The company language keeps:

How distributional cues influence statistical learning for language

Katja Stärk

The company language keeps:

How distributional cues influence statistical learning for language

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 8 april 2024
om 10.30 uur precies

door

Katja Sophia Stärk
geboren op 9 juni 1992
te Leipzig, Duitsland

The company language keeps:

How distributional cues influence statistical learning for language

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, April 8, 2024
at 10.30 am

by

Katja Sophia Stärk
born on June 9, 1992
in Leipzig, Germany

Für meine Eltern

Table of contents

Chapter 1

General Introduction

"Already know you that which you need."
Master Yoda – George Lucas

Humans are exquisitely adept at finding patterns in their environment, including their language input. Take the following two examples: First, imagine a baby listening to their mother speaking. Even though the mother might adjust her language to the child, speaking in child-directed speech, she will not mark word boundaries in any obvious way comparable to blank spaces between words in written language. In the beginning, the baby does not understand anything, and in order to learn their first words, they first need to identify the words within their continuous speech input; but neither does the baby know anything about how the language works or to what they should pay attention. However, they can rely upon their cognitive abilities to detect patterns, and sure enough, they will start to recognise words themselves within their first year of life (Bergelson & Swingley, 2012).

Second, imagine yourself living in a foreign country where the community language is a language you do not know. Since you have not attended a formal class, you will not have any knowledge of the formal rules (i.e., patterns) of the language. Consequently, you do not understand much in the beginning. Yet, after a little while, you pick up on some words and phrases such as, for instance, *hallo* 'hello' and *Guten Morgen* 'good morning'; later you notice that verbs end in *-e* when they describe actions of the speaker (*ich* 'I') but in *-st* when they describe actions of the listener (*du* 'you').

In both examples, statistical learning (SL) likely plays a key role in detecting frequent usage patterns and in generalising across multiple individual instances to begin the process of constructing a language. SL can be described as the process of learning temporal or spatial co-occurrence patterns in the environment (cf. Frost et al., 2019). In other words, when two or more things often occur together, we pick up on this and learn the pattern. This has been found from birth and across the

human lifespan (e.g., Aslin et al., 1998; Saffran, Aslin, et al., 1996; Saffran et al., 1997; Teinonen et al., 2009). In the first example above, the baby listens to speech in which the boundaries between words are not clear. Therefore, the baby might, for instance, first implicitly chunk together syllable pairs that frequently co-occur to form likely word candidates in the language. On the basis of these first word candidates, they can then draw conclusions about other patterns (e.g., words mostly being stressed on the first syllable), which can then help to extract more words from the speech stream.

In the second example, the same happens to you in an unfamiliar language, showing that SL remains relevant when processing and learning language throughout our lifetime (Alexander et al., 2022; and see e.g., Batterink & Paller, 2017, Perruchet & Desaulty, 2008, and Saffran, Newport, et al., 1996, for further examples of adults acquiring a second, but artificial, language via SL). It is also found at different levels within our input. For example, two syllables can be chunked together based on their frequent co-occurrence in the word *hallo*, two words can be chunked together based on their frequent co-occurrence in the phrase *Guten Morgen*, and two non-adjacent morphemes can be associated with one another in the person and number agreement of the pronouns with the morphological verb markers *-e* and *-st*.

The two examples demonstrate the power of SL for language learning, which has held promise as a key mechanism underlying language acquisition and use since the publication of Saffran et al.'s (1996) seminal paper on SL in eight-month-old infants. Despite over a quarter of a century of research on SL, there are many aspects of SL and its role in language that are still unknown. The aim of this thesis was twofold: first, to examine the availability and reliability of word segmentation cues in natural child-directed speech (Chapter 2) and second, to investigate how prior knowledge which participants gained in their natural environment (as studied in Chapter 2) influences their subsequent language processing and word segmentation (Chapters 3 to

5). In the remainder of this chapter, I will address each of these topics in further detail.

Word segmentation in infancy

One of the first puzzles infants need to solve in order to acquire language is identifying the boundaries between words. That is, before a baby can start to learn their first words, they first need to find those words within their speech input. This is no easy feat since word boundaries are not marked in any one completely reliable way. However, infants can rely on a variety of cues to help them break into the speech stream. For instance, if words in a language are almost always stressed on the first syllable then stress can be considered to provide a reliable cue to word onset in that language. Crucially, the availability and reliability of cues varies between languages (Cutler, 2012). It is therefore important to gain an overview of the availability and reliability of a broad range of cues in a language in order to better understand the input babies receive. Establishing the landscape of cues for a given language sheds light on the learners' pathway to language acquisition by permitting theorising on how children draw on these cues during learning and by opening the door to cross-linguistic comparison which can critically shape our understanding.

In this thesis, I studied five potential word segmentation cues in German child directed speech: word stress, transitional probabilities (TPs), lexical and sublexical frequencies, word length, and single word utterances. I discuss these cues below.

Word stress

In stress-timed languages such as English, word stress can be a good indicator for word boundaries. In these languages, at least one syllable per word often stands out by being louder, higher in pitch, or longer in duration. When this prominent marker follows a regular pattern, it can be considered a reliable cue for word segmentation. According to the World Atlas of Language Structure (WALS), Irish and

Czech, for example, follow a word-initial stress pattern while Hebrew and Berber follow a word-final stress pattern (Goedemans & van der Hulst, 2013a). This means that in Irish or Czech, where the first syllable of a word is stressed, a stressed syllable indicates to the child that a new word starts. In Hebrew and Berber, on the other hand, where the last syllable of a word is stressed, a stressed syllable indicates to the child that a word ends.

Research has shown that babies are indeed sensitive to word stress (Sansavini et al., 1997; Spring & Dale, 1977). That is, they can distinguish between different stress patterns and develop a preference for the pattern found in their native language within their first year of life (Jusczyk et al., 1993). Furthermore, they actively use this stress pattern to establish word boundaries, with infants initially also inferring these boundaries in the wrong position when words do not follow the dominant pattern (Jusczyk et al., 1999). This shows how heavily young (American-English) infants seem to rely on stress as a cue to word segmentation, assuming word boundaries also in the wrong location until they learn to rely on multiple cues, helping them to correctly assign word boundaries in atypically stressed words at the age of 10.5 months (Jusczyk et al., 1999).

Most of the research on language acquisition has investigated phenomena in the English language (Kidd & Garcia, 2022). Even though English and German are closely related, one cannot assume that empirical findings in the English language also apply to the German language, including the availability and reliability of segmentation cues. However, research suggests that German five-month-old infants can also discriminate between iambic and trochaic stress patterns, equally showing a preference for trochaic over iambic stress, similar to the English infants (Friederici et al., 2007; Höhle et al., 2009; Tippmann, 2015; Weber et al., 2004), and they have also been shown to use stress to segment words from speech (Höhle et al., 2001).

This suggests that word stress is an important cue to word segmentation in stress-timed languages such as English or German. In English, a study revealed that 90% of content words are stressed on their initial syllable (Cutler & Carter, 1987), supporting the children's intuition to heavily rely on this cue as it proves to be highly reliable. In Chapter 2, I investigated how reliable the stress pattern is in German child-directed speech.

Transitional probabilities

A second potential cue to word segmentation are TPs[1]. TPs between syllables indicate how often one syllable occurs in combination with another syllable compared to how often it occurs overall. TPs within words are generally higher than TPs at word boundaries (Saksida et al., 2017), meaning that infants can chunk together syllable pairs with high TPs or assume word boundaries at low-TP transition points. Take, for instance, the phrase *pretty baby* – the syllable *pre* is often followed by the syllable *ty* in English, resulting in a high TP of .8, while the syllable *ty* is only rarely followed by the syllable *ba* in English, resulting in a low TP of .0003 (Saffran, 2003). The high TP of the syllable pair *pre ty* is an indication that the two syllables form a word while the low TP of the syllable pair *ty ba* is an indication that the two syllables might belong to different words.

This is an important cue to word segmentation because it does not require any prior knowledge to break into the speech stream. To use word stress, for example, the baby first needs to know which stress pattern their language follows. To be able to form a hypothesis about the pattern, they first need to know a few words over which they can generalise. In comparison, TPs can be used without any initial hypotheses. Babies can implicitly chunk together syllable pairs that

---

[1] Note that I refer to TPs throughout this thesis as a measure of how frequently syllables co-occur. I do not intend to suggest that humans consciously track those TPs but rather that the frequent co-occurrence strengthens the connection between those syllables, with high TPs indicating such frequent co-occurrence (see e.g., Perruchet, 2019, for a review).

often occur in combination. These first chunks can then help to form hypotheses about other cues to wordhood, which might then be more reliable or prominent in their language.

As already established, learners of all ages have been found to be sensitive to TPs as a word segmentation cue (e.g., Aslin et al., 1998; Saffran, Aslin, et al., 1996; Saffran et al., 1997; Teinonen et al., 2009). In these experiments, the participants were able to segment words from continuous speech relying solely on TPs to guide their segmentation. Remarkably few studies have been conducted on SL in German. It has been shown that German adults can segment words from speech based on TPs (Matzinger et al., 2021), but adults as well as six-month-old infants rely more on word stress than TPs when the two cues collide (Marimon Tarter, 2019). However, TPs might still be an important cue for German babies to break into the speech stream and acquire the stress pattern of the language before stress becomes the dominant cue.

Importantly, TPs can not only be determined within and between words but also in two different directions: Forwards TPs describe how likely the present syllable is followed by the next one (e.g., how likely the syllable *ba* is followed by the syllable *by* in the word *baby*) while backwards TPs describe how likely the present syllable is preceded by the previous one (e.g., how likely the syllable *by* is preceded by the syllable *ba* in the word *baby*). Both directions have been shown to be informative (Perruchet & Desaulty, 2008); however, which direction is more informative differs between languages. Infants have been shown to develop a preference for the direction which is more informative in their native language by the age of 13 months (Thiessen et al., 2019). It is therefore crucial to establish the precise numbers of within-word and between-word, forwards and backwards TPs in a language before investigating SL in that language (see also Monaghan & Rowland, 2017, for an account on how corpus analyses, experimental and computational studies can go hand in hand). In consequence, I established the precise TP distributions in German child-directed

speech (cf. Chapter 2) before investigating SL in German populations (cf. Chapters 3 to 5).

Lexical and sublexical frequencies

The frequency of syllable co-occurrences is, of course, not the only frequency that influences language acquisition. The influence of the frequency of a word on how quickly we learn it and how well we remember it is unquestionable (e.g., Ebbinghaus, 1885, 1913), a fact that has been directly incorporated into theories of child language acquisition (see Ambridge et al., 2015, and Lieven, 2010, for reviews). In natural languages, some words occur more frequently than others, following Zipf's law (Zipf, 1935, 1949). Zipf's law states that a word's token frequency and its rank are inversely related, which means that the most frequent word in a corpus occurs approximately twice as often as the second most frequent word, approximately three times as often as the third most frequent word, and so on. Such frequency distributions have been shown to facilitate word segmentation (Kurumada et al., 2013; but see Lavi-Rotbain & Arnon, 2022, for evidence that it is rather the predictability than the skew of the distribution which facilitates segmentation). Due to their salience, the frequent words stand out and can be segmented in the early exposure phases. Subsequently, they can act as anchors in the speech stream to aid segmentation of adjacent words (see e.g., Bortfeld et al., 2005; Peters, 1983; Pinker, 1984).

Zipf's observations were based on corpus data of written German (Kaeding, 1897). Since language changes over time and because spoken and child-directed speech might differ from written adult-directed speech, it is essential to study the frequency distribution of the child's input in German, which I did in Chapter 2. Additionally, I investigated the frequency distribution at the syllable level and also extended my analysis to syllable structures (i.e., consonant-vowel patterns underlying the syllables). The latter was included to examine whether some patterns are predominantly found word-initially or word-finally, guiding segmentation in a similar way as phonotactics.

Word length

Another factor that might influence word segmentation is word length. In natural languages, words normally have different word lengths (see e.g., Alonso et al., 2011, for an analysis of Spanish word length; and Li & Shirai, 2000, for an analysis of English word length). This can also be explained by Zipf's law, which states that languages are optimised for efficient communication (Zipf, 1935, 1949). This means that frequently repeated words are usually short, reducing the communicative effort for the speaker, while less frequently repeated words might be longer, potentially to reduce the input rate of new information for the listener (i.e., if all words were monosyllabic the listener would receive a lot of information within a short amount of time while longer words reduce this input rate).

In German, too, approximately 50% of words in a written language corpus have been found to be monosyllabic, approximately 30% disyllabic, and approximately 20% tri- or multisyllabic (Kaeding, 1897; Sigurd et al., 2004). However, variation in word length has been found to be detrimental for statistical word segmentation (Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012; but see Perruchet & Vinter, 1998, for computational counter-evidence). Therefore, child-directed speech might consist of shorter words, standardising the word lengths to potentially facilitate speech segmentation and language acquisition. This has been suggested for Norwegian child-directed speech in comparison to Norwegian adult-directed speech (Garmann et al., 2019). On the other hand, different word lengths have been suggested to guide word category assignment (i.e., suggesting that child-directed speech does not differ from adult-directed speech in word length), with monosyllabic words indicating that a word might belong to the closed class and multisyllabic words indicating that a word might belong to the open class (see e.g., Segal et al., 2009, for an analysis of Hebrew child-directed speech). In both cases it is relevant to know the exact distribution of word lengths in the speech input to German children, which I investigated in Chapter 2.

Single-word utterances

The final cue to word segmentation which I investigated for this thesis was single-word utterances. Sometimes words are produced in isolation, for instance, when we greet other people (*hello*), call them (*John*), or express our agreement or disagreement (*yes/no*). Because these words are detached from the continuous speech stream, they stand out to the listener and can be learnt more easily (Junge et al., 2012). Afterwards, these words can act as anchors in the subsequent input to aid the segmentation of adjacent words similar to the frequent words in the Zipfian distribution above (see e.g., Bortfeld et al., 2005; Peters, 1983; Pinker, 1984).

Previous research has shown that adults use single-word utterances in child-directed speech and even repeat approximately a third of those utterances in close temporal proximity, enhancing the probability that the child will pick up on those words (Aslin et al., 1996; Brent & Siskind, 2001). Corpus analyses of English and Hebrew child-directed speech suggest that approximately 15% of utterances comprise single words (MacWhinney & Snow, 1985; Segal et al., 2009) while another analysis of English gives an even higher estimate of approximately 26% of utterances (Monaghan & Christiansen, 2010). I expanded this investigation to German child-directed speech in Chapter 2, asking how many utterances comprised single words, how often words were repeatedly produced in isolation, and how often the child's name (as an especially prominent cue) was produced in isolation.

Statistical learning in child- and adulthood

Does the output of statistical learning influence future statistical learning?

After studying the availability and reliability of word segmentation cues in child-directed speech (cf. Chapter 2), the second part of my thesis is concerned with SL in child- and adulthood (cf. Chapters 3 to 5). We are already born with the ability to extract patterns from the

environment and can use this ability to segment words from continuous speech as described above. It has been proposed that during SL, syllables that often occur together are being chunked together, with this chunk then being stored in long-term memory (Batterink & Paller, 2017). However, little is known about how this knowledge influences future processing and learning.

That prior knowledge facilitates learning has long been known (Ebbinghaus, 1885, 1913). With regard to SL, studies showed that different kinds of prior knowledge already affect future learning in infancy (Lai & Poletiek, 2011; Lew-Williams & Saffran, 2012). For example, Lew-Williams & Saffran (2012) found that when nine- to ten-month-old infants expected to hear disyllabic words they could not segment trisyllabic words from continuous speech and vice versa. Studies in adults have found that experience with tone helps to segment other tonal languages (Potter et al., 2017; but see Wang & Saffran, 2014) and learning of simpler non-adjacent dependency structures facilitates learning of more complex non-adjacent dependency structures later on (Lany & Gómez, 2008; Zettersten et al., 2020).

These studies showed that a variety of prior expectations and experiences influences SL. Other studies investigated how the direct output of SL in long-term memory affects future SL, however, mostly in terms of phonotactics. Phonotactics describe linguistic patterns regarding the positions of phonemes within a syllable. For example, the phonemes /mn/ do not appear in this combination in English while they do in Russian (e.g., много 'a lot'). As the example shows, phonotactic patterns differ between languages and can be learned via SL. The output of this SL has been shown to guide future SL. For instance, Finn and Hudson Kam (2008) found that English adults relied on the phonotactics of their native language (i.e., SL knowledge from natural language) to segment novel words from continuous speech, even though the speech stream contained statistical cues suggesting different word boundaries (see also Mersad & Nazzi, 2011, for related findings in French; and Toro et al., 2011, for related findings in Catalan).

While the acquisition of phonotactics takes place at the phoneme level, SL is also often studied at the syllable level (see studies described above) because the chunking of syllables leads to the acquisition of words and the formation of a lexicon. However, little is known about how the learning unfolds and how the learned units are subsequently influencing the processing and learning of new linguistic material. In a recent EEG study, Batterink and Paller (2017) played participants continuous speech which comprised either novel words (structured condition) or completely random syllable combinations (unstructured condition). During the early phases of the exposure, participants' brains entrained to the syllable frequency of the speech. During the later phases of the exposure, however, participants of the structured condition became continuously more entrained to the word level frequency (which was not the case in the unstructured condition where the speech did not contain any words). This finding provides important insights into how SL unfolds over time and suggests that syllables get chunked together as learners identify frequently occurring syllable combinations, with the chunks (i.e., words) most likely being stored in long-term memory.

Regarding the question of whether the output of SL influences future SL, two highly relevant papers were published during my PhD. Siegelman et al. (2018) investigated why SL of visual and auditory material did not show a high positive correlation, even though SL was supposed to be a domain-general process. They conducted a series of experiments but most importantly, they found that participants' performance in SL of auditory linguistic material was influenced by how much individual experimental items resembled words in the participants' native language, as measured by independent ratings of each word. This provided indirect evidence that participants' prior knowledge gained via SL in the real world (the acquisition of Hebrew) influences subsequent SL of new linguistic input.

Building on this finding, Elazar et al. (2022) developed a more direct test of how frequent syllable combinations in the participants'

native language influence participants' subsequent SL of new linguistic material. They tested Spanish participants in either a Spanish-like condition or a Spanish-unlike condition. Participants of both conditions were exposed to a continuous speech stream containing experimental words and were afterwards tested on a lexical decision task. Importantly, the experimental words in the Spanish-like condition consisted of syllable combinations found frequently in natural Spanish while the experimental words in the Spanish-unlike condition consisted of syllable combinations rarely found in natural Spanish. Participants' performance on the lexical decision task was highly influenced by their familiarity with the syllable co-occurrences, such that participants in the Spanish-like condition were better at accepting (Spanish-like) experimental words than participants in the Spanish-unlike condition were at accepting (Spanish-unlike) experimental words. Furthermore, participants in the Spanish-like condition were worse at rejecting (Spanish-like) foils than participants in the Spanish-unlike condition were at rejecting (Spanish-unlike) foils. This finding provides more direct evidence that participants' SL is influenced by their prior knowledge of syllable co-occurrences acquired via SL in their native language.

In parallel to these studies, I examined in this thesis how prior knowledge of syllable co-occurrences (acquired via SL in the participants' native language) influences participants' subsequent language processing and learning. Crucially, I studied how the learning unfolded incrementally over the course of the exposure in adult (cf. Chapter 3) and child populations (cf. Chapter 4) and investigated how the children's language proficiency influenced their SL performance.

Does the frequency distribution of the new input matter?

After exploring how previous knowledge of syllable co-occurrences alone affects future SL (cf. Chapters 3 and 4), I expanded the research question and asked how the factor of previous knowledge interacts with another factor influencing SL, namely the frequency

distribution in which the new stimuli are presented (cf. Chapter 5). As described above, previous studies have found that the type of input distribution might affect statistical word segmentation (Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2022). Kurumada et al. (2013) concluded that learning from a Zipfian frequency distribution facilitates SL in comparison to learning from a uniform frequency distribution (though see Lavi-Rotbain & Arnon, 2022, for evidence that it is the predictability of the words rather than their precise skew which influences word segmentation). In Chapter 5, I asked whether German adults' word segmentation benefits from a combination of participants' prior knowledge of syllable co-occurrences and the stimuli being presented in a Zipfian frequency distribution.

Thesis outline

In this thesis, I investigated the availability and reliability of word segmentation cues in German child-directed speech and tested whether prior knowledge of German syllable co-occurrences influences subsequent learning of input adhering to those familiar syllable distributions.

Chapter 2 lays the foundation for my thesis by examining the availability and reliability of five word segmentation cues in a corpus analysis of German child-directed speech. For this chapter, I analysed approximately one day worth of speech input to a child (approximately 15,000 words). I obtained the words from utterances on the CHILDES database (MacWhinney, 2000) and coded for lexical stress, syllable TPs within and between words, word and syllable frequencies as well as syllable structure frequencies, word length, and single-word utterances.

Chapter 3 investigates how adults' prior knowledge of syllable co-occurrences influences their subsequent learning and processing of new input based on those familiar syllable distributions. The adults performed a serial recall task, which required them to repeat sequences of eight syllables. Three different types of sequences were presented auditorily: naturalistic sequences, non-naturalistic sequences, and

unstructured foil sequences. Naturalistic and non-naturalistic sequences contained disyllabic experimental words. Crucially, the words in the naturalistic sequences were frequently co-occurring syllable pairs in natural German (i.e., they occur with high TPs in German, without forming a German word by themselves). Words in the non-naturalistic sequences, on the other hand, comprised syllable pairs not found in a corpus of natural German (i.e., they could only be learned without prior knowledge during the experiment). I hypothesised that the adults had formed long-term memory representations of highly frequent syllable pairs in natural German, which would boost their repetition of the naturalistic sequences, with their knowledge helping them to chunk the syllables into pairs (i.e., experimental words) in the early phases of the experiment, reducing the memory load from remembering eight syllables to four words per sequence. Unstructured foil sequences did not contain any learnable patterns.

Chapter 4 turns from adults' language processing in the previous chapter to how seven- to nine-year-old children's prior knowledge of syllable co-occurrences influences their subsequent learning. The children performed the same serial recall task as the adults but with shorter sequences of six syllables. I hypothesised that children, too, would benefit from their prior knowledge of syllable co-occurrences in the naturalistic condition from the early stages of the experiment. Additionally, the children's language proficiency was assessed, and I explored whether their language proficiency predicted the children's performance on the serial recall task.

Chapter 5 studies the interaction of two factors influencing SL, namely participants' prior knowledge of syllable co-occurrences and the kind of frequency distribution in which the experimental words are presented. The chapter reports three online experiments with German adults. In the first experiment, the adults were exposed to an artificial language for three minutes, followed by a two-alternative forced choice (2AFC) segmentation task. The experiment had a 2x2 between-participants design, with participants being randomly assigned to one

of four groups: Naturalistic + Zipfian, Naturalistic + Uniform, Non-naturalistic + Zipfian, or Non-naturalistic + Uniform. The two naturalistic conditions contained the six experimental words from the naturalistic sequences in Chapters 3 and 4, and the non-naturalistic conditions contained the six experimental words from the non-naturalistic sequences in Chapters 3 and 4. These words were either presented in a Zipfian distribution in the two Zipfian conditions or in a uniform distribution in the two uniform conditions. I hypothesised that participants in the Naturalistic + Zipfian condition would outperform participants in the other conditions on the 2AFC segmentation task, given that they could benefit from both prior knowledge of the syllable co-occurrences and the Zipfian distribution of their experimental input. Likewise, I expected participants in the Non-naturalistic + Uniform condition to reach the lowest 2AFC scores since they had no facilitatory factors influencing their SL. However, unexpected results required two follow-up experiments. Potential reasons for the unexpected findings are discussed.

Finally, **Chapter 6** summarises and discusses the findings of Chapters 2 to 5 and outlines potential future research.

Chapter 2

Word segmentation cues in
German child-directed speech:
A corpus analysis

Abstract[2]

To acquire language, infants must learn to segment words from running speech. A significant body of experimental research shows that infants use multiple cues to do so; however, little research has comprehensively examined the distribution of such cues in naturalistic speech. We conducted a comprehensive corpus analysis of German child-directed speech (CDS) using data from the Child Language Data Exchange System (CHILDES) database, investigating the availability of word stress, transitional probabilities (TPs), and lexical and sublexical frequencies as potential cues for word segmentation. Seven hours of data (~15,000 words) were coded, representing around an average day of speech to infants. The analysis revealed that for 97% of words, primary stress was carried by the initial syllable, implicating stress as a reliable cue to word onset in German CDS. Word identity was also marked by TPs between syllables, which were higher within than between words, and higher for backwards than forwards transitions. Words followed a Zipfian-like frequency distribution, and over two-thirds of words (78%) were monosyllabic. Of the 50 most frequent words, 82% were function words, which accounted for 47% of word tokens in the entire corpus. Finally, 15% of all utterances comprised single words. These results give rich novel insights into the availability of segmentation cues in German CDS, and support the possibility that infants draw on multiple converging cues to segment their input. The data, which we make openly available to the research community, will help guide future experimental investigations on this topic.

Introduction

One of the first puzzles that children must solve during language acquisition is finding boundaries between individual words in speech. However, this is no easy feat, since there are no perfectly reliable cues that learners can draw upon (Aslin et al., 1996; Lehiste, 1970). Instead, children must look to a broad range of imperfect, probabilistic cues (e.g., stress patterns, phonotactic and allophonic regularities, and information about syllable co-occurrences), and use these in combination (Monaghan, 2017). Importantly, each language differs in the availability and likely combination of cues for segmentation, meaning each solution will necessarily be language-specific (see Cutler, 2012). Studying the distribution of cues to segmentation in a variety of different languages is therefore critical for shaping our understanding of whether and how they aid infants' language acquisition.

There is a substantial literature documenting the prevalence of various particular segmentation cues across different languages, most prominently in European languages such as English (see e.g., Aslin et al., 1996; Brent & Siskind, 2001; Cutler & Carter, 1987; Piantadosi, 2014) and French (e.g., Shi & Lepage, 2008; see Aslin et al., 1996, and Kabak et al., 2010, for related results from Turkish and French, and see Saksida et al., 2017, for a larger cross-linguistic comparison), though comparatively less is known about the way such cues occur in German. Moreover, there is a notable absence of comprehensive corpus studies seeking to quantify the availability of individual cues in combination in the input. In the current chapter, we present one such study of German child-directed speech (CDS). Building upon past research that has focused on single prominent cues to segmentation (e.g., *word stress:* Cutler & Carter, 1987; *transitional probabilities (TPs):* Saksida et al., 2017; *single-word utterances:* Brent & Siskind, 2001), we provide a rare comprehensive assessment of a broad range of cues that have been shown to help learners to locate word boundaries in speech, giving a

rich overview of the way these cues exist in German CDS. We address each cue that we study in turn below.

Word stress

One well-established cue to word segmentation is stress; the emphasis of a particular syllable within a word over the others. Regular stress patterns in a given language can help mark particular positions within words, and thus can provide a strong indication of word boundaries. For instance, in English, words are typically stressed on the first syllable (Cutler, 1996; Cutler & Norris, 1988), whereas in Hebrew stress usually occurs in a word-final position (Glinert, 1989) – flagging word onset and offset, respectively. Infants' use of stress as a cue for speech segmentation has been shown to be guided by the basic rhythm of the language being acquired; infants acquiring syllable-timed languages such as French, Italian, and Cantonese (i.e., languages in which syllables tend to have similar durations) start with segmentation based on the syllable, while infants acquiring stress-timed languages such as English and German (i.e., languages in which stressed syllables are longer and more emphasised than unstressed syllables) might break into the speech stream by assuming a trochaic foot (see e.g., Goyet et al., 2010; Nazzi et al., 2006).

Cutler and Norris (1988) proposed that the occurrence of a strong syllable triggers word segmentation in English, with English speakers interpreting this as the onset of a new word (Curtin et al., 2005; Echols et al., 1997; Houston et al., 2004; Jusczyk et al., 1999; Norris et al., 1995). In English, this strategy promises a high success rate, as 90% of content words begin with a strong syllable (Cutler & Carter, 1987). Jusczyk et al. (1999) reported developmental evidence in support of this claim: in a series of experiments, they showed that 7.5-month-old English infants treated strong syllables as indicators for word onset (e.g., interpreting "guiTAR is" as "gui TARis"), and only learnt to segment words following an atypical stress pattern at a later point in development (10.5 months).

Studies in a range of languages have documented infants' sensitivity to prosodic cues from a very young age (Bull et al., 1984, 1985; Eilers et al., 1984; Spring & Dale, 1977) – perhaps even from birth (Nazzi et al., 1998; Sansavini et al., 1997), with infants developing a preference for the stress pattern of their native language over the course of development (Jusczyk et al., 1993). For German, there is evidence that young infants (around five months old) can discriminate between trochaic and iambic stress, showing a preference for trochaic stress over the less common iambic stress pattern (Friederici et al., 2007; Höhle et al., 2009; Tippmann et al., 2015; Weber et al., 2004). Critically, research has shown that infants can use this information to guide word segmentation (Höhle et al., 2001). Here, we examined the precise way in which lexical stress cues are distributed across words in German CDS, providing key evidence for the widely assumed dominant trochaic stress pattern in German.

Transitional probabilities

Another likely cue to word segmentation is the TP between syllables (Saffran, Aslin, et al., 1996; Saksida et al., 2017). TPs express the likelihood that particular syllables will occur alongside each other in speech, given their prior co-occurrence in the input (both together, and with other items). Languages typically have higher TPs within than between words, such that word boundaries can be inferred at the point at which the subsequent syllable is hard to predict, given the prior syllable. For instance, in the sequence *pretty baby* the within-word syllable transitions from *pre* to *ty* and from *ba* to *by* have higher TPs (and are therefore easier to predict) than the between-word transition from *ty* to *ba* (Saffran, 2003, reports a TP of .8 for the transition from *pre* to *ty* compared to a TP of .0003 from *ty* to *ba*).

In an extensive body of research, learners of all ages have been found to be highly sensitive to the transitional information contained within speech (e.g., Saffran, Aslin, et al., 1996; Saffran et al., 1997), and from an early age, infants can use this co-occurrence information to

calculate the likely locations of word boundaries in speech (Aslin et al., 1998; Teinonen et al., 2009; see Black & Bergmann, 2017, for a meta-analytic review). This process, termed *statistical learning*, has been investigated with speakers of a variety of languages (e.g., *German:* Marimon Tarter, 2019; Matzinger et al., 2019; *English:* Saffran et al., 1996; *Finnish:* Teinonen et al., 2009; *French:* Franco et al., 2015; *Hebrew:* Siegelman et al., 2018). Moreover, TPs have been found to be informative in both directions, for both forwards transitions (i.e., a subsequent syllable being predictable based on the preceding syllable, e.g., predicting *by* from *ba* in the word *baby*) and backwards transitions (e.g., predicting *ba* from *by*; Perruchet & Desaulty, 2008). Critically though, as a cue, TPs have significant language-specific properties. Notably, languages differ on whether forwards or backwards TPs are most informative (Onnis & Thiessen, 2013). In the current study, we determined the strength of both forwards and backwards TPs as cues to word identity in German.

Lexical and sublexical frequency

Frequency has been found to play an important role in language acquisition (see Ambridge et al., 2015, for a review). In natural language, word frequency follows Zipf's law (Zipf, 1935, 1949), whereby a small number of words occur very frequently, whereas the vast majority of words are only rarely used. Zipfian distributions have been found to aid word segmentation in adult statistical learning studies, especially for larger lexica (Kurumada et al., 2013), presumably because highly frequent sequences enable rapid segmentation, which can act as anchors in subsequent utterances. This *anchor effect* has been found to benefit word segmentation in infant (Altvater-Mackensen & Mani, 2013; Bortfeld et al., 2005; Mersad & Nazzi, 2012; Shi & Lepage, 2008) and adult learners (Cunillera et al., 2010; Valian & Coulson, 1988), and in recent work, Cunillera et al. (2016) documented the neural signature of this effect – demonstrating that anchor words elicited greater stimulus-preceding negativity (a marker of expectation) in adults' electroencephalography (EEG) data

compared to less frequent words. Further support for the role of high frequency words in segmentation comes from the computational modelling literature; Monaghan and Christiansen (2010) demonstrated that their PUDDLE model of speech segmentation could quickly extract high frequency words from utterances contained within corpora of CDS, and use them to segment the remainder of the input.

Since frequency has been found to play a pivotal role in language acquisition, it follows that the benefits of highly frequent items may extend beyond word frequency, to the frequency of the syllables that words contain, and their syllabic structure. Syllable structures describe the patterns of consonants and vowels within a syllable (e.g., the syllable *ba* consists of a consonant and a vowel, abbreviated as a CV structure). These structures might follow a certain distribution, which might help segmentation in a similar way to the phonotactics of a language (see e.g., Boll-Avetisyan, 2018). That is, certain combinations of consonants and vowels might occur more often in specific positions and provide cues to word-hood. Consequently, we examined the frequency distributions of word types, word tokens, syllables, and syllable structures in German CDS.

Word length

For many of the world's languages, the length of individual words can vary quite substantially. However, Zipf's law (Zipf, 1935, 1949) states that word length is optimised for efficient communication, such that the most frequent words in a language are typically short. Support for this notion can be found for a range of languages, including English (see e.g., Li & Shirai, 2000, frequency counts for CDS corpora comprising 2.6 million words), Spanish (e.g., Alonso et al., 2011), and Swedish (Sigurd et al., 2004). In German, prior analyses revealed that approximately 50% of (written) words were monosyllabic, whereas around 30% were disyllabic, and approximately just 20% were longer still (Kaeding, 1897; Sigurd et al., 2004). While heterogeneity among word lengths is commonplace within language, a number of studies

have demonstrated that having a variety of word lengths in speech poses a significant challenge to speech segmentation (Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012; Kurumada et al., 2013; but see Perruchet & Vinter, 1998, for computational counter-evidence) – though this difficulty may be eased when speech contains additional cues (Frost et al., 2020; Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012). Conceivably, caregivers may remove some of the complexity associated with varying word length by providing a more uniform signal in CDS (Garmann et al., 2019; but see Segal et al., 2009). We investigated this possibility here.

Single-word utterances

Finally, another potential cue for identifying word boundaries is the occurrence of words in isolation, in single-word utterances. Research has found that most caregivers use single-word utterances in conversations with their infants, repeating around a third of these within close temporal proximity (Aslin et al., 1996; Brent & Siskind, 2001). Previous studies have estimated that up to 26% of utterances in English CDS comprise single words (Monaghan & Christiansen, 2010; but see MacWhinney & Snow, 1985, for a more conservative estimation of 14%). These single-word utterances have been suggested to help segmentation by first facilitating learning of these items (Junge et al., 2012), then flagging the boundaries of neighbouring items in subsequent multi-word utterances (Peters, 1983; Pinker, 1984) – similar to the way in which high frequency words have been proposed to assist segmentation. In the present study, we examined how many single-word utterances occurred in German CDS and how many single-word utterances were repeatedly produced, supposedly boosting the facilitated segmentation effect.

Aims and hypotheses

Past research has revealed that infants are sensitive to a range of cues to speech segmentation, and that the prevalence of these cues within the speech that children hear is subject to marked cross-

linguistic variation. However, much remains to be done to determine the relative weighting of these cues across the world's languages. In the current study, we adopt a corpus-based approach to determine cue availability in German CDS. Using High German as our target language, we took the equivalent of one day's worth of input to a German-acquiring infant, and coded it for primary word stress, TPs, word frequency, word length, and the occurrence of words in single-word utterances. We hypothesised that we would find a dominant trochaic stress pattern for German similar to the one found in English (Cutler & Carter, 1987). In addition, we expected to see higher within-word than between-word TPs, and higher backwards than forwards TPs, similar to the results found in English (another right-branching language; Onnis & Thiessen, 2013; Saksida et al., 2017). With regard to word frequency, we expected to find a Zipfian-like distribution of word types, word tokens, and syllables (Zipf, 1935, 1949), as has been found for a variety of the world's languages, with a small number of words occurring with comparatively higher frequency than the remainder of words in the corpus. In terms of word length, we expected to find a greater proportion of shorter than longer words (Piantadosi et al., 2011; Zipf, 1935, 1949). Based on corpus analyses of English, we hypothesised that the corpus may contain a large proportion of single-word utterances (MacWhinney & Snow, 1985; Monaghan & Christiansen, 2010), with a large amount of these occurring repeatedly (Brent & Siskind, 2001; Monaghan & Christiansen, 2010).

Method

Data

Our data are openly available on the Open Science Framework (OSF): https://osf.io/vpdu6/. Our corpus comprised 20 German datasets from the Child Language Data Exchange System (CHILDES) database (MacWhinney, 2000). All datasets contained CDS spoken to children under two years of age. In order for our corpus to contain a representative sample of speech, we included files from a number of

different children, recorded in different contexts (e.g., playing with toys, reading books, eating, or bathing). This reduced the likelihood that speaker-specific patterns would influence our results. In total, we included data from 19 individual speakers talking to ten different children, taken from the Caroline (Von Stutterheim, 2010), Manuela (Wagner, 2006), Miller (Miller, 1979), Rigol (Rigol, 2007), and Wagner (Wagner, 1974, 1985) corpora, with the age of the children at the time of recording ranging from 00;06.13 to 01;08.13 years. Together this totalled 07:32 hours of recording, during which caregivers (and occasionally siblings or researchers) provided 3967 utterances of CDS input, comprising an overall total of 16,474 words, and 14,660 words after filtering out proper names, sounds, and unintelligible speech (see the Appendix for further information on the included datasets). We estimate that this represents approximately one day's worth of input[3].

Coding

We coded the data by word tokens, that is, individual occurrences of words in CDS (so, with one entry for each of the 16,474 words). We defined a word as a unit that the child needs to segment to assign meaning[4]. For each word, we coded for information at the word and syllable levels (for the full coding scheme see our OSF page). At the word level, we coded for word type (i.e., grouping different pronunciations of words, which result in different word tokens, into one word type), parts of speech (i.e., whether a word is a noun or a verb, etc.) and resulting categorisation as content or function words, with content words comprising nouns, verbs and adjectives, and function words comprising all remaining word categories. We also coded for word length (number of syllables), and word stress (the position of

---

[3] A recent study by Donnelly and Kidd (2021) using daylong recordings found that, at twelve months, the average number of words a child hears is 14,572 (SD = 6826), based on a large sample of over 100 children acquiring Australian English as a first language. This number increases slightly across the next year, to 16,827 words at 24 months. Note that this estimates the words in the environment, only a subset of which is likely to be child-directed speech.

[4] This is different from a phonological word as those comprise chunks such as *haste* for *has(t) de [: du]* ('have you'), which we treated as two different words.

stress within words). At the syllable level, we coded the phonetic representation and syllable structure for each syllable of the word (i.e., describing the pattern of consonants and vowels which the syllable comprised, e.g., CV for the syllable [ba]). Sounds and unintelligible material were excluded from the analyses. Proper names were excluded from all analyses, except for the analysis which sought to establish the occurrence of proper names in single-word utterances.

Results

We will first outline the results for our analyses of word stress, TPs, and word and syllable frequencies. We will then present our findings for word length, and finally for those cues that can facilitate segmentation by flagging word boundaries (i.e., highly frequent words and single-word utterances). Our analyses and results are openly available on OSF: https://osf.io/vpdu6/. Additional analyses (such as analyses on subsets of the data, for instance, excluding monosyllabic words) can be found in the "Additional Analyses" section in our analysis file. All analyses were performed in R 3.6.3 (R Core Team, 2022).

Word stress

We examined the position of primary within-word stress, to establish how reliable the widely assumed dominant trochaic stress pattern is as a potential cue for segmentation in German. This analysis was performed on the whole corpus, excluding proper names and sounds.

The vast majority of words in our corpus of CDS were found to carry word-initial stress; in total, approximately 97% of words were stressed on the first syllable, whereas around 3% were stressed on the second, and less than 1% on the third to seventh syllables, but with no words being stressed on the sixth syllable (see Table 2.1). In addition to this primary analysis, which used the entire corpus (thereby providing the closest approximation to the full input), we ran two further iterations – the first of which excluded repetitions (examining unique

Table 2.1. *Frequency of primary word stress at each syllable position.*

Syllable position	Primary word stress: All word tokens		Primary word stress: Unique word tokens		Primary word stress: Word tokens excluding monosyllabic words	
	Count	%	Count	%	Count	%
1	14,206	96.90	1,536	87.03	2,771	85.92
2	398	2.71	191	10.82	398	12.34
3	47	0.32	31	1.76	47	1.46
4	6	0.04	5	0.28	6	0.19
5	1	0.01	1	0.06	1	0.03
6	0	–	0	–	0	–
7	2	0.01	1	0.06	2	0.06
8	0	–	0	–	0	–
Total	14,660		1,765		3,225	

word tokens only), and the second of which was run on word tokens but excluded monosyllabic words (which can only be stressed on their first and only syllable). This was vital for establishing whether the observed stress pattern is generalisable, and is not reliant on particular tokens.

For both of these iterations, we analysed the resulting corpus in the same way as before. Both analyses yielded the same pattern of results: excluding repetitions, 87% of words were stressed on the first syllable, 11% on the second, and 2% on the third to seventh syllables (with no words carrying stress on the sixth syllable). Excluding monosyllabic words, 86% of words were stressed on the first syllable, 12% on the second, and 2% on the third to seventh syllables (again with no words carrying sixth-syllable stress). Thus, these data provide strong evidence to suggest that German CDS has a dominant trochaic stress pattern (i.e., word-initial stress).

Transitional probabilities

We next examined the way in which TPs between syllables varied according to two key aspects: context (i.e., for transitions within versus between words); and direction (i.e., probabilities of syllable co-occurrence for *forwards* versus *backwards* transitions). To do this, we

extracted pairs of syllables from either within or between words within utterances only (i.e., not crossing utterance boundaries, which are typically indicated by a pause or a switch in speakers), and calculated forwards and backwards TPs for both contexts. Forwards TPs were calculated following Equation (1a), and backwards TPs following Equation (1b). That is, the forwards TPs within the word *baby*, for instance, were calculated by dividing the number of times the two syllables *ba* and *by* co-occurred by the total number of times the syllable *ba* occurred:

(1a)   probability of B, given A = $\dfrac{\text{\# occurrences A + B}}{\text{\# occurrences A}}$

(1b)   probability of A, given B = $\dfrac{\text{\# occurrences A + B}}{\text{\# occurrences B}}$

Figure 2.1 shows that both backwards and forwards TPs are higher within than between words. To test whether the TPs varied according to context and direction, we fitted a linear mixed-effects model using the *lme4* 1.1-23 package (Bates et al., 2015). The dependent variable was TP, and context and direction were entered as fixed effects. We used deviation contrasts for context (*within words: –0.5, between words: 0.5*) and direction (*forwards: –0.5, backwards: 0.5*). We fitted the maximal model supported by the data (Barr et al., 2013), controlling for the syllable pair as a random intercept with direction[5] as a random slope. To examine the effects of the model predictors, we used likelihood-ratio (χ^2) comparisons to obtain *p*-values (through serial decomposition), and bootstrap simulations (Runs = 1000) to calculate 95% confidence intervals for the beta estimates. The marginal and conditional R^2 effect sizes are also reported as goodness-of-fit estimates. These denote the proportion of the variance explained by the model both with (conditional R^2) and without (marginal R^2) controls for sources of random variance (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013).

---

[5] Direction is included as a random slope because forwards and backwards TPs are calculated for each syllable pair. Context, however, is not included as a random slope because within-word and between-word TPs are almost exclusively calculated on different syllable pairs.
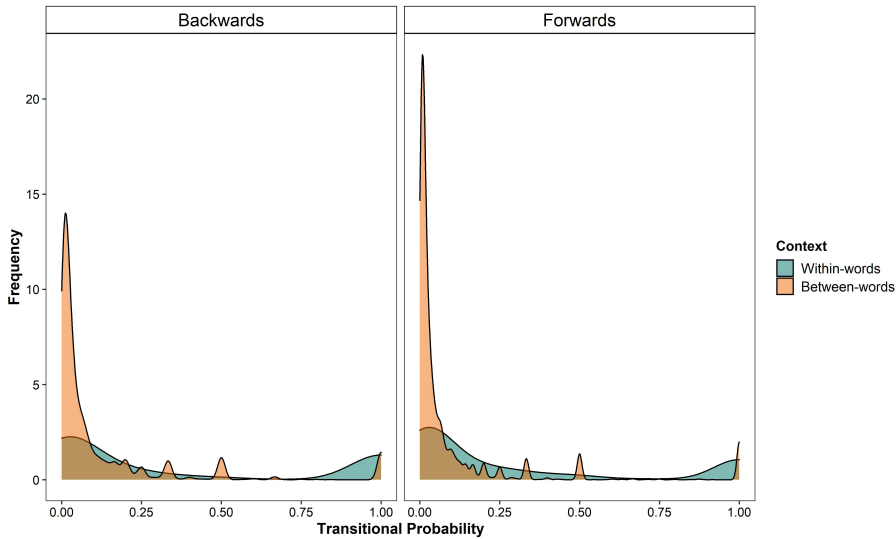
Figure 2.1. *Density plot of transitional probabilities (TPs) between syllables in the corpus. The panels on the left and right show the frequency data for backwards and forwards transitions, respectively. TPs within words are indicated in green, whereas TPs between words are indicated in orange.*

There was a significant main effect of context, with TPs being higher within words than between words (*within words: M* = 0.33, *SD* = 0.41; *between words: M* = 0.11, *SD* = 0.21). There was also a significant effect of direction, with TPs being higher for backwards transitions than for forwards transitions (*backwards: M* = 0.17, *SD* = 0.29; *forwards: M* = 0.13, *SD* = 0.26; see Figure 2.1 and Table 2.2). There was a significant interaction between context and direction, driven by a larger difference between the two contexts for the backwards TPs (*forwards: within words: M* = 0.30, *SD* = 0.39; *between words: M* = 0.10, *SD* = 0.21; *backwards: within words: M* = 0.36, *SD* = 0.44; *between words: M* = 0.12, *SD* = 0.22). The maximal model with context and direction as fixed predictors accounted for approximately 10% of the variance in the data without the random effects structure, and 46% of the variance with the random effects structure.

Table 2.2. *Summary of the linear mixed-effects model for transitional probabilities (TPs).*

Parameter	*b*	95% CI	*SE*	χ^2	*df*	*p*
(Intercept)	0.220	[0.215, 0.225]	0.003	–	–	–
Context	–0.219	[–0.229, –0.209]	0.005	1556.53	1	< .001
Direction	0.047	[0.035, 0.059]	0.006	52.46	1	< .001
Context × Direction	–0.041	[–0.063, –0.018]	0.012	11.24	1	< .001

Notes: Model fit: AICc = 2151; BIC = 2211; $R^2_{marginal}$ = 0.098; $R^2_{conditional}$ = 0.464.

Frequency

Word frequency

We examined the frequency distribution of words in the input, in the light of the suggestion that highly frequent words and a Zipfian-like frequency distribution (Zipf, 1935, 1949) can support segmentation (Kurumada et al., 2013).

There was a Zipfian-like frequency distribution (Zipf, 1935, 1949) for both word tokens and word types (see Figure 2.2 for a density plot of word token frequencies; see our OSF repository for the same plot but with word types), with the corpus containing a large amount of low frequency words (i.e., open class words such as nouns, which were high in quantity, but were rarely repeated, amounting to 31% of words in the corpus), and a small amount of words with much higher frequencies (i.e., closed class words such as determiners, which were used in combination with all nouns, e.g., *ein Fuchs* "a fox", *ein Häschen* "a rabbit", and *ein Laster* "a lorry", amounting to 69% of words in the corpus). This was in line with our hypothesis. A summary of the word frequency density (i.e., the percentage of individual words in the corpus occurring once, twice, three times, etc.) is provided in Table 2.3, alongside the analogous results from Kaeding's (1897) study of written German. Both studies revealed a Zipfian-like frequency distribution (i.e., in both sets of input, approximately half of the words occurred just
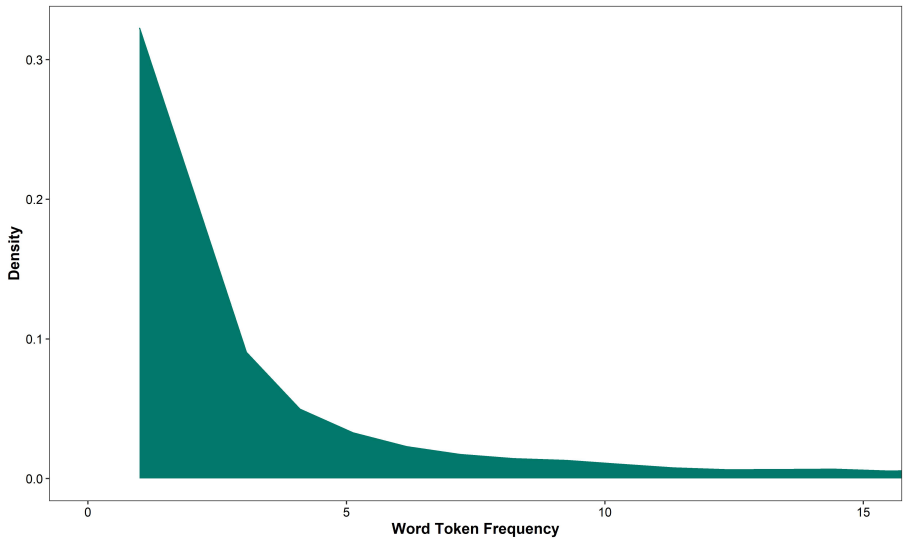
Figure 2.2. *Density plot of word token frequencies, indicating the extent to which words occur with particular frequencies in the corpus.*

Table 2.3. *Summary of word token frequencies in Kaeding's (1897) study of written German and in the present corpus of German child-directed speech.*

Word token frequency	Kaeding (1897)	Current dataset
1	49.14%	49.76%
2	13.37%	15.12%
3	6.61%	7.65%
4	4.31%	4.34%
5	3.04%	3.37%
6–10	7.76%	7.47%

once; 49% of words in Kaeding's study, 50% of words in the current corpus; and approximately 15% of words occurred twice, etc.).

We focused our subsequent frequency analyses on the 50 most frequent items (computing analyses for both word tokens and word types), to shed light on the properties of the words that infants were hearing the most. For word tokens, the 50 most frequent items constituted 54% of the corpus (7896 out of 14,660 words; see Figure
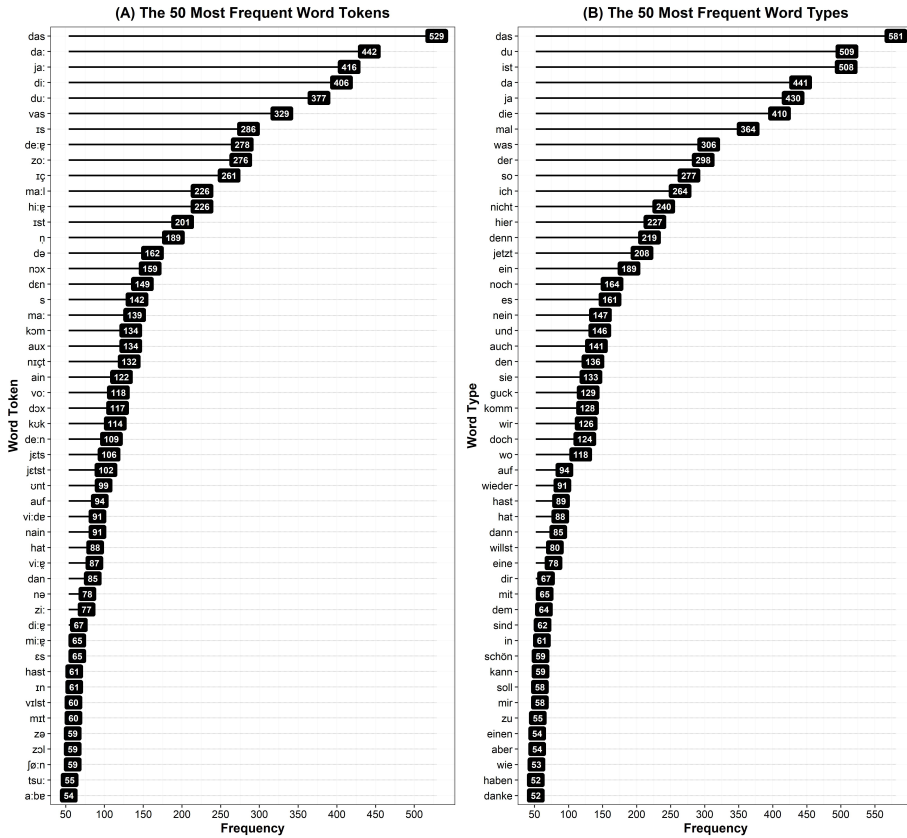
Figure 2.3. *Frequencies for the 50 most frequent words in the corpus. Panel A (left) shows word tokens, and Panel B (right) shows word types.*

2.3 Panel A), and were almost exclusively monosyllabic, with *wieder* ("again", 91 occurrences) and *aber* ("but", 54 occurrences) being the only multi-syllabic exceptions. For word types, the 50 most frequent items constituted 59% of the corpus (8602 out of 14,660 words; see Figure 2.3 Panel B). As with word tokens, the vast majority of word types were monosyllabic, with six exceptions; *wieder* ("again", 91 occurrences), *eine* ("a", 78 occurrences), *aber* ("but", 54 occurrences), *einen* ("a", 54 occurrences), *danke* ("thanks", 52 occurrences) and *haben* ("have", 51 occurrences), which were all disyllabic. Thus, these data suggest that the vast majority of the most frequent words in

45

German CDS are monosyllabic, with a small number of disyllabic exceptions.

To investigate which *kind* of words were most frequent, we distinguished between function and content words. Of the 50 most frequent words, 41 tokens (82%) or 39 types (78%) were function words (e.g., *das* "the" or *ja* "yes"), whereas just nine tokens (18%) or eleven types (22%) were content words (e.g., *guck* "look" or *schön* "nice"). These highly frequent function words accounted for approximately 47% of word tokens in the entire corpus (6834/14,660), and 50% of word types (7290/14,660). The vast majority were monosyllabic (39/41 tokens, and 35/39 types). These highly frequent monosyllabic function words accounted for approximately 46% of word tokens in the entire corpus (6689/14,660), and 48% of word types (7039/14,660).

Syllable and syllable structure frequency

We examined the frequencies of individual syllables, and particular syllable structures. For instance, the word *Baby* consists of two syllables, [be:] and [bi] with the respective syllable structures CVV and CV. Our corpus comprised 18,736 syllable tokens in total, and particular syllables were seen to occur with a Zipfian-like distribution (Zipf, 1935, 1949; see OSF for a density plot of syllable frequencies). Because of the large quantity of monosyllabic words within the corpus, the most frequent syllables were identical to the most frequent word tokens (see Figure 2.4 Panel A). Of particular interest, then, are syllables occurring in multisyllabic words. The results of our additional analyses excluding monosyllabic words, as well as focusing particularly on disyllabic and trisyllabic words (as used in most artificial language learning studies) can be found on OSF in Section 5 of the analysis file. Since children, however, encounter the monosyllabic words in their input, we draw our conclusions from the complete dataset, reporting only the results for the whole corpus (including all word lengths) here. We summarise our findings for

(A) The 50 Most Frequent Syllables

(B) The 45 Syllable Structures
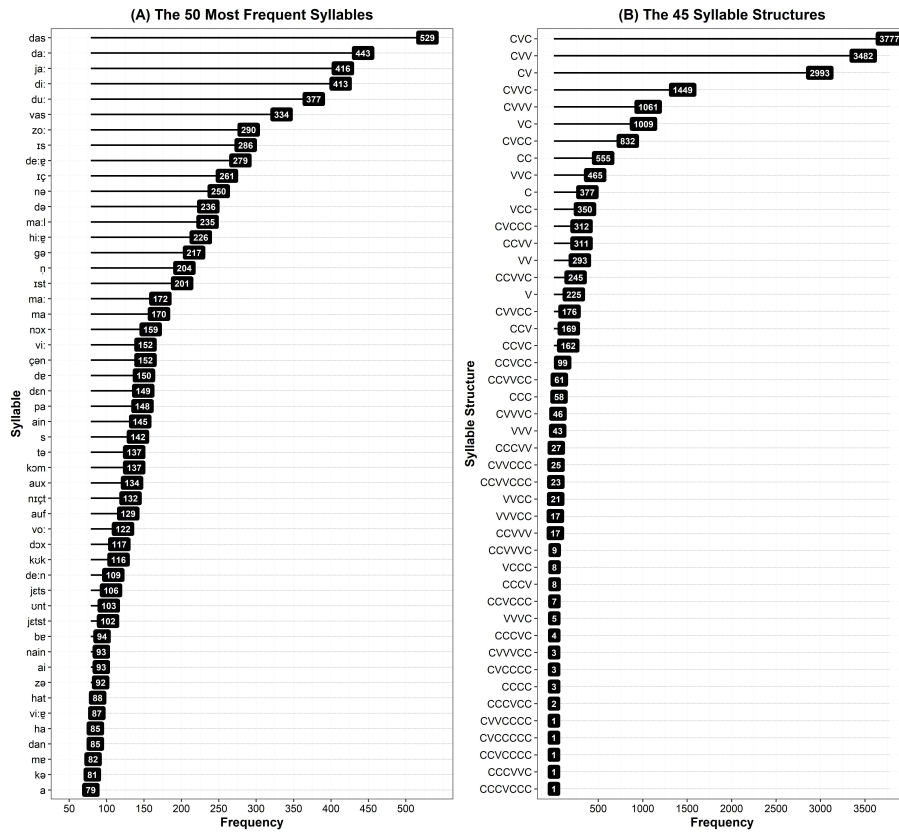
Figure 2.4. *Syllable and syllable structure frequencies in the corpus. Panel A (left) shows the 50 most frequent syllables, and Panel B (right) shows all 45 different syllable structures. Because we consider syllabic consonants such as [n̩] as consonants, it is possible to have syllables with multiple consonants but no vowels (e.g., the second syllable of the verb* putzen *[pʊtsn̩] "clean" consists of three consonants); similarly, because we code long vowels as VV, it is possible to have syllables with multiple vowels (e.g., the word* er *[eːɐ̯] "he" consists of three vowels).*

multisyllabic words, as well as disyllabic and trisyllabic words in the Supplementary material folder on OSF.

For syllable structure, there was again a Zipfian-like distribution (Zipf, 1935, 1949; see OSF for a density plot of syllable structure

frequencies), with a small number of structures occurring much more frequently than others. We examined the frequencies with which particular syllable structures occurred at different positions within words, to explore the possibility that patterns of regularity may indicate word boundaries (for instance, if certain structures are mostly found at word edges).

There were 45 different syllable structures within our corpus (see Figure 2.4 Panel B). In initial and final positions, there were 42 different structures; in medial positions, there were 24. We observed slight differences dependent on syllable position; the most common structure in medial positions was an open syllable (CV as in [gə]; comprising 40% of medial syllables), whereas the most common structures in initial and final positions were closed (CVV as in [da:] or CVC as in [das]). Word-initial syllables ended most often in a long vowel (i.e., CVV; comprising 22% of initial syllables), whereas syllables in word-final positions ended most often in a consonant (i.e., CVC; comprising 23% of final syllables). However, these three structural types (CV, CVV, and CVC) were found to occur in all positions within words with a high degree of frequency, constituting the most frequent syllable structures for all three locations – limiting the extent to which these structures may serve to cue segmentation. The difference between structure occurrence in initial versus final positions is particularly subtle (*initial:* CVV 22%, CVC 21%; *final:* CVV 18%, CVC 23%), possibly because of the large amount of monosyllabic words in the corpus. Again, syllable structures occurring in multisyllabic words can provide further insights. We summarise our findings for multisyllabic words, as well as disyllabic and trisyllabic words in the Supplementary material folder on OSF.

Word length

Next, we examined word length, and the frequency with which different word lengths occurred in the input. Table 2.4 lists this for the number of word tokens, the number of unique word tokens, and the

Table 2.4. *Frequency statistics for word length (measured in number of syllables).*

Number of syllables	Number of word tokens		Number of unique word tokens		Number of unique word types	
	Count	%	Count	%	Count	%
1	11,435	78.00	655	37.09	553	34.33
2	2,550	17.39	715	40.49	672	41.71
3	545	3.72	293	16.59	285	17.69
4	98	0.67	78	4.42	76	4.72
5	21	0.14	17	0.96	17	1.06
6	9	0.06	7	0.40	7	0.43
7	0	–	0	–	0	–
8	2	0.01	1	0.06	1	0.06
Total	14,660		1,766		1,611	

number of unique word types. Word tokens provided the raw frequency counts of every word in the corpus. Unique word tokens represent the number of different words in the corpus regardless of the number of repetitions of this item (e.g., a list containing: "one, one, two" would count three word tokens but only two unique word tokens). The unique word types column combines different pronunciations of the same word (e.g., a list containing: "not, not, n't," would count three word tokens, two unique word tokens but only one unique word type). We computed the word length of unique word tokens and unique word types as a measure of robustness to ensure the reliability of the findings and to control for potential correlations with other effects such as word frequency.

The words in our corpus were between one and eight syllables long (with the longest words being nominal compounds). In total, 11,435 (78%) of all words were monosyllabic, 2550 (17%) disyllabic, 545 (4%) trisyllabic, and 130 (1%) between four and eight syllables long (with seven-syllable words never occurring). After controlling for frequency (via excluding repetitions) there was a slight shift in this pattern, with disyllabic words occurring slightly more often (40%) than monosyllabic words (37%), followed by trisyllabic words (17%), and words with four to eight syllables (6%). This pattern shift indicates that

shorter (monosyllabic) words were subject to a greater degree of repetition in the corpus. Interestingly, although German allows significant compounding, only 2% of word tokens in our corpus were compounds.

Single-word utterances

Finally, we examined the corpus for single-word utterances, which may aid segmentation by subsequently flagging the boundaries of adjacent words in multi-word utterances. Of the 3513 utterances (excluding proper names and sounds), 527 utterances (or 15%) comprised single words (898 of 3967 utterances, or 23%, including proper names and sounds). Although we excluded proper names and sounds from all of our prior analyses, proper names – particularly the child's – have been found to be highly salient anchors for infants' segmentation of multi-word utterances (Bortfeld et al., 2005). Thus, we examined how often proper names occurred in single-word utterances. Across the whole corpus, single-word utterances comprising proper names occurred just 42 times – amounting to 7% of single-word utterances, and 1% of all utterances (including proper names, but excluding sounds).

The remainder of the single-word utterances were found to largely comprise function words (71% of single-word-utterances, excluding proper names and sounds). The most frequent words were particles such as *ja* ("yes"), which amounted to 17% of single-word utterances (3% of all utterances in the corpus), *nein* ("no", 6% of single-word utterances), *danke* ("thanks", 4% of single-word utterances), and *bitte* ("please", 3% of single-word utterances), adverbs such as *so* ("like this", 10% of single-word utterances), and *da* ("there", 9% of single-word utterances), the pronoun *was* ("what", 5% of single-word utterances), and the interjection *hallo* ("hello", 2% of single-word utterances). 29% of single-word utterances were content words such as the imperative *komm* ("come", 3% of single-word utterances), and the noun *Baby* ("baby", 3% of single-word utterances). A list of all single-

word utterances, including the remaining ones which occurred less than 10 times, can be found in the Supplementary material folder on OSF.

Discussion

This study offers the first corpus analysis investigating the availability of word segmentation cues in German CDS, and the first to combine an analysis of a broad range of possible cues. We analysed approximately one day's worth of input data from the CHILDES database (MacWhinney, 2000), examining a variety of potential word segmentation cues in German CDS: word stress, TPs, word and syllable frequencies, syllable structures, word length, and single-word utterances. We discuss the results for each of the cues in turn.

Word stress

Analyses of the corpus revealed a dominant and reliable trochaic stress pattern, with almost all words (97%) being stressed on the first syllable – providing strong evidence for the widely assumed trochaic stress pattern in German (Friederici et al., 2007; Höhle et al., 2001, 2009; Tippmann et al., 2015; Weber et al., 2004). Crucially, the trochaic stress pattern persisted even when monosyllabic words were withheld from the analysis – indicating that infants may be able to use stress to inform segmentation of words of various lengths. These findings are comparable to findings on English word stress, where 90% of content words contained word-initial stress (Cutler & Carter, 1987; compared to 93% in the current study).

In the German linguistics literature there is still somewhat of a controversy about the rules underlying the predominant stress pattern in German, with some researchers claiming a universal rule assigning stress from the right word-edge (e.g., Giegerich, 1985; Vennemann, 1990; Wiese, 1996), and others claiming a different rule for stress assignment in words of Germanic origin (word-initial stress) versus more recent borrowings (right-edge stress) (Benware, 1980; Braches, 1987; Féry, 1986; Wurzel, 1970, 1980; see Goedemans & van der

Hulst, 2013a, 2013b, for a classification; and Jessen, 1999, for a discussion). We note, though, that since 95% of the words in our corpus were monosyllabic or disyllabic, establishing whether primary stress occurred on the first versus the penultimate syllable would not be possible. That is, for infants segmenting the speech detailed in our corpus, both of these possible stress patterns would be interpreted as containing word-initial stress.

Nevertheless, given the data reported here, we can assume that German CDS largely adheres to a word-initial stress pattern, which children can draw upon with high return given its ubiquity in the input (potentially with a small number of exceptions due to affixation – 3% in our corpus). This is consistent with experimental work on infant segmentation, which has shown that children make use of stress cues early in development (e.g., Höhle et al., 2001; Houston et al., 2000).

Transitional probabilities

The analysis of the TPs provided support for another cue to word segmentation in German CDS, with TPs being significantly higher within than between words. This finding builds on prior demonstrations that TPs are informative cues to word-hood in a variety of languages (e.g., Saffran, 2003; Saksida et al., 2017) – extending this to German. Together with the many experimental demonstrations of TP-based segmentation in experiments (Saffran, Aslin, et al., 1996; see Black & Bergmann, 2017, for a review), the naturalistic data lend credence to the possibility that infants draw on these statistics during language acquisition.

There are two additional features of the TP results that deserve discussion. The first concerns the magnitude of within-word TPs, which appear to be rather small compared to experimental studies, where words are typically defined by TPs that are much higher (indeed, in psycholinguistic experiments these are often perfect, i.e., TP = 1.0). Thus, if children draw on TPs to aid segmentation "in the wild," they are doing so in a much noisier channel. Nevertheless, there is good

reason to believe that they do. For instance, Pelucchi et al. (2009) showed that eight-months-old infants can segment words from a foreign natural language (English-acquiring infants segmenting Italian speech) under experimental conditions following a short exposure phase.

The quasi-regular nature of this cue is a necessary outcome of the generative nature of language, and might actually be a key to learning. Kidd et al. (2012) argue that infants demonstrate a *Goldilocks effect*, such that they prefer to attend to events that are neither highly predictable nor unpredictable, thus avoiding making generalisations that are either too simple or too complex. A recent computational model of word learning suggests that cue variability may indeed serve to help, rather than hinder, learning – guiding the creation of a robust, canalised language system that is resistant to noise in the input (Monaghan, 2017). This possible utility of noise in learning is underpinned by the principle that variation in the availability and reliability of distributional cues may encourage learners to seek guidance from multiple possible information sources, reducing the importance of a particular individual cue, and increasing the resilience of the language system to noise. Variability within various distributional statistics has been found to have advantages for segmentation (Kurumada et al., 2013), word learning (Hendrickson & Perfors, 2019; Monaghan et al., 2017), semantic category learning (Lany, 2014), and acquisition of syntactic structure (Gómez, 2002). Here, we raise the possibility that this may also extend to variability among TPs.

Interestingly, we found backwards TPs to be significantly higher (and thus, more informative) than forwards TPs, and the magnitude of the difference between within-word and between-word TPs was larger in the backwards direction. This finding provides further support for the notion that TPs are informative in both directions, as has been observed in English (Perruchet & Desaulty, 2008). Further, these data lend critical support to the idea that backwards TPs are more informative than forwards TPs in right-branching languages such as English (Onnis

& Thiessen, 2013), extending this to German. This is in contrast to the distributional patterns observed for left-branching languages such as Korean, where forwards TPs are held to be more informative (Onnis & Thiessen, 2013). The generalisation here is that there are distinct influences of typology on probabilistic distributions in language; in particular, head-direction creates conditions in which one prominent element grounds a dependent one (e.g., compare *red wine* in English to *vino rosso* in Italian). Ultimately, this means that any statistical learning mechanism useful for segmentation must rapidly attune to the target language (Onnis & Thiessen, 2013; Thiessen et al., 2019).

Taken together, the results concerning stress and TPs suggest that, in German, stress is a dominant segmentation cue; consistent with recent experimental work by Marimon Tarter (2019), who found that German-acquiring six-month-old infants and German-speaking adults preferentially attended to stress over statistical information during a segmentation task. This is the opposite pattern than has been observed for English (Thiessen & Saffran, 2003), suggesting an unexpected cross-linguistic difference in the two, highly-related, languages. Further research will be necessary to unpack these differences.

Lexical and sublexical frequency

With regard to frequency, we found Zipfian-like distributions (Zipf, 1935, 1949) for every feature that we analysed; words, syllables, and syllable structures, replicating Kaeding's (1897) work on written German. These findings provide further evidence for the well-established ubiquity of Zipfian distributions in natural language. In recent work, such distributions have been suggested to help speech segmentation (Kurumada et al., 2013). In terms of word frequency, highly frequent items have been proposed to aid segmentation by acting as anchor points for subsequent segmentation to occur around; these words are believed to undergo early extraction from the speech stream, before flagging the boundaries of the words they appear alongside in subsequent speech (Altvater-Mackensen & Mani, 2013; Bortfeld et al.,

2005; Kurumada et al., 2013; Mersad & Nazzi, 2012; Monaghan & Christiansen, 2010; Shi & Lepage, 2008). The precise utility of Zipfian distributions among syllables and syllable structures remains to be established; however, it is conceivable that these may serve segmentation in a similar way. This possibility requires empirical investigation.

Word length

Regarding word length, we found the majority of words to be monosyllabic, with only 22% of words having more than one syllable, and only 5% of words having more than two syllables. The amount of monosyllabic words reported here was considerably higher than that described in previous reports on German (78% here, versus 50% in Kaeding, 1897 – which Zipf's calculations were based upon). This discrepancy may be traced back to the contrast between spoken and written language, with spoken language shortening words by the use of contractions; or the difference may be due to the contrast between child-directed and adult-directed speech, with CDS potentially being defined not only by the use of a higher pitch and shorter utterances (e.g., Cristia, 2013), but also by the use of shorter words in general (Garmann et al., 2019; but see Segal et al., 2009). This resulted in a much larger proportion of monosyllabic words here than in Kaeding's (1897) frequency dictionary (though a comparison of our data with data for more recent adult-directed speech, collected in a similar manner, would be necessary to draw more firm conclusions). In any case, data from our corpus indicate that caregivers may optimise word length (via simplification) for efficient communication to a greater extent in child- compared to adult-directed speech (see Garmann et al., 2019), with even more monosyllabic words than would be predicted by Zipf's law (Zipf, 1935, 1949). This in turn offers an interesting new perspective on the finding that a variety of word lengths adds difficulty to segmentation (Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012). It appears that, if word length is a problem for segmentation, for German infants, this may well be circumvented by a fairly uniform input consisting of

mostly monosyllabic words (see Perruchet & Vinter, 1998, for computational evidence in support of this proposal).

Our analyses of word length also provide another instance where cues appear to converge. We found the vast majority of the 50 most frequent words, and almost two-thirds of the whole corpus to be monosyllabic function words. Importantly, those words are more stressed in German than in English, and therefore perfectly detectable by the infant (Höhle & Weissenborn, 2003; but see also Gerken, 1994; Gerken & MacInthosh, 1993; Shafer et al., 1992, for evidence of the detection of function words in English). In consequence, infants can detect and segment those highly frequent function words – and subsequently use them as anchors to facilitate acquisition of the words surrounding them (Bortfeld et al., 2005; Mersad & Nazzi, 2012).

Single-word utterances

Finally, 15% of the utterances in our corpus were single words, 85% of which were words that were repeated in isolation at least once, and 62% occurred in isolation between ten and 90 times. The amount of single-word utterances in German CDS is similar to that observed for English CDS, although it falls towards the lower boundary of the estimations made in prior research (estimated at around 14% by MacWhinney & Snow, 1985; and 26% by Monaghan & Christiansen, 2010). Nevertheless, this yields a fairly substantial amount of isolated words, which can potentially be segmented more easily, and in turn subsequently aid segmentation of adjacent words in multi-word utterances (Peters, 1983). Previous research found that approximately 33% of single-word utterances were repeated in close temporal proximity (Brent & Siskind, 2001). Even though we did not examine temporal proximity here, we can add that 85% of single-word utterances were indeed repeated within the corpus.

In addition, we found that 1% of all utterances comprised proper names presented in isolation. This was mostly the child's own name (74%), but also included names of siblings (17%), and others (9%). The

number of occurrences concerning children's names here is comparable to prior observations in English CDS (1%, Monaghan & Christiansen, 2010), and accounts for 20% of all the times a child's name occurred in the speech (compared to 24% of instances in English; Monaghan & Christiansen, 2010). These isolated occurrences of names may help increase their prominence to young learners, with names being suggested to enjoy a privileged position as salient anchor words that lend significant benefits to segmentation (Bortfeld et al., 2005; Mersad & Nazzi, 2012), operating in a similar way to high frequency words. Thus, these findings indicate that single-word utterances (Brent & Siskind, 2001; Monaghan & Christiansen, 2010), and particularly isolated incidences of children's names, may serve segmentation to a similar degree in German CDS as has been previously suggested for other languages.

Limitations and future directions

Despite addressing a broad variety of segmentation cues, there are a number of potential cues not addressed here that may be valuable during language acquisition. For instance, we did not assess phonotactics or allophonic variation. Future assessments may wish to include these features to provide an extensive overview of the potential segmentation cues in German CDS. Additionally, while our results paint a strong picture of the prevalence of several individual cues in German CDS, indicating their potential importance for speech segmentation, determining how these cues interact requires further exploration. Moreover, establishing the way in which learners draw on these cues together during learning requires much empirical investigation. One way to address this topic is to combine cross-linguistic research, including corpus studies as well as experimental studies, with computational modelling approaches (cf. Monaghan & Rowland, 2017).

We note too that the syllable serves as the segmentation unit for many of these cues, which raises the question of how infants come to

identify the precise boundaries of a given syllable (which would be necessary in order for it to inform subsequent learning). This capability is likely the outcome of several distinct, and perhaps converging, sources of information – such as the phonotactics of a language, in addition to its prosody, as well as broader distributional properties (e.g., permissible syllable structures and TPs). Since the majority of words in our corpus of German CDS were monosyllabic and stressed word-initially, it is difficult to speculate on the relative contributions of other cues for this task, but this would be an insightful avenue for future research.

We also acknowledge that our results are based on what may be considered a relatively small amount of data, particularly given the recent surge in studies using day-long recordings (e.g., Casillas et al., 2019; Donnelly & Kidd, 2021; Weisleder & Fernald, 2013). However, there is evidence to suggest that corpus size does not lead to significant changes in distributional statistics (see Gambell & Yang, 2006; and see Saksida et al., 2017, for TP analyses on nine different languages using similar-sized corpora). Thus, it is unlikely that the results we observed would vary significantly with a larger corpus. We note, too, that the kinds of in-depth, fine-grained analyses we conducted are atypical of studies using day-long recordings, which are based on fairly course estimates of language, computed via automated algorithms or through transcription of small subsets of the data. Rather, our focus on the minutiae of lexical and sublexical distributional information required a good degree of hand-coding.

Finally, it is important to acknowledge that we are generalising over a large age range. While we restricted our analyses to speech directed at infants aged six to 20 months, it is likely that at least some properties of CDS change across this time frame (see e.g., Kunert et al., 2011; Raneri et al., 2020; Vosoughi & Roy, 2012). For instance, Kunert et al. (2011) reported evidence to suggest that two of the cues investigated in the current study (syllable structure and word length) become more complex in English CDS as children get older and start to use more

complex syllables and words themselves. Therefore, longitudinal research of the type we have reported here would be a valuable addition to the literature.

Conclusion

We conducted the first corpus analysis investigating a broad range of word segmentation cues in German CDS, finding a highly reliable word-initial stress pattern, higher within-word and backwards TPs, and a Zipfian-like distribution (Zipf, 1935, 1949) of word and syllable frequencies. We also found slight differences of syllable structures between positions within a word, a prevalence of monosyllabic words, and especially highly frequent, short function words, and finally, a significant amount of single-word utterances. All of the cues we examined have the potential to aid word segmentation, and of course, might boost the effect when infants can draw on a combination of cues, as is the case in natural language (Brent & Cartwright, 1996; Matzinger et al., 2019).

Acknowledgements

Appendix: Metadata on the CHILDES corpora

Table A2.1. *Background information on the data included in the corpus.*

Corpus	Child	Age	Length of recording	Total number of utterances	Total number of utterances excl. unintelligible utterances	Speaker	Number of utterances per speaker	Number of utterances per speaker excl. unintelligible utterances
Caroline	Caroline (female (f))	00;10.01	00:06:44	22	22	MOT	22	22
		00;10.02	00:06:37	20	20	MOT	20	20
		00;11.25	00:18:22	81	79	MOT	63	63
						FAT	18	16
		01;00.19	00:11:04	37	37	MOT	37	37
		01;00.23	00:16:37	69	69	MOT	53	53
						FAT	16	16
		01;01.02	00:12:06	61	61	MOT	61	61
		01;01.04	00:01:46	10	10	MOT	10	10
Manuela	Dasca (male (m))	00;06.13	00:00:34	5	5	MOT	5	5
	Nibra (m)	00;10.24	00:00:24	5	4	MOT	5	4
	Oskoa (m)	00;10.12	00:01:05	28	28	MOT	28	28
		00;10.12	00:00:50	18	18	MOT	18	18
	Viala (m)	00;06.13	00:01:08	21	21	MOT	21	21
	Viwia (m)	00;10.20	00:00:15	7	7	MOT	7	7
Miller	Kerstin (f)	01;03.22	Approx. 00:15:00*	214	211	MOT	199	198
						FAT	13	11
						OBS	2	2

		Age	Length			Speaker		
		01,04.13	Approx. 00:30:00*	421	416	MOT	284	283
						FAT	22	18
						OBS	79	79
						VIS	36	36
Rigol	Corinna (f)	01;00.09	00:31:22	714	657	MOT	379	356
						FAT	322	289
						OBS	13	12
		01;00.23	00:32:38	629	577	MOT	279	257
						FAT	348	318
						OBS	2	2
		01;01.08	00:33:44	487	479	MOT	463	455
						OBS	24	24
	Cosima (f)	01;08.13	00:30:38	449	443	MOT	354	350
						FAT	14	12
						OBS	81	81
Wagner	Katrin (f)	1;05.15	03:22:00	855	803	MOT	607	570
						FAT	248	233
Total			07:32:54	4153	3967			

Notes: The Caroline corpus contains utterances consisting of several sentences which makes the utterances longer than the ones in the other corpora; that is, to work with comparable numbers, the number of utterances in the Caroline corpus should be increased. * The length of recording of the Miller corpus is an estimate based on the number of utterances in the datasets. The speaker abbreviations are: MOT = mother, FAT = father, OBS = observer/researcher, and VIS = visitor.

Chapter 3

Close encounters of the word kind: Attested distributional information boosts statistical learning

Abstract[6]

Statistical learning, the ability to extract regularities from input (e.g., in language), is likely supported by learners' prior expectations about how component units co-occur. In this study, we investigated how adults' prior experience with sublexical regularities in their native language influences performance on an empirical language learning task. Forty German-speaking adults completed a speech repetition task in which they repeated eight-syllable sequences from two experimental languages: one containing disyllabic words comprised of frequently occurring German syllable transitions (naturalistic words) and the other containing words made from unattested syllable transitions (non-naturalistic words). The participants demonstrated learning from both naturalistic and non-naturalistic stimuli. However, learning was superior for the naturalistic sequences, indicating that the participants had used their existing distributional knowledge of German to extract the naturalistic words faster and more accurately than the non-naturalistic words. This finding supports theories of statistical learning as a form of chunking, whereby frequently co-occurring units become entrenched in long-term memory.

Introduction

Humans are exquisitely sensitive to the regularities in their environment. Statistical learning (SL), the ability to draw on these regularities, is hypothesised to underlie learning across all sensory domains. Although it is indisputable that humans are capable of SL (which might rely upon multiple interacting mechanisms, see Frost et al., 2015), the totality of the parameters influencing SL are still yet to be mapped out. In our study, we examined the degree to which SL of linguistic stimuli is influenced by prior knowledge of attested syllable transitions present in natural language. That is, we asked whether and how adults' prior experience with the sublexical regularities in their native language in the form of syllable bigrams would permeate into the laboratory, such that it would enhance the adults' performance on an empirical language learning task when the distributional properties of the to-be-learned material aligned with those of the natural language.

In a canonical auditory SL task using linguistic stimuli, participants listen to a stream of speech that contains to-be-discovered words that are defined by statistical regularities (e.g., Saffran, Newport, et al., 1996). The discovery of the statistical segmentation effect heralded great promise for non-nativist approaches to language acquisition because it suggested the existence of a powerful learning mechanism (or *mechanisms*) that can induce structure from the input and thus questioned the need to postulate innately specified language-specific knowledge. That even very young infants are capable of SL is not controversial; however, the parameters that influence the process are still not well understood. This is partly due to the fact that much of the research on the topic has been conducted independently from other fields in cognitive psychology (Frost et al., 2019), such that connections to older disciplines concerned with learning and memory have not always been made. Yet, any task concerning learning of linguistic stimuli should be expected to conform to well-known properties of verbal memory, with SL being no exception (Isbilen et al., 2020; Vlach & DeBrock, 2017; Vlach & Sandhofer, 2014).

Since as far back as Ebbinghaus (1885, 1913), researchers have known that verbal learning is most effective when learners build upon prior experience. Accordingly, if researchers are to take the results of SL research to be ecologically valid, they should not expect participants to come into the laboratory without prior implicit assumptions about how linguistic stimuli like phonemes and syllables are ordered (Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Siegelman et al., 2018) and should instead expect participants to learn best when the target language is consistent with those assumptions. Such *historically-contingent* and in many instances *top-down* influences on performance suggest that the output of SL shapes future learning.

Background literature

A growing number of studies have shown that prior knowledge and expectations derived from a speaker's native language shape subsequent SL in a number of ways. This process begins very early. For example, Lew-Williams and Saffran (2012) found that infants' statistical segmentation of novel words from continuous speech was guided by their experience with words of the same versus a different length in a pre-training phase such that segmentation was only possible when words were the same length in both exposure phases. Similarly, research has revealed a significant benefit of starting small during incremental SL, with learners bootstrapping upon initial experience with simpler structures. Zettersten et al. (2020) demonstrated that adults' prior experience with a simplified nonadjacent dependency-learning task boosted later learning of a more complex instantiation of the same structure (see also Lany & Gómez, 2008, for similar findings with infants). In related work, Lai and Poletiek (2011) found that exposure to simple AB dependencies helped subsequent learning of longer, more complex strings containing centre-embeddings. Together, these findings provide converging evidence that prior experience scaffolds for future learning of related material.

Importantly, similar transfer effects have been found to emerge through experience with natural as well as artificial languages and across different learning contexts. For instance, Potter et al. (2017) documented a language-experience effect in novice learners of Mandarin after just two semesters of study. In Potter et al.'s study, participants completed a SL task in which the artificial language overlapped with Mandarin insofar as it was tonal in nature (see also Wang & Saffran, 2014) to see whether participants' experience with related material would impact learning. Participants completed the task at two time points separated by an interim learning period of six months. Although participants' performance was initially at chance, there were significant improvements at Time 2, with participants achieving accuracy scores of 66% on a two-alternative forced-choice (2AFC) segmentation test, indicating that participants' SL performance had been critically shaped by their experience with relevant linguistic input. Non-Mandarin-learning controls exhibited no such improvements, performing at chance on both occasions.

Other studies have investigated how statistical distributions in naturalistic language constrain SL in laboratory settings, with a large focus on phonotactic probabilities (for reviews of how phonotactics impact on early acquisition see Johnson, 2016; Jusczyk, 2002). For instance, Finn and Hudson Kam (2008) showed that participants could only successfully segment statistically defined novel words from continuous speech when the words contained syllables that followed phonotactic constraints of English (see also Toro et al., 2011). Mersad and Nazzi (2011) showed that the presence of words containing high phonotactic probability served as anchors that successfully aided segmentation compared to a condition in which all words had a uniform but lower phonotactic probability. Dal Ben et al. (2021) replicated this latter effect using a more narrowly defined difference in phonotactic probability across experimental conditions. Overall, these studies provide strong evidence for the suggestion that fine-grained features of natural language, in this case phonotactic probability, shape participants' subsequent expectations about how their input is shaped.

This is consistent with results reported by Siegelman et al. (2018), who found that performance on an auditory SL segmentation task was predicted by post hoc ratings of how word-like test items and foils were.

These findings provide converging support for the notion that prior experience can shape future learning at multiple levels of description, boosting performance when the properties of the input align. Building on this, Elazar et al. (2022)[7] investigated the specific hypothesis that entrenched memory traces for syllable co-occurrences in natural language boost SL. They tested two groups of Spanish-speaking participants on an auditory SL task. One group listened to a Spanish-like speech stream in which transitional probabilities (TPs) of the experimental words were highly attested in Spanish while the other group listened to a Spanish-unlike speech stream in which TPs of the experimental words were rarely attested in Spanish. Participants were tested on a lexical decision task for experimental words and respective Spanish-like or Spanish-unlike foils. Elazar et al. found that participants in the Spanish-like condition were better at accepting words than participants in the Spanish-unlike condition, indicating that participants' prior knowledge of Spanish syllable trigrams facilitated their SL. Furthermore, participants in the Spanish-like condition were worse at rejecting (Spanish-like) foils than participants in the Spanish-unlike condition were at rejecting (Spanish-unlike) foils, suggesting that participants' knowledge of Spanish also (mis)led them to accept familiar foils. Overall, the results suggest that participants indeed entered the experiment with entrenched memory traces for syllable co-occurrences on which they drew to process and learn new input.

The present study

In our study, on which we worked in parallel to Elazar et al.'s (2022) study, we tested the almost identical hypothesis that entrenched memory traces for syllable bigrams in natural language boost SL.

[7] Elazar et al.'s (2022) paper was published during the review process of the paper, upon which this chapter is based.

However, Elazar et al. (2022) used a between-participants design following the typical exposure-phase–test-phase structure, whereas we used a within-participants design using verbal repetition. This within-participants design allowed for a more stringent test of the entrenchment hypothesis because differences between conditions could not be attributed to differences between participants, in addition to allowing us to track the emergence of learning across the course of the experiment. In using verbal repetition, we built upon recent developments in the measurement of SL that have been inspired by the verbal learning literature. Participants' recall on verbal tests of short-term memory has been shown to be both sensitive to newly learned material (e.g., Majerus et al., 2004) and mediated by their long-term lexical knowledge (e.g., Kowialiewski & Majerus, 2018; Majerus et al., 2004, 2012). In recent work building upon Majerus et al. (2004), Isbilen et al. (2020) investigated the utility of verbal recall as a measure of auditory SL in a triplet segmentation task. Isbilen et al. showed that, after familiarisation with continuous speech, adult participants were better able to repeat syllable sequences that followed the statistical distribution of the familiarisation stream than they were to repeat random and unattested syllable sequences (for similar results from children see Kidd et al., 2020). In some cases, performance was predicted by distributional statistics derived from spoken and written corpora. Isbilen et al. suggested that the results were consistent with chunking models of SL (e.g., Christiansen & Chater, 2016; Jones, 2012; Perruchet & Vinter, 1998; Robinet et al., 2011) in which the repetition of syllable sequences creates word-like phonological units via their association strength; specifically, their high TPs.

These processes and their explanation seem functionally equivalent to another effect in the literature – the *Hebb repetition effect* (Hebb, 1961; see also Page & Norris, 2009; Smalle et al., 2016; Szmalec et al., 2012). However, the one difference between auditory SL tasks and Hebbian learning tasks is that, although SL tasks typically measure the outcome of learning following familiarisation, Hebbian learning tasks track learning of sequence regularities across time. This is an important

gap in SL research, with researchers not yet knowing *how* learning proceeds during familiarisation. The evidence that exists has suggested that learners gradually come to recognise structured sequences as containing higher level chunks over the course of exposure, suggesting that learners engage in the dual processes of (i) binding/chunking adjacent syllables together and (ii) storing them in long-term memory (Batterink & Paller, 2017).

In our study, we used a sequence-repetition method common in Hebbian learning studies to also investigate how existing knowledge of sublexical regularities influences the trajectory of SL over time. Our article makes two contributions to the literature: (i) we report detailed corpus data on syllable transitions in German, and (ii) we determine how these attested transitions contribute to SL across the course of learning. Thus, building on previous investigations of the effect of prior knowledge on SL (e.g., Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Siegelman et al., 2018; Toro et al., 2011), we examined how knowledge of the statistical properties of participants' native language – focusing on syllable bigrams – influences subsequent processing and learning of an artificial language that is built with those properties in mind. We extracted the TPs between syllable pairs in natural German and used this information to create artificial language sequences containing words that were either based on the natural German TPs (i.e., *naturalistic sequences*) or not (i.e., *non-naturalistic sequences*), examining learning of these sequences relative to scrambled foils.

Under the assumption that SL for language involves the tracking *and* subsequent long-term registration of distributional information, we hypothesised that learners would use their existing distributional knowledge of German to shape their processing of new input. To test this hypothesis, we measured learning using a speech production task in which the participants repeated either unstructured sequences of random syllable combinations (foils) or structured sequences containing novel words – with these words either adhering to German

syllable distribution (i.e., *naturalistic* sequences), or not adhering to German syllable distribution (i.e., *non-naturalistic* sequences). We predicted that, overall, participants' repetition (and therefore learning) of the structured sequences would be better than their repetition of the foils, but that participants' repetition of the naturalistic sequences would be better than their repetition of the non-naturalistic sequences. An advantage of our method was that, in contrast to past research measuring learning via 2AFC and repetition after familiarisation, it enabled us to track learning across the three conditions across the course of the experiment. We also predicted that, over time, participants' repetitions would improve for both types of structured sequences. Importantly, we expected to see the strongest improvements for *naturalistic* sequences and predicted that performance would improve more rapidly for *naturalistic* than for *non-naturalistic* sequences because naturalistic sequences better aligned with German syllable distributions.

Method

All materials, data, analyses, and results for this article are openly available via the Open Science Framework (OSF; https://osf.io/4dsmy); the results of the experiment testing the validity of the stimuli can also be accessed via OSF (https://osf.io/p9fcm).

Participants

Forty native German-speaking adults (28 self-identified female, 12 self-identified male; M_{age} = 23.9 years, *SD* = 5.58) without any known hearing, speech, or language disorders participated in the experiment. The participants registered via the Max Planck Institute's internal database; we made additional announcements at Radboud University Nijmegen and on social media, which also allowed participants to register via email. The sample size of 40 participants was informed by a power analysis conducted in R 4.0.2 (R Core Team, 2022) using the package *simr* 1.0.5 (Green & MacLeod, 2016). We based the simulations on data collected by Isbilen et al. (2017), who had

compared two conditions similar to our non-naturalistic and unstructured foil sequences in a serial recall task following an exposure phase. Our simulations indicated that a sample of 16 participants would be sufficient to detect an effect size of a –0.1 syllable recall difference between naturalistic and non-naturalistic sequences as well as between non-naturalistic and foil sequences during the later stages of our experiment, which is comparable to the test phase in Isbilen et al.'s study (for more details, see the Analysis folder on OSF). We increased the sample size to 40 because the participants in our experiment were exposed to multiple experimental languages while performing the serial recall task (i.e., without prior exposure phase), which would decrease the effect and also make the model more complex (because we included the additional variable block, which was not present in Isbilen et al.'s study). We decided to not perform an analysis modelling our entire experiment (including block) because this would have entailed a considerably more complex simulation that would have been based purely on our own intuitions rather than on previous data.

The study was approved by the Ethical Committee of the Faculty of Social Sciences, Radboud University Nijmegen, and was carried out in accordance with the World Medical Association Declaration of Helsinki. All participants gave written informed consent prior to their participation in the study. They were free to withdraw at any time and were compensated (€8) upon completing the 45-minute session.

Design

We employed a serial repetition task based on studies of the Hebb repetition effect (Hebb, 1961; Page & Norris, 2009), which required the participants to repeat sequences of syllables aloud, with these repetitions then being scored for accuracy. The study had a within-participants design, with all the participants receiving three different types of sequences: (i) naturalistic sequences, (ii) non-naturalistic sequences, and (iii) unstructured foils. The naturalistic and non-naturalistic sequences were structured, with each containing four

disyllabic experimental words, whereas foils were unstructured, containing the same syllables as the structured sequences but in a scrambled order.

Materials

Corpus analysis

We created the speech stimuli from a pool of 12 German syllables (*fa, ge, gei, mi, mo, nu, pa, sa, su, ti, ver, zu*) obtained from a corpus analysis of the 1,000 most frequent German words in the CHILDES database (MacWhinney, 2000)[8], which corresponded to over three million word tokens. We chose to draw our materials from child-directed language for two reasons. First, because words that are highly frequent in child-directed language will also have an early age-of-acquisition, we logically deduced that these words would have sublexical transitions (i.e., bigrams) that would have the greatest likelihood of being entrenched. Second, this study was part of a larger project that tested the effects under investigation in developmental populations (cf. Chapter 4). We chose the syllables from syllable pairs (i.e., bigrams) occurring with high within-word backwards TPs, relying on backwards TPs because our corpus analysis of child-directed speech in Chapter 2 showed that backwards TPs were significantly more reliable cues to wordhood than forwards TPs in German speech (Stärk et al., 2022; for a similar cross-linguistic analysis see Saksida et al., 2017)

We then used the syllables to form 12 disyllabic "words": six words for each of the two structured sequence types (naturalistic: *gefa, minu, moti, pagei, versu, zusa*; non-naturalistic: *fazu, geimi, nuver, samo, suge, tipa*). As summarised in Table 3.1, the extracted bigrams yielded the six naturalistic words in which the two syllables co-occurred with

---

[8] We included the following corpora from the CHILDES database (MacWhinney, 2000) in our analysis: Caroline (Von Stutterheim, 2010), Grimm (Grimm, 2006, 2007), Leo (Behrens, 2006), Manuela (Wagner, 2006), Miller (Miller, 1979), Rigol (Rigol, 2007), Stuttgart (Lintfert, 2010), TAKI (Lintfert, 2010), and Wagner (Wagner, 1974, 1985).

Table 3.1. *Syllable frequencies, pair frequencies, and forwards and backwards transitional probabilities of the stimuli derived from the corpus analysis.*

Pair	Frequency			Transitional probability	
	Syllable 1	Syllable 2	Pair	Forwards	Backwards
mi nu	6,472	454	454	.070	1.000
pa gei	46,359	368	368	.008	1.000
ver su	14,010	344	344	.025	1.000
ge fa	1,839,597	3,133	1,586	.001	.506
zu sa	14,460	4,670	1,994	.138	.427
mo ti	1,748	2,467	525	.300	.213

relatively high backwards TPs in natural German speech but importantly were not recognisable alone as words (TP > .20, M_{TP} = .69, range = .21–1.00)[9]. To create the non-naturalistic words, we concatenated the same 12 syllables in a different order, such that their syllable pairs did not co-occur in natural German (TP = 0). Each syllable occurred once in each set of words, and we counterbalanced the position of syllables within words such that, if a syllable appeared word-initially in the naturalistic set of words, it was word-final in the non-naturalistic set, and vice versa. For the unstructured foil sequences, we scrambled the syllables, such that these sequences contained no learnable regularities. We carefully constructed the foils to avoid inadvertently creating words from both German and the experimental languages. Because all three sequence types comprised the same 12 syllables, the frequencies of the syllables in natural German presented in Table 3.1 applied to all conditions. However, the non-naturalistic words comprised syllable pairs which did not occur in our corpus sample of natural German (i.e., their pair frequencies as well as their forwards and backwards TPs were 0). Likewise, the unstructured foils did not comprise any patterns found in the corpus.

---

[9] In an analysis of TPs in child-directed speech across nine languages, Saksida et al. (2017) reported a mean between-word TP of .11, compared to a mean within-word TP of .25, whereas for German, we reported a mean between-word TP of .11 and a mean within-word TP of .33 in Chapter 2 (Stärk et al., 2022). Thus, our naturalistic words were, on average, more indicative of word-like units than between-word transitions.

Our design involved the explicit assumption that high TPs are more word-like and, thus, that the participants would require less exposure to chunk adjacent syllables into words. The implicit assumption of our sequence repetition method was that these transitions would thus be easier to repeat. In order to collect independent evidence in support of the explicit assumption that the naturalistic words would be more word-like, we conducted a separate experiment in which we asked German-speaking participants to select our naturalistic or non-naturalistic words for wordiness in comparison to foils in a 2AFC task without familiarisation (i.e., the participants had no prior training on the words). The participants successfully identified the naturalistic words at above chance levels in comparison to foils but did not do so for the non-naturalistic words. These results were consistent with the argument that our naturalistic words, when presented in isolation, were more identifiable as German-like than our non-naturalistic words (for full details, see Chapter 5's Experiment 3 or Appendix S1 in the Supporting Information online).

Stimuli characteristics

The stimuli were recorded by a female native speaker of German, who recorded individual unaccentuated syllables in isolation. We adjusted the syllable recordings using the sound editing program Audacity (Audacity Team, 2018) to ensure uniformity in length, resulting in an average syllable duration of 377 ms (range = 352–416).

Within the context of the experiment, each structured sequence type contained perfect within-word TPs (*structured sequences:* within-word TPs = 1.00, between-word TPs ≤ .25; compared to *unstructured sequences* where TPs between all syllables were generally low, with TPs ≤ .125). Note, however, that participants were tested on all three sequence types. Thus, across the whole experiment, accounting for the repeated use of syllables across each type of sequence, within-word TPs for *both* structured sequences were .33, and TPs for all other syllable pairs were less than or equal to .125.

Syllables were combined into 72 sequences, 24 of each sequence type. Each sequence was eight syllables long, which equated to four experimental words (i.e., bigrams). Within sequences, each syllable was followed by 500 ms of silence, and the final syllable of a sequence was followed by a beep to indicate the beginning of the repetition portion of the trial. Because the syllables had an inter-stimulus interval of 500 ms, we emphasise that our study was not a segmentation task in the classical sense. Rather, our choice of method allowed us to determine (i) whether attested syllable bigrams are more naturally grouped during recall and (ii) how this attested knowledge influences learning incrementally across time. In order to track the participants' incremental learning, we divided the experiment into 12 blocks of six sequences, with each block containing two sequences of each type. Within each block, sequences were presented pseudo-randomly, with no direct repetition of a particular sequence type. Across the whole experiment, each word occurred 16 times in total, with words appearing equally often in each position within a sequence (for more information on the stimuli and their creation, see the Materials folder on OSF).

Procedure

We sent the participants an informed consent document one day prior to the day that the study took place. Upon arrival in the laboratory, they were reminded of the task instructions and were told that the study was to investigate adults' repetition of language, but no mention was made of the learnable patterns contained within the input. The participants completed the study in isolation in a sound-attenuated booth, with sequences being played over closed-cup headphones using the software *Presentation* (Neurobehavioral Systems, 2014). The participants repeated the sequences into a microphone, and these were recorded by the computer for offline coding.

Before the experiment began, the participants first received three (unstructured) practice sequences that were six syllables long, comprising a different scrambled set of syllables (*ba, fun, gi, re, se, to*).
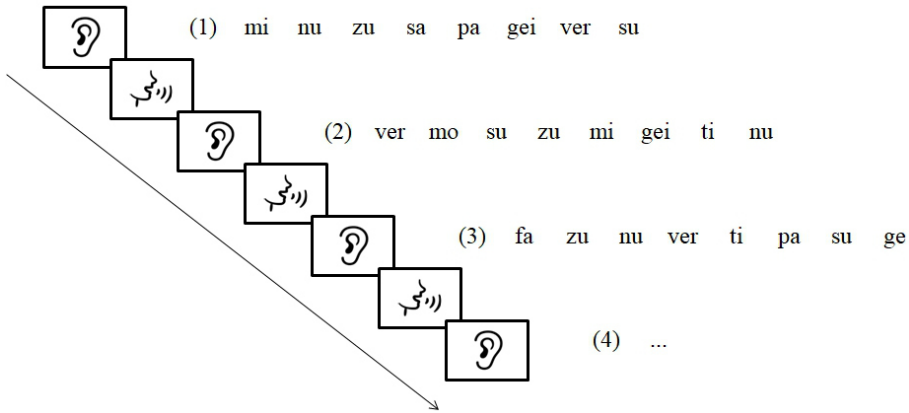
(1) mi nu zu sa pa gei ver su

(2) ver mo su zu mi gei ti nu

(3) fa zu nu ver ti pa su ge

(4) ...

Figure 3.1.   *Three example experimental sequences. On each trial, participants listened to an eight-syllable sequence and then repeated it. (1) = one naturalistic sequence; (2) = one unstructured foil sequence; (3) = one non-naturalistic sequence.*

After completing the practice sequences, the participants proceeded to the main experiment. In each trial, the participants heard a sequence of eight syllables followed by a beep (see Figure 3.1). Upon hearing the beep, they were required to repeat the sequences as best they could. At the halfway point, the participants were given the opportunity to take an optional break. At the end of the session, they were debriefed and paid for their time.

Data preparation

To prepare the data for our analyses, we first transcribed the recordings of the participants' verbal responses. All responses were transcribed by the experimenter, and two naïve coders each transcribed 10% of the recordings (i.e., data for four participants) for the purpose of performing reliability checks. Inter-transcriber reliability analyses revealed strong reliability between transcribers, following the more conservative interpretation of the kappa statistic suggested by McHugh (2012) (*syllable level:* observed agreement = 83.0%; $\kappa = 0.87$ with 95%

CIs of [0.84, 0.89]; *bigram level:* observed agreement = 87.2%; κ = 0.88 [0.84, 0.92]).

We coded the accuracy of participants' responses sequence-by-sequence, comparing the verbal response against the sequence that the participants had heard. We computed scores at the syllable level and at the bigram level. At the syllable level, the participants received 1 point for each syllable repeated correctly in the correct position (for a maximum of 8 points per sequence). At the bigram level, the participants received 1 point for each bigram (i.e., syllable pair) repeated correctly in the correct position (for a maximum of 4 points per sequence). A bigram denoted an experimental word in the structured sequences. The participants' performance at this level, therefore, provided crucial information about whether they had recalled sequences better because of learning the experimental words, rather than indirectly assessing the learning solely at the syllable level. In the unstructured sequences, the four bigrams per sequence were the random syllable pairs in Positions 1 and 2, 3 and 4, 5 and 6, and 7 and 8 (with different syllable combinations in each position across each sequence). Table 3.2 illustrates how we applied the coding scheme to potential repetitions by the participants.

This strict coding scheme was conservative in the sense that it required the participants to recall elements in the correct position and thus did not give any credit for syllables recalled in the correct order after only one syllable was missed or added in the repetition (e.g., as in the "B C D E F G H A" and "A X B C D" cases in Table 3.2). We also used two further coding schemes, including a serial order coding scheme based on Isbilen et al.'s (2017) study, which relaxed the strict positional requirement and which were thus more lenient and gave more credit to the participants. However, because all three analyses converged in the same direction, we have chosen to report only the most conservative scheme here. The analyses for all three coding schemes can be found in the Analysis folder on OSF.

Table 3.2. *Example scorings of participants' repetitions of the sequence A B C D E F G H where each letter represents one syllable.*

Repetition	Syllable score	Bigram score
A B C D F H –	4 (A B C D)	2 (AB CD)
A X B C D –	1 (A)	0
A B –	2 (A B)	1 (AB)
X Y C D X Z G H	4 (C D G H)	2 (CD GH)
B C D E F G H A	0	0

Results

The aims of our analyses were twofold: (i) to examine performance on each type of sequence and (ii) to examine the time course of learning. We predicted that the participants would recall naturalistic sequences better than non-naturalistic sequences and non-naturalistic sequences better than unstructured foil sequences. We also predicted that the participants would *improve faster* on the naturalistic sequences than on the non-naturalistic sequences. To test the hypotheses regarding the incremental learning throughout the study, we ran our analyses by experimental block and exposure phase, that is, we combined blocks to determine *early, intermediate,* and *late* exposure phases, respectively.

Analysis by experimental block

We analysed the data in R 4.1.3 (R Core Team, 2022) using generalised linear mixed-effects models. We specified a Poisson distribution because the dependent variables (i.e., syllable and bigram recall) were count data. The models were computed using the package *lmerTest* 3.1-3 (Kuznetsova et al., 2017; based upon *lme4* 1.1-28 by Bates et al., 2015). We computed the same models with syllable recall and bigram recall as the dependent variables to test overall recall and recall of the experimental words. Models were fit with a fixed effect of sequence type using sliding contrasts (*naturalistic: .5* vs. *non-naturalistic: –.5*, and *non-naturalistic: .5* vs. *foil: –.5*) to examine whether learning differed across the experimental conditions and with a fixed effect of block entered as a centred continuous variable to

examine learning over the course of the experiment as well as the interaction of the two variables. We fit the maximal model supported by the data (Barr et al., 2013; Bates et al., 2018), controlling for participants and items as random intercepts, with sequence type and block as random slopes of participants (due to our within-participants design, with participants being exposed to all sequence types and blocks) but not as random slopes of items (because sequences differed between sequence types and blocks).

We checked the models for evidence of singularity in the variance-covariance matrix and for evidence of overfitting the random effects structure by conducting a principal component analysis. Models showing evidence of singularity or overfitting were simplified (for the documentation, see the Analysis folder on OSF). To determine significance, we used an alpha level of .05. Furthermore, we have reported bootstrapped 95% confidence intervals for the beta estimates of the model predictors, based on 1,000 iterations, as well as the marginal and conditional R^2 effect sizes of the models as goodness-of-fit estimates. These R^2 values denote the proportion of the variance explained by the model both with (conditional R^2) and without (marginal R^2) controls for sources of random variance (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013).

Crucially, there was a significant main effect of sequence type at both the syllable and bigram level, with participants displaying better recall for naturalistic than non-naturalistic sequences (see Table 3.3), in line with our experimental hypothesis. Recall was also better for non-naturalistic sequences than for unstructured foil sequences (for a visualisation of participants' syllable and bigram recall accuracy, see Figure 3.2 and Figure 3.3, respectively).

Regarding participants' performance over time, there was a main effect of block, with participants improving over the course of the experiment. Critically, there was also a significant interaction of sequence type and block, with participants' recall improving more

Table 3.3. *Summary of the linear mixed-effects models investigating the influence of sequence type and block on participants' syllable and bigram recall.*

Parameter	b	95% CI	SE	t	p
Syllable level					
(Intercept)	1.00	[0.87, 1.12]	0.06	15.90	< .001
Nat. vs. Non-nat.	0.10	[0.06, 0.14]	0.02	4.70	< .001
Non-nat. vs. Foils	0.04	[0.00, 0.07]	0.02	2.20	.03
Block	0.11	[0.07, 0.14]	0.02	6.65	< .001
Nat. vs. Non-nat. × Block	0.03	[0.01, 0.06]	0.01	2.28	.02
Non-nat. vs. Foils × Block	0.03	[0.01, 0.05]	0.02	1.72	.09
Bigram level					
(Intercept)	–0.12	[–0.29, 0.04]	0.09	–1.43	.15
Nat. vs. Non-nat.	0.17	[0.10, 0.23]	0.03	4.97	< .001
Non-nat. vs. Foils	0.09	[0.03, 0.15]	0.03	3.02	.003
Block	0.14	[0.11, 0.18]	0.02	7.35	< .001
Nat. vs. Non-nat. × Block	0.04	[0.00, 0.09]	0.02	1.88	.06
Non-nat. vs. Foils × Block	0.03	[–0.03, 0.08]	0.03	1.11	.27

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Model fit syllable level: AIC = 10,975; BIC = 11,052; $R^2_{marginal}$ = .079; $R^2_{conditional}$ = .091; model fit bigram level: AIC = 7,047; BIC = 7,125; $R^2_{marginal}$ = .058; $R^2_{conditional}$ = .313.
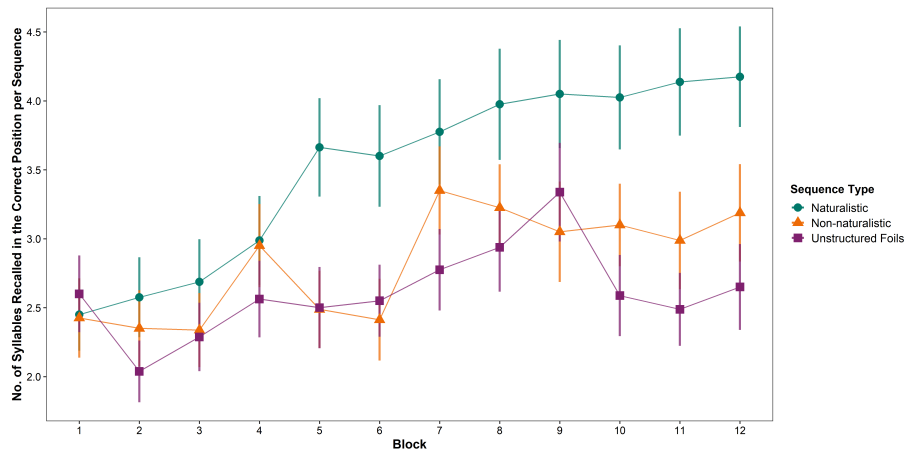

Figure 3.2. *Mean recall of syllables (out of eight per sequence) for the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.*

Figure 3.3. *Mean recall of bigrams (out of four per sequence) in the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.*
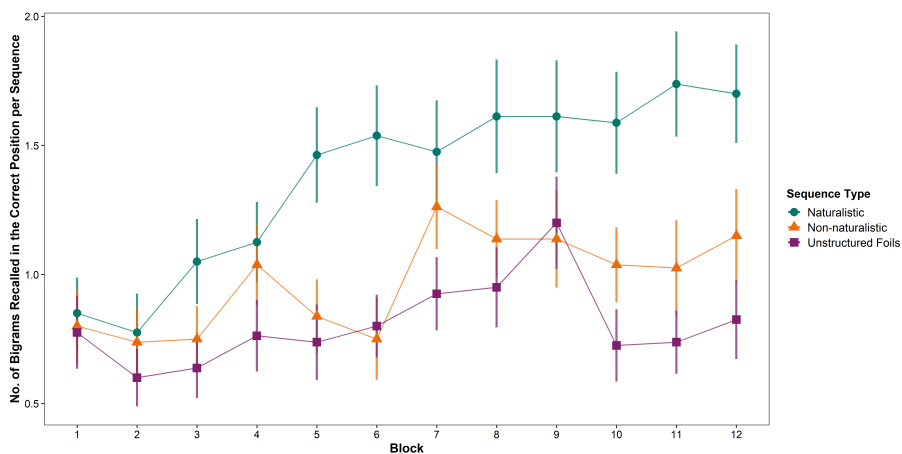
rapidly over the course of the experiment for naturalistic sequences than for non-naturalistic sequences. However, at the bigram level, this did not meet the alpha level that we had chosen for determining significance. Participants did not improve significantly over time when recalling the non-naturalistic sequences in comparison to the unstructured foil sequences; however, participants' recall of non-naturalistic sequences numerically improved after the break at the halfway point between Blocks 6 and 7. The random-effects structure improved the model-fit in both cases.

Analysis by exposure phase

Although the above results depicted the participants' overall improvement throughout the entire experiment, they did not reveal at which point learning began to emerge within the task. To unpack this, we divided the experiment into three phases (*early exposure:* Blocks 1–4; *intermediate exposure:* Blocks 5–8; *late exposure:* Blocks 9–12), testing the hypothesis that learning in the naturalistic condition would

be faster than in the non-naturalistic condition. The variable exposure phase was added as a fixed effect into a new analysis instead of block. We fit the maximal model supported by the data (Barr et al., 2013; Bates et al., 2018) with sequence type (sliding contrast: *naturalistic:* .5 vs. *non-naturalistic:* −.5, and *non-naturalistic:* .5 vs. *foil:* −.5) and exposure phase (sliding contrast: *early exposure:* −.5 vs. *intermediate exposure:* .5, and *intermediate exposure:* −.5 vs. *late exposure:* .5) as well as their interaction as fixed effects, and random intercepts and slopes for participants and items, where appropriate (as described previously).

In addition to a significant main effect of sequence type, there was a main effect of exposure phase, with the participants improving between the early and intermediate exposure phase (see Table 3.4 for the analysis at the syllable level and Table 3.5 for the analysis at the bigram level; for figures illustrating the syllable and bigram recall accuracy over the three phases see the Analysis folder on OSF). The participants also improved numerically between the intermediate and late exposure phase, but this did not meet the level that we had set for significance. Importantly, the interaction of sequence type and exposure phase was significant, with greater improvements on naturalistic relative to non-naturalistic sequences between the early and intermediate exposure phase. There was no difference in improvement between naturalistic and non-naturalistic sequences between the intermediate and late exposure phase. Improvement in recall of the non-naturalistic sequences did not differ from the improvement in recall of unstructured foil sequences, either between the early and intermediate exposure phase or between the intermediate and late exposure phase.

Discussion

Prior knowledge of syllable co-occurrences facilitates statistical learning

SL is assumed to underlie learning across many fundamental domains of cognition, most prominently language (e.g., Christiansen &

Table 3.4. *Summary of the linear mixed-effects model investigating the influence of sequence type and exposure phase on participants' syllable recall.*

Parameter	b	95% CI	SE	t	p
(Intercept)	1.00	[0.88, 1.13]	0.06	15.95	< .001
Nat. vs. Non-nat.	0.10	[0.06, 0.14]	0.02	4.73	< .001
Non-nat. vs. Foils	0.03	[0.00, 0.07]	0.02	2.09	.04
Early vs. Intermediate	0.09	[0.05, 0.12]	0.02	5.04	< .001
Intermediate vs. Late	0.03	[0.00, 0.06]	0.02	1.89	.06
Nat. vs. Non-nat. × Early vs. Intermediate	0.05	[0.02, 0.09]	0.02	3.18	.001
Nat. vs. Non-nat. × Intermediate vs. Late	0.00	[−0.03, 0.04]	0.02	0.22	.83
Non-nat. vs. Foils × Early vs. Intermediate	0.00	[−0.03, 0.04]	0.02	0.12	.91
Non-nat. vs. Foils × Intermediate vs. Late	0.01	[−0.02, 0.05]	0.02	0.68	.50

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Model fit: AIC = 10,986; BIC = 11,135; $R^2_{marginal}$ = .047; $R^2_{conditional}$ = .344.

Table 3.5. *Summary of the linear mixed-effects model investigating the influence of sequence type and exposure phase on participants' bigram recall.*

Parameter	b	95% CI	SE	t	p
(Intercept)	−0.12	[−0.29, 0.06]	0.09	−1.44	.15
Nat. vs. Non-nat.	0.16	[0.10, 0.23]	0.03	4.96	< .001
Non-nat. vs. Foils	0.09	[0.03, 0.15]	0.03	3.18	.001
Early vs. Intermediate	0.14	[0.09, 0.18]	0.02	5.79	< .001
Intermediate vs. Late	0.04	[−0.01, 0.08]	0.02	1.53	.13
Nat. vs. Non-nat. × Early vs. Intermediate	0.07	[0.02, 0.13]	0.03	2.61	.009
Nat. vs. Non-nat. × Intermediate vs. Late	0.00	[−0.05, 0.05]	0.03	−0.08	.93
Non-nat. vs. Foils × Early vs. Intermediate	−0.01	[−0.07, 0.06]	0.03	−0.17	.86
Non-nat. vs. Foils × Intermediate vs. Late	0.02	[−0.05, 0.07]	0.03	0.58	.56

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Model fit: AIC = 7,048; BIC = 7,143; $R^2_{marginal}$ = .061; $R^2_{conditional}$ = .313.

Chater, 2016; Lidz & Gagliardi, 2015; Saffran, Newport, et al., 1996; Saffran & Kirkham, 2018). Although the existence of a human capacity for SL is clear, precisely how SL both depends and builds upon existing knowledge is still unclear (but see Elazar et al., 2022). Past research has shown that phonotactic probability constrains SL (Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Toro et al., 2011). In our study, we asked whether participants would draw on their prior knowledge of statistical distributions of syllables to inform their learning and processing of new linguistic input. We created two experimental languages for our native German-speaking participants; one informed by the naturally occurring TPs in German, as extracted from corpora, and another that was completely devoid of attested TPs. Breaking away from the classic format of SL paradigms that typically comprise separable training and testing phases, we presented these languages using a sequence-repetition speech-production task and tracked learning across the experiment. We hypothesised that the participants' repetitions would be more accurate and would improve more rapidly for naturalistic than for non-naturalistic sequences.

As we had predicted, recall accuracy was higher for naturalistic than non-naturalistic sequences, suggesting that the participants had drawn on their prior distributional knowledge of German to process the new experimental input. Additionally, the participants' prior experience boosted further learning of the naturalistic words in the initial stages of the experiment, increasing the recall advantage of the naturalistic sequences compared to the non-naturalistic sequences between the early and intermediate exposure phases. These findings are consistent with the idea that learners not only track syllable co-occurrences but that they also draw on this knowledge when processing subsequent input (Elazar et al., 2022; Siegelman et al., 2018), which in our study led to accelerated learning of naturalistic sequences from the beginning of the experiment. Thus, what we observed could be described as a kind of *Matthew Effect* for SL concerning syllable transitions (Merton, 1968; for similar arguments regarding literacy, see Stanovich, 1986), where those bigrams that were attested in the participants' native language

provided an advantage for future learning. This interpretation is consistent with older claims from the verbal learning literature, which has long argued that learning is a historically-contingent process that builds upon past experience (Ebbinghaus, 1885, 1913).

Overall, the results support the suggestion that participants draw upon their rich repository of existing knowledge during learning (Bertels et al., 2015; Finn & Hudson Kam, 2008; Lew-Williams et al., 2011; Lew-Williams & Saffran, 2012; Mersad & Nazzi, 2011; Onnis & Thiessen, 2013; Potter et al., 2017). An important issue concerns exactly how this existing knowledge is both represented and how it subsequently aids learning. Although many details are still to be ironed out, SL for language logically involves the discovery and registration of perceptual regularities that are then re-described into higher level representations based on existing knowledge and generalisation processes. Thus, in classic domains of enquiry like speech segmentation, TPs act as initial local cues alongside others like stress to help the listener bootstrap into the language (Cutler, 2012; Mattys & Bortfeld, 2016), after which lexical knowledge provides crucial anchors and top-down expectations about new to-be-learned material (e.g., Bortfeld et al., 2005; Lew-Williams et al., 2011; Mersad & Nazzi, 2012; for further evidence of top-down influence on the learning of adjacent dependencies, see Wang et al., 2020). We did not study segmentation per se, although we have no reason to postulate a different learning mechanism to explain our results. Accordingly, we suggest that the advantage that we observed for attested bigrams derived from this existing well-entrenched lexical knowledge providing expectations about how the input is structured, acting as local attractors through which the participants could chunk the stimuli better than when they had no, or indeed *incorrectly biasing*, expectations such as in the non-naturalistic condition (comparable to the bias for Spanish-like foils observed by Elazar et al., 2022).

Accordingly, we suggest that the results provide support for the idea that syllable co-occurrences are tracked and become more entrenched

with each encounter (Isbilen et al., 2017, 2020; Jost & Christiansen, 2017; Siegelman et al., 2018). Such entrenchment can be seen as a form of chunking that facilitates subsequent processing and production because participants can draw on stored chunks instead of individual syllables (e.g., Christiansen & Chater, 2016; Jones, 2012; Jones et al., 2021; Perruchet & Vinter, 1998; Robinet et al., 2011). The learning advantage seen for the sequences comprising attested TPs exemplified this further, with higher accuracy and faster learning seen for sequences that adhered to a distribution that should already have been well-entrenched within the participants due to their prior experience with German. This is in line with previous studies showing that participants drew on their long-term lexical knowledge to guide recall in short-term memory tasks (e.g., Jones & Macken, 2018; Kowialiewski & Majerus, 2018; Majerus et al., 2012; for neuroimaging evidence, see Tremblay et al., 2016). Together with Elazar et al.'s (2022) study, our study demonstrated that long-term linguistic knowledge guides future learning at the level of syllable transitions, thus complementing work on phonotactic probability (e.g., Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011).

There are potential parallels between our data and those from electroencephalogram (EEG) studies that have tracked SL across familiarisation. Notably, Batterink and colleagues have demonstrated that SL is a gradual, two-staged process of chunking adjacent syllables and storing them in long-term memory (Batterink, 2020; Batterink & Paller, 2017). The properties of the EEG signal suggested that participants initially treated the speech signal as a stream of syllables. However, across familiarisation participants entrained to higher levels of linguistic organisation as the syllables in the stream became more familiar to them, that is, participants were able to identify that some adjacent syllables frequently co-occurred and treated them as word-like, storing these frequently co-occurring syllable combinations in long-term memory. With this in mind, one interpretation of our data is that our participants were building upon their attested knowledge of German syllabic regularities to implicitly treat naturalistic syllable

pairs as word-like sooner than they did in the non-naturalistic condition, thus accounting for the difference in learning rate during the early and intermediate exposure phases of the experiment. Acquiring word-like representations of the non-naturalistic sequences compared to the foil sequences was more difficult for the participants and did not interact with the exposure phase. There are two likely reasons for this. First, the non-naturalistic sequences contained unattested TPs, and so, given the assumption that these matter, the participants were starting from the lowest of bases. Second, the non-naturalistic sequences contained the same syllables as the naturalistic sequences, which means that there could have been some interference from long-term knowledge of German.

Related to this latter point, the comparative difficulty that the participants experienced with learning the non-naturalistic sequences might have been due to the difficulty of simultaneously learning multiple languages, especially with the same syllables occurring in multiple words across sequence types. This is consistent with work by Page et al. (2013), who have shown that Hebbian learning is slower when structured sequences have item overlap (see also Antovich & Graf Estes, 2018, for evidence that bilingual but not monolingual infants can extract words from multiple experimental languages when these languages are presented interleaved). In previous studies of SL, adults learned only the first of two subsequently presented artificial languages, unless (i) there were contextual cues indicating the change between languages or (ii) the exposure to the second language was either tripled or initiated before a certain level of entrenchment was reached for the words of the first language (Bulgarelli & Weiss, 2016; Gebhart et al., 2009). In our study, this level of entrenchment had presumably already been reached for the naturalistic words when the participants entered the experiment, such that participants' predisposition for (and enhanced learning of) the naturalistic words might have biased learning in favour of the naturalistic sequences at the expense of the others. Ultimately, however, there was significant evidence that the participants *did* learn in the non-naturalistic condition

compared to the foil condition, and so acquiring multiple syllable transitions across different sequence types, even when the sequence types were drawn from the same syllable inventory, was not impossible in the context of the task.

The serial recall task as a window into statistical learning

On a methodological note, this study offers an alternative behavioural method to track SL in real time, with participants' training and testing being critically intertwined. This method builds on the classic Hebbian repetition paradigm (Hebb, 1961; see also Page & Norris, 2009; Smalle et al., 2016; Szmalec et al., 2012), as well as on more recent studies that have used recall tasks to examine learning after a period of exposure to a new artificial language (Isbilen et al., 2018, 2020; Kidd et al., 2020; Majerus et al., 2004). Here, we have shown that recall tasks of this nature, in the absence of an initial exposure phase, can serve as an insightful window onto learning and may be an advantageous method for future studies of SL. Accordingly, we believe that the task can serve as a valuable addition to the toolkit of methods used to study SL. One advantage of the task that we have already discussed is the ability to use it to track learning across the course of an experiment. Another notable benefit of the repetition paradigm is that it is, in the words of Christiansen (2019), a *processing*-based measure of learning. This contrasts with reflection-based measures of SL, such as traditionally used measures of SL like the 2AFC task. The difference between the two is that processing-based tasks require less meta-cognitive effort because, unlike reflection-based measures, they do not ask participants to reflect upon and choose between two possible candidate words. Although 2AFC tasks have their advantages, there are circumstances under which they are not always optimal, including when testing auditory SL in developmental populations and when the aim is to measure individual differences (see Arnon, 2019; Isbilen et al., 2020, 2022; Kidd et al., 2020). Our suggestion is that verbal repetition may be particularly useful in circumstances where researchers are

interested in the course of learning or when reflection-based measures such as 2AFC do not yield reliable results.

With this in mind, one obvious question concerns exactly how verbal recall relates to other measures of SL and to the bigger question of how it relates to the mechanism underlying SL (or the multiple interacting mechanisms underlying it, see Frost et al., 2015). These questions are not mutually exclusive, and we cannot hope to provide a compelling answer to them here. What is clear is that there are many different measures of SL, going from verbal repetition to sequence reproduction (e.g., Conway et al., 2010) to 2AFC following familiarisation (e.g., Saffran, Newport, et al., 1996) to reaction times to structured sequences, as in the serial reaction time task (Nissen & Bullemer, 1987). It is interesting that, although all the measures capture learning of probabilistic distributions and thus are billed as measures of SL, performances on these tasks are often unrelated (see e.g., Siegelman & Frost, 2015). There are likely to be many reasons for this. One obvious methodological reason is that any mode of measurement is an imperfect way of tapping a psychological concept, and so any one task will have non-overlapping measurement error that it does not share with other tasks. More interestingly, the processes underlying SL have been argued to be complex and multi-componential (Arciuli, 2017; Frost et al., 2015), and thus different tasks may differentially implicate different components. This lack of understanding of these individual components limits the understanding of the mechanism(s) underlying SL.

What we see as the value of verbal sequence repetition is in its potential for elucidating the role of SL in language learning. Repetition has had a long history of use in the verbal learning literature beginning with Hebb (1961) and has also been used to measure linguistic proficiency. For instance, non-word repetition is highly sensitive to speakers' distributional knowledge of their language (e.g., Jones et al., 2007, 2014; Szewczyk et al., 2018), and sentence repetition reliably taps grammatical parsing procedures underlying sentence production

and comprehension (Acheson & MacDonald, 2009; Potter & Lombardi, 1990). Thus, verbal sequence repetition appears to be a relatively direct way of observing both (i) existing knowledge and, as we have shown here, (ii) how that knowledge may result in different learning trajectories across time. Studying verbal repetition in a learning paradigm, as we have done in our study (see also Isbilen et al., 2020), is one way to study the dynamics of SL across time (see also Batterink, 2020; Batterink & Paller, 2017).

Limitations and future directions

Our results, alongside those of Elazar et al. (2022), reveal positive evidence in favour of the argument that humans identify frequently occurring linguistic units and encode them as long-term memory representations that are subsequently used for future learning. A key promise of this effect is that it captures what is assumed to be the output of SL; participants are better at learning naturalistic distributions because they have prior experience with them, distributions that they have presumably discovered via SL. However, as with most laboratory-based studies of SL, we have only tested the learning of simple statistical computations. How this scales up to the acquisition of language proper, with all of its complexities, is unclear. Domain-general processes like chunking no doubt play an important role in acquisition and in processing (e.g., Bannard & Matthews, 2008; Christiansen & Chater, 2016; Jones et al., 2021; Lieven et al., 1997). Indeed, Isbilen et al. (2022) have recently shown that adults' chunking of syllables in verbal repetition is related to their recall of highly frequent sequences of words, suggesting a partially shared basis for learning and processing across the different linguistic levels. However, it is important to be mindful of the limits of such effects as they relate to the entirety of language. In particular, because studies of SL typically limit themselves to formal aspects of language (i.e., relationships between linguistic elements devoid of meaning), how a process like SL works within the maelstrom of natural language and how it works in

concert with other key learning mechanisms is still very much an open question and thus a matter for future research.

Conclusion

To conclude, in this study, we demonstrated that adult participants' prior knowledge of TPs derived from their native language forms a robust foundation upon which subsequent learning and processing occur. Our data thus lend further support to the notion that prior knowledge can have a critical impact on future learning (Bertels et al., 2015; Dal Ben et al., 2021; Ebbinghaus, 1885, 1913; Finn & Hudson Kam, 2008; Lew-Williams et al., 2011; Lew-Williams & Saffran, 2012; Mersad & Nazzi, 2011; Onnis & Thiessen, 2013; Potter et al., 2017), providing further evidence that laboratory-based learning is shaped by the (mis/)alignment between the properties of the input and participants' prior expectations. Implementing a sequence repetition task in the absence of a familiarisation phase provided a rich real-time behavioural assessment of SL (though see e.g., Batterink, 2020; Batterink & Paller, 2017, for related online assessments using EEG). We suggest that dynamic speech-production measures may serve as a useful vehicle for further exploring the nature and time course of SL in future research.

Acknowledgements

Chapter 4

The effect of children's prior knowledge and language abilities on their statistical learning

Abstract[10]

Statistical learning (SL) is assumed to lead to long-term memory representations. However, the way that those representations influence future learning remains largely unknown. We studied how children's existing distributional linguistic knowledge influences their subsequent SL on a serial recall task, in which 49 German-speaking seven- to nine-year-old children repeated a series of six-syllable sequences. These contained either (i) disyllabic words based on frequently occurring German syllable transitions (naturalistic sequences), (ii) disyllabic words created from unattested syllable transitions (non-naturalistic sequences), or (iii) random syllable combinations (unstructured foils). Children demonstrated learning from naturalistic sequences from the beginning of the experiment, indicating that their implicit memory traces derived from their input language informed learning from the very early stages onward. Exploratory analyses indicated that children with a higher language proficiency were more accurate in repeating the sequences and improved most throughout the study compared to children with lower proficiency.

Introduction

Statistical learning (SL), the ability to use probabilistic co-occurrence to group elements present in the environment, has been argued to support development across multiple sensory and cognitive domains. One domain in which SL has been identified as playing a role is language development (Romberg & Saffran, 2010; Saffran, Aslin, et al., 1996), where it has been linked to learning across the sub-domains of phonology (e.g., Kuhl, 2004), segmentation and word learning (e.g., Hay et al., 2011; Lany & Saffran, 2010), and grammar (Kidd & Arciuli, 2016). A common assumption of research in this field is that SL results in detailed long-term memory representations for learned stimuli; however, how these representations endure and how they influence children's performance in experiments is largely unknown. In the current chapter, we explore this issue by asking how children's knowledge of their native language influences their performance on an auditory SL task, and whether variation in performance is linked to their language proficiency.

There is significant evidence in favour of the argument that participants' prior linguistic experience guides their SL of linguistic stimuli. Notably, many studies have shown that adult participants' segmentation of words and grammatical patterns (e.g., non-adjacent dependencies) from running speech is constrained both by language-specific phonotactic constraints (i.e., rules) and, within the class of legal sequences, phonotactic probabilities within and across languages (e.g., Bonatti et al., 2005; Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Onnis et al., 2005; Trecca et al., 2019). Although infants and children have much less experience with language, their prior knowledge also influences SL. For instance, Lew-Williams and Saffran (2012) found that nine- and ten-month-old infants' statistical speech segmentation was guided by their expectations about word length, which were forged through pre-exposure to words that were either consistent or inconsistent in length with items in the to-be-segmented stream. These findings are echoed in

work by Thiessen et al. (2019), who demonstrated that 13-month-old, but not seven-month-old, native English-learning infants showed a preference for attending to backwards versus forwards transitional probabilities (TPs) in an ambiguous artificial language. That is, older infants attended to the direction which aligned with the structural properties of their native language (i.e., the head-initial nature of English, which favours backwards processing; see Onnis & Thiessen, 2013, for related findings with adults). Taken together, these data provide converging evidence that learners' expectations about linguistic input emerge over experience with language and can permeate into the laboratory to shape participants' performance on SL tasks.

In the current study, we ask a related though crucially different question: do children draw upon their distributional knowledge of syllable transitions when processing new statistically defined linguistic input? Specifically, does the presence of attested and therefore highly probable bigrams from participants' native language boost SL? Attention to TPs has featured prominently in explanations of SL, since this is how higher units of linguistic organisation (i.e., "words" in studies of segmentation) have been defined (Saffran, Aslin, et al., 1996). Successful learning on these tasks indicates that participants either track TPs or chunk co-occurring syllables into bigger units (Aslin et al., 1998; Perruchet, 2019; Perruchet & Vinter, 1998). The underlying assumption that the output of SL – the acquisition of statistical structure as defined by TPs – leads to long-term representations that aid future learning is less frequently tested. Accordingly, we ask whether participants use these prior expectations about TPs in their ambient language, which they have presumably acquired across development, to process new input.

There is growing evidence from adults in support of this proposal. In a series of studies investigating the interrelationships between visual SL and auditory SL for linguistic and non-linguistic stimuli, Siegelman et al. (2018) reported that, while visual SL and auditory SL for non-

linguistic stimuli were related (suggesting a common underlying capacity for performance across modalities), auditory SL for linguistic stimuli was not related to the other two. Indeed, they reported poor internal consistency for their linguistic auditory SL task, suggesting that participants showed little overlap in performance across individual test items within the task. However, Siegelman et al. found that performance on the linguistic SL task was predicted by independent ratings of "wordness" for the test items. That is, the more closely a test word resembled a real word in the participants' native language (Hebrew), the more likely they would be to segment it from the speech stream and recognise it as a word at test, suggesting that participants entered the experiment with entrenched linguistic knowledge affecting their performance on the SL tasks based on linguistic items.

Two more recent studies with adult populations provide more direct evidence for this entrenchment effect. In a between-participants design, Elazar et al. (2022) tested Spanish-speaking participants on an auditory SL task using one of two speech streams. In a "Spanish-like" condition, participants listened to a stream of continuous speech containing test words defined by TPs that were highly attested in Spanish corpora. In a "Spanish-unlike" condition, the test words were defined by TPs that were rarely attested. Overall, the authors reported a pattern of results that was suggestive of a learning advantage for Spanish-like stimuli, thus providing evidence in favour of the argument that participants build expectations of likely syllable transitions, which they use as priors when parsing new input.

In a study that we build upon in the current chapter (cf. Chapter 3), Stärk et al. (2023) investigated whether native German-speaking adults drew on their prior knowledge of German syllable co-occurrences (i.e., syllable pairs with high TPs) to acquire new but distributionally consistent linguistic input. Unlike Elazar et al. (2022), the authors tested the influence of attested TPs in a within-participants design, a more stringent test of the entrenchment hypothesis since any across-condition differences could not be attributed to differences across

participants. Doing so required a different method of measuring SL. Accordingly, participants heard and repeated three different kinds of sequences in a serial recall task: two of which were structured, while the other sequences were unstructured foils, with the dependent measure being recall accuracy. The structured sequences contained disyllabic experimental words that were either based on likely German syllable transitions (naturalistic sequences) or were devoid of attested transitions (non-naturalistic sequences). The unstructured sequences were scrambled combinations of syllables, serving as baseline. Findings indicated that adults drew on their knowledge of German syllable transitions during the task and did so from the early phases of the experiment onward, showing higher recall accuracy and faster improvement for the naturalistic sequences over the other two sequence types. The participants also performed significantly better on recall of the non-naturalistic sequences in comparison to foils, demonstrating that they were able to acquire knowledge of two sequences within the context of the task, despite the fact that there was likely a degree of interference across conditions because all sequences were constructed from the same syllable inventory.

Overall, the work of Elazar et al. (2022) and Stärk et al. (2023) provides converging evidence in support of the hypothesis that adults draw on their prior linguistic knowledge to process and learn subsequent input, which raises the question of whether children differ from adults in their weighing of prior knowledge and the input statistics, given their ability to pick up on new statistics present in a SL task. This may or may not differ from adults, depending on the paradigm (see Raviv & Arnon, 2018; Shufaniya & Arnon, 2018). Notable differences between adults and children have been found in studies of verbal learning. For instance, Smalle et al. (2017) showed that nine- to ten-year-old children require substantially less exposure to implicitly acquire phonotactic restrictions on novel words than adults do. Additionally, Smalle et al. (2018) have reported that eight- to nine-year-old children retain implicitly learned phonological sequences better than adults, as demonstrated through Hebbian Repetition

Learning. This raises the possibility that, while still within the sensitive period for language acquisition, children differ from adults. Consequently, we investigate whether children draw on their prior distributional knowledge when processing subsequent linguistic input, as previously reported in adults (Elazar et al., 2022; Stärk et al., 2023).

Furthermore, we address the issue concerning whether and how performance in laboratory-based SL tasks relates to children's real-world language development. Typically developing children's language proficiency (as measured with assessments of their expressive and receptive vocabulary) has been found to be correlated with their SL ability (e.g., Evans et al., 2009; Frost, Jessop, et al., 2020; Kidd & Arciuli, 2016; Lany, 2014; though the relationship is not always observed and appears to be task-dependent, see West et al., 2021). Here we take a slightly different approach and ask: do individual differences in language proficiency result in different learning trajectories throughout the course of SL (e.g., as measured over the course of our experiment)? If SL constitutes an individual ability that supports natural language learning (Siegelman et al., 2017), we should expect children of different language abilities to be differentially sensitive to statistically defined sequences, under the assumption that superior SL abilities (and its component processes, Arciuli, 2017) have supported acquisition throughout development.

Thus, in the present study, we examined the effect of prior distributional knowledge on the learning of statistically defined linguistic input in seven- to nine-year-old German-speaking children. We utilised the repetition paradigm from Chapter 3 (Stärk et al., 2023) as described above, adjusting it for children's lower working memory capacity by shortening the sequence length. We hypothesised that children would recall naturalistic sequences better than non-naturalistic sequences and show faster improvement for the naturalistic than non-naturalistic sequences, similar to the findings with adults (Elazar et al., 2022; Stärk et al., 2023). Additionally, we investigated whether children's language proficiency was related to their SL. While past

studies have reported correlational analyses between SL performance and language proficiency, including studies utilising sequence repetition (e.g., Smalle et al., 2018), we were interested in how differences in language proficiency influenced the dynamics of learning throughout the task. Since no similar study exists, this analysis was exploratory.

Method

This study was preregistered on AsPredicted: https://aspredicted. org/546hk.pdf. All of our materials, data, analyses, and results are openly available on the website of the Open Science Framework (OSF): https://osf.io/t5qf4/.

Participants

Forty-nine seven- to nine-year-old native German-speaking children (29 female, 20 male; mean age = 8;7 years; months, SD = 0;6, range: 7;6–9;11) without any known hearing, speech, or language disorders were included in the final sample (N = 49). We had planned to test 60 children, but recruitment and testing were hindered by the COVID-19 pandemic. Despite not being able to recruit our originally planned sample, our power analysis suggests that a sample of 49 is sufficient to detect an effect size of 0.2 syllables recall difference between naturalistic and non-naturalistic sequences as well as between non-naturalistic and foil sequences (i.e., an effect size half the size of the one found in adults, compensating for children's cognitive development). The analysis was conducted in R 4.1.2 (R Core Team, 2022) using the package *simr* 1.0.6 (Green & MacLeod, 2016) and relied upon data from Chapter 3 (Stärk et al., 2023) of which we adjusted the design for the current study (see the Analysis folder on OSF for a detailed description of this and two post hoc power analyses).

Participants were recruited from a German primary school in Leipzig, Germany. Invitations for participation were sent to the parents of all second- and third-graders (with German second-graders being 7–

8 years old and third-graders being 8–9 years old). Of these, 55 consented to participating in our experiment. Data for six children were excluded from analyses: two did not fulfil the inclusion criteria regarding language proficiency (see Design section for more details), two found the serial recall task (i.e., our SL measure) too difficult to complete, and two were excluded due to technical failure. The study was approved by the Ethical Committee of the Faculty of Social Sciences, Radboud University Nijmegen, and was carried out in accordance with the World Medical Association Declaration of Helsinki. Children were free to withdraw at any time. They received a certificate and stickers for their participation.

Design

The experiment utilised a serial recall task, in which participants were presented with sequences of six syllables and repeated them out loud (see Figure 4.1). The task is based on studies of the Hebb repetition effect (Hebb, 1961; Page & Norris, 2009), and was modified from Chapter 3's related study with adults (Stärk et al., 2023) to be suitable for children. Specifically, we reduced the length of the sequences (sequences comprised 6 syllables, rather than 8), thereby making allowances for the fact that working memory (Cowan, 2016) and other cognitive processes supporting performance (e.g., processing speed, Kail & Salthouse, 1994) improve across childhood. The study had a within-participants design, in which all participants received all three sequence types: naturalistic, non-naturalistic, and unstructured foils. The naturalistic and non-naturalistic sequences were structured, with each containing three disyllabic experimental words, whereas unstructured sequences comprised the same syllables as the structured sequences but in a scrambled order (see the Materials section for further details). Participants' repetitions of the sequences were scored for accuracy, and we examined performance across the task to gain insights into learning over the course of exposure.
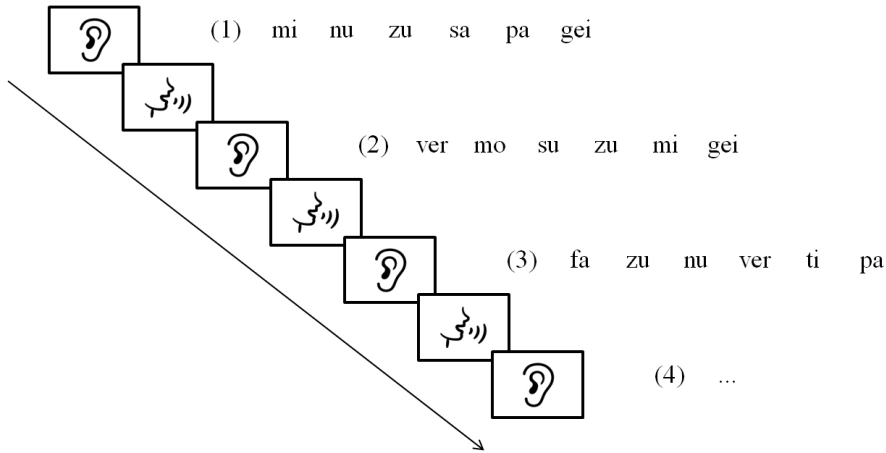
(1) mi nu zu sa pa gei

(2) ver mo su zu mi gei

(3) fa zu nu ver ti pa

(4) ...

Figure 4.1. *Three example experimental sequences. On each trial, participants listened to a six-syllable sequence and then repeated it. (1) = one naturalistic sequence; (2) = one unstructured foil sequence; (3) = one non-naturalistic sequence. Design adapted from Chapter 3 (Stärk et al., 2023).*

Materials

Language proficiency assessments

We assessed the children's language proficiency to ensure that both their lexical and morphosyntactic knowledge was representative of the average German-speaking seven- to nine-year-old. To measure their lexical knowledge, we used the German Peabody Picture Vocabulary Test version IV (PPVT; Lenhard et al., 2015). This is a test of receptive vocabulary, in which children are shown four pictures per trial and have to identify the picture that best matches a word they were given. Their correct responses are counted and converted into an age-dependent score. Children whose performance was 1.5 SDs below the average of their age group were excluded from the main analyses.

To measure the children's morphosyntactic knowledge, we used the German LITMUS sentence repetition task (SRT; Abed Ibrahim et al., 2018; Hamann et al., 2013; Hamann & Abed Ibrahim, 2017). This test consists of 45 sentences with varying degrees of difficulty, which children listen to and repeat. Their correct responses are counted and form their score. If performance was 2.5 SDs below the average estimated from our own sample, children were excluded from the main analyses. A more lenient cut-off for the SRT was chosen because the test is not standardised.

Experimental stimuli

For the serial recall task (i.e., our SL measure), participants repeated sequences of six syllables. There were three different sequence types presented pseudo-randomly throughout the exposure: (i) naturalistic sequences, (ii) non-naturalistic sequences, and (iii) unstructured foil sequences. The first two sequence types were structured, meaning that they contained experimental words which participants could segment, while the unstructured foil sequences did not contain any learnable patterns. Importantly, the words contained in the naturalistic sequences differed from the words contained in the non-naturalistic sequences, in that the naturalistic words comprised frequently co-occurring German syllable pairs while the non-naturalistic words comprised unattested syllable pairs. That is, neither sequence type contained natural German words but the naturalistic words might appear familiar to German speakers because of their high TPs in natural German. They represent the outcome of participants' SL in natural German. In the current study, we investigate whether seven- to nine-year-old children draw on this prior knowledge (i.e., the outcome of previous SL in natural language) to process a new experimental language (as modelled by the six naturalistic words). This is compared to participants' SL in a language without prior knowledge (as modelled by the six non-naturalistic words).

The experimental words were created via a corpus analysis, which analysed the TPs of the 1000 most frequent words of German child-directed speech on the CHILDES database (MacWhinney, 2000)[11]. Six syllable pairs (i.e., bigrams) occurring with high within-word backwards TPs were selected for use as the naturalistic words (naturalistic words: *gefa, minu, moti, pagei, versu, zusa*; see Chapter 2 or Stärk et al., 2022, for evidence that backwards TPs are more reliable cues to 'wordness' than forwards TPs in German speech). These were in turn used to create the naturalistic sequences. Importantly, the experimental naturalistic words did not comprise existing German words, but the high TPs (TP > .20, with an average TP = .69 and a range of .21–1) between the syllables within each pair make them potentially familiar to German speakers. There was no repetition of syllables across the words. The non-naturalistic words were created by switching the first and second syllables of the naturalistic words and combining each final syllable with a different first syllable, such that the syllable pairs neither form German words nor have high TPs (TP = 0) in natural German (non-naturalistic words: *fazu, geimi, nuver, samo, suge, tipa*). Unstructured foil sequences consisted of the same 12 syllables presented in a scrambled order, such that these sequences contained neither learnable patterns nor German or experimental words.

Within the context of the experiment, both types of structured sequences had perfect within-word TPs (*structured sequences:* within-word TPs = 1, between-word TPs ≤ .25), while the TPs between syllables in the unstructured foil sequences were generally low (*unstructured sequences:* TPs ≤ .17). Since participants repeated all three sequence types, with syllables being repeated across types, within-word TPs for both structured sequences were TP = .33, and TPs for all other syllable pairs were TPs ≤ .14, calculated over the entire exposure. Forwards and backwards TPs were equal within the

---

[11] We included the following corpora from the CHILDES database (MacWhinney, 2000) in our analysis: Caroline (Von Stutterheim, 2010), Grimm (Grimm, 2006, 2007), Leo (Behrens, 2006), Manuela (Wagner, 2006), Miller (Miller, 1979), Rigol (Rigol, 2007), Stuttgart (Lintfert, 2010), TAKI (Lintfert, 2010), and Wagner (Wagner, 1974, 1985).

experimental sequences (see Stimuli file on OSF for the precise numbers and the details of the stimuli creation).

Each syllable was recorded in isolation by a female native speaker of German. The recordings were adjusted using Audacity (Audacity Team, 2018) to ensure uniformity of length, resulting in an average syllable duration of 377ms (range: 352ms–416ms). The experiment comprised 72 sequences in total, 24 of each sequence type. Each sequence consisted of six syllables (i.e., three experimental words in the structured sequences), separated by 500ms of silence. Thus, while we refer to our structured sequences as containing "words," it is important to remember that to the participant, the stimuli were lists of syllables; thus, any grouping of the syllables based on TPs, be it from existing naturalistic knowledge or knowledge gained through the course of the task, is evidence of SL, which in this case can be interpreted as the chunking of co-occurring syllables. The final syllable of each sequence was followed by a beep, which indicated that participants could start their repetition.

The experiment was divided into 12 blocks of six sequences, which contained two items of each sequence type. Within each block, sequences were presented pseudo-randomly, with no direct repetition of a particular sequence type (e.g., there was no adjacent presentation of naturalistic sequences). There were four experimental lists, which differed in the order of blocks (and with which sequence type the experiment started). Participants were randomly assigned to one list. Across the entire experiment, each word occurred 12 times in total, appearing equally often in each position within a sequence (for more information on the stimuli and their creation see the Materials folder of the project on OSF).

Procedure

Children whose parents signed the consent form were invited for two experimental sessions at their after-school club. The children were tested over two sessions of approximately 30 minutes. In the first

session, children completed the German PPVT (Lenhard et al., 2015) and the SRT (Abed Ibrahim et al., 2018; Hamann et al., 2013; Hamann & Abed Ibrahim, 2017). In the second session, they completed the serial recall task (i.e., our main task, measuring children's SL). Both sessions took place in a quiet, private room at the school. Stimuli were played over closed-cup headphones via a laptop, and the children repeated the sequences into a microphone, with repetitions being recorded for subsequent offline coding. The serial recall task was conducted using the software *Presentation* (Neurobehavioral Systems, 2014).

If the children fulfilled the language proficiency inclusion criteria in the first session, they were invited for the second session on the following day. The serial recall task was introduced as a game, which had the aim of helping an alien to repair their broken spaceship and fly back to their home planet. The children were told that they could help the alien by repeating what it says. First, they received six unstructured practice trials: three four-syllable sequences and three six-syllable sequences, comprising a different set of syllables than the experimental sequences (*ba, fun, gi, re, se, to*), before receiving the 72 experimental sequences. The children received a sticker and were told more components of the alien story after the practice trials, during the two breaks (after Sequences 24 and 48), and after completing the experiment. At the end of the session, the children were debriefed and received a certificate.

Data preparation

Language proficiency assessments

The PPVT was coded online by the experimenter, while the SRT was transcribed and coded offline upon completion by the experimenter and a trained assistant. The SRT data were coded for identical repetition following Hamann and Abed Ibrahim (2017). That is, children received a point for each sentence repeated entirely verbatim. Five participants' SRT recordings (10%) were transcribed by both coders for inter-transcriber reliability analyses, which revealed a strong reliability

between the two transcribers (observed agreement = 94.7%; κ = 0.83 with 95% CIs of [0.74, 0.93]; following the more conservative interpretation of the kappa statistic suggested by McHugh, 2012). Furthermore, Cronbach's alpha (α) was calculated for the SRT to determine task-internal reliability, which revealed good internal consistency (α = .81 [.74, .88]). This justifies its use as individual differences measure in our exploratory analysis (see OSF for almost identical values of McDonald's omega).

Serial recall task

Sequence repetitions were transcribed by the experimenter and a trained assistant. Participants' repetitions were scored at both the syllable and the bigram level, with a bigram being an experimental word in the two structured conditions and random syllable pairs in the unstructured condition (therefore referred to as "bigram level" rather than "word level" here). At the syllable level, participants received one point for each syllable repeated correctly in the correct position (max. 6 points per sequence). At the bigram level, participants received one point for each bigram (i.e., "word" or syllable pair) repeated correctly in the correct position (max. 3 points per sequence), providing valuable information about whether participants recalled sequences better because of learning the experimental words (i.e., bigrams) in the structured sequence types. In the unstructured sequences, the three bigrams per sequence were the random syllable pairs in Syllable positions 1 and 2, 3 and 4, and 5 and 6 (with syllable combinations varying between sequences).

Recall scores at the syllable and bigram level were highly correlated for all three sequence types (*naturalistic sequences: r(47) = .96, p < .001; non-naturalistic sequences: r(47) = .94, p < .001; unstructured foil sequences: r(47) = .88, p < .001*). Data for five participants were transcribed by both coders for inter-transcriber reliability analyses, revealing a moderate to strong reliability between the two transcribers (*syllable level:* observed agreement = 79.2%; κ = 0.74 [0.69, 0.79];

Table 4.1. *Task-internal consistency measures for the serial recall task.*

	Syllable level		Bigram level	
Sequence type	α	95% CI	α	95% CI
Overall	.95	[.93, .97]	.93	[.90, .96]
Naturalistic sequences	.85	[.80, .91]	.81	[.74, .89]
Non-naturalistic sequences	.89	[.84, .93]	.87	[.82, .92]
Unstructured foil sequences	.85	[.79, .91]	.77	[.67, .86]

Note: α = Cronbach's alpha.

bigram level: observed agreement = 88.3%; $\kappa = 0.80$ [0.74, 0.86]; again following the more conservative interpretation of kappa suggested by McHugh, 2012). Furthermore, Cronbach's alpha (α) was calculated for the serial recall task as measure of task-internal reliability, revealing acceptable to excellent internal consistency (see Table 4.1, and see OSF for almost identical values of McDonald's omega).

We preregistered two different coding schemes for the serial recall task, the one described above and a serial-order coding scheme, which grants participants points more generously for repeating a syllable or bigram in the correct serial order rather than in the exact position within the sequence (see Isbilen et al., 2020, and Kidd et al., 2020). We report the results of the former coding scheme only, since the strict scheme is more conservative. However, the results of all coding schemes can be found in the Analysis folder of our OSF project. The results do not differ substantially between the different coding schemes.

Results

All of the analyses presented in this chapter were performed in R 4.1.3 (R Core Team, 2022) using RStudio (RStudio Team, 2022). Data preprocessing and visualisation were performed using the package *tidyverse* 1.3.1 (Wickham, 2017; Wickham et al., 2019).

Language proficiency assessments

On the PPVT, the children scored on average 54.63 (SD = 9.07, range: 38–73), which is slightly higher than the normed average of 50. One child performed below our inclusion threshold of 1.5 SDs below the norming average (PPVT = 35) and was thus excluded from the analyses. On the SRT, the children scored on average 34.89 (SD = 5.36, range: 23–44). One child scored 2.2 SDs below the group average on the standard scoring criteria, but below the inclusion threshold of 2.5 SDs on more sensitive coding schemes (see preregistration, coding scheme, and analysis files on OSF). This participant was excluded from the analyses.

Serial recall task

Analysis by experimental block

We analysed the data using generalised linear mixed-effects models (package *lmerTest* 3.1-3; Kuznetsova et al., 2017; based on *lme4* 1.1-28; Bates et al., 2015), with syllable recall and bigram recall as the dependent variables to test overall recall and recall of the experimental words, respectively. We specified a Poisson distribution with a log-link, since the dependent measures were count variables. Models were fitted with a fixed effect of sequence type using sliding contrasts (*naturalistic:* 0.5 vs. *non-naturalistic:* −0.5, and *non-naturalistic:* 0.5 vs. *foils:* −0.5) to examine whether recall differed across the experimental conditions, and a fixed effect of block was entered as a centred continuous variable to examine learning over the course of the experiment. Additionally, the interaction between the two factors was included as a fixed effect. We fitted the maximal model supported by the data (Barr et al., 2013), controlling for participants and items as random intercepts, with sequence type and block (as well as their interaction) as random slopes for participants, but not items as those differed between sequence types and blocks. The marginal and conditional R^2 effect sizes are also reported as goodness-of-fit estimates. These denote the proportion of the variance explained by the model both with

Table 4.2. *Summary of the linear mixed-effects models investigating the influence of sequence type and block on the children's syllable and bigram recall.*

Parameter	b	95% CI	SE	t	p
Syllable level					
(Intercept)	0.29	[0.14, 0.44]	0.08	3.74	< .001
Nat. vs. Non-nat.	0.10	[0.04, 0.17]	0.03	3.21	.001
Non-nat. vs. Foils	0.05	[–0.01, 0.10]	0.03	1.66	.10
Block	0.01	[–0.04, 0.05]	0.02	0.25	.80
Nat. vs. Non-nat. × Block	0.02	[–0.01, 0.06]	0.02	1.17	.24
Non-nat. vs. Foils × Block	0.01	[–0.03, 0.05]	0.02	0.49	.63
Bigram level					
(Intercept)	–1.15	[–1.39, –0.91]	0.13	–9.17	< .001
Nat. vs. Non-nat.	0.24	[0.11, 0.36]	0.06	4.07	< .001
Non-nat. vs. Foils	0.14	[0.04, 0.23]	0.05	2.90	.004
Block	0.01	[–0.06, 0.09]	0.04	0.37	.71
Nat. vs. Non-nat. × Block	0.03	[–0.04, 0.09]	0.03	0.82	.42
Non-nat. vs. Foils × Block	0.02	[–0.05, 0.09]	0.04	0.65	.52

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Model fit syllable level: AIC = 10,961; BIC = 11,023; $R^2_{marginal}$ = 0.027; $R^2_{conditional}$ = 0.088; ICC = 0.063; RMSE = 1.248; σ = 1; model fit bigram level: AIC = 5661; BIC = 5742; $R^2_{marginal}$ = 0.046; $R^2_{conditional}$ = 0.415; ICC = 0.387; RMSE = 0.599; σ = 1.

(conditional R^2) and without (marginal R^2) controls for sources of random variance (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013).

There was a significant main effect of sequence type at both the syllable level and the bigram level, with participants displaying better recall for naturalistic than non-naturalistic sequences (see Table 4.2). At the bigram level, participants also showed better recall for the non-naturalistic than the unstructured foil sequences, indicating that the non-naturalistic words were also learnt, even though this did not facilitate the overall syllable recall for those sequences (for a visualisation of participants' syllable and bigram recall accuracy, see Figure 4.2 and Figure 4.3, respectively). There was no significant main effect of block, and no significant interaction between block and sequence type, suggesting that the conditional differences across
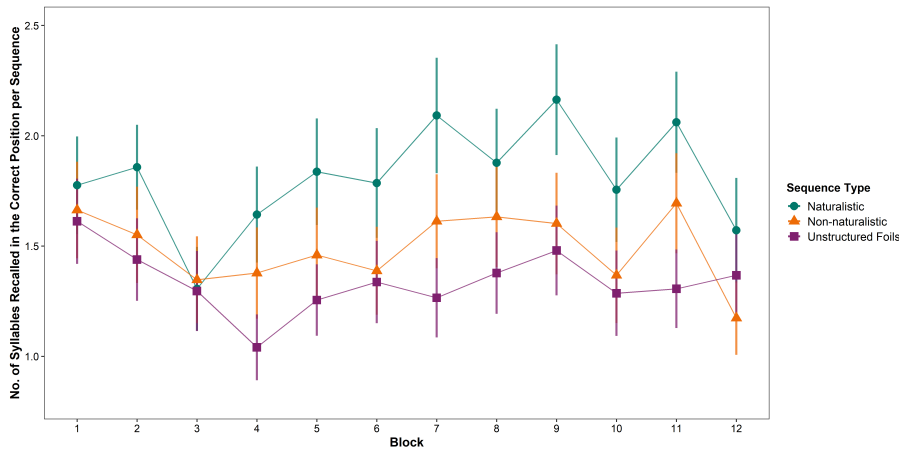
Figure 4.2. *Mean recall of syllables (out of six per sequence) for the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.*

Figure 4.3. *Mean recall of bigrams (out of three per sequence) in the three sequence types given by experimental Blocks 1–12. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. Error bars indicate ±1 standard error.*
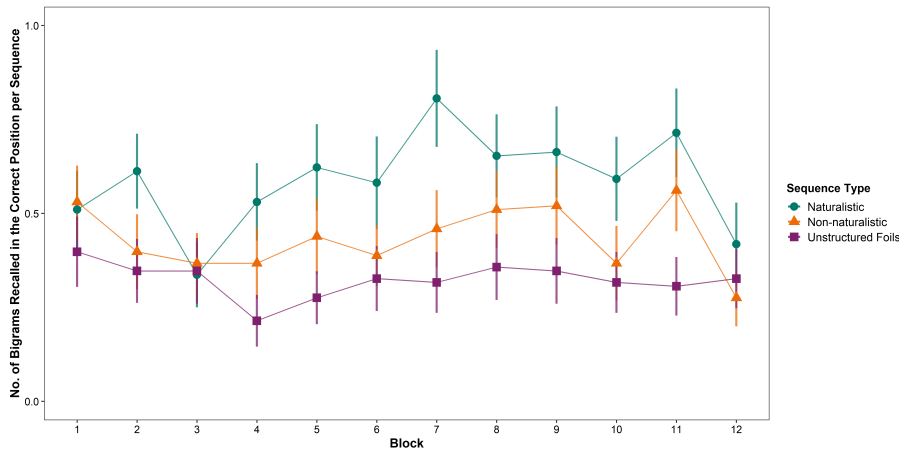
sequence types reflected the overall performance of children across the experiment.

Analysis by exposure phase

While there were no effects of block in the previous analyses, performance across the experiment was not totally even. Chapter 3 found that adult performance in the current task varied across phases of the experiment (early, intermediate, and late; Stärk et al., 2023). We conducted the same analysis, dividing the exposure into three phases, separated by breaks during the data acquisition (*early exposure:* Blocks 1–4; *intermediate exposure:* Blocks 5–8; *late exposure:* Blocks 9–12). Accordingly, we added exposure phase instead of block as a fixed effect in the models examining children's syllable and bigram recall. We fitted the maximal model supported by the data (Barr et al., 2013) with sequence type (sliding contrast: *naturalistic:* 0.5 vs. *non-naturalistic:* −0.5, and *non-naturalistic:* 0.5 vs. *foils:* −0.5), exposure phase (sliding contrast: *early exposure:* −0.5 vs. *intermediate exposure:* 0.5, and *intermediate exposure:* −0.5 vs. *late exposure:* 0.5), and their interaction as fixed effects, and random intercepts and slopes for participants and items, where appropriate.

Comparable to the results of the previous analysis, we found significant main effects of sequence type but neither a main effect of exposure phase nor an interaction between the two factors (see Table 4.3). For an illustration of syllable and bigram recall accuracy over the three phases see Figure 4.4 and Figure 4.5.

The relationship between language proficiency and statistical learning

We next report our exploratory analyses investigating whether differences in children's performance on the SL task were related to their native-language proficiency. We obtained two measures of language proficiency – vocabulary size, as measured by the PPVT, and morphosyntactic processing, as measured by the SRT. Table 4.4 shows

Table 4.3. *Summary of the linear mixed-effects models investigating the influence of sequence type and exposure phase on the children's syllable and bigram recall.*

Parameter	b	95% CI	SE	t	p
Syllable level					
(Intercept)	0.29	[0.12, 0.44]	0.08	3.64	< .001
Nat. vs. Non-nat.	0.12	[0.05, 0.19]	0.03	3.60	< .001
Non-nat. vs. Foils	0.05	[–0.02, 0.10]	0.03	1.61	.11
Early vs. Intermediate	–0.01	[–0.06, 0.04]	0.02	–0.52	.61
Intermediate vs. Late	0.03	[–0.02, 0.07]	0.02	1.30	.19
Nat. vs. Non-nat. × Early vs. Intermediate	0.01	[–0.03, 0.05]	0.02	0.41	.68
Nat. vs. Non-nat. × Intermediate vs. Late	0.02	[–0.02, 0.06]	0.02	1.03	.30
Non-nat. vs. Foils × Early vs. Intermediate	0.02	[–0.03, 0.06]	0.02	0.89	.37
Non-nat. vs. Foils × Intermediate vs. Late	–0.02	[–0.07, 0.03]	0.02	–0.79	.43
Bigram level					
(Intercept)	–1.13	[–1.36, –0.88]	0.12	–9.47	< .001
Nat. vs. Non-nat.	0.17	[0.06, 0.28]	0.05	3.16	.002
Non-nat. vs. Foils	0.14	[0.04, 0.24]	0.05	2.92	.003
Early vs. Intermediate	0.05	[–0.03, 0.13]	0.04	1.21	.23
Intermediate vs. Late	0.01	[–0.06, 0.08]	0.04	0.20	.84
Nat. vs. Non-nat. × Early vs. Intermediate	0.03	[–0.05, 0.10]	0.04	0.76	.45
Nat. vs. Non-nat. × Intermediate vs. Late	0.01	[–0.07, 0.08]	0.04	0.21	.84
Non-nat. vs. Foils × Early vs. Intermediate	0.03	[–0.05, 0.12]	0.04	0.67	.50
Non-nat. vs. Foils × Intermediate vs. Late	0.00	[–0.09, 0.09]	0.05	–0.05	.96

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Exposure phase: Early, Intermediate, Late. Model fit syllable level: AIC = 10,962; BIC = 11,085; $R^2_{marginal}$ = 0.024; $R^2_{conditional}$ = 0.394; ICC = 0.379; RMSE = 1.246; σ = 1; model fit bigram level: AIC = 5674; BIC = 5761; $R^2_{marginal}$ = 0.045; $R^2_{conditional}$ = 0.084; ICC = 0.041; RMSE = 0.598; σ = 1.

the bivariate correlations between all variables, in addition to partial correlations between repetition of structured sequences and language proficiency, controlling for foil repetition. All bivariate correlations between repetition of syllable sequences and the SRT and PPVT were
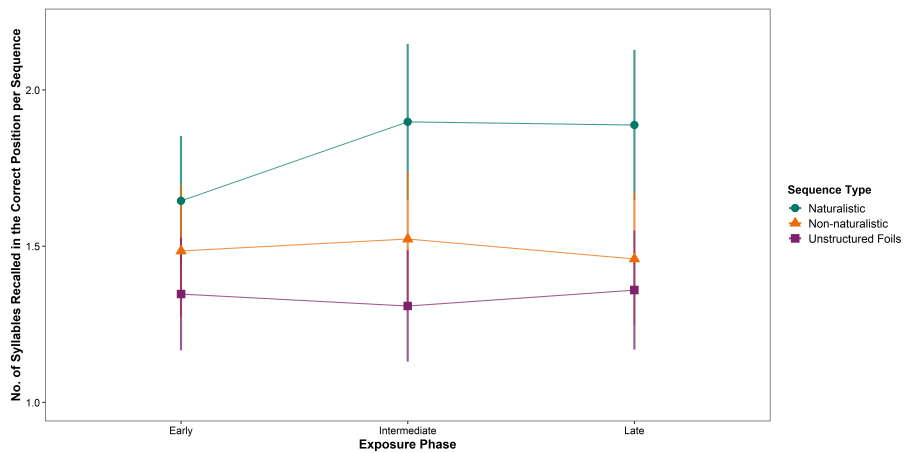
Figure 4.4. *Mean recall of syllables (out of six per sequence) for the three sequence types given by exposure phase. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. The three exposure phases were early, intermediate, and late. Error bars indicate ±1 standard error.*

Figure 4.5. *Mean recall of bigrams (out of three per sequence) for the three sequence types given by exposure phase. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. The three exposure phases were early, intermediate, and late. Error bars indicate ±1 standard error.*
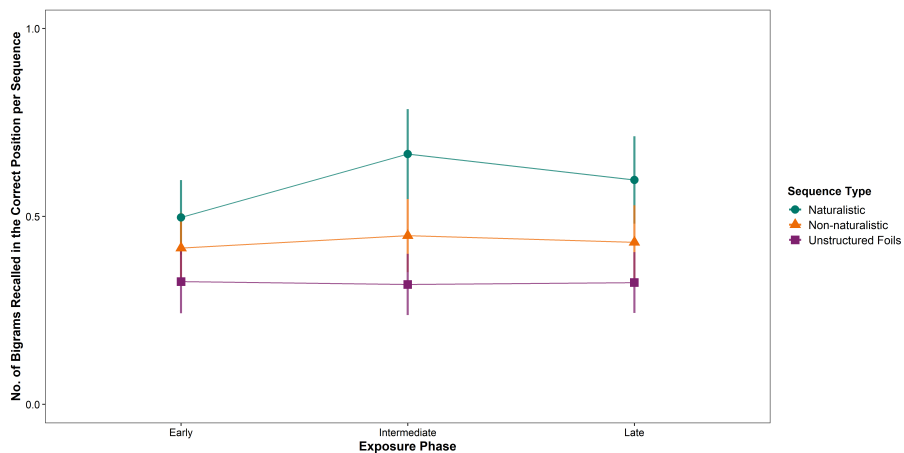
Table 4.4. *Pearson bivariate correlations between children's recall in the three sequence types (naturalistic, non-naturalistic, and foil sequences) and their performance on the language proficiency assessments (SRT and PPVT) at the syllable (left) and bigram level (right). Partial correlations between language assessments and structured sequences, controlling for foil repetition, appear in brackets.*

	Syllable level				Bigram level			
	Non-nat.	Foils	SRT	PPVT	Non-nat.	Foils	SRT	PPVT
Nat.	.74***	.83***	.70*** (.24)	.38** (.07)	.64***	.74***	.62*** (.30*)	.30* (.05)
Non-nat.	–	.85***	.67*** (.13)	.27 (–.16)	–	.81***	.58*** (.17)	.16 (–.23)
Foils		–	.74***	.41**		–	.63***	.36*
SRT			–	.40**			–	.40**
PPVT				–				–

*Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. N = 49; * $p < .05$, ** $p < .01$, *** $p < .001$*

positive and all but two were significant. The partial correlations testing the relationship between syllable recall and language proficiency controlling for foil repetition were lower. Only the partial correlation between naturalistic repetition of bigrams and SRT performance was significant.

While the simple correlations between sequence recall and PPVT or SRT were significant, we chose the latter measure as our individual differences variable because the recall task and the SRT both involve sequencing linguistic units in the same modality, but with different units (syllables versus morphemes). Finding that the two are systematically related would provide evidence in support of the idea that sequencing of syllables and morphemes share an underlying common and statistically sensitive mechanism (Isbilen et al., 2022). Thus, we next analysed whether SRT performance influenced children's performance on the experiment. We fitted the maximal model supported by the data (Barr et al., 2013) with sequence type (sliding contrast: *naturalistic:* 0.5 vs. *non-naturalistic:* −0.5, and *non-naturalistic:* 0.5 vs. *foils:* −0.5),

block (centred continuous variable) and SRT (centred continuous variable), as well as the interactions between the three factors, as fixed effects, and random intercepts and slopes for participants and items, where appropriate. In addition to the main effects of sequence type described above, there was a significant main effect of SRT score (*syllable level: β* = 0.41 [0.32, 0.50], *t* = 8.34, *p* < .001; *bigram level: β* = 0.59 [0.41, 0.76], *t* = 6.55, *p* < .001), with children who scored higher on the SRT displaying better syllable and bigram recall on the experimental task. Furthermore, there was a significant interaction between sequence type and SRT performance (*syllable level: β* = −0.05 [−0.09, 0.00], *t* = −2.04, *p* = .04; *bigram level: β* = −0.11 [−0.20, −0.02], *t* = −2.36, *p* = .02), with a greater recall difference between naturalistic and non-naturalistic sequences for children with lower SRT scores compared to their higher SRT scoring peers. At the syllable level, there was an additional significant interaction between block and SRT performance (*syllable level: β* = 0.05 [0.00, 0.09], *t* = 2.25, *p* = .03), with children with higher SRT scores showing greater improvement in syllable recall throughout the experiment, although the effect was not significant at the bigram level (see analysis file on OSF for full details).

Given that linguistic proficiency was related to performance across the experiment, we scrutinised this further by replacing the variable of block with the more interpretable variable of phase (*early, intermediate, late*). We fitted the maximal model supported by the data (Barr et al., 2013) with sequence type (sliding contrast: *naturalistic:* 0.5 vs. *non-naturalistic:* −0.5, and *non-naturalistic:* 0.5 vs. *foils:* −0.5), exposure phase (sliding contrast: *early exposure:* −0.5 vs. *intermediate exposure:* 0.5, and *intermediate exposure:* −0.5 vs. *late exposure:* 0.5), and SRT performance (centred continuous variable), as well as the interactions between the three factors, as fixed effects, and random intercepts and slopes for participants and items, where appropriate. The outcomes of the model analysing the data at the syllable level are given in Table 4.5, and the outcomes of the model analysing the data at the bigram level are given in Table 4.6. As in the previous models, we found significant main effects of sequence type (with syllable recall only differing

Table 4.5. *Summary of the linear mixed-effects model investigating the influence of sequence type, exposure phase, and SRT score on the children's syllable recall.*

Parameter	b	95% CI	SE	t	p
(Intercept)	0.29	[0.18, 0.39]	0.05	5.46	< .001
Nat. vs. Non-nat.	0.12	[0.05, 0.19]	0.03	3.65	< .001
Non-nat. vs. Foils	0.05	[−0.02, 0.10]	0.03	1.52	.13
Early vs. Intermediate	−0.01	[−0.05, 0.04]	0.02	−0.49	.63
Intermediate vs. Late	0.02	[−0.02, 0.07]	0.02	1.04	.30
SRT	0.41	[0.31, 0.50]	0.05	8.33	< .001
Nat. vs. Non-nat. × Early vs. Intermediate	0.01	[−0.04, 0.05]	0.02	0.21	.83
Nat. vs. Non-nat. × Intermediate vs. Late	0.03	[−0.02, 0.08]	0.02	1.35	.18
Non-nat. vs. Foils × Early vs. Intermediate	0.02	[−0.04, 0.06]	0.02	0.66	.51
Non-nat. vs. Foils × Intermediate vs. Late	0.00	[−0.06, 0.05]	0.03	−0.17	.87
Nat. vs. Non-nat. × SRT	−0.05	[−0.10, 0.00]	0.02	−2.06	.04
Non-nat. vs. Foils × SRT	0.01	[−0.03, 0.05]	0.02	0.39	.69
Early vs. Intermediate × SRT	0.06	[0.02, 0.11]	0.02	2.85	.004
Intermediate vs. Late × SRT	0.00	[−0.04, 0.04]	0.02	−0.05	.96
Nat. vs. Non-nat. × Early vs. Intermediate × SRT	0.01	[−0.04, 0.05]	0.02	0.47	.64
Nat. vs. Non-nat. × Intermediate vs. Late × SRT	−0.03	[−0.07, 0.02]	0.02	−1.23	.22
Non-nat. vs. Foils × Early vs. Intermediate × SRT	0.01	[−0.04, 0.06]	0.02	0.38	.71
Non-nat. vs. Foils × Intermediate vs. Late × SRT	−0.04	[−0.08, 0.01]	0.03	−1.53	.13

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Exposure phase: Early, Intermediate, Late. Model fit: AIC = 10,924; BIC = 11,103; $R^2_{marginal}$ = 0.225; $R^2_{conditional}$ = 0.395; ICC = 0.219; RMSE = 1.246; σ = 1.

between naturalistic and non-naturalistic sequences, and bigram recall additionally differing between non-naturalistic and foil sequences) and SRT performance. The main effect of exposure phase was not significant. At the syllable level, we found two significant interactions. The interaction between SRT performance and sequence type indicated that children with lower SRT scores showed a greater recall difference between naturalistic and non-naturalistic sequences than children with

Table 4.6. *Summary of the linear mixed-effects model investigating the influence of sequence type, exposure phase, and SRT score on the children's bigram recall.*

Parameter	b	95% CI	SE	t	p
(Intercept)	−1.13	[−1.30, −0.94]	0.09	−12.97	< .001
Nat. vs. Non-nat.	0.19	[0.08, 0.29]	0.06	3.49	< .001
Non-nat. vs. Foils	0.14	[0.03, 0.25]	0.05	2.75	.006
Early vs. Intermediate	0.02	[−0.06, 0.10]	0.04	0.43	.67
Intermediate vs. Late	0.02	[−0.06, 0.10]	0.04	0.48	.63
SRT	0.57	[0.39, 0.73]	0.09	6.71	< .001
Nat. vs. Non-nat. × Early vs. Intermediate	0.02	[−0.07, 0.11]	0.04	0.48	.63
Nat. vs. Non-nat. × Intermediate vs. Late	0.03	[−0.06, 0.11]	0.04	0.67	.50
Non-nat. vs. Foils × Early vs. Intermediate	0.05	[−0.04, 0.16]	0.05	1.13	.26
Non-nat. vs. Foils × Intermediate vs. Late	−0.01	[−0.11, 0.10]	0.05	−0.13	.89
Nat. vs. Non-nat. × SRT	−0.08	[−0.17, 0.01]	0.04	−1.84	.07
Non-nat. vs. Foils × SRT	0.00	[−0.08, 0.09]	0.04	0.08	.93
Early vs. Intermediate × SRT	0.09	[0.00, 0.17]	0.04	1.96	.05
Intermediate vs. Late × SRT	−0.04	[−0.11, 0.04]	0.04	−0.89	.37
Nat. vs. Non-nat. × Early vs. Intermediate × SRT	0.03	[−0.06, 0.11]	0.04	0.60	.55
Nat. vs. Non-nat. × Intermediate vs. Late × SRT	−0.06	[−0.14, 0.02]	0.04	−1.37	.17
Non-nat. vs. Foils × Early vs. Intermediate × SRT	−0.06	[−0.16, 0.03]	0.05	−1.28	.20
Non-nat. vs. Foils × Intermediate vs. Late × SRT	0.01	[−0.09, 0.11]	0.05	0.20	.84

Notes: Sequence type: Nat. = naturalistic sequences, Non-nat. = non-naturalistic sequences, Foils = unstructured foil sequences. Exposure phase: Early, Intermediate, Late. Model fit: AIC = 5647; BIC = 5789; $R^2_{marginal}$ = 0.222; $R^2_{conditional}$ = 0.254; ICC = 0.041; RMSE = 0.599; σ = 1.

higher SRT scores. The interaction between SRT performance and exposure phase indicated that children with higher SRT scores showed stronger improvement between the early and intermediate exposure phase than children with lower SRT scores (see Figure 4.6, which plots performance for "High" and "Low" SRT performers, determined by median split).
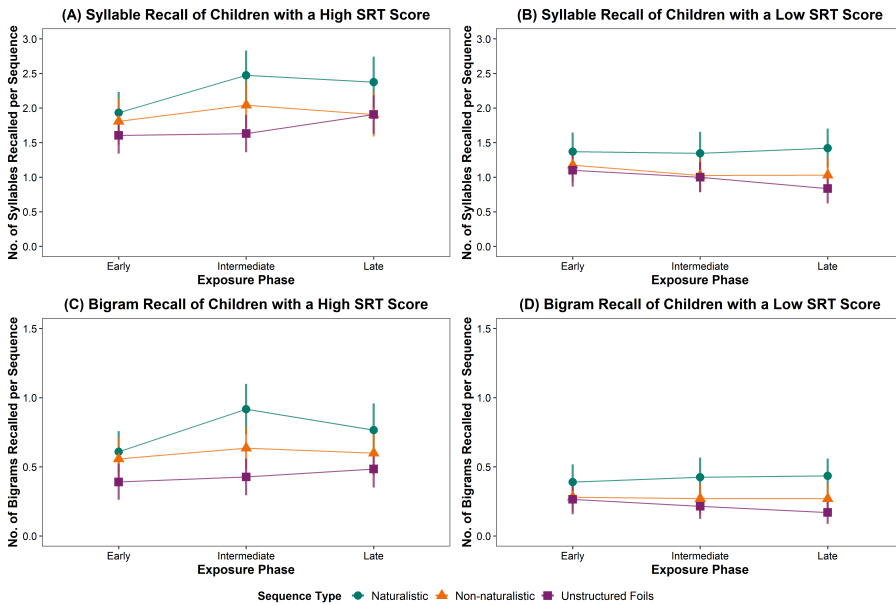
Figure 4.6. *Mean recall of syllables (top) and bigrams (bottom) per sequence of the children with an SRT score higher than the median split (left) and lower than the median split (right) for the three sequence types given by exposure phase. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. The three exposure phases were early, intermediate, and late. Error bars indicate ±1 standard error.*

Children's performance showed a dramatic decrease in the final block, most likely driven by fatigue, which could potentially obscure learning effects across the later part of the experiment. To rule out this possibility, we ran one final exploratory analysis without Block 12, reducing the late exposure phase to Blocks 9 to 11, and fitted the same four models predicting participants' syllable or bigram recall by sequence type, block or exposure phase, and SRT performance (as well as their interactions). Importantly, all previously observed effects remained significant, with most effects becoming even slightly larger than in the original models including all 12 blocks. The two models including the factor block showed the same overall outcome whether Block 12 was included or not. In the model predicting bigram recall by
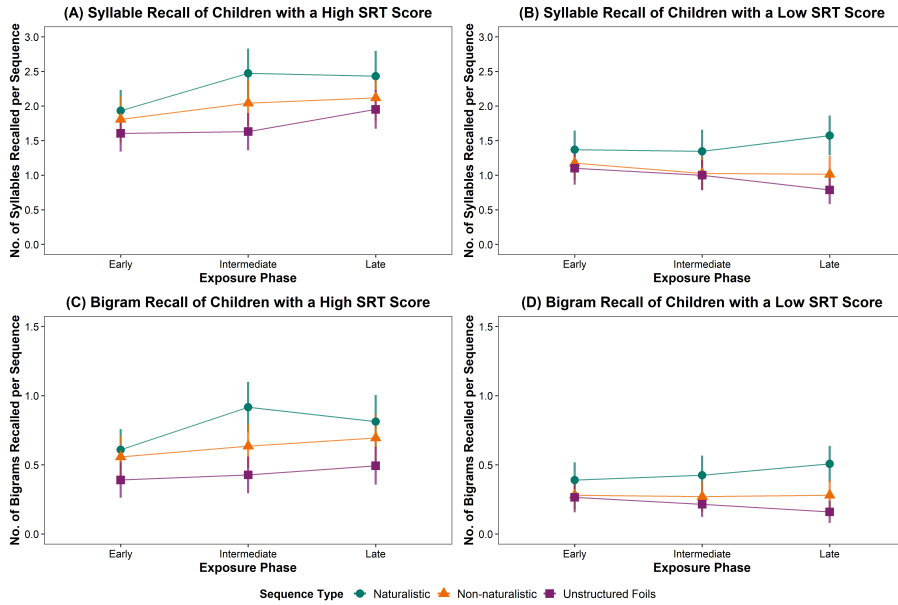
Figure 4.7. *Mean recall of syllables (top) and bigrams (bottom) per sequence of the children with an SRT score higher than the median split (left) and lower than the median split (right) for the three sequence types given by exposure phase when removing the final block. The three sequence types were naturalistic, non-naturalistic, and unstructured foils. The three exposure phases were early, intermediate, and late (with the late exposure phase = Blocks 9–11). Error bars indicate ±1 standard error.*

sequence type, phase, and SRT performance, the two interactions already observed in the other models now also reached the traditional significance threshold. These were the interactions between sequence type and SRT performance, indicating that children with a lower SRT score showed a higher recall difference between naturalistic and non-naturalistic sequences than children with a higher SRT score ($\beta = -0.14$ [-0.23, -0.03], $t = -2.83$, $p = .005$), and the interaction between exposure phase and SRT performance, indicating that children with a higher SRT score improved more in the early stages of the experiment than children with a lower SRT score ($\beta = 0.09$ [0.00, 0.18], $t = 1.97$, $p = .049$). Finally, an interaction between sequence type, phase, and SRT

performance became significant at the syllable level, indicating that the recall difference between naturalistic and non-naturalistic sequences decreased for children with higher SRT scores as compared to children with lower SRT scores towards the end of the experiment ($\beta = -0.05$ [-0.10, 0.00], $t = -2.02$, $p = .04$). Figure 4.7 illustrates high-SRT and low-SRT children's recall across the three exposure phases when Block 12 is removed.

Discussion

There is a growing body of evidence to suggest that SL is influenced by related prior knowledge – including phonotactic knowledge (e.g., Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Onnis et al., 2005), expectations about word length (Lew-Williams & Saffran, 2012), and knowledge of language structure (Thiessen et al., 2019). Since learners have been shown to form enduring memory representations for statistically defined input (see e.g., Batterink & Paller, 2017), it is conceivable that the influence of prior knowledge on SL also extends to prior knowledge of linguistic TPs (e.g., in the form of chunked syllables), obtained via exposure to natural language. There is promising early evidence for this possibility in adults (Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023); however, little is known about the way in which knowledge of syllable co-occurrences might influence SL in children. Here, we investigated whether seven- to nine-year-old German-speaking children used existing distributional knowledge to recall sequences of syllables and whether this was related to their language proficiency.

Children's recall was significantly better for naturalistic sequences (i.e., those which complemented the distributional properties of their native language) than non-naturalistic sequences, in line with our experimental hypothesis. This echoes the findings of related work with adults (Elazar et al., 2022; Stärk et al., 2023) and indicates that children too (in this case, seven- to nine-year-old native speakers of German) have well-entrenched memory traces of frequently co-occurring

syllables, forged through experience with their native language and which shape their subsequent learning of related material (Siegelman et al., 2018). Past research has demonstrated that both infants and adults draw upon prior experience with language structure to guide their use of forwards versus backwards TPs when processing new input (Onnis & Thiessen, 2013; Thiessen et al., 2019). The present study goes further by demonstrating that children use the precise TPs previously encountered in their native language to both recall and hierarchically group (or chunk) syllable sequences. That is, we demonstrated that children had acquired the syllable co-occurrence patterns in natural German and implicitly brought these to bear to process new linguistic input.

Children also showed better recall for non-naturalistic sequences over unstructured foil sequences at the bigram level, demonstrating once more how children readily acquire new words from input regularities (Evans et al., 2009; Kidd et al., 2020; Saffran et al., 1997). The absence of the effect at the syllable level suggests that, while the children were sensitive to the in-experiment bigrams in the non-naturalistic condition, this advantage did not extend to a greater window as measured by the syllable recall measure. We suspect that the effect was not visible at the syllable level because of the difficulty of learning simultaneous sequences constructed from the same syllable inventory. Thus, with participants receiving sequences of all three sequence types simultaneously, the immediate advantage of the naturalistic sequences might have drawn participants' attention to these patterns first, leading to a disadvantage for any other learnable pattern in the input, such as in the non-naturalistic sequences (cf. Antovich & Graf Estes, 2018; Bulgarelli & Weiss, 2016; Gebhart et al., 2009).

Contrary to our expectations, at the group level, children's recall did not improve over the course of the task. Thus, we did not replicate the pattern of within-experiment learning observed in adults in Chapter 3 (cf. Stärk et al., 2023). There are two potential explanations for this result. First, this could be explained by an implicit learning advantage

for children relative to adults. That is, that the effect emerged early and did not change thereafter could be because children rapidly identified (parts of) the patterned sequences and did not learn anything more thereafter. There is some recent evidence in support of children's more rapid acquisition of statistical patterns in comparison to adults. Smalle et al. (2017) showed that nine- to ten-year-old children implicitly acquire phonotactic restrictions on novel words faster than adults do. Additional work by Smalle et al. (2018) using the Hebbian repetition paradigm suggests that eight- to nine-year-olds show greater retention of implicitly learnt syllable sequences than adults four hours and one week after they were first tested. Thus, it appears that, in repetition learning tasks at least, children can learn faster and retain verbal information better than adults, general differences in cognition like working-memory span notwithstanding. This finding is consistent with the generally held belief that there is a critical period for language learning (Hartshorne et al., 2018; Newport, 1990). One tentative possibility is that the difference between the current data and that of Chapter 3 (Stärk et al., 2023), where children seize upon the naturalistic sequences earlier, reflects children's greater ability to identify distributional patterns.

While possible, we think this explanation might be less likely than a second explanation that appeals both to the complex nature of the task and to individual differences in children's performance on the task. As outlined above, the fact that all three sequences came from the same syllable inventory introduced a complexity to the task that is not seen, for instance, in studies of Hebbian learning, which typically have patterned and foil sequences without syllable overlap, and which typically observe improvement in the repetition of the patterned sequence relative to foil repetition across time (e.g., Smalle et al., 2018). This interference across sequence conditions may have prevented us from observing any interaction with block or phase, with the end result being that we only observed overall differences across the conditions.

Individual differences seem to have also impacted on learning across the experiment, with the children's language proficiency, as measured by the SRT task, interacting with learning over the course of the experiment. We observed a significant correlation between the SRT and performance in the naturalistic condition, even after partialling out variance attributable to foil repetition, suggesting that higher language proficiency was associated with better performance in that condition. Intriguingly, in our statistical models, we found that language proficiency interacted with the overall magnitude of the difference between the naturalistic and non-naturalistic conditions: children with lower SRT scores had a greater difference in sequence recall than children with higher SRT scores. One possible explanation for this effect is that the magnitude of the difference was smaller in the more proficient speakers because, in addition to learning the transitions in the naturalistic condition, they were also building more robust knowledge of the non-naturalistic sequences than were the children with lower proficiency. This interpretation is consistent with the fact that the high-SRT children showed a decreasing trend in their recall of the naturalistic sequences relative to the non-naturalistic sequences in the latter part of the experiment, which emerged as an effect when we removed Block 12 from the analyses. If the children were also acquiring more robust knowledge of the non-naturalistic sequences during the latter half of the experiment, this newly acquired knowledge may have interfered with their performance in the naturalistic condition, given that the sequences contained the same syllables but different transitions. That is, the simultaneous learning of naturalistic and non-naturalistic sequences may have come at a price, with the acquisition of non-naturalistic sequences interfering with the naturalistic ones as their exposure to both sequences increased across time.

Our individual differences analyses revealed some suggestive patterns, but we stress that they should be interpreted with caution and treated as preliminary. Although our measures had good internal consistency (cf. Arnon, 2020), our sample size was not large and was

further restricted due to circumstances outside of our control. Future studies could improve upon ours by recruiting a larger sample and by recruiting a larger age range. Language proficiency is correlated imperfectly with age, which may be due to the fact that the mechanisms underlying language acquisition vary across individuals in ways that are not completely age-dependent (Kidd & Donnelly, 2020; Kidd et al., 2018). Thus, more high-powered studies with wider age ranges may better elucidate the role of language proficiency in auditory SL of syllable sequences. At the moment, our data suggest that these individual differences exist and may be related to proficiency in nontrivial ways.

One final discussion point concerns the mechanism by which children are learning the syllable transitions in the task (and presumably, how they are learning the knowledge of syllable transitions they bring to the task). While the statistical structure of our sequences was defined in terms of TPs, we suspect that the process underlying learning is more likely to be the chunking of frequently co-occurring syllables into higher units – in our case, disyllabic words – following models of SL and language that identify chunking as a basic learning mechanism (Christiansen & Chater, 2016; Frank, Goldwater, et al., 2010; Perruchet, 2019; Perruchet & Vinter, 1998). One benefit of conceiving of SL as chunking is that it builds natural bridges to related areas of the literature on verbal learning (e.g., Isbilen et al., 2020; Jones, 2012; Jones et al., 2007), thus grounding the field within the broader cognitive domain of learning and memory. Tasks like serial recall are well-suited to exploring the overlap in these literatures since recall is sensitive to long-term knowledge and learning effects (e.g., Kidd et al., 2020; Smalle et al., 2018; Szewczyk et al., 2018).

Limitations and future directions

In the current study, we showed that seven- to nine-year-old children draw on their prior knowledge of syllable co-occurrences in their native language when processing new linguistic input, which has

previously been found in adults (Elazar et al., 2022; Stärk et al., 2023). We also found that children's learning of attested and unattested syllable transitions was related to their language proficiency, as measured by sentence recall. The general conclusion from these data is that children form long-term representations for distributional information acquired over language development and use this knowledge to process new input. This skill may vary across individuals. We see these results as consistent with the general observation that language acquisition involves attending to and inducing abstract knowledge from regularly occurring sequences of linguistic units (e.g., Arnon, 2021; Bannard & Matthews, 2008; Saffran, Aslin, et al., 1996). However, we stress that we are not reducing acquisition purely to SL. From a very young age, children begin to build abstract knowledge at multiple levels of description. Thus, our demonstration that seven- to nine-year-olds are sensitive to distributional information does not entail that only this information is represented, but rather shows that this is one source of information that likely matters for learning. We note that almost every theory of language acquisition incorporates SL to some degree (e.g., Chang et al., 2006; Lidz & Gagliardi, 2015; Pearl, 2021; Tomasello, 2003), but exactly how SL contributes in these theories differs substantially. An important future direction is accurately placing SL within the broader enterprise of language acquisition.

Conclusion

To conclude, we investigated the influence of children's prior distributional knowledge on their auditory SL performance and examined whether this effect was related to the children's language proficiency. Children drew upon their prior knowledge of TPs of their native language (i.e., highly frequent syllable co-occurrences) to process and learn new linguistic input, demonstrating that, like adults (cf. Chapter 3; Stärk et al., 2023), German-speaking children had indeed developed entrenched knowledge of German syllable co-occurrences, which permeated through into the experiment to shape subsequent learning. In exploratory analyses, we found that children's

performance on the SL task interacted with their language proficiency, with the results suggesting that children with higher proficiency were more sensitive to both the naturalistic and non-naturalistic patterned sequences. This is consistent with the idea that there are meaningful individual differences in SL that are related to language acquisition (Kidd et al., 2018; Siegelman et al., 2017), although significant additional work in this space is required to determine the exact nature of the effect.

Acknowledgements

Chapter 5

The influence of prior knowledge and a Zipfian frequency distribution on statistical learning

Abstract[12,13]

Past research has shown that statistical word segmentation can be facilitated by different variables, such as the learners' prior knowledge of syllable co-occurrences and words being presented in a Zipfian rather than a uniform frequency distribution. Here, we investigated how these two variables, previously studied individually, interact. In Experiment 1, participants were tested on an auditory statistical learning (SL) task. They were assigned to one of four conditions, which manipulated (i) their familiarity with the syllable pairs that the words contained ('naturalistic' attested bigrams versus 'non-naturalistic' unattested bigrams) and (ii) the frequency distribution of these words within the speech stream (Zipfian versus uniform distribution). We hypothesised that naturalistic and Zipfian conditions would improve participants' SL performance. However, while participants were above chance in all conditions, we found the *opposite* pattern. Experiment 2 examined whether the advantage of non-naturalistic over naturalistic words was driven by the length of the familiarisation phase. However, participants in the non-naturalistic condition still outperformed participants in the naturalistic condition. Consequently, we tested whether the effect was driven by the experimental words (Experiment 3). This did not seem to be the case: participants demonstrated a significant preference for attested over unattested bigrams when they received no exposure stream at all. Overall, the data are consistent with the suggestion that existing language knowledge and skewed distributions influence statistical word segmentation; however, the direction of the results are difficult to definitively explain in light of past research. We discuss possible reasons for the unexpected results.

---

[12] This chapter is based on Stärk, K., Kidd, E., & Frost, R. L. A. (in prep.). The influence of prior knowledge and a Zipfian frequency distribution on statistical learning.

[13] We originally planned to conduct an EEG study, investigating the influence of infants' prior knowledge of syllable distributions on the infants' statistical learning performance. However, due to the COVID-19 pandemic, we had to adjust those plans and conduct this online study with adult participants instead.

Introduction

Speech comprises a variety of cues that help the learner to successfully segment and process their language input. One reliable segmentation cue is the frequency with which syllables co-occur: transitional probabilities (TPs; see e.g., Perruchet & Desaulty, 2008; Saffran, Newport, et al., 1996; and see Saksida et al., 2017, and Chapter 2, Stärk et al., 2022, for corpus analyses investigating TPs in natural speech). High TPs can indicate that syllables belong to the same word while low TPs can indicate that syllables belong to different words. That is, the learner can group together frequently co-occurring syllables and assume word boundaries when TPs are low. While ample studies of statistical learning (SL) have examined this process of grouping syllables into words based on high TPs (e.g., Aslin et al., 1998; Batterink & Paller, 2017; Raviv & Arnon, 2018; Saffran, Aslin, et al., 1996; Saffran et al., 1997; Teinonen et al., 2009), recent studies have started to investigate how prior knowledge of syllable co-occurrences – acquired through exposure to the speaker's native language – influences their subsequent SL (Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023, i.e., Chapter 3). However, it is not yet clear how the facilitatory effect of prior knowledge interacts with other factors influencing SL, such as the frequency distribution of word tokens in a language.

In natural languages, words follow a Zipfian frequency distribution, which is a power law probability distribution stating that a word's token frequency and its rank are inversely related (Kaeding, 1897; Piantadosi, 2014; Zipf, 1935, 1949). This means that the most frequent word (e.g., the article "the") occurs approximately twice as often as the second most frequent word (e.g., the preposition "of"), approximately three times as often as the third most frequent word (e.g., the conjunction "and"), and so on. In consequence, there are only a few high frequency words but many low frequency words (e.g., the adjective "magniloquent"). Such skewed distributions have been shown to facilitate speech segmentation (Kurumada et al., 2013; see also Lavi-

Rotbain & Arnon, 2022). In the current chapter, we set out to investigate how a combination of familiar or unfamiliar syllable co-occurrences and a Zipfian or uniform frequency distribution influence participants' SL.

Prior knowledge of syllable co-occurrences

TPs are an important cue to word segmentation since they are often assumed to not require any prior knowledge of the to-be-learnt language, allowing learners to break into the speech stream and detect other language-specific patterns (e.g., Saffran, Aslin, et al., 1996). However, this does not mean that learners do not build upon prior knowledge of the language. For instance, the reliability of TPs as well as which direction of TP prediction is more informative (*forwards TPs:* e.g., predicting the succeeding syllable *by* from the syllable *ba* in *baby*, or *backwards TPs:* e.g., predicting the preceding syllable *ba* from *by* in *baby*; Perruchet & Desaulty, 2008) is language-dependent (Onnis & Thiessen, 2013), which means that learners first need to establish these reliabilities from the input. Importantly, once learners have established these reliabilities in their language, they are influenced by this knowledge in their way of processing new input (e.g., English participants rely on backwards TPs while Korean participants rely on forwards TPs to segment new, ambiguous language input; Onnis & Thiessen, 2013; and see Thiessen et al., 2019, for related findings in infants acquiring their languages' preferences).

The influence of prior knowledge on SL has been investigated in a variety of studies (e.g., Elazar et al., 2022; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Onnis & Thiessen, 2013; Siegelman et al., 2018; Stärk et al., 2023; Toro et al., 2011). While earlier studies showed that prior phonotactic knowledge (i.e., knowledge of phoneme co-occurrences) influences participants' word segmentation (Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Toro et al., 2011), more recent studies have shown that prior knowledge of syllable co-

occurrences also influences participants' word segmentation and SL (Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023).

Frequently co-occurring syllables likely resonate as chunks in long-term memory, with recent studies showing that experimental stimuli are more readily learnt if the patterns of syllable co-occurrence are more similar to those in participants' native language. For instance, Siegelman et al. (2018) showed that participants' performance on a SL task was better on experimental items rated as more 'word-like' by a separate set of Hebrew speakers. This is consistent with two subsequent experimental studies that directly manipulated syllable transitions in their materials. In the first study, Elazar et al. (2022) familiarised Spanish-speaking participants with one of two experimental languages. In the 'Spanish-like' condition, participants listened to a speech stream comprising experimental words based on high TPs in Spanish (i.e., syllable combinations that frequently occur in Spanish). In the 'Spanish-unlike' condition, participants listened to a speech stream comprising experimental words based on low TPs in Spanish (i.e., syllable combinations that rarely occur in Spanish). Participants in the 'Spanish-like' condition were better at accepting target words in a subsequent lexical decision task than participants in the 'Spanish-unlike' condition but worse at rejecting foils that had a high resemblance with Spanish words. This provides further evidence that participants' prior knowledge of syllable co-occurrences in their native language influences subsequent processing and learning of new language input.

In the second study, Stärk et al. (2023; cf. Chapter 3) tested German-speaking participants in a serial recall task that required them to repeat eight-syllable sequences. The sequences belonged to one of three sequence types: (i) naturalistic sequences, (ii) non-naturalistic sequences, or (iii) unstructured foil sequences. The naturalistic and non-naturalistic sequences contained experimental words, such that participants could chunk the syllables into words with sufficient exposure. This facilitated the sequence repetition and became

detectable in participants' responses. Importantly, the composition of words was critically different across the sequence types: in the *naturalistic* sequences, words were based on German syllable transitions (i.e., syllables co-occurring in the participants' native language, comparable to Elazar et al.'s 'Spanish-like' condition), while the words in the *non-naturalistic* sequences were completely devoid of any attested transitions (comparable to Elazar et al.'s 'Spanish-unlike' condition). As hypothesised, native speakers of German drew on their prior knowledge of German syllable co-occurrences to chunk the new linguistic input from the early phases of the experiment, with higher recall and faster improvement in the naturalistic sequences compared to the other two sequence types. In this chapter, we build on this finding by examining how prior knowledge of syllable co-occurrences interacts with the frequency distribution of the to-be-learnt words in statistical speech segmentation.

Zipfian frequency distribution

A second segmentation cue is word frequency. In natural language, words follow a Zipfian frequency distribution, which means that a small amount of words occurs very frequently while the remainder of words occurs increasingly less often (Kaeding, 1897; Piantadosi, 2014; Stärk et al., 2022; Zipf, 1935, 1949). For example, an article such as "the" occurs very frequently, followed by different nouns, while each noun itself occurs less often (e.g., "dog", "shoe", etc.).

Previous studies on adults' SL have found that languages in which words follow a Zipfian distribution provide better learning contexts than languages in which words follow a uniform distribution (i.e., when each word is equally frequent; Kurumada et al., 2013; see also Lavi-Rotbain & Arnon, 2022). Kurumada et al. (2013) argued that segmentation is facilitated in a Zipfian distribution by the frequent repetition of some words, which can be segmented early on due to the frequent repetitions hindering a decay of the memory representation (Ebbinghaus, 1885, 1913; Perruchet & Vinter, 1998). Subsequently,

these frequent words can aid the segmentation of adjacent words by serving as anchor points in the speech stream (Altvater-Mackensen & Mani, 2013; Bortfeld et al., 2005; Cunillera et al., 2010, 2016; Kurumada et al., 2013; Mersad & Nazzi, 2012; Shi & Lepage, 2008; Valian & Coulson, 1988; for computational evidence see Monaghan & Christiansen, 2010; but see Lavi-Rotbain & Arnon, 2022, for evidence that the effect is driven by predictability of the words in a language rather than an anchoring effect).

To test their hypothesis, Kurumada et al. (2013) examined the influence of the lexicon size (i.e., the number of word types presented in the exposure stream: 6, 9, 12, or 24 items) and the type of input distribution (i.e., Zipfian or uniform) on participants' speech segmentation performance. Their results suggest that highly frequent words in a Zipfian distribution are indeed acquired more easily and act as anchor points to aid segmentation of adjacent words (i.e., *contextual facilitation*). Qualitatively, their findings seem to point into the direction that a Zipfian distribution could be increasingly more helpful for languages with larger lexica where anchor words become increasingly more beneficial.

Interestingly though, Lavi-Rotbain and Arnon (2022) found that a Zipfian distribution can also facilitate segmentation in a small lexicon of only four words. They investigated whether the predictability of words in a distribution rather than the distributions' precise skew (e.g., Zipfian or binary, which is mostly a uniform frequency distribution, except for one highly frequent word) facilitates word segmentation. To examine this, they exposed participants to a speech stream comprising four trisyllabic word types that were presented in either a uniform, binary, or Zipfian frequency distribution. The results suggest that language-like efficiency (i.e., greater word predictability than in a uniform distribution because of some words occurring repeatedly) leads to better segmentation performance than reduced or perfect efficiency (as found in a uniform distribution where the next word is difficult to predict; see their Experiment 2). Whether the distribution

was Zipfian or binary in nature did not affect the outcome above the main effect of efficiency (see their Experiment 3). Importantly, this suggests that a Zipfian distribution can equally facilitate word segmentation in a small lexicon. In this chapter, we build on these findings by further investigating how a Zipfian distribution in comparison to a uniform distribution affects speech segmentation, especially in interaction with participants' prior knowledge of syllable co-occurrences within those distributions.

The present study

The past literature on SL shows that both prior knowledge of syllable co-occurrences (i.e., attested TPs) and skewed distributions facilitate learning. In the present study, we went a step further and investigated how these two cues interact by employing a 2x2 design, with two levels of prior knowledge: *naturalistic* and *non-naturalistic* experimental words, and two levels of word frequency distribution: *Zipfian* and *uniform*. We implemented a SL task with a familiarisation and test phase, using the experimental words from Stärk et al.'s (2023) study on the effect of prior knowledge of TPs on participants' subsequent language learning (i.e., Chapter 3). Words were presented in either Zipfian or uniform distributions. We predicted that participants who were familiarised with the naturalistic words would score higher on the two-alternative forced-choice (2AFC) segmentation task than participants who were exposed to the non-naturalistic words (cf. Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023). We also predicted that participants who were familiarised with words presented in a Zipfian distribution would score higher than participants who were exposed to words presented in a uniform distribution (cf. Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2022). Finally, we expected participants in the Naturalistic + Zipfian condition to perform best on the segmentation task because they could benefit from both facilitatory effects.

Experiment 1

Method

All of our materials, data, analyses, and results are available on the website of the Open Science Framework (OSF): https://osf.io/eq7xk/.

Participants

All experiments presented in this chapter were approved by the Ethical Committee of the Faculty of Social Sciences, Radboud University Nijmegen, and were carried out in accordance with the World Medical Association Declaration of Helsinki. Participants gave informed consent prior to their participation by checking the consent box provided within the experimental software *Gorilla* (Anwyl-Irvine et al., 2020). They were free to withdraw at any time and were compensated £3.55 upon completing the 15-minute session.

In Experiment 1, 240 native German-speaking adults were included in the analysis (89 female, 147 male, 4 non-binary; mean age = 25.7 years, SD = 4.5 years, range = 18–35 years), with approximately 60 per condition (see Table 5.1 for demographic information by condition). The sample size was informed by a previous study investigating the effect of prior distributional knowledge on adults' SL in a serial recall task (cf. Chapter 3; Stärk et al., 2023) which, in a sample of 40 participants, found a syllable recall difference of 10% between the condition in which participants could build on prior distributional knowledge and the condition in which they could not. We increased the sample size to 60 participants per condition to accommodate the differences in design and testing mode, applying a 2AFC task between-participants instead of the serial recall task within-participants while testing unsupervised and online instead of supervised in the laboratory. Participants were recruited via the online recruitment platform *Prolific* (Prolific, 2021). They were native German speakers who grew up monolingually in Germany, with normal hearing and linguistic abilities, living in Germany. Seven additional participants were tested but

Table 5.1. *Participant information for Experiment 1 by condition.*

Condition	Number of participants	Gender			Age			
		Female	Male	Non-binary	*M*	*SD*	*Min*	*Max*
Naturalistic + Uniform	61	22	38	1	25.26	4.55	18	35
Naturalistic + Zipfian	58	26	31	1	25.38	4.11	18	33
Non-naturalistic + Uniform	61	21	40	0	26.18	4.53	18	35
Non-naturalistic + Zipfian	60	20	38	2	25.93	4.80	18	35

excluded from the final sample for either failing to meet the inclusion criteria (i.e., currently living outside of Germany with less regular exposure to the German language; N = 1) or failing the auditory attention check during the familiarisation phase (N = 6).

Design

We used a between-participants 2x2 design to study the influence of stimuli naturalness (i.e., whether experimental words comprised syllable transitions present in natural German or not) and word frequency distribution on auditory SL. The two levels of the factor "naturalness" were *naturalistic* and *non-naturalistic*. Naturalistic words were based on syllable pairs occurring with high TPs in natural German, while non-naturalistic words were based on syllable pairs not occurring in this combination in a corpus of natural German. The two levels of the factor "frequency distribution" were *Zipfian* and *uniform*. In the Zipfian conditions, experimental words followed a Zipfian frequency distribution while in the uniform conditions, experimental words appeared with equal frequency. The dependent variable was participants' performance on a 2AFC speech segmentation test (see the Materials section for further details regarding the exposure streams and the segmentation task).

Materials

Exposure streams. The experimental stimuli were adapted from Chapter 3 (Stärk et al., 2023). Words were created from a pool of 12 German syllables (*fa, ge, gei, mi, mo, nu, pa, sa, su, ti, ver, zu*), obtained from a corpus analysis of the 1000 most frequent German words in the CHILDES database (MacWhinney, 2000). These syllables were concatenated to form six disyllabic *naturalistic* words (*gefa, minu, moti, pagei, versu, zusa*) and six disyllabic *non-naturalistic* words (*fazu, geimi, nuver, samo, suge, tipa*). Critically, the naturalistic words were extracted from naturally co-occurring German syllable pairs (i.e., syllable pairs occurring in German speech with high backwards TPs, which were found to be more reliable than forwards TPs in natural German; cf. Chapter 2; Stärk et al., 2022; TP > .20) while the non-naturalistic words consisted of non-co-occurring syllable pairs (TP = 0).

In the present study, we used the six naturalistic words in the Naturalistic + Zipfian and Naturalistic + Uniform conditions and the six non-naturalistic words in the Non-naturalistic + Zipfian and Non-naturalistic + Uniform conditions. For each of the four conditions, we concatenated the words into a three-minute-long speech stream containing 300 word tokens, with words presented in a pseudo-random order, avoiding direct repetition of words. In the two uniform conditions, each word occurred exactly 50 times, while in the two Zipfian conditions, words occurred with different frequencies following a Zipfian-like distribution (i.e., 130, 65, 40, 30, 20, and 15 times, respectively). To control for item-specific effects in the Zipfian conditions, we created twelve different Zipfian languages (six naturalistic and six non-naturalistic languages), with each word occurring with a different frequency in every language (i.e., every word occurred 130 times in one language, 65 times in another language, etc.; see the Stimuli folder on OSF for more details). Participants in the Zipfian conditions were automatically assigned to one of the languages by the experiment platform *Gorilla* (Anwyl-Irvine et al., 2020). There

was a five-second fade-in and fade-out to mask word boundaries at the beginning and end of the exposure streams.

Since the experiment was conducted online, we included a beep-detection attention check in the speech streams to ensure that participants were attentive throughout the familiarisation phase. All streams contained six beeps, which occurred either 20s or 30s apart from each other. The location of beeps was counterbalanced, such that three occurred within words and three occurred at word boundaries, to avoid cueing segmentation. To pass the attention check, participants were required to correctly press the space bar when they heard a beep at least five out of six times, ensuring that they were listening to the entire exposure stream.

Test stimuli. The 2AFC segmentation task comprised 24 test pairs. Each target word (i.e., the six words of the experimental language) was presented four times, paired with four different part words. Part words were constructed to share one syllable with the target word, by either adding a syllable before the initial syllable of the target word (e.g., the target word *zusa* was paired with the part words *tizu* and *geizu*, sharing the first syllable of the target word), or by adding a syllable after the final syllable of the target word (e.g., the target word *zusa* was paired with the part words *samo* and *sapa*, sharing the second syllable of the target word). The order of target word and part word in a test pair was counterbalanced, with each target word occurring once before and once after a word with which it shared its first syllable, and once before and once after a word with which it shared its second syllable. More information about the exposure streams and test items can be found in the Materials folder on OSF.

Procedure

Participants were recruited via the online recruitment platform *Prolific* (Prolific, 2021), which screened for the required inclusion criteria (see Participants section) and automatically redirected participants to the experiment on *Gorilla* (Anwyl-Irvine et al., 2020).

They first received an informed consent form, which they had to sign by checking a tick box in order to participate. After filling in a short demographic questionnaire, participants were given the chance to adjust their volume and ensure that the automatic play of audio files was enabled in their browser. Because of the online format of the study, we included a headphone check to control for participants' audio setup (a German translation of the headphone check used in Milne et al., 2021), which required participants to correctly identify which of three snippets of white noise contained a beep. There were six trials, and participants were required to achieve 100% accuracy to ensure that they were wearing headphones with a decent sound quality in a reasonably quiet environment. After passing the headphone check, participants proceeded to the main experiment, where they were automatically assigned to one of the experimental groups. Participants who failed the headphone check were asked to terminate their participation but were free to rerun the entire session by contacting the experimenter (since they had not entered the main experiment yet). The main experiment consisted of a familiarisation and a test phase. During the familiarisation phase, participants were asked to listen carefully and press the space bar when hearing a beep to ensure that they were continuously attentive throughout the familiarisation. During the 2AFC test phase, participants heard two words per trial, an experimental word and a foil, and were asked to decide which one best resembles the language to which they were previously exposed. Finally, participants were debriefed and paid via electronic transfer.

Analysis

All analyses were performed in R 4.2.2 (R Core Team, 2022) using RStudio (RStudio Team, 2022). Data pre-processing and visualisation were carried out using the package *tidyverse* 1.3.2 (Wickham, 2017; Wickham et al., 2019). Mixed-effects models were calculated using the package *lmerTest* 3.1-3 (Kuznetsova et al., 2017; based on *lme4* 1.1-31; Bates et al., 2015) while the models' performance was assessed and the model fits calculated using the package *performance* 0.10.2 (Lüdecke
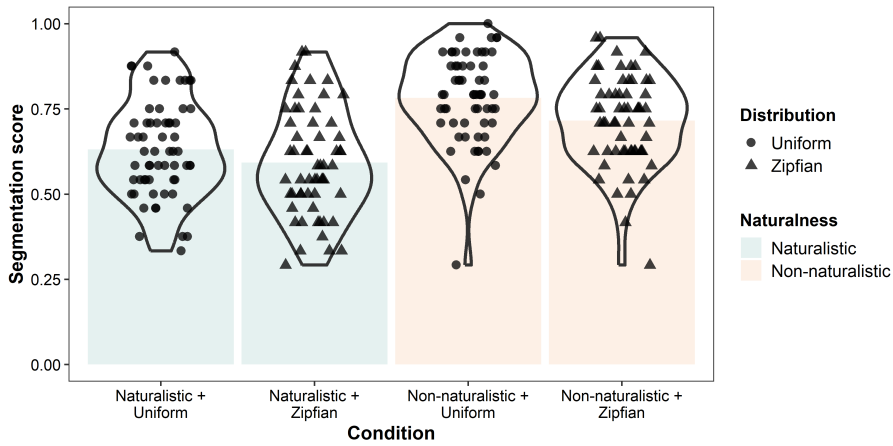
Figure 5.1. *Pirate plot illustrating the proportion of correct responses in the four experimental conditions: the naturalistic conditions (left in green) and the non-naturalistic conditions (right in orange), as well as the Zipfian conditions (triangles) and the uniform conditions (dots). Columns indicate overall means in the segmentation task, while dots and triangles represent individual participant means, with the outline indicating the distribution.*

et al., 2021). The model details are given in the Results section below (and see OSF for the analysis scripts).

Results

Contrary to our predictions, participants in the non-naturalistic conditions had a higher segmentation score than participants in the naturalistic conditions, and participants in the uniform conditions had a higher segmentation score than participants in the Zipfian conditions, with participants in the Non-naturalistic + Uniform condition scoring the highest and participants in the Naturalistic + Zipfian condition scoring the lowest (see Figure 5.1).

Participants' responses differed from chance across all conditions, indicating learning was independent of the experimental words'

Table 5.2. *Descriptive statistics and comparisons to chance for the four conditions.*

Condition	Descriptive statistics		Comparison to chance ($\mu = 0.5$)		
	M	*SD*	*t*	*df*	*p*
Naturalistic + Uniform	0.63	0.48	10.40	1463	< .001
Naturalistic + Zipfian	0.59	0.49	6.98	1391	< .001
Non-naturalistic + Uniform	0.78	0.41	26.14	1463	< .001
Non-naturalistic + Zipfian	0.72	0.45	18.10	1439	< .001

naturalness or frequency distribution (see Table 5.2 for the descriptive statistics as well as chance comparisons of all four conditions). To test whether the observed differences between the conditions were meaningful, we analysed the data using a generalised linear mixed-effects model, specifying a binary distribution with a log-link to account for the binary response of the 2AFC test. The fixed effects were the naturalness (effects coding: *naturalistic:* +1, and *non-naturalistic:* –1) and the frequency distribution (effects coding: *Zipfian:* +1, and *uniform:* –1) of the words contained in the speech stream, as well as their interaction. The only random effect added to the model was the random intercept of participants, which constituted the maximal model supported by the data (Barr et al., 2013; Bates et al., 2018; Matuschek et al., 2017) since we used a between-participants design, with participants of each condition being exposed to a different speech stream and different test items (i.e., ruling out a random intercept of items and any random slopes of participants or items). The model output showed the two observed main effects to be significant, with participants correctly choosing the target word more often in the non-naturalistic conditions than in the naturalistic conditions and more often in the uniform conditions than in the Zipfian conditions (see Table 5.3 for the model results). There was no interaction between the two factors.

Discussion

We found clear evidence of statistical segmentation, with participants' performance differing from chance on the 2AFC task in all

Table 5.3. *Summary of the generalised linear mixed-effects model investigating the influence of naturalness and frequency distribution of the words contained in the exposure streams on participants' segmentation scores.*

Parameter	*b*	95% CI	*SE*	*t*	*p*
(Intercept)	0.82	[0.73, 0.92]	0.04	18.53	< .001
Naturalness	–0.34	[–0.43, –0.26]	0.04	–7.66	< .001
Distribution	–0.14	[–0.23, –0.06]	0.04	–3.11	.002
Naturalness × Distribution	0.05	[–0.04, 0.14]	0.04	1.18	.24

Notes: Model fit: AIC = 6973; BIC = 7006; $R^2_{marginal}$ = 0.037; $R^2_{conditional}$ = 0.109; ICC = 0.075; RMSE = 0.443; σ = 1.

conditions – in line with a vast body of literature on adults' ability to segment speech via SL (e.g., Saffran et al., 1997; and see Frost et al., 2019, for a review). That is, participants showed segmentation in both naturalistic and non-naturalistic conditions (i.e., whether they had prior knowledge of syllable co-occurrences or not) and in both Zipfian and uniform conditions. However, unexpectedly, participants in the non-naturalistic conditions showed better segmentation performance than participants in the naturalistic conditions (in contrast to previous findings regarding the effect of prior knowledge on SL; cf. Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023), and participants in the uniform conditions showed better segmentation performance than participants in the Zipfian conditions (in contrast to previous findings regarding the effect of a Zipfian frequency distribution on SL; cf. Kurumada et al., 2013).

We expected participants in the Zipfian conditions to outperform participants in the uniform conditions due to the salience of the frequent word, which can be segmented early on and subsequently serve as an anchor point to facilitate segmentation of less frequent words (Kurumada et al., 2013; and see Bortfeld et al., 2005, for anchor word effects; but see also Lavi-Rotbain & Arnon, 2022, for arguments in favour of word predictability rather than anchor words guiding segmentation). However, we found the opposite effect, although there are some findings in the literature that are similar. While Kurumada et

al. (2013) found evidence for facilitated segmentation from Zipfian compared to uniform distributions, they rather found it in the form of *contextual facilitation*. That is, the more often participants had heard the word *preceding* the target word before, the more likely they were to correctly segment the target word itself. Kurumada et al. (2013) did not find a significant main effect of distribution type, but their results numerically pointed towards Zipfian distributions being more beneficial for larger lexica of at least nine word types (while our experimental languages only comprised six word types). On the other hand, another study indirectly showed that participants' segmentation of a lexicon as small as four word types can be facilitated by a Zipfian distribution as compared to a uniform distribution (Lavi-Rotbain & Arnon, 2022). We will discuss this point further in the General Discussion.

With respect to our second variable, learning from naturalistic and non-naturalistic stimuli, we expected participants to show better segmentation performance in the naturalistic conditions, in which participants could benefit from their prior knowledge of German syllable co-occurrences to segment and store the experimental words. Even though we used the same experimental words as in Chapter 3 (Stärk et al., 2023), where we found the expected effect in a serial recall task with German adults, we unexpectedly found the opposite effect in the current study, with participants in the non-naturalistic conditions reaching higher segmentation scores than participants in the naturalistic conditions. The explanation for this discrepancy likely lies within the differences in design of the two studies, with the current study using a classical familiarisation and test phase (cf. Saffran et al., 1997) and the study in Chapter 3 (Stärk et al., 2023) using a serial recall task. Crucially, we tested the effect of naturalistic versus non-naturalistic syllable transitions *within*-participants in Chapter 3: the simultaneous exposure to several experimental languages in the serial recall task might have delayed participants' learning of the non-naturalistic words because they were less salient in comparison to the naturalistic words. However, participants in the current study were only familiarised with

one language, meaning that while participants can be assumed to draw on prior knowledge to acquire the naturalistic words, they can still acquire the non-naturalistic words fairly quickly, relying on pure SL without prior knowledge. Therefore, the opposite finding might have its origins in the method of exposure, with the familiarisation phase potentially not only being long enough for participants in the non-naturalistic condition to catch up with participants in the naturalistic condition, but also to favour the acquisition of the non-naturalistic words. Hence, it is possible that a shorter familiarisation phase could yield the originally expected results, with participants in the naturalistic condition already benefitting from their prior knowledge but participants in the non-naturalistic condition not having enough time to catch up (and outperform) participants in the naturalistic condition yet. We test this hypothesis in Experiment 2.

Experiment 2

In Experiment 2, we set out to test the hypothesis that the unexpected effect of prior knowledge hindering segmentation in Experiment 1 was driven by the amount of familiarisation that participants had received (i.e., the length of the exposure stream). In Experiment 1, participants were exposed to an artificial language for three minutes (300 word tokens of six word types). Participants in the non-naturalistic condition (pure SL without prior knowledge) performed better than participants in the naturalistic condition. We hypothesised that the familiarisation phase of three minutes was long enough for participants in the non-naturalistic conditions to reach an excellent level of segmentation, concealing any initial advantage of the naturalistic conditions (cf. Chapter 3; Stärk et al., 2023) or somehow even favouring acquisition in the non-naturalistic conditions. The hypothesis that the familiarisation phase was long enough is supported by previous findings in the literature where a SL effect is even observed after a shorter familiarisation phase than in Experiment 1. In the landmark study by Saffran and colleagues, for instance, infants already showed a SL effect after a 180-word exposure of two minutes (Saffran,

Aslin, et al., 1996) while adults, too, have been found to show an early SL effect after an approx. 125-word exposure of two minutes, even in a potentially more difficult task with varying word lengths (Giroux & Rey, 2009).

We therefore shortened the familiarisation phase in Experiment 2 to one minute (or 100 word tokens, one third of the original length) and only considered the uniform distribution to investigate the effect of prior knowledge more closely. This way, each word was presented around 17 times. We hypothesised that this would be sufficient exposure for statistical segmentation in both contexts (with or without prior knowledge of syllable co-occurrences) but not too much exposure to override the advantage of the naturalistic stimuli (as in Experiment 1). Finally, we also halved the amount of test trials in the 2AFC task. An analysis of Experiment 1 showed that 12 instead of 24 test trials were sufficient to reveal any effects. Importantly, this allowed us to present target words only twice during the test, reducing the amount of possible learning that could take place in the test phase (rather than the familiarisation phase). We hypothesised that participants in the naturalistic condition would show better segmentation performance than participants in the non-naturalistic condition because they could draw on their prior knowledge of syllable co-occurrences in natural German to segment their input.

Method

Participants

A different set of participants was recruited for Experiment 2 via *Prolific* (Prolific, 2021). Eighty (N = 80) native German-speaking adults were included in the analysis (23 female, 55 male, 2 non-binary; mean age = 26.6 years, SD = 4.5 years, range = 18–35 years), 40 per condition (see Table 5.4 for demographic information by condition). As in Experiment 1, participants were monolingual, native German-speakers, with normal hearing and linguistic abilities, living in Germany. The sample size was determined via a power analysis

Table 5.4. *Participant information for Experiment 2 by condition.*

Condition	Number of participants	Gender			Age			
		Female	Male	Non-binary	*M*	*SD*	*Min*	*Max*
Naturalistic	40	10	29	1	26.77	4.60	18	35
Non-naturalistic	40	13	26	1	26.40	4.29	19	35

conducted in R 4.0.2 (R Core Team, 2022) using the package *simr* 1.0.5 (Green & MacLeod, 2016), for which data of the two uniform conditions (*Naturalistic + Uniform* and *Non-naturalistic + Uniform*) of Experiment 1 were entered into a generalised linear mixed-effects model with analogous specifications as described in the Results section below. The sample size was increased from N = 50 to N = 150 in increments of 10 participants, with 1000 Monte Carlo simulations being performed at each step (see the Analysis folder on OSF for the details). The results showed that a sample of 80 participants completing 12 trials (i.e., half of the amount of 2AFC segmentation trials presented in Experiment 1) would provide 91% power (95% CI: [89%, 93%]) to detect a log odds ratio of –0.363 (i.e., a small effect). Following this justification, we reduced the sample size to 40 participants per condition (in comparison to 60 in Experiment 1) and 12 2AFC trials (in comparison to 24 in Experiment 1; see Materials for more information on the stimuli). Participants received a compensation of £3.55 upon completion of the session.

Design

We adjusted the design from Experiment 1 by reducing it to a single between-participants comparison, only studying the influence of naturalness of the words contained in the exposure stream on participants' auditory SL. The two levels of the factor "naturalness" remained *naturalistic* and *non-naturalistic*, presented exclusively in the uniform frequency distribution. Importantly, we shortened the familiarisation phase from three minutes to one minute (i.e., from 50 to approximately 17 occurrences of each word) to test whether the length

of exposure influenced participants' SL in Experiment 1. The dependent variable was the participants' segmentation score.

Materials

Exposure streams. We adjusted the materials of Experiment 1 by shortening the familiarisation phase to one minute, or 100 words (i.e., one third of the original familiarisation phase). We only used the two uniform languages (naturalistic and non-naturalistic) and selected the part of the original exposure streams where the six words occurred with the most balanced distribution (*naturalistic:* 18 times *gefa* and *versu*, 17 times *pagei*, 16 times *minu* and *zusa*, and 15 times *moti*; *non-naturalistic:* 18 times *fazu* and *suge*, 17 times *samo*, 16 times *geimi* and *tipa*, and 15 times *nuver*). The attention check during the familiarisation phase was adjusted to the new length of the speech stream and required participants to correctly detect at least one out of two beeps.

Test stimuli. We halved the amount of trials in the 2AFC task (to 12 trials, two per target word) by reducing the part words to one sharing the target words' first syllable and one sharing the target words' second syllable, such that, for example, *zusa* would be paired with *sapa* and *tizu* (see the Materials folder of the OSF project for further details).

Procedure

The procedure was identical to that of Experiment 1, only with a shortened familiarisation and test phase.

Analysis

The analysis was identical to that of Experiment 1, with the adjusted model description given in the Results section below.

Results

Participants in the non-naturalistic condition reached a higher segmentation score than participants in the naturalistic condition (see
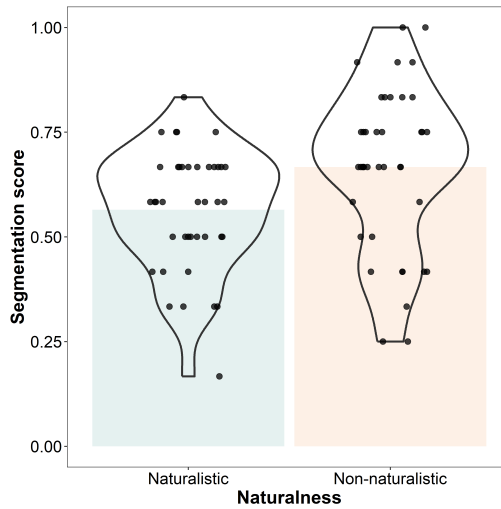
Figure 5.2. *Pirate plot illustrating the proportion of correct responses in the naturalistic (left in green) and the non-naturalistic condition (right in orange). Columns indicate overall means in the segmentation task, while dots represent individual participant means, with the outline indicating the distribution.*

Figure 5.2), which is in line with our observations in Experiment 1 but the opposite pattern we predicted. Participants' responses differed from chance in both conditions (*naturalistic: M* = .56, *SD* = .50; *t*(479) = 2.85, *p* = .005; *non-naturalistic: M* = .67, *SD* = .47; *t*(479) = 7.74, *p* < .001). As in Experiment 1, we analysed the data using a generalised linear mixed-effects model with participants' segmentation performance as the dependent variable. Due to the simplified design, only naturalness (effects coding: *naturalistic:* +1, and *non-naturalistic:* −1) was added as a fixed effect, with the random intercept of participants added as the only random effect. The model output showed the main effect of naturalness to be significant, with participants correctly choosing the target word more often in the non-naturalistic condition than in the naturalistic condition (see Table 5.5 for the model results).

Table 5.5. *Summary of the generalised linear mixed-effects model investigating the influence of naturalness of the words contained in the exposure streams on participants' segmentation scores.*

Parameter	b	95% CI	SE	t	p
(Intercept)	0.50	[0.33, 0.65]	0.08	6.06	< .001
Naturalness	−0.23	[−0.40, −0.06]	0.08	−2.77	.006

Notes: Model fit: AIC = 1268; BIC = 1283; $R^2_{marginal}$ = 0.015; $R^2_{conditional}$ = 0.065; ICC = 0.051; RMSE = 0.468; σ = 1.

Discussion

As in Experiment 1, participants' performance differed from chance in both conditions, indicating successful statistical segmentation (cf. Saffran, Newport, et al., 1996; Saffran et al., 1997; and see Frost et al., 2019, for a review). That is, participants segmented both naturalistic and non-naturalistic conditions, independent of whether they had prior knowledge of syllable co-occurrences in the experimental languages or not. However, even though the effect of naturalness decreased in comparison to Experiment 1, participants in the non-naturalistic condition still showed better segmentation performance than participants in the naturalistic condition (the opposite of our hypothesised finding, based on previous literature reporting prior knowledge facilitating SL; cf. Elazar et al., 2022; Onnis & Thiessen, 2013; Siegelman et al., 2018; Stärk et al., 2023).

We expected participants in the naturalistic condition to show better segmentation performance than participants in the non-naturalistic condition because the naturalistic condition provided the opportunity to draw on prior knowledge of syllable co-occurrences to segment the experimental words. Participants already showed the opposite effect in Experiment 1, where we speculated that the length of the familiarisation phase might have helped participants in the non-naturalistic condition to catch up with, and even outperform, participants in the naturalistic condition. After shortening the familiarisation phase in Experiment 2 to one minute (one third of the

original length), we still found the same unexpected result of a non-naturalistic advantage. This is surprising since the same stimuli previously helped participants to recall sequences of the same naturalistic words better than sequences of the same non-naturalistic words (cf. Chapter 3; Stärk et al., 2023; see also Elazar et al., 2022; Onnis & Thiessen, 2013; and Siegelman et al., 2018, for further evidence of prior linguistic knowledge facilitating SL). However, it might be possible that the task's simultaneous exposure to naturalistic and non-naturalistic words in Chapter 3 (Stärk et al., 2023), paired with the higher salience of the familiar naturalistic words, led to the observed advantage (because the salient naturalistic words received participants' attention first). It might further be that the non-naturalistic words on their own were also perceived as word-like by native German speakers, leading to a similar advantage when tested between-participants (without the direct comparison to the naturalistic words), as in the current study. We therefore conducted Experiment 3 to test whether the naturalistic or non-naturalistic words sounded more word-like to native speakers.

Experiment 3

In Experiment 3, we examined whether the non-naturalistic words were biased in appearing word-like by design, making them easier to segment and learn than the naturalistic words. This would explain the better performance in the non-naturalistic conditions compared to the naturalistic conditions in Experiments 1 and 2. We therefore asked participants in Experiment 3 to judge which word of two options sounded more word-like. That is, we omitted the familiarisation phase and directly presented participants with the 2AFC task. They either heard naturalistic or non-naturalistic target words paired with part words, as in the previous experiments (giving them the full 24 test items as in Experiment 1). We expected participants to select the naturalistic words more often as word-like in comparison to their part words, providing evidence for their sound creation as words based on high-transition syllable combinations in natural German. On the other hand,

Table 5.6. *Participant information for Experiment 3 by condition.*

Condition	Number of participants	Gender			Age			
		Female	Male	Non-binary	*M*	*SD*	*Min*	*Max*
Naturalistic	20	19	1	0	23.55	4.74	18	34
Non-naturalistic	22	18	4	0	22.00	4.03	18	35

we expected participants to select non-naturalistic words equally often as more word-like as their part words, providing evidence that the non-naturalistic words were created as plausible but neutral words (with low syllable transitions in natural German) and the part words were created as good distractors. However, if participants selected non-naturalistic words more often as word-like in comparison to their part words, it would suggest that the participants in Experiments 1 and 2 may have perceived them as equally or even more word-like than their naturalistic counterparts, potentially explaining the segmentation advantage.

Method

Participants

A new set of participants was recruited for Experiment 3 via *Prolific* (Prolific, 2021). The final sample comprised 42 native German-speaking adults (37 female, 5 male, 0 non-binary; mean age = 22.7 years, SD = 4.5 years, range = 18–35 years), with approximately 20 per condition (see Table 5.6 for demographic information by condition). As before, participants were monolingual, native German speakers, with normal hearing and linguistic abilities, living in Germany. They received a compensation of ₤3.55 upon completion of the session.

Design

We adjusted the design from Experiment 1 such that there was no familiarisation phase, and thus, we tested participants' judgement of the experimental words to identify potential biases (as expected in the naturalistic condition but unwanted in the non-naturalistic condition). The two levels of the factor "naturalness" remained *naturalistic* and

non-naturalistic. The dependent variable was the participants' 2AFC wordiness score.

Materials

Test stimuli. We adjusted the materials of Experiment 1 by removing the exposure streams. The 2AFC segmentation task comprised the same 24 test pairs as in Experiment 1, however, without the prior familiarisation, no longer testing segmentation but rather whether target words sounded more word-like than foils.

Procedure

Recruitment and screening were identical to Experiments 1 and 2, with the exception of the removal of the familiarisation phase. Upon passing the headphone check, participants immediately completed the 2AFC task, with the instruction to pick the word which sounded most word-like. Finally, participants were debriefed and paid as before.

Analysis

The analysis was identical to that of Experiment 1, with the adjusted model description given in the Results section below.

Results

Naturalistic words were rated more often as word-like in comparison to part words than non-naturalistic words were (see Figure 5.3). Notably, participants' responses differed from chance only in the naturalistic condition but not in the non-naturalistic condition, indicating that naturalistic target words resembled German words while non-naturalistic target words did not (*naturalistic: M* = .68, *SD* = .47; *t*(479) = 8.51, *p* < .001; *non-naturalistic: M* = .52, *SD* = .50; *t*(527) = 1.04, *p* = .30). As in Experiments 1 and 2, we additionally analysed the data using a generalised linear mixed-effects model with participants' performance on the 2AFC task as the dependent variable. Naturalness (effects coding: *naturalistic:* +1, and *non-naturalistic:* –1) was added
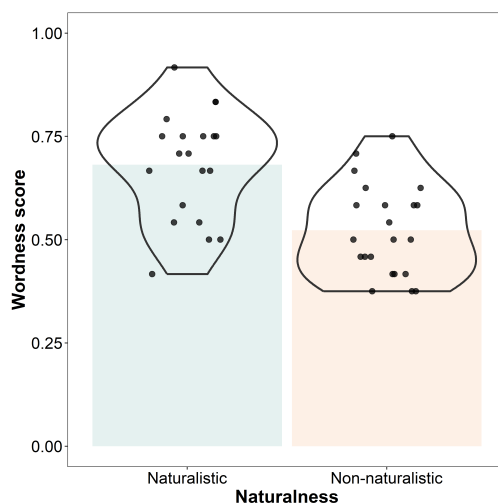
Figure 5.3. *Pirate plot illustrating the proportion of correct responses in the naturalistic (left in green) and the non-naturalistic condition (right in orange). Columns indicate overall means in the wordiness task, while dots represent individual participant means, with the outline indicating the distribution.*

as the only fixed effect. Here, this refers to the naturalness of the 2AFC target words only (since there was no familiarisation phase). The random intercept of participants was added as the only random effect. The main effect of naturalness was significant, with participants' preference for words over part words being stronger in the naturalistic condition than in the non-naturalistic condition (see Table 5.7 for the model results). This is in line with our hypothesis that naturalistic words would be perceived as more word-like based on their creation from high German syllable transitions, therefore bearing a close resemblance to German words.

Discussion

Participants chose naturalistic words more often as being "word-like" than part words, indicating that the naturalistic words contained word-like properties to native German speakers. This provides

Table 5.7. *Summary of the generalised linear mixed-effects model investigating the influence of naturalness of the target words on participants' wordiness scores.*

Parameter	b	95% CI	SE	t	p
(Intercept)	0.43	[0.28, 0.58]	0.08	5.50	< .001
Naturalness	0.34	[0.19, 0.50]	0.08	4.32	< .001

Notes: Model fit: AIC = 1335; BIC = 1349; $R^2_{marginal}$ = 0.033; $R^2_{conditional}$ = 0.056; ICC = 0.024; RMSE = 0.476; σ = 1.

evidence for sound stimulus creation of the naturalistic words, which are based upon high German syllable transitions (i.e., syllable combinations which frequently occur in natural German). Participants in the non-naturalistic condition did not show a preference for either non-naturalistic words or part words, indicating that both sounded equally word-like to native German speakers. This provides evidence that the non-naturalistic words and part words fulfilled their function well, with participants not being biased to select either one as more word-like – since both should be equally likely experimental words (with non-naturalistic words serving as condition of SL without prior knowledge in Experiments 1 and 2). While this strengthens the validity of our experimental words, it does not explain the results found in Experiments 1 and 2. We discuss this point further in the General Discussion.

General Discussion

Speech segmentation via auditory SL has previously been shown to be facilitated by learners' prior knowledge of the syllable co-occurrences of the to-be-learnt stimuli (Elazar et al., 2022) as well as by the presentation of these stimuli in a Zipfian frequency distribution (Kurumada et al., 2013). In the current chapter, we set out to study the interaction of these two cues in three experiments. In Experiment 1, German adults were randomly assigned to one of four conditions of a 2x2 design (*Naturalistic + Zipfian, Naturalistic + Uniform, Non-naturalistic + Zipfian, Non-naturalistic + Uniform*). As in a traditional

SL study, participants were familiarised with an experimental language and subsequently tested on a 2AFC segmentation task using word/part word comparisons. We hypothesised that participants in the Naturalistic + Zipfian condition would show the highest 2AFC scores since they could benefit from both cues, while participants of the Non-naturalistic + Uniform condition would show the lowest scores because neither cue was present.

However, we found the opposite result: both non-naturalistic words and words presented in a uniform distribution yielded higher 2AFC scores than naturalistic words and words presented in a Zipfian distribution. These results contradict previous findings reporting a facilitatory effect of both linguistic entrenchment (Elazar et al., 2022; Siegelman et al., 2018; Stärk et al., 2023) and a Zipfian frequency distribution (Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2022). To find an explanation for these unexpected results, we ran two follow-up experiments. In Experiment 2, we shortened the familiarisation phase to see whether the long exposure facilitated learning of the non-naturalistic words, masking an initial advantage of the naturalistic words, which was not the case. In Experiment 3, we administered the 2AFC task without an initial familiarisation phase to see whether there was a bias for the non-naturalistic words in the design of the experimental words, which was also not the case.

Before speculating on the surprising aspects of our findings, we will first draw conclusions from the expected findings. Experiments 1 and 2 showed clear SL effects, with participants picking the target words over the part words more often than chance in the 2AFC tasks of all four conditions. The Non-naturalistic + Uniform condition, in which participants showed the strongest performance, is comparable to SL experiments without any additional cues aiding the learning (e.g., Batterink & Paller, 2017; Perruchet & Desaulty, 2008; Saffran, Newport, et al., 1996). This condition can therefore be regarded as a baseline, against which to compare the other conditions. Furthermore, Experiment 3 showed that the stimuli were well designed, with

participants picking target words over foils more often only in the naturalistic condition, where target words reminded them of German, but not in the non-naturalistic condition, where target words and foils were equally artificial.

Prior knowledge of syllable co-occurrences

Why was performance in the naturalistic conditions worse than in the non-naturalistic conditions? Previous studies clearly showed an advantage of prior knowledge for SL, whether the prior knowledge was defined as the legal or frequent phonotactics of the native language (Dal Ben et al., 2021; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011; Toro et al., 2011), the reliability of forwards compared to backwards TPs (Onnis & Thiessen, 2013), the participant-rated familiarity with the experimental material (Siegelman et al., 2018), or the familiarity with the specific syllable co-occurrences directly (Elazar et al., 2022; Stärk et al., 2023). Even though we used the experimental words from Chapter 3 (Stärk et al., 2023), we did not find the same effect of naturalistic stimuli boosting SL compared to non-naturalistic stimuli, which poses the question of what differed between the two studies. Two points seem to stand out. Firstly, the current study was conducted online while the study reported in Chapter 3 (Stärk et al., 2023) was conducted in the lab, potentially suggesting that the unsupervised nature of the current experiments affected the current results. However, since participants showed clear evidence of speech segmentation in all conditions (i.e., were following the instructions, leading to puzzling but clear results), this seems an unlikely explanation.

Secondly, the current study used a between-participants design while the study reported in Chapter 3 (Stärk et al., 2023) familiarised participants with both naturalistic and non-naturalistic stimuli in interleaved sequences, with participants' performance measured throughout the familiarisation instead of in a subsequent 2AFC test as in the current study. One could argue that the interleaved presentation (especially with words of both conditions being created from the same

set of syllables) enhanced the advantage of the naturalistic sequences in Chapter 3's study because participants implicitly focused more on the familiar naturalistic items, neglecting other learnable patterns in the familiarisation. However, while this could explain an advantage of the naturalistic items, which might cancel out when applying a between-participants design exposing participants to only one condition, this cannot explain an advantage of the non-naturalistic items in the current study.

Moreover, Elazar et al. (2022) also employed a between-participants design testing whether Spanish adults would show better segmentation of Spanish-like experimental words (i.e., words comprising syllables frequently co-occurring in natural Spanish) compared to frequency-matched (i.e., Spanish-like) foils, on the one hand, and Spanish-unlike experimental words (i.e., words comprising syllables rarely co-occurring in natural Spanish) compared to frequency-matched (i.e., Spanish-unlike) foils, on the other. As measure served an old/new word decision task (i.e., a lexical decision task for experimental words) following a familiarisation phase. Participants in the Spanish-like condition were better at accepting target words compared to rejecting foils than participants in the Spanish-unlike condition, suggesting that participants' prior knowledge of syllable co-occurrences in Spanish boosted their segmentation in the Spanish-like condition. While Elazar et al. (2022) used a lexical decision task, we used a 2AFC task to measure segmentation performance. The former requires a decision on a single item while the latter requires a decision between two items. However, both measures are reflection-based (Christiansen, 2019), and there is no reason to suspect that this difference led to the opposite pattern of results.

One clue as to the explanation for the unexpected result comes from a comparison in 2AFC scores across the three experiments. It is noteworthy that, in the naturalistic conditions, participants in Experiment 3 reached higher 2AFC scores (without familiarisation) than participants in Experiments 1 and 2, while in the non-naturalistic

conditions, participants' performance in Experiment 3 did not differ from chance, unlike in Experiments 1 and 2. Consequently, the familiarisation phase seems to play a negative role in the naturalistic conditions, influencing participants' segmentation in Experiments 1 and 2 in the opposite way than we expected. We can only speculate on the driving force behind this influence. One potential reason might be that the presence of attested bigrams in the naturalistic conditions meant that participants started to look for bigger units (i.e., words containing more than two syllables), which could have led to them benefitting less from the attested statistics. In comparison, the unattested syllable co-occurrences in the non-naturalistic conditions were discoverable from the local statistics, which did not provide any misleading cues as to word length.

This may also be an explanation as to the difference between the serial recall task (cf. Chapter 3, Stärk et al., 2023) and the 2AFC segmentation task (current study). That is, the method might have influenced the results. In the serial recall task the rationale strategy is to chunk neighbouring syllables to reduce the working memory load (McCauley & Christiansen, 2019). Accordingly, the familiar syllable co-occurrences would have boosted the chunking and led to the observed facilitatory effect of prior knowledge. Participants in the 2AFC segmentation task, on the other hand, first received an exposure phase, which provided more linguistic context than the eight-syllable sequences in the serial recall task. Listening to the speech stream might have led participants to look for larger units because the attested transitions in the naturalistic conditions were derived from longer German words (five three-syllable words and one four-syllable word).

If this were true it would raise the question why Elazar et al.'s (2022) segmentation task did not lead to similar effects. One explanation might lie within the differences of Elazar et al.'s stimuli and the stimuli used in the current study. For the current study, we extracted syllable pairs with high TPs from multisyllabic German words to form the naturalistic words within the experiment. Even though these

syllable pairs might occur in several German words (e.g., *gefa* in *gefallen* or *gefangen*), they were always missing part of the original word, potentially leading participants to look for longer words in the experimental speech stream. On the other hand, Elazar et al.'s Spanish-like words were trisyllabic words of the form ABC (where each letter stands for a syllable), with AB and BC occurring with high TPs in natural Spanish. Importantly, the two syllable pairs were extracted from different words, such that their concatenation potentially pre-empted any misguidance. Note, however, that this remains highly speculative and that our unexpected findings require further investigation.

Zipfian frequency distribution

In previous studies, the presentation of stimuli in a Zipfian frequency distribution facilitated SL in comparison to the presentation of stimuli in a uniform frequency distribution, which has been found across modalities (*auditory SL:* Kurumada et al., 2013; *visual SL:* Lavi-Rotbain & Arnon, 2021). For instance, Kurumada et al. (2013) investigated the influence of a Zipfian frequency distribution on adults' auditory SL in an online study, comparable to the non-naturalistic conditions in the current study. However, even though they found evidence for facilitated word segmentation from Zipfian distributions compared to uniform distributions, they found it rather indirectly in the form of *contextual facilitation*. That is, the more often participants had encountered the word preceding the target word, the better they were at correctly segmenting the target word itself. There was indeed no main effect of distribution type, neither using a 2AFC segmentation task (as in the current study) nor using an orthographic segmentation task, adjusted from Frank et al. (2010, 2013).

Importantly, Kurumada et al.'s (2013) orthographic segmentation task allowed for the incremental tracking of participants' learning over the course of the familiarisation, and even though the main effect of distribution type was not significant, the numerical findings draw an interesting picture with regard to our present results. In the

orthographic segmentation task, participants listened to one sentence at a time. After each audio sequence, the sentence was presented on a computer screen as a string of syllables (e.g., "go lah bu pa doh ti"), and participants were asked to click on the spaces where they believed the word boundaries to be. Participants numerically benefitted from a Zipfian distribution when the lexicon contained at least nine word types or when the familiarisation was shorter than approximately 80 words (approximately 20 sequences containing three to five words each; see Kurumada et al.'s Figure 5). That is, the more words the lexicon contained, the more participants benefitted from a Zipfian frequency distribution, with highly frequent words facilitating the entrance into the new language and subsequently aiding the segmentation of adjacent words.

However, when the lexicon only contained six word types (as in the current study), Kurumada et al. (2013) observed the advantage of the Zipfian condition only within the first 80 presented words, which was then superseded by an advantage of the uniform condition upon longer familiarisation. In such a small lexicon, frequent words seem to still facilitate the entrance into the language in the early phases of the familiarisation but soon hinder it in comparison to the uniform distribution because the less frequent words do not appear often enough to be acquired as quickly. Even though these results only showed numerically and were not significant, the results of our current study (Experiment 1) further support these findings. Learning the six words in the current study is a less demanding task, in which participants seem to benefit from the equal exposure to all word types (i.e., a uniform rather than a Zipfian distribution) after the 300-word exposure.

While this suggests that the advantage of the uniform over the Zipfian conditions in the current study might be explained by the length of the familiarisation and the size of the lexicon, another study indirectly found an advantage of the Zipfian over the uniform condition in an even smaller lexicon of four words using a 2AFC segmentation task following a 128-word exposure. In this study, Lavi-Rotbain and

Arnon (2022) investigated whether the predictability of words in a distribution rather than the distributions' precise skew (e.g., Zipfian or binary) facilitates word segmentation. Before taking a closer look at the study and discussing potential reasons for their findings in contrast to Kurumada et al.'s (2013) and our own findings, we need to properly understand the concept of predictability as investigated by Lavi-Rotbain and Arnon (2022).

The predictability of a language quantifies the ease of predicting an upcoming word in the given language and is measured using the inverse concept of efficiency in information theory. For example, a language in which each word occurs equally often (i.e., a uniform language) is perfectly efficient ($\eta = 1$) but completely unpredictable. Natural languages, on the other hand, are less efficient ($\eta = .64$ being the average efficiency of 16 natural languages calculated by Lavi-Rotbain & Arnon, 2022) because certain words such as the article "the" get repeated frequently, making the language more predictable. The efficiency of a language is calculated using another concept in information theory, entropy, which measures the level of uncertainty or surprisal of a given event (see Equations (1a) and (1b)). For example, in a uniform language consisting of two words, each word occurs with a probability of .5. The entropy of that language (i.e., the observed entropy) can be calculated using the dividend of Equation (1b): $H_{obs} = -\sum_{i=1}^{2} 0.5 \times \log_2 0.5 = 1$. In a uniform distribution, the observed entropy equals the maximal entropy: $H_{max} = \log_2 2 = 1$. Efficiency is the ratio between the two and normalises entropy by set size, such that efficiency will only take values between 0 (least efficient/most predictable) and 1 (most efficient/least predictable). Consequently, the efficiency of a uniform distribution is always $\eta = 1$ while the efficiency of any skewed distribution is $\eta < 1$.

(1a) $\text{Efficiency} = \dfrac{\text{observed entropy}}{\text{maximal entropy}}$

(1b) $\eta(X) = \dfrac{-\sum_{i=1}^{N} p(x_i) \times \log_2 p(x_i)}{\log_2 N}$

Lavi-Rotbain and Arnon (2022) conducted one corpus analysis and two experiments. In the corpus analysis, they analysed the predictability of 16 natural languages in corpora from the CHILDES database (counting British English and American English separately; MacWhinney, 2000). The languages' predictability ranged from .59 (British English) to .70 (Estonian), with an average of .64 ($SD = 0.03$). In their second and third experiment, they built on this knowledge and created experimental languages that differed in their predictability from perfect ($\eta = 1$) over reduced (range = .83–.85) to language-like (range = .54–.65). They tested adults (and nine- to twelve-year-old children) in a 2AFC segmentation task following a familiarisation phase. The experiments were conducted between participants, with Experiment 2 comparing two uniform conditions and two binary conditions and Experiment 3 comparing two binary conditions and two Zipfian conditions. The results showed that adults (and children) were better at segmenting words from the conditions with language-like efficiency, regardless of the skew of the distribution (i.e., whether it was binary or Zipfian). Participants' performance in the conditions with reduced efficiency did not differ from (other) participants' performance in the uniform conditions.

Even though this seems to suggest that Lavi-Rotbain and Arnon (2022) found an advantage of a Zipfian over a uniform condition, the picture is more complex than that because they additionally included predictability as a variable, which was not part of the research question in Kurumada et al.'s (2013) or the current study. To be able to better compare the experimental languages within the three studies, we calculated the efficiency of Kurumada et al.'s and our own experimental conditions (see Table 5.8). Following these calculations, we must conclude that all of Kurumada et al.'s and our own Zipfian languages have a predictability comparable to Lavi-Rotbain and Arnon's languages with reduced efficiency. Based on this comparison and Lavi-Rotbain and Arnon's findings, we would expect to see no difference between Kurumada et al.'s uniform and Zipfian conditions, which is indeed in line with their findings. On the other hand, we would also

Table 5.8. *Efficiency measures of experimental languages given by condition within studies.*

Experimental languages by condition within studies			η
Uniform			1.00
Binary			
Lavi-Rotbain & Arnon (2022)			
Reduced efficiency			0.85
Language-like efficiency	Experiment 2		0.54
	Experiment 3		0.65
Zipfian			
Kurumada et al. (2013)			
6 word types	Experiment 1		0.84
	Experiment 2		0.86
...			
36 word types	Experiment 1		0.79
Lavi-Rotbain & Arnon (2022)			
Reduced efficiency			0.83
Language-like efficiency			0.61
Current study			0.85

Note: Kurumada et al. (2013) do not provide the exact number of times with which words are repeated in their conditions. The information in this table is based on estimates from their Figure 2.

expect to see no difference between our uniform conditions and our Zipfian conditions in Experiment 1, although we found the uniform distributions to facilitate segmentation in comparison to the Zipfian distributions. This unexpected finding is inconsistent with Lavi-Rotbain and Arnon's findings.

Leaving aside the issue of predictability and returning to our current research question, two questions remain regarding Lavi-Rotbain and Arnon's (2022) results. Firstly, they claimed that the skew of the distribution (i.e., whether the distribution was Zipfian or binary compared to a uniform distribution) did not affect segmentation performance on top of the language's predictability. However, since Lavi-Rotbain and Arnon only used four word types, one could argue that their binary distribution (with one word being highly frequent and

three words having the same, lower frequency) hardly differed from their Zipfian distribution (with one word being more than twice as frequent as the next most frequent word, which differs less in frequency to the remaining two words), which might explain why they did not find an effect of distribution type. We therefore suggest that further research is necessary to draw conclusions about how the skew of a distribution affects learning.

Secondly, following our discussion above, the question remains why Lavi-Rotbain and Arnon (2022) found an advantage of a Zipfian over a uniform distribution in a lexicon of only four word types using a 2AFC task following a relatively long familiarisation (i.e., conditions comparable to the ones under which Kurumada et al. (2013) found no effect while we found the opposite effect). The only possible explanation why they might have found this effect is because they compared words to non-words rather than part words in the 2AFC task, simplifying the task and bringing effects to light which were hidden in the more difficult design (using part words as foils) employed by Kurumada et al. (2013) and ourselves.

In conclusion, our results are in line with numerical findings by Kurumada et al. (2013) suggesting that in a lexicon of six word types, a uniform distribution might facilitate segmentation in comparison to a Zipfian distribution after sufficient exposure (approximately 80 word tokens). Lavi-Rotbain and Arnon (2022) claim that the picture is more complex than that, with the predictability of a distribution rather than its skew influencing segmentation. However, further research is necessary to uncover when a Zipfian distribution aids (or hinders) segmentation and how predictability interacts with the precise skew of a distribution in influencing segmentation.

Conclusion

We investigated the interaction of two types of influences, previously found to facilitate SL, namely participants' prior knowledge of syllable co-occurrences (Elazar et al., 2022; Stärk et al., 2023, cf.

Chapter 3) and the presentation of the stimuli in a Zipfian rather than a uniform frequency distribution (Kurumada et al., 2013). Unexpectedly, we found both prior knowledge and a Zipfian distribution to hinder rather than facilitate SL. We speculated that the attested bigrams in the naturalistic exposure streams (mis)led participants to look for bigger units, hindering their segmentation in these conditions. However, further research is necessary to investigate this unexpected finding. Furthermore, we suggested that the post-familiarisation segmentation test in combination with the length of the familiarisation and the size of the lexicon might have prevented us from observing the Zipfian advantage, with participants displaying the advantage only in the early phases of familiarisation (Kurumada et al., 2013; but see Lavi-Rotbain & Arnon, 2022, for findings suggesting otherwise). Future research in this area might test several Zipfian languages of different efficiencies while shortening the familiarisation phase as in Experiment 2 of the current study or adopting a different test. Kurumada et al. (2013) used the orthographic segmentation task but we would like to suggest the use of an auditory task for the investigation of auditory SL (see e.g., Isbilen et al., 2020, for a post-familiarisation test; and Stärk et al., 2023, for a serial recall or incidental learning test).

Chapter 6

General Discussion

In this thesis, I investigated two research questions. Firstly, I examined the availability and reliability of five word segmentation cues in German child-directed speech (CDS), in order to shed light on the input German-acquiring infants receive and to deepen our understanding of the possible utility of these cues during language acquisition (Chapter 2). Secondly, I studied how children and adults build on their prior knowledge of syllable co-occurrences in their native language, knowledge acquired via statistical learning (SL) in the natural world, when processing new language input based on these syllable distributions (Chapters 3 to 5). In this final chapter, I will summarise my main findings before discussing their implications for the two main research questions asked in this thesis. I also identify potential directions for future research.

Summary of findings

In **Chapter 2**, I addressed the first research question regarding the availability and reliability of word segmentation cues in German CDS. To answer this question, I conducted a corpus analysis investigating approximately 4000 utterances or 15,000 words of CDS from 20 German datasets on the CHILDES database (MacWhinney, 2000), which equals approximately one day worth of input to a child (Donnelly & Kidd, 2021). I coded the corpus at the word and syllable level to cover a range of five different segmentation cues: word stress, transitional probabilities (TPs), lexical and sublexical frequencies, word length, and single-word utterances. The aim of this analysis was to gain both an overview of a broad range of segmentation cues available in German CDS (whereas previous studies usually focussed on a single cue) and precise knowledge about German TPs, which laid the foundations for the remainder of this thesis.

The results of Chapter 2 showed that all five of these cues could be useful for word segmentation in German but had different degrees of availability and reliability. The most reliable segmentation cue was word stress, with 97% of words carrying word-initial stress. TPs proved

to be higher within than between words, making them suitable as segmentation cue. Backwards TPs were higher than forwards TPs, showing that backwards TPs are more informative in German. Furthermore, words followed a Zipfian-like frequency distribution, which has previously been suggested to facilitate word segmentation (Kurumada et al., 2013), suggesting that the frequency landscape of words could also serve as a cue. Regarding word length, 78% of word tokens were monosyllabic and the majority of the most frequent words were monosyllabic function words, which might point to their importance in flagging neighbouring words after being picked up early due to their salience (Frost et al., 2019; Shi & Lepage, 2008). Similarly, isolated words (i.e., words in single-word-utterances) could be learnt more easily and subsequently aid segmentation of adjacent words (Brent & Siskind, 2001; Peters, 1983). These accounted for 15% of all utterances.

In Chapters 3 to 5, I addressed the second research question, regarding the effect of prior distributional knowledge on SL. In **Chapter 3**, I investigated whether German adults would draw on their prior knowledge of syllable co-occurrences when processing new language input. Participants performed a serial recall task which required them to repeat sequences of eight syllables. Those sequences belonged to one of three types: naturalistic sequences, non-naturalistic sequences, or unstructured foil sequences. Naturalistic and non-naturalistic sequences were structured, such that they contained experimental words that were statistically defined via TPs, which participants could acquire over the course of the experiment. Crucially, the experimental words in the naturalistic condition comprised highly frequent syllable pairs from the participants' native language (German). That is, participants had prior knowledge in this condition, gained via exposure to their native language. The experimental words in the non-naturalistic condition comprised syllable pairs not found in a corpus analysis of German, meaning that participants started learning from a lower base (i.e., as in traditional SL experiments). Unstructured foil sequences did not contain any learnable patterns.

The results of Chapter 3 showed that participants learnt in both naturalistic and non-naturalistic sequences compared to unstructured foil sequences. Crucially, participants were better at repeating naturalistic sequences than non-naturalistic sequences. They even built further on this naturalistic advantage and improved at repeating naturalistic sequences during the early phases of the experiment in comparison to the non-naturalistic sequences.

After investigating the effect of prior knowledge on adults' SL in Chapter 3, I asked whether and how prior knowledge of syllable co-occurrences affected the SL of seven- to nine-year-old German children in **Chapter 4**. Additionally, I explored in this chapter how the children's language abilities affected their SL. The children performed an adapted version of the serial recall task used in Chapter 3, shortening the sequences to six syllables. On a second day, their language proficiency was tested with the German PPVT and a German sentence repetition task. Like the adults, the children learnt in both naturalistic and non-naturalistic sequences compared to unstructured foil sequences. Importantly, they were also better at repeating naturalistic than non-naturalistic sequences but, unlike the adults, the children did not improve further over the course of the experiment. There was some evidence in that children with higher language proficiency also showed a better SL performance. Overall, children with higher language proficiency improved more during the early phases of the experiment compared to children with lower language proficiency. Additionally, the difference between children's repetition of naturalistic and non-naturalistic sequences was smaller in children with higher language proficiency, suggesting that these more proficient children may have experienced more competition between the two trained sequences, raising the possibility that they were better at learning the two sequences in parallel.

In Chapters 3 and 4, I showed that prior knowledge of syllable co-occurrences facilitates SL of new linguistic input based on these familiar patterns. In **Chapter 5**, I went a step further and investigated

how the factor of prior knowledge interacted with another factor previously shown to facilitate word segmentation, namely the presentation of words in a Zipfian rather than a uniform frequency distribution (Kurumada et al., 2013). This study had a between-participants design in which participants were assigned to one of four conditions (Experiment 1): Naturalistic + Zipfian, Naturalistic + Uniform, Non-naturalistic + Zipfian, or Non-naturalistic + Uniform. The study was conducted online, where participants were exposed to an unfamiliar language for three minutes, after which they completed a two-alternative forced-choice (2AFC) segmentation task. The words in the language either comprised syllable pairs found in natural German (naturalistic words) or syllable pairs unattested in a corpus of German (non-naturalistic words). The words were either presented in a Zipfian distribution (with one word being highly frequent and the other words being less and less frequent) or in a uniform distribution (with each word occurring equally often).

Participants in all four conditions performed above chance, indicating segmentation across the board. However, unexpectedly, participants in the non-naturalistic conditions outperformed participants in the naturalistic conditions, and participants in the uniform conditions outperformed participants in the Zipfian conditions. To explain these findings, I conducted two follow-up experiments. Experiment 2 shortened the exposure phase from three minutes to one minute and only compared the naturalistic to the non-naturalistic condition, keeping the word frequency uniform. Participants in both conditions performed above chance. However, as in Experiment 1, participants in the non-naturalistic condition outperformed participants in the naturalistic condition. Experiment 3 tested the validity of the experimental words by asking participants in a 2AFC task without prior exposure to pick the option which sounded more word-like. Participants in the naturalistic condition picked the naturalistic words over the foils above chance while participants in the non-naturalistic condition did not pick the non-naturalistic words more often than the foils.

Interacting word segmentation cues: Bringing the five studied cues back together

In the following sections, I further interpret the findings of my dissertation, starting with the first main research question regarding the availability and reliability of word segmentation cues in German CDS. Overall, my results are in line with suggestions that children use multiple cues to break into the speech stream (see e.g., Brent & Cartwright, 1996; Cairns et al., 1997; Christiansen et al., 1998; Lalonde & Werker, 1995; Mattys et al., 1999; Monaghan, 2017; Morgan & Saffran, 1995), providing evidence for the reliability of five potential cues.

The most reliable cue meets the most available cue: Word stress and transitional probabilities

I found word stress to be the most reliable word segmentation cue in the corpus, having almost perfect reliability. The result is similar to findings in English, where 90% of content words have been found to carry word-initial stress (Cutler & Carter, 1987; compared to 93% in the current study). Infants acquiring these languages might only need to segment a few content words to hypothesise that their language follows a word-initial stress pattern. Once they have established that the stress pattern is highly reliable, they can rely mostly on this cue, potentially supported by other cues to avoid missegmenting words which are not following the dominant stress pattern. TPs (or syllable co-occurrence frequencies), on the other hand, are always available and do not require any prior knowledge, but they were less reliable than stress in German CDS. That is, TPs are a promising first cue into a language in which an infant needs to find out which other cues to rely upon and which patterns those other cues follow (see e.g., Aslin et al., 1998; Saffran, Aslin, et al., 1996; Saksida et al., 2017), but TPs are likely used in combination with other cues (see e.g., Yang, 2004).

There is evidence that word stress and TPs might interact in German and English such that TPs help extracting first word candidates from

speech, over which the infant can generalise that the languages follow a word-initial stress pattern. English seven-month-old infants have been shown to rely more on TPs than word stress when the two cues contradict one another but to change their preference in favour of word stress at approximately eight or nine months of age, suggesting that they first used TPs but later acquired the language's stress pattern (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; see also Jusczyk et al., 1993, for evidence that English-acquiring infants develop a preference for a strong/weak stress pattern between six and nine months of age). In German, there is evidence that six- to seven-month-old infants rely more on word stress than on TPs when the two cues contradict one another (Marimon Tarter, 2019), indicating that infants at that age have already learnt that stress is more reliable than TPs. Future research is necessary to show whether younger German infants prefer TPs over stress.

Importantly, children's preference for one cue over another has to be interpreted in the context of contradicting cues. That is, it suggests that the children have, for example, acquired the stress pattern of their language and learnt that it is more reliable than TPs. In natural language, however, cues will more often go hand-in-hand rather than contradict one another, and there is evidence that TPs remain relevant even after stress has been established as a more reliable cue. English 7.5-month-old infants have been shown to correctly segment words following the dominant trochaic stress pattern but to treat all strong syllables as word onsets, therefore missegmenting words following an iambic stress pattern (e.g., segmenting "TAR is" from "guiTAR is"; Jusczyk, Houston, et al., 1999). The authors suggested that before learning to correctly segment words following the non-dominant iambic stress pattern at 10.5 months of age, infants might use TPs alongside stress to segment words. That is, once the infants have established that a stressed syllable likely indicates a word onset in English, they use this cue to identify the beginning of a new word but might still rely on TPs to determine when the word ends (e.g., "TAR is" was segmented because the syllable "TAR" is stressed and often

followed by the syllable "is"). Taken together, these findings demonstrate how linguistic cues rarely work in isolation in natural language across development.

Finally, it seems worth mentioning that not only forwards but also backwards TPs can be used to segment speech and that backwards TPs are even more reliable than forwards TPs in languages such as German or English (cf. Chapter 2; Onnis & Thiessen, 2013; Perruchet & Desaulty, 2008; Thiessen et al., 2019). Thiessen et al. (2019) showed that English-acquiring infants develop a preference for backwards over forwards TPs between seven and 13 months of age. Given the results discussed above, this might suggest that German and English infants first use forwards and backwards TPs to segment the first word candidates of the language and to decide where a word ends following a stressed word onset (cf. Jusczyk et al., 1999). They then learn that backwards TPs are more reliable, though again, that does not mean that the infants stop using forwards TPs at that point but rather that they learnt which cue to trust when there is contradicting input.

How do other cues fit into the picture?

In the last section, I pointed out how combining word stress and TPs could aid word segmentation, two cues standing out in German due to their availability and reliability in the input. In this section, I discuss how the remaining three cues studied in Chapter 2 might fit into the picture. Highly frequent words and words presented in isolation can both be segmented early due to their salience (see e.g., Ambridge et al., 2015, for a review of the role that frequency plays in first language acquisition; and for accounts on the role of single-word-utterances in first language acquisition see e.g., Brent, 1999; Brent & Cartwright, 1996; Brent & Siskind, 2001; Monaghan & Christiansen, 2010; Peters, 1983; Pinker, 1984). Children might discover these words combining information from TPs (i.e., syllable co-occurrence frequencies) with information from word frequency and pauses (before and after words presented in isolation). That is, words might stand out from speech due

to their frequency or their prominent position between pauses, but it might be the frequency with which two syllables of a disyllabic word, for example, occur in combination that leads to them being chunked and stored as a potential word candidate. These high-frequency or isolated words might be the first word candidates over which children generalise to determine the language's stress or phonotactic pattern. Subsequently, all these cues might interact in guiding future segmentation, with the previously segmented high-frequency and isolated words acting as anchor points in the speech stream (Bortfeld et al., 2005; Kurumada et al., 2013; Mersad & Nazzi, 2012) and word stress and TPs interacting as described above.

An interesting question is how word length influences segmentation. Most words in the corpus were monosyllabic, which do not have any within-word TPs (at the syllabic level). However, uniformity in word lengths could potentially circumvent the problem a variety of word lengths seems to pose on word segmentation (Johnson & Tyler, 2010; Lew-Williams & Saffran, 2012). On the other hand, a variety of word lengths might only pose a problem for small lexica where word length is controlled artificially (see also Perruchet & Vinter, 1998, for computational evidence that variable word length does not lead to segmentation difficulties). In natural language, a variety of word lengths might instead facilitate word segmentation. For example, the most frequent words in the present corpus were almost exclusively monosyllabic function words such as the article "the", which can precede all nouns, leading to high backwards TPs between the nouns and the article.

Highly frequent, monosyllabic function words play an important role in segmentation and simultaneous grammatical categorisation (Frost et al., 2019). Infants have been shown to segment pseudo-words when they are preceded by the article "the" but not when they are preceded by a pseudo-article (Shi et al., 2006; Shi & Lepage, 2008), indicating that the infants segmented the article in natural language and relied on it for subsequent language processing (see also Shafer et al.,

1998, for evidence that even 10.5-month-old infants notice when function words are missing or mispronounced). In German, for instance, learning a highly frequent article can help segmenting and categorising the succeeding nouns, not only regarding the word category but also the noun's gender (see Höhle et al., 2004, for evidence that German twelve- to 14-month-old infants use articles to categorise the following word as a noun; and see also van Heugten & Shi, 2009, for evidence that French toddlers know which gender article precedes familiar nouns). It has also been suggested that function words may help infants to determine the word order of their language (Gervain et al., 2008), which might simultaneously help the infant to determine which direction of TPs is more informative in their language (such as backwards TPs in the example above; Thiessen et al., 2019).

Interim summary on German word segmentation cues

To summarise, my findings provide evidence for the availability of multiple potential word segmentation cues in German, which has previously been found to be advantageous for learning natural and artificial languages (Brent & Cartwright, 1996; Cairns et al., 1997; Christiansen et al., 1998; Cunillera et al., 2006; Lalonde & Werker, 1995; Mattys et al., 1999; Matzinger et al., 2021; Monaghan, 2017; Morgan & Saffran, 1995). Word stress is the most reliable cue in German but it likely interacts with TPs, isolated words, word frequencies, and specifically monosyllabic, highly frequent function words, not only in segmenting but simultaneously categorising words. In Chapter 5, I investigated how a language's word frequency distribution (Zipfian vs. uniform) interacts with learners' prior knowledge of the language's syllable distributions in influencing word segmentation (see discussion regarding the second research question below), but further research will be needed to investigate the interaction of different and multiple segmentation cues (e.g., whether learning words of two word lengths can simultaneously be facilitated by having monosyllabic function words).

How does prior knowledge influence subsequent language learning and processing?

After gaining insights into which word segmentation cues are available and reliable in German CDS (cf. Chapter 2), Chapters 3 to 5 dived deeper into SL, the grouping of elements which often co-occur in the environment such as the syllables of a word. The output of SL is assumed to be stored as long-term memory representations; however, how these representations endure and whether speakers draw on these to process new input was largely unknown. Such prior knowledge had been shown to affect subsequent learning at the phoneme level (i.e., in terms of phonotactics, see e.g., Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011). At the syllable level, however, only the most recent studies started to investigate whether and how prior knowledge affected subsequent SL (Elazar et al., 2022; Siegelman et al., 2018). I investigated this in Chapters 3 to 5 of this thesis.

Prior knowledge of syllable co-occurrences facilitates statistical learning

In Chapters 3 and 4, prior knowledge facilitated adults' as well as seven- to nine-year-old children's SL in the naturalistic condition. The two experiments (adults vs. children) are not completely comparable because the children received a simplified version of the task, with sequences containing only six instead of eight syllables, but the advantage of the naturalistic condition showed clearly in both participant groups. Both adults and children recalled more syllables and bigrams (i.e., experimental words) in the naturalistic condition where they could rely on prior knowledge than in the non-naturalistic condition where they could not rely on prior knowledge. This is in line with recent findings in Hebrew and Spanish adults (Elazar et al., 2022; Siegelman et al., 2018), as well as comparable findings with SL at the phoneme level (i.e., phonotactics; Finn & Hudson Kam, 2008; Mersad & Nazzi, 2011), and the broader verbal learning literature where it has

long been known that learning builds on prior knowledge (Ebbinghaus, 1885, 1913).

This result also fits well into descriptions of SL as a form of chunking and entrenchment (Jost & Christiansen, 2017; Perruchet & Pacton, 2006; Perruchet & Vinter, 1998; Robinet et al., 2011) where syllables get chunked into larger units (e.g., *ba + by → baby*), with those becoming more entrenched with each encounter. In terms of the findings in Chapters 3 and 4, this can be interpreted as participants already starting with well entrenched syllable chunks when entering the experiments. During the exposure to new linguistic input, participants could then quickly access those chunks, which explains the advantage observed for the naturalistic condition. Additionally, the adult participants in Chapter 3 were able to further build on this knowledge, which resulted in a boost of learning (i.e., further entrenchment leading to an improvement in recall scores), especially during the early phases of the experiment. The children in Chapter 4 did not show further learning during the experiment, probably due to the difficulty of the task for their age.

While the results of Chapters 3 and 4 are clearly in favour of this interpretation and my hypotheses, the results of Chapter 5 seem to contradict those. In Experiments 1 and 2 of Chapter 5, participants in the non-naturalistic conditions reached higher segmentation scores than participants in the naturalistic conditions. I ruled out several potential driving forces behind this result, such as the online nature or the between-participants design of the experiments in Chapter 5. Most importantly, the experimental words were well designed. Participants in Experiment 3 rated naturalistic words as more word-like than foils, demonstrating that the syllable pairs sounded familiar to German speakers due to the bigrams' attested co-occurrence in natural German. Furthermore, participants did not show a preference for non-naturalistic words over foils, demonstrating that both were likely word candidates, without any preference driven by potential similarity to German.

The most likely explanation for the non-naturalistic advantage in Chapter 5 is that the exposure phase in the naturalistic conditions had an inhibitory effect on participants' segmentation. Without the exposure phase, participants in the naturalistic condition of Experiment 3 rated the naturalistic words as more word-like than the foils. However, after the exposure phases in Experiments 1 and 2, participants selected naturalistic words over foils less often. I speculated that participants might have been implicitly looking for longer words in Experiments 1 and 2 of Chapter 5 because the syllable pairs were derived from longer German words and were presented in continuous speech during the exposure phase, making it possible that the word continued.

Individual differences in children's statistical learning

I have described above how prior knowledge acquired via previous SL facilitates subsequent SL. Additionally, I investigated in Chapter 4 how the children's language proficiency affects their SL performance. The theory behind this is based on a slightly more complex feedback loop, with superior SL abilities (or its component processes, Arciuli, 2017) leading to better language proficiency, which could in turn feed back into better SL performance. Previous studies had found a correlation between language proficiency and SL performance (see e.g., Evans et al., 2009; Frost, Jessop, et al., 2020; Kidd & Arciuli, 2016; Lany, 2014). Here, I could not only show that children with higher language proficiency performed better at the SL task but also that there were interesting, different learning trajectories for children with higher and children with lower language proficiency. Children with higher language proficiency were better at building on their prior knowledge and therefore performed better at recalling naturalistic sequences. Additionally, they improved more during the early phases of the experiment and also picked up more of the non-naturalistic sequences than children with lower language proficiency. These findings provide interesting insights into potential future research avenues.

Prior knowledge of syllable co-occurrences in interaction with the word frequency distribution

Apart from the effect of participants' prior knowledge, I also studied the effect of a Zipfian frequency distribution on participants' SL in Chapter 5. I did not find an interaction between participants' prior knowledge and the frequency distribution of the experimental language. However, I unexpectedly found a uniform distribution to be more beneficial than a Zipfian distribution. This seems to be in line with previous observations by Kurumada et al. (2013), showing that in a small lexicon of only six word types a uniform distribution facilitates segmentation in comparison to a Zipfian distribution after a certain amount of exposure (after approximately 80 words). Their finding was not statistically significant but the current results of Experiment 1 provide further evidence for it. Adding another level of complexity, Lavi-Rotbain and Arnon (2022) found a Zipfian distribution to facilitate word segmentation in comparison to a uniform distribution in a small lexicon of only four word types but only when the experimental language had a language-like efficiency (i.e., was more predictable than our experimental language). Future research will therefore be necessary to disentangle the precise contributions of the different factors of lexicon size, skew, and predictability of a language on participants' word segmentation.

Interim summary on factors influencing statistical learning

To summarise, my findings are in line with descriptions of SL as a form of chunking and entrenchment where syllables that often occur together form a chunk, which becomes stronger with each activation (cf. Perruchet & Pacton, 2006; Perruchet & Vinter, 1998; Robinet et al., 2011). Participants entered the experiments with well entrenched syllable chunks, facilitating their SL performance in Chapters 3 and 4 as well as influencing their word segmentation in Chapter 5. The adult participants in Chapter 3 additionally showed signs of further entrenchment during the task. Children's language abilities influenced

their SL performance, and a Zipfian frequency distribution influenced participants' word segmentation, with both of these findings offering interesting new research avenues.

Future directions

Having discussed the findings of my dissertation, I will mention a few potential directions for future research in this section. In Chapter 2, I analysed data which approximates to one day worth of input. One next obvious step would be to generalise over a bigger dataset or investigate whether the availability and reliability of cues changes over time when caregivers adjust their CDS to the more advanced linguistic needs of the child. Investigating such a large dataset requires the automation of coding and analysis processes, in which case it would be only logical to include even more potential word segmentation cues into the analysis, such as cues on the phonemic level (e.g., phonotactics). A more interesting question than how the reliability of cues changes over time in one language (which might not be that much) is how different languages compare to one another (i.e., which cues are comparable between which languages in terms of availability and reliability, and which cues differ; see e.g., Saksida et al., 2017).

Building on findings from such corpus analyses, experimental studies can investigate the cues on which children rely in specific languages and at specific time points during language development. For instance, I discussed above how TPs could be more relevant in the first months of life while word stress might become more dominant in English and German once the child has established a reliable stress pattern (Johnson & Jusczyk, 2001; Marimon Tarter, 2019; Thiessen & Saffran, 2003). However, in syllable-timed languages such as French, infants rely more on TPs than on stress (Marimon et al., 2019). Such cross-linguistic studies between languages with different patterns of cue reliabilities might reveal important language-specific trajectories in word segmentation strategies. For instance, which cues do children use in combination, and are there individual differences (see also Marimon

et al., 2022), or does this change with age? The language environment might influence the availability and reliability of a certain cue but individual cognitive abilities might then influence whether the cue is easy to process for a child (potentially in comparison to other only slightly less reliable cues). This line of research would provide valuable insights into typical language development within the first year of life.

Building upon the work in Chapters 3 to 5, one next step to take here would be to run further follow-up experiments on Chapter 5 to investigate whether a shorter exposure phase would yield a Zipfian advantage (cf. Kurumada et al., 2013). Repeating Experiment 1 using the serial recall task from Chapters 3 and 4 might reveal the underlying factors driving the unexpected findings (I had to switch to the 2AFC task due to COVID-related testing limitations). As discussed in Chapter 5, it would be good to also include the languages' predictability (and, if possible, different lexicon sizes) into the study. All these points would help moving the field to more naturalistic experimental settings. In line with this, future studies should try to implement more naturalistic stimuli (e.g., using different syllable structures typically found in certain positions within a word) and combine different segmentation cues in a variety of languages. Finally, computational models could help gain further insights into multiple interacting cues.

Conclusion

As complex systems, languages present a seemingly insurmountable problem to the learner. However, languages contain many probabilistic cues, and humans excel at spotting and learning from these patterns in their linguistic environment. In this thesis, I have shown that there are a multiplicity of statistical cues in children's linguistic input, and that children and adults build upon prior knowledge when processing and learning new linguistic material. That is, establishing long-term representations of statistically regular patterns sets the learner onto a path further into their language. However, such knowledge could also be detrimental in

contexts where prior expectations are misleading. Overall, this thesis shows that speakers are exquisitely attuned to their linguistic input, showing that SL is a core component of the toolkit humans use to acquire and use language.

References

Abed Ibrahim, L., Hamann, C., & Öwerdieck, D. (2018). *Identifying specific language impairment (SLI) across different bilingual populations: German sentence repetition task (SRT): Vol. Proceedings of the 42nd annual Boston University Conference on Language Development (BUCLD)* (A. B. Bertolini & M. J. Kaplan, Eds.; pp. 1–14). Cascadilla Press.

Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin*, *135*(1), 50–68. https://doi.org/10.1037/a0014411.

Alexander, E., Van Hedger, S. C., & Batterink, L. J. (2022). Learning words without trying: Daily second language podcasts support word-form learning in adults. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02190-1.

Alonso, M. A., Fernandez, A., & Díez, E. (2011). Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods*, *43*(2), 449–458. https://doi.org/10.3758/s13428-011-0062-3.

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990. https://doi.org/10.1111/desc.12071.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273. https://doi.org/10.1017/S030500091400049X.

Antovich, D. M., & Graf Estes, K. (2018). Learning across languages: Bilingual experience supports dual language statistical word segmentation. *Developmental Science*, *21*(2). https://doi.org/10.1111/desc.12548.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160058. https://doi.org/10.1098/

rstb.2016.0058.

Arnon, I. (2019). Statistical learning, implicit learning, and first language acquisition: A critical evaluation of two developmental predictions. *Topics in Cognitive Science*, *11*(3), 504–519. https://doi.org/10.1111/tops.12428.

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68–81. https://doi.org/10.3758/s13428-019-01205-5.

Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language*, *48*(5), 937–958. https://doi.org/10.1017/S0305000921000386.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063.

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Psychology Press.

Audacity Team. (2018). *Audacity(R): Free audio editor and recorder* (2.4.2). https://audacityteam.org/.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear

mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Batterink, L. J. (2020). Syllables in sync form a link: Neural phase-locking reflects word knowledge during language learning. *Journal of Cognitive Neuroscience*, *32*(9), 1735–1748. https://doi.org/10.1162/jocn_a_01581.

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes*, *21*(1–3), 2–24. https://doi.org/10.1080/01690960400001721.

Benware, W. A. (1980). Zum Fremdwortakzent im Deutschen. *Zeitschrift Für Dialektologie Und Linguistik*, *47*(3), 289–312.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109.

Bertels, J., Destrebecqz, A., & Franco, A. (2015). Interacting effects of instructions and presentation rate on visual statistical learning. *Frontiers in Psychology*, *6*. Scopus. https://doi.org/10.3389/fpsyg.2015.01806.

Black, A., & Bergmann, C. (2017). *Quantifying infants' statistical word segmentation: A meta-analysis*. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci)*, 124–129. http://hdl.handle.net/11858/00-001M-0000-002D-DEDE-0.

Boll-Avetisyan, N. (2018). The role of phonological structure in speech segmentation by infants and adults: A review and methodological considerations. In C. Ulbrich, A. Werth, & R. Wiese (Eds.), *Empirical approaches to the phonological structure of words*. Walter de Gruyter.

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451–459. https://doi.org/10.1111/j.0956-7976.2005.01556.x.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x.

Braches, H. (1987). *German word stress and lexical phonology* [Doctoral Thesis]. National Library of Canada.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125. https://doi.org/10.1016/S0010-0277(96)00719-6.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. https://doi.org/10.1016/S0010-0277(01)00122-6.

Bulgarelli, F., & Weiss, D. J. (2016). Anchors aweigh: The impact of overlearning on entrenchment effects in statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1621–1631. https://doi.org/10.1037/xlm0000263.

Bull, D., Eilers, R. E., & Oller, D. K. (1984). Infants' discrimination of intensity variation in multisyllabic stimuli. *The Journal of the Acoustical Society of America*, *76*(1), 13–17. https://doi.org/10.1121/1.391110.

Bull, D., Eilers, R. E., & Oller, D. K. (1985). Infants' discrimination of final syllable fundamental frequency in multisyllabic stimuli. *The Journal of the Acoustical Society of America*, *77*(1), 289–295. https://doi.org/10.1121/1.392270.

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. https://doi.org/10.1006/cogp.1997.0649.

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early language experience in a Tseltal Mayan village. *Child Development*. https://doi.org/10.1111/cdev.13349.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/

10.1037/0033-295X.113.2.234.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468–481. https://doi.org/10.1111/tops.12332.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268. https://doi.org/10.1080/016909698386528.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X.

Conway, C. M., Baurnschmidt, A., Huang, S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009.

Cowan, N. (2016). Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on Psychological Science*, *11*(2), 239–264. https://doi.org/10.1177/1745691615621279.

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass*, *7*(3), 157–170. https://doi.org/10.1111/lnc3.12015.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*(2), 260–274. https://doi.org/10.1080/17470210902888809.

Cunillera, T., Laine, M., & Rodríguez-Fornells, A. (2016). Headstart for speech segmentation: A neural signature for the anchor word effect. *Neuropsychologia*, *82*, 189–199. https://doi.org/10.1016/j.neuropsychologia.2016.01.011.

Cunillera, T., Toro, J. M., Sebastián-Gallés, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research*, *1123*(1), 168–178. https://doi.org/10.1016/j.

brainres.2006.09.046.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, *96*(3), 233–262. https://doi.org/10.1016/j.cognition.2004.08.005.

Cutler, A. (1996). Prosody and the word boundary problem. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 87–99). Lawrence Erlbaum Associates Inc.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3), 133–142. https://doi.org/10.1016/0885-2308(87)90004-0.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121.

Dal Ben, R., Souza, D. de H., & Hay, J. F. (2021). When statistics collide: The use of transitional and phonotactic probability cues to word boundaries. *Memory & Cognition*, *49*, 1300–1310. https://doi.org/10.3758/s13421-021-01163-4.

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, *92*(2), 609–625. https://doi.org/10.1111/cdev.13511.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* Teachers College, Columbia University (Original work published 1885).

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, *36*(2), 202–225. https://doi.org/10.1006/jmla.1996.2483.

Eilers, R. E., Bull, D. H., Oller, D. K., & Lewis, D. C. (1984). The

discrimination of vowel duration by infants. *The Journal of the Acoustical Society of America*, *75*(4), 1213–1218. https://doi.org/10.1121/1.390773.

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the "Tabula" is anything but "Rasa:" What determines performance in the auditory statistical learning task? *Cognitive Science*, *46*(2), e13102. https://doi.org/10.1111/cogs.13102.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189).

Féry, C. (1986). Metrische Phonologie und Wortakzent im Deutschen. *Studium Linguistik*, *20*, 16–43.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation. *Experimental Psychology*, *62*(5), 346–351. https://doi.org/10.1027/1618-3169/a000295.

Frank, M. C., Arnon, I., Tily, H., & Goldwater, S. (2010). *Beyond transitional probabilities: Human learners impose a parsimony bias in statistical word segmentation*. *Proceedings of the 31st annual meeting of the Cognitive Science Society (CogSci)*, 760–765.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. https://doi.org/10.1016/j.cognition.2010.07.005.

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PLoS ONE*, *8*(1), e52500. https://doi.org/10.1371/journal.pone.0052500.

Friederici, A. D., Friedrich, M., & Christophe, A. (2007). Brain

responses in 4-month-old infants are already language specific. *Current Biology*, *17*(14), 1208–1211. https://doi.org/10.1016/j.cub.2007.06.011.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010.

Frost, R. L. A., Dunn, K., Christiansen, M. H., Gómez, R. L., & Monaghan, P. (2020). Exploring the "anchor word" effect in infants: Segmentation and categorisation of speech with and without high frequency words. *PLoS ONE*, *15*(12), e0243436. https://doi.org/10.1371/journal.pone.0243436.

Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C. F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology*, *120*, 101291. https://doi.org/10.1016/j.cogpsych.2020.101291.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: High frequency marker words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000683.

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty* [Unpublished Manuscript].

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (2019). The phonology of children's early words: Trends, individual variation, and parents' accommodation in child-directed speech. *Frontiers in Communication*, *4*(10), 1–11. https://doi.org/10.3389/fcomm.2019.00010.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science*, *33*(6), 1087–1116. https://doi.org/10.1111/

j.1551-6709.2009.01041.x.

Gerken, L. A. (1994). Young children's representation of prosodic phonology: Evidence from English-speakers' weak syllable productions. *Journal of Memory and Language*, *33*, 19–38.

Gerken, L. A., & MacInthosh, B. J. (1993). Interplay of function morphemes and prosody in early language. *Developmental Psychology*, *29*(3), 448–457.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. *Cognitive Psychology*, *57*(1), 56–74. https://doi.org/10.1016/j.cogpsych.2007.12.001.

Giegerich, H. J. (1985). Metrical phonology and phonological structure. German and English. *Cambridge Studies in Linguistics London*, *43*, 1–301.

Giroux, I., & Rey, A. (2009). Lexical and Sublexical Units in Speech Perception. *Cognitive Science*, *33*(2), 260–272. https://doi.org/10.1111/j.1551-6709.2009.01012.x.

Glinert, L. (1989). *The grammar of Modern Hebrew*. Cambridge University Press.

Goedemans, R., & van der Hulst, H. (2013a). Fixed stress locations. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/14.

Goedemans, R., & van der Hulst, H. (2013b). Weight-sensitive stress. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/15, Accessed on 2019-12-12.).

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 6.

Goyet, L., de Schonen, S., & Nazzi, T. (2010). Words and syllables in fluent speech segmentation by French-learning infants: An ERP study. *Brain Research*, *1332*, 75–89. https://doi.org/10.1016/j.brainres.2010.03.047.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power

analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504.

Grimm, A. (2006). *Intonational patterns and word structure in early child German.* (D. Bamman, T. Magnitskaia, & C. Zaller, Eds.; pp. 237–248).

Grimm, A. (2007). *The development of early prosodic word structure in child German: Simplex words and compounds.* [PhD Thesis]. University of Potsdam.

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication*, *2*, 16. https://doi.org/10.3389/fcomm.2017.00016.

Hamann, C., Chilla, S., Ruigendijk, E., & Abed Ibrahim, L. (2013). *A German sentence repetition task: Testing bilingual Russian/German children* [Poster]. COST Meeting, Krakow, Poland.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007.

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. https://doi.org/10.1016/j.cogpsych.2011.06.002.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). Blackwell.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22. https://doi.org/10.1016/j.cognition.2019.03.005.

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. https://doi.org/10.1016/j.infbeh.2009.03.004.

Höhle, B., Giesecke, D., & Jusczyk, P. W. (2001). Word segmentation in a foreign language: Further evidence for crosslinguistic strategies. *The Journal of the Acoustical Society of America*, *110*(5), 2687–2687. https://doi.org/10.1121/1.4777233.

Höhle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, *6*(2), 122–127. https://doi.org/10.1111/1467-7687.00261.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, *5*(3), 341–353. https://doi.org/10.1207/s15327078in0503_5.

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review*, *7*(3), 504–509. https://doi.org/10.3758/BF03214363.

Houston, D. M., Santelmann, L., & Jusczyk, P. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes*, *19*(1), 97–136. https://doi.org/10.1080/01690960344000143.

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018). *Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*, 1856–1861.

Isbilen, E. S., McCauley, S. M., & Christiansen, M. H. (2022). Individual differences in artificial and natural language statistical learning. *Cognition*, *225*, 105123. https://doi.org/10.1016/j.cognition.2022.105123.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). *Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 564–569.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020).

Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, *44*(7), e12848. https://doi.org/10.1111/cogs.12848.

Jessen, M. (1999). German. In H. van der Hulst (Ed.), *Word prosodic systems in the languages of Europe* (pp. 515–544). Mouton de Gruyter.

Johnson, E. K. (2016). Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics*, *2*(1), 391–412. https://doi.org/10.1146/annurev-linguistics-011415-040616.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567. https://doi.org/10.1006/jmla.2000.2755.

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2 GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225.

Jones, G. (2012). Why chunking should be considered as an explanation for developmental change before short-term memory capacity and processing speed. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00167.

Jones, G., Cabiddu, F., Andrews, M., & Rowland, C. F. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, 104232. https://doi.org/10.1016/j.jml.2021.104232.

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. *Developmental Science*, *17*(2), 298–310. https://doi.org/10.1111/desc.12111.

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: A computational model of the learning of new words. *Developmental Science*, *10*(6), 853–873. https://doi.org/10.1111/j.1467-7687.2007.00638.x.

Jones, G., & Macken, B. (2018). Long-term associative learning predicts verbal short-term memory performance. *Memory & Cognition*, *46*(2), 216–229. https://doi.org/10.3758/s13421-017-0759-3.

Jost, E., & Christiansen, M. H. (2017). Statistical learning as a domain-general mechanism of entrenchment. In H.-J. Schmid (Ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge* (1st ed., pp. 227–244). Walter de Gruyter.

Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development. *Developmental Science*, *15*(4), 463–473. https://doi.org/10.1111/j.1467-7687.2012.1144.x.

Jusczyk, P. W. (2002). How infants adapt speech-processing capacities to native-language structure. *Current Directions in Psychological Science*, *11*(1), 15–18. https://doi.org/10.1111/1467-8721.00159.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*(3), 675–687. https://doi.org/10.1111/j.1467-8624.1993.tb02935.x.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*(3), 159–207. https://doi.org/10.1006/cogp.1999.0716.

Kabak, B., Maniwa, K., & Kazanina, N. (2010). Listeners use vowel harmony and word-final stress to spot nonsense words: A study of Turkish and French. *Laboratory Phonology*, *1*(1), 207–224. https://doi.org/10.1515/labphon.2010.010.

Kaeding, F. W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Selbstverlag des Herausgebers. https://archive.org/details/hufigkeitswrter00kaedgoog/page/n8/mode/2up.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*(2–3), 199–225. https://doi.org/10.1016/0001-6918(94)90003-5.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184–193. https://doi.org/10.1111/cdev.12461.

Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, *200*, 104964. https://doi.org/10.1016/j.jecp.2020.104964.

Kidd, E., & Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics*, *6*(1), 319–340. https://doi.org/10.1146/annurev-linguistics-011619-030326.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, 01427237211066405. https://doi.org/10.1177/01427237211066405.

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64–83. https://doi.org/10.1016/j.jml.2018.03.005.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533.

Kunert, R., Fernandez, R., & Zuidema, W. (2011). *Adaptation in child directed speech: Evidence from corpora. Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, 112–119. https://hdl.handle.net/11245/1.354246.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. https://doi.org/10.1016/j.cognition.2013.02.002.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*, *118*(2), 265–273. https://doi.org/10.1016/j.cognition.2010.11.011.

Lalonde, C. E., & Werker, J. F. (1995). Cognitive influences on cross-language speech perception in infancy. *Infant Behavior and Development*, *18*(4), 459–475. https://doi.org/10.1016/0163-6383(95)90035-7.

Lany, J. (2014). Judging words by their covers and the company they keep: Probabilistic cues support word learning. *Child Development*, *85*(4), 1727–1739. https://doi.org/10.1111/cdev.12199.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, *19*(12), 1247–1252. https://doi.org/10.1111/j.1467-9280.2008.02233.x.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284–291. https://doi.org/10.1177/0956797609358570.

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. *Cognition*, *206*, 104492. https://doi.org/10.1016/j.cognition.2020.104492.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. *Cognition*, *223*, 105038. https://doi.org/10.1016/j.cognition.2022.105038.

Lehiste, I. (1970). *Suprasegmentals*. MIT Press.

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody picture vocabulary test: Revision IV (German adaptation)*.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329. https://doi.org/10.1111/j.1467-7687.2011.01079.x.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–246. https://doi.org/10.1016/j.cognition.2011.10.007.

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect*. Walter de Gruyter.

Lidz, J., & Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, *1*(1), 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236.

Lieven, E. V. M. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*(11), 2546–2556. https://doi.org/10.1016/j.lingua.2010.06.005.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219. https://doi.org/10.1017/S0305000996002930.

Lintfert, B. (2010). *Phonetic and phonological development of stress in German* [PhD Thesis]. University of Stuttgart.

Lüdecke, D., Makowski, D., Waggoner, P., Patil, I., & Ben-Shachar, M. S. (2021). *Package 'performance': Assessment of Regression Models Performance*. https://easystats.github.io/performance/.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295. https://doi.org/10.1017/S0305000900006449.

Majerus, S., Martinez Perez, T., & Oberauer, K. (2012). Two distinct

origins of long-term learning effects in verbal short-term memory. *Journal of Memory and Language*, *66*(1), 38–51. https://doi.org/10.1016/j.jml.2011.07.006.

Majerus, S., Van der Linden, M., Mulder, L., Meulemans, T., & Peters, F. (2004). Verbal short-term memory reflects the sublexical organization of the phonological language network: Evidence from an incidental phonotactic learning paradigm. *Journal of Memory and Language*, *51*(2), 297–306. https://doi.org/10.1016/j.jml.2004.05.002.

Marimon, M., Dos Santos, M., Nazzi, T., & Höhle, B. (2019). *Word segmentation cues: Prosody or statistics? Evidence from French* [Poster]. Workshop on Infant Language Development (WILD), Potsdam, Germany.

Marimon, M., Höhle, B., & Langus, A. (2022). Pupillary entrainment reveals individual differences in cue weighting in 9-month-old German-learning infants. *Cognition*, *224*, 105054. https://doi.org/10.1016/j.cognition.2022.105054.

Marimon Tarter, M. (2019). *Word segmentation in German-learning infants and German-speaking adults: Prosodic and statistical cues* [PhD Thesis]. University of Potsdam.

Mattys, S. L., & Bortfeld, H. (2016). Speech segmentation. In M. G. Gaskell & J. Mirković (Eds.), *Speech perception and spoken word recognition* (pp. 65–85). Psychology Press.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465–494. https://doi.org/10.1006/cogp.1999.0721.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001.

Matzinger, T., Ritt, N., & Fitch, W. T. (2021). The influence of different prosodic cues on word segmentation. *Frontiers in Psychology*, *12*, 458. https://doi.org/10.3389/fpsyg.2021.622042.

Matzinger, T., Schwarz, M., Ritt, N., & Fitch, W. T. (2019). *The effects*

of prosodic cues on word segmentation in an artificial language learning task [Talk]. International Cognitive Linguistics Conference (ICLC), Nishinomiya, Japan.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*, 1–51. https://doi.org/10.1037/rev0000126.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition*, *39*(6), 1085–1093. https://doi.org/10.3758/s13421-011-0074-3.

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. https://doi.org/10.1080/15475441.2011.609106.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56.

Miller, M. (1979). *The logic of language development in early childhood*. Springer-Verlag.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01514-0.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, *9*(1), 21–34. https://doi.org/10.1111/tops.12239.

Monaghan, P., Brand, J., Frost, R. L. A., & Taylor, G. (2017). *Multiple variable cues in the environment promote accurate and robust word learning*. Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci), 817–822. http://hdl.handle.net/11858/00-001M-0000-002E-670F-6.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of

sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564. https://doi.org/10.1017/S0305000909990511.

Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, *67*(S1), 14–39. https://doi.org/10.1111/lang.12221.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*(4), 911–936.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), Article 20170213. https://doi.org/10.1098/rsif.2017.0213.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development*, *21*(4), 779–784. https://doi.org/10.1016/S0163-6383(98)90044-3.

Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, *54*(3), 283–299. https://doi.org/10.1016/j.jml.2005.10.004.

Neurobehavioral Systems, Inc. (2014). *Presentation* (18.0). www.neurobs.com.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1207/s15516709cog1401_2.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1–32. https://doi.org/10.1016/0010-

0285(87)90002-8.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1209–1228.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225–237. https://doi.org/10.1016/j.jml.2005.02.011.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. https://doi.org/10.1016/j.cognition.2012.10.008.

Page, M. P. A., Cumming, N., Norris, D., McNeil, A. M., & Hitch, G. J. (2013). Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, *69*(4), 506–526. https://doi.org/10.1016/j.jml.2013.07.001.

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173.

Pearl, L. (2021). Theory and predictions for the development of morphology and syntax: A Universal Grammar + statistics approach. *Journal of Child Language*, *48*(5), 907–936. https://doi.org/10.1017/S0305000920000665.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–247. https://doi.org/10.1016/j.cognition.2009.07.011.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning. *Topics in Cognitive Science*, *11*(3), 520–535. https://doi.org/10.1111/tops.12403.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305. https://doi.org/10.3758/MC.36.7.1299.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge University Press.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108.

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.

Potter, C. E., Wang, T., & Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, *41*, 913–927. https://doi.org/10.1111/cogs.12473.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X.

Prolific. (2021). (2021.01-04). https://www.prolific.co.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Raneri, D., Von Holzen, K., Newman, R., & Bernstein Ratner, N. (2020). Change in maternal speech rate to preverbal infants over the first two years of life. *Journal of Child Language*, *47*(6), 1263–1275. https://doi.org/10.1017/S030500091900093X.

Raviv, L., & Arnon, I. (2018). The developmental trajectory of

children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, *21*(4), e12593. https://doi.org/10.1111/desc.12593.

Rigol, R. (2007). German Rigol corpus [Data set]. *TalkBank*. https://doi.org/10.21415/t50s34.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*(7), 1352–1389. https://doi.org/10.1111/j.1551-6709.2011.01188.x.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78.

RStudio Team. (2022). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114. https://doi.org/10.1111/1467-8721.01243.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x.

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, *20*(3), e12390. https://doi.org/10.1111/desc.12390.

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

Segal, O., Nir-Sagiv, B., Kishon-Rabin, L., & Ravid, D. (2009). Prosodic patterns in Hebrew child-directed speech. *Journal of Child Language*, *36*(3), 629–656. https://doi.org/10.1017/S030500090800915X.

Shafer, V., Gerken, L. A., Shucard, J., & Shucard, D. (1992). *'The' and the brain: An electrophysiological study of infants' sensitivity of English function morphemes*. *Proceedings of the 16th Annual Boston University Conference on Language Development (BUCLD)*.

Shafer, V. L., Shucard, D. W., Shucard, J. L., & Gerken, L. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language, and Hearing Research*, *41*(4), 874–886. https://doi.org/10.1044/jslhr.4104.874.

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, *11*(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x.

Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy*, *10*(2), 187–198. https://doi.org/10.1207/s15327078in1002_5.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, *42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160059. http://dx.doi.org/10.1098/rstb.2016.0059.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical

learning performance. *Cognition*, *177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001.

Sigurd, B., Eeg-Olofsson, M., & van Weijer, J. (2004). Word length, sentence length and frequency—Zipf revisited. *Studia Linguistica*, *58*(1), 37–52. https://doi.org/10.1111/j.0039-3193.2004.00109.x.

Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M. G., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01925.

Smalle, E. H. M., Muylle, M., Szmalec, A., & Duyck, W. (2017). The different time course of phonotactic constraint learning in children and adults: Evidence from speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1821–1827. https://doi.org/10.1037/xlm0000405.

Smalle, E. H. M., Page, M. P. A., Duyck, W., Edwards, M., & Szmalec, A. (2018). Children retain implicitly learned phonological sequences better than adults: A longitudinal study. *Developmental Science*, *21*(5), e12634. https://doi.org/10.1111/desc.12634.

Spring, D. R., & Dale, P. S. (1977). Discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research*, *20*(2), 224–232. https://doi.org/10.1044/jshr.2002.224.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech*, *65*(1), 3–27. https://doi.org/10.1177/0023830920979016.

Stärk, K., Kidd, E., & Frost, R. L. A. (2023). Close encounters of the word kind: Attested distributional information boosts statistical learning. *Language Learning*, *73*(2), 341–373. https://doi.org/

10.1111/lang.12523.

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition*, *179*, 23–36. https://doi.org/10.1016/j.cognition.2018.06.002.

Szmalec, A., Page, M. P. A., & Duyck, W. (2012). The development of long-term lexical representations through Hebb repetition learning. *Journal of Memory and Language*, *67*(3), 342–354. https://doi.org/10.1016/j.jml.2012.07.001.

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*, 1–8. https://doi.org/10.1186/1471-2202-10-21.

Thiessen, E. D., Onnis, L., Hong, S.-J., & Lee, K.-S. (2019). Early developing syntactic knowledge influences sequential statistical learning in infancy. *Journal of Experimental Child Psychology*, *177*, 211–221. https://doi.org/10.1016/j.jecp.2018.04.009.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706.

Tippmann, J. (2015). *Neuronale Korrelate der Verarbeitung phonotaktischer und prosodischer Cues bei drei Monate alten Säuglingen – Eine Studie mit elektrophysiologischen Signalen* [Bachelor thesis]. University of Leipzig.

Tippmann, J., Pötzsch, A., Vignotto, M., Obrig, H., & Rossi, S. (2015). *Verarbeitung phonotaktischer und prosodischer Hinweisreize bei dreimonatigen Säuglingen mittels Elektroenzephalografie (EEG) und funktioneller Nahinfrarotspektroskopie (fNIRS)* [Poster]. Jahrestagung der Gesellschaft für Aphasieforschung und -behandlung (GAB), Innsbruck, Austria.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Toro, J. M., Pons, F., Bion, R. A. H., & Sebastián-Gallés, N. (2011). The

contribution of language-specific knowledge in the selection of statistically-coherent word candidates. *Journal of Memory and Language*, *64*(2), 171–180. https://doi.org/10.1016/j.jml.2010.11.005.

Trecca, F., McCauley, S. M., Andersen, S. R., Bleses, D., Basbøll, H., Højen, A., Madsen, T. O., Ribu, I. S. B., & Christiansen, M. H. (2019). Segmentation of highly vocalic speech via statistical learning: Initial results from Danish, Norwegian, and English. *Language Learning*, *69*(1), 143–176. https://doi.org/10.1111/lang.12325.

Tremblay, P., Deschamps, I., Baroni, M., & Hasson, U. (2016). Neural sensitivity to syllable frequency and mutual information in speech perception and production. *NeuroImage*, *136*, 106–121. https://doi.org/10.1016/j.neuroimage.2016.05.018.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*(1), 71–86.

van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Developmental Science*, *12*(3), 419–425. https://doi.org/10.1111/j.1467-7687.2008.00788.x.

Vennemann, T. (1990). Syllable structure and simplex accent in Modern Standard German. *CLS*, *26*(2), 399–412.

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217–230. https://doi.org/10.1016/j.jml.2016.10.001.

Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science*, *38*(4), 757–774. https://doi.org/10.1111/cogs.12092.

Von Stutterheim, C. (2010). German Caroline corpus [Data set]. *TalkBank*. https://doi.org/10.21415/t5ns5s.

Vosoughi, S., & Roy, D. K. (2012). *A longitudinal study of prosodic exaggeration in child-directed speech. Speech Prosody, 6th International Conference. Speech Prosody Special Interest Group*

(SProSIG).

Wagner, K. R. (1974). *Die Sprechsprache des Kindes. Teil 1: Theorie und Analyse*. Präger.

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language*, *12*(2), 475–487. https://doi.org/10.1017/S0305000900006565.

Wagner, M. (2006). *First steps to communication: A pragmatic analysis*. Narr Verlag.

Wang, F. H., Zevin, J. D., Trueswell, J. C., & Mintz, T. H. (2020). Top-down grouping affects adjacent dependency learning. *Psychonomic Bulletin & Review*, *27*(5), 1052–1058. https://doi.org/10.3758/s13423-020-01759-y.

Wang, T., & Saffran, J. R. (2014). Statistical learning of a tonal language: The influence of bilingualism and previous linguistic experience. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00953.

Weber, C., Hahne, A., Friedrich, M., & Friederici, A. D. (2004). Discrimination of word stress in early infant perception: Electrophysiological evidence. *Cognitive Brain Research*, *18*(2), 149–161. https://doi.org/10.1016/j.cogbrainres.2003.10.001.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145.

West, G., Melby-Lervåg, M., & Hulme, C. (2021). Is a procedural learning deficit a causal risk factor for developmental language disorder or dyslexia? A meta-analytic review. *Developmental Psychology*, *57*(5), 749–770. https://doi.org/10.1037/dev0001172.

Wickham, H. (2017). *tidyverse: Easily install and load the 'Tidyverse'*. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome

to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.

Wiese, R. (1996). *The phonology of German*. Clarendon Press.

Wurzel, W. U. (1970). Der Fremdwortakzent im Deutschen. *Linguistics*, *8*(56), 87–108. https://doi.org/10.1515/ling.1970.8.56.87.

Wurzel, W. U. (1980). Der deutsche Wortakzent: Fakten—Regeln—Prinzipien: Ein Beitrag zu einer natürlichen Akzenttheorie. *Zeitschrift Für Germanistik*, *1*(3), 299–318.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456. https://doi.org/10.1016/j.tics.2004.08.006.

Zettersten, M., Potter, C. E., & Saffran, J. R. (2020). Tuning in to non-adjacencies: Exposure to learnable patterns supports discovering otherwise difficult structures. *Cognition*, *202*, 104283. https://doi.org/10.1016/j.cognition.2020.104283.

Zipf, G. K. (1935). *The psycho-biology of language*. MIT Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Research data management plan

Personal data

I collected the following personal data for this thesis:

- names of participants (Chapters 3 and 4) and Prolific IDs (Chapter 5), respectively
- age (Chapters 3 and 5) and birthdays (Chapter 4), respectively
- gender, language background, as well as information about language, speech, or hearing disorders (Chapters 3 to 5), and school classes (Chapter 4)
- audio recordings (Chapters 3 and 4) and language assessments (Chapter 4), respectively

It was necessary to collect these personal data to achieve the goals of my project. Names and Prolific IDs were required to arrange the testing sessions and coordinate the payment. The background information (language background as well as information about disorders) and the language assessments were used to exclude participants from analyses if they did not meet the studies' inclusion criteria. The age and birthdays as well as participants' gender were necessary for the descriptive statistics while the design of my experiments required participants to repeat speech sequences auditorily. I ensured that I did not collect more personal data than necessary for achieving the goals of my research project.

Privacy

Personal data were anonymised where possible. The participants' names (Chapters 3 and 4) and Prolific IDs (Chapter 5), respectively, as well as their birthdays (Chapter 4) were kept separately from the research data. They were matched with anonymous participant numbers in a password-protected file, to which only I have access.

As it is not possible to fully anonymise the audio recordings (Chapters 3 and 4), these are stored under restricted access in the MPI

Archive – not publicly available. Anonymous transcriptions of the audio recordings are publicly available on OSF and at the MPI Archive.

The informed consent forms (Chapters 3 and 4) are stored in a secured cupboard by the lab manager of the Language Development Department at the Max Planck Institute for Psycholinguistics. I have no longer access to these forms.

Ethical approval and informed consent

There was a blanket ethical approval for the studies in Chapters 3 to 5 granted to the Language Development Department by the Ethics Committee of the Faculty of Social Sciences at Radboud University (ECSW2017-3001-474 Manko-Rowland;Language Development).

The participants in Chapter 3 were either registered at the MPI database or recruited at Radboud University and via social media, respectively. They registered for the study via the database or contacted the experimenter via email. Prior to their participation, the participants received an informed consent sheet via email. They could withdraw from both the study and the MPI database at any time.

The participants in Chapter 4 were recruited at their school. Parents of all second- and third-graders were sent informed consent forms. Children whose parents signed the informed consent form were asked whether they wanted to participate in the experiment. The experiment was explained to them by the experimenter. Both parents and children could withdraw their consent at any time.

The participants in Chapter 5 were recruited via the online recruitment platform Prolific. They registered for the study online and received an informed consent as part of the online procedure prior to their participation in the experiment. They could withdraw their consent at any time.

Data storage

The research data and analysis scripts of Chapters 3 to 5 are stored in the MPI Archive with varying access levels, depending on the sensitivity of the data. Anonymised data and analysis scripts are additionally shared on the Open Science Framework (OSF). All relevant links can be found below.

- The whole project available at the MPI Archive: https://hdl.handle.net/1839/40ba3498-63c7-48ec-b909-2aed2f28355a
- Chapter 2: https://osf.io/vpdu6/
- Chapter 3: https://osf.io/4dsmy/
- Chapter 4: https://osf.io/t5qf4/
- Chapter 5: https://osf.io/eq7xk/

English summary

How do infants learn language? It may look easy because they seem to learn their mother tongue without difficulties, but there are challenges. For instance, the language input infants receive does not contain any obvious cues to indicate word boundaries, but children still manage to build a broad vocabulary before they start school. In this thesis, I investigated three important questions about human language development: (1) Which linguistic cues can infants use to learn their first words? (2) Do we use patterns from our mother tongue, which we acquired as infants, to learn another language as we get older? and (3) Do we exploit various linguistic cues to help us learn and understand new languages?

In the first part of my thesis, I investigated linguistic cues that infants could potentially use to find words in their speech input. When you listen to an unfamiliar language it sounds like a constant stream and it is difficult to tell where a word starts and where it ends. Infants, however, seem to identify words in this constant speech stream without difficulties. There are a variety of linguistic cues that could help the infants find the words, such as words always being stressed on the first syllable, therefore marking the beginning of words in a language. I analysed transcripts of the natural speech that infants acquiring German would typically hear from their parents and found that the cues I examined were all available in the speech, such that German-acquiring infants could exploit them to segment their input. Word stress was the most reliable indicator of word boundaries in German, with almost all words being stressed on the first syllable. This means that German-acquiring infants could identify the beginning of a word by paying attention to the stressed syllables. This cue most likely interacts with the other cues in the input to aid the acquisition of the language, such as how often two syllables occur in combination. If, for example, the syllables "ba" and "by" frequently occur together our brain implicitly notices this and memorises them as the unit "baby". Other cues were, for example, word frequency and words occurring in isolation. That is, if a word occurs very frequently (such as "the") or in isolation (such as "yes") it stands out and our brain memorises it more easily.

Next, I conducted experiments with German adults and seven- to nine-year-old children to test whether their ability to detect patterns in an artificial "alien" language was enhanced when these sequences were statistically similar to German. This way, I could explore whether people transfer the linguistic patterns in their mother tongue when learning a new (in this case, made-up) language. For example, in German some syllable combinations are very frequent, such as "ge fa". Other combinations are not very frequent, such as "fa zu". If we use patterns from our mother tongue when learning new languages, languages with syllable combinations similar to those in our mother tongue would be easier to learn than languages with different syllable combinations. In these studies, the participants would listen to sequences from the alien language and then repeat them out loud as accurately as possible. I tested the participants on three different types of sequences. The first were *naturalistic* sequences, which were designed to include frequent German syllable combinations, like "ge-fa mi-nu mo-ti". These were compared against *non-naturalistic* sequences, which were built from syllable combinations that are not very common in German, like "fa-zu nu-ver ti-pa". Finally, there were some foil sequences with randomly scrambled syllables, like "fa ge mo mi ver nu". Because of their randomness, these were supposed to be very difficult for the participants to learn compared to the naturalistic and non-naturalistic sequences. Apart from the difficulty of the foil sequences, the studies showed that both adults and children were better at repeating naturalistic sequences than non-naturalistic sequences. This suggests that adults and children store linguistic information of their mother tongue, such as syllable combinations, in their memory and use this knowledge when processing new language input.

I showed that participants will intuitively draw on their knowledge of linguistic patterns to process new language input; but when learning a new language, adults – like infants – can also use other cues to identify words in the speech stream. For example, if some words occur very frequently in the input, you may notice and learn them more easily. Evidence from previous studies indeed suggests that it may be easier to

identify the words of an artificial language when some of these words are repeated very frequently. I conducted three more experiments to explore whether additionally increasing the frequency of certain words in the alien language helps participants learn the language. I expected that participants would benefit from both their prior knowledge of linguistic patterns (i.e., syllable combinations found in German, such as in the naturalistic word "ge fa") and the frequent repetition of some words in the alien language. Unexpectedly, I observed the *opposite* pattern; prior knowledge of linguistic patterns in the mother tongue as well as variations in word frequency hindered participants' learning of the alien language. I ran two follow-up experiments which provided evidence that this unexpected effect was not driven by the characteristics of the alien language, but they could not explain these surprising effects. Collectively, this suggests that language knowledge and word frequency can affect language learning, but the direction of the results are difficult to reconcile with past research.

In this thesis, I found that (1) language contains several cues that infants could use to learn their first words. In German, this includes word stress, frequent syllable combinations, frequent words, and words occurring in isolation. (2) I also found that we appear to memorise syllable co-occurrences in our mother tongue and use this information to process subsequent language input. (3) However, more research is needed to disentangle how different sources of linguistic information may work together to help us learn new languages.

Nederlandse samenvatting

Hoe leren kinderen taal? Het lijkt alsof dit heel gemakkelijk gaat en dat kinderen ogenschijnlijk zonder veel problemen hun moedertaal lijken te leren, maar dat betekent niet het niet moeilijk is! Bedenk bijvoorbeeld dat de taal die kinderen horen een constante stroom van spraak is, en dat de afzonderlijke woorden in die spraakstroom nauwelijks te onderscheiden zijn. Toch slagen kinderen erin om een brede woordenschat op te bouwen voordat ze naar de basisschool gaan. Hoe kan dat? In dit proefschrift heb ik drie belangrijke vragen over de menselijke taalverwerving onderzocht: (1) Welke taalkundige informatie kunnen kinderen uit de spraakstroom van hun ouders benutten om hun eerste woorden te leren? (2) Gebruiken we patronen uit onze moedertaal, die we als kind hebben verworven, om een andere taal te leren als we ouder zijn? en (3) Welke aspecten van onze taalkennis gebruiken we om nieuwe talen te verwerken en te leren?

In het eerste deel van mijn proefschrift onderzocht ik verschillende taalaspecten die kinderen mogelijk benutten om afzonderlijke woorden te onderscheiden in de spraakstroom van hun ouders. Het is je vast bekend dat als je een onbekende taal hoort, het als een constante klankenstroom klinkt. Als je de taal niet spreekt, is het moeilijk te zeggen waar het woord begint en waar het eindigt. Kinderen leren echter losse woorden te herkennen in deze spraakstroom zonder enige instructie. Er zijn verschillende aspecten die hen kunnen helpen om de losse woorden te vinden. Zo is het bijvoorbeeld zo dat in sommige talen, zoals het Duits en Nederlands, de klemtoon vaak op de eerste lettergreep van het woord valt. Een beklemtoonde lettergreep geeft dus vaak het begin van een woord aan. In mijn onderzoek analyseerde ik transcripties van opnames van pratende Duitsers. Zo kreeg ik een goed beeld van de spraak die kinderen van Duitssprekende ouders normaal om zich heen horen. Vervolgens kon ik onderzoeken welk taalaspecten voorkwamen in het Duits en door Duitse kinderen gebruikt konden worden om losse woorden in het Duits te herkennen en te leren. Het bleek dat er in het Duits verschillende taalaspecten aanwezig waren die kinderen kunnen benutten om losse woorden in de spraakstroom te identificeren. Beklemtoning gaf de meeste informatie over of iets in de

spraakstroom een los woord was of niet. Kinderen die Duits leren kunnen dus het begin van een woord het beste herkennen door te letten op de beklemtoonde lettergrepen, want die geven vaak het begin van een woord aan. Dit taalaspect speelt waarschijnlijk samen met andere taalaspecten een rol in de taalverwerving van kinderen. Een ander taalaspect dat bijvoorbeeld een rol kan spelen bij taalverwerving is hoe vaak bepaalde combinaties van lettergrepen in een taal voorkomen. Als lettergrepen samen vaak voorkomen, is het waarschijnlijk dat die samen één woord vormen. Ook wanneer een klank heel vaak voorkomt (bijvoorbeeld "de"), of als een klank vaak in z'n eentje voorkomt, zijn dat aanwijzingen dat het hier om een los woord gaat (bijvoorbeeld "ja").

Vervolgens voerde ik experimenten uit met Duitse volwassenen en zeven- tot negenjarige kinderen. Ik testte of zij woorden konden leren in een kunstmatige "alientaal" wanneer de taalkundige patronen van de alientaal overeenkwamen met het Duits. Zo kon ik onderzoeken of mensen de taalkundige patronen in hun moedertaal gebruiken bij het leren van een nieuwe (in dit geval verzonnen) taal. In het Duits komen sommige lettergreepcombinaties heel vaak voor, zoals "ge fa". Andere combinaties komen niet zo vaak voor, zoals "fa zu". Als we taalpatronen uit onze moedertaal gebruiken bij het leren van nieuwe talen, zouden talen met lettergreepcombinaties die lijken op die in onze moedertaal makkelijker te leren zijn dan talen met andere, vreemde lettergreepcombinaties. In deze onderzoeken luisterden de deelnemers naar opnames van een stroom lettergrepen in de alientaal. Ze moesten deze vervolgens zo nauwkeurig mogelijk hardop herhalen. De deelnemers kregen drie verschillende soorten lettergreepreeksen te horen. Sommigen waren *naturalistische* reeksen. Deze bevatten lettergreepcombinaties die in het Duits vaak voorkomen, zoals "ge-fa mi-nu mo-ti". Deze werden vergeleken met *niet-naturalistische* reeksen met lettergreepcombinaties die niet vaak voorkomen in het Duits, zoals "fa-zu nu-ver ti-pa". Als laatste waren er reeksen met willekeurig door elkaar gehusselde lettergrepen, zoals "fa ge mo mi ver nu". Vanwege hun willekeurigheid zou het voor de deelnemers erg moeilijk moeten

zijn om deze reeksen na te zeggen vergeleken met de naturalistische en niet-naturalistische lettergreepreeksen die in meer of mindere mate de regels van het Duits volgen. Ik concludeerde dat zowel volwassenen als kinderen beter waren in het herhalen van naturalistische reeksen dan niet-naturalistische reeksen. De deelnemers vonden het herhalen van de willekeurige lettergreepreeksen het moeilijkst. Deze experimenten suggereren dat volwassenen en kinderen taalkundige patronen van hun moedertaal, zoals lettergreepcombinaties, opslaan in hun geheugen en deze kennis gebruiken bij het verwerken van een andere, nieuwe taal.

In mijn experimenten toonde ik dus aan dat mensen onbewust gebruik maken van hun kennis van taalkundige patronen van hun moedertaal, zoals lettergreepcombinaties, om informatie in een andere, nieuwe taal te verwerken. Maar bij het leren van een nieuwe taal kunnen volwassenen – net als kinderen – ook andere taalkennis gebruiken om losse woorden in de spraakstroom te identificeren. Als bepaalde woorden bijvoorbeeld heel vaak voorkomen in een spraakstroom, valt dat op. Hierdoor kan je ze gemakkelijker herkennen en leren. Eerdere studies suggereren inderdaad dat mensen makkelijker losse woorden in een kunstmatige taal herkennen als deze woorden zeer vaak herhaald worden. Ik voerde nog drie experimenten uit om te onderzoeken of het verhogen van de frequentie van bepaalde woorden in de alientaal deelnemers zou helpen om losse woorden in de spraakstroom te leren. Ik verwachtte dat deelnemers zowel zouden profiteren van hun kennis van taalkundige patronen die overeenkwamen in het Duits en de alientaal (d.w.z. naturalistische lettergreepcombinaties zoals "ge fa") als van de frequente herhaling van sommige woorden in de vreemde taal. Onverwacht zag ik het juist het tegenovergestelde patroon: kennis van taalkundige patronen in de moedertaal en variaties in woordfrequentie belemmerden het leren van de vreemde taal. Ik voerde twee vervolgexperimenten uit om te controleren of het onverwachte effect niet werd veroorzaakt door de kenmerken van de kunstmatige alientaal, maar dat bleek niet het geval. Deze experimenten suggereren dat taalkennis van de moedertaal en woordfrequentie het leren van talen kunnen beïnvloeden. De richting

van de effecten die ik observeerde is echter nog moeilijk te rijmen met resultaten uit eerder onderzoek.

In dit proefschrift toonde ik aan dat (1) talen verschillende kenmerken hebben die kinderen kunnen benutten om losse woorden in de spraakstroom te identificeren en zo hun eerste woorden te leren. In het Duits zijn dit onder andere de beklemtoning, de frequentie van woorden, frequentie waarmee bepaalde lettergrepen samen voorkomen om één woord te vormen, en of bepaalde woorden voornamelijk in hun eentje voorkomen. (2) Ik heb ook aangetoond dat we kennis over taalkundige patronen in onze moedertaal – in dit geval lettergreepcombinaties – gebruiken bij het verwerken van een andere, nieuwe taal. (3) Er is echter meer onderzoek nodig om te ontrafelen hoe we verschillende aspecten van onze taalkennis gebruiken om nieuwe talen te leren.

Translated from English by Merel Wolf

Deutsche Zusammenfassung

Wie erlernen Kinder Sprache? Es sieht vielleicht einfach aus, da sie ihre Muttersprache offenbar ohne Schwierigkeiten erwerben, aber es gibt durchaus Herausforderungen. Beispielsweise enthält die Sprache, die Kinder hören, keine offensichtlichen Hinweise darauf, wo ein Wort endet und das nächste beginnt. Dennoch erwerben Kinder bis zum Schulanfang ein breites Vokabular. In dieser Doktorarbeit habe ich drei wichtige Fragen zum Spracherwerb untersucht: (1) Welche sprachlichen Hinweisreize können Kinder benutzen, um ihre ersten Wörter zu lernen? (2) Benutzen wir Strukturen aus unserer Muttersprache, die wir als Kinder gelernt haben, wenn wir später eine weitere Sprache lernen? und (3) Benutzen wir verschiedene sprachliche Hinweisreize, um neue Sprachen zu verstehen und zu lernen?

Im ersten Teil meiner Doktorarbeit habe ich verschiedene sprachliche Hinweisreize untersucht, die Kinder eventuell benutzen, um ihre ersten Wörter im Sprachstrom ihrer Eltern aufzuspüren. Man kann es sich so vorstellen, als würde man einer unbekannten Sprache zuhören. Eine unbekannte Sprache klingt wie ein konstanter Strom, in dem es unmöglich erscheint, zu sagen, wo ein Wort beginnt und wo es endet. Kinder scheinen jedoch ohne Schwierigkeiten die Wörter in diesem konstanten Sprachstrom ausfindig machen zu können. In der Tat gibt es eine Reihe von sprachlichen Hinweisreizen, die Kindern dabei helfen, die Wörter aufzuspüren. Wenn in einer Sprache beispielsweise alle Wörter auf der ersten Silbe betont werden, würde eine betonte Silbe in dieser Sprache eindeutig auf den Anfang eines Wortes hinweisen. Für meine Arbeit habe ich Sprachaufnahmen untersucht, die den Alltag von Deutsch lernenden Kindern mit ihren Deutsch sprechenden Eltern widerspiegeln. Dabei hat sich gezeigt, dass alle untersuchten Hinweisreize tatsächlich in der Sprache vorhanden waren, sodass Deutsch lernende Kinder sie benutzen könnten, um den Sprachstrom in Wörter zu zerlegen. Betonung war dabei im Deutschen der zuverlässigste Hinweisreiz. Fast alle Wörter waren auf der ersten Silbe betont, was bedeutet, dass Deutsch lernende Kinder den Anfang eines Wortes dadurch erkennen können. Vermutlich wird dieser Hinweisreiz gemeinsam mit anderen Hinweisreizen beim Hören verwendet, um den

Spracherwerb zu erleichtern, beispielsweise wie oft zwei Silben zusammen auftreten. Wenn zum Beispiel die Silben „ba" und „by" häufig zusammen vorkommen, bemerkt unser Gehirn dies implizit und merkt sie sich als Einheit „Baby". Weitere Hinweisreize waren häufige Wörter und einzeln auftretende Wörter. Wenn ein Wort sehr häufig (wie „der") oder isoliert (wie „ja") vorkommt, fällt es uns auf und unser Gehirn merkt es sich leichter.

Anschließend habe ich Experimente mit deutschen Erwachsenen und sieben- bis neunjährigen Kindern durchgeführt. Dabei habe ich getestet, ob das Lernen einer künstlichen „Aliensprache" einfacher ist, wenn ihre Struktur dem Deutschen gleicht. Auf diese Weise konnte ich untersuchen, ob Menschen die sprachlichen Strukturen ihrer Muttersprache übertragen, wenn sie eine neue, in diesem Fall erfundene, Sprache lernen. Beispielsweise kommen im Deutschen manche Silbenkombinationen sehr häufig vor (wie „ge fa"). Andere Kombinationen sind nicht sehr häufig (wie „fa zu"). Wenn wir beim Erlernen einer neuen Sprache Strukturen aus unserer Muttersprache verwenden, wären Sprachen mit ähnlichen Silbenkombinationen wie in unserer Muttersprache leichter zu lernen als Sprachen mit unterschiedlichen Silbenkombinationen. In diesen Studien wurden den Teilnehmenden Sequenzen aus der Aliensprache vorgespielt und sie mussten diese so genau wie möglich nachsprechen. Dabei gab es drei verschiedene Arten von Sequenzen. Die erste Art waren *naturalistische* Sequenzen, welche Silbenkombinationen enthielten, wie sie häufig im Deutschen vorkommen (wie „ge-fa mi-nu mo-ti"). Diese wurden mit *nicht-naturalistischen* Sequenzen verglichen, die Silbenkombinationen enthielten, wie sie im Deutschen nicht sehr häufig vorkommen (wie „fa-zu nu-ver ti-pa"). Schließlich gab es noch Sequenzen, die willkürlich durcheinander gewürfelte Silben enthielten (wie „fa ge mo mi ver nu"). Aufgrund ihrer Willkürlichkeit sollten diese für die Teilnehmenden im Vergleich zu den naturalistischen und nicht-naturalistischen Sequenzen sehr schwer zu erlernen sein. Abgesehen von der Schwierigkeit des Erlernens letzterer Sequenzen zeigten die Studien, dass sowohl Erwachsene als auch Kinder naturalistische

Sequenzen besser wiederholen konnten als nicht-naturalistische Sequenzen. Dies deutet darauf hin, dass Erwachsene und Kinder sprachliche Informationen ihrer Muttersprache, wie etwa Silbenkombinationen, in ihrem Gedächtnis speichern und dieses Wissen bei der Verarbeitung neuen sprachlichen Inputs nutzen.

In den beiden Experimenten habe ich gezeigt, dass die Teilnehmenden intuitiv auf ihr Wissen über sprachliche Strukturen ihrer Muttersprache zurückgreifen, um neuen sprachlichen Input zu verarbeiten. Doch beim Erlernen einer neuen Sprache können Erwachsene – genau wie Kinder – auch andere Hinweisreize nutzen, um Wörter im Sprachstrom einer unbekannten Sprache aufzuspüren. Frühere Studien deuten etwa darauf hin, dass es einfacher ist, Wörter im Sprachstrom einer unbekannten Sprache zu identifizieren, wenn einige dieser Wörter sehr häufig wiederholt werden. Für meine Doktorarbeit habe ich drei weitere Experimente durchgeführt, um herauszufinden, ob es Teilnehmenden beim Erlernen einer fremden Sprache hilft, wenn in der „Aliensprache" manche Wörter häufiger vorkommen als andere. Meine Erwartung war, dass die Teilnehmenden sowohl von ihrem Vorwissen über sprachliche Strukturen in ihrer Muttersprache (d. h. Silbenkombinationen im Deutschen, wie zum Beispiel im naturalistischen Wort „ge fa") als auch von der häufigen Wiederholung einiger Wörter in der fremden Sprache profitieren würden. Allerdings zeigten die Ergebnisse das genaue Gegenteil; mit der Muttersprache übereinstimmende sprachliche Strukturen sowie verschiedene Worthäufigkeiten erschwerten den Teilnehmenden das Erlernen der fremden Sprache. Daraufhin habe ich zwei Folgeexperimente durchgeführt, die gezeigt haben, dass dieser unerwartete Effekt nicht auf die Eigenschaften der fremden Sprache zurückzuführen ist. Insgesamt zeigt dies, dass Sprachkenntnisse und Worthäufigkeit das Sprachenlernen beeinflussen können, jedoch lässt sich die Richtung der Beeinflussung vorerst nur schwer mit früheren Forschungsergebnissen in Einklang bringen.

In dieser Doktorarbeit habe ich gezeigt, dass (1) Sprache verschiedene Hinweisreize enthält, die Kinder benutzen können, um Wörter im Sprachstrom zu identifizieren und dadurch ihre ersten Wörter zu lernen. Dazu gehören im Deutschen beispielsweise die Betonung von Wörtern, häufig gemeinsam auftretende Silbenkombinationen, häufige Wörter und isoliert vorkommende Wörter. (2) Außerdem habe ich gezeigt, dass wir uns merken, welche Silben in unserer Muttersprache häufig gemeinsam auftreten und diese Information verwenden, um neuen sprachlichen Input zu verarbeiten. (3) Es bedarf allerdings weiterer Forschung, um herauszufinden, wie verschiedene sprachliche Hinweisreize in Kombination genutzt werden, um neue Sprachen zu erlernen.

Acknowledgements

It is a long way to submitting a thesis, along which I was lucky enough to meet countless invaluable people who supported me in reaching this goal. First of all, I would like to thank my supervisors, Rebecca, Evan, and Caro R., for setting out on this exciting journey with me and helping me to stay on course until the very end. **Caro R.**, we first met after the Nijmegen Lectures when you were just setting up the new department. Thank you for choosing me for your first cohort of PhD students. It was very exciting to see LaDD develop and become this great community of people who are not only passionate about language development but also unicorns and fairies, cakes, and so many other things. Thank you for providing us with plenty of opportunities at work but also for inviting us to Sinterklaas parties and pub quizzes, making us feel at home in a strange country.

Evan, thank you for choosing me as your PhD student and guiding me safely across the PhD ocean from one shore to the other, especially when there was no land in sight. Thanks for your kindness and patience, especially when I did not immediately see your point. As you know, I greatly admire your extraordinary knowledge of the literature and your scientific experience. Thanks for sharing your wisdom with me and for establishing our little core family within the department (i.e., FEK). And thank you so much for always responding within a heartbeat and being there to help me with every little problem. I greatly appreciate it.

Rebecca, I remember being intrigued by the clever design with which you came up for the paper I was asked to read for the interview, which was one of the reasons why I was already looking forward to working with you before my PhD even started. During my PhD, I did not only get the chance to use some of those clever designs but you also patiently showed me how to make my thoughts intelligible to the reader – thank you. I also highly appreciate that I could always rely on swift and detailed replies from you – with a nice, flattering British-English wording, which provided a good template on how to phrase reviews in a friendlier way. I know a lot of people were jealous of me having a supervisor who would show us their dog during (online) meetings and

think of you whenever coming across anything cat-related, were it bookmarks or sweets (Katjes). In short, thanks very much for everything.

I would also like to thank the members of my manuscript committee, **James**, **Paula**, and **Judith**, for happily accepting this additional task, carefully reading through the entire manuscript, and asking important questions to improve my work. Thanks also to the other members of my corona, **Jill**, **Louisa**, and **Hans Rutger**, for joining me at my defence. I am looking forward to discussing my research with all of you.

Special thanks go to all the people behind the scenes who equally supported me throughout my PhD journey. **Nanjo**, I think you were the reason why LaDD became our second family; thank you for all your kind words and for always being there for us. You will always have a special place in my heart as the first person who welcomed me to the MPI, who smiled at me every morning from the office opposite mine, who invited me to my first Dutch birthday party, and who always listened to and laughed with us; *heel erg bedankt*. **Ilse**, we did not have the chance to get to know each other as well but you surely are the new heart of LaDD. I admire the positive energy you send out, even at work, and enjoy getting to know you via WA photos from the distance. Thanks very much for your immediate replies even in busy times and all your support. Thanks also to **Ina**, **Edith**, **Martina**, **Evelyn**, **Carolin**, **Thy**, and **Lianne** for our conversations and/or your support in one way or another; your offices are always safe havens in the corridors of a busy building.

Pati, danke für deinen Enthusiasmus und deine entspannte Haltung, selbst in stressigen Zeiten. Leider war dein Büro recht abgelegen, aber eigentlich passt es sehr gut, dass du neben Elfen im Wald arbeitest. Danke, dass ich mich immer mit allem an dich wenden konnte und dass ich immer sicher sein konnte, dass du eine gute Lösung finden würdest. **Nienke**, thanks for your calm and kind personality. You always seemed

to take the stress out of a situation, which made it a pleasure to work with you. Thanks for your support and also for the great team work organising our Kletskoppen demo. **Kevin**, you are the heart of our research school, taking care of everyone and connecting people, while also securing the future of the IMPRS. Thanks for organising all the fun events, which gave us invaluable opportunities to network, and thanks for all your countless support over the years.

I would also like to thank the **Operations** and **TG** departments. Thank you all for your amazing work, which keeps the institute and its research running at all times. You make sure that the researchers can actually concentrate on their research, knowing that you truly love your work and do your utmost to support them. Thank you that we can always rely on you! Special thanks to **Karin** and **Meggie** for your expertise in anything literature-related and your passion for open science. Thanks to **Maarten** for your help programming in Presentation and to **Johan** for your help with the experiment setups. Thanks also to **Angela** and **Anique** for looking after us and helping us in any way and to **Luise** for making sure everything goes smoothly with the bills. And thanks to **Jeroen** for having a special eye on our department.

Many thanks to all the **LaDD** and **FEK** members. I learned a lot from you. It was truly inspiring to watch the different ways in which you approached and succeeded at your work. Thanks also for the fun LaDD outings, lunch breaks, and encounters at *the* table. Similarly, I would like to thank my fellow **committee members** organising the **lunch talks**, **IMPRS conference**, and **Kletskoppen**. I learnt a lot about collaboration, organisation, writing, and many more things with and from you.

A big thank you also to my amazing LaDD siblings: **Ingeborg**, thanks for being my office mate during the first half of my PhD when I was still finding my way into the whole process. Thanks for hosting my birthday party, your advice, and your kindness. **Shanthi**, thanks for being my office mate during the second half of my PhD when I was

already more stressed and exhausted. It was very refreshing to share the office with someone who was still very excited and energetic. Thanks for your positivity and the fun outings. **Merel W.**, you were such an inspiration with your wonderful personality and your efficient way of working. Thanks for the fun outings (e.g., Burgers' zoo or dinners) and all your support throughout the years. **Melis**, thanks so much for taking any risks to give me a surprise farewell gift. Thanks for the fun times together, your support, and your kindness. **Yevheniy**, thanks for your humour and the insightful conversations. And finally, thanks to **Clara**, **Jenni**, **Chih**, and **Eleni** for continuing our traditions and making sure that there is always a friendly LaDDie to come home to.

Thanks also to all the other **IMPRSive people** who embarked on this journey with me. I absolutely loved our community, all the breaks and evenings spent together, all the events and conversations, your support and the feeling of togetherness. Some of you deserve a special mention: **Federica**, you were the first person to join me at the MPI still before our journey really began. Your dedication, discipline, and perseverance are really inspiring. Thanks for all the time spent together, both organising things and having fun. **Lara**, you were the third in our guesthouse team. Thanks so much for all the outings (e.g., *schaatsen* or Jane Goodall) and your positivity, even in difficult situations, putting things back into perspective for me. **Chén**, thanks for all the fun evenings, especially my very first hotpot, the lunch breaks, outings, and great conversations. Time spend with you is never wasted. **Elly**, thanks for our dinners and walks as well as all the conversations. You are such a wonderful person. **Gaby**, thanks for your joyful and energetic personality as well as all our conversations and outings. Things always seem easier and more fun around you. **Joery**, thanks for introducing me to the beautiful nature around Nijmegen. I often went back to the hills and moors to relax after a stressful week. **Limor**, thanks for welcoming me into our IMPRSive family when I first arrived, the evenings spent together and our conversations. **Merel P.**, thanks for all your kindness and understanding as well as our conversations and the Dutch dinners.

Nico B., vielen Dank für deine Freundschaft und deine Unterstützung zu Beginn dieses schwierigen Projektes. Es ist wundervoll, wie du dich für so viele Dinge begeistern kannst und deine ganze Seele hineinsteckst. Es hat mir sehr geholfen, mich mit dir über die deutsche Sprache (d.h., den Korpus) oder kulturelle Unterschiede auszutauschen. So kam alles erst richtig in Gang.

Vielen herzlichen Dank meiner Praktikantin **Caro D. B.** für dein Interesse an meiner Arbeit sowie deine Unterstützung bei allen erdenklichen Aufgaben. Es war immer schön, mich mit dir auszutauschen. Ich hoffe, du hast trotz der Pandemie ein paar gute Einblicke erhalten. Thanks also to **Nico P.** and **Liz** for your help with my experiments. **Naomi**, thanks for drawing the alien for the story I told the children. And **Raquel**, thanks for drawing so many aliens for the cover until we found a meaningful version I liked.

Vielen Dank auch an meine früheren Supervisor, Antje, Tanja und Andrea. **Antje**, du begleitest mich nun schon länger als mein halbes Leben. Vielen Dank für dein ungebrochenes Interesse an meiner Person und meinem Werdegang. Du hast mir schon in der Schule wichtige Herangehensweisen für das Studium und das Schreiben von Papern an die Hand gegeben und ich habe deine Hinweise noch heute im Kopf, wenn ich einen Korpus analysiere oder ein Paper schreibe. Es ist immer schön, dich wiederzusehen und mich von deiner Energie und Effizienz inspirieren zu lassen. **Tanja**, vielen Dank für das Praktikum und die Betreuung meiner Bachelorarbeit. Von dir habe ich vieles über psychologische Methoden und Statistik gelernt sowie interessante erste Einblicke in das Leben von IMPRSive PhD Studenten erhalten. **Andrea**, danke dass du mir den ersten Schritt in die große, weite Welt so leicht gemacht hast. Du hast eine wunderbare Art, Leute zur Arbeit zu motivieren und gleichzeitig dafür zu sorgen, dass sie genügend Auszeiten nehmen. Deine ruhige, kompetente Art zu arbeiten und deine empathische Art im Umgang mit Mitarbeitern inspiriert mich bis heute.

Thanks also to my Dutch language group people: **Leslie**, thanks for great conversations about stats and life; **Olly**, thanks for super tasty British and German dinners with sides of deep conversations and postdoc advice; and **Ileana**, thanks for your German-American love.

Vielen Dank schließlich auch meinen Leipziger Freunden, dass ihr mich nicht nur nicht vergessen habt, sondern dass es vielmehr bei jedem Wiedersehen so ist, als wäre keine Zeit verstrichen. **Andreas** und **Andreas**, danke für eure Gastfreundschaft, eure Empathie, die tollen Unterhaltungen, Abendessen und Spieletreffen. **Felix**, danke auch für deine Gastfreundschaft, deine Offenheit, deine Verrücktheit sowie die tollen Unterhaltungen und Erlebnisse. **Isi**, danke für deine Sicht der Dinge, deine Gelassenheit, deine Empathie und Weisheit sowie deine Ablenkung mit Kindergeschichten. **Ana**, thanks for your love, openness, craziness, and courage. You are a true inspiration, and your positive energy is always contagious.

Very special thanks to my best friend, **Andrew**, I would not have come to this point without you. Thanks for being such a great friend. Thanks for making my life more fun but also, and maybe more importantly, for being there when it was less fun. Of course, I would also like to say thank you for everything you taught me. You did not only make me a better scientist but a better person.

Thanks to my wonderful paranymphs, Julia, Weng, and Teun. **Julia**, du bist praktisch meine PhD-Schwester. Wir haben zusammen angefangen und sind gemeinsam in die LaDD-Familie und alle Herausforderungen hineingewachsen. Danke für das Organisieren meiner Geburtstags- und Abschiedsfeier, danke für die gemeinsame Zeit beim Backen, Kochen, Organisieren, Austauschen, etc., und danke für deine immer offenen Ohren und permanente Bereitschaft zu helfen, wo es nur möglich war. Es war toll, dich immer nebenan zu wissen. **Weng**, we also met right at the beginning of my PhD, and little did I know how important you would become to me when we first spoke at your poster. Thank you so much for your love and support, thanks for

our wonderful days out – I can always trust that there will be plenty of amazing food when I am with you, and thanks for always being there and supporting me. *Maraming salamat.* **Teun**, we met half way through my PhD and you immediately filled it with more colours (or "puppy love"). Thanks for the fun times, inspiring conversations, and all your understanding. You gave me the Dutch experience of *schaatsen* and *fietsen* (the time in the forest, you might not even remember, but it was special to me) and making *oliebollen*. Thank you so much for your excitement for fun things and your support on less good days alike.

Liebe **Familie**, vielen Dank für euer Interesse an meiner Arbeit und eure Unterstützung bei dieser oder Ablenkung von dieser. Vielen Dank auch, dass ihr das Einreichen meiner Arbeit mit mir gemeinsam gefeiert und euch an meinen Erfolgen mit mir erfreut habt. Besonderer Dank gilt natürlich meinen Eltern, denen dieses Buch gewidmet ist. **Mama** und **Papa**, ihr habt mich von klein auf bestens auf diese schwierige Aufgabe vorbereitet, mir immer dabei geholfen, die Welt zu erforschen und zu ergründen. Während des PhDs, insbesondere während der Pandemie und während der letzten Phase, standet ihr mir noch einmal besonders tatkräftig zur Seite. Ohne euch wäre ich niemals so weit gekommen. Ich freue mich schon, diesen gemeinsamen Erfolg mit euch zu feiern.

Marcel, du meinst zwar, ich solle mir nicht die Mühe machen, noch einen weiteren Absatz zu schreiben, da wir uns praktisch erst nach meinem PhD kennen gelernt haben, aber ich danke dir von ganzem Herzen für deine Liebe, deine Unterstützung, deine Offenheit und alles andere. Danke, dass du mich für die letzten Hürden motiviert hast. Ich freue mich schon riesig, in Zukunft noch viele weitere mit dir gemeinsam zu nehmen.

Curriculum vitae

Katja Stärk was born in Leipzig, Germany, in 1992. She obtained her Bachelor's and Master's degree in General Linguistics from the University of Leipzig in 2013 and 2016, respectively. During this time, Katja completed several internships at the Max Planck Institute for Human Cognitive and Brain Sciences and worked as a teaching assistant in the Department of Linguistics at the University of Leipzig. She completed an Erasmus traineeship in the Department of Psychosocial Science at the University of Bergen, Norway, before joining the Language Development Department at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, as one of the department's first doctoral researchers in 2017.

Publications

Thesis chapters

Stärk, K., Kidd, E., & Frost, R. L. A. (2023). Close encounters of the word kind: Attested distributional information boosts statistical learning. *Language Learning*, *73*(2), 341–373. https://doi.org/10.1111/lang.12523.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). The effect of children's prior knowledge and language abilities on their statistical learning. *Applied Psycholinguistics*, *43*(5), 1045–1071. https://doi.org/10.1017/S0142716422000273.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word segmentation cues in German child-directed speech: A corpus analysis. *Language and Speech*, *65*(1), 3–27. https://doi.org/10.1177/0023830920979016.

Other publications

Parhammer*, S. I., Ebersberg*, M., Tippmann*, J., **Stärk*, K.**, Opitz, A., Hinger, B., & Rossi, S. (2019). The influence of distraction on speech processing: How selective is selective attention? *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2019)*, 3093–3097.

*shared first authorship