WORD SEGMENTATION IN GERMAN-LEARNING INFANTS AND GERMAN-SPEAKING ADULTS: PROSODIC AND STATISTICAL CUES

by

Mireia Marimon Tarter

Submitted to the

Faculty of Human Sciences of the

University of Potsdam

2019

PredictAble

The research reported in this dissertation has been supported by the European Union's Horizon 2020 Research and Innovation Program under the grant agreement No 641858, PredictAble Project.

Under the supervision of

Prof. Barbara Höhle, University of Potsdam

Dr. Thierry Nazzi, Université Paris Descartes

Date of submission: 29th April 2019

Published online at the
Institutional Repository of the University of Potsdam:
https://doi.org/10.25932/publishup-43740
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-437400

Declaration of authorship

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works by any other author, in any form, is properly acknowledged at their point of use.

Acknowledgements

First and foremost I would like to thank my supervisor Professor Dr. Höhle for giving me the opportunity to work in such a great project. Thank you for your support, guidance and commitment. I am truly grateful for that. Secondly, I would like to thank all the PredictAble professors, who gathered together an amazing group of scientists and made the whole project possible. Thank you for all the valuable feedback during these years in the summer schools and in the skype meetings. Special thanks to Núria Sebastián-Gallés for introducing me into infant research. Thanks to Prof. Dr. René Kager for his helpful and thorough revision of this work.

Golm can be a grey far-away place, but the people in the group made it a second home and I never felt alone. Thanks to all the members of the Psycholinguistics Group and special thanks to Tom, Natalie, Antonia and Alan for all their knowledge, feedback and help with my experiments and having so much patience with me. There is a bit of each of you in this thesis. Thank you to Anikó, Annika and Anja for making German bureaucracy easier. Thank you to the Babylab Team for their assistance and to all the participants and families who came to the lab. To the team in Paris: thank you for your constant support during my stays in Paris. Special thanks to Dr. Thierry Nazzi for his guidance, thousands of e-mails and discussions.

To the PredictAble people: you were a wonderful support group. It was great to share all the summer schools together. Thank you for staying together in this journey and I wish you a lot of luck in your personal and professional life projects. Special thanks to Astrid, I couldn't think of a better colleague and friend next to me during this journey. Thank you for your invaluable friendship.

Thank you to my family for your love and support, and always putting a smile on my face. Thank you, Marc, for your brilliant thoughts on statistics. Thank you to Lucas for his endless support on a daily basis, for making me laugh and bearing with me through my stressful moments. Thank you for always being there for me and making every day better than the day before.

Contents

List of figures

List of tables

1. General introduction 1

2. Theoretical Background: Prosody and statistics as cues for word segmentation 5

2.1. Prosody 5

2.1.1 Introduction 5

2.1.2 Universal trochaic bias 7

2.1.3 Early prosodic segmentation 9

2.1.4 Prosody and later language development 12

2.1.5 Specific methodological considerations in speech perception experiments 14

2.1.6 Models for early prosodic segmentation 17

2.1.7 Summary 19

2.2 Statistical learning 20

2.2.1 Introduction 20

2.2.2 Early segmentation with statistical learning 22

2.2.3 Statistical learning and later language development 24

2.2.4 Specific methodological considerations in statistical learning experiments 26

2.2.5 Models for early statistical learning 29

2.2.6 Summary 31

2.3 The weighting of cues for word segmentation 32

2.3.1 Introduction 32

2.3.2 When cues conflict 33

2.3.3 Summary 36

2.4 Summary and research questions 38

3. Experiment 1: The Headturn Preference Procedure 41

3.1 Experiment 1: Test-reliability test of the Headturn Preference Procedure (HPP) 41
3.1.1 Introduction 41
3.1.2 Participants 42
3.1.3 Stimuli 42
3.1.4 Procedure 43
3.1.5 Results 44
3.1.5.1 Rhythmic preferences 44
3.1.5.2 Test-retest reliability 46
3.1.5.3 Correlations to later language development 47
3.1.6 Discussion 48

4. Experiment 2: Weighting of prosodic and statistical cues in German adults 53
4.1 Experiment 2a: German adult segmentation 53
4.1.1 Introduction 53
4.1.2 Participants 55
4.1.3 Stimuli 55
4.1.4 Procedure 59
4.1.5 Results 60
4.1.5.1 Behavioral results 61
4.1.5.2 Pupillometry results 63
4.2 Experiment 2b: German adult segmentation with amplitude ramp 68
4.1.1 Participants 68
4.1.2 Stimuli 69
4.1.3 Procedure 69
4.1.4 Results 70
4.3 General discussion 72

5. Experiment 3: Weighting of prosodic and statistical cues in 9-month-old German infants 77
5.1 Experiment 3a: Word segmentation at 9 months 77

5.1.1 Introduction 77
5.1.2 Participants 81
5.1.3 Stimuli 81
5.1.4 Procedure 81
5.1.5 Results 83
5.2 Experiment 3b: Word segmentation at 9 months with pupillometry 86
5.2.1 Introduction 86
5.2.3 Participants 87
5.2.3 Stimuli 88
5.2.4 Procedure 88
5.2.5 Results 89
5.3 Experiment 3c: Word segmentation at 9 months without familiarization 92
5.3.1 Participants 92
5.3.2 Stimuli and procedure 92
5.3.3 Results 93
5.4 Experiment 3d: Word segmentation at 9 months with double familiarization 94
5.4.2 Participants 94
5.4.3 Stimuli and procedure 94
5.4.5 Results 95
5.5. General discussion 96

6. Experiment 4: Weighting of prosodic and statistical cues in 6-month-old German infants 101
6.1 Experiment 4a: Word segmentation at 6 months 101
6.1.1 Participants 101
6.1.2 Stimuli and Procedure 101
6.1.3 Results 102
6.2 Experiment 4b: Word segmentation at 6 months without familiarization 105
6.1.1 Participants 105
6.1.2 Stimuli and Procedure 105

6.1.3 Results 105
6.3 General discussion 106

7. Conclusion 113

8. References 119

List of Figures

Figure 1. Experiment 1: Mean looking times in the three sessions 45

Figure 2. Experiment 1: Negative correlation between age and iambic looking times in the second session 46

Figure 3. Experiment 1: Correlation between difference scores and ELFRA-1 total scores 47

Figure 4. Experiment 2a: Percentages of 'yes' responses in each condition 61

Figure 5. Experiment 2a: Time course of the pupil size changes in the different conditions 63

Figure 6. Experiment 2a: Time course of the pupil size changes in the different conditions by button 66

Figure 7. Experiment 2b: Percentages of 'yes' responses in each condition when the string had a ramp 71

Figure 8. Experiment 3a: Mean looking times at test for the 9-month-olds 84

Figure 9. Experiment 3b: Time course of the pupil size in the different conditions 90

Figure 10. Experiment 3c: Mean looking times at test for the 9-month-olds without familiarization 93

Figure 11. Experiment 3d: Mean looking times at test for the 9-month-olds with double familiarization time 95

Figure 12. Experiment 4a: Mean looking times at test for the 6-month-olds 102

Figure 13. Experiment 4b: Mean looking times at test for the 6-month-olds without familiarization 106

List of Tables

Table 1. Experiment 1: Correlations between the looking times across the three sessions 46

Table 2. Experiment 1: Correlations between the raw scores and the total test scores in ELFRA-1 47

Table 3. Experiment 1: Correlations between the raw scores and the total test scores in ELFRA-2 48

Table 4. Experiment 2a: Acoustic properties of the stressed and unstressed syllables in the familiarization string 57

Table 5. Experiment 2a: Acoustic properties of the test trial syllables 57

Table 6. Experiment 2a: Possible segmentations of the string 58

Table 7. Experiment 2a: Model against chance 62

Table 8. Experiment 2a: Maximal Model 62

Table 9. Experiment 2a: Maximal Model for the 2000-2500 ms window 65

Table 10. Experiment 2a: Maximal Model for the 2500-3000 ms window 65

Table 11. Experiment 2a: Maximal Model for the 2000-2500 ms window (non-word condition as baseline) 67

Table 12. Experiment 2a: Maximal Model for the 2500-3000 ms window (non-word condition as baseline) 67

Table 13. Experiment 2b: Model from Experiment 2a and 2b with the ramp effect 71

Table 14. Experiment 3a: Correlations between the looking times and the test scores in ELFRA-1 (9-month-olds) 85

Table 15. Experiment 3a: Correlations between the looking times and the test scores in FRAKIS (9-month-olds) 85

Table 16. Experiment 3b: Maximal Model from 500 - 2500 window (9-month-olds) 91

Table 17. Experiment 4a: Correlations between the looking times and the test scores in ELFRA-1 (6-month-olds) 103

Table 18. Experiment 4a: Correlations between the looking times and the test scores in FRAKIS (6-month-olds) 104

1. GENERAL INTRODUCTION

Any adult who has attempted to learn a further language might know how challenging and confusing it can be, but infants master this process in a very short time. A main aim in language acquisition research has been to understand how infants are capable of achieving such a complex task and how this learning process changes at different stages of development. A wide range of studies have aimed to describe the emergence of word learning as well as the mechanisms that are involved in this process.

One of the first steps in word learning is word segmentation. To learn new words, infants start to encode word forms by segmenting continuous speech and distinguishing word boundaries. Since infants seem to start extracting words from fluent speech between 6 and 7.5 months of age (e.g., Jusczyk & Aslin, 1995), word segmentation has generated a great deal of empirical and theoretical interest. There is evidence that infants use at least two mechanisms to segment words forms from fluent speech: (1) using prosodic information (e.g., Jusczyk, Cutler & Redanz, 1993; Weber, Hahne, Friedrich & Friederici, 2005; Junge, Kooijman, Hagoort & Cutler, 2012) and (2) using statistical information (e.g., Saffran, Aslin & Newport, 1996b; Aslin, Saffran & Newport, 1998). However, how these two mechanisms interact and whether they change during development is still not fully understood.

Studies that support a prosodic bootstrapping account (Gleitman & Wanner, 1982; Mazuka, 1996; Morgan & Demuth, 1987) suggest that infants might begin using prosodic cues to start segmenting words from fluent speech and for further language development (Christophe, Gout, Peperkamp & Morgan, 2003; Soderstrom, Seidl, Kemler Nelson & Jusczyk, 2003). However, other studies have shown that infants rely more strongly on statistical information and the authors argue that statistical cues support the first steps in word segmentation (e.g., Saffran et al., 1996b; Aslin et al., 1998, Thiessen & Saffran, 2003). Thus, to distinguish and clarify which cues infants use for word

segmentation and whether these change with language experience, further investigation is necessary.

A handful of studies have also focused on the relation between word segmentation skills and later language development (e.g., Newman, Ratner, Jusczyk, Jusczyk & Dow, 2006; Singh, Reznick & Xuehua, 2012; Mainela-Arnold & Evans, 2014). Recent research has suggested that both statistical learning skills and the use of prosodic cues for word segmentation are related to later language development. For instance, a link has been reported between the preference for the dominant word stress pattern in German at the age of 4 months and the linguistic performance of children at the age of 5 years (Höhle, Pauen, Hesse & Weissenborn, 2014). Regarding statistical learning, it has been found that the sensitivity to regularities in the speech is associated with later language outcomes (e.g., Arciuli & Simpson, 2012; Kidd & Arciuli, 2016). However, the reliability of the methods used in the field has been little investigated and the specific links to language development have been previously addressed only to a limited extent. This thesis documents key contributions to the reliability of early indicators of a potential risk in language development as well as to the relation between early word segmentation and later language outcomes.

The main aim of the present work is to understand in what way different cues to word segmentation are exploited by infants when learning the language in their environment, as well as to explore whether the ability of exploiting cues to word segmentation is related to developing segmentation skills. Thus, the research questions of the present thesis are as follows:

a) Do German infants and adults rely more strongly on prosodic or statistical information when segmenting words from fluent speech? How do infants integrate knowledge of the cues at particular points in their development? Are these two cues used differently depending on age and/or language experience?

b) Can the use and weighting of these cues in a word segmentation task predict later language skills? Is the Headturn Preference Procedure reliable enough to obtain predictive measures for later language development?

The experiments presented in this thesis aim to find answers to the above questions. Experiment 1 was pursued to determine the reliability of the method used in most of the experiments in the present thesis (the Headturn Preference Procedure), as well as to examine correlations and individual differences between infants' performance and later language outcomes. Experiments 2a and 2b were designed to investigate how German-speaking adults weight statistical and prosodic information for word segmentation. In Experiment 2a we familiarized adults with a string in which statistical and prosodic information indicated different word boundaries. We obtained both behavioral and pupillometry responses. Experiment 2b was a control experiment that added an amplitude ramp in the string. Then, we continued to explore the weighting of these two cues in infancy.

Experiments 3a–d were conducted to understand in what way different cues to word segmentation are exploited by 9-month-old German-learning infants. From the literature, we expected 9-month-olds to rely more strongly on prosodic information. Experiment 3a tested 9-month-olds in a word segmentation task in which infants were familiarized with the same string used in the adult experiments. Because of the null results obtained, we conducted three more experiments before drawing any strong conclusions. We tested infants on the same task but without familiarization (Experiment 3c) and with double time of exposure (Experiment 3d). In Experiment 3b, we obtained pupillometry responses.

In Experiments 4a and 4b we explore whether there are changes during development. We tested 6-month-olds on the same task as in Experiment 3a. Since a goal of the present dissertation is to explore the link between the weighting of cues for word segmentation and later language

development, in Experiments 3a and 4a we conducted follow-up questionnaires with the infants and obtained language outcomes at later stages of development.

To answer the research questions specified above, the present thesis is organized as follows. In the first part, I will lay out the theoretical background of the two main cues to word segmentation (statistical and prosodic information) and present the method that will repeatedly be used throughout the thesis (Chapter 2). In the second part, I will present experimental data on the reliability of this method related to later language outcomes in Chapter 3. In Chapter 4 I will present studies that examined cue weighting in German-speaking adults, and in Chapters 5 and 6 I will present infant data on cue weighting. Finally, I will discuss the data presented in this thesis and gathered in the experiments in light of the current theories on word segmentation and will suggest potential further research.

2. THEORETICAL BACKGROUND: PROSODY AND STATISTICS AS CUES FOR WORD SEGMENTATION

2.1 Prosody

2.1.1 Introduction

Speech to infants –like speech to adults– is a continuous sound stream in which word boundaries are not marked by a set of unique and reliable phonetic cues (Cole & Jakimik, 1980). Therefore, an important requirement for lexical development is to divide this fluent speech stream into units that correspond to the words of the language. Previous studies have shown that infants rely on several types of information to solve this segmentation problem. One of the main sources is prosodic information. Prosody is the rhythm and melody of speech, including patterns of tone, stress, and intonation. Prosodic features are supra-segmental properties of speech units and typically associated to the syllable (or mora) unit or to higher levels in the prosodic hierarchy (Beckman & Pierrehumbert, 1986; Nespor & Vogel, 1986; Pierrehumbert & Beckman, 1988). In speech, prosodic variation is signaled by pitch (fundamental frequency), duration, intensity (loudness), and segmental reduction (Bolinger, 1989; Lehiste, 1970; Shattuck-Hufnagel & Turk, 1996).

Newborns are equipped with highly efficient capacities to process specific information from speech. From birth onwards they are already sensitive to prosodic information, which is assumed to be crucial for starting to segment and recognize words from fluent speech (Moon, Cooper & Fifer, 1993; Nazzi, Bertoncini & Mehler, 1998; Christophe, Mehler & Sebastián-Gallés, 2001).

One of the topics that have been studied in infant speech perception research is the sensitivity to the language rhythm. Research has shown that from birth infants can discriminate between two rhythmically distant languages even if neither of them is their native language (Moon et al., 1993; Ramus, Nespor & Mehler, 1999; Nazzi et al., 1998), probably only relying on prosodic cues

(Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Dehaene-Lambertz & Houston, 1998; Ramus, Hauser, Miller, Morris & Mehler, 2000). At around 5 months of age, infants can discriminate their native language from another rhythmically similar language, but they cannot distinguish two non-native rhythmically similar languages (e.g., Bosch & Sebastián-Gallés, 1997, 2001; Nazzi, Jusczyk & Johnson, 2000; Ramus et al., 2000; see Jusczyk, 1997, for a review). Regarding stress pattern recognition, there is evidence that from birth on English-learning infants are sensitive to the acoustic correlates of stress location and are able to distinguish between stressed and unstressed syllables (Spring & Dale, 1977; Jusczyk & Thompson, 1978). This sensitivity to stress has also been attested in Italian newborns (Sansavini, Bertoncini & Giovanelli, 1997), for 6-month-old Spanish-learning infants (Skoruppa, Cristia, Peperkamp & Seidl, 2013), and for French-learning infants from 4 to 10 months of age (Friederici, Friedrich & Christophe, 2007; Höhle, Bijeljac-Babic, Herold, Weissenborn & Nazzi, 2009; Skoruppa, Pons, Christophe, Bosch, Dupoux, Sebastián-Gallés, Alves Limissuri & Peperkamp, 2009; Skoruppa et al., 2013).

In the case of German-learning infants, the discrimination between strong and weak syllables has been found very early (Weber et al., 2005; Höhle et al., 2009). Friederici, Friedrich & Weber (2002) tested German-learning infants as young as 2 months in their ability to discriminate CV-syllables varying in vowel duration (long vs. short) in an oddball paradigm. Infants were able to discriminate a long syllable among short syllables, but not vice versa. The authors explain this by a greater perceptual saliency of a longer element in the context of shorter elements than of a shorter element in the context of longer elements. Weber et al. (2005) tested 4- and 5-month-olds on a mismatch paradigm with trochaic and iambic items and 4-month-olds did not show discrimination for either the trochaic or the iambic deviant item. In contrast, 5-month-olds discriminated the items when the deviant one was trochaic (and the standard iambic), but not vice versa, which still suggests a discrimination between the two patterns. Höhle et al. (2009) tested 4- and 6-month-old German-

learning infants and 6-month-old French-learning infants in the Headturn Preference procedure (HPP), where infants were presented with trochaic and iambic disyllabic CVCV strings. German-learning infants showed a listening preference for the trochaic pattern at 6 months of age, but not at 4 months.

2.1.2 Universal trochaic bias

The ability to learn the typical rhythm and stress pattern of the native language has been claimed to play a central role in the earliest steps of speech segmentation. It has been documented that stressed syllables are treated as the beginnings of words by infants learning languages with stress-based rhythmic properties such as English, German, or Dutch (English: Jusczyk et al., 1993; Turk, Jusczyk & Gerken, 1995; German: Höhle, 2002; Dutch: Kuijpers, Coolen, Houston & Cutler, 1998; Junge et al., 2012; Kooijman, Hagoort & Cutler, 2009). In these languages, evidence for a metrical-based segmentation was obtained, in the sense that stressed syllables are preferred as the beginnings of words with following unstressed syllables being attached to them. For example, 9-month-old English-learning infants were presented with weak-strong-weak non-words which contained a 250 ms silent pause either before or after the strong syllable. Infants preferred configurations in which the pause was before the strong syllable, which was therefore the onset of the word (Echols, Crowhurst & Childers, 1997). This is the so-called "trochaic bias." Such a strategy is in accordance with the dominant trochaic foot structure in these languages, which is the basis of the dominant trochaic stress pattern in disyllabic words.

However, it remains an open question when the preference for the native stress pattern appears, whether it is present across languages, and whether the timing of the onset of the preference differs between languages. Evidence for the trochaic bias has been reported for 9-month-old English-learning infants, who listened significantly longer to disyllabic words with a trochaic stress pattern than to iambic words (Jusczyk et al., 1993). However, 6-months-olds did not show any preference,

suggesting that it is between these ages that English infants find out the predominant stress pattern of their native language. Likewise, in a more recent study, 9-month-old Hebrew-learning infants (an iambic language) preferred to listen to lists of iambic words rather than trochaic words, but did not generalize this preference to English, a foreign language with a trochaic stress pattern (Segal & Kison-Rabin, 2016). In Spanish, a language with a greater proportion of trochees than iambs (60% vs. 40%), 9-month-old infants failed to reveal a clear trochaic bias. However, a trochaic preference could be elicited when items contained an initial heavy syllable (CVC.CV) (Pons & Bosch, 2010).

In German, it has been suggested that the trochaic bias is already present at 4 months of age. Friederici et al. (2007) tested 4-month-old French- and German-learning infants in an ERP experiment and showed that language experience affected infants' brain responses. Each infant language group displayed a processing advantage for the rhythmic structure of their native language. The first clear evidence of a preference for trochaic over iambic words by German infants has been provided for 6-month-old infants (Höhle et al., 2009). In their study, Höhle et al. tested 4- and 6-month-old German-learning infants and 6-month-old French-learning infants in the Headturn Preference procedure (HPP), where infants were presented with trochaic and iambic disyllabic CVCV strings. German-learning infants showed a listening preference for the trochaic pattern at 6 months of age, but not at 4 months, suggesting an emergence of this preference between the two ages (for further evidence of this trochaic bias in German infants, see Herold, Höhle, Walch, Weber & Obladen, 2008). In contrast, French-learning 6-month-old infants showed no preference for either the trochaic or the iambic pattern and their ability to discriminate the stress pattern was found to decrease between the ages of 6 and 10 months (Skoruppa et al., 2009; Bijeljac-Babic, Höhle & Nazzi, 2012). Presumably, this is because French is a language that does not use lexically contrastive stress. In line with these results, 6-month-old German-French bilinguals also showed a preference for trochaic sequences and did not show a delay compared to their monolingual peers

(Bijeljac-Babic, Höhle & Nazzi, 2016). German infants' preference for the trochaic pattern is best explained as an effect of regular exposure to a language in which this pattern dominates and can thus be considered a result of building first representations about prosodic properties of the ambient language. These findings with German infants motivated Experiment 4 (Chapter 6), where we further explore this bias and examine the possible link with later language outcomes.

In sum, it has been suggested that the trochaic bias is innate and universal (Allen & Hawkins, 1980), but other authors advocate that it develops from the linguistic input (Juszcyk et al., 1993). Whereas there is no sufficient linguistic data to conclusively decide between the two options, there is evidence that seems to favour the development of the trochaic bias as a result of language experience (Höhle et al., 2009; Bijeljac-Babic et al., 2012; Skoruppa et al., 2009). Independently of the nature of the trochaic bias, there are consistent findings in the literature that this sensitivity can be observed very early in the prosodic domain – for German learning infants this familiarity emerges between 4 and 6 months of age – and that it provides infants an additional cue to word boundaries and facilitates subsequent word recognition and word segmentation.

2.1.3 Early prosodic word segmentation

Demonstrations of segmentation of word forms from fluent speech have been found in infants as young as 5.5 months and throughout early infancy (Jusczyk, Houston & Newsome, 1999; Johnson & Jusczyk, 2001; Johnson, Seidl & Tyler, 2014), and it is thought that it develops around 6 to 8 months of age in several languages, including Canadian English, German, Dutch, Catalan, Spanish, and European and Canadian French (for a recent review, see Goyet, Millotte, Christophe & Nazzi, 2016; for a recent meta-analysis, see Bergmann & Cristia, 2016). Infants' segmentation abilities seem to be influenced by several cues to word boundaries such as phonotactic regularities (Mattys, Jusczyk, Luce & Morgan, 1999), lexical constraints (Jusczyk, Cutler & Norris, 2003), rhythmic structure (Houston, Jusczyk, Kuijpers, Coolen & Cutler, 2000; Nazzi, Iakimova, Bertoncini,

Frédonie & Alcantara, 2006), and prosodic cues (Jusczyk et al., 1993, 1999; Johnson & Jusczyk, 2001). The prosodic cues that support infants' speech segmentation are commonly assumed to be language specific and consist of rhythmic grouping cues (e.g., Abboub et al., 2016), intonational contours (e.g., Shukla, White & Aslin, 2011), and stress patterns (e.g., Echols et al., 1997; Jusczyk et al., 1999).

Stress is mainly characterized by three acoustic features: fundamental frequency (F0), timing (duration), and intensity. Languages vary on how these acoustic cues are weighted and even within a language these same cues are used to serve different kinds of linguistic (e.g. indicating suprasegmental, pragmatic relationships) and non-linguistic functions (e.g. physiological functions related to breathing), including aiding in word segmentation (Cutler & Mehler, 1993; Jusczyk et al., 1999). Thus, the acquisition of the prosodic properties and the attunement to the predominant native stress pattern are likely to have strong impact on word segmentation. It has been demonstrated that English-learning infants can segment disyllabic words with a trochaic stress pattern from the age of 7 months on (Morgan & Saffran, 1995; Jusczyk et al., 1999; Houston et al., 2000) and Dutch-learning infants from the age of 10 months (Kuijpers et al., 1998; Kooijman Hagoort et al., 2009). German infants are able to segment disyllabic words from fluent speech at roughly the same age as their English and Dutch peers (Jusczyk et al., 1999; Houston et al., 2000; Höhle & Weissenborn, 2003; Bartels, Darcy & Höhle, 2009). However, the segmentation of iambic units seems to develop only at a later age for Dutch- and English-learning infants. Evidence from Jusczyk et al. (1999) showed that English-learning infants failed to segment iambic words from continuous passages at 7.5 months of age. It is not until 10 months of age that English-learning infants can segment iambs in a similar manner to trochees (Jusczyk et al., 1999; Gerken & Aslin, 2005). However, 8-month-old Canadian-French-learning infants showed segmentation of iambs more readily from French passages (Polka & Sundara, 2012).

This prosodic knowledge might also be used to segment words from non-native but rhythmically similar languages. Höhle, Giesecke & Jusczyk (2001) tested English- and German-learning infants on bisyllabic trochaic word segmentation at 9 months of age. Both groups were successful in extracting trochaic German words from German text passages. English-learning infants were able to extract words from Italian (Pelucchi, Hay & Saffran, 2009) and Dutch (Houston et al., 2000), both trochaic languages. In contrast, 8-month-old Canadian-French infants (a predominantly phrase-final stress language) could not segment English words from Canadian-English passages and Canadian-English infants could not segment words from Canadian-French passages (Polka & Sundara, 2012). Both of these findings support the relevance of prosodic information for segmentation, since cross-linguistic segmentation does not work when the languages are rhythmically dissimilar.

Furthermore, infants' prosody-based segmentation strategies are not limited to disyllabic words. Segmentation of the input into higher units such as phrases and clauses also seems to be initially affected by prosodic cues (Hirsh-Pasek et al., 1987; Jusczyk, Kemler Nelson, Hirsh-Pasek, Kennedy, Woodward & Piwoz, 1992; Nazzi et al., 2000; Seidl, 2007; Soderstrom et al., 2003, Wellmann, Holzgrefe, Truckenbrodt, Wartenburger & Höhle, 2012; Männel, Schipke & Friederici, 2013) by the syntactic categorization of words (Shi, Werker & Morgan, 1999), and by the detection of word order regularities (Nespor, Guasti & Christophe, 1996).

Altogether, it is likely that infants are sensitive to prosodic information from fluent speech and that this information plays a key role in word segmentation and in further processes related to the acquisition of at least some languages. In the case of German, prosody has been shown to be quite prominent and a key in word segmentation processes (Höhle et al., 2001; Höhle & Weissenborn, 2003) and syntax processing (Soderstrom et al., 2003; 2005; Wellmann et al., 2012). Hence, German adults and infants are an interesting population for further exploring prosody's importance in development and in relation to other word segmentation cues. This motivated a large part of the

current work, more specifically Experiments 2a and 2b with German adults and Experiments 3 and 4 with 6- and 9-month-old German-learning infants, which are described in Chapters 5 and 6.

2.1.4 Prosody and later language development

One of the goals of the present dissertation is to explore how well infants are able to take advantage of prosodic cues and whether this ability has an impact on later language development and could thus be an early predictor of language disorders. It is possible that an infant who has trouble segmenting words from fluent speech may not have many word forms stored in memory to associate with referents in the real world and, as a consequence, her vocabulary might be reduced. It is likely that the discovering of the prosodic properties of the ambient language and the development of a trochaic bias could be an indicator of infants' language abilities. In this section I first review the relation found between segmentation abilities and later language development. After that, I describe the studies that have suggested a specific link between speech perception measures at an early age and later language outcomes.

Although more longitudinal research is needed, evidence for relations of early speech perception in the first year of life and later language achievements has been obtained with several methods and in several linguistic areas (for a recent review and meta-analysis, see Cristia, Seidl, Junge, Soderstrom & Hagoort, 2014). Such predictive relations have been documented for the perceptual attunement to the native sound system (Conboy, Rivera-Gaxiola, Klarman, Aksoylu & Kuhl, 2005; Kuhl, Conboy, Padden, Nelson & Pruitt, 2005; Rivera-Gaxiola, Klarman, Garcia-Sierra & Kuhl, 2005; Conboy, Rivera-Gaxiola, Silva-Pereyra & Kuhl, 2008; Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola & Nelson, 2008; Tsao, Liu & Kuhl, 2004), word segmentation skills (Newman et al., 2006; Junge et al., 2012; Singh et al., 2012), and the processing of prosodically relevant information (Weber et al., 2005; Friedrich, Weber & Friederici, 2004; Höhle et al., 2014; Seidl & Cristia, 2012).

In perceptual attunement to the native sound system, Kuhl et al. (2005) reported a negative relation between non-native sound perception at 7.5 months of age and later language abilities at 14, 18, 24, and 30 months of age. Their results suggested relations in both directions: better native language discrimination predicted accelerated later language abilities and better non-native language discrimination predicted reduced later language abilities.

Concerning word segmentation skills, Newman et al. (2006) tested infants between 7.5 and 12 months of age on speech perception tasks and assessed their linguistic and cognitive skills at 4-6 years of age. They found that children who had been able to segment words from fluent speech had higher language measures outcomes, but not a higher general IQ. These findings were later replicated by Newman, Rowe and Ratner (2016) and Singh et al. (2012), who also observed a relation between word segmentation abilities at 7.5 months of age and vocabulary size at 2 years. These results are in accordance with a study by Junge and colleagues (2012), who reported a correlation between the latency of a negative ERP component (evoked by the presentation of an isolated word previously presented in a sentence) at 7 months of age and the CDI receptive vocabulary scores at 12 and 24 months in Dutch-learning infants. In addition, segmentation abilities have been found to be delayed in children with cognitive and/or linguistic deficits (Nazzi, Paterson & Karmiloff-Smith, 2003).

Regarding the processing of prosodically relevant information, Weber et al. (2005) investigated the prosodic abilities of 5-month-old German-learning infants at risk for Specific Language Impairment (SLI)[1] using ERPs in a passive oddball design. Infants at risk showed a significantly reduced amplitude of the discrimination response (Mismatch Negativity), suggesting that a reduced stress pattern discrimination at 5 months of age could be a marker of risk for later language impairment.

[1] Infants were considered to be at risk when they produced fewer than 6 out of 164 language-related items at the age of 12 months (ELFRA-1) and when they produced fewer than 50 out of 260 words at the age of 24 months (ELFRA-2).

Also using ERPs and a similar design, Friedrich et al. (2004) showed that 2-month-old infants at risk for SLI already have a delayed mismatch response to CV-syllables differing in vowel duration compared to typically developing (TD) infants. With behavioral methods, Höhle et al. (2014) reported a relation between prosodic perception at 4 months and later language outcomes at the age of 5 years. German-learning infants were tested in the HPP procedure on a discrimination task between iambic and trochaic sequences. In addition, results show that children with a family risk for SLI might have language problems that may be evident at a very young age. Seidl and Cristia (2012) also observed a link between early processing of prosodic information and later language outcomes in English-learning infants. They obtained infants' response to two versions of the same sequence of words, only one of which had been uttered as a well-formed prosodic unit. Results showed that sensitivity to prosody at 6 months of age predicts vocabulary size at 24 months of age.

Taking into account all the evidence above, we can establish that there is a relation between prosodic word segmentation and later language development. In addition, as the prosodic bootstrapping account predicts, a weakness in exploiting prosodic information can have a broad range of consequences in language development and can play a key role in the early prediction of language disorders. However, more longitudinal studies are needed to determine the strength of this link and which specific language areas can be an early predictor of language development. In the present thesis we contribute to this research with three longitudinal studies at the ages of 6 and 9 months of age (Experiments 1, 3, and 4).

2.1.5 Specific methodological considerations in speech perception experiments

Speech perception abilities have been a focus of research over the past five decades and different methodologies have been used to discover how infants tune their speech perception abilities to language-specific properties of the speech they hear. One of the most commonly used methods in speech perception experiments with infants is the HPP, which is the primary method used in the

experiments of the present thesis (see Section 3.1.4 for more information about the experimental setting). The HPP was chosen because it can be used to investigate infants' ability to memorize and recognize speech. Therefore, it is a well approved method used for segmentation studies in young infants and it has also been successfully used to explore the relation to later language outcomes (e.g. Höhle et al., 2014). In addition, studies that posed similar research questions as ours like Höhle et al. (2009) and Thiessen and Saffran (2003) also used this method.

The outcome measure of the HPP is the attention to different auditory stimuli measured as the time that the child looks to a visual attractor during the presentation of speech (listening time). Infants' preferences can thus be expressed as an enhanced attention to one type of stimulus over another type of stimulus. A common stimulus contrast found in the literature is the novel vs. familiar contrast. Stimuli might draw more attention if they sound familiar or unfamiliar to the infants. The direction of the preference (i.e., whether the infants show longer listening times to the familiar or to the novel stimuli) is determined by several factors like stimulus complexity, duration of exposure during the experiment, or the infant's individual developmental status (Hunter, Ames & Koopman, 1983; Roder, Bushnell & Sasseville, 2000). Therefore, the same stimuli can elicit familiarity and novelty effects in infants of different ages (Colombo & Bundy, 1983) or even in infants of the same age depending on their lexical development (DePaolis, Portnoy & Vihman, 2016). In most experiments using this technique, the direction of the preference is not relevant as both effects reveal the ability to discriminate between stimuli from the experimental conditions. However, it matters when comparing the weighting of different input cues, which is relevant to some experiments in the present thesis (see Chapters 5 and 6).

Evidence for relations of early speech perception in the first year of life and later language achievements has been obtained with several methods and in several linguistic areas (see Section 2.1.4). These studies used behavioral assessments of infants' speech perception like the Conditioned

Headturn Paradigm (e.g., Gout, Christophe & Morgan, 2004) or the HPP (e.g., Höhle et al., 2014), or obtained neurocognitive measures via ERPs (e.g., Weber et al., 2005; Junge et al., 2012). These methods are well approved and widely used in infant research. However, experiments applying them are typically designed as group studies and are not suited to testing individual differences. Thus, data analyses from these experiments are restricted to the group level since the inter- as well as the intra-individual variation in infants' data is typically rather large. Evaluation of the reliability of these measures is required in order to use these measures as a tool to predict later language achievements on an individual level.

One measure of the reliability of a test instrument is the test-retest reliability, which characterizes the intra-individual stability of performance across several measurements. So far, only a handful of published studies have assessed the reliability of infant speech perception measurements. A first one comes from Houston, Horn, Qi, Ting and Gao (2007), who examined test-retest reliability for a Visual Habituation Procedure[2] testing 9-month-olds' discrimination of two phonemically very different pseudo-words after habituation with one of them. Ten participants were tested twice with the same stimuli on two separate days (1 to 3 days apart). A significant, moderately strong correlation between the performances across the two test sessions was observed.

In another study by Cardillo (2010) 20 English-learning infants were retested on a similar task with the purpose of assessing changes in sound perception over time. Infants were tested at 7 and 11 months of age on the vowel contrast /u-y/, using a variant of the Conditioned Head Turn paradigm. Her results revealed a nonsignificant correlation between the two discrimination measures. Cristia, Seidl, Singh and Houston (2016) carried out a meta-analysis addressing the test-retest reliability of

[2] In this procedure, infants were first habituated to audiovisual repetitions of a non-word (*seepug*) before entering the test phase. The test phase consisted of old (*seepug*) and novel trials (*boodup*).

the Central Fixation Paradigm[3] and the HPP with data collected in 13 different experiments in 3 different laboratories. Infants of different ages (5 to 12 months) and in different sample sizes (10 to 89 participants) were tested in sound and word form discrimination tasks after a previous familiarization or habituation phase. All infants were tested twice with the same procedure at intervals ranging between 1 and 7 days. Rather weak evidence for test-retest reliability was obtained, with only 5 of the 13 experiments showing correlations of infants' performance across the test sessions. No systematic pattern of the occurrence of these correlations could be detected, concerning neither their direction, nor the linguistic level tested, nor the experimental method. However, given the heterogeneity of the studies considered in this analysis, this result is not surprising but underlines the necessity for more targeted research on this issue.

The present thesis addresses this issue and investigates the test-retest reliability of the HPP, which has not been meticulously addressed so far (see Experiment 1). The goal of the experiment was two-fold: we address the reliability of the HPP procedure and we explore the relation between speech perception measures obtained with this procedure and later language outcomes.

2.1.6 Models for early prosodic segmentation

The literature about early prosodic word segmentation presented previously in this chapter is a result of a large theoretical framework that has proposed prosody as the central learning mechanism to access early language acquisition. Several theories have tried to explain how children effortlessly acquire the words and the structure of their native language even though speech provides no direct information about underlying structure. The so-called Bootstrapping Accounts in language acquisition have tried to identify the mechanisms that help children to learn properties of their

[3] In this procedure infants sit on a caregiver's lap and look at neutral visual stimuli presented on a screen. At the same time, the infant hears auditory stimuli from speakers placed near the screen. Throughout all phases, the infant's attention is directed to the screen before each trial using an attention-getting visual stimulus. When the infant fixates the screen, the trial begins and continues until the infant looks away for longer than 1 s or the maximum trial duration is reached.

native language with the information that the child can access from her input, namely how the child might find the linguistically relevant units that can serve as constraints for further learning. Although several kinds of bootstrapping mechanisms have been proposed (distributional bootstrapping, syntactic bootstrapping, typological bootstrapping, etc.; for a review, see Höhle, 2009), I will focus on prosodic bootstrapping theories.

Prosodic bootstrapping theories (Morgan, 1986; Gleitman, 1990; Jusczyk et al., 1992; Morgan & Demuth, 1987; Weissenborn & Höhle, 2001; for a recent overview see De Carvalho, Dautriche, Millote & Christophe, 2018) assume that the ability to process prosodic information in the speech signal (stress, rhythm, intonation) might be crucial to detecting lexical and syntactic boundaries, and might consequently affect language development in these domains (Morgan & Demuth, 1987; Christophe, Nespor, Guasti & Van Ooyen, 2003; Soderstrom et al., 2003). Therefore, unlike in statistical learning models (see Section 2.2.5), models for early prosodic segmentation maintain that infants must first learn about the native prosody before they can begin to compute distributional regularities.

Evidence supporting prosodic bootstrapping theories comes from different levels. At the lexical level, prosodic cues such as word length or stress pattern seem to help infants to differentiate between word classes (lexical words or functional words) and subclasses like verbs and nouns (Kelly & Bock, 1988; Christophe, Guasti, Nespor, Dupoux & Ooyen, 1997; Shi et al., 1999). At the syntactic level, it is suggested that infants make use of prosodic information to acquire the basic word order rules (Guasti, Nespor, Christophe & van Ooyen, 2001; Nespor, Mehler, Shukla, Peña & Gervain, 2007) and to identify syntactically relevant units (Morgan & Demuth, 1987; Jusczyk 1997). Further evidence comes from the fact that early measures of processing prosodic information are linked to later language skills (see Section 2.1.4).

2.1.7 Summary

In this section I have reviewed evidence regarding the early processing and acquisition of prosodic properties with an emphasis on the lexical level. Overall, the literature suggests that infants come equipped with sensitivity to prosodic information that allows them to learn properties of the language in their environment. I have shown that this early sensitivity changes according to the prosodic properties of the language that infants are learning, such as the acquisition of the trochaic pattern in English and German infants. This is relevant not only for Experiment 1, where we further explore the trochaic bias, but also for Experiments 3 and 4, which add evidence to the importance of the early acquisition of the native prosody. I have given an overview of the different studies that have found a relation between word segmentation and later language outcomes because we further explore this relation in Experiments 3a and 4a with 6- and 9-month-olds. Additionally, I have outlined the importance of the reliability of the methods used. In particular, I have focused on speech perception methods like the HPP, because its reliability is analyzed in the present thesis. Finally, I have discussed prosodic bootstrapping approaches, which consider prosody to be the first cue in accessing and processing language.

2.2 Statistical learning

2.2.1 Introduction

Statistical learning (SL) refers to the sensitivity to regularities in the input and it has been described as "automatic,"[4] "dynamic," "incidental," and "spontaneous" (Saffran et al., 1996b; Fiser & Aslin, 2001; Turk-Browne, Jungé & Scholl, 2005). A large body of research indicates that both infants and adults can segment words from fluent speech or an artificial language into words based on conditional statistical information, i.e., transitional probabilities (e.g., Hayes & Clark, 1970; Saffran et al., 1996b; Aslin et al., 1998; Johnson & Jusczyk, 2001; Pelucchi et al., 2009; Johnson & Tyler, 2010; Thiessen & Erickson, 2013; Bulgarelli, Benitez, Saffran, Byers-Heinlein & Weiss, 2017; for a review, see Krogh, Vlach & Johnson, 2012 or Saffran & Kirkham, 2017). Transitional probability (TP) is the conditional probability of *Y* given *X* in the sequence *XY.*

$$\text{Probability of Y| X} = \frac{\text{frequency of XY}}{\text{frequency of X}}$$

The ability to compute TPs relates to a central learning mechanism that supports the processing of statistical information. Even rats and cotton-top tamarins (a species of monkey) could track TPs in the same speech stream as used in Saffran et al. (1996b) (Toro & Trobalón, 2005; Hauser, Newport, & Aslin, 2001). There is also evidence that human adults and infants can track not only forward TPs, but also backward TPs in fluent speech (Jones & Pashler, 2007; Perruchet & Desaulty, 2008; Pelucchi et al., 2009).

SL has been documented across different domains (non-verbal auditory: Endress & Mehler, 2009a; Gebhart, Newport & Aslin, 2009; visual: Kirkham, Slemmer & Johnson, 2002; verbal: Pelucchi et

[4] Some authors assert that SL is a form of implicit learning (e.g., Olson & Chun, 2001; Kim, Seitz, Feenstra & Shams, 2009; Arciuli & Simpson, 2012), but others argue that SL cannot proceed in the absence of attention (e.g., Baker, Olson & Behrmann, 2004; Toro et al., 2005).

al., 2009). It is certain that different domains of knowledge place distinct demands on perception. However, whether one single mechanism operates across all domains or whether different ones are involved is not yet clear. Some researchers suggest that computations in different modalities might be quite similar (Perruchet & Pacton, 2006; Saffran, 2008), but others argue that different mechanisms are involved in the processing of different kinds of statistical information (Amso, Davidson, Johnson, Glover & Casey, 2005). A few comparable experiments have provided evidence for SL as a unified capacity. For example, 8-month-olds' performance in detecting TPs of non-linguistic tone sequences paralleled that in speech segmentation within infants (Saffran, Johnson, Aslin & Newport, 1999).

In the language domain, SL refers to the sensitivity to distributional regularities in the speech input, e.g., TPs between syllables or the frequency of occurrence of specific units like syllables or words. In speech, low TPs between syllables are likely to be associated with word boundaries, whereas high TPs between syllables are more likely to occur within a word. For example, in English the sequence *ele* is likely to be followed by *vator* or *phant*, but after that a lot of different words can follow. Such regularities in the speech input provide information about the language structure and are considered to be useful in learning word meanings, word endings and beginnings, lexical categories, and grammatical structure (e.g., Saffran, 2001; Graf Estes, Evans, Alibali & Saffran, 2007).

One of the main questions in the SL research refers to which types of units are tracked. TPs are exploited by both adults and infants to segment speech and extract words (e.g., Hayes & Clark, 1970; Saffran et al., 1996a, 1996b; Aslin et al., 1998). Among other statistical cues that are successfully exploited by adults are phonotactic regularities (Onishi, Chambers & Fisher, 2002; Finn & Hudson Kam, 2008), non-adjacent TPs (Peña, Bonatti, Nespor & Mehler, 2002), the relative frequency of functors and lexical items (Gervain, Nespor, Mazuka, Horie & Mehler, 2008), and

distributional properties of phonemes and allophones (Brent & Cartweight, 1996; Batchelder, 2002; also by infants: Maye, Werker & Gerken, 2002; Maye, Weiss & Aslin, 2008).

SL can be assessed in a number of ways but the most common procedure is the one first used in Saffran et al. (1996b) and in Aslin et al. (1998).[5] Infants were first exposed to an artificial familiarization speech stream and then tested on single words, which were formed from syllables that had been presented either with high TPs or with low TPs in the familiarization stream (see Section 2.2.4 for more information about this methodology). Since then, a huge replication literature has followed introducing new controlled variables in the stimuli such as different speakers, other cues to word segmentation, words of different lengths, etc. (e.g. Graf Estes & Lew-Williams, 2015; Hay, Pelucchi, Graf Estes & Saffran, 2011; Mersad & Nazzi, 2012). Further studies have explored the SL ability across development and demonstrated that SL is present from as early as 5.5 months of age and throughout infancy (e.g., Thiessen & Saffran, 2003; Romberg & Saffran, 2010; Thiessen & Erickson, 2013; Bulgarelli et al., 2017). Some of the research has also investigated the importance of SL segmentation cues compared to other cues (Mattys et al., 1999; Johnson & Jusczyk, 2001; see Section 2.3).

2.2.2 Early word segmentation with statistical learning

The sensitivity to and the ability to use statistical information in the input is already present in the first year of life and seems to play an important role in language acquisition (Thiessen & Saffran, 2003; Graf Estes, Evans & Else-Quest, 2007). There is a great deal of evidence that infants have very powerful statistical learning mechanisms at their disposal. Newborns are already sensitive to statistical structure in speech (Teinonen, Fellmann, Näätänen, Alku & Huotilainen, 2009) and non-speech streams (Bulf, Johnson & Valenza, 2011; Kudo, Nonaka, Mizuno, Mizuno & Okanoya,

[5] Both studies use the exact same procedure but the latter added a control for the frequency of words appearing in the familiarization string.

2011). From their fifth month of life, English-learning infants were shown to use TPs to segment words, not only from an artificial speech stream (Saffran et al., 1996b; Aslin et al., 1998; Thiessen & Saffran, 2003, 2007; Johnson & Jusczyk, 2003; Johnson & Tyler, 2010; Thiessen & Erickson, 2013; Graf Estes & Lew-Williams, 2015; Graf Estes, Gluck & Bastos, 2015), but also from natural speech (Johnson & Jusczyk, 2001; Thiessen, Hill & Saffran, 2005; Pelucchi et al., 2009; Johnson & Seidl, 2009; Lew-Williams, Pelucchi & Saffran, 2011; for a meta-analysis, see Black & Bergmann, 2017). This ability has also been attested in infants learning other languages such as Dutch (Johnson & Tyler, 2010) or French (Mersad & Nazzi, 2012).

To sum up, infants are sensitive to frequency patterns and other regularities in the input and learn from the speech stream they are exposed to. However, the fact that infants can distinguish between sound sequences of different internal coherence (low TPs vs. high TPs) does not necessarily mean that infants are treating and storing the output of SL like coherent word-like units that exhibit word-like properties. However, if they do, the output of the SL process would provide representations that serve as good word candidates available for mapping to meaning, since establishing a link between sound and meaning is an essential aspect of language acquisition. Early word segmentation evidence supporting these assumptions is reviewed in this section.

Prior research has further explored the role of SL in word learning and provided evidence supporting the hypothesis that infants use the distributional information for word learning and word mapping (Saffran, 2001; Graf Estes et al., 2007; Erickson, Thiessen & Graf Estes, 2014). Graf Estes et al. (2007) familiarized 17-month-old infants with an artificial language string and then tested them on an association task using the Switch Procedure[6] (the sound sequences were mapped to

[6] In this task infants are habituated to two different label-object combination, presented one at a time. An object moves from side to side while its associated label plays. Once the habituation criterion is reached, the test trials begin. Two types of trials are presented: same trials and switch trials. During same trials, the infant views the label-object combinations from the habituation phase. During switch trials, the labels for the two objects are switched.

novel objects). Infants only acquired the words when the labels were statistical words in the speech, but not when they were part-words. Thus, SL generated new representational units that infants mapped more readily to meanings. Hay et al., (2011) built on these results conducting a similar experiment with real Italian language instead of synthesized speech. Eight-month-old English-learning infants were successful in mapping words with high forward and/or backward TPs as labels for objects, but failed when no TPs were present or TPs in both directions were low. However, these results could not be replicated (Newsom, 2018). Consistent with Hay et al. (2011), adult participants learned word labels more quickly than part-word labels (Mirman, Magnuson, Graf Estes & Dixon, 2008).

2.2.3 Statistical learning and later language development

Taking into account all the above evidence, there are reasons to believe that SL plays an important role in the early stages of language development, specifically in word segmentation and word learning. If this is true and if word segmentation is related to later language development (as previously described in Section 2.1.4), then a relationship between SL and later language outcomes should be observable. If we assume that SL contributes to lexical development and other language processes, it is possible that delays and dysfunction in SL are observed in populations with language disorders. In this section I will outline the evidence from previous research about the relation between SL abilities and later language outcomes, which motivated us to consider the link between vocabulary outcomes and SL performance in Experiments 3a and 4a.

Recent research has investigated individual differences in SL performance and its correlation with later language outcomes. The general hypothesis is that individuals who perform better on SL tasks should also achieve superior language learning and processing outcomes. Evidence supporting this hypothesis comes from certain studies with children and adults that report significant correlations

between SL and language outcomes (reading abilities in children and adults: Arciuli & Simpson, 2012; comprehension of syntactic structures in children: Kidd & Arciuli, 2016).

Overall, despite the heterogeneity and the great variability faced in this kind of research, there is consistency in the prediction that individuals with SLI will show impairments in SL tasks. Two meta-analyses (Graf Estes et al., 2007; Obeid, Brooks, Powers, Gillespie-Lynch & Lum, 2016) have confirmed this general pattern: children with SLI perform worse on SL tasks than TD children. Evans, Saffran and Robe-Torres (2009) tested SLI and TD elementary-school-aged children in an SL word segmentation task. Whereas performance for the TD control group was significantly greater than chance after 21 min of familiarization, SLI children performed better than chance only after 42 min of familiarization, but not after 21 min. Remarkably, expressive and receptive vocabulary (measured by standardized vocabulary tests) were correlated with the performance in the SL word segmentation task in the TD control group, but not in SLI children. However, in the second experiment (42 min of exposure) there was a correlation of SLI children's performance and receptive vocabulary. The study suggests that children with SLI do not use statistical information as effectively as their peers. Importantly, as observed by Mainela-Arnold and Evans (2014), SL might be predictive for later language development in children with SLI. In their study, SL (TPs) abilities predicted lexical-phonological abilities in TD and SLI children, but not lexical-semantic knowledge.

The research on populations with language difficulties or delays clearly shows that SL is impaired relative to that observed in TD populations. How SL abilities are used by infants will help us to understand the developmental trajectories characterizing children with different disorders and individual differences more generally. Although it is likely that a relation between SL and language development exists, more research is needed to reach a conclusion as to whether SL can be used as a potential predictor of later language development and language disorders. Few studies have

assessed the reliability of methods for assessing SL abilities in infants and the relation of these measures with language outcomes. Therefore we wanted to shed more light on this relation by conducting SL experiments with young infants and obtaining later language outcomes data (Experiments 3a and 4a).

2.2.4 Specific methodological considerations in statistical learning experiments

SL can be assessed in a number of ways (visual scene base pairs[7]: Fiser & Aslin, 2001; Artificial Grammar Learning: Reber, 1967; Serial Reaction Task: Nissen & Bullemer, 1987; Hebb repetition task: Hebb, 1961; contextual cueing: Chun & Jiang, 1998; cross-situational learning, Yu & Smith, 2007), but the most common SL test is the one on word segmentation originally proposed by Saffran et al. (1996b). Infants are usually exposed to an artificial familiarization stream consisting of non-words with controlled TPs within and between words for around 2-3 minutes and then tested on single words from the string. The TPs between syllables within the words in the string are 1.0 and the order of occurrence of these words in the string is varied such that the TPs across the words are lower than the TPs within the words, usually ranging between 0.4 and 0.2. This procedure usually takes place in a HPP booth, where infants are sitting on their caregiver's lap and their attention (looking times) to the specific single words is measured. This is the task described in Section 3.1.4 and used in the infant experiments in this thesis (Chapters 5 and 6). It is important to note that, in some artificial language studies, the string starts and ends with an amplitude ramp to avoid any salience effects of the initial syllables. In the current thesis we added a ramp and did not find any significant effect on the results. However, we only explored this variable with adults (see Chapter 4, Experiment 2b).

---

[7] This task consists of two phases: familiarization and test. During the familiarization phase certain shapes are organized into base pairs, which consist of two given shapes in a particular spatial relation. If one of the elements of a base pair appears in a given scene during familiarization, the other element always appears in an invariant spatial relation to it (TPs). The joint probability of the two shapes in each of the base pairs is .50, whereas the probability of non-base pairs is typically less than .02. Infants are tested on base and non-base pairs.

SL experiments have received criticism in recent years and research is and has been trying to improve the ecological validity of such experiments. One of the main problems is that the stimuli used in these experiments are unnatural and insufficiently complex. Compared to real language, they are relatively simple in their acoustic properties and in the distribution of the words. Whereas natural speech contains high variability, regularities like rhythm and stress cues, as well as phonotactic regularities and pauses, these cues are usually removed from the experimental stimuli. Contrary to what we might think, increased variability might not pose a challenge to infants' discovery of words, but may actually facilitate word segmentation. In the laboratory experimental condition, the stimuli are highly controlled: the input is extremely concentrated (presentation of only a few words for no more than 3 minutes) and the TPs of syllables are perfect. However, while the amount of language input that infants (or adults) receive in the real world is vastly greater, the conditional probabilities as cues to word boundaries are much noisier.

Importantly, recent research has responded to this criticism by focusing on trying to find solutions and improvements to SL experiments by increasing the complexity of the familiarization stream. For example, several studies have recorded a human speaker, thus obtaining more natural stimuli, instead of using synthesized speech, in order to add more acoustic variation (e.g., Hay et al., 2011; Johnson & Jusczyk, 2001; also the experiments presented in this thesis). In some cases, different results were obtained than when synthesized speech was used. In fact, according to Black and Bergmann (2017), natural speech is more likely to show familiarity preferences, which might be caused by the increased complexity and/or variability of the stimuli compared to synthesized speech. Furthermore, there is evidence that using Infant Directed Speech (IDS) in an SL word segmentation task improves infants' performance (Thiessen et al., 2005). In their study, 6.5- to 7.5-month-old English-learning infants were successful in distinguishing words from syllable sequences

spanning word boundaries after hearing IDS, but not after exposure to Adult Directed Speech (ADS).

Word length has also been one of the controlled variables that some studies have explored. However, the results are inconsistent. Johnson and Tyler (2010) included a condition where the familiarization string was a mixture of disyllabic and trisyllabic words. Both 5.5- and 8-month-olds successfully segmented the words from the artificial language in the uniform condition, but neither of the groups succeeded when the words varied in length. Similarly, Lew-Williams and Saffran (2012) exposed 9- and 10-month-old infants to a list of either disyllabic or trisyllabic words (pre-exposure), followed by a speech stream composed of disyllabic or trisyllabic words (familiarization). Infants failed to segment the words when the pre-exposure and the speech stream contained words of different lengths (e.g., when the pre-exposure contained trisyllabic words and the speech stream disyllabic words). In contrast, Thiessen et al. (2005) showed that 6.5- to 7.5-month-old English-learning infants were successful in segmenting a string with varying word lengths when it was produced with IDS prosody.

Previous research has evaluated the reliability of SL measures in order to explore individual differences and eventually use these measures as a tool to predict later language achievements. However, SL measures do not correlate highly with each other (different SL tasks) nor with other measures of cognitive ability. Siegelman and Frost (2015) looked at the test-retest reliability of SL and report that the outcomes are quite variable, with correlations ranging from just below 0.7 to 0.2. Erickson, Kaschak, Thiessen and Berry (2016) obtained results in line with Siegelman and Frost (2015). Participants were tested on a range of SL performance tasks repeatedly at two points in time. The test-retest reliability of the tasks was generally low, even after trying to improve the reliability of the task by adding more trials. Overall, the individual SL measures showed significant test-retest correlations but the correlations were generally low.

Therefore, if SL measures are to be used as instruments to explore individual differences in language learning, it is critical to ensure that the task has the appropriate properties needed to yield meaningful correlations at an individual level. The task must be valid (large enough variance of the output scores, internal validity, etc.) and must be reliable enough to discriminate "good" learners from "bad" learners (for more validity criteria suggestions, see Siegelman, Bogaerts and Frost, 2017). In the present thesis we aim to assess the test-retest reliability of an SL task and to explore whether individual differences in such a task are related to later language development.

2.2.5 Models for early statistical learning segmentation

As with early prosodic acquisition, SL has been proposed as one bootstrapping mechanism. This mechanism is assumed to compute statistical properties on different language levels and is used to find syntactically relevant units in the input. For example, inflectional endings and function words are highly frequent and typically occur at the edges of words or syntactic phrases (Gerken, 1996; Mintz, Newport & Bever, 2002; Pelzer & Höhle, 2006). Numerous models have been proposed to explain the SL mechanism in word segmentation and word learning in a language acquisition frame. Although examining all the SL accounts and models is beyond of the scope of this dissertation, I will try to provide an overview of the main principles.

It is not yet clear which type of model provides the most valid account of human learning processes across tasks (Frank, Goldwater, Griffiths & Tenenbaum, 2010), but the main principle assumed in these models is that infants rely on language-universal cues, such as conditional statistical information, as a first step in segmenting words from speech. Therefore, SL would work without any previous knowledge about the native language. Such accounts suggest that language-specific cues like prosodic cues are part of a second step in language acquisition: to make use of language-specific cues, infants must already know something about the sound patterning of their native language with respect to correlations between sound patterns and word boundaries. For example,

SL accounts argue that simply hearing the alternation of stressed and unstressed syllables is not enough evidence for infants to acquire a rhythmic segmentation bias (trochaic or iambic), because infants have not yet discovered how stress is correlated with word boundaries. Therefore, these accounts contend that SL plays an important role in the development of a rhythmic segmentation bias and that SL is a word segmentation strategy that precedes attention to stress cues. It is only after infants have learned an inventory of words that they can discover that stress predicts word onsets.

SL accounts are supported by two main lines of evidence: computational and experimental. The first one is that distributional information (TPs) can provide good cues to word boundaries independently of the language that the infant is learning (Brent & Cartwright, 1996; Saffran et al., 1996b; Aslin et al., 1988; Barchelder, 2002; Swingley, 2005). Swingley (2005) analyzed both Dutch and English infant-directed corpora and concluded that TPs can provide enough information about word boundaries. Nevertheless, recent studies show that there are cross-linguistic differences, since co-occurrence statistics are not equally informative in all languages (Saksida, Langus & Nespor, 2016). The second line of evidence is that infants can actually make use of these TPs to segment words, and treat them like words (e.g., Graf Estes et al., 2007; see the previous Section 2.2.2 for more details). However, while it is unambiguously clear that infants are sensitive to the regularities in the input, some authors claim that there is no clear evidence that infants use distributional information for language acquisition and that such mechanisms might not be suitable for all languages (Yang, 2004; Endress & Mehler, 2009b; Endress & Hauser, 2010).

Overall, SL accounts have focused on the acquisition of language regularities, postulating that infants start the acquisition process using statistical information from fluent speech. In these accounts it is assumed that infants rely on this kind of information as a first step in word segmentation.

2.2.6 Summary

In this section I have provided an overview of the evidence supporting infants' sensitivity to the conditional statistical information in their linguistic environment and their use of these patterns to facilitate subsequent learning. I have reviewed the evidence of word segmentation through the SL mechanism and its relation with later language development and explained the principles of the models that claim SL to be the main mechanism for early language acquisition. The literature shows that more research is necessary to characterize the nature of potential causal links between statistical learning and later language outcomes. In addition, I have described the most common statistical experiment methodology and its limitations, which recent research has been trying to improve.

2.3 The weighting of cues for word segmentation

2.3.1 Introduction

One of the goals of this thesis is to understand how both infants and adults use acoustic information to identify linguistic structure and to explore the cues they use for word segmentation. As reported in the previous sections (2.1 and 2.2), infants' ability to track TPs across speech segments and their sensitivity to prosodic and rhythmic information seem to be involved in the capacity to detect word boundaries. Although sensitivity to both prosodic and statistical cues has been shown very early, research indicates that not all the cues are equal, neither in the frequency in which they appear, in the degree of reliability, nor in their reliability across different languages.

Recall that English-learning infants are sensitive to distributional regularities in the input from the age of 5.5 months (e.g., Johnson & Jusczyk, 2001; Thiessen & Erickson, 2013; Bulgarelli et al., 2017) and are able to use this information to segment words from fluent speech. In parallel, infants also develop sensitivity to other potential prosodic word boundary cues such as word stress patterns (e.g., Jusczyk et al, 1993; Morgan & Saffran, 1995). However, it is difficult to determine which cues infants rely on to solve specific language learning problems at any given point in development. Studies that examine cues in isolation cannot reveal their particular development or how a particular cue is weighted with respect to other cues. Therefore, some recent research has focused on the roles of the different cues in word segmentation and on whether the use of one of the cues occurs developmentally earlier than the others, since what is important or attended to the most might be different at different points in development, for different tasks and also for different languages. In this chapter I provide an overview of the research on cue weighting for word segmentation in both adults and infants, but I mainly focus on how infants respond when different cues provide conflicting information about possible word boundaries.

2.3.2 When cues conflict

Some studies have investigated whether statistical information shows any dominance over other segmentation cues and have set up experimental situations where TPs conflict with other cues that indicate different word boundaries. However, results with adult participants do not provide a homogeneous picture. On the one hand, some studies have shown that there is a dominance of statistical information when both prosodic and statistical cues are present. For example, Mattys, White and Melhorn (2005) showed that English diphones with low phonotactic probability are interpreted as word boundaries regardless of stress pattern and that diphones with high within-word phonotactic probabilities suppress the perception of word onsets signaled by stress cues. On the other hand, other studies have claimed that prosodic cues easily override TPs in the segmentation of speech in English, Finnish, and Italian speakers (Vroomen, Tuomainen & De Gelder, 1998; Gambell & Yang, 2006; Shukla, Nespor & Mehler, 2007; Fernandes, Ventura & Kolinky, 2007; Langus, Marchetto, Bion & Nespor, 2012). For example, Fernandes et al. (2007) tested Portuguese listeners in an artificial-language learning setting and showed that coarticulation overruled TPs. In Vroomen et al. (1998), Finnish, Dutch, and French adult listeners performed best when the phonological properties of the artificial language matched those of the native one (speakers of Finnish profited from vowel harmony and word-initial stress, speakers of Dutch from word-initial stress, and French speakers from neither of these). Interestingly, prosodic cues also outweigh statistics in acoustically impoverished conditions such as a degraded signal with white-noise superimposition (Smith, Cutler, Butterfield & Nimmo-Smith, 1989; Liss, Spitzer, Caviness, Adler & Edwards, 1998; Fernandes et al., 2007).

It is also likely that more weight may be given to remaining segmentation cues when some sources of information are absent. For example, stress was of minor importance for speakers of Dutch or English when alternative cues like phonotactic cues were available (Cairns, Shillcock, Chater &

Levy, 1997). Also, Italian speakers were capable of segmenting an artificial string based on TPs when no prosodic cues were present (Langus et al., 2012). However, in the same series of experiments, participants successfully exploited non-native prosody to learn about the statistical properties of the speech stream. Likewise, Sohail and Johnson (2016) tested English speakers on an artificial speech stream that contained either (1) TPs to word boundaries, (2) silences marking utterance boundaries, or (3) a combination of both cues. Participants performed equally well in conditions 2 and 3, but performed at chance in condition 1, showing that participants failed to compute TPs or to use this information to segment the speech stream when no other cues were present.

Previous studies have reported an interaction between prosody and statistics in word segmentation. Shukla et al. (2007) explored the interaction between phrasal prosodic cues (intonational phrases) and TPs between syllables in Italian adult speakers. Participants recognized statistically well-formed items only when they were consistent with prosodic phrase boundaries. Therefore, the authors argue that, although participants could compute TPs independently of prosody, prosodic cues might act as a filter and constrain the lexical search.

Infants have been tested in similar experimental conditions to adult participants. When different cues offer conflicting information about word boundaries, prosodic cues like lexical stress seem to modulate word segmentation in most of the studies that tested such scenarios. The first evidence comes from Mattys et al. (1999), who tested 9-month-olds using the HPP with a familiarization string containing stress and phonotactic cues pitted against each other. The string consisted of strong-weak CVC-CVC disyllabic non-words (C-C was a consonant cluster) and the words were stressed either on the first or on the second syllable. Infants listened significantly longer to stimuli with strong-weak patterns that violated phonotactic cohesion than to weak-strong stimuli that did not.

Further evidence was obtained by Johnson and Jusczyk (2001), who investigated which of the types of information has a stronger effect on infants' speech segmentation when TPs are pitted against prosodic cues. In their study, infants were familiarized for 2 minutes with a syllable string that contained stress as well as TP information as cues for segmentation. After familiarization, infants were tested with words based on prosodic cues and words based on TPs. They found that English-learning 8-month-olds' speech segmentation was affected more strongly by the prosodic than by the statistical cues. Jusczyk et al. (1999) familiarized 7.5-month-old English-learning infants with passages in which weak-strong targets were always followed by the same monosyllabic unstressed words. Infants followed the trochaic bias, taking the strong syllable of the iambic word and the following weak syllable as one single word, suggesting that both kinds of information seem to play a role.

Using a similar experimental design as Johnson and Jusczyk (2001), Thiessen and Saffran (2003) provided evidence for a developmental shift in cue reliance during the second half of the first year of life: while English-learning 7-month-olds relied more strongly on statistical cues in their segmentation performance, 9-month-olds were more strongly guided by the prosodic cues. Based on these findings, the authors argue for an initial dominance of statistical cues over prosodic cues, which turns into a stronger weight of prosodic cues with growing language experience. They explain this change in cue relevance by a crucial difference in the status of the cues. According to their reasoning and supporting SL accounts (see Section 2.2.5), the exploitation of prosodic cues requires the previous acquisition of the language dominant word stress pattern, which does not make prosody an optimal candidate for a bootstrapping mechanism. In contrast, detecting co-occurrence patterns and computing TPs in the speech input needs no specific language knowledge and therefore may serve as an initial gateway to speech segmentation. Their proposal for this developmental shift was supported by further findings showing that 5-month-old English infants

relied more strongly on statistical cues (Thiessen & Erickson, 2013) and that 11-month-olds relied more strongly on prosodic cues (Johnson & Seidl, 2009).

However, the preference for statistical or prosodic cues observed at 7 and 9 months of age (Thiessen & Saffran, 2003) can easily be influenced by previous exposure to stress patterns, as reported by Thiessen and Saffran (2007). In this study infants heard a previous iambic or trochaic word list (pattern-induction material) and were then familiarized with an artificial speech stream which contained only statistical cues. Infants' learning seemed to be influenced from the pattern-induction materials and segmented according to their prosodic pattern. This suggests that infants are capable of learning the prosodic structure of a language within a very short time period and use it for segmentation. In addition, this evidence supports the claim that infants can easily develop a trochaic bias with exposure to their native language.

To the best of our knowledge, the relation of statistical and prosodic cues has not been tested in infants in languages other than English before. However, if Thiessen and Saffran's (2003) argument about statistical cues being language independent is correct, a similar developmental shift in cue reliance should be observed across languages –at least across languages in which the word stress pattern provides reliable cues for word segmentation. German is an interesting case to test this hypothesis, because German-learning infants have shown indications of already being sensitive to relevant language-specific prosodic properties at the age of 4 months (Friederici et al., 2007; Herold et al., 2008). The main aim of the present thesis is to fill this research gap by providing experimental data from both German infants and adults.

2.3.3 Summary

In the previous section I have discussed the weighting of cues for word segmentation in both adults and infants through the different experimental settings in the literature. Although infants and adults

are sensitive to both TPs and prosodic information, research has shown that not all the cues are equal. Exploring the strength of the different cues has been possible by adding conflicting cues marking different word boundaries. When cues collide, divergent results have been reported in the literature. However, there is a tendency to believe that prosodic cues might play a more important role in early word segmentation, at least at the early stages and for languages like English or German. This was the motivation for testing German adults and infants in an SL word segmentation task, German being a language which has not yet been tested in such an experiment.

2.4 Summary and research questions

Taken together, the findings from the previous research suggest that the use of prosody and statistical cues are central mechanisms during the early steps of language acquisition. Infants are highly sensitive to prosodic and statistical cues which –among other cues– will allow them to segment words from fluent speech and map words to real word referents. In fact, previous research has already documented an early impact of the native prosody on early language development such as with word stress pattern preferences (trochaic bias) in languages like English and German.

As outlined in Sections 2.1.4 and 2.2.3, a link between later language development and early word segmentation has been found for both prosody and statistical learning. However, critically, more longitudinal studies are needed to explore this relation. Although statistical and prosodic cues are likely to be early indicators of later language development, as shown by several studies, the methods used in early speech perception do not always provide reliable measures to assess individual variability. To help with this matter, one chapter of this thesis is focused on the methodology widely used for obtaining measurements of early speech perception (Chapter 3) and two of the experiments are longitudinal studies which address the link between speech perception measures and later language outcomes (Experiments 3a and 4a).

Although infants use both statistical computations and prosody to segment words from fluent speech, it is difficult to determine which cues infants rely on to solve specific language learning problems at any given point in development. Interestingly, recent research has focused on the roles of the different cues to word segmentation and on whether the use of one of the cues is developmentally earlier than the other. As found by Johnson and Jusczyk (2001) and Thiessen and Saffran (2003), 7-month-old English-learning infants rely more on statistical cues whereas 8- and 9-month-olds show a stronger reliance on prosodic cues. Based on the previous literature on German infants showing that prosody is a highly salient cue for word segmentation, we were interested in

exploring the weighting of potential segmentation cues by German infants in their early stages of language acquisition.

From a theoretical point of view, different models (statistical and prosodic bootstrapping accounts) have tried to explain the mechanisms used in early word segmentation as well as the developmental shift observed in English-learning infants. Actually, it is a chicken-and-egg problem: does the infant use statistical cues to discover prosodic regularities in her native language or does she use prosodic cues to isolate chunks upon which TPs are computed? A further goal of the current work is to shed more light on this issue by providing data from German infants and adults and trying to provide an explanation within the different frameworks.

I will try to answer the following questions:

a) Do German infants and adults rely more strongly on prosodic or statistical information when segmenting words from fluent speech? How do infants integrate knowledge of the cues at particular points in development? Do these two cues interact with each other and are they used differently depending on age and/or language experience?

b) Can the use and weighting of these cues in a word segmentation task predict later language skills? Is the Headturn Preference Procedure reliable enough to obtain predictive measures for later language development?

The first research question will be investigated and discussed based on experimental data collected within this dissertation project in Chapter 4 (adult data), Chapter 5, and Chapter 6 (infant data). First, an experiment with German adults was conducted to gain insights into which cues adults would base their segmentation on when presented with our stimuli (Experiments 2a and 2b). Second, we conducted two experiments with 6-month-old German infants to explore how they weight statistical and prosodic cues (Experiments 3a and 3b). The purpose of Experiment 3a was to

compare the potential segmentation strategies and Experiment 3b was a control experiment to check for any spontaneous preferences for the stimuli presented during the test phase of Experiment 3a. Experiments 3c and 3d were a replication of Experiments 3a and 3b with 9-month-olds. Because of the obtained null results, 9-month-olds were tested without familiarization (Experiment 3c) and with double familiarization exposure (Experiment 3d). A further goal of the experiments with 6- and 9-month-olds is related to the second research question. Therefore, data regarding infants' later language outcomes was obtained by using parental questionnaires. The link between the weighting of these mechanisms and later language development is discussed.

A few methodological issues are also addressed in the current thesis. First, reliability data for the HPP as well as its relation to later language outcomes is presented and discussed in Chapter 3. Secondly, we explore the issue of having an amplitude ramp in the artificial language string used for familiarization. Thus, Experiment 2b with adults is a replication of Experiment 2a but with the addition of an amplitude ramp. Finally, we also extended the research question to another methodology: pupillometry. Both adults (Experiment 2a) and infants (Experiment 3b) were tested in such a procedure. The motivation for this was to compare behavioral and online data (for adults) and to obtain time continuous data about the word segmentation process (for adults and infants).

3. THE HEADTURN PREFERENCE PROCEDURE

3.1 Experiment 1: Test-reliability test of the Headturn Preference Procedure (HPP)

3.1.1 Introduction

Infants are equipped with highly efficient capacities to process specific information from speech. One instance of this is a high sensitivity to prosodic information, which has even been observed in newborns and which is assumed to be crucial for bootstrapping certain aspects of the lexical and syntactic acquisition (for an overview, see De Carvalho et al., 2018). Recall that there is growing evidence that the abilities that infants show in their speech perception can be predictive of their later language achievements (see Section 2.1.4), and thus that early signs of a developmental risk in language acquisition may already be detectable at a young age. However, to apply measures of early speech perception in such a way, they must be reliable indicators of individual performance. The present experiment investigates the reliability of the HPP –a behavioral paradigm that is widely used in infant speech perception research– in a test-retest-reliability study by repeatedly testing German 6-month-old infants for a listening preference for trochaic or iambic disyllabic sequences.

The stimuli used in the current experiment have been used in other HPP experiments (Höhle et al. 2009; Bijeljac-Babic et al., 2016) and have already shown a predictive value for later language performance (Höhle et al., 2014). In Höhle et al.'s (2014) longitudinal study, German 4-month-old infants were tested on their responses to trochaic and iambic items after being familiarized to trochaic patterns using the HPP. Overall, longer looking times to the familiarized trochaic patterns were found. Most importantly, the amount of decrease in looking times for the iambic items was correlated to the children's performance in a language assessment when they were 5 years old. A high decrease in looking time was associated with higher scores in tests on morphological rules and sentence comprehension. This suggests a specific relation between early prosodic development and later language skills.

In the present study, infants' response to iambic and trochaic disyllabic sequences using the HPP without any familiarization was repeatedly tested during their sixth month of life. This age was chosen since Höhle et al. (2009) found that German 6-month-olds show a spontaneous listening preference for trochaic items (trochaic bias), which indicates that they have acquired a basic property of the German prosodic system. Individual variation in this developmental achievement may thus be a potential predictor of later language performance and is therefore specifically interesting for our test-retest-reliability study.

3.1.2 Participants

Thirty-eight 6-month-old German monolingual infants (18 girls) were tested in three test sessions. The mean age at the first test session was 6 months and 10 days (range 6;01 – 6;18), the mean at the second session was 6 months and 18 days (range 6;08 – 6;25), and at the last session the mean age was 6 months and 30 days (range 6;13 – 7;01). All infants were born full-term without apparent health problems. Four additional infants were tested but excluded due to fussiness (2) and not completing all the testing sessions (2). Written informed consent was obtained from all participating families.

3.1.3 Stimuli

The stimuli were those used in Höhle et al. (2009), which consisted of CVCV /gaba/ sequences, stressed either on the first (trochaic pattern) or on the second syllable (iambic pattern), and recorded by a female German speaker. The first syllables of the trochaic sequences had a mean duration of 283 ms (*SD* = 20.8) and an average pitch of 195 Hz (*SD* = 3.9). The corresponding values for the second syllable were 308 ms (*SD* = 25.0) and 163 Hz (*SD* = 15.9). The first syllables of the iambic sequences had a mean duration of 173 ms (*SD* = 11.0) and an average pitch of 186 Hz (*SD* = 5.2); the values for the second syllables were 430 ms (*SD* = 21.2) and 183 Hz (*SD* = 5.9). Five audio files

for each stress pattern were created, each containing the same set of tokens of the same stress pattern, separated by 600 ms pauses. They differed in the order of presentation of the different tokens. The trochaic speech files contained 16 tokens and had an average duration of 18.39 s (range: 18.28 – 18.51). The iambic files contained 15 tokens and the average duration was 18.01 s (range: 18.00 – 18.07). The difference in the number of tokens was due to the fact that the iambic sequences were longer than the trochaic ones because of the long duration of the second syllables in the iambic stress pattern.

3.1.4 Procedure

We used the HPP as introduced by Hirsh-Pasek et al. (1987). The procedure and apparatus were the same as in Höhle et al. (2009) in all three test sessions except that the experiment was run without familiarization. The test sessions were planned to be separated by 7 days. However, depending on the parents' availability the timing varied slightly across infants. The mean period between the first and the second test sessions was 7.23 days (range: 6 – 10) and between the second and the third sessions 8.39 days (range: 4 – 11).

During the experiment, infants were seated on a caregiver's lap in the center of a test booth. The caregiver listened to music over headphones to prevent influences on the infant's behavior. Furthermore, he or she was instructed not to interfere with the infant during the experiment. Inside the booth, three lights were fixed: a green one at the center, and a red one on each side. On the outside of the test booth, two loudspeakers were mounted just below the red lights. Each trial started with the blinking of the green light to attract the infant's attention to the center. When the infant oriented to the light, it went out and one of the side red lights started to blink. When the infant turned her head towards it, the speech stimulus was started. The speech stimulus was either played until completion or was stopped when the infant turned her head away for more than 2 consecutive seconds. If the infant turned her head for less than 2 s, the presentation of the speech

file continued but the time spent looking away was not included in the total looking time. Looking times were coded by an experimenter outside the testing room with a push-button control.

The first two speech files (one trochaic and one iambic) served as warm-up trials and were not included in the analysis. The eight experimental speech files were presented in four different versions, which differed in trial order and were counterbalanced across participants. Infants were always tested with the same version across the three test sessions. Each experimental session lasted between 3 and 5 minutes, depending on the infant's behavior. For each infant, language outcome at 12 and at 24 months of age was assessed by two standardized German parental questionnaires (ELFRA-1 and ELFRA-2, Grimm & Doil, 2006). The questionnaire ELFRA-1 consists of four subtests: speech production (productive vocabulary and production of sounds and word combinations), speech perception (receptive vocabulary and reaction to language), gestures, and fine motor skills. ELFRA-2 consists of three subtests: productive vocabulary, syntax, and morphology.

3.1.5 Results

As in Höhle et al. (2009), all individual looking times longer than 18 s were reduced to 18 s to account for the difference in length of the audio files between the two conditions (0.38 % of the total number of trials). The data were not normally distributed (Shapiro Test, $W = 0.88$, $p < .001$). Therefore, non-parametric tests (Wilcoxon Signed-Rank Test and Spearman's rho correlation) were used for the data analysis. A second experimenter (blind to the experiment) recoded offline 14% of the videos chosen randomly to determine the inter-rater reliability (agreement was 94%).

3.1.5.1 Rhythmic preferences

The mean looking times for each stress pattern and for each infant were calculated separately for each session (see Figure 1). In the first session, the mean looking times were 7.93 s ($SD = 4.69$) for

the trochaic sequences and 7.63 s (*SD* = 4.3) for the iambic sequences. This difference was not significant (*V* = 304, *p* = .17, *r* = -.22). In the second session, infants oriented to the trochaic sequences for 4.77 s (*SD* = 3.46) and to the iambic sequences for 4.8 s (*SD* = 3). The difference was again not significant (*V* = 405, *p* = .69, *r* = -.06). In the last test session, infants oriented to the trochaic sequences for 4.76 s (*SD* = 3.4) and to the iambic sequences for 4.23 s (*SD* = 3.12). Here the difference was significant with a medium size effect (*V* = 222, *p* = .01, *r* = -.39). Twenty-six out of 38 infants had longer looking times to the trochaic than to the iambic sequences. In addition, a significant decline in looking times for both iambs and trochees was observed between the first and the second test sessions (*V* = 687, *p* < .01; *V* = 703, *p* < .01, respectively). In an exploratory analysis we observed a negative correlation (*r* = -.55, *p* < .01) between age and iambic looking times in the second test session (see Figure 2).

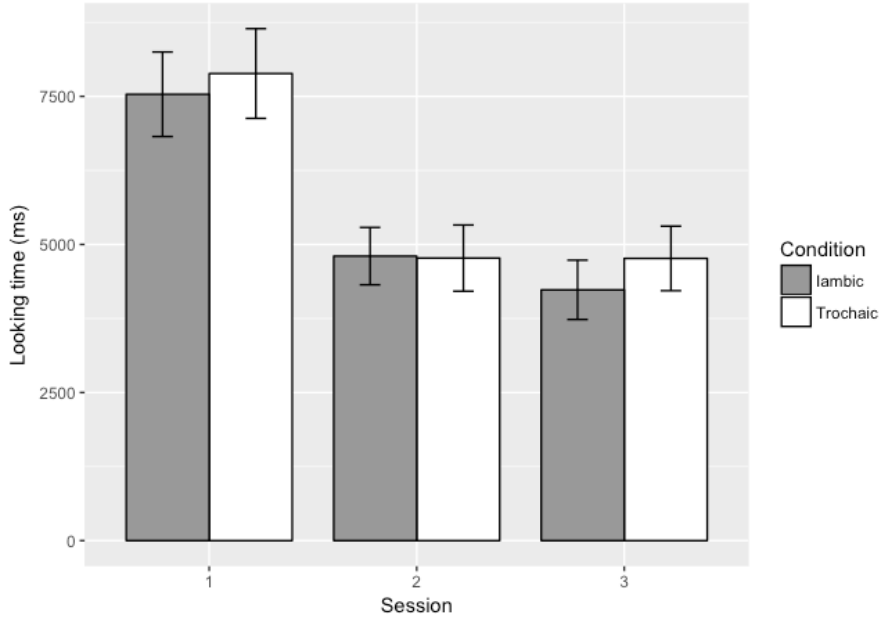

Figure 1: Mean looking times in the three sessions.
The error bars represent the standard error.

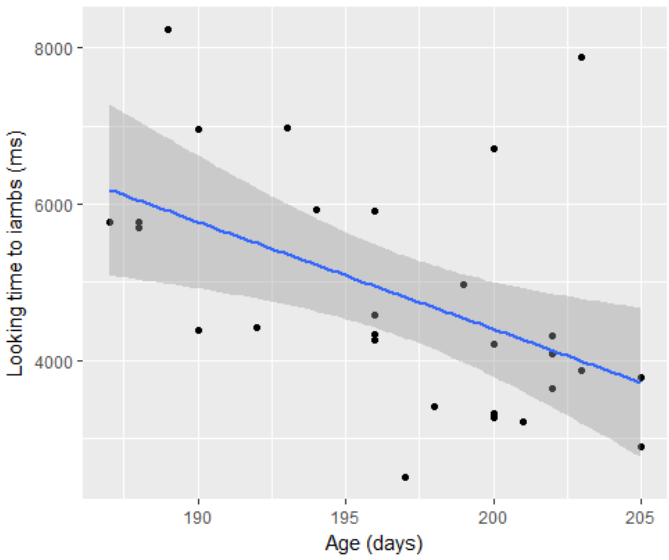

Figure 2: Negative correlation between age and iambic looking times in the second session

3.1.5.2 Test-retest reliability

Regarding the individual performance stability, 10 out of 38 infants had the same pattern of preference across all three sessions with 6 infants showing numerically longer looking times to trochaic sequences and 4 infants to iambic sequences. To examine the test-retest reliability of the procedure, correlations were calculated for the looking time raw scores obtained in the three sessions (see Table 1). Significant and medium-size correlations were found in looking times for iambic sequences across sessions 2 and 3 while the looking times for trochaic sequences were correlated across all three sessions. Correlations were also calculated for the difference scores (trochaic looking times minus iambic looking times) between the sessions but none of these was significant.

Table 1. Correlations between the looking times across the three sessions

Session	**Iambs**		**Trochees**		**Difference scores**	
	Rho	p-value	Rho	p-value	Rho	p-value
Session 1	.25	.92	.03	.41	-.24	.16
Session 2	-.15	.18	.30	.03*	.37	.02*
Session 3	.19	.87	.17	.15	.02	.44

** = significant, $p < .05$*

3.1.5.3 Correlations with later language development

Correlations between the looking times for each rhythmic pattern and infants' ELFRA-1 and ELFRA-2 scores were calculated. Thirty-five of the infants that were tested in the HPP were included in the ELFRA-1 analysis. Three infants were excluded because the parents did not complete the ELFRA-1 questionnaire. Of the infants that were tested in the HPP, 29 were included in the ELFRA-2 analysis. Nine infants were excluded because the parents did not complete the ELFRA-2 questionnaire. The overall mean ELFRA-1 score was 74.2 (*SD* = 37.39) out of 370 points. Infants scored 34.85 (*SD* = 27.19) out of 171 possible points in the receptive vocabulary subtest, and 3.91 (*SD* = 6.7) out of 181 possible points in the productive vocabulary subtest. The trochaic raw scores as well as the difference scores in session 2 were positively correlated with the overall scores in ELFRA-1 (see Table 2 and Figure 3).

Table 2. Correlations between the raw scores and the total test scores in ELFRA-1

Session	**Iambs**		**Trochees**		**Difference scores**	
	Rho	p-value	Rho	p-value	Rho	p-value
Session 1	-.06	.37	-.07	.65	-.09	.62
Session 2	-.008	.48	.022	.45	.12	.54
Session 3	.24	.89	.021	.45	-.03	.86

** = significant, p < .05*

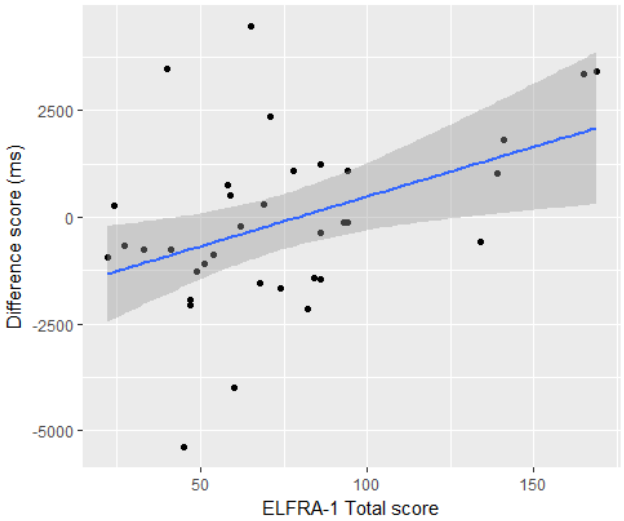

Figure 3: Correlation between difference scores and ELFRA-1 total scores

The overall mean ELFRA-2 score was 135.7 (*SD* = 64.3) out of 323 points. Infants scored 113.1 (*SD* = 52.52) out of 260 possible points in the productive vocabulary subtest, 17.53 (*SD* = 9.23) out of 47 possible points in the syntax subtest, and 5.14 (*SD* = 4.45) out of 16 possible points in the morphology subtest. No significant correlations were observed between the ELFRA-2 outcomes and the looking times to either iambic or trochaic sequences (see Table 3).

Table 3. Correlations between the raw scores and the total test scores in ELFRA-2

Session	**Iambs**		**Trochees**		**Difference scores**	
	Rho	p-value	Rho	p-value	Rho	p-value
Session 1 - 2	.29	.07	.30	.03*	.25	.11
Session 2 - 3	.44	< .01*	.47	< .01*	.31	.06
Session 1 - 3	.22	.18	.33	.02*	.08	.63

** = significant, p < .05*

3.1.6 Discussion

The aim of this study was to check whether the HPP is a robust instrument in terms of test-retest reliability. The long-term goal was to give some insight into whether the HPP could be implemented as a tool for the early detection of children at risk for developmental language problems. Therefore, a test-retest-reliability study was conducted as a first step. We tested 6-month-old German monolingual infants for a listening preference for trochaic items (the trochaic bias) at three different test points with the same procedure. We then followed up with the infants tested in that first task and obtained their language development scores at 12 and 24 months of age through a standardized parental questionnaire. Our main research question was whether the speech perception outcomes from the HPP procedure are consistent and thus reliable across multiple testing sessions. We found that the looking times for the trochaic sequences significantly correlated (with medium-size effects) between all three sessions. In contrast, the looking times for the iambic sequences only correlated between the second and third sessions. These data suggest that there is some degree of stability in

the looking times to the two conditions across the test sessions: infants with long looking times to trochaic items showed this pattern across all sessions. The fact that the looking times for the iambic patterns did not correlate across all sessions but only between the second and third sessions allows two conclusions. First, the observed correlations do not simply reflect the fact that children are quite stable with respect to their overall attention to the stimulation in this experimental paradigm, i.e., that they are not either long or short listeners across the board. Second, the finding that the looking times for the iambic items were less stable across the test sessions could be taken as an indication that the emergence of the trochaic bias (which was statistically significant only in the last test session) is mainly related to a change in the looking times to iambic items. This would fit the typical pattern of perceptual reorganization: infants develop a preference for the trochaic stress pattern of their native language (German), which is more familiar to them. An observed negative correlation between age and iambic looking times in the second test session further supports this assumption (Figure 2). These differences in the stability of the looking times to the two test conditions are probably also the reason why the difference scores are not significantly correlated across the test sessions. Interpreted in this way, our results indicate that repeated measures can also provide detailed insights into an ongoing developmental change.

Our findings are in contrast with Höhle et al. (2009), who tested 6-month-old German-learning infants for a trochaic bias with the exact same stimuli and method. Infants in their study showed a significant preference for trochaic over iambic sequences. The only difference between their experiment and the first test session of our study was the age range of the infants tested, slightly broader in their study. Infants' mean age range in Höhle et al. (2009) was 6 months 1 day to 6 months 28 days (mean: 6 months 12 days). In our study it was 6 months 1 day to 6 months 18 days (mean: 6 months 10 days). The infants in our study were slightly younger and therefore might not have had enough experience with the trochaic stress pattern of their environment to show a stronger

preference for the trochaic sequences. However, we would have expected them to show this preference in the second test session. As we point out later in this discussion, the absence of preference also in the second session may be caused by an interaction between familiarity with the stimuli, procedure and infants' memory.

Our findings are in line with Houston et al. (2007), who observed significant correlations between two testing sessions using a visual fixation paradigm (interval between test sessions was 1-3 days). However, in our experiment we increased the number of testing sessions to 3, as well as the number of days between the sessions (6–10 days). This might have caused our correlations to not be as strong as in the study by Houston et al. (2007). They are also partly in line with the study by Cristia et al. (2016), who reported rather weak evidence in the test-retest reliability for the HPP tested in one of the three labs (interval between test sessions was 13-15 days), with 2 out of 3 experiments showing negative correlations (9-month-old infants) and the other one close to zero (7-month-olds). It is interesting to note that in these studies they also tested an age range in which a potential developmental change takes place (in this case for word recognition).

Turning our attention to the rhythmic preferences, we did not observe a group preference for the trochaic sequences in session 1 or in session 2. The group preference appears in the third session. The results in session 2 were unexpected for our predictions. From previous studies (Friederici et al., 2007; Herold et al., 2008; Höhle et al., 2009; 2014) we expected infants to show a familiarity preference for the trochaic sequences across the three ages. However, the repeated testing and familiarity with the stimuli and procedure may interact in a complex manner with the infants' development and memory. It is possible that infants remembered the stimuli from the first test session given the strong priming context of the same test booth, the darkness, sitting on mother's lap, etc. (Rovee-Collier, 1999). This may cause infants to be less interested in either set of stimuli (recall that a significant decline in looking times is observed from the first to the second test

session). Thus, perhaps not all infants showed a familiarity preference in the second session, but a novelty effect towards iambic sequences or equal interest in both stimuli.

The last research question of this study addressed the relation between early speech perception measures and the later language development. We only found two significant correlations between the infants' HPP task performance and their later language development scores at 12 months of age: the difference scores and the trochaic looking times in session 2 correlated positively with the ELFRA-1 total test scores, suggesting that (a) infants who had a longer looking time to trochaic sequences scored better at the ELFRA-1 test and that (b) the larger the difference between trochaic and iambic items, the better the ELFRA-1 scores were. Thus, our results add evidence to the linkage between individual performance in speech perception tasks and later language development.

However, no correlation was found for the ELFRA-2 test. This suggests that infants who had higher vocabularies at 24 months of age were not better than their peers at discriminating between stress patterns at the age of 6 months. Our results are somewhat similar to the ones of Höhle et al. (2014), who also did not find a direct correlation between the difference scores and later language outcomes. However, they found a correlation between mean decrease iambic scores at 5 months and later language development, which we did not find in our sample in a post-hoc analysis. Our findings are in contrast with Weber et al. (2005), who found that infants who showed reduced neurophysiological responses to stress differences at 5 months had lower word production at the ELFRA-2 test. However, they did not correlate individual differences, but split infants into a group with low word production and a group with higher word production. We suggest that the individual differences observed at 12 months balance out at a later age, in the sense that infants who did not show a preference for trochees in the HPP also end up being successful language learners within or above the normal range.

Using a laboratory task as a tool in a diagnostic implies that the task must obtain reliable measures within individuals and that these measures can be interpreted as an indicator of the infant's capacities. In this respect, we provide some evidence that an HPP task might be reliable enough as a stable measurement for later language abilities, at least 12 months later. The fact that we found some evidence for test-retest reliability with a procedure and with stimuli that have shown correlations with later language performance in a previous study (Höhle et al. 2014) encourages further attempts to make these measurements suitable for diagnostic use. Efforts need to be taken to enhance the tools for analyzing individual data statistically, to combine several measures and dependent variables (e.g., EEG data with eye-tracking data), and to establish norms for these measures. There is still a long way to go and this requires cooperation among the various groups of researchers.

4. EXPERIMENT 2: WEIGHTING OF PROSODIC AND STATISTICAL CUES IN GERMAN ADULTS

4.1 Experiment 2a: German adult segmentation

4.1.1 Introduction

Research indicates that adults can segment words from fluent speech or an artificial language into words based on conditional statistical information, i.e., transitional probabilities (for a review, see Krogh, et al., 2012, or Saffran & Kirkham, 2017). However, according to the literature, prosodic cues also play a role in adult word segmentation. For example, Saffran, Newport and Aslin (1996a) found that English speakers were significantly better when the final vowel of the words in an artificial string was lengthened than when the lengthening occurred in the first syllable. Similar results were obtained with Spanish, French, and Dutch listeners (Toro, Sebastián-Gallés & Mattys, 2009; Tyler & Cutler, 2009). However, the same artificial string might be parsed differently according to the prosodic properties of the native language of the speaker. In Vroomen et al. (1998), Finnish, Dutch, and French adult listeners performed best when the phonological properties of the artificial language matched those of the native one (Finnish speakers profited from vowel harmony and word-initial stress, Dutch speakers from word-initial stress, and French speakers from neither of these). These outcomes suggest that prosodic cues interact and might easily override TPs in the segmentation of speech in English, Finnish, and Italian speakers (Vroomen et al., 1998; Gambell & Yang, 2006; Shukla et al., 2007; Fernandes et al., 2007; Langus et al., 2012). In contrast, when prosodic cues were absent, Italian speakers were capable of segmenting an artificial string based only on TPs (Langus et al., 2012).

The main aim of this chapter is to investigate how German-speaking adults weight statistical and prosodic information for word segmentation. We used a combination of methods in the present

experiment: after familiarizing adults with a string in which statistical and prosodic information indicated different word boundaries, we obtained behavioral responses as well as continuous online data of the participants' pupil dilation when doing the task. Thereby, we intended to gain further insights into the speech segmentation process. Pupillometry involves measuring the diameter change in the pupil, which not only regulates the influx of light but also constantly oscillates in response to activity of the nervous system resulting from psycho-sensory stimulation (Loewenfeld, 1958). A larger pupil diameter has been linked to a greater cognitive effort (Beatty & Lucero-Wagoner, 2000). In adults, pupil diameter is modulated by attention and cognitive load (Hess & Polt, 1960; Kahneman & Beatty, 1966; Laeng, Sirois & Gredebäck, 2012) and has been associated with cognitive processing and violations of expectation (Karatekin, 2007; Jackson & Sirois, 2009; Vogelzang, Hendriks & van Rijn, 2014; Fritzsche & Höhle, 2015; Tromp, Hagoort & , 2016).

A relevant study for our experiment is the one by Engelhardt, Ferreira and Patsenko (2010), who investigated processing effort by measuring participants' pupil diameters as they listened to sentences containing a temporary syntactic ambiguity. Interestingly, when prosodic structure conflicted with syntactic structure, pupil diameter increased. However, to our knowledge, there have been no statistical segmentation studies that have used pupillometry as a dependent measure, and therefore we do not have an informed hypothesis for the pupil reaction of the participants. Taking into account the information above, we can predict that adults will have a larger pupil dilation when presented with words that they have not segmented from the speech stream compared to words that they have extracted from the speech stream (familiar words).

Overall, the experiment was designed to gain insights into which cues German adults would base their segmentation on when presented with a string in which prosodic cues are pitted against statistical cues. To our knowledge, this is the first study that (a) tests German listeners in such a task and (b) tests this ability with both pupillometry and behavioral methods. Previous research has

shown that English speaking adults can make use of TPs to find word boundaries in a speech stream (Aslin et al., 1998) as well as prosodic cues like final lengthening (Saffran et al., 1996a). However, previous studies have shown that the trochaic bias seems to be a rather powerful mechanism in speakers of German (Bhatara, Boll-Avetisyan, Unger, Nazzi & Höhle, 2013). Note that it is likely that German speakers –like speakers of English (Cutler & Norris, 1988)– are biased toward treating a stressed syllable as the onset of a word. Therefore, we expected German monolingual adults to rely more strongly on prosodic cues than on TPs.

4.1.2 Participants

A total of 38 adult native speakers of German (4 males) recruited at the University of Potsdam were included in the sample. The age range was 19–40 years. Two additional adults were tested but not included due to a technical problem. For the pupillometry analysis, 5 of the 38 participants were excluded due to calibration problems (3) and not enough good data (2). Participants reported German as their first language and no history of hearing or speech problems. Participants who had been regularly exposed to more than one language while growing up were not included in the sample. This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participants.

4.1.3 Stimuli

A familiarization language string consisting of four disyllabic sequences (*gobu, tade, bido, puda*) was created. The syllables were taken from Thiessen and Saffran (2003) but adapted to the German phonotactics. Further, it was made sure that none of the syllables or their combinations formed a real German word. We used natural speech recordings for the stimuli to have more ecological validity concerning the prosodic information compared to synthesized speech. Thus, our stimuli were recorded in a sound attenuated booth by a female German speaker. She was asked to record

the stimuli in a lively voice as if she were talking to an infant (mild infant directed speech). To elicit each syllable with the acoustic properties of a stressed or an unstressed position, the speaker produced the syllables combined with the carrier syllable 'ke,' both in a stressed and in an unstressed position within a trochaic word (i.e., stressed position: *goke, take, bike, puke*; unstressed position: *kebu, kede, kedo, keda*). The syllables for the test trials were recorded separately with a monotonous voice. We tried to diminish coarticulation effects by cutting the recordings at zero crossing points with PRAAT. Coarticulatory effects are most prominent within syllables and may be present across boundaries, but to a reduced extend, which is the case of our stimuli (Rubertus & Noiray, 2018). Anticipatory coarticulation, for example, between /e/ and /b/ in a recording like /kebu/ is minimal in adult speech production compared to the coarticulatory effect between /k/ and /e/ (Noiray, Abakarova, Rubertus, Krüger & Tiede, 2018). In addition, in our material none of the crucial syllables contained a /k/ such that no specific type of segmenting the string could have been supported by coarticulation cues or their missing. The specific acoustic details of the stressed and unstressed syllables used to create the artificial language string are presented in Table 4. The acoustic properties of the syllables used to create the test trials are presented in Table 5.

Stressed syllables in German are typically associated with higher intensity and longer duration, and vowels have a larger relevance for carrying these stress cues than consonants (Dogil & Williams, 1999). Pitch is also increased in stressed syllables. The stressed syllables in our stimuli had a higher F0 mean and pitch peak than their unstressed counterparts (F0: 268 vs. 218 Hz, pitch peak: 283 vs. 228 Hz). Regarding intensity, the stressed syllables were on average 13.2 dB louder but 20 ms shorter than their unstressed counterparts. The fact that the unstressed syllables had a slightly longer duration than the stressed ones (285 vs. 251 ms) is due to final lengthening as the unstressed syllables were all produced in the second position of the string. This is a typical pattern if a disyllabic trochaic string is produced in isolation (see also Höhle et al., 2009). The syllables used in

the test phase were similar to the unstressed syllables in the string in duration, F0 mean, and pitch peak, but had similar intensity to the stressed syllables.

Table 4: Acoustic properties of the stressed and unstressed syllables in the familiarization string

Stressed syllables				
Syllable	**Duration (ms)**	**Intensity mean (dB)**	**Mean F0**	**Pitch Peak**
da	230,7	69,6	250,4	271,2
bu	234	71,2	283,8	303,8
de	227,9	71,4	257,3	264,8
do	231,2	71,05	280,6	298
Average	*230,9*	*70,8*	*268*	*284,5*
ke	*285*	*66,8*	*272,7*	*283,5*
Unstressed syllables				
Syllable	**Duration (ms)**	**Intensity mean (dB)**	**Mean F0**	**Pitch Peak**
pu	298	57,6	281	241,9
bi	219	59,2	212,2	253,9
go	230	59,5	183,2	194,4
ta	265	54	199,5	223
Average	*251,75*	*57,6*	*218,9*	*228,3*

Table 5: Acoustic properties of the test trial syllables

Syllable	**Duration (ms)**	**Intensity mean (dB)**	**Mean F0**	**Pitch Peak**
da	268	68,5	181,7	203,5
bu	203	73,3	220,8	238,2
de	251	69,8	185,8	211,3
do	291	68,4	189,2	220
pu	311	69,1	275,7	241,4
bi	235	68,9	211,8	254
go	232	69,3	183,3	194,2
ta	286	67,1	199,2	223,6
Average	*259,6*	*69,3*	*205,9*	*223,2*

All target syllables were cut at zero crossings to be merged into the artificial language string for the familiarization. There were no pauses and no coarticulation between the single syllables in the string. The string had a duration of 2 min 11 s and started with the stressed dummy syllable /ke/, which was then followed by the first syllable of the first word (see Table 4 for this syllable's acoustic properties). The dummy syllable only occurred once in the string and was used to prevent the segmentation of the string from being started with its initial syllable.

In the artificial language created, statistical cues (TPs between syllables) conflicted with prosodic cues (trochaic stress pattern). The TPs between syllables within the four disyllabic sequences considered as words (see above) were 1.0. The order of occurrence of these words in the string was varied such that the TPs across the four words were lower than the TPs within these words, ranging between 0.4 and 0.2. No immediate repetitions of the same word were allowed in the string. As for the prosodic cue, the second syllable of the words was consistently stressed throughout the string. Therefore, participants would segment the four words correctly if they relied on the TPs (we call these words statistical words in the following). In contrast, if the participants attended to the prosodic cues following a trochaic segmentation, they would segment words that adhere to the prosodically dominant trochaic pattern but that cross the TP boundaries (we call these prosodic words). See Table 6 for the two possible segmentations.

Table 6: Possible segmentations of the string

Expected segmentation based on TPs:	Expected segmentation based on prosodic information:
taDE/puDA/goBU/taDE/biDO/taDE/puDA/taDE…	ta/DEpu/DAgo/BUta/DEbi/DOta/DEpu/DAta/DE….

To compensate for potential differences in item frequency between statistical and prosodic words in the string, two of the statistical words in the familiarization string (*tade* and *gobu*) occurred twice as

often (90 times each) as the other two statistical words (*puda* and *bido*), which occurred 45 times each. Therefore, the prosodic words formed from the two frequent statistical words (*buta* and *dego*) occurred 45 times each in the string, just as often as the infrequent statistical words (*puda* and *bido)*.

Each test trial consisted of a word from one of three conditions. In the statistical condition, the two frequent statistical words (*tade*, *gobu*) and the two infrequent statistical words (*puda*, *bido*) were presented. In the prosodic condition, the prosodic words *buta*, *dego*, *depu*, and *dogo* were presented. In the non-word condition four disyllabic sequences were presented. These were combined from syllables that never occurred adjacently in the string (i.e., their TPs were 0.0: *bugo*, *pude*, *dobi*, and *tada*). The words had a duration of 500 ms on average. The test trials did not contain any prosodic information.[8] The test phase had a total of 36 trials and the total duration of the experiment was approximately 7 minutes.

4.1.4 Procedure

The experiment was conducted in a test booth in front of a computer screen and an eye-tracker. All participants filled out a consent form and a questionnaire about their linguistic background before taking the seat in front of the monitor of the eye-tracker. Participants had a button box in their hands throughout the whole experiment; they sat approximately 60–70 cm away from the display and tracking was remote. The experimental session started with an eye calibration using a 5-point sequence, which consisted of a grey background with white points. After the calibration, the written instructions for the task were presented on a grey screen. Participants were told that they were going to listen to a string of words for two minutes and that after exposure they would have to answer some questions about the words in the language. After the participant had read the instructions, there was the possibility to ask the experimenter questions. The participant had to press a button to

[8] However, they might not be completely free of prosodic cues. They were naturally recorded so they might contain onset or offset cues. Here we refer to prosodic cues related to stressed or unstressed syllables.

start the experiment. In the familiarization phase, the screen was grey and a black loudspeaker icon was shown in the center. Participants were instructed simply to listen to the speech and to focus on the screen all the time. All the participants were familiarized with the same string. Immediately after the familiarization, the test phase started, which consisted of a total of 36 trials (each word was presented three times) and had a total duration of approximately 5 minutes. The order of the test trials was randomized for each participant. Each test trial consisted of a single word, played while a loudspeaker icon was shown on the screen. After the presentation of each word participants had to decide whether the disyllabic word that was presented acoustically had been present in the previous familiarization string. They were encouraged to answer as quickly as possible, and to make their best guess if unsure. To provide the answer, the words "yes" and "no" were shown (in black on a grey background) on either side of the screen and the participants gave their response by pressing one of two possible buttons (right and left buttons) on the button box. The next test trial started (1) when the participant had pressed a button and (2) when the participant was looking at the screen (the eye-tracker was detecting the eyes). If this was not the case, an attention getter appeared at the center of the screen.

Stimulus presentation was programmed using PsyScope software, which collected both the behavioral and the pupillometry responses. All visual stimuli were shown on a 17" (1280 x 1024) TFT screen with a resolution of 300 x 300 pixels. Pupil diameter was recorded with a Tobii 1750 binocular corneal reflection eye-tracker with a temporal resolution of 50 Hz.

4.1.5 Results

To control for frequency of occurrence in the familiarization string, only the two infrequent statistical words (*puda*, *bido*), the two prosodic words formed by the boundaries of the two frequent statistical words (*buta*, *dego*), and the four non-words (*bugo*, *pude*, *dobi*, *tada*) were included in the analysis. The higher number of non-words was controlled for in the analysis.

4.1.5.1 Behavioral results

The number of 'yes' responses (i.e., the decision that the presented item had been part of the familiarization string) was used as the outcome measure. Participants responded 'yes' to 46.5% of the non-word trials, to 68% of the prosodic word trials, and to 47.4% of the statistical word trials (see Figure 4).

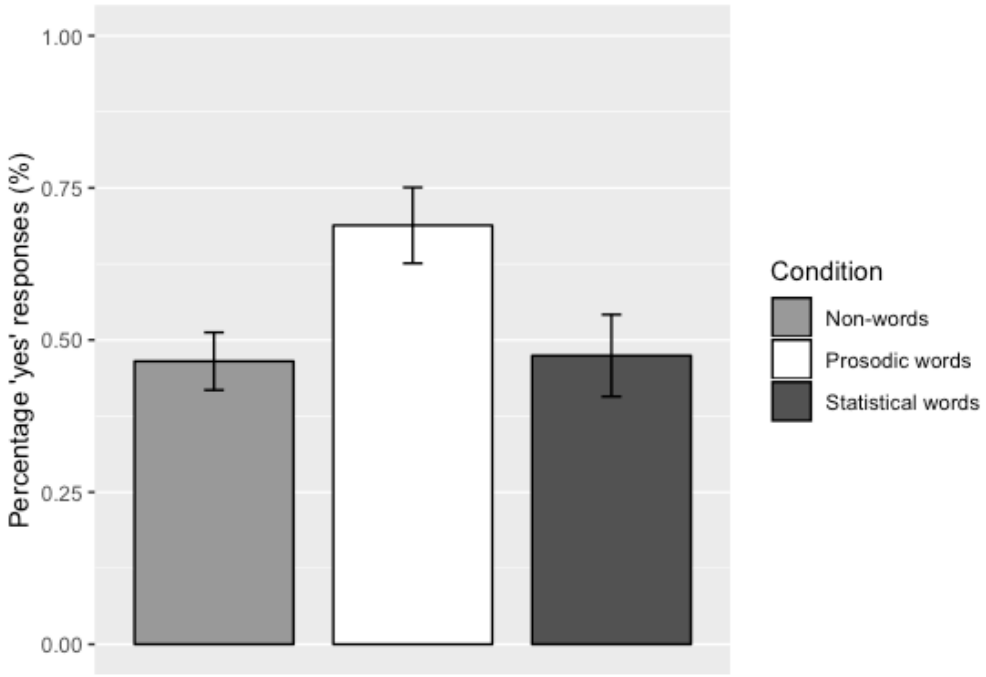

Figure 4: Percentages of 'yes' responses in each condition. The error bars represent the standard error.

We employed general linear mixed effects models with random factors for participants and items using the *glmer* function in the *lme4* R package for statistical analyses. Graphs were generated using the package *ggplot2* (Wickham, 2009) and the contrasts were coded with the *MASS* package (Venables & Ripley, 2002). In the model, condition (*Condition*) was entered as a fixed effect with three levels: prosodic word (*Prosodic*), non-word (*Nonword*), and statistical word (*Statistical*). We used a sliding contrast for successive comparisons between the conditions. We coded the contrast so that the prosodic condition was compared to the two other conditions, while non-words and statistical words were not compared (no difference was expected between these two conditions).

Participants and *items* were included as random effects in the model. First, we ran a model against chance level performance. Secondly, we fit the maximal model to the data. The complete output of the two models is provided in Table 7 (against chance) and Table 8 (maximal model). The estimates (*ß*) indicate the logit-transformed number of 'yes' responses. The analysis against chance level indicates a significant effect only for the prosodic word condition (*Prosodic*, *ß* = 0.87, p = .02), showing that participants only performed above chance level when presented with a prosodic word. The results provided in Table 8 show a significant difference between non-words and prosodic words (*Nonword - Prosodic*, *ß* = -1.04, *p* = .02), the negative *ß* suggesting that participants gave fewer 'yes' responses in the non-word condition than in the prosodic word condition. Moreover, there was also a tendency for a significant difference between the prosodic and the statistical words (*Prosodic - Statistical*, *ß = 1*, *p* = .052). The positive *ß* reflects that the participants tended to give more 'yes' responses when presented with prosodic words compared to statistical words.

Table 7: Model against chance

Fixed Effects	***ß***	**SE**	***z*-score**	***p*-value**
Statistical	-0.13	0.38	-0.34	.72
Prosodic	0.87	0.38	2.28	.02 *
Nonword	-0.16	0.27	-0.61	.53
Random Effects	**Variance**		**SD**	
id (Intercept)	0.30		0.54	
item (Intercept)	0.23		0.47	

Table 8: Maximal model

Fixed Effects	***ß***	**SE**	***z*-score**	***p*-value**
Grand mean (intercept)	0.19	0.21	0.89	.37
Prosodic - Statistical	1	0.52	1.91	.052
Nonword - Prosodic	-1.04	0.45	-2.29	.02 *
Random Effects	**Variance**		**SD**	
id (Intercept)	0.30		0.54	
item (Intercept)	0.23		0.47	

** = significant, p < .05*

4.1.5.2 Pupillometry results

Both eyes were tracked but only data obtained when at least one eye could be recorded entered the analysis (92 % of the trials). Blinks were eliminated (0.10 % of the total data points). The Task-Evoked Pupillary Response (TEPR)[9] was calculated and taken as the main dependent variable. Then, the TEPR measure was corrected using a 200 ms baseline[10] for each individual trial for each item. TEPR measures were averaged across all trials within each condition. A 3-second window starting at the onset of each word was investigated (the pupil takes 1.2 seconds on average to reach its maximum diameter; Just & Carpenter, 1993). Successful trials were defined as those containing pupil measures from at least half the length of the trial. Those participants who did not reach a threshold of 50% of successful trials were excluded from the analysis (2 participants). A total of 33 participants were included in the analysis. Figure 5 illustrates the response dynamics of the pupil during the 3 seconds averaged across trials. Importantly, the pupil size increases in response to the acoustic information in all three conditions.

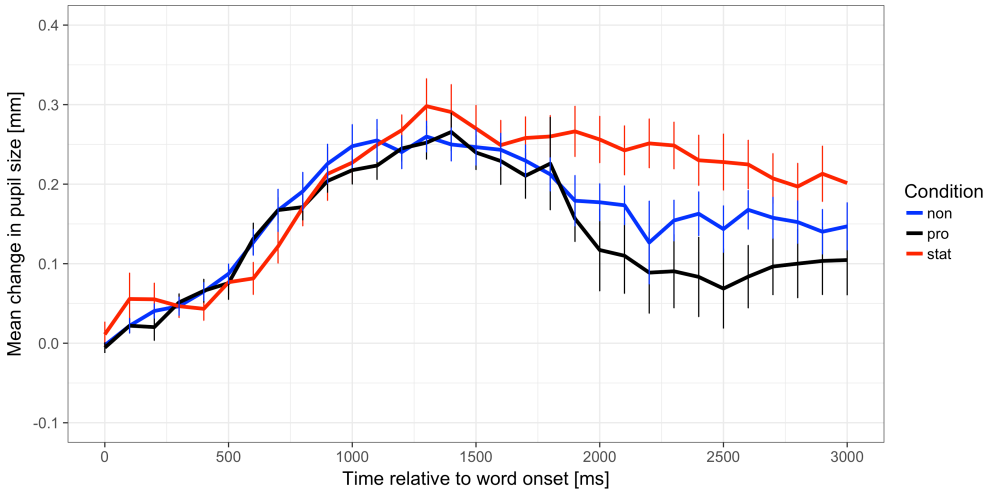

Figure 5: Time course of the pupil size changes in the different conditions

[9] TEPR is defined as subtle changes in pupil size (typically less than .05 mm) which indicate intensity and online resource demands of numerous cognitive processes (Beatty & Lucero-Wagner, 2000).

[10] Given that the baseline pupil size for each participant can vary between participants and trials, a common approach is to baseline-correct all values. For this purpose, pupil size during a short time interval before the onset of the experimental manipulation is averaged to form the baseline (Hepach & Westermann, 2016).

The differences between the pupil size changes from all trials in the different conditions were analyzed using a linear mixed effects model using the *lmer* function in the *lme4* R package. Graphs were generated using the package *ggplot2* (Wickham, 2009) and the contrasts were coded with the *MASS* package (Venables & Ripley, 2002). The windows of analysis could not be decided *a priori* because of the lack of previous studies. Therefore, we applied the model in different 500 ms windows during the 3-second time window investigated. There were a total of 6 windows (0–500 ms, 500–1000 ms, 1000–1500 ms, 1500–2000 ms, 2000–2500 ms, and 2500–3000 ms).

The model we fitted followed the recommendation by Matuschek, Kliegl, Vasisth, Baayen and Bates (2017) to specify a maximal random effects structure for confirmatory hypothesis testing without losing power. We checked for the random component structure with the *RePsychLing* R Package (Bates, Kliegl, Vasishth & Baayen, 2015) and fitted the maximal model that best explains our data. In the model, condition (*Condition*) was entered as a fixed effect with three levels: prosodic word (*Prosodic*), non-word (*Nonword*), and statistical word (*Statistical*). Following the behavioral analysis, we used a sliding contrast for successive comparisons between the conditions. We coded the contrast so that the prosodic condition was compared to the two other conditions, while non-words and statistical words were not compared. *Participant* was included as a random factor. The factors *Age* and *Gender* were excluded from the model because they did not improve the model fit to the data. To explore whether there were any effects of the button press in our sample, we added the interaction between *Condition* and *Button Response.* The factor *Button Response* was coded as "yes/no" depending on the button that the participant had pressed. The same model was applied to all time windows.

The model revealed significant results in two time windows (2000–2500 ms and 2500–3000 ms). The output of the two models is presented in Tables 9 and 10. In the time window between 2000

and 2500 ms (Table 9), the prosodic condition elicited significantly smaller pupil size changes compared to the statistical condition (*Prosodic - Statistical*, *ß* = -0.11, *t* = -3.87, *p* < .001).

Table 9: Maximal Model for the 2000-2500 ms window

Fixed Effects	***ß***	**SE**	***t*-value**	***p*-value**
Intercept	0.16	0.02	8.26	< .001*
Prosodic - Statistical	-0.11	0.03	-3.87	< .001*
Nonword - Prosodic	0.04	0.02	1.77	.07
Button response No - Yes	0.01	0.02	0.55	.58
Pro - Stat* Button	0.11	0.06	1.87	.06
Non - Pro* Button	-0.08	0.05	-1.59	.11
Random Effects	**Variance**		**SD**	
id (Intercept)	0.09		0.09	
Residual	0.07		0.27	

** = significant, p < .05*

Table 10: Maximal Model for the 2500-3000 ms window

Fixed Effects	***ß***	**SE**	***t*-value**	***p*-value**
Intercept	0.14	0.02	6.92	< .001*
Prosodic - Statistical	-0.11	0.03	-3.46	< .001*
Nonword - Prosodic	0.06	0.02	2.20	.02*
Button response No - Yes	0.09	0.02	0.40	.68
Pro - Stat* Button	0.08	0.06	1.33	.18
Non - Pro* Button	-0.06	0.05	-1.07	.28
Random Effects	**Variance**		**SD**	
id (Intercept)	0.01		0.10	
Residual	0.08		0.28	

** = significant, p < .05*

The estimate *ß* between the prosodic and the non-word condition was positive, which indicates that overall there were larger pupil size changes in the non-word condition than in the prosodic condition. However, this difference did not reach significance (*Nonword - Prosodic*, *ß* = 0.04, *t* = 1.77, *p* = .07). Interestingly, there was a tendency toward significance for the interaction between *Button Response* and the difference between prosodic and statistical condition (*ß* = 0.11, *t* = 1.87, *p*

= .06), suggesting that the statistical condition elicited greater pupil size changes when the answer pressed was "no" than when participants pressed "yes" (see Figure 6).

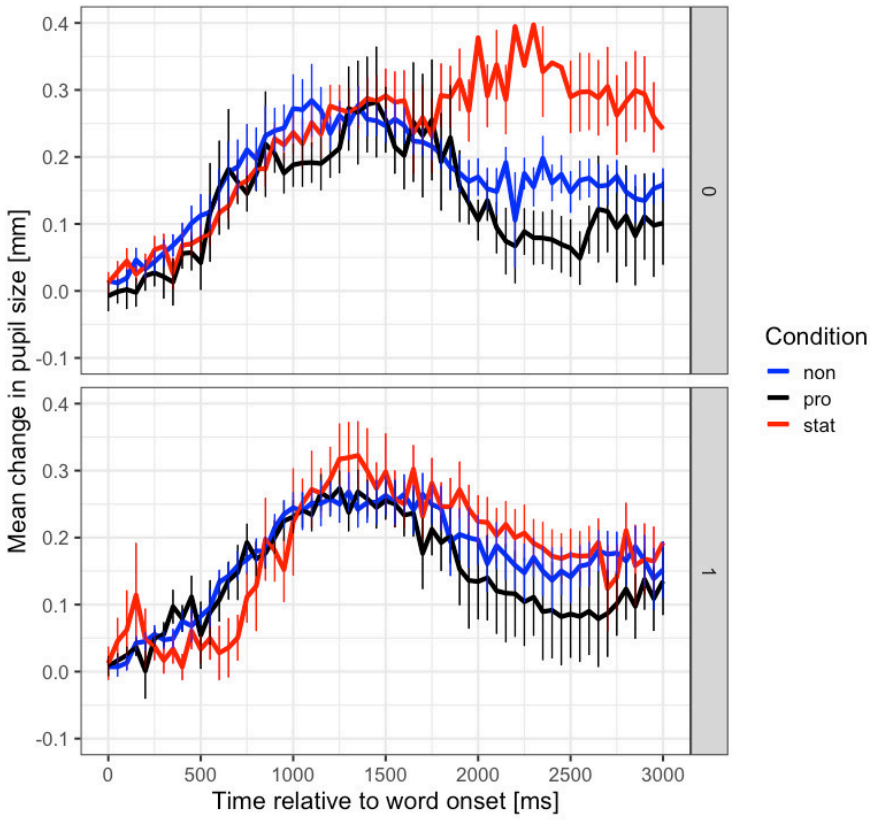

Figure 6: Time course of the pupil size changes in the different conditions by button (top graph shows "No" responses and bottom graph "yes" responses).

In the time window between 2500 and 3000 ms (Table 10), the output shows that there is a significant difference between the prosodic and the statistical condition (*Prosodic - Statistical*, β = -0.11, t = -3.46, $p < .001$), which indicates that overall there were less changes in pupil size in the prosodic condition compared to the statistical condition. Additionally, there was a significant difference between the prosodic and the non-word condition (*Nonword - Prosodic*, β = 0.06, t = 2.2, p = .02). In contrast to the previous time window, the interaction between the button press and the conditions was not significant (β = 0.08, t = 1.33, p = .18).

In addition, we ran a further analysis with the non-word condition as baseline instead of the prosodic condition so that the non-word trials were compared to the prosodic and statistical word

trials. The reason why we did this was to assess the status of the statistical words, namely whether they were considered as non-words or words by the participants. The same model as before was applied to all time windows. The model revealed significant results in two time windows (2000–2500 ms and 2500–3000 ms). The output of the two models is presented in Tables 11 and 12. In the time window between 2000 and 2500 ms (Table 11), the non-word condition elicited significantly smaller pupil size changes compared to the statistical condition (*Nonword - Statistical*, *ß* = -0.07, *t* = -2.79, *p* < .01). Similarly, in the later time window, between 2500 and 3000 ms (Table 12), the difference between statistical and non-words showed a tendency to significance (*Nonword - Statistical*, *ß* = -0.04, *t* = -1.84, *p* = .06).

Table 11: Maximal Model for the 2000-2500 ms window (non-word condition as baseline)

Fixed Effects	***ß***	**SE**	***t*-value**	***p*-value**
Intercept	0.16	0.02	8.26	< .001*
Nonword - Statistical	-0.07	0.02	-2.79	< .01*
Prosodic - Nonword	-0.04	0.02	-1.77	.07
Button response No - Yes	0.01	0.02	0.55	.58
Non - Stat* Button	0.02	0.05	0.55	.57
Pro - Non* Button	0.08	0.05	1.59	.11
Random Effects	**Variance**		**SD**	
id (Intercept)	0.009		0.09	
Residual	0.07		0.27	

Table 12: Maximal Model for the 2500-3000 ms window (non-word condition as baseline)

Fixed Effects	***ß***	**SE**	***t*-value**	***p*-value**
Intercept	0.14	0.02	6.92	< .001*
Nonword - Statistical	-0.04	0.02	-1.86	.06
Prosodic - Nonword	-0.06	0.02	-2.20	.02*
Button response No - Yes	0.00	0.02	0.4	.68
Non - Stat* Button	0.02	0.05	0.45	.64
Pro - Non* Button	0.06	0.05	1.07	.28
Random Effects	**Variance**		**SD**	
id (Intercept)	0.01		0.10	
Residual	0.08		0.28	

** = significant, p < .05*

In short, our results show a higher number of 'yes' responses to prosodic words compared to the other two conditions, suggesting that the prosodic words were more often recognized as having appeared in the string than the statistical and the non-words. Additionally, the pupillometry data of the participants showed that German listeners had greater changes in pupil size for the statistical words compared to the prosodic words. In a later time window, participants also showed greater changes in pupil size for non-words compared to prosodic words. This suggests that adults treated prosodic words different from non-words and statistical words, which is consistent with our behavioral results. Furthermore, we showed that participants are treating statistical words differently than non-words, which suggests that participants have somehow tracked the TPs and segmented the statistical words. However, before drawing any strong conclusions, there is a further consideration that we need to address. In previous similar studies, a string with an amplitude ramp was often used at the beginning and end of the string, because the first two syllables of the string can have an impact on participants' segmentation. Given the fact that our familiarization string started with the dummy syllable /ke/, the question arises whether this syllable might have had an effect on the segmentation strategy used by the participants, namely whether it prevented segmentation strategies that made use of the first syllable in the string. Thus, Experiment 2b was conducted to rule out any possible effects of this methodological question. In the following experiment we added an amplitude ramp at the beginning and the end of the string.

4.2 Experiment 2b: German adult segmentation with amplitude ramp

4.2.1 Participants

A total of 38 adult native speakers of German (3 males) recruited at the University of Potsdam were included in the sample. The age range was 18–33 years. Two additional adults were tested but not included because they reported being bilingual. All participants reported German as their first

language and no history of hearing or speech problems. None of them participated in Experiment 2a. Written informed consent and detailed information about language background was obtained from all participants.

4.2.2 Stimuli

The stimuli consisted of the same familiarization string as used in Experiment 2a, but two 5-second amplitude ramps were added with the Audacity functions “Fade in” and “Fade out” (Audacity Team, 2012): an increasing one at the beginning of the string and a decreasing ramp at the end. The dummy syllable /ke/ was again the first syllable of the string. The test phase stimuli were identical to those in the test phase of Experiment 2a.

4.2.3 Procedure

The experimental procedure differed slightly from the previous experiment (Experiment 2a) because pupillometry data were not obtained. The experiment was conducted in a test booth in front of a computer. Participants wore headphones throughout the whole experiment. The written instructions for the task were presented on the screen before starting the experiment. Participants were told that they were going to listen to a string of words for two minutes and that after exposure they would have to answer some questions about the words in the language. After the participant had read the instructions, there was the possibility to ask the experimenter questions. The participant had to press a key to start the experiment. In the familiarization phase, a loudspeaker icon was shown in the center of the screen. Participants were instructed simply to listen to the speech. All the participants were familiarized with the same string. Immediately after the familiarization the test phase started, which consisted of a total of 36 trials (each word was presented three times) and had a total duration of approximately 5 minutes. The order of the test trials was randomized for each participant. Each test trial consisted of a single word, played while a

loudspeaker icon was shown on the screen. After the presentation of each word, participants had to decide whether the disyllabic word that was presented acoustically had been present in the previous familiarization string. They were encouraged to answer as quickly as possible, and to make their best guess if unsure. To provide the answer, the words "yes" and "no" were on either side of the screen and the participants gave their response by pressing one of two possible keys (right and left Alt keys) on the keyboard. The next test trial started once the participant had pressed a key or after 400 ms after word-offset.

4.2.4 Results

As in Experiment 2a, only the two infrequent statistical words (*puda*, *bido*), the two prosodic words formed by the boundaries of the two frequent statistical words (*buta*, *dego*), and the four non-words (*bugo*, *pude*, *dobi*, *tada*) were included in the analysis. The higher number of non-words was controlled for in the analysis. The number of 'yes' responses (i.e., the decision that the presented item had been part of the familiarization string) was used as the outcome measure. Participants responded 'yes' to 55.9 % of the non-word trials, to 76.7 % of the prosodic word trials, and to 57.6% of the statistical word trials (see Figure 7). 1.53 % of the total responses were time-out responses (participants did not press a response key within 400 ms after word-offset) and were not included in the analysis.

We employed general linear mixed effects models with random factors for participants and items using the *glmer* function in the *lme4* R package for statistical analyses. Graphs were generated using the package *ggplot2* (Wickham, 2009). The contrasts were coded with the *MASS* package (Venables & Ripley, 2002). The main purpose of this experiment was to evaluate whether having an amplitude ramp at the beginning and at the end has an impact on participants' segmentation performance. Therefore, in the following analysis we included the data from both experiments in the same model.

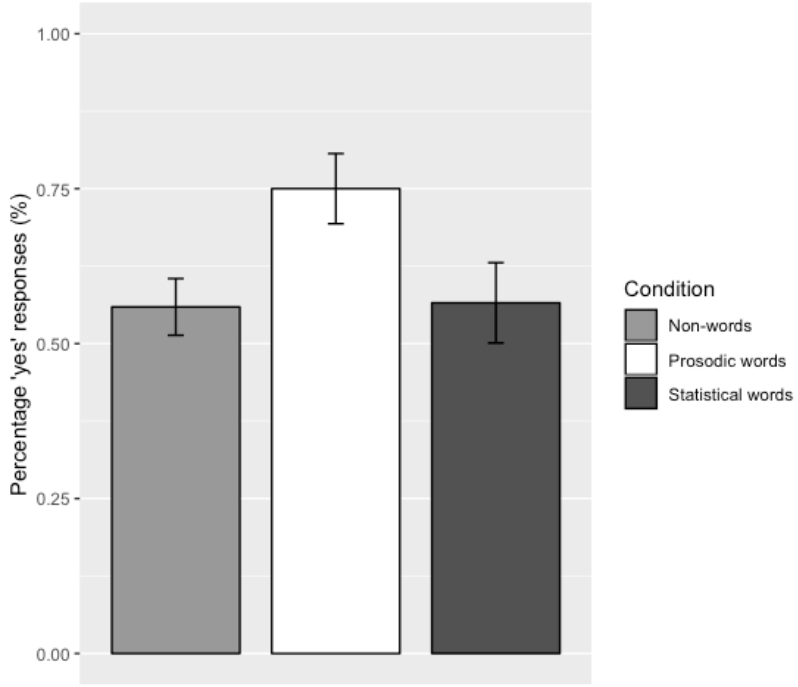

Figure 7: Percentages of 'yes' responses in each condition when the string had a ramp. The error bars represent the standard error.

We fitted the same maximal model to the data as in the previous analysis of the behavioral data (Section 4.1.5.1), but we added the factor *Ramp* as a fixed effect with two levels: *Ramp* and *No Ramp*. These two levels corresponded to the two different experiments: Experiment 2a without a ramp and Experiment 2b with a ramp. Thus, we basically checked whether there is a significant difference between the two experiments. The complete output of the model is provided in Table 13. The estimates (*ß*) indicate the logit-transformed number of 'yes' responses. The analysis shows no effect of ramp (*Ramp - No Ramp*, $ß = 0.41$, $p = .13$) and no interaction between *Ramp* and *Condition* ($p = .95$), meaning that there is no significant difference between the participants' performance in the two experiments.

Table 13: Model from Experiment 2a and 2b with the the ramp effect

Fixed Effects	***ß***	**SE**	***z***	**p**
Grand mean (intercept)	0.41	0.15	2.6	.009*
Prosodic - Statistical	1	0.40	2.49	.01*
Nonword - Prosodic	-1.02	0.34	-2.94	.003*
No Ramp - Ramp	0.41	0.28	1.48	.13
Random Effects	**Variance**		**SD**	
id (Intercept)	0.11		0.33	
item (Intercept)	0.27		0.52	

** = significant, $p < .05$*

Analyzing both experiments together increased the statistical power of the model. Results from this analysis show that there is not only a significant difference between the prosodic word condition and the non-word condition (*Nonword - Prosodic*, *ß = -1.02*, *p < .01*), but the difference between the prosodic word condition and the statistical word condition is also significant (*Prosodic - Statistical*, *ß = 1*, *p* = .01). Thus, this confirms the difference between non-words and prosodic words observed in Experiment 2a and strengthens the tendency reported between prosodic and statistical words.

4.3. General discussion

In these experiments we first wanted to investigate whether adult German speakers favor statistical or prosodic cues in their segmentation of a continuous speech string. Secondly, we wanted to validate our stimuli and the procedure, which included three conditions in the test phase. Our results showed a higher number of 'yes' responses to prosodic words compared to the other two conditions. This suggests that German monolingual adults rely more strongly on prosodic cues, namely stress, compared to TPs when segmenting a continuous syllable string. In the second experiment we checked whether an intensity ramp would have an impact on participants' segmentation performance. Hence, we added an amplitude ramp of 5 s at the beginning and at the end of the string from the previous experiment. Again, participants showed a higher number of 'yes' responses to prosodic words compared to the other two conditions. As in Experiment 2a, the prosodic words were better recognized as having appeared in the string than the statistical and the non-words, meaning that the first and last items of the string did not have a significant effect on participants' segmentation. Although the two experiments were not exactly equal in terms of methodology, analyzing both experiments together increased the statistical power. Our findings again show that German monolingual adults rely more strongly on stress than on TPs when segmenting the presented continuous syllable string.

Our results corroborate the findings from other experimental settings that show that a trochaic bias seems to be a rather powerful perceptual mechanism in speakers of German (Bhatara et al., 2013). It is likely that German speakers –like speakers of English (Cutler & Norris, 1988)– are biased toward treating a stressed syllable as the onset of a word. In addition, these results are in line with previous studies with speakers from other languages which show that prosodic cues aid segmentation (Saffran et al., 1996a; Toro et al., 2009; Tyler & Cutler, 2009) and that statistical cues seem to be easily overridden by prosodic cues in adults (Shukla et al., 2007; Fernandes et al., 2007; Langus et al., 2012).

Additionally we obtained pupillometry data of the participants. To our knowledge, this is the first pupillometry study that compares the exploitation of statistical and prosodic cues for word segmentation by adults. German listeners showed greater changes in pupil size when hearing the statistical words compared to both prosodic and to non-words. In addition, prosodic words elicited significantly greater changes in pupil size compared to non-words (in the later time window). Considering the previous literature, we interpret our results in terms of cognitive effort, meaning that it was easier and cognitively less demanding for the participants to make a decision about the prosodic words compared to the statistical or the non-words which is consistent with our behavioral results. The significant difference in pupil size changes between the statistical and the non-words can be explained in terms of TPs. Since non-words consisted of syllables appearing in the string, it is possible that the syllables were still recognized, but the decision (behavioral response) was easier compared to the statistical words because the TPs between the syllables were 0. This is consistent with Endress and Mehler (2009b), who showed that Italian speakers could not segment words from fluent speech using distributional information even if they could demonstrably track it.

It is important to highlight the fact that the adult pupillary responses, as well as the behavioral decisions, could have been influenced by word frequency. Although the different words at test were

presented the same number of times (each of the words appeared three times), the words in the string had different frequencies because of the control of the TPs within- and between-words. Two of the statistical words, *tade* and *gobu,* occurred twice as often (90 times) as the two other statistical words (45 times), *puda* and *bido.* This way, the prosodic words were formed from the between-word boundaries of the two statistical frequent words. Although only the infrequent statistical words were included in the analysis, the frequent statistical words were also presented at test. A potential saliency of the frequent statistical words in the string (90 times) could have affected the perception of the statistical words compared to the other two conditions in the sense that frequent statistical words might have been noticed more readily during familiarization and been memorized better than the less frequent ones. This may have negatively affected the recognition of the statistical words during the test phase. However, we ruled out this possibility in a post-hoc analysis: there was no significant difference between frequent and infrequent statistical words (t = -0.57, p = .56).

The fact that all words were presented more than once during the test phase could have also had a boosting effect of participants' familiarity perception in the different conditions. In fact, there is evidence that sequential effects are present in Lexical Decision Tasks[11] (Meyer & Schvaneveldt, 1971) and that there is a local influence of the item frequency of consecutive trials. Perea & Carreiras (2003) found evidence in a lexical decision task that responses for both low-frequency words and non-words were influenced by the frequency of the precursor word, but that high-frequency words were less affected. The authors conclude that participants shift their response criteria on a trial-by-trial basis, depending on the characteristics (item frequency and lexical status) of the immediate preceding trial. Recall that in our experiment both infrequent and frequent statistical words were included in the test phase and that non-words did not occur in the string (TPs were 0.0). Thus, following the argument of Perea & Carreiras (2003), it might be that the test trials

[11] A lexical decision task is a behavioral method where the participant needs to make a decision about whether combinations of letters are words or not.

containing statistical words were less affected by a sequential effect compared to the non-words trials, because the total amount of test trials for the statistical condition was larger.

Finally, it is important to note that the button presses could have had an influence on the pupil dilation of the participants, as observed in Table 9 and Figure 6. In fact, some studies have found that the requirement of a button press can increase pupil dilation (Privitera, Renniger, Carney, Klein & Aguilar, 2010). In our results, the degree of pupil change was modulated by the type of button press only in the statistical condition, i.e., it was larger with 'no' responses. It has been argued that in lexical decision experiments, subjects have difficulty in responding "no" to non-words or written pseudo-words which are pronounced exactly like English words, for example "brane" (Coltheart, Besner, Jonasson & Davelaar, 1979). Following this reasoning, it is possible that the participants in our study also experienced difficulty with the statistical words. Our participants responded "no" to 53.5 % of the test trials –in which a statistical word was presented– and at the same time they showed larger pupil size changes compared to the prosodic words in the pupillometry data. This suggests that the syllables were familiar to the participants, but according to the segmentation strategy they used (prosodic information), statistical words were considered non-words or pseudo-words. A similar pattern of results was obtained with non-words, which were formed by syllables that appeared in the string but never occurred together. Participants also showed a high percentage of 'no' responses and larger pupil size changes compared to the prosodic words.

Given the fact that Experiments 2a and 2b have revealed that adults show a strong weight of prosodic cues in segmenting the materials used in this study, we now ask how these results from adult subjects compare to the mechanisms available to infant learners, for whom word segmentation is a critical component of native language acquisition. Importantly, the fact that adults use prosodic cues does not mean that infants will show the same behavior. We tested German-learning 9- and 6-month-olds in a similar design to that used by Thiessen and Saffran (2003) and Johnson and

Juszcyk (2001). If infants rely more strongly on TPs, as English-learning infants do in early stages of word segmentation, they will behave differently from the German adults and segment the four statistical words from the string. However, if infants follow the German dominant trochaic stress pattern, they will perform like the German adults by segmenting the prosodic words that cross the TP boundaries from the string. A third possibility is that infants treat both prosodic and statistical information as important and they show no preference at all.

5. EXPERIMENT 3: WEIGHTING OF SEGMENTATION CUES IN 9-MONTH-OLD GERMAN INFANTS

5.1 Experiment 3a: Word segmentation at 9 months

5.1.1 Introduction

A number of cues have been identified that support infants' speech segmentation, but extensive research has focused on two types of information as playing a central role in the earliest steps of speech segmentation: transitional probabilities (TPs) (e.g., Saffran et al., 1996b) and prosodic cues (e.g., Jusczyk et al., 1993). Evidence for statistical learning has been found in infants as young as 5 months (Thiessen & Erickson, 2013) and throughout early infancy (e.g., Thiessen & Saffran, 2003, 2007). However, infant word segmentation is also affected by prosodic cues like lexical stress or phrasal prosody (e.g., Johnson & Jusczyk, 2001; Höhle et al., 2009). Recall that stressed syllables are preferred as the beginnings of words with following unstressed syllables by infants learning languages with stress-based rhythmic properties such as German.

Some studies have investigated whether one of the cues shows any dominance over the other and whether one of the cues is used earlier in the development than the other (e.g., Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). Crucially, Thiessen and Saffran (2003) observed a developmental shift in cue reliance during the second half of the first year of life: while English-learning 7-month-olds relied more strongly on statistical cues in their segmentation performance, 9-month-olds were more strongly guided by the prosodic cues. Based on these findings the authors argue for an initial dominance of statistical cues over prosodic cues, which turns into a stronger weight of prosodic cues with growing language experience. They explain this change in cue relevance by a crucial difference in the status of the cues: TPs are present across languages and can be used without previous language knowledge but prosodic cues are language-specific. Their proposal for this

developmental shift was supported by further findings showing that 5-month-old English-learning infants rely more strongly on statistical cues (Thiessen & Erickson, 2013) and that 11-month-olds rely more strongly on prosodic cues (Johnson & Seidl, 2009). To the best of our knowledge, the relation of statistical and prosodic cues has not been tested in infants in languages other than English before. However, if Thiessen and Saffran's (2003) argument about statistical cues being language independent is correct, a similar developmental shift in cue reliance should be observed across languages –at least across languages in which word stress patterns provide reliable cues for word segmentation. German is an interesting case to test this hypothesis since German infants have shown indications of being sensitive to relevant language-specific prosodic properties already at the age of 4 months (Friederici et al., 2007; Höhle et al. 2009). It is relevant to note that German may have more frequent trochaic patterns than English (Delattre, 1963). Further, the inflectional system of German is richer than that of English and many inflectional endings that are added to monosyllabic words in German lead to disyllabic trochees. Altogether, it is possible that German infants become sensitive to prosodic cues earlier than their English peers.

As seen in Section 2.1.4 and Section 2.2.3, both SL and prosodic word segmentation have been associated with later language outcomes (e.g., Evans et al., 2009; Junge et al., 2012; Seidl & Cristia, 2012; Arciuli & Simpson, 2012). We wanted to further explore this link and examine whether the use of one cue or the other is related to further language acquisition. Thus, we followed up with the infants that participated in this experiment and obtained language outcome measures at later ages.

The main goal of the next two chapters is to shed more light on the initial reliance on TPs for segmenting words from speech as a universal stage by testing German-learning infants. Do German-learning infants also show an initial dominance of statistical cues over prosodic cues? If so, is there a shift into a stronger weight of prosodic cues with growing language experience? Is the use of one of the two types of information related to later language development? Based on the previous

results that German-learning infants show a trochaic bias at a very early age (Friederici et al., 2007; Höhle et al. 2009, 2014), our hypothesis was that German-learning 9- and 6-month-old infants would weight prosodic cues more heavily than TPs when the two types of cues are in conflict and that this reliance would be related to later language outcomes.

This question was investigated in an experiment that was methodologically similar to Johnson and Juszcyk (2001) and to the Thiessen and Saffran studies (2003, 2007). However, it departs from this previous work in an important aspect since three conditions were used at test (instead of two): words based on TP information, words based on prosodic information, and non-words (disyllabic sequences combined from syllables that never occurred adjacently in the familiarization string). This third condition was included to help with the interpretation of infants' direction of preference, which is important when making inferences about infants' dominant processing mechanism.

As considered in Section 2.1.5, predicting infants' direction of preference in the HPP *a priori* is quite difficult. When comparing the strength of different input cues the direction of the effect matters: without knowing which direction of preference (novelty or familiarity) infants are likely to show in a given experiment, it is impossible to detect which cue was most heavily weighted. In the research on statistical learning with artificial languages a novelty effect was usually found (e.g., Saffran et al., 1999; Curtin, Mintz & Christiansen, 2005; for a meta-analysis, see Black & Bergmann, 2017), meaning that the infants showed longer listening times during the test to those items that had low TPs in the familiarization string compared to items with high TPs. Remarkably, Black and Bergmann (2017) reported in their meta-analysis that more mature infants might show a different direction of preference (e.g., from a preference for non-words to a preference for words) in word segmentation experiments.

Adding conflicting cues (stress cues against TPs) can change the direction of preference. Thiessen and Saffran (2003) found a shift in the preference direction: 9-month-olds showed shorter looking

times for words over part-words[12] (familiarity preference) when the language string had no prosody cues, but longer looking times for words (novelty preference) when stress cues were added. The authors argue that the familiarity preference observed in the first experiment was due to difficulty in making the match between the stressed familiarization syllables and the monotonic test syllables. The same shift in the preference direction also occurred in the study by Johnson and Jusczyk (2001). When only TPs were present, 8-month-olds showed shorter looking times for words over part-words (familiarity preference) and when stress cues were added to the string infants showed longer looking times for words over part-words (novelty preference). In our experiment, the solution to this issue was the addition of the non-word condition at test, which could help to interpret the direction of the results. The looking times to the non-words, which are novel for the infants, can serve as a baseline for infants' looking behavior to novel stimuli and can be compared to the two cued conditions. A limitation of the additional third condition at test is the reduction of the number of trials per condition, and therefore loss of statistical power. In our experiment 12 trials were presented at test (4 per condition), whereas in other studies with only two conditions at test, 6 trials per condition could be presented.

In sum, the present chapter aims to shed more light on the weighting of potential segmentation cues by German infants. We conducted four experiments with 9-month-old German-learning infants to explore whether they show a dominance of prosodic cues over statistical cues, as observed in Thiessen and Saffran (2003) for English-learning 9-month-olds. Further, to investigate whether their performance is related to later language development, we obtained language measures via parents' questionnaires (ELFRA-1, Grimm & Doil, 2006; FRAKIS, Szagun, Stumper & Schramm, 2009) at the ages of 14 and 18 months, respectively. The purpose of Experiment 3a was to explore the potential segmentation strategies at this age. Experiment 3b was a replication of Experiment 3a but

[12] As defined by the statistical structure of the language.

obtained pupillometry data. Experiment 3c was a control experiment to check for any spontaneous preferences for the stimuli presented during the test phase of Experiment 3a. Experiment 3d was a modified experiment with double familiarization time to examine whether infants needed more input to segment words from the string. From the previous literature, we expected German-learning 9-month-olds to rely more strongly on prosodic cues than on TPs and this reliance to be linked to later language outcomes.

5.1.2 Participants

Twenty-five 8- to 9-month-old German monolingual infants were tested (13 girls, 12 boys). The mean age was 8 months and 26 days (range 8;15 – 9;12). All infants were born full-term without apparent health problems. Six additional infants were tested but excluded due to crying (1), fussiness (3), and technical problems (2). This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

5.1.3 Stimuli

The same familiarization string as in the adult Experiment 2a was used. However, the test trials differed in two ways. First, the number of words presented per condition was reduced to two: the two infrequent statistical words (*puda* and *bido*), the two prosodic words with comparable frequency to the infrequent statistical words (*buta* and *dego*), and two non-words (*dabi* and *bide*). Second, a test trial consisted of 12 repetitions of the same word with an 800 ms pause between each token for a total duration of 18 s. As in Experiment 2a, three conditions were presented at test: prosodic words, statistical words, and non-words.

5.1.4 Procedure

We used the Headturn Preference Procedure as introduced by Hirsh-Pasek et al. (1987). During the experimental session, infants were seated on a caregiver's lap in the center of a test booth. The

caregiver listened to music over headphones to prevent influences on the infant's behavior. Furthermore, he or she was instructed not to interfere with the infant during the experiment. Inside the booth, three lights were fixed: a green one at the center and a red one on each side. On the outside of the test booth, two loudspeakers were mounted just below the red lights. The procedure during the familiarization was as follows: the green light started to blink to attract the infant's attention to the center. When the infant oriented to the light, the experimenter pressed a key attached to a button box, initiating the flashing of a randomly chosen side light. The light flashed until the infant looked away for two consecutive seconds or after 30 seconds. Then the light extinguished and the center light began flashing again. During the familiarization the speech string was played continuously and was not contingent on the infant's looking behavior to avoid uncontrolled breaks in the familiarization strings. The purpose of using lights during this phase of the experiment was to keep the infant's attention high, as well as to familiarize infants with the contingency between their looking behavior and the light activity, which is important in the testing phase.

Immediately after the completion of the familiarization string, the test phase began. In the test phase the presentation of the acoustic stimuli was contingent on the infant's looking behavior. Each trial started with the green light blinking to attract the infant's attention to the center. When the infant oriented to the green center light, this light went out and one of the side red lights started to blink. When the infant turned her head towards the now blinking light, the speech stimulus was started and presented until completion (18 s) or until the infant turned away from the target side for more than two consecutive seconds. If the infant briefly turned her head for less than two seconds, the presentation of the speech file continued but the time spent looking away was not included in the total listening time. The information about the duration of looking time was coded by an experimenter outside the testing room with a push-button control. The coder was blind to the

experimental condition that was presented. The same experimenter coded all the sessions from all the infants that came to the lab.

All infants were familiarized with the same language and tested on the same words in the test trials. Each of the trials was presented two times during the test phase resulting in a total of 12 trials. There were four different versions of the experiment, which differed in the order of stimulus presentation: three blocks of four trials selected from the three experimental conditions were created. These blocks differed in the order and distribution of the items across conditions. Between infants, the order of presentation of these three blocks was counterbalanced. The total duration of the experimental session was between 3 and 5 minutes, depending on the infant's behavior.

For each infant, the language outcome at 14 and at 18 months of age was assessed by two standardized German parental questionnaires (ELFRA-1 at 14 months, Grimm & Doil, 2006; FRAKIS at 18 months, Szagun, Stumper & Schramm, 2009). The questionnaire ELFRA-1 consists of four subtests: speech production (productive vocabulary and production of sounds and word combinations), speech perception (receptive vocabulary and reaction to language), gestures, and fine motor skills. FRAKIS consists of three subtests: vocabulary (receptive and productive), syntax, and morphology. Two different questionnaires were used because each test is valid for a specific age range. The ELFRA-1 questionnaire is valid for children around 12 months and the FRAKIS questionnaire is valid for children from 1;6 to 2;6 years of age.

5.1.5 Results

Infants listened for 7.8 s (*SD* = 2.3) on average to the prosodic words during the test trials, for 8 s (*SD* = 2.9) to the statistical words, and for 7.9 s (*SD* = 2.7) to the non-words (see Figure 8). All trials were included in the analysis. We performed the statistical analysis with non-parametric tests (Wilcoxon Signed-Rank Test and Spearman's rho correlation) because the data were not normally

distributed (Shapiro Test, W = 0.86, p < .001). The statistical analysis revealed no significant difference between statistical and prosodic words (V = 165, z = -0.05, p = .95), as well as no significant difference between non-words and prosodic words (V = 144, z = -0.47, p = .63). The difference between non-words and statistical words was also not significant (V = 152, z = -0.26, p = .79).

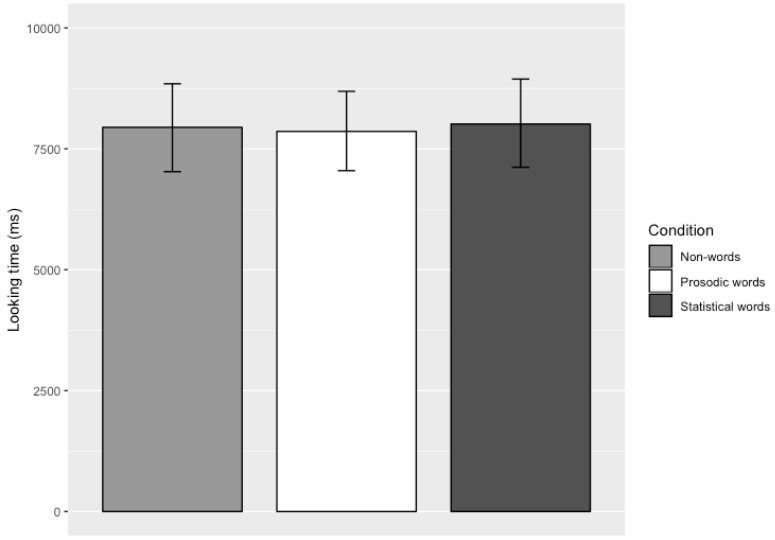

Figure 8: Mean looking times at test for the 9-month-olds.
The error bars represent the standard error.

Correlations between the language test scores (ELFRA-1 and FRAKIS) and the mean looking times for the prosodic condition and for the statistical condition were calculated. In addition, we also calculated the correlations between the ELFRA-1 and FRAKIS outcomes with two difference scores (prosodic condition minus statistical condition; prosodic condition minus non-word condition). We could obtain the ELFRA scores for all infants tested in the HPP (n = 25). The FRAKIS scores were available for 21 infants. The overall mean ELFRA-1 score was 196.3 (SD = 80) out of 370 points. Infants scored 66.4 (SD = 37) out of 171 possible points in the receptive vocabulary subtest, and 10 (SD = 13.1) out of 181 possible points in the productive vocabulary subtest. We correlated the overall ELFRA-1 score as well as the receptive and productive

vocabulary subtests with the looking times in the HPP test. No significant correlations were observed (see Table 14).

Table 14: Correlations between the looking times and the test scores in ELFRA-1

	Prosodic words		Statistical words		Non-words		DF statistical - prosodic		DF non-words - prosodic	
	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value
Total score	.17	.42	.24	.25	.05	.80	.10	.64	.06	.74
Productive vocabulary	-.06	.75	.04	.82	.35	.10	.02	.9	.38	.07
Receptive vocabulary	.28	.18	.22	.29	.08	.69	-.02	.9	.02	.9

DF = Difference score (A minus B)

** = significant, p < .05*

The overall mean FRAKIS score was 33.4 (*SD* = 23.6) out of 674 points. Infants scored 32 (*SD* = 21.9) out of 600 possible points in the productive vocabulary, 1.28 (*SD* = 2.61) out of 32 possible points in the syntax subtest, and 0.04 (*SD* = 0.21) out of 42 possible points in the morphology subtest. We correlated the overall FRAKIS score as well as the vocabulary, syntax and morphology subtests with the looking times in the HPP test. No significant correlations were observed (see Table 15).

Table 15: Correlations between the looking times and the test scores in FRAKIS

	Prosodic words		Statistical words		Non-words		DF statistical - prosodic		DF non-words - prosodic	
	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value
Total score	.09	.69	-.13	.57	.08	.72	-.25	.27	.13	.56
Vocabulary	.13	.58	-.10	.67	.09	.68	-.25	.27	.11	.62
Morphology	.09	.67	-.17	.45	-.33	.14	-.26	.27	-.33	.14
Syntax	-.43	.97	-.27	.24	-.32	.16	.06	.79	.21	.36

DF = Difference score (A minus B)

** = significant, p < .05*

The results of this experiment show that 9-month-olds have no preference for any of the three different test conditions, suggesting that infants did not segment the words from the artificial string. Furthermore, their performance in the HPP procedure was not related to later language outcomes at the ages of 14 and 18 months.

If infants treated both cues equally strong, it is possible that the conflicting cues canceled each other out and therefore no consistent preference is observed. However, infants may be learning but their learning outcome might not be visible at a behavioral level. If this is the case, online measures might be a better methodological option to reflect the learning process that is underway during the test phase because they measure the cognitive processing throughout the course of the test and may be less susceptible to factors affecting the direction of infants' preference. We address this issue in Experiment 3b and tested 9-month-olds in a similar task with pupillometry.

We address two further issues with slight variations of the method of Experiment 3a. The first concern is that infants may show spontaneous preferences for single specific words presented during the test phase, which could have masked a potential learning effect, and thus they showed no learning at test. To address this issue we tested a group of 9-month-olds without familiarization in Experiment 3c. Another possible concern is that infants did not have enough input from the artificial string to process the different cues and to use them for word segmentation. We addressed this possibility in Experiment 3d by doubling the exposure time to the string.

5.2 Experiment 3b: Word segmentation at 9 months with pupillometry

5.2.1 Introduction

Pupillometry is a minimally demanding online measure. In contrast to behavioral methods like the HPP, pupillometry is a response that can be evoked in a passive listening procedure and does not need an overt behavioral response. This method has been used as a speech perception measure at

very early ages, with infants as young as 3 months of age (Hochmann & Papeo, 2014) as well as with 30-month-old infants (Fritzsche & Höhle, 2015; Tamási, McKean, Gafos, Fritzsche & Höhle, 2017). In young children, increased pupil dilation has been found to be related to surprise (Jackson & Sirois, 2009), novelty (Hochmann & Papeo, 2014), violation of expectation (Gredebäck & Melinder, 2010; Hepach & Westermann, 2013), and cognitive effort (Karatekin, 2007). The demonstrated sensitivity of pupillometry when testing very young infants shows that it can provide new insights into infants' abilities, since it has been shown to be more sensitive to differences in experimental conditions than measures of looking time (Jackson & Sirois, 2009; Hepach & Westermann, 2013; for an overview, see Hepach & Westermann, 2016). These findings and the outcomes with adults in Experiment 2a motivated us to test 9-month-olds in a similar task with pupillometry measurements.

Although the results with the German adult listeners (see Section 4.1.5.2) showed that a larger pupil size change would be expected in the statistical condition compared to the prosodic condition, it might not be the case that infants show the same pattern. So far, no infant pupillometry studies have conducted a similar segmentation experiment. However, note that Tamási et al. (2017) found that 30-month-old infants had larger pupil size when presented with a mispronounced word. Therefore, in our case it might be that infants show larger pupil sizes when faced with a word that they have not previously heard or segmented. Hence, if there is any stronger reliance on prosody for the 9-month-olds, as we predict, we could expect infants to show a similar pattern to that of the adult German listeners.

5.2.2 Participants

Twenty-eight 8- to 9-month-old German monolingual infants were tested (14 girls, 14 boys). The mean age was 9 months and 4 days (range 8;16 – 9;16). All infants were born full-term without

apparent health problems. Eight additional infants were tested but excluded due to calibration failure (4), fussiness (2), and technical problem (2). None of them participated in Experiment 3a. This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

5.2.3 Stimuli

The same familiarization string as in Experiment 3a was used. However, the test trials differed to relate the timing of the stimulus presentation directly to the timing of the pupil dilation, which would not so easily be possible if repetitions of the test words were used. Therefore a test trial consisted of a single repetition of the word (around 500 ms) followed by 2.5 s of silence (the trial lasted 3 s). The same test words were used as in Experiment 3a.

5.2.4 Procedure

During the experimental session, infants were seated on a caregiver's lap in a test booth in front of a computer screen and an eye-tracker. A loudspeaker was placed behind the screen. The caregiver listened to music over headphones to prevent influences on the infant's behavior. Furthermore, he or she was instructed not to interfere with the infant during the experiment. The distance between the infant and the display was approximately 60–70 cm and tracking was remote. The experimental session started with an eye calibration using a 5-point sequence, in which a grey background with white points was presented on the screen. If the calibration failed, the experimenter started it again for a maximum of 3 times. After a successful calibration (at least 4 points out of 5), the familiarization phase started, which consisted of auditory and visual stimuli. The artificial string was played continuously and was not contingent on the infant's looking behavior. Simultaneously, a video of an aquarium with swimming fish was played on the screen. Immediately after the completion of the familiarization string, the test phase began. In the test phase the presentation of

the acoustic stimuli was contingent on the infant's looking behavior. Each trial started with an attention getter: a silent cartoon figure moving his legs and hands on a grey background. When the infant oriented to the screen, the attention getter disappeared and the test trial was played while the same blank grey screen was shown. The trial did not start if the infant was not looking toward the screen. This procedure was repeated for all 12 trials. If the infant looked away from the screen for more than two consecutive seconds, the attention getter appeared until the infant reoriented to the screen. The experimenter could start an acoustic stimulus like a baby laugh or a bird sound to redirect the infant's attention to the screen. This was only used if the infant would not reorient to the screen. The experimenter was blind to the experimental condition that was presented.

All infants were familiarized with the same language and tested on the same words in the test trials. Each of the trials was presented two times during the test phase, resulting in a total of 12 trials. There were four different versions of the experiment, which differed in the order of stimulus presentation: three blocks of four trials selected from the three experimental conditions were created. These blocks differed in the order of the items of the conditions. Between infants, the order of presentation of these three blocks was counterbalanced. The total duration of the experimental session was between 3 and 5 minutes, depending on the infant's behavior.

Stimulus presentation was programmed using PsyScope software, which collected both the behavioral and the pupillometry responses. All visual stimuli were shown on a 17" (1280 x 1024) TFT screen with a resolution of 300 x 300 pixels. Pupil diameter was recorded with a Tobii 1750 binocular corneal reflection eye-tracker with a temporal resolution of 50 Hz.

5.2.5 Results

Both eyes were tracked but only data obtained when at least one eye could be recorded entered the analysis. Blinks were eliminated (13.1% of the total data points). The TEPR was calculated and

taken as the main dependent variable. The TEPR measure was baseline-corrected for each individual trial using a 200 ms period for each item. All trials were averaged within each condition. A 3 s window starting at the onset of each word was investigated. Successful trials were defined as those containing pupil measures from at least half the length of the trial. Those participants who did not reach a threshold of 50% (following Fritzsche & Höhle, 2015) of successful trials were excluded from the analysis (2 participants). A total of 26 participants were included in the analysis. Figure 9 illustrates the response dynamics of the pupil during the 3 seconds of all trials. Importantly, the pupil size increased in response to the acoustic information in all three conditions (from 500 to 1000 ms).

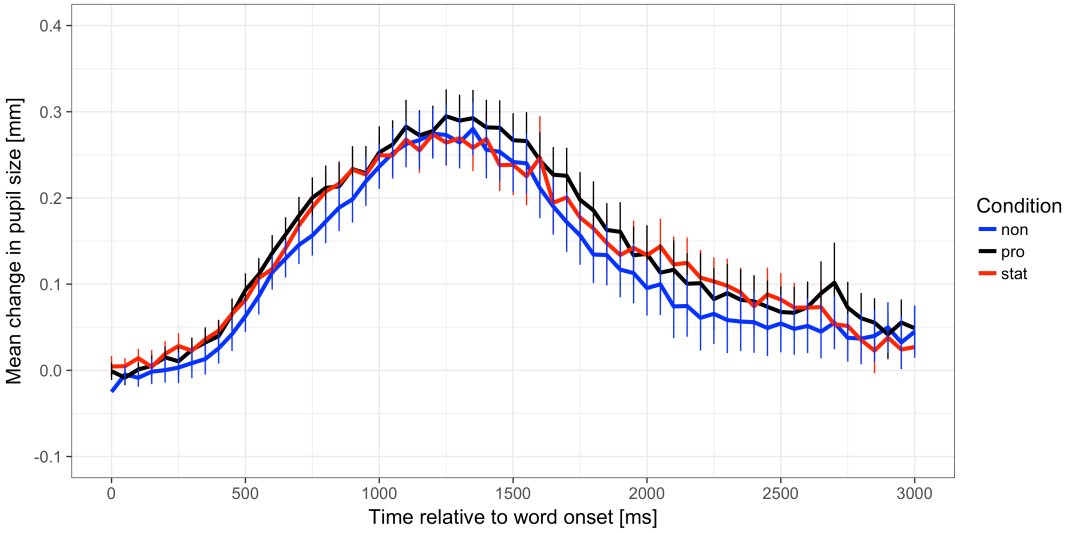

Figure 9: Time course of the pupil size changes in the different conditions

The differences between the pupil size changes in the different conditions were analyzed using a linear mixed effects model using the *lmer* function in the *lme4* R package. Graphs were generated using the package *ggplot2* (Wickham, 2009) and the contrasts were coded with the *MASS* package (Venables & Ripley, 2002). Although the outcomes with German adults showed significant effects on the last two time windows, we could not expect infants to have the same physiological pupillary response. Therefore, we applied the model to the data collected from the end of the word (500 ms) until 2500 ms, where the pupil went back to baseline.

The model we fitted followed the recommendation by Matuschek et al. (2017) to specify a maximal random effects structure for confirmatory hypothesis testing without losing power. We checked for the random component structure with the *RePsychLing* R Package (Bates et al., 2015) and fitted the maximal model that best explained our data. In the model, condition (*Condition*) was entered as a fixed effect with three levels: prosodic word (*Prosodic*), non-word (*Nonword*), and statistical word (*Statistical*). We used a sliding contrast for successive comparisons between the conditions. We coded the contrast so that the prosodic condition was compared to the two other conditions, while non-words and statistical words were not compared. *Participant* and *Trial Position* were included as random factors. The factor *Gender* was left out of the model because it did not improve the model fit to the data. The output of the model is presented in Table 16. The model revealed no significant differences between the conditions (*Prosodic - Statistical*, β = -2.42, t = -1.62, p = .10; *Nonword - Prosodic*, β = -2.28, t = 1.52, p = .12), suggesting that responses to the three different conditions were not different from each other. As is the case in most infant studies, *Trial Position* turned out to be significant (β = -3.69, t = -3.01, p = .001), the negative β suggesting a decrease in pupil size changes over time.

Table 16: Maximal Model from 500 - 2500 window (9-month-olds)

Fixed Effects	***ß***	**SE**	***t***	***p***
Grand mean (intercept)	1.85	2.03	9.14	< .001*
Prosodic - Statistical	-2.42	1.49	-1.62	.10
Nonword - Prosodic	-2.28	1.49	1.52	.12
Trial Position	-3.69	1.19	-3.01	.001*
Random Effects	**Variance**		**SD**	
id (Intercept)	0.01		0.1	
Residual	0.06		0.25	

** = significant, p < .05*

In this experiment we tested 9-month-old German-learning infants in a word segmentation task similar to that in Experiment 3a but obtaining pupil dilation data. Infants showed no difference in pupil size between the three conditions at test, suggesting that infants did not segment the words from the artificial string. Although we expected to obtain further insights into infants' weighting of prosodic and statistical cues for word segmentation through pupillometry, the pupil size changes between conditions were not different. We further discuss this outcome and the lack of preference observed in the previous experiments in the general discussion (Section 5.5).

As mentioned previously in Experiment 3a, two further control experiments were conducted. In Experiment 3c we checked for spontaneous preferences for single specific words presented during the test phase by testing the infants with only the test phase of Experiment 3a. In Experiment 3d, we doubled the familiarization exposure time to ensure that infants got enough exposure to the artificial string to process the different cues and to use them for word segmentation.

5.3. Experiment 3c: Word segmentation at 9 months without familiarization

5.3.1 Participants

Ten[13] 8- to 9-month-old German monolingual infants were tested (4 girls, 6 boys). The mean age was 8 months and 25 days (range 8;16 – 9;14). All infants were born full-term without apparent health problems. One additional infant was tested but excluded due to interaction with the caregiver in more than 3 trials. None of them participated in Experiments 3a and 3b. This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

5.3.2 Stimuli and procedure

[13] This was a pilot experiment. Therefore, the sample size is small.

The stimuli and procedure in this experiment were identical to those used in the test phase of Experiment 3a. However, no familiarization was presented. Later language outcomes were not assessed.

5.3.3 Results

Infants listened for 8.5 s (*SD* = 3.3) on average to the prosodic words during the test trials, for 8.8 s (*SD* = 2.5) to the statistical words, and for 8.8 s (*SD* = 2.3) to the non-words (see Figure 10). We performed the same statistical analysis as in Experiments 3a and 3b. The statistical analysis revealed no significant differences between the conditions: statistical vs. prosodic words ($V = 30$, $z = -0.19$, $p = .84$), non-words vs. prosodic words ($V = 24$, $z = -0.29$, $p = .76$), and non-words vs. statistical words ($V = 26$, $z = -0.09$, $p = .92$).

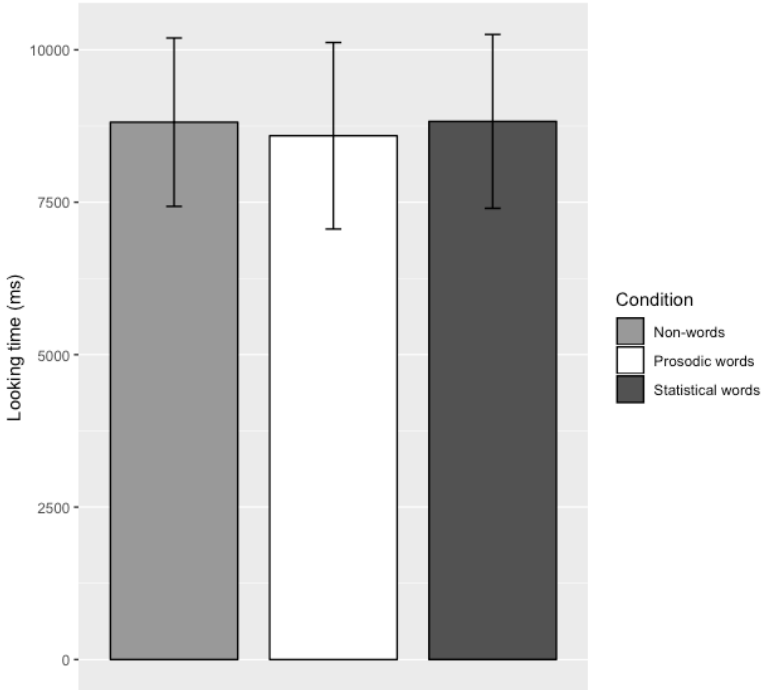

Figure 10: Mean looking times at test for the 9-month-olds without familiarization. The error bars represent the standard error.

This experiment aimed to examine whether the familiarization had an effect and whether infants actually learned from the string by checking for any spontaneous preferences for the stimuli

presented during the test phase of Experiment 3a. The findings are similar to those in Experiment 3a: 9-month-olds seem to have no preference for any of the three different test conditions. Thus it seems that the familiarization did not have any effect on the infants' performance in the test phase in Experiment 3a. However, this finding needs to be taken with caution due to the small sample size (n = 10) and the lack of statistical power. As we mentioned in the discussion of Experiment 3a, there are several other possibilities to explain these results. One possible explanation is the amount of input. In real life, infants receive a high amount of exposure and are able to learn and extract the words from speech. In contrast, in the laboratory situation the input was restricted to around 2 minutes. It might be the case that infants need more time to learn from the artificial string and to process the different cues. To investigate this issue, we doubled the exposure time to the familiarization string in the following experiment and tested a new small group of 9-month-olds.

5.4 Experiment 3d: Word segmentation at 9 months with double familiarization

5.4.1 Participants

Thirteen[14] 8- to 9-month-old German monolingual infants were tested (6 girls, 7 boys). The mean age was 8 months and 26 days (range 8;15 – 9;12). All infants were born full-term without apparent health problems. Two additional infants were tested but excluded due to crying (2). None of them participated in Experiments 3a, 3b, or 3c. This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

5.4.2 Stimuli and procedure

The stimuli and procedure in this experiment were identical to those used in Experiment 3a. However, the time of the familiarization phase was doubled (4 min 22 s instead of 2 min 11 s). The longer exposure was divided into two parts as follows: when infants entered the test booth with

[14] This was a pilot experiment. Therefore, the sample size is small.

their caregiver, the experimenter started to play the first familiarization string (2 min 11 s). Meanwhile, the experimenter explained the experimental setting instructions to the caregiver until the familiarization string finished. This first exposure was incidental. Infants heard the second familiarization string (2 min 11 s) in the experimental session as in Experiment 3a. We followed this procedure (similar to Evans et al., 2009) to make the long familiarization less monotonic for the infant and to thus avoid potentially high dropout rates. Later language outcomes were not assessed.

5.4.3. Results

Infants listened for 7.9 s (*SD* = 2.5) on average to the prosodic words during the test trials, for 8.1 s (*SD* = 3.5) to the statistical words, and for 7.6 s (*SD* = 2) to the non-words (see Figure 11). We performed the same statistical analysis as in the previous experiments. The statistical analysis revealed no significant differences between the conditions: statistical vs. prosodic words ($V = 47$, $z = -0.06$, $p = .94$), non-words vs. prosodic words ($V = 47$, $z = -0.20$, $p = .94$), and non-words vs. statistical words ($V = 50$, $z = -0.33$, $p = .78$).

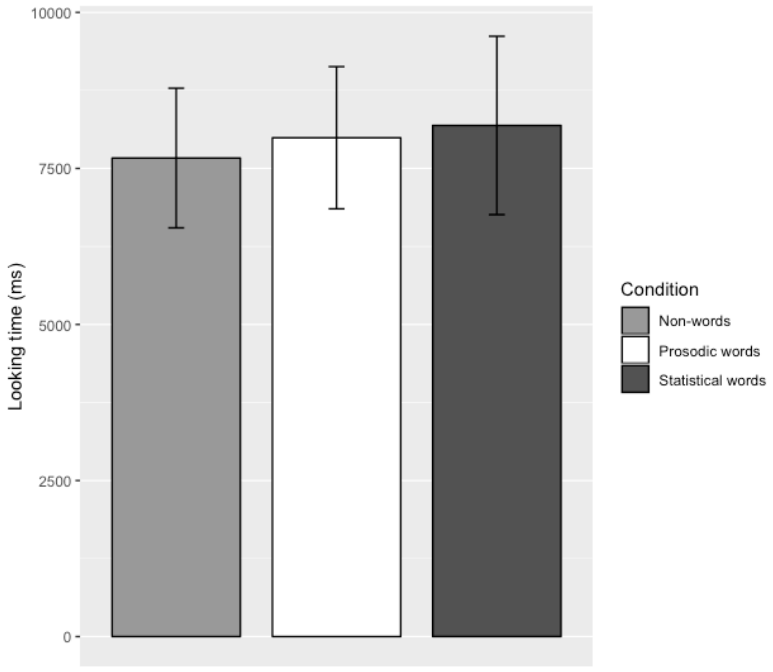

Figure 11: Mean looking times at test for the 9-month-olds with double familiarization time. The error bars represent the standard error.

Again, the results of this experiment are similar to those in Experiment 3a: 9-month-olds show no preference for any of the three different test conditions. This suggests that, although infants received double exposure to the artificial string, they did not show any learning effect at test. Note, however, that this was a control experiment with a small sample size (n = 13) and therefore the results need to be taken with caution.

5.5 General discussion

In the experiments presented in this chapter we tested 9-month-old German-learning infants in a word segmentation task in which statistical cues and prosodic cues indicated different word boundaries. The main aim was to explore whether infants at this age weight one cue more strongly than the other when segmenting words from fluent speech. In the first experiment, Experiment 3a, infants were familiarized with a 2 min string and then tested in the HPP with three conditions. Infants did not show any indication that they had segmented the speech stream. To obtain measures other than only looking times we tested a group of 9-month-old infants with pupillometry (Experiment 3b). However, we obtained again null results. Because infants did not show any indication that they had segmented the speech stream in the previous experiments (they responded in the same way to the non-words compared to both word types), we tested two smaller groups of infants with minimal variations in the experimental procedure: one without the familiarization (Experiment 3c) and one with double familiarization time (Experiment 3d). In Experiment 3c we checked for spontaneous preferences for specific words at test and in Experiment 3d we addressed the possibility that infants did not get enough input and we again obtained null results. Note that, although it might be an effect of insufficient statistical power (Experiment 3c and 3d), the results we obtained in these two experiments are especially important because they provided us with first insights about the segmentation strategies at 9 months of age and suggest that no clear segmentation effect was present (null results).

We will now address several possible reasons why we did not find any effect at the age of 9 months. The first possibility is that infants do not learn in this experimental setting or with these specific stimuli. This is rather unlikely, taking into account that our stimuli worked for adult participants and that the HPP worked with 9-month-olds in Thiessen and Saffran's (2003) study, which is the closest to our experimental set-up. However, the innovative third condition (the non-word condition) implemented in the test phase may have introduced some additional cognitive complexity that we cannot measure. Therefore, we cannot rule out the possibility that the specific test procedure did not work for the 9-month-old infants. An exact replication of Thiessen & Saffran (2003) would allow us to gain more insights into the uncertainty created by the introduction of a third condition at test.

The second explanation, also considered in the discussion of Experiment 3a, is that infants learn from the familiarization but do not show learning at test. Perhaps infants became confused when matching between the stressed familiarization syllables and the prosodically flat test syllables or got bored during the test phase or got confused due to the fact that prosody and statistics pointed in different directions. This was the motivation for testing infants with pupillometry. We expected this method to be more sensitive, giving us better insight into whether words are recognized during the test phase and thus which cues are being used to process the speech stream, but infants still showed no difference between the conditions.

The third explanation is that infants treated both cues equally strong and the conflicting cues cancelled each other out, resulting in a segmentation failure. Hence, no consistent preference for any of the conditions should occur. It may be that infants at this age perceived both cues and are paying attention and attempting to use them simultaneously for segmentation (maybe they cannot ignore one of the cues). For example, Hay & Saffran (2012) found that none of the cues present in a stream received sufficient weight to override the other. In their study they tested 9-month-olds in a SL task with the HPP and aimed to investigate how the acoustic characteristics of sounds (intensity

and duration) interact with sequential statistical cues to word boundaries. Infants were familiarized with two languages (trochaic and iambic) with statistics and acoustic cues (either intensity or duration) pointing to the same word boundaries or not. Infants were misled by acoustic cues in fluent iambic speech (as our string), treating stressed syllables as word onsets although those stressed syllables were actually the second syllables of words (marked by duration and statistical information). Furthermore, infants successfully discriminated words from part-words when the first syllable of each word was marked by greater intensity, suggesting that sequential statistical cues were constrained in that case. The authors conclude that acoustic cues and sequential statistics interact and that infants may be sensitive to both statistical and rhythmic grouping cues to word boundaries, with neither cue receiving sufficient weight to override the other. This may be a plausible explanation for our results. Since they may segment the string on different levels (one level based on prosody and the other based on statistics), the segmentation is hindered. However, if this was the case, then they should have at least responded in a different way to the non-words compared to statistical and prosodic cues –a difference we did not observe– because either statistical or prosodic words might have sounded more familiar than the non-words.

Finally, the last explanation is related to infants' direction of preference. We cannot rule out the possibility that the stimuli presented elicited a different direction of preference to infants depending on the stage of their development, i.e., some infants can show novelty and some infants familiarity towards the same stimuli (DePaolis et al., 2016). Black and Bergmann (2017) reported in their meta-analysis that more mature infants might show a different direction of preference (e.g., from a preference for words to a preference for non-words). However, this was not the case in our sample. In our analysis we found no correlations between infants who showed longer looking times for one of the conditions and later language outcomes.

Following the previous argument related to the individual development, it is plausible that not all infants weighted the cues in the same way, and therefore no group effect emerged. Recall that a shift in the exploitation of word segmentation cues has already been observed in English-learning infants (Thiessen & Saffran, 2003). In their study, 9-month-olds showed longer looking times for statistical words (novelty preference) when the string contained both stress and TPs. The individual infants' preferences in our results from Experiment 3a show that 11 infants looked longer for non-words, 7 for statistical words and 6 for prosodic words. However, these outcomes do not allow us to infer any further details about the individual development of each infant because infants were only tested once and there was no correlation between the looking times and their later vocabulary scores. The individual preferences for the different conditions might have cancelled each other out. The fact that pupillometry revealed no effects also suggests that a mixture of infants with different directions of preference may be responsible for the null result. It might be that a shift from a stronger reliance on statistical cues to prosodic cues also exists for German-learning infants, happening at around the same age but at a different developmental pace for each infant.

Our results contrast with those of Thiessen and Saffran (2003) in further ways. They found a developmental shift in English-learning infants between the ages of 7 and 9 months: whereas younger infants relied more strongly on statistical cues, older infants used prosody to segment an artificial speech stream. One possible explanation for this difference in outcomes is the nature of the stimuli. Thiessen and Saffran (2003) used synthesized speech in which lexical stress was artificially modified in terms of pitch, duration, and intensity. It is possible that the more natural variation in prosodic properties found in the materials used for our study was favouring a prosodic segmentation compared to Thiessen and Saffran's experiment and therefore the infants in the present study treated the task differently. Interestingly, Black & Bergmann (2017) report a significant effect of stimuli naturalness in their meta-analysis. While studies using synthesized speech yield reliable novelty

preferences, studies using naturally recorded speech fail to find reliable effects, perhaps because a more complex signal takes more time to process. Another plausible explanation is that 9-month-old German infants rely on TPs like 7-month-old English-learning infants, but at an earlier stage of development because of the prominence of prosody in German, which has an impact on early language development. Following Thiessen and Saffran's (2003) study, we continued to explore the weighting of statistical and prosodic cues to word segmentation in younger infants (Experiment 4).

Turning our attention to the later language development outcomes, it is possible that the use of a segmentation cue (specifically prosodic information) over another does not predict language skills as predicted. However, we suggest that the absence of significant correlations was due to the low variance and lack of effect in the HPP performance. If there is no segmentation, it is conceivable that no correlation with later language development can be observed.

We believe that further research is needed to explore whether the observed word segmentation failure (or weighting of cues for word segmentation) in such a laboratory setting represents a true weakness in the ability to segment or analyze fluent speech and whether it has implications for later language development. In short, while our data do not allow us to draw any further conclusions regarding what cues 9-month-old German-learning infants use to segment words from fluent speech, we hope that these findings will be useful and open future research questions.

6. EXPERIMENT 4: WEIGHTING OF SEGMENTATION CUES IN 6-MONTH-OLD GERMAN INFANTS

6.1 Experiment 4a: Word segmentation at 6 months

6.1.1 Introduction

We conducted two experiments with 6-month-old German infants to explore whether they show an initial dominance of statistical cues over prosodic cues in speech segmentation. Furthermore, to investigate whether this ability is related to later language development, we obtained the ELFRA-1 and FRAKIS tests at the ages of 12 and 18 months, respectively. The purpose of Experiment 4a was to compare the potential segmentation strategies and Experiment 4b was a control experiment to check for any spontaneous preferences for the stimuli presented during the test phase of Experiment 4a. Based on the previous research, we expected German-learning 6-month-olds to rely more strongly on prosodic cues than on TPs and this reliance to be linked to later language outcomes.

6.1.2 Participants

Twenty-four 6- to 7-month-old German monolingual infants were tested (12 girls, 12 boys). The mean age was 6 months and 21 days (range 6;12 – 7;0). All infants were born full-term without apparent health problems. Eight additional infants were tested but excluded due to fussiness (4), crying (2), experimenter error (1), and technical problems (1). This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

6.1.3 Stimuli and procedure

The stimuli and procedure in this experiment were identical to those of Experiment 3a. For each infant, the language outcome at 12 and at 18 months of age was assessed by two standardized

German parental questionnaires (ELFRA-1, Grimm & Doil, 2006; FRAKIS, Szagun, Stumper & Schramm, 2009). The questionnaire ELFRA-1 consists of four subtests: speech production (productive vocabulary and production of sounds and word combinations), speech perception (receptive vocabulary and reaction to language), gestures, and fine motor skills. FRAKIS consists of three subtests: vocabulary (receptive and productive), syntax, and morphology. Two different language questionnaires were used because each test is valid for a specific age range. The ELFRA-1 questionnaire is valid for children around 12 months and the FRAKIS questionnaire is valid for children from 1;6 to 2;6 years of age.

6.1.4 Results

Infants listened for 7.55 s (SD = 2.97) on average to the prosodic words during the test trials, for 8.59 s (SD = 3.01) to the statistical words, and for 8.52 s (SD = 2.76) to the non-words (see Figure 12). All trials were included in the analysis. Since the looking times were not normally distributed (Shapiro Test, $W = 0.83$, $p < .001$), we performed the statistical analysis with the non-parametric tests (Wilcoxon Signed-Rank Test and Spearman's rho correlation).

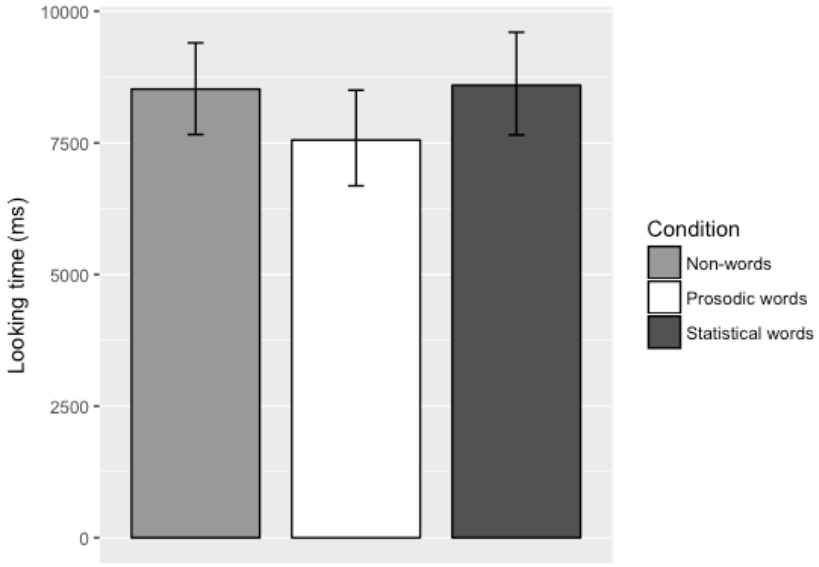

Figure 12: Mean looking times at test for the 6-month-olds. The error bars represent the standard error.

The statistical analysis revealed that the difference between statistical and prosodic words was significant (V = 230, z = -2.3, p = .021), as well as the difference between non-words and prosodic words (V = 72, z = -2.24, p = .024). However, the difference between non-words and statistical words was not significant (V = 147, z = -0.07, p = .81).

Correlations between the language test scores (ELFRA-1 and FRAKIS) and the looking times for all three conditions were calculated. In addition, we also calculated the correlations between the looking time of two difference scores (prosodic condition minus statistical condition; prosodic condition minus non-word condition) and the ELFRA-1 and FRAKIS tests. Due to the sampling problems associated with longitudinal designs, the ELFRA scores were available for only 20 infants and the FRAKIS scores for 16 infants (out of 24 tested in the HPP procedure). The overall mean ELFRA-1 score was 83.35 (*SD* = 39.97) out of 370 points. Infants scored 45.05 (*SD* = 33.58) out of 171 possible points in the receptive vocabulary subtest, and 2.75 (*SD* = 2.97) out of 181 possible points in the productive vocabulary subtest. We correlated the overall ELFRA-1 score as well as the receptive and productive subtests with the looking times in the HPP test. There were significant negative correlations between the productive vocabulary subtest and the mean looking times for the three conditions (prosodic words: r = -.42, p = .05; non-words: r = -.37, p = .04 ; statistical words: r = -.59, p < .01), which suggests that the longer looking times during the test phase, the lower the scores were in the productive vocabulary test. No further significant correlations were observed (see Table 17).

Table 17: Correlations between the looking times and the test scores in ELFRA-1

	Prosodic words		**Statistical words**		**Non-words**		**DF statistical - prosodic**		**DF non-words - prosodic**	
	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value
Total score	-.13	.57	-.36	.11	-.001	.99	-.36	.11	.11	.61
Productive vocabulary	-.42	.05*	-.59	.005*	-.37	.04*	-.05	.89	.10	.66
Receptive vocabulary	-.06	.77	-.29	.20	-.03	.88	-.37	.10	.06	.78

DF = Difference score (A minus B)

** = significant, p < .05*

The overall mean FRAKIS score was 57.93 (*SD* = 69.35) out of 674 points. Infants scored 54.81 (*SD* = 64.5) out of 600 possible points in the productive vocabulary, 2.62 (*SD* = 4.03) out of 32 possible points in the syntax subtest, and 0.5 (*SD* = 1.31) out of 42 possible points in the morphology subtest. We correlated the overall FRAKIS score as well as the vocabulary (receptive and productive), syntax and morphology subtests with the looking times in the HPP test. There was a significant negative correlation between the looking time to statistical words and the morphology subtest. This suggests that the longer the infants looked to the statistical words, the worse they scored in the morphology subtest. No further significant correlations were observed (see Table 18).

Table 18: Correlations between the looking times and the test scores in FRAKIS

	Prosodic words		**Statistical words**		**Non-words**		**DF statistical - prosodic**		**DF non-words - prosodic**	
	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value	Rho	p-value
Total score	-.08	.76	-.20	.44	-.10	.72	-.25	.34	-.12	.63
Vocabulary	-.07	.76	-.20	.44	-.10	.71	-.25	.34	-.13	.63
Morphology	-.28	.29	-.51	.04*	-.41	.10	-.25	.34	-.05	.83
Syntax	-.27	.29	-.32	.21	-.20	.45	-.13	.62	.07	.77

DF = Difference score (A minus B)

** = significant, p < .05*

Our results show that the infants listened longer to the non-words and the statistical words compared to the prosodic words with no differences between the non-words and the statistical words. However, given the fact that we used the same familiarization strings and the same test trials for all infants, the question arises whether the preferences that infants showed during the testing phase are in fact a consequence of having segmented the familiarization strings or whether they simply reflect a preference for some syllable combinations over others. Thus, Experiment 4b was conducted to investigate whether the same preference for the test items occurs when the experiment is run without a familiarization phase.

6.2 Experiment 4b: Word segmentation at 6 months without familiarization

6.2.1 Participants

Twenty-six 6- to 7-month-old German monolingual infants were tested (14 girls, 12 boys). The mean age was 6 months and 23 days (range 6;15 –7;0). All infants were born full-term without apparent health problems. Five additional infants were tested but excluded due to crying (1), fussiness (2), technical problems (1), and looking times longer than 2 *SD* above the mean in one of the three conditions (1). None of them participated in Experiment 4a. This study was approved by the Ethics Committee of the University of Potsdam. Written informed consent was obtained from all participating families.

6.2.2 Stimuli and procedure

The stimuli and procedure in this experiment were identical to those in the test phase of Experiment 4a. However, no familiarization was presented. Later language outcomes were not assessed.

6.2.3 Results

Infants listened to the prosodic words for 8.65 s (SD = 2.01) on average during the test trials, to the statistical words for 9.07 (SD = 2.62), and to the non-words for 8.84 s (SD = 2.73) (see Figure 13). All trials were included in the analysis. Since the looking times were not normally distributed (Shapiro Test, $W = 0.80$, $p < .001$), we performed the statistical analysis with the non-parametric Wilcoxon Signed-Rank Test. Results revealed that the difference between statistical and prosodic words did not reach significance ($V = 209$, $z = -0.82$, $p = .40$), nor did the difference between non-words and prosodic words ($V = 165$, $z = -0.24$, $p = .80$) or the difference between non-words and statistical words ($V = 193$, $z = -0.42$, $p = .60$). The results of this experiment show no preference for any of the three different test conditions. This suggests that the differences in listening times

obtained between the conditions in Experiment 4a are probably not merely the result of the infants' preferences for certain syllable combinations over others.

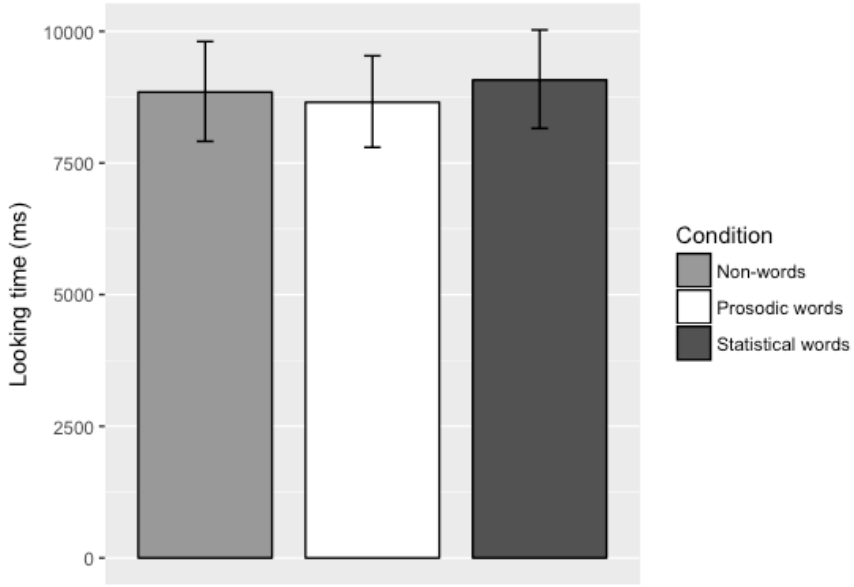

Figure 13: Mean looking times at test for the 6-month-olds without familiarization. The error bars represent the standard error.

6.3 General discussion

Our results from Experiments 4a and 4b show that the infants listened longer to the non-words and the statistical words compared to the prosodic words, with no differences between the non-words and the statistical words. In line with previous studies, this effect can be interpreted as a novelty effect for the statistical and the non-words, indicating that infants' attention was more attracted by those trials that contained test items that were less familiar to them from the familiarization string. In turn, this suggests that infants segmented the familiarization string into prosodic words. Therefore, we propose that, unlike the 9-month-olds, German-learning 6-month-old infants rely more strongly on prosodic cues than on TPs when segmenting the string. Interestingly, no difference was found between the statistical and the non-words, which could suggest that TP information is ignored by German-learning 6-month-olds as soon as prosodic cues are available –which again mirrors the results from the German adults.

Adding a third condition at test helped us to interpret the direction of preference in the segmentation experiments: a novelty effect was found as a preference for items that had low TPs in statistical learning tasks of previous experiments with infants (e.g., Saffran et al., 1996b; Johnson & Jusczyk, 2001; Curtin et al., 2005). However, in our study infants' listening times for disyllabic test items that had high TPs in the familiarization string were as high as listening times for disyllabic test items composed of syllables that had never co-occurred in the familiarization string. In contrast, listening times were lower for prosodic words, i.e., disyllabic items that had formed trochaic forms in the familiarization string. This suggests that the former two types of test items sounded less familiar to the infants than the prosodic word test items, suggesting that they segmented the familiarization string according to the prosodic cues and thus weighted the prosodic cues more strongly than the statistical cues. If we had found differences between the statistical and the prosodic words without the information from the non-word condition, these differences would have been much harder to interpret with respect to the potential greater weight of one cue over the other.

Note that it is possible that a short exposure as in our experiment (2 min and 11 s) might not be sufficient to build up a reliable statistical representation of the words. Infants probably need to hear several instances of a word to learn that the syllables contain some statistical coherence. We tested 9-month-olds with double familiarization exposure, who showed no effect, but we did not test this condition with 6-month-old infants. Although it seems unlikely, we cannot rule out the possibility that infants may rely more on statistical cues with a longer exposure. Nevertheless, the TPs present in our stimuli should have been reliable enough for infants to use for word segmentation, since they were perfect (1.0 within-words) and less complex than in natural languages. In addition, infants were exposed to a string with a similar duration as in Thiessen & Saffran (2003).

Our results raise further questions when we relate them to previous findings. Firstly, our previous results with 9-month-olds did not show any preference between the three conditions using the exact

same stimuli and procedure. As shown in the previous chapter, whereas 9-month-olds do not show an effect for a stronger reliance for statistical or prosodic cues, 6-month-olds relied more on prosodic cues. Thus, our findings suggest that the relative weighting given to statistical and prosodic cues might differ depending on the age and/or the language acquired. These outcomes are consistent with previous studies (Morgan & Saffran, 1995; Hay & Saffran, 2012) that showed a higher sensitivity to rhythmic properties of the input at 6 months of age compared to 9-month-olds, who showed no preference for either of the cues.

Secondly, recall that English-learning infants rely more on statistical cues at the ages of 5 and 7 months (e.g., Thiessen & Erickson, 2013; Thiessen & Saffran, 2003), but change their reliance to prosodic cues by the ages of 8, 9, and 11 months (Johnson & Juszcyk, 2001; Thiessen & Saffran, 2003; Johnson & Seidl, 2009). Thus, a developmental shift occurs in English-learning infants between the ages of 7 and 9 months. However, our results show that German-learning infants already rely on prosodic cues at 6 months of age, which is in contrast with those from Thiessen and Saffran (2003)[15]. Similar to Experiment 3a with 9-month-olds, one possible explanation for this difference in outcomes is the nature of the stimuli. Thiessen and Saffran (2003) used synthesized speech in which lexical stress was artificially modified in terms of pitch, duration, and intensity. Thus, their syllables had less variability (310 ms in duration, 4 to 8 dB difference between stressed and unstressed syllables, and a pitch range of 255 to 270 Hz). It is possible that the more natural variation in prosodic properties as found in the materials used for the study with the German infants was more conducive to the use of a segmentation mechanism that relies on prosodic information.

Another plausible explanation for the difference to the results from Thiessen & Saffran (2003) is that German infants, just like English learners, go through a phase with higher reliance on TPs, but

[15] Although we name our sample as 6-month-olds, the ages of both studies are comparable: in both studies infants from 6.5 to 7 months of age were tested.

earlier. This possibility is supported by the cross-linguistic differences between English and German, which might have an impact on early language development. According to Delattre (1963), the proportion of initially stressed words is higher in German than in English (89 % vs 74 % of disyllabic words). In addition, the inflectional system of German is prosodically determined and much richer than that of English, with many syllabic inflectional endings being added to monosyllabic nouns, verbs, adjectives, and even determiners, leading to disyllabic trochaic word forms. In contrast, word stems which are already disyllabic typically take non-syllabic inflectional endings, thus retaining their disyllabic trochaic word form. Therefore, adding inflected word forms to the count by Delattre (1963) would probably even enhance the advantage in the proportion of trochaic word forms for German as compared to English.

In sum, our results call into question whether an initial reliance on TPs for segmenting words from speech is a universal stage in language development and/or whether German infants leave such a stage earlier than their English peers. English and German show some differences in terms of consistency in their word stress patterns, which may help German-learning infants to master stress patterns more rapidly. Therefore, we suggest that the properties of German have an early impact on early word segmentation and support the development of a trochaic bias. This is in line with the findings that the trochaic bias appears earlier in German-learning infants than in English infants (Jusczyk et al., 1993; Echols et al., 1997; Höhle et al., 2009). Although there is a still a debate about the origin of the trochaic bias, the cross-linguistic difference to English-learning infants that we observed provides evidence that it most probably develops.

However, our results do not principally cast doubt on the developmental trajectory as it has been proposed by Thiessen and Saffran (2003), who assume an initial higher reliance on statistical cues compared to prosodic cues with a shift to prosodic cues at a later age that is valid across different languages. To evaluate this proposal for German-learning infants, younger children need to be

tested. However, recent research shows associations between the distribution of specific acoustic cues and dominant prosodic patterns (Nespor, Shukla, Van de Vijver, Avesani, Schraudolf & Donati, 2008) and that different acoustic cues lead to different ways of segmenting speech even in young infants (Bion et al., 2011; Abboub, Nazzi & Gervain, 2016). These findings may suggest that the exploitation of prosodic cues requires less language-specific knowledge than assumed and therefore there is no disadvantage of prosodic cues compared to statistical cues as candidates for bootstrapping. Further experiments on the relations of these different cues across different languages are needed to shed more light on these questions.

Additionally to the experimental data, we obtained later language outcomes at the ages of 12 and 18 months from standard assessments. We observed significant negative correlations with the productive vocabulary at the age of 12 months. Infants who had longer looking times to the three conditions, scored worse in the productive vocabulary subtest in the ELFRA-1 questionnaire (see Table 17). At the age of 18 months, only the morphology test was significantly and negatively correlated with the looking times to statistical words (see Table 18). These correlations are partly against our expectations. Recall that in the HPP we observed shorter looking times for prosodic words and we interpreted the difference to non-words and statistical words as a novelty effect. Hence, we expected better language outcomes for infants who showed shorter looking times for prosodic words. This is actually the case of the correlation between prosodic words and the productive vocabulary subtest (r = -.42, p = .05). However, the fact that this correlation also holds for the other two conditions (non-words: r = -.37, p = .04; statistical words: r = -.59, p < .01) goes against our interpretation.

It could be that only the productive vocabulary is generally related to the word segmentation skills at the age of 6 months. However, if word segmentation is indeed correlated with vocabulary size, we would also expect the receptive vocabulary to be significantly correlated. Furthermore, word

production this young may not be a stable measure for language proficiency, because the variability in productive vocabulary size in infants under 13 months is not equivalent to the variability in receptive vocabulary (Bates, Dale & Thal, 1995). In our sample, the productive vocabulary scores at 12 months of age were rather low (2.75 out of 181 possible points in the productive vocabulary subtest, *SD* = 2.97) and therefore minimal individual differences could have lead these correlations to be significant. Similarly, the significant correlation found between the longer looking times to statistical words and the morphology subtest at 18 months of age is also against our interpretation of the HPP results. Again, the general scores were not very high (0.5 out of 42 possible points, *SD* = 1.31) and subtle differences between infants could have lead to significance. Besides, the sample size at the age of 18 months was considerably reduced compared to the HPP sample size (16 vs. 24 infants). It may be that infants' performance at both ages and their language profiles may have been determined and/or affected by more generalized cognitive abilities (e.g. Singh et al., 2012; but see Newman et al., 2006), such as attention, or environmental and emotional factors that cannot be controlled for. The most plausible explanation for the negative correlations between the word segmentation performance and the later productive vocabulary skills is a general effect of information processing. It is possible that infants who looked longer are slower in information processes and therefore scored worse in the productive vocabulary subtest at 12 and at 18 months of age. This explanation is supported by Tsao et al. (2004), who showed that the number of trials to reach the criterion phase for habituation was significantly correlated with language comprehension and production. Infants who required fewer trials to pass the habituation criterion phase at 6 months developed significantly better lexical abilities by 13 months. In short, infants who processed the information of the habituation phase faster scored better at later language abilities.

Altogether, a stronger reliance on prosodic cues over statistical cues at the age of 6 months does not seem to be related to later language outcomes, which is in contrast to the previous research that

found a link between early speech perception and later language outcomes (e.g. Newmann et al., 2006; Junge et al., 2012). It may be that this link exists but could not be captured at these ages by the selected language questionnaires. On the one hand, there are some common correlates shared between word segmentation and word learning such as auditory acuity or family socioeconomic status that could have influenced infants' language development and that we did not take into account. On the other hand, this seems unlikely because many of the studies that tested a link between early speech perception and later language outcomes used parental questionnaires. We speculate that the use of prosodic cues is not very robust at the age tested and with the procedure used in this experiment, and therefore very few correlations appear to later language outcomes. Thus, it remains unclear from our results whether the ability to use prosodic cues to word segmentation relates to later language vocabulary.

7. CONCLUSION

The main goal of this thesis was to understand how both German infants and adults use acoustic information to identify linguistic structure and to explore the cues they use for word segmentation. We aimed to investigate in what way different cues to word segmentation are exploited by German adults as well as by infants when learning the language of their environment. In summary, we have argued that when both prosodic and statistical cues are available, both German adults and German infants seem to rely more on prosody.

In the first chapter of this dissertation I introduced and put into context the main research questions. In the second chapter I presented the theoretical background of early word segmentation using prosodic and statistical cues, as well as its relation to later language development. Evidence on the different weighting of these two cues by infants and adults from different languages was also presented in this chapter. The third chapter focused on the HPP and reported a study that tested it test-retest-reliability in German 6-month-old infants. In chapter 4, I presented behavioral and pupillometry data showing that, German adults rely more on prosody when both cues are available and in conflict in a speech stream. Chapter 5 aimed to test 9-month-olds in a similar experimental situation as the previous adult experiment. However, null results were obtained in the several experiments conducted. In line with the adult data, in chapter 6 I provided evidence that 6-month-old German-learning infants rely more strongly on prosodic than on statistical cues when segmenting speech.

The key research question of this thesis was whether German adults and infants rely more on prosodic or statistical information when segmenting words from fluent speech. To answer this question we conducted two experiments with adults (Experiments 2a and 2b) and six infant experiments (Experiments 3a - 4b). We tested both adults and infants with a speech stream in which

prosodic cues and statistical cues were pitted against each other. Our behavioral findings revealed that German adults show a strong weight of prosodic cues, at least for the materials used in this study (Experiments 2a and 2b). Interestingly, our pupillometry results fit the behavioral data, showing that for German adults it was apparently easier to make a decision about the prosodic words (Experiment 2b), which sets future research directions using pupil dilation to understand the mechanisms underlying word segmentation.

Regarding infants, the main conclusion that can be drawn from our results is that German infants might weight these two cues differently depending on (a) age and (b) language experience. Regarding (a) age, we observed that whereas 6-month-old infants relied more strongly on prosodic cues (Experiments 4a and 4b), 9-month-olds failed to segment words from the speech stream, showing no preference for either of the cues (Experiments 3a, 3b, 3c, and 3d). We speculate that prosody provides infants with their first window into the specific acoustic regularities in the signal, which enables them to master the specific stress pattern of German rapidly. Our results with 6-month-olds are consistent, not only with our previous adult results, but with previous research that already documented an early impact of the native prosody on early language development such as word stress pattern preferences (trochaic bias) in languages like German (Höhle et al., 2009). This is therefore an important and novel finding for the understanding of the cues used in word segmentation in early stages of development in German. However, research is needed to replicate these results, as well as to compare prosodic cues with other cues to word segmentation in such an early developmental stage.

Regarding (b) language experience, we showed that there are cross-linguistic differences between German- and English-learning infants. We also obtained different results than Thiessen and Saffran (2003) at both ages. One fundamental question that we raised at this point is whether German infants go through this phase with a higher reliance on TPs but shift to prosodic cues earlier than

English-learning infants, or do not go through a period of stronger reliance on TPs at all. To evaluate this proposal for German-learning infants, younger children need to be tested.

According to statistical bootstrapping accounts infants rely on TPs as a first step in word segmentation. Therefore, the shift in reliance found in Thiessen & Saffran (2003) should be observed across languages –at least across languages in which the word stress pattern provides reliable cues for word segmentation. Although we cannot rule out the possibility that German infants rely on this kind of information in a very first stage, we propose that the properties of German have an early impact on word segmentation and support the development of the trochaic bias, and therefore cause a stronger reliance on prosody. Thus, our data with 6-month-old infants showing such an early reliance on prosodic cues support the prosodic bootstrapping theories, which argue that the ability to process prosodic information in the speech signal might be crucial to detecting lexical boundaries and might consequently affect language development. Whether infants use statistical cues to discover prosodic regularities in their native language or whether they use prosodic cues to isolate chunks upon which TPs are computed is still not a fully answered question. However, our findings are a step forwards in the understanding of an early impact of the native prosody compared to SL in early word segmentation.

The difference observed between English- and German-learning infants casts a new light on the fact that language-specific properties or language experience have an effect on early word segmentation. We believe that further research is needed to confirm this novel finding and extend it to other languages. Although it was beyond the scope of this dissertation, future studies should consider testing German-learning infants with our stimuli as synthesized speech as in Thiessen and Saffran (2003). This might contribute to the understanding of the observed differences between the two languages and rule out the possibility that the differences arose because of the different stimuli properties, as mentioned in the discussion of the experiment. In addition, future research should

continue to address the criticisms of the SL stimuli and test the limits of SL in situations where statistics approximate the noise found in real language.

The second research question of this dissertation was whether the use and weighting of prosodic and statistical cues in a word segmentation task can predict later language development skills. To provide an answer to this question, we additionally followed up with 6- and 9-month-old infants and obtained later language outcomes at two different points in development (Experiments 3a and 4a). The outcomes revealed significant correlations mainly in productive vocabulary for the 6-month-olds at the ages of 12 and 18 months. The lack of correlations in the other language areas suggests that the strength of this capacity may be more fragile than expected or that this relation could not be captured at these ages by the language tests or by the HPP. We also did not find any evidence for a link between SL and later language outcomes.

Our results are in contrast with previous literature (e.g. Newman et al., 2006; Junge et al., 2012) that reported positive correlations between speech perception measures and later language outcomes. However, when comparing our results to these older studies, it must be pointed out that in most of these studies older infants were tested in less cognitive demanding tasks. For instance, Junge and colleagues (2012) tested 10-month-olds with ERPs in a stress pattern discrimination task where short familiarization phases were followed by test trials. A study that tested younger infants (4-month-olds) was Höhle et al. (2014), but infants were tested in a discrimination of the typical native trochaic pattern, which can be considered an easier task in our experiments, because infants are not required to learn anything in the laboratory setting, but use their already acquired language knowledge. Newman et al. (2006) retrospectively analysed a large sample of infants and thus statistical power was higher than in our study. In addition, infants were older (age range 7.5-12 months) and were tested in different speech perception tasks. In general, our results add evidence to the link between individual performance in speech perception tasks and later language

development, but we strongly believe that more longitudinal studies are needed to further examine how early word segmentation capacities are related to language proficiency.

In the present thesis we additionally addressed some specific methodological considerations that are worth mentioning. To answer the third and last research question, we explored the reliability of one of the most common infant speech perception measures: the HPP. We conducted a test-retest-reliability test for the HPP testing German-learning infants three times on the same experiment within 3 weeks (Experiment 1). Our results suggest that the HPP is a reliable method to obtain speech perception measures. As we have observed, the HPP might also be a reliable enough tool to measure individual differences and predict later language outcomes, at least at age of 12 months. However, we encourage further studies to make these measurements suitable for diagnostic use. Effort needs to be made to enhance the tools for analyzing individual data statistically, to combine several measures and dependent variables (e.g., EEG data with eye-tracking data), and to establish norms for these measures.

Furthermore, we used the HPP in our SL word segmentation experiments (Experiments 3 and 4) with an innovative characteristic: three conditions were presented at test instead of two. We are aware that this might not have worked for the 9-month-olds, but we demonstrated that 6-month-olds were successful in such a task. These outcomes open the possibility of extending the test phase of three conditions to further infant experiments, allowing the researchers to control for infants' direction of preference, add more conditions, or explore further issues. Furthermore, as observed in the present thesis, it is highly interesting to investigate our research questions not only with behavioral methods, but also with other methods such as pupillometry, because they can provide new insights into the question under study. Thus, we recommend that future studies extend the research to different methods combining several measures and dependent variables.

In short, the aim of the present work was to understand in what way prosodic and statistical cues to word segmentation are exploited by German infants when learning the language in their environment, as well as to explore whether this ability is related to later language development. This thesis provides novel evidence on how German-learning infants use prosodic and statistical cues at the ages of 6 and 9 months, suggesting that infants start to use prosody as a cue from very early on. In addition, the present work contributed evidence to adult segmentation research showing that German listeners mainly use prosodic cues to segment words from speech.

8. REFERENCES

Abboub, N., Nazzi, T & Gervain, J. (2016). Prosodic grouping at birth. *Brain and Language, 162*. 46-59.

Amso, D., Davidson, M.C., Johnson, S.P., Glover, G. & Casey, B.J. (2005). Contributions of the hippocampus and the striatum to simple association and frequency-based learning. *Neuroimage, 27,* 291-298.

Arciuli, J. & Simpson, I. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36(2),* 286-304.

Aslin, R., Saffran, J. & Newport, E. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science, 9(4)*, 321-324.

Audacity Team (2012). Audacity®. Version 2.0.0. Audio editor and recorder. Available from: http://audacityteam.org/.

Baker, C., Olson, C. & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15(7),* 460-466.

Bartels, S., Darcy, I. & Höhle, B. (2009) Schwa syllables facilitate word segmentation for 9-month-old German-learning infants. In J. Chandlee, M. Franchini, S. Lord & Rheiner, G. (Eds.) *BUCLD 33: Proceedings of the 33rd Annual Boston University Conference on Language Development*. 73-84. Somerville M.A.: Cascadilla Press.

Batchelder, E. (2002). Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition, 82(2),* 167-206.

Bates, E., Dale, P. & Thal, D. (1995). Individual differences and their implications for theories of language development. In Fletcher, P. & MacWhinney, B. (Eds), *Handbook of child language.* Oxford: Basil Blackwell.

Bates, D., Kliegl, R., Vasishth, S. & Baayen, R.H. (2015). Parsimonious mixed models. *RePsychLing* R Package.

Beatty, J. & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of Psychophyscology, 2,* 142-162.

Beckman, M. & Pierrehumbert, J. (1986). Intonational Structure in Japanese and English. *Phonology, 3,* 255-309.

Bhatara A., Boll-Avetisyan N., Unger A., Nazzi, T. & Höhle B. (2013). Native language affects rhythmic grouping of speech. *Journal of the Acoustic Society of America, 134,* 3828-3843.

Bijeljac-Babic, R., Höhle, B. & Nazzi, T. (2012). Effect of bilingualism on lexical stress pattern discrimination in French-learning infants. *PLoS One, 7(2)*.

Bijeljac-Babic, R., Höhle, B. & Nazzi, T. (2016). Early Prosodic Acquisition in Bilingual Infants: The Case of the Perceptual Trochaic Bias. *Frontiers in Psychology,* 7.

Bion, R., Benavides-Varela, S. & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and Speech, 54,* 123-140.

Black, A. & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the*

39th Annual Meeting of the Cognitive Science Society (124-129). Austin, TX: Cognitive Science Society.

Bolinger, D. (1989). Intonation and Its Uses: Melody in Grammar and Discourse. Stanford University Press.

Brent, M. & Cartweight, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93-125.

Bulf, H., Johnson, S. & Valenza, E. (2011) Visual statistical learning in the newborn infant. *Cognition, 121,* 127-132.

Bulgarelli, F., Benitez, V., Saffran, J., Byers-Heinlein, K., Weiss, D. (2017). Statistical learning of multiple structures by 8-month-old infants. *Proceedings of the 41st annual BUCLD,* LaMendola, M. & Scott, J., (Eds), 128-139. Somerville, MA: Cascadilla Press.

Cairns, P., Shillcock, R., Chater, N. & Levy, J. (1997). Bootstrapping Word Boundaries: A Bottom-up Corpus-Based Approach to Speech Segmentation. *Cognitive Psychology, 33(2),* 111-153.

Cardillo, G. C. (2010). *Predicting the predictors.* University of Washington.

Christophe, A., Gout, A., Peperkamp, S. & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics, 31,* 585-598.

Christophe, A., Guasti, T., Nespor, M., Dupoux, E. & Van Ooyen, B. (1997). Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition. *Language and Cognitive Processes, 12,* 585-612.

Christophe, A., Mehler, J. & Sebastián-Gallés, N. (2001). Perception of Prosodic Boundary Correlates by Newborn Infants. *Infancy, 2(3),* 385-394.

Christophe, A., Nespor, M., Guasti, T. & Van Ooyen, B. (2003). Prosodic structure and syntactic acquisition: The case of the head-direction parameter. *Developmental Science, 6(2),* 211-220.

Chun, M. M. & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36(1),* 28-71.

Cole, A. & Jakimik, J. (1980). How are syllables used to recognize words? *The Journal of the Acoustical Society of America, 67(3)*, 965-970.

Colombo, J. & Bundy, R. (1983). Infant response to auditory familiarity and novelty. *Infant Behavior & Development, 6(3),* 305-311.

Coltheart, M., Besner, D., Jonasson, J. & Davelaar, E. (1979). Phonological encoding in the lexical decision task. *The Quarterly Journal of Experimental Psychology, 31(3)*, 489-507.

Conboy, B., Rivera-Gaxiola, M., Klarman, L., Aksoylu, E. & Kuhl, P. (2005) Associations between native and nonnative speech sound discrimination and language development at the end of the first year. In: Brugos A., Clark-Cotton, M., Ha S. (Eds). *Supplement to the Proceedings of the 29th Boston University Conference on Language Development*.

Conboy, B., Rivera-Gaxiola, M., Silva-Pereyra, J. & Kuhl, P.K. (2008). "Event-related potential studies of early language processing at the phoneme, word, and sentence levels." In A.D. Friederici, A. & Thierry, G. (Eds.), *Early language development: bridging brain and behavior, Trends in language acquisition research series*, 5, 24-64. Amsterdam/The Netherlands: John Benjamins.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M. & Hagoort, P. (2014) Infant predictors of language. *Child Development, 85(4),* 1330-1345.

Cristia, A., Seidl, A., Singh, L. & Houston, D. (2016). Test-Retest Reliability in Infant Speech Perception Tasks. *Infancy, 21,* 648-667.

Curtin, S., Mintz, T. & Christiansen, M. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96,* 233-262.

Cutler, A. & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics, 21(1)*, 103-108.

Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 113-121.

De Carvalho, A., Dautriche, I., Millote, S. & Christophe, A. (2018). Early perception of phrasal prosody and its role in syntactic and lexical acquisition. In *The Development of Prosody in First Language Acquisition*, Edition: Trends in Language Acquisition Research. John Benjamins, 17-35.

Delattre, P. (1963). Comparing the Prosodic Features of English, German, Spanish and French. *IRAL, 1(3),* 193-210.

DePaolis, R., Portnoy, T. & Vihman, M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Frontiers in Psychology, 7,* 715.

Dogil, G. & Williams, B. (1999). The phonetic manifestation of word stress. In Harry van der Hulst (ed.), *Word Prosodic Systems in the Languages of Europe.* Berlin: de Gruyter. 273-334.

Echols, C., Crowhurst, M. & Childers, J. (1997). The Perception of Rhythmic Units in Speech by Infants and Adults. *Journal of Memory and Language, 36(2),* 202-225.

Endress, A. & Hauser, M. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology, 61(2),* 177-199.

Endress, A. & Mehler, J. (2009a). Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology,* 1-21.

Endress, A. & Mehler, J. (2009b). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60,* 351-367.

Engelhardt, P., Ferreira, F. & Patsenko, E. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology, 63 (4),* 639-645.

Erickson, L., Kaschak, M.,Thiessen, E. & Berry, C. (2016). Individual Differences in Statistical Learning: Conceptual and Measurement Issues. *Collabra, 2(1),* 1-17.

Erickson, L., Thiessen, E. & Graf Estes, K. (2014). Statistically coherent labels facilitate categorization in 8-month-olds. *Journal of Memory and Language, 72,* 49-58.

Evans, J., Saffran, J. & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 2,* 321-335.

Fernandes, T., Ventura, P. & Kolinky, R. (2007). Statistical information and coarticulation as cues to word boundaries: A matter of quality of the signal. *Perception & Psychophysics 69,* 856-864.

Finn, A. & Hudson Kam, C. (2008). The curse of knowledge: first language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108(2)*, 477-499.

Fiser, J. & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499-504.

Frank, M., Goldwater, S., Griffiths, T. & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117(2)*, 107-125.

Friederici A., Friedrich, M. & Christophe A. (2007). Brain responses in 4-month-old infants are already language specific. *Current Biology, 17*, 1208-1211.

Friederici, A., Friedrich, M. & Weber, C. (2002) Neural manifestation of cognitive and precognitive mismatch detection in early infancy. *Neuroreport, 13,* 1251-1254.

Friedrich, M., Weber, C. & Friederici, A. D. (2004). Electrophysiological evidence for delayed mismatch response in infants at-risk for specific language impairment. *Psychophysiology, 41*(5), 772-782.

Fritzsche, T. & Höhle, B. (2015). Phonological and lexical mismatch detection in 30-month-olds and adults measured by pupillometry. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow.

Gambell, T. & Yang, C. (2006). Word Segmentation: Quick but not Dirty. *34th Northeaster Linguistic Society meeting*, Yale University.

Gebhart, A., Newport, E. & Aslin, R (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review, 16(3),* 486-490.

Gerken L. (1996). Prosody's role in language acquisition and adult parsing. *Journal of Psycholinguist Research, 25(2),* 345-56.

Gerken, L. & Aslin, R. (2005). Thirty years of research on infant speech perception: The legacy of Peter W. Jusczyk. *Language Learning and Development, 1*(1), 5-21.

Gervain, J., Nespor, M., Mazuka, R., Horie, R. & Mehler, J. (2008) Bootstrapping word order in prelexical infants: A Japanese-Italian crosslinguistic study. *Cognitive Psychology, 57,* 56-74.

Gleitman, L. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In *E. Wanner & L. R. Gleitman (Ed.), Language acquisition: State of the art*. New York: Cambridge University Press, 3-48.

Gleitman, L. (1990). The Structural Sources of Verb Meaning. *Language Acquisition, 1(1),* 3-55.

Gout, A., Christophe, A. & Morgan (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language, 51*, 548-567.

Goyet, L., Millotte, S., Christophe, A. & Nazzi, T. (2016). Processing Continuous Speech in Infancy: From Major Prosodic Units to Isolated Word Forms. In Jeffrey, L., Synder, W. & Pater, J. (Eds.). *The Oxford Handbook of Developmental Linguistics:* Oxford.

Graf Estes, K. & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental psychology, 51(11)*, 1517.

Graf Estes, K., Evans, J. & Else-Quest, N. (2007). Differences in the Nonword Repetition Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech Language and Hearing Research ,50(1),* 177-195.

Graf Estes, K., Evans, J., Alibali, M. & Saffran, J. (2007). Can infants map meaning to newly segmented words? *Psychological Science, 18,* 254-260.

Graf Estes, K., Gluck, S. & Bastos, C. (2015). Flexibility in Statistical Word Segmentation: Finding Words in Foreign Speech. *Language Learning and Development, 11(3),* 252-269.

Gredebäck, G. & Melinder, A. (2010). Infants' understanding of everyday social interactions: a dual process account. *Cognition, 114(2)*, 197-206.

Grimm, H. & Doil, H. (2006). *ELFRA. Elternfragebögen für die Früherkennung von Risikokindern*. Göttingen: Hogrefe.

Guasti, T., Nespor, M., Christophe, A. & van Ooyen, B. (2001). Prosodic structure and syntactic acquisition: the case of the head-direction parameter. *Developmental Science 6(2),* 213-222.

Hauser, M., Newport, E. & Aslin, R. (2001). Segmentation of the speech stream in a nonhuman primate: Statistical learning in cotton top tamarins. *Cognition, 78,* B53-B64.

Hay J. & Diehl, R. (2007). Perception of Rhythmic Grouping: Testing the Iambic/Trochaic Law. P*erception & Psychophysics, 69(1),* 113-122.

Hay J. & Saffran J. (2012). Rhythmic grouping biases constrain infant statistical learning. *Infancy, 17,* 610-641.

Hay, P., Pelucchi, B., Graf Estes, K. & Saffran, J. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology, 63(2),* 93-106.

Hayes, J. & Clark, H. (1970). Experiments on the segmentation of an artificial speech analog. In Hayes, J. (Ed.)*, Cognition and the development of language,* 221-234. New York: Wiley.

Hebb, D. (1961). Distinctive features of learning in the higher animal. In Delafresnaye, B. (Ed.), *Brain mechanisms and learning,* 37-46. Oxford: Blackwell.

Hepach, R. & Westermann, G. (2013). Infants' sensitivity to the congruence of others' emotions and actions. *Journal of Experimental Child Psychology, 115,* 16-29.

Hepach, R. & Westermann, G. (2016). *Pupillometry in infancy research.* 359-377.

Herold, B., Höhle, B., Walch, E., Weber, T. & Obladen, M. (2008). Impaired word stress pattern discrimination in very-low-birthweight infants during the first 6 months of life. *Developmental Medicine and Child Neurology, 50 (9),* 678-83.

Hess, E. & Polt, J. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132,* 349-350.

Hirsh-Pasek, K., Kemler Nelson, D., Jusczyk, P., Cassidy, K., Druss, B. & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26(3),* 269-286.

Hochmann, J.-R. & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological Science, 25(11),* 2038-2046.

Höhle, B. (2002). Der Einstieg in die Grammatik: Die Rolle der Phonologie/Syntax-Schnittstelle für Sprachverarbeitung und Spracherwerb. Habilitationsschrift Freie Universität Berlin.

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics, 47 (2),* 359-382.

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J. & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior and Development, 32(3),* 262-274.

Höhle, B., Giesecke, D. & Jusczyk, P. (2001). Word segmentation in a foreign language: Further evidence for crosslinguistic strategies. *The Journal of the Acoustical Society of America 110,* 2687.

Höhle, B., Pauen, S., Hesse, V. & Weissenborn, J. (2014). Discrimination of Rhytmic Pattern at 4 Months and Language Performance at 5 Years: A Longitudinal Analysis of Data From German-Learning Children. *Language Learning, 64,* 141-164.

Höhle, B. & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed class elements in continuous speech. *Developmental Science 6(2).*

Houston, D., Horn, Qi, Ting & Gao (2007). Assessing speech discrimination in individual infants. *Infancy, 12,* 119-145.

Houston, D., Jusczyk, P., Kuljpers, C., Coolen, R. & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review, 7 (3),* 504-509.

Hunter, M., Ames, E. & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology, 19(3),* 338-352.

Jackson, I. & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental Science, 12(4)*, 670-679.

Johnson, E. & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44(4)*, 548-567.

Johnson, E. & Jusczyk, P. (2003). Exploring possible effects of language--specific knowledge on infants' segmentation of an artificial language. Jusczyk lab final report, 141-148.

Johnson, E. & Seidl, A. (2009). At 11 months, prosody still outranks statistics. *Developmental Science, 12(1),* 131-141.

Johnson, E. & Tyler, M. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science, 13(2),* 339-345.

Johnson, E., Seidl, A. & Tyler, M. (2014). The Edge Factor in Early Word Segmentation: Utterance-Level Prosody Enables Word Form Extraction by 6-Month-Olds. *PLoS One, 9(1).*

Jones, J. & Pashler, H. (2007). Is the Mind Inherently Forward-Looking? Comparing Prediction and Retrodiction. *Psychonomic Bulletin & Review, 14,* 295-300.

Junge C., Kooijman V., Hagoort P. & Cutler A. (2012). Rapid recognition at 10 months as a predictor of language development. *Developmental Science 15,* 463-473.

Jusczyk, P. & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 28,* 1-23.

Jusczyk, P. & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by two-month-old infants. *Perception & Psychophysics, 23,* 105-109.

Jusczyk, P. W. (1997). *Language, speech, and communication. The discovery of spoken language.* Cambridge, MA, US: The MIT Press.

Jusczyk, P., Cutler, A. & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology, 46 (1),* 65-97.

Jusczyk, P., Cutler, A. & Redanz, N. (1993). Infants' Preference for the Predominant Stress Patterns of English Words. *Child Development, 64(3),* 675-687.

Jusczyk, P., Houston, D. & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159-207.

Jusczyk, P., Kemler-Nelson, D., Hirsh-Pasek, K., Kennedy, L., Woodward, A. & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24,* 252-293.

Just, M. & Carpenter, P. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47(2),* 310-330.

Kahneman, D. & Beatty, J. (1966). Pupil diameter and load memory. *Science, 1966,* 154, 1583-1585.

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Developmental Review, 27(3),* 283-348.

Kelly, M. & Bock, J. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance, 14(3),* 389-403.

Kidd, E. & Arciuli, J. (2016). Individual Differences in Statistical Learning Predict Children's Comprehension of Syntax. *Child Development, 87(1),* 184-193.

Kim, R. Seitz, A., Feenstra, H. & Shams, L. (2009). Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience Letters, 461(2),* 145-149.

Kirkham, N., Slemmer, J. & Johnson, S. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83(2),* 35-42.

Kooijman, V., Hagoort, P. & Cutler, A. (2009). Prosodic Structure in Early Word Segmentation: ERP Evidence From Dutch Ten-Month-Olds. *Infancy, 14 (6),* 591- 612.

Krogh, L., Vlach, H. & Johnson, S. (2012). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology, 3,* 598.

Kudo, N., Nonaka, Y., Mizuno, N., Mizuno, K. & Okanoya, K. (2011). On-line statistical segmentation of a non-speech auditory stream in neonates as demonstrated by event-related brain potentials. *Developmental Science, 14(5)*, 1100-1106.

Kuhl, P., Conboy, B., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. &Nelson T. (2008). Early phonetic perception as a pathway to language: New data and native language magnet theory, expanded. *Philosophical Transactions of the Royal Society,* 979-1000.

Kuhl, P., Conboy, Padden, Nelson & Pruitt (2005). Early Speech Perception and Later Language Development: Implications for the "Critical Period". *Language Learning and Development, 1*, 237-264.

Kuijpers, C., Coolen, R., Houston, D. & Cutler, A. (1998). Using the head-turning technique to explore cross-linguistic performance differences. In Rovee-Collier, C., Lipsitt, L. & Hayne, H. (Eds.). *Advances in infancy research,* 205-220. London: Ablex.

Laeng, B., Sirois, S. & Gredebäck, G. (2012). Pupillometry: A Window to the Preconcious? *Perspectives on Psychological Science, 7(1),* 18-27.

Langus, A., Marchetto, E., Bion, R. & Nespor, M. (2012). an prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language, 66(1),* 285-306.

Lehiste, I. (1970). Suprasegmentals. Cambridge: MIT Press.

Lew-Williams, C. & Saffran, J. (2012). All words are not created equal: expectations about word length guide infant statistical learning. *Cognition, 122,* 241-246.

Lew-Williams, C., Pelucchi, J. & Saffran, J. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science, 14(6),* 1323-1329.

Liss, J., Spitzer, S., Caviness, J., Adler, C. & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *Journal of the Acoustical Society of America, 104(4),* 2457-2566.

Mainela-Arnold, E. & Evans, J. (2014). Do statistical segmentation abilities predict lexical-phonological and lexical-semantic abilities in children with and without SLI? *Journal of Child Language, 41 (2),* 327-351.

Männel, C., Schipke, C. S. & Friederici, A. (2013). The role of pause as a prosodic boundary marker: Language ERP studies in German 3- and 6-year-olds. *Developmental Cognitive Neuroscience, 5,* 86-94.

Mattys, S., Jusczyk, P., Luce, P. & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465-494.

Mattys, S., White, L. & Melhorn, J. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134,* 477-500.

Matuschek, H., Kliegl, R., Vasisth, S., Baayen, H. & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305-315.

Maye, J., Weiss, D. & Aslin, R. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science, 11,* 122-134.

Maye, J., Werker, J. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

Mersad, K. & Nazzi, T. (2012). When Mommy Comes to the Rescue of Statistics: Infants Combine Top-Down and Bottom-Up Cues to Segment Speech. *Language Learning and Development 8(3),* 303-315.

Meyer, D. & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Mintz, T., Newport, E. & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26,* 393-425.

Mirman, D., Magnuson, J., Graf Estes, K. & Dixon, J. (2008). The link between statistical segmentation and word learning in adults. *Cognition, 108(1)*, 271-280.

Moon, C., Cooper, R. & Fifer, W. (1993). Two-Day-Olds Prefer Their Native Language. *Infant Behavior and Development, 16,* 495-500.

Morgan, J. & Demuth, K. (Eds) (1987). *Signal to syntax: Bootstrapping from speech togrammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Morgan, J. & Saffran, J. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66,* 911-936.

Morgan, J. (1986). *From simple input to complex grammar*. Cambridge, MA: MIT Press.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24(3),* 756-66.

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S. & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *Journal of Memory and Language, 54*, 283-299.

Nazzi, T., Jusczyk, P. & Johnson, E. (2000). Language Discrimination by English-Learning 5-Month-Olds: Effects of Rhythm and Familiarity. *Journal of Memory and Language, 43*, 1-19.

Nazzi, T., Paterson, S. & Karmiloff-Smith, A. (2003). Early word segmentation by infants and toddlers with Williams syndrome. *Infancy, 4(2),* 251-271.

Nespor, M. & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris Publications.

Nespor, M., Guasti, T. & Christophe, A. (1996). Selecting word order: the rhythmic activation principle. In Kleinhenz, U. (Ed.), *Interfaces in phonology,* 1-26. Berlin: Akademie Verlag.

Nespor, M., Mehler, J., Shukla, M., Peña, M. & Gervain, J. (2007). On different mechanisms involved in the acquisition of language. In: N.C.D. Khuong and Riche, S. Sinha (eds.) *The Fifth Asian GLOW. Conference Proceedings.* Mysore. Central Institute of Indian Languages. 291-314.

Nespor, M., Shukla, M., Van de Vijver, R., Avesani, C., Schraudolf, H. & Donati, C. (2008). Different phrasal prominence realizations in VO and OV languages. *Lingue e Linguaggio, VII(2)*, 1-19.

Newman, R., Ratner, N., Jusczyk, A., Jusczyk, P. & Dow, K. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology, 42 (4),* 643-655.

Newman, R., Rowe, M. & Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed-speech and infant processing skills in language development. *Journal of Child Language, 43(5),* 1158-1173.

Newsom, M. (2018). Generalizing across gender during early word learning: Evidence from a statistical learning paradigm. University of Tennessee Honors Thesis Projects. Knoxville.

Nissen, M. & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19(1)*, 1-32.

Noiray, A., Abakarova, D., Rubertus, E., Krüger, S. & Tiede, M. (2018). How children organise their speech in the first years of life? Insight from ultrasound imaging. *Journal of Speech, Language, and Hearing Research, 61,* 1355-1368.

Obeid, R., Brooks, P., Powers, K., Gillespie-Lynch, K. & Lum, J. (2016). Statistical learning in Specific Language Impairment and Autism Spectrum Disorder: A meta-analysis. *Frontiers in Psychology, 7,* 1245.

Olson, I. & Chun, M. (2001). Temporal contextual cuing of visual attention. J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 27(5),* 1299-1313.

Onishi, K., Chambers, K. & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition. 2002, 83,* B13-B23.

Pelucchi, B. Hay, J. & Saffran, J. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80(3),* 674-685.

Pelzer, L. & Höhle, B. (2006) Processing of morphological markers as a cue to syntactic phrases by 10 month-old German-learning infants. In Belletti, A., Bennati, E., Chesi, C., DiDomenico, E. & Ferrari, I. (Eds.) *Language Acquisition and Development: Proceedings of GALA2005*, 411-422. Cambridge: Cambridge Scholars Press.

Peña, M., Bonatti, L., Nespor, M. & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604-607.

Perea, M. & Carreiras, M. (2003). Sequential effects in the lexical decision task: The role of the item frequency of the previous trial. *The Quarterly Journal of Experimental Psychology, 56A (3)*, 385-401.

Perruchet, P. & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36(7),* 1299-1305.

Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Science, 10(5),* 233-238.

Pierrehumbert, J. & Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Polka, L. & Sundara, M. (2012). Word Segmentation in Monolingual Infants Acquiring Canadian English and Canadian French: Native Language, Cross-Dialect, and Cross-Language Comparisons. *Infancy, 17(2),* 198-232.

Pons, F. & Bosch, L. (2010). Stress Pattern Preference in Spanish-Learning Infants: The Role of Syllable Weight. *Infancy 15 (3),* 223-245.

Privitera, C., Renniger, L., Carney, T., Klein, S. & Aguilar M. (2010). Pupil dilation during visual target detection. *Journal of Vision, 10(3).*

Ramus, F., Hauser, M., Miller, C., Morris, D. & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288,* 349-351.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73(3)*, 265-292.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6(6)*, 855-863.

Rivera-Gaxiola, Klarman, Garcia-Sierra & Kuhl, P. (2005). Neural patterns to speech and vocabulary growth in American infants. *NeuroReport*, *16*, 495-498.

Roder, B., Bushnell, E. & Sasseville A. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy, 1*, 491-508.

Romberg, R. & Saffran, J. (2010). Statistical learning and language acquisition. Wires *Cognitive Science, 1(6),* 906-914.

Rovee-Collier, C. (1999). The Development of Infant Memory. *Current directions in Psychological Science, 8(3),* 80-85.

Rubertus, E. & Noiray, A. (2018). On the development of gestural organization: A cross-sectional study of vowel-to-vowel anticipatory coarticulation. *PLOS One*.

Saffran, J. & Kirkham, N. (2017). Infant Statistical Learning. *Annual Review of Psychology, 69,* 181-203.

Saffran, J. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition, 81,* 149-69.

Saffran, J. (2008). What Can Statistical Learning Tell Us About Infant Learning? In Woodward, A. & Needham, A. (Eds). *Learning and the Infant Mind.* Oxford Scholarship Online.

Saffran, J., Johnson, E., Aslin, R. & Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70(1)*, 27-52.

Saffran, J., Newport, E. & Aslin, R. (1996a). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606-621.

Saffran, J., Aslin, R. & Newport, E. (1996b). Statistical learning by 8-month-olds. *Science, 274*, 1926-1928.

Saksida, A., Langus, A. & Nespor, M. (2016). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science,* 1-11.

Sansavini, A., Bertoncini, J. & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology, 33(1),* 3-11.

Segal, O. & Kison-Rabin, L. (2016). Evidence for language-specific influence on the preference of stress patterns in infants learning an Iambic language (Hebrew). *Journal of Speech, Language and Hearing Research, 55(5),* 1329-1341.

Seidl, A. & Cristia, A. (2012). Infants' learning of phonological status. *Frontiers in Psychology, 3,* 448.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57,* 24-48.

Shattuck-Hufnagel, S. & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25 (2),* 193-247.

Shi, R., Werker, J. & Morgan, J. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72,* B11-B21.

Shukla, M., Nespor, M. & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology, 54,* 1-32.

Shukla, M., White, K. & Aslin, R. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-month-old infants. *PNAS, 108(15)*, 6038-6043.

Siegelman, N. & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 1(81),* 105-120.

Siegelman, N., Bogaerts, L. & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods, 49(2),* 418-432.

Singh, L., Reznick, S. & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Developmental Science, 15*, 482-495.

Skoruppa, K., Cristia, A., Peperkamp, S. & Seidl, A. (2011). English-learning infants' perception of word stress patterns. *Journal of the Acoustical Society of America, 130,* 50-55.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., Limissuri, R. & Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. *Developmental Science, 12(6),* 914-919.

Smith, M., Cutler, A., Butterfield, S. & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech & Hearing Research, 32(4)*, 912-920.

Soderstrom, M., Seidl, A., Kemler Nelson, D. & Jusczyk, P. (2003). The Prosodic Bootstrapping Of Phrases: Evidence From Prelinguistic Infants. *Journal Of Memory And Language, 49 (2),* 249-267.

Soderstrom M., Kemler Nelson D. & Jusczyk P. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development 28*, 87-94.

Sohail, J. & Johnson, E. (2016). How Transitional Probabilities and the Edge Effect Contribute to Listeners' Phonological Bootstrapping Success. *Language Learning and Development, 12(2),* 105-115.

Spring, D. & Dale, P. (1977). Discrimination of Linguistic Stress in Early Infancy. *Journal of Speech, Language, and Hearing Research, 20,* 224-232.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86-132.

Szagun, G., Schramm, S & Stumper, B. (2009). *Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform)*. Pearson Assessment & Information.

Tamási, K., McKean, C., Gafos, A., Fritzsche, T. & Höhle, B. (2017). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *Journal of Experimental Child Psychology,* 140-148.

Teinonen, T., Fellmann, R., Näätänen, R., Alku, P. & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *Neuroscience, 10(1),* 21.

Thiessen, E. & Erickson, L. (2013). Discovering Words in Fluent Speech: The Contribution of Two Kinds of Statistical Information. *Frontiers in Psychology, 3(590),* 1-10.

Thiessen, E. & Saffran, J. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39(4),* 706-16.

Thiessen, E. & Saffran, J. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3,* 73-100.

Thiessen, E., Hill, E. & Saffran, J. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 49-67.

Toro J. M., Sebastian-Gallés N. & Mattys S. (2009). The role of perceptual salience during the segmentation of connected speech. *European. Journal of Cognitive Psychology, 21,* 786-800.

Toro, J. M. & Trobalón, J. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics, 67(5),* 867-875.

Tromp, J., Haagort, P. & Meyer, A. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. *Quarterly Journal of Experimental Psychology, 69,* 1093-1108.

Tsao, F., Liu, H. & Kuhl, P. (2004). Speech perception in infancy predicts language development in the second year of life: a longitudinal study. *Child Development, 75(4),* 1067-1084.

Turk, A., Jusczyk, P. & Gerken, L. (1995). Do English-learning infants use syllable weight to determine stress? *Language and Speech, 38,* 143-158.

Turk-Browne, N., Jungé J. & Scholl, B. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134,* 552-564.

Tyler, M. & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America, 126(1)*, 367-376.

Venables, W. & Ripley, B. (2002). Modern Applied Statistics with S. Springer Science + Business Media, LCC: New York.

Vogelzang, M., Van Rijn, H. & Hendriks, P. (2016) Pupillary responses reflect ambiguity resolution in pronoun processing. *Language, Cognition and Neuroscience*, *31(7),* 876-885.

Vroomen, J., Tuomainen, J. & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language, 38(2),* 133-149.

Weber, C., Hahne, A., Friedrich, M. & Friederici, A. (2005). Reduced stress pattern discrimination in 5-month-olds as a marker of risk for later language impairment: neurophysiological evidence. *Brain Research: Cognitive Brain Research, 25(1),* 180-7.

Weissenborn, J. & Höhle, B. (2001). (Eds). *Approaches to Bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition*. John Benjamins Publishing Company.

Wellmann, C. Holzgrefe, J., Truckenbrodt, H., Wartenburger, I. & Höhle, B. (2012). How each prosodic boundary cue matters: Evidence from German infants. *Frontiers in Psychology, 3,* 1-13.

Wickham, H. (2009). *ggplot2, Use R.* Springer Science + Business Media, LCC: New York.

Yang, C. (2004). Universal grammar, statistics, or both. *Trends in Cognitive Sciences, 8,* 451-456.