universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Neurological entrainment in an artificial grammar learning paradigm

–

Synchronisation neurologischer Prozesse beim Erlernen einer künstlichen Grammatik

verfasst von | submitted by

Hieronymus-Benedict Lanfer BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 867
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Allgemeine Linguistik: Grammatiktheorie und kognitive Sprachwissenschaft
Betreut von | Supervisor:	Univ.-Prof. Dr. Jutta Müller

Zusammenfassung

Vorherige Studien zum Erlernen künstlicher Grammatiken haben gezeigt, dass Kinder in der Lage sind, aufgrund der statistischen Eigenschaften eines Silbenstroms auf hierarchische Strukturen innerhalb des Stroms zu reagieren. Eine der offenen Fragen bei der Untersuchung, wie Erwachsene solche Grammatiken verarbeiten, ist, inwiefern das Erlernen künstlicher Grammatik allein durch statistische Eigenschaften der Grammatik begründet werden kann, und inwiefern bereits bestehendes linguistisches Vorwissen nötig ist, um das Erlernen einer künstlichen Grammatik durch Erwachsene zu ermöglichen. Die vorliegende Studie erforscht mit einer Elektroenzephalogramm- (EEG) und Verhaltensstudie die Reaktion gesunder Erwachsener beim Erlenen einer künstlichen Grammatik. Dabei handelt es sich um eine Grammatik, die ausschließlich auf statistischen Eigenschaften beruht und nicht-statistische Elemente wie Prosodie ausschließt. In dieser Studie wurden deutsche Muttersprachler*innen einem Silbenstrom ausgesetzt, der nicht-adjazente Dependenzen (NAD) beinhaltet und hierarchische Strukturen auf drei Ebenen (Silbe, Wort, Phrase) vorweist. Da diese hierarchischen Strukturen in regelmäßigen Intervallen auftreten, werden neurophysiologische Reaktionen der Partizipant*innen an der Frequenz dieser Intervalle erwartet. Während einer Zuhörphase wurden mittels EEG die neurophysiologischen Reaktionen der Partizipant*innen auf den Silbenstrom mit den darin enthaltenen hierarchischen Strukturen gemessen. Anschließend mussten die Partizipant*innen eine Verhaltensaufgabe ausführen, bei der sie Silbensequenzen mit intakter NAD von Silbensequenzen unterscheiden sollten, bei denen die NAD verletzt wurde. Die EEG-Daten wurden mittels Steady-State Auditory Evoked Potential (SSAEP) und Cluster-Based Permutation Test (CBPT) analysiert, während die Antworten in der Verhaltensaufgabe mittels t-Tests und ANOVA analysiert wurden. Die Ergebnisse von SSAEP und CBPT zeigen Reaktionen der Partizipant*innen an den Frequenzen aller drei Ebenen hierarchischer Strukturen. Die Verhaltensergebnisse zeigen, dass die Partizipant*innen signifikant über dem Zufallsniveau in der Lage waren, die Sequenzen mit intakter NAD von den Sequenzen mit verletzter NAD zu unterscheiden. Diese Ergebnisse liefern Belege dafür, dass statistische Informationen ausreichend sind, um das Lernen einer künstlichen Grammatik aus einem Silbenstrom zu ermöglichen. Allerdings schränkt die Struktur des verwendeten Silbenstroms die Interpretierbarkeit der Ergebnisse ein, weshalb mehr Forschung benötigt wird, um nähere Erkenntnisse darüber zu erlangen, inwiefern eine rein auf statistischem Lernen begründete künstliche Grammatik für Erwachsene zu erlernen ist und welche Aspekte der Grammatik die Reaktion der Partizipant*innen beeinflussen.

Abstract

Previous studies on artificial grammar learning have shown that children are able to respond to hierarchical structures within a syllable stream based on the statistical properties of that stream. An open question in the investigation of how adults process artificial grammars is to what extent the learning of artificial grammar can be justified solely by the statistical properties of the grammar, and to what extent pre-existing linguistic knowledge is necessary to enable the learning of an artificial grammar by adults. This thesis uses an electroencephalogram (EEG) and behavioural study to investigate the learning behaviour of healthy adults when learning an artificial grammar. This artificial grammar is based exclusively on statistical properties and excludes non-statistical elements such as prosody. In the present study, native German speakers were exposed to a syllable stream containing non-adjacent dependencies (NAD) and hierarchical structures on three levels (syllable, word, phrase). As these hierarchical structures occur at regular intervals, neurophysiological reactions of the participants at the frequency of the structures are expected. During a listening phase, the neurophysiological reactions of the participants to the syllable stream with the hierarchical structures it contains were measured using EEG. The participants then had to perform a behavioural task in which they had to distinguish syllable sequences with an intact NAD from syllable sequences in which the NAD was violated. The EEG data were analysed using steady-state auditory evoked potential (SSAEP) analysis and a cluster-based permutation test (CBPT), while the responses in the behavioural task were analysed using t-tests and ANOVA. The SSAEP and CBPT results show participant responses at the frequencies of all three levels of hierarchical structure. The behavioural results show that the participants were able to discriminate intact from violated sequences significantly above chance level. These results provide evidence that statistical information was sufficient to enable the learning of an artificial grammar from a syllable stream. However, the structure of the syllable stream used in the present study limits the generalizability of the results. More research is needed to gain further insight into the extent to which an artificial grammar based purely on statistical learning can be learned by adults and which aspects of the grammar influence participant responses.

Table of contents

1 Introduction 5
1.1 Historical and theoretical background of AGL 7
1.2 Modern AGL research 8
1.2.1 AGL at the word level 8
1.2.2 AGL beyond the word level 10
1.3 Neurophysiological investigations of adult AGL 14
1.4 Prior studies of the same design 17
1.5 The present study 21
2 Methods and materials 22
2.1 Participants 22
2.2 Experimental setup and procedure 23
2.3 Stimulus material 25
2.4 EEG setup and recording 28
2.5 Statistical methods for analysis 29
3 Results 30
3.1 EEG Results 30
3.2 Detailed SSAEP results 32
3.2.1 Response by grammar type 34
3.3 Behavioural results 36
4 Discussion 38
4.1 Main effect 39
4.2 Performance by grammar 41
4.3 Theoretical implications 47
4.3.1 Suppression effect 47
4.3.2 Simultaneous processing of hierarchical structures 49
4.4 Limitations 50
4.5 Connections to future research and follow-up studies 53
5 Conclusion 56
6 References 58
7 Appendix 61
7.1: Histograms of accuracy 61

1 Introduction

Segmentation is one of the fundamental problems of language processing. Anyone who has listened to a conversation in a foreign language will have realized that finding sentence boundaries, let alone individual words, is exceedingly difficult, despite the fact that native speakers can do it effortlessly in their own native language. This is because human speech occurs as a continuous speech stream (Lehiste, 1960) that is segmented based on language-specific segmentation cues. Segmenting the speech stream into meaningful units is critical for deriving any type of meaning from an utterance. While there are many different strategies employed by speakers of a language to segment the speech stream into units (cf. Gilbert et al., 2021, p. 1f; Tyler & Cutler, 2009, p. 367 for a short overview), knowledge of the internal structure of the specific language is required to apply those cues to a speech stream. For example, using phonotactic constraints to identify whether two adjacent sounds are part of the same word or part of different words requires knowledge of the specific phonotactic constraints of the target language. This then begs the question how that knowledge can be acquired in the first place, before a prospective speaker knows how to segment utterances. The answer to this question is one of the key points of contention in the debate on how much of language learning is purely statistical learning and how much is based on a facilitative language faculty unique to language (Frank & Bod, 2011; Hauser et al., 2002; Pinker, 1984; Saffran, 2002). Besides investigations into the strategies used by infants when and before they first acquire language, the last 30 years have seen an increased investigation into the ability of adults to acquire artificial languages.

Artificial grammar learning (AGL) describes the process of acquiring the rules underlying regularities in a set of linguistic stimuli which do not belong to any naturally occurring language, typically one that is purpose built for the investigation of a specific aspect of language. Because these stimuli are created independently of any pre-existing language, AGL paradigms have a unique advantage compared to other language learning tasks: they allow for complete control over the linguistic phenomena that are investigated in an experiment. It is therefore possible to isolate a specific aspect of language, like distal syntactic dependencies, and investigate the learning mechanisms that are specific to that aspect of language, without having to control for interference through the phonological, semantic, or even pragmatic aspects of a natural language. Through this property, AGL paradigms are ideally suited for foundational research into the basic nuts and bolts of language learning.

AGL has been demonstrated extensively in infants as one of the earliest skills that are acquired prior to primary language acquisition (Müller et al., 2018, pp. 5f). It is, however, much less clear to what degree this ability is retained into adulthood after a primary language has been acquired. AGL could offer insight into the hypothesis that syntax acquisition may behave

similarly to categorical perception in phoneme perception. There, once the internal structure of a primary language has been rudimentarily acquired, the ability to freely perceive gradual changes in phonemes is lost (cf. Kronrod et al., 2016; Kuhl, 1991; Liberman et al., 1957). Does a syntactic equivalent exist, where in the process of growing up, children lose some of the uninhibited ability of infants to process linguistic input freely and construct new syntactic structures out of it? And to what degree are statistical distributional cues sufficient to elicit these syntactic structures? These are some of the questions that adult AGL research may answer, giving deeper insight into the development of the ability to acquire new linguistic structures as life progresses.

One of the mechanisms that AGL is ideally suited to investigate in isolation is the non-adjacent dependency (NAD). It describes the phenomenon where one element of an utterance is in a direct relation with another element at a distal position in the utterance, meaning that some other material is intervening between the two dependent elements. These dependencies occur in almost every sentence. For example, even though the sentence “Peter sleeps” only contains two words, the agreement relation between the subject Peter and the person and tense marker -s spans over the root of *sleep*.

In semantics, an NAD occurs every time an anaphora or cataphora is made, e.g. when using pronouns to refer to a prior or following discourse referent. Here, the realisation of the pronoun is dependent on the co-indexed discourse referent, as its *φ*-features are copied onto the pronoun, even though the discourse referent may occur very far from the pronoun, even across sentence boundaries. An example would be “Peter is tired. He’ll fall asleep soon” where the *He* in the second sentence is co-referent and co-indexed with *Peter* from the first sentence.

In phonology, this is seen in cases of vowel harmony, where the phonological features of one vowel may influence the phonetic realisation of a following vowel, even with intervening consonants (or in some languages even transparent vowels) in between the harmonizing vowels. An example for this is Hungarian, where a word-initial +back vowel will spread its +back feature to subsequent vowels, even with intervening consonants between the vowels (cf. Finley, 2015, pp. 49f).

It is perhaps best seen in syntax, where a plethora of phenomena such as agreement, case marking, split verbs and movement all require interplay between two non-adjacent elements of a sentence. Because the NAD is such a fundamental requirement for the construction of a sentence, investigating how and under which circumstances adults process these dependencies is a key question in the understanding of human language processing. Therefore, the present study aims to investigate these capabilities, conducting an AGL study with EEG measurement and a behavioural task to examine the ability of adults to process an NAD in a continuous speech stream based only on distributional information of the syllables.

In order to give some background for the design decisions made in the present study, some historical and current background on previous studies is provided in the following sections.

1.1 Historical and theoretical background of AGL

AGL is a long-established method in psycholinguistic research. Reber (1967) first introduced implicit learning of a linguistic grammar in an artificial language as an experimental paradigm for the investigation of language learning processes in adults. In his experiment, he showed cards with single letters written on them to participants. Depending on which letter they saw on the previous card, the letter on the next card could be determined by a statistical process, with different probabilities given to different letters. Thus, a grammar is created, by first selecting which letter may follow which, and also statistically determining the likelihood of any given letter following another (cf. Reber, 1967, p. 856). While his experimental setup of showing cards with written letters to his participants is quite different from modern AGL research which focuses mainly on auditory stimuli, he first established three key aspects for AGL paradigms which are still used today:

First, he used linguistic stimuli which are able to be understood by the participants. Crucially, he does not use a foreign alphabet or an artificially created set of symbols as stimuli, even though the abstract grammatical operations which he investigates would have behaved in the same way using the Cyrillic alphabet, or a set of numbers, or simple dots and lines. This ensures that the stimuli are parsed in a comparable way to ordinary language, and not as a purely abstract informational system.[1] Subsequent research has focused even more on naturalistic language stimuli, moving away from the information theoretical approach that was often used in early linguistic theory (cf. Chomsky, 1957, 1965) and was still present in Reber (1967).

Second, he separates the task into a learning phase and a test phase. During the learning phase, the participants are exposed to novel stimuli and instructed to pay close attention to the stimuli to possibly identify a pattern. During the test phase, the participants are asked to perform a task which tests whether they identified the pattern. Although the shape of the learning phase in modern AGL research varies from single sentences separated by button press tasks (Ding et al., 2016, p. 14, experiment 7) to several minutes of uninterrupted exposure to repeating stimuli (e.g. Batterink & Paller, 2017; Buiatti et al., 2009; Thompson & Newport, 2007), the separation into learning phase and subsequent test phase has been retained.

Third, he used statistical properties to establish grammatical rules. Reber (1967, p. 856f) uses information theory and entropy in order to define which sentences are possible, and how much

[1] For a recent investigation of non-linguistic statistical learning using tones, see Moser et al. (2021).

information each additional symbol conveys. While modern research uses statistical transitional probabilities between stimulus units (e.g. Ding et al., 2016; Getz et al., 2018; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin 1996; Thompson & Newport 2007;), the use of transitional probabilities follows the same logic as information theory. Transitional probability (TP) describes the likelihood with which a specific stimulus may follow next, given the previous stimulus. For example, in the grammar of a hypothetical college student, in which the string “Tonight, I’m going to the…” may only be followed by the three equally likely options “gym”, “bar”, or “club”, the TP from the first string to any of the three options to complete the sentence is $^1/_3$. This means that, given the rules of that grammar and having already encountered the first part of the string up to “the”, one can predict with $^1/_3$ probability which word is going to come next given the previous input. This statistical measure is the basis of current day statistical learning studies, including Bulca et al. (2019) and Weyers et al. (2022), on which the present study is based.

1.2 Modern AGL research

The currently employed paradigm for conducting AGL research was introduced in a pair of papers by Saffran, Aslin and Newport (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Instead of letters written on cards as used by Reber (1967), they shift the focus to auditory stimuli, exposing participants to a continuous speech stream constructed out of individual syllables. As auditory learning is the primary mode of primary language acquisition (Kuhl, 2010) starting even before birth (Choi et al., 2017), this shift more closely matches the modality in which language is actually acquired. It also allows research into the AGL capabilities of infants, which are investigated in Saffran, Aslin, & Newport (1996).

1.2.1 AGL at the word level

As Jusczyk & Aslin (1995) showed, 7½-month-old (7½M) infants are capable of identifying individual words within fluent speech. Saffran, Aslin, & Newport (1996) expand on this finding, using synthesized CV-syllables to investigate the capabilities of 8M infants to differentiate familiarized tri-syllabic pseudowords from unfamiliar tri-syllabic pseudowords. In their study, a continuous speech stream of concatenated syllables is played to the infants for 2 minutes. This speech stream consists of sets of fixed triplets of syllables, meaning the same three syllables always follow each other. There are four different triplets which are distributed randomly across the speech stream, not allowing for any direct repetition of a triplet. These triplets form fixed pseudowords (hereafter: words), with the TP from one syllable to the next being 1 within any word, but only 0.33 across words. Crucially, synthesized speech is used in order to eliminate natural linguistic clues such as pitch variation, intonation, pauses etc. from the speech stream.

Thus, the ability of the infants to differentiate the familiar and unfamiliar syllable sequences can only be caused by the different TPs between the syllables.

In two experiments, they test the infants' ability to differentiate the familiarized words from other unfamiliar words. In a preferential listening task, the infants are presented with repetitions of two of the four words from the speech stream. They are also presented with two novel words consisting of the same syllables as the words from the listening phase but ordered differently than in the listening phase (experiment 1), or with syllable triplets that occurred in the speech stream, but only across word boundaries and thus much less frequently than the familiarized words (experiment 2). They find that in both variations of this setup, infants are able to differentiate between the familiar and unfamiliar words, listening significantly longer to the novel words, as shown in Table 1 below.

Experiment	Mean listening times (s)		Matched-pairs *t* test
	Familiar items	Novel items	
1	7.97 (SE = 0.41)	8.85 (SE = 0.45)	$t(23) = 2.3, P < 0.04$
2	6.77 (SE = 0.44)	7.60 (SE = 0.42)	$t(23) = 2.4, P < 0.03$

Table 1 Mean listening times to familiar vs novel stimuli, Saffran et al. (1996a, p. 1927)

These longer listening times are caused by a dishabituation effect in the infants and indicate the ability to differentiate the novel from the familiar stimuli (cf. p. 1927).

In Saffran, Newport, & Aslin (1996), the AGL capabilities of adults are investigated. Like in Saffran, Aslin, & Newport (1996), synthesized speech is used to create CV-syllables, which are then arranged into sets of syllable triplets, creating tri-syllabic words. However, in this variant, some syllables are re-used across multiple words, leading to TPs between 0.3 and 1 after each syllable within any word. There are also significantly more words included in the syllable stream, meaning the TP between words ranges between 0.1 and 0.2 (cf. Saffran, Newport, & Aslin, 1996, p. 612). Here, the familiarization time was significantly longer than in Saffran, Aslin, & Newport (1996), consisting of three blocks of seven minutes each. After familiarization, the participants were asked to perform a lexical decision task in a 2 alternative forced choice paradigm (2AFC), where they were auditorily presented with two tri-syllabic words. Only one of the words was familiar to them from the familiarization phase, and they were tasked to pick which of the two was the familiar word. Half of the participants received completely novel words constructed out of the same syllable inventory as the speech stream, while the other half received part words, meaning that the initial two syllables from a word from the familiarization phase were taken and combined with a third syllable which was not presented with that syllable pair during familiarization. Participants were able to identify the familiarized word significantly above chance level, with the performance in the novel condition being better than in the part

word condition (cf., p. 613f). They also conducted a version of this experiment which included vowel lengthening, where the vowel of the first or the last syllable of each word in the familiarization phase was lengthened by 100ms. They were then given the same identification task as in the novel condition of the first experiment. Performance for all conditions was significantly above chance level, while performance for the condition with lengthened vowel on the first syllable did not differ significantly from performance in the unlengthened condition. Performance in the final vowel lengthened condition, however, was significantly better than in either of the other conditions. (cf., p. 617). Across both experiments, participants were able to identify the familiarized word significantly above chance level in all conditions, even in the conditions without any prosodic cues, suggesting that they were indeed able to recognize the familiarized words purely based on distributional statistics.

Several features of these two studies still play an important role in current day AGL studies.

First, the use of auditory stimuli and a concatenated speech stream with or without prosodic cues is the basic method which most AGL studies since then have adopted. While there may be considerable variation in the exact properties of the speech stream used, the idea of concatenating syllables and distributing them such that their TPs form a pattern has become the primary method of conducting AGL research. Using specifically designed syllables and syllable combinations instead of natural speech allows granular control over the properties of the speech stream, isolating the research topic from the confounding factors present in natural speech.

Second, synthesized speech allows for precise control over the prosodic properties of the speech stream. This not only eliminates phonetic confounds but also enables comparative research in very different areas of the world, eliminating the influence of the unique phonetic features of a language or speaker on the stimuli and thus enabling the exact replication of the same stimuli in any context[2].

1.2.2 AGL beyond the word level

Beyond the scope of the word level, the present study aims to investigate the basic mechanisms used in the learning of hierarchical structures above the word level. In order to address this issue, Thompson and Newport (2007) first employed an AGL paradigm with a continuous speech stream to investigate the ability of adults to identify structures above the word level. They call this level the phrasal level, as the words in the speech stream are combined into bigger units, analogous to how words in natural language may form phrases with other words. In their paradigm, each phrase consists of a pair of words. This chunking into

[2] The use of different speech synthesizers may introduce similar unique properties to the synthesised speech, analogous to how different human speakers may produce differences in their pronunciation of the same stimuli. It is therefore imperative to report the specific settings of the synthesizer used.

bigger phrases is again solely based on TPs, but this time based on the TPs at the word rather than the syllabic level.

Thompson and Newport use TPs at the phrasal level to emulate four grammatical operations which are also found in natural language: optional phrases, repeated phrases, phrasal movement and word class size change (p. 16). In all conditions, participants were trained on a continuous speech stream consisting of CVC syllables. These syllables were arranged into trisyllabic words based on TPs. The words were then arranged into a baseline sequence of six word classes, ABCDEF, with each letter representing one word class. Each class may be filled using one of three possible trisyllabic words, so that each of these word triplets forms a word class based on the position it may appear in and its surrounding elements. This yields a total pool of 18 different words. The word pairs AB, CD, and EF form phrases, as the phrasal grammatical operations always operate on both words of the phrase, but never across a phrase boundary. Because the syntactic operations only apply to whole phrases but not across phrase boundaries, a pattern of “peaks and dips” (p. 6) in TP emerges when the canonical sequence is modified.

During the learning phase, in which the participants simply listen to the continuous speech stream, 50% of the sentences are canonical ABCDEF sentences. This is true for both the experimental and the control condition.

In the experimental condition, the other half of the sentences is modified by applying a syntactic operation. In the optional phrase condition, one of the three phrases (AB, CD, or EF) is omitted. In the repeated phrase condition, a phrase may be repeated at the end of the sentence. In the phrasal movement condition, one of the phrases may be moved into a different position in the sentence, changing the linear order of the phrases. In the word class size condition, the baseline sequence of ABCDEF is retained, but the assignment of words to each word class is varied. It is no longer the case that each word slot may be filled by one of three candidate words, but instead slots A, C, and E are reduced to two candidates each, while words B, D, and F now have four possible candidate words.

Finally, Thompson and Newport include a fifth condition, in which all variations from the previous conditions are combined. They also added a variant of this all-combined condition, in which the number of canonical ABCDEF sentences (without syntactic modification) in the training phase was reduced to 5%, with the remaining 95% of the sentences subject to all syntactic operations described above. They named this variant the all-combined 5% condition.

In the control condition, the non-canonical half of the sentences consists of sentences which feature the same syntactic procedure as used in the experimental condition, but applied both within and across phrase boundaries, so that the pattern of TPs is as uniform as possible, in contrast to the pattern of peaks and dips in TP in the experimental condition. While the TPs

between words in the control sentences are uniform unlike in the experimental condition, half of the sentences that participants in the control condition listen to are still canonical sentences. Participants in the control condition therefore listen to the same number of canonical sentences as in the experimental condition (cf., p. 11).

The purpose of all these variations is to vary the TPs between phrases, but not within words. The syntactic variations create variance in the TPs, with the probabilities within words and phrases remaining constant at 100%, while the probabilities between phrases vary. Thus, the experiments aim to investigate the ability of the participants to identify units above the word level. Table. 2 shows the change in TPs for each condition, as well as the control conditions.

Transitional Probabilities From Experiments 1 Through 4

	$A \rightarrow B$	$B \rightarrow C$	$C \rightarrow D$	$D \rightarrow E$	$E \rightarrow F$
Optional phrases	1.00	0.80	1.00	0.80	1.00
Optional control	0.90	0.90	0.90	0.90	0.90
Repeated phrases	1.00	0.86	1.00	0.86	1.00
Repeated control	0.92	0.94	0.92	0.94	0.93
Moved phrases	1.00	0.60	1.00	0.60	1.00
Moved control	0.78	0.78	0.78	0.78	0.78
Class size variation	1.00	1.00	1.00	1.00	1.00
Class size control	1.00	1.00	1.00	1.00	1.00
All-combined	1.00	0.56	1.00	0.52	1.00
All-combined control	0.80	0.83	0.74	0.77	0.74
All-combined 5%	1.00	0.33	1.00	0.22	1.00
All-combined 5% control	0.67	0.71	0.58	0.59	0.47

Table 2 TPs by experimental condition, Thompson & Newport (2007, p. 10)

Participants were auditorily presented with the stimuli of the condition they participated in, for 20 consecutive minutes per day, for five consecutive days, yielding a total of 100 minutes of exposure. After day 1 and day 5, participants were asked to complete two tests, the Sentence test and the Phrase test. In the Sentence test, participants were auditorily presented with two sentences. One sentence followed the baseline sequence of ABCDEF, but with a novel combination of word class realizations, while the other sentence also followed the baseline sequence, except for one word being inserted from the wrong word class, resulting in sentences like ABCFEF or EBCDEF. Participants were then given a 2AFC task of selecting which of the sentences followed the rules of the grammar.

In the Phrase test, participants were again asked to perform a 2AFC task, this time choosing between two phrases. The first phrase followed the canonical phrase structure from the learning phase, i.e. word pairs like AB or CD, while the second phrase was also a valid sequence of phrases, but across a phrase boundary, like BC or DE.

The results for the Phrase test showed that in all conditions except for the word class size condition, participants showed a significantly higher accuracy in the experimental condition compared to the control condition.

In the Sentence test, participants again showed significantly higher accuracy in the experimental condition, except for the repeated phrase and the all-combined condition. The performance in both tests increased significantly between the tests on day 1 and the tests on day 5 for all conditions. However, in the conditions in which the performance in the Sentence test differed significantly from the control condition, this significant difference was already present on day 1. The only exception to this was the all-combined 5% condition, in which there was no significant difference between experimental and control condition on day 1, but a significant difference on day 5[3].

As Thompson and Newport discuss, these results suggest that introducing higher-level grammatical operations increases, not decreases the ability of the participants to identify and discriminate phrasal structures in a continuous speech stream. Especially notable are the results of the Sentence test in the all-combined 5% condition, where participants showed a clear learning effect and increased ability to identify canonical sentences, even though only 5% of the stimulus material actually consisted of canonical sentences and the material they were presented with was grammatically complex, featuring several different grammatical operations simultaneously. Thompson and Newport take this as evidence that variations in TP are sufficient to help learners identify phrases above the word level in a continuous speech stream (p. 32). A post-hoc frequency analysis also showed that this learning effect was not significantly correlated with the pure frequency of the target material in the learning stimuli (pp. 35f). This finding also suggests that participants may have been able to build hierarchical abstraction of the material they were exposed to and may have reconstructed the canonical sequence out of the grammatically modified sequences.

While they present strong evidence to support this basic claim, there are a couple of aspects of their experimental design and analysis which are worth discussing. First, unlike the studies discussed above, Thompson and Newport opted not to use synthesised speech and instead used recordings of naturally spoken words performed by a female speaker. They also intentionally did not exclude phonological confounds which may have influenced the way the participants parsed the speech stream. The authors discuss this themselves, stating that “the exposure materials clearly sound like sequences of already-segmented words, not like unsegmented sequences of syllables. Each word is spoken with primary stress, there is minimal coarticulation between words, and there are long pauses between sentences. In

[3] For exact reference of the evaluation of the statistical tests used for this analysis, cf. pp. 13f, 20-23, 28f, 31f.

addition, for our native English-speaking participants, phonotactic constraints would disfavour combining many of the CVCs into multisyllabic words, such as *hoxjes*." (p. 37)[4]. Therefore, the materials of this study are much closer to natural language than in other AGL studies. While this increases the direct applicability of the results for natural language learning, it clearly limits the degree to which the performance of the participants can be attributed to purely statistical properties of the stimuli. While the results strongly suggest some ability of hierarchical processing in the participants, the experimental design means that it is not clear whether or how much of this was based in pre-existing linguistic knowledge of the participants and how much is attributable to the statistical properties of the speech stream.

1.3 Neurophysiological investigations of adult AGL

Besides the behavioural studies discussed so far, the underlying neural mechanisms of adult AGL have also been investigated. Both word-level as well as phrase-level learning can be tracked using neurophysiological methods such as electroencephalography (EEG) or magnetoencephalography (MEG) in a steady-state design. In contrast to long-established neurolinguistic measurements such as event-related potentials (ERPs) which are used to investigate the effect of individual stimuli, steady-state designs measure the brain response at specific frequency or time ranges during continuous and/or repeated stimulus presentation, making it an ideal method for the investigation of speech stream paradigms.

Buiatti et al. (2009) used Steady-State Responses (SSRs) to investigate the ability of adult French native speakers to identify syllables and words in a continuous speech stream. They constructed a speech stream consisting of trisyllabic words, with no structure above the word level. The syllables were arranged such that the first and third syllable within a word were dependent, meaning the first syllable predicted the third syllable with a probability of 1. In contrast, the second syllable varied, and across words no dependency existed, yielding TPs between 0.33 and 0.5 after the first and third syllables of each word. Because there is an intervening syllable between the first and third syllable which varies independently of the other syllables, the dependency between the first and third syllable can be analysed as a NAD.

In order to investigate the participants' ability to process this dependency and use it to identify word boundaries, they employed a frequency response analysis which they called frequency-tagging. As Picton et al. (2003) demonstrated, presenting an auditory stimulus at regular intervals evokes a spike in brain potential at the exact frequency of the presentation of the stimulus[5]. This feature of the brain allows for direct analysis of the learned underlying

[4] The superlative *foxiest* is an example of a phonotactically similar word.
[5] To clarify, this response is explicitly NOT at the frequency of the stimulus, but at the frequency of the presentation of the stimulus. If you present a 1,000Hz tone at a 40Hz frequency, the SSR brain response will occur at 40Hz and not at 1,000Hz (cf. Picton et al., 2003, pp. 179f).

structures: If the difference between the TP of the word-internal NAD and the TP between words allows the participants to identify word boundaries, a frequency response at the frequency of the words is expected. They investigated this by measuring the EEG response while exposing participants to four different speech streams: First, the speech stream described above, which they called the word condition. Second, a control condition featuring a speech stream consisting of the same syllable material, but randomly arranged without any dependencies, which they called the random condition. The other two streams were variants of the word and random streams with subliminal pauses of 25ms interspersed between each word in the word condition and simply after every six syllables in the random condition in order to match the frequency of the word condition. In contrast to previous studies which measured the participants' response either by subsequent behavioural tests as in the studies discussed above or using ERPs in response to specific violations (Friederici, 2002; Neville et al., 1991), the SSR measurement allows measurement already during the initial learning phase when participants are first familiarized with the stimuli and start to identify the underlying structures using the statistical properties of the speech stream. This eliminates interference due to memory constraints during a later recall task, and directly measures the learning of the participants as it happens.

They found that in both random conditions, a frequency response at the presentation frequency of the syllables occurred but found no such response in either word condition. Instead, they found a frequency response at the presentation frequency of words in the word condition with pauses and no stimulus specific frequency response at all in the word condition without pauses (Buiatti et al., 2009, p. 513).

These results show that participants required extra-statistical markers, in this case the 25ms pauses, in order to attune to the frequency of the words. The absence of a significant power response at the syllable frequency in both word conditions also suggests that the presence of higher-level structures may inhibit the response to lower-level structures. As the response of the participants in the no-pause condition showed, the presence of higher-level structures seems to suppress the response to lower-level structures even when the higher-level structure itself does not evoke a significant power response. This is also why the response can be attributed to identifying the structure of the words rather than just responding at the frequency of the pauses. Because the same suppression effect occurred in the word condition without the pauses and no response at the pause frequency occurred in the random condition with the interjected pauses, it is clear that participants responded directly to the presence of the word-level structures and not just to the identically timed pauses. Buiatti et al. therefore suggest a suppression effect, which causes a diminished response to lower-level elements when they are embedded in a higher-level structure (p. 516).

Adding to this investigation of hierarchical structures, Getz et al. (2018) adopted the design from Thompson and Newport's (2007) all-combined 5% condition and measured participants' brain response during the learning and test phase in order to investigate the brain's response to syntactic operations like movement, optionality etc. They employ an MEG power analysis, tracking the power response of the participants during the learning phases. They find that at the frequency of the phrase, the experimental group has a significantly increased MEG power response compared to the control group, and this difference increases over time. Both of this is only true for the phrase frequency, not the word frequency. The maximal level of separation between the response of the experimental and the control group was achieved after 4 minutes (p. 139). One possible explanation for the absence of a response at the word level, besides the suppression due to the presence of a higher-level structure at the phrasal level, was that the two words that made up a phrase overwhelmingly occurred together. Of the grammatical operations modifying the sequences of the speech stream, only the word class size condition changed anything below the phrasal level. This condition was present in less than a quarter of the sequences participants were exposed to. All other operations modified entire phrases, which made segmentation at the word level considerably less important for identifying the underlying structure of the sentences. As a result, identifying the dependencies within each sequence may have skipped the word level for most sentences. A factor in this may also have been the inclusion of prolonged pauses. Following each sentence, a pause of the same duration as a sentence was introduced. Therefore, and because of the limitations of the materials from Thompson and Newport (2007) discussed above, it is ultimately unclear how much of the participant response to the hierarchical structure was caused by the statistical properties of the speech stream compared to the extra-statistical factors like phonotactic constraints and pauses.

As the results of these experiments show, more research into the interplay of different hierarchical levels is needed. Ding et al. (2016) investigated this dynamic in natural language using synthesized English and Mandarin stimuli and found that the presence of higher-level structures like phrases and sentences did not significantly impact the response at the frequency of the lower-level structures. Specifically, they showed that long-distance dependencies can be tracked on multiple hierarchical levels simultaneously, showing the underlying processes that enable natural language processing in adults. Like the studies above, they focused on investigating natural language processes, using natural language stimuli, which restricts the area of application to natural language processing.

The key question arising from these results is whether adults can perform the higher-level NAD processing shown in Getz et al. (2018) and Ding et al. (2016) purely based on TPs in an artificial language, without the aid of extra-statistical information gathered from pre-existing knowledge of natural language or strategically placed pauses. Answering this question may

provide additional insight into the tension between constituent structure based and statistics-based analysis of language acquisition as discussed in Ding et al. (2017). There, the authors discuss how a purely statistically trained N-gram prediction model can be differentiated from a rule-based constituent structure model, citing the inclusion of context and long-distance dependencies between different elements of a discourse as features that a statistical prediction model cannot accommodate (p. 3). Purely probability-based studies such as Weyers et al. (2022) as well as the present study contribute to the answering of that question by exploring the boundary of purely statistical learning of linguistic structures in paradigms without any extra-statistical information.

Additionally, the existence and scope of a hypothesized suppression effect on lower-level structures by higher-level structures remains unclear. While both of these phenomena were investigated in Buiatti et al. (2009), Ding et al. (2016) and Getz et al. (2018), none of these experiments were intended for investigating exactly these questions relying only on statistical properties. Buiatti et al. (2009) did not feature structures above the word level, making no claims about the processing of higher-level structures. Ding et al. (2016) employed natural language stimuli in a variety of formats and found that different hierarchical levels can be processed and integrated at the same time. The use of natural language stimuli, however, introduced different variables besides pure TPs by using linguistic units that were already known to the participants and associated with pre-existing information. Getz et al. (2018) used the all-combined 5% condition from Thompson and Newport (2007), which does feature NADs above the word level, but mixes several different dependencies into one dataset. Part of the dataset also features a dependency at the word level, meaning that this study cannot investigate NADs in isolation. Their inclusion of prolonged pauses after each sentence also practically eliminates the task of segmenting the speech stream into individual sentences. This limits the applicability of the results to natural language processing, as segmentation into sentences is a core task of both AGL and natural language processing. Whether TPs alone are sufficient for the brain to keep track of syntactic NADs was previously investigated in Bulca et al. (2019), which was expanded upon in Weyers et al. (2022). As the present study is an adapted replication of that study, it is explained in detail in the following section.

1.4 Prior studies of the same design

While research into purely statistical AGL is well-established in infants (Culbertson et al., 2016; Höhle et al., 2006; Van Heugten & Shi, 2010), research into adult responses to these paradigms has been limited. Thompson and Newport's (2007) Movement condition featured an NAD, as the participants had to track the phrasal constituents in order to identify the underlying structure. However, the speech stream still featured canonical sentences for half of the stimuli, and participant performance was only measured behaviourally. Buiatti et al. (2009)

featured an NAD, but it was only at the word-internal level, and the intervening element only varied between three different options. Studies like Müller et al. (2009) or Ding et al. (2016) investigated responses to NADs in adults, but used natural language stimuli. Weyers et al. (2022) aimed to investigate the questions asked above, namely the neurophysiological analysis of adults' ability to identify and track NADs at a phrasal level, entirely based on distributional statistics of the syllable stream. It is an extension of the work done in Bulca et al. (2019).

In Weyers et al. (2022), 24 adult German native speakers were exposed to two blocks of four learning and test phases, during which their EEG was recorded. During each learning phase, participants were presented with a speech stream and were asked to identify an underlying rule in the ordering of the syllables.

For the learning phases, two grammars were constructed. Each grammar consisted of concatenated six-syllable sequences. These sequences were separated into words A, X, and B in an AXB arrangement. Each word consisted of two syllables, resulting in three words per sequence. Each word could vary only in the first syllable, with the second syllable remaining constant throughout all conditions of each grammar. The first syllable of A and B could vary between two options, while the first syllable of X varied between 20 options. This variation in the first syllables of A and B featured a fixed dependency, meaning that *FI* as the first syllable of A was always followed by *SE* as the first syllable of B, and *PE* was always followed by *TO*. Fig. 1 shows two example sequences, featuring both variations for the first syllable of A and B from one grammar, with the TPs between syllables as well as the NAD indicated by arrows.

Fig. 1 TPs of artificial grammar sequences, Weyers et al. (2022)

The MBROLA speech synthesizer (Dutoit et al., 1996) was used to synthesize the syllable inventory of 27 syllables per grammar. 40 sequences were concatenated to form one learning phase, lasting a total of 1:28min. Each learning phase was followed by a test phase, in which the participants had to perform a rating task, where they were presented with 16 individual sequences and had to decide whether or not the sequences followed the same pattern as the stimuli they heard during the learning phase. The rating was performed on a five-point Likert scale, indicating their certainty of whether or not the sequences they just heard were correct, on a scale ranging from certainly correct to certainly incorrect (cf. Bulca et al., 2019, p. 11). Post-hoc, participant responses were mapped into a binary judgement on grammaticality. Half

of the sequences in each test block violated the grammatical pattern of the learning phase by swapping the first syllable of the B word to the first syllable of the B word of the other sentence type. For the test items, a separate set from the one used in the learning phase was used for the first syllable of the X word. After four learning and test phases, the first block was concluded, and a break followed. Afterwards, participants again went through four learning and test phases, this time with stimuli constructed from the same syllable inventory as in the first block but ordered randomly (avoiding multiple consecutive repetitions of the same syllable as well as German and English words).

In the rating task, participants overall did not significantly differ in their response to the grammatical sequences compared to their response to the ungrammatical sequences, identifying the sequences correctly with a mean accuracy of 52.47% (SD = 15.25). However, a sub-group of participants was able to identify the correct sequence with significantly above-chance accuracy (mean: 73.7%, SD = 11.33). The EEG measurements were analysed using ERPs as well as Steady-State Auditory Evoked Potentials (SSAEPs), similar to the steady-state responses analysed in Ding et al. (2016) and Buiatti et al. (2009). While there was no significant difference between the ERPs evoked in the grammatical vs random condition, the participants showed significant increases in power during the grammatical condition at the phrasal, 2nd harmonic of the phrasal as well as word frequency compared to the random condition, as depicted in the spectral analysis in Fig. 2.

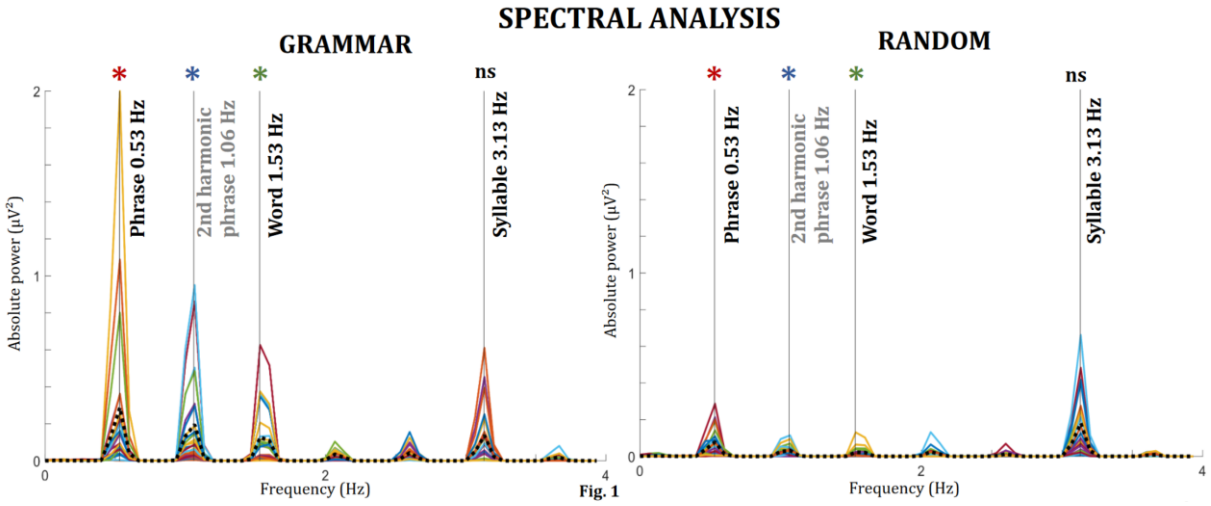


Fig. 2 Spectral analysis of power response, Weyers et al. (2022)

The results indicate that participants were able to simultaneously track the hierarchical structure of the sequences on both the word and phrasal level. Participants were able to do this despite the absence of any form of extra-statistical assistance like pauses in the speech-stream (as in Buiatti et al., 2009) or the use of natural language (as in Ding et al., 2016 or Müller et al., 2009).

In order to adapt this design for the present study, the following points were focused on:

First, the learning phases were quite short at 1:28min. As Getz et al. (2018, p. 139) showed, the difference in power level response at the phrasal level between experimental and control group continued to differentiate until roughly four minutes of exposure time, possibly indicating a continued entrainment effect. As such, the shorter learning phases in Weyers et al. (2022) did not fully take advantage of the entrainment effect of prolonged exposure to the stimuli. A longer learning phase, which makes full use of this effect, may therefore lead to an increase in performance in the rating task.

Second, test phases being added after each learning phase may introduce a confound. As there are three learning phases after the first test phase, the presentation of individual sequences during the test phases means that participants are clued into the length and structure of the individual utterances. As such, the material presented during these intermittent test phases may have influenced the response of the participants in subsequent learning and test phases by offering both additional exposure as well as aiding in segmentation of the speech stream. This makes interpreting the results less clear, as the general training effect from additional exposure during the learning phase is confounded with the additional information gained during the test phases. The learner group did show an improvement in their accuracy in the rating task in later test phases, but the mixture of learning and test phases means that we cannot attribute this improvement unequivocally to the entrainment effect of the learning phases.

Third, the violation of the NAD was introduced by swapping the first syllable of the B word with the first syllable of the other variant of the B word. Because both B words use the same second syllable, this is equivalent to swapping the entire words. If one were to interpret the two sequences from each grammar as sentences, then simply swapping a word from one sentence to another sentence of the same language may be interpreted as simply constructing a new, felicitous sentence. In the same way, swapping words between the sentences “Mary likes dogs.” and “Peter likes cats.” forms the new, but grammatical sentences “Mary likes cats.” and “Peter likes dogs.” While these sentences differ from the ones learned during the learning phase, they are perfectly grammatical. Participants may have reasoned about the swapped B words in the same way, in the sense that no grammatical violation has taken place, as a previously learned, valid word was simply put in a new context while even staying in the same syntactic position. They could also have constructed ad-hoc “word classes” similar to the analysis proposed in Thompson and Newport (2007), where participants learn that e.g. in grammar 1, either of *PE* or *FI* may appear in the first position for word A, and either *TO* or *SE* may appear in the first position of word B, without constructing the NAD between words A and B as a necessary condition of the grammar. Under this analysis, swapping a syllable in the first position of the B word would not constitute a grammatical violation, as the “new” syllable is one

of the two syllables learned in that position during the learning phase, and the resulting word is a valid member of the word class for that word position. Therefore, participants may have recognized the NAD violation, but not classed it as a violation in the rating task, leading to a rating accuracy that does not directly reflect the ability of the participants to differentiate between the canonical and non-canonical sequences of the test phase.

1.5 The present study

To remedy these aspects of Weyers et al. (2022), the present study was designed as an adapted replication. It uses the same syllable material and overall design, with the learning phases being constructed according to an identical grammar as in Weyers et al. (2022). However, some adaptions to the procedure have been made.

First, the number of learning and test phases has been reduced to one phase each per block. The total amount of exposure during the now single learning phase is the same as the sum of the four individual learning phases in Weyers et al. (2022), now presented in one single uninterrupted learning phase[6]. This also means that only one test phase is used to check the results of the learning phase. This does eliminate the confounding factor of the intermittent test phases clueing in the participants to the underlying structure of the stimuli. Unfortunately, this also makes a direct comparison between the results from Weyers et al. (2022) and the present study less clear, as the no longer available additional information from the intermittent learning phases may be counteracted by a stronger entrainment effect along the lines of what Getz et al. (2018, p. 139) showed. Therefore, while the increased uninterrupted exposure is expected to lead to an increased response, the exclusion of the intermittent test phases may also have an opposite effect by no longer affording the participants with additional clues to the structure of the speech stream. Therefore, there is no clear prediction on the effect of this change on the performance of the participants, both behaviourally as well as neurophysiologically, although the fully measured differentiation effect from Getz et al. (2018) appears to be a more reliable effect than the speculative effect of the intermittent test phases which has not been investigated in isolation. Thus, a stronger entrainment in the power response due to the increased continuous exposure is plausible. Eliminating the intermittent test phases introduces another issue. As Müller et al. (2012, p. 15956f) showed, participant performance significantly decreased during purely passive listening exposure, with an active task being required to facilitate any rule learning. Although that study featured an oddball design not entirely comparable to the present study, the longer period of pure exposure with no interactive element

[6] Due to differences in the auditory equipment used in the different labs, syllables in the present study were slightly shorter at 304ms compared to 320ms for Weyers et al (2022). This results in a slightly longer total runtime (but not additional exposure to syllable material) in Weyers et al. (2022) as well as slightly different frequencies of interest (I. Weyers, personal communication, July 29, 2024).

to maintain participant attention may still deteriorate performance compared to the shorter exposure periods followed by active tasks for the participants in Weyers et al. (2022).

Second, to test whether the participants learned the NAD between the A and B words, the nature of the violation during the test phase is changed. Instead of swapping a syllable from one sequence to the other, the NAD is violated by swapping the order of words A and B, yielding a BXA structure instead of the canonical AXB. This is similar to a variant of the moved phrases condition from Thompson and Newport (2007) where the first and third phrase were swapped, with the middle phrase remaining in position. The only difference to that condition is the X element in the present study, which continues to vary independently of the A and B elements. It also avoids the issue of the participants building word classes and interpreting the violated condition from Weyers et al. (2022) as grammatical because all the syllables individually were felicitous in the position they occured in. Unfortunately, the present study being conducted with German native speakers as opposed to the English native speakers of Thompson and Newport (2007) makes a word order violation less impactful. Due to easily available scrambling as well as the generally relatively free constituent order in German (Müller et al., 2021, ch. 2), German native speakers may process a word order violation like this differently from native speakers of languages with a more rigid word order like English.

Third, the rating scale was changed from a five-point to a six-point Likert scale. This was done in order to avoid having a neutral option for participants and force them into a commitment to a correct or incorrect judgement. An even scale is also more suited for later re-mapping onto a binary judgement, as the neutral option on an odd-numbered Likert scale has no equivalence in a binary judgement task.

In order to maintain comparability to the original study, the rest of the present study was kept identical to Weyers et al. (2022).

2 Methods and materials

2.1 Participants

In total, 33 German native speakers were recruited, all right-handed. Of those, two were excluded from the analysis as they were used as pilots, and two due to technical difficulties during the testing procedure (no clear signal in the ground electrode could be established), leaving a total of 29 participants included in the analysis. As the method of recruiting and selecting participants may introduce selection biases into the experiment, the method used to recruit the participants as well as some demographic features of the participants are briefly discussed in this section.

Participants were recruited using CogSciHub, a platform of cognitive science and cognitive neuroscience researchers at the University of Vienna. One of its features is the Study Participant Platform (SPP), which serves as a database for prospective participants in cognitive science experiments. Recruiting from this database introduces some biases, which are reflected in the demographic features of the participants included in the analysis: Participants were overwhelmingly female, with 22 of 28 participants self-identifying as female. The platform is also skewed towards young people, with participants having a mean age of 25 (SD = 3.85, range = 18-34). The level of education was also higher than in the general population, with 13 (46%) participants having completed at least secondary education (Matura, Abitur or equivalent), 14 (50%) having some sort of higher education degree, and only one person (4%) having completed primary education as their highest level of completed education. This compares to 63% having completed secondary education, 19.7% having completed tertiary education and 17.3% having completed primary education as their highest level of education in the Austrian general population (Statistik Austria, 2022). Additionally, all participants spoke at least one other language besides German to a self-assessed level of 3/5 or higher.

As such, it cannot be claimed that the participants recruited for this experiment are representative of the Austrian general population. However, this is only of limited concern, because investigating the possible impact of demographic factors like different genders, ages and socio-economic statuses on the reaction to the stimuli was not the aim of this study. The previous literature discussed in Chapter 1 also focused on participants of similar demographic backgrounds. Furthermore, the variation in the demographic factors like age, level of education etc. will be included in the statistical analysis of the behavioural data.

2.2 Experimental setup and procedure

Participants were tested in the test rooms of the Psycholinguistisches Labor Babelfisch at the University of Vienna. The experiment was conducted by two experimenters.
In the control room, participants were asked to fill out the form of consent and the demographic questionnaire[7]. If the participant provided their informed consent and were not excluded from participation for medical reasons, they were escorted into the soundproof test room, where they were seated in a chair centred between two speakers and in front of a desk. There, a rough outline of the experimental procedure was explained to the participant. The EEG was then set up as described in section 2.4. The screen and keyboard were arranged centrally in front of the participant, the screen being about an arm's length from the participant (about 60-70cm depending on the arm length of the participant). The keyboard was placed in front of the

[7] Both forms are provided in the supplementary materials.

participant so that they could reach the bottom row of keys comfortably, where stickers were used to label the keys x through m with the numbers 1 through 6 from left to right. The participant was then handed the detailed instruction sheet[8], and one of the experimenters remained in the room to answer potential questions while the other experimenter set up the EEG recording in the control room. In the instruction sheet, participants were alerted to the fact that they were about to hear a continuous stream of syllables, and that these syllables followed a specific pattern. They were instructed that it was their task to try and identify this pattern. The part of the instructions which referred to the judgement task of the participants was kept the same as in Weyers et al. (2022), asking participants "Does [the syllable sequence you just heard] follow the same rules as the syllable sequences heard in the listening phase?". The labelling of the rating scale was also kept the same, asking participants to judge how sure they were that the sequence they just heard followed the same rules as those in the listening phase on a scale from *Surely wrong* to *Surely right*.

The experiment was programmed using MATLAB version 2023a (MathWorks, 2023) and Psychtoolbox. version 3.0.18 (Kleiner et al., 2007). It consisted of two blocks, a grammatical block and a random block. The order of the blocks was randomized across participants. Even though the random block did not offer any input that participants could learn the artificial grammar from, it offered an opportunity for participants to adapt to the specific task of identifying syllables and integrating them into larger structures, as well as getting more used to the synthesized voice, the experimental setting etc. Therefore, the order of the blocks was randomized, and the order of the blocks is included in the statistical analysis of the results.

Each block consisted of a listening phase and a test phase.

During the listening phase, the participants simply sat as still as possible and listened to the continuous speech stream being played to them over the loudspeakers, with only a fixation cross being displayed on the screen in front of them.

During the test phase, participants were presented auditorily with individual sequences of the grammar they had heard in the previous listening block, with half of the sequences following the canonical structure, and the other half of the sequences following a BXA structure as discussed in section 1.5. After each sequence, they were asked to rate on a scale from 1 to 6 how sure they were that the single sequence they just heard followed the same pattern as the material in the previous listening block. In the test phase of the random block, they were presented with a random string of syllables. The graphic that was displayed on the screen for the selection task is presented in Fig. 3, with a rating of 1 indicating definitely wrong, and a rating of 6 indicating definitely correct. In order to control for any possible left-right bias, half of

[8] The instruction sheet is provided in the supplementary materials.

the participants were presented with the flipped task and graphic, with a rating of 1 indicating definitely correct, and a rating of 6 indicating definitely wrong.

Fig. 3 Rating task graphic displayed to the participants during the test phase

Participants were asked to rate 20 sequences in total.

After the first test phase was finished, a break of five minutes was inserted before the start of the next listening phase. At the start of this break, participants were given a short questionnaire asking them whether they felt like they identified any specific pattern in the previous block, and if so, asked them to describe or spell out the pattern they thought they identified.

Apart from filling out the questionnaire, the break also served two purposes: First, it gave participants an opportunity to rest after a relatively demanding listening and rating task. Due to the nature of the stimulus design, the rapid and repetitive stream of syllables may become fatiguing over time. Second, the break helped participants forget the pattern from the previous block, ensuring that the two blocks are truly independent and don't interfere with each other. To further facilitate this, at least one of the experimenters spent the break in the room with the participants, chatting with them in order to help clear their mind from the previous task. No feedback on the experiment was given to the participants during the break, and no questions regarding the content of the experiment were answered. After the conclusion of the break, the second block started, following the same procedure described above for the first block. After the conclusion of the second block, the EEG was disassembled and the participants were given the opportunity to wash their hair and clean up. Afterwards, participants were compensated for their time with 15€.

2.3 Stimulus material

The present study re-used the same stimuli as Bulca et al. (2019) and Weyers et al. (2022). All auditory stimuli were synthesized CV-syllables. The syllables were synthesized using MBROLA (Dutoit et al., 1996), voice DE7, with a pitch of 200Hz. The sampling rate was 44,100Hz, which was sampled down to 22,050Hz for playback. The length of each syllable was set to 304ms by adding or truncating silence. The syllables were played immediately following each other, meaning no pause of any kind was introduced into the speech stream. No change in intonation or pitch was introduced. With the syllable length of 304ms, the frequencies of interest for the different hierarchical levels are 3.29Hz for the syllable frequency,

1.65Hz for the word frequency and 0.55Hz for the phrasal frequency. With a total of 160 sequences per learning phase, a learning phase lasted 4:52min.

For the grammatical block, two grammars were constructed. The only difference between the two grammars consists in the choice of the syllable inventory; both grammars follow the same distributional rules. Graphs showing the TPs for both grammars are shown in Fig. 4A, B. Continuation from the last syllable of each sequence (*BA / KI*) to the beginning of the following sequence (*PE*|*FI / SI*|*GO*) is shown to better illustrate TPs between sequences. The NAD between first and fifth syllable is indicated in orange. The full stimulus material is provided in the supplementary materials.

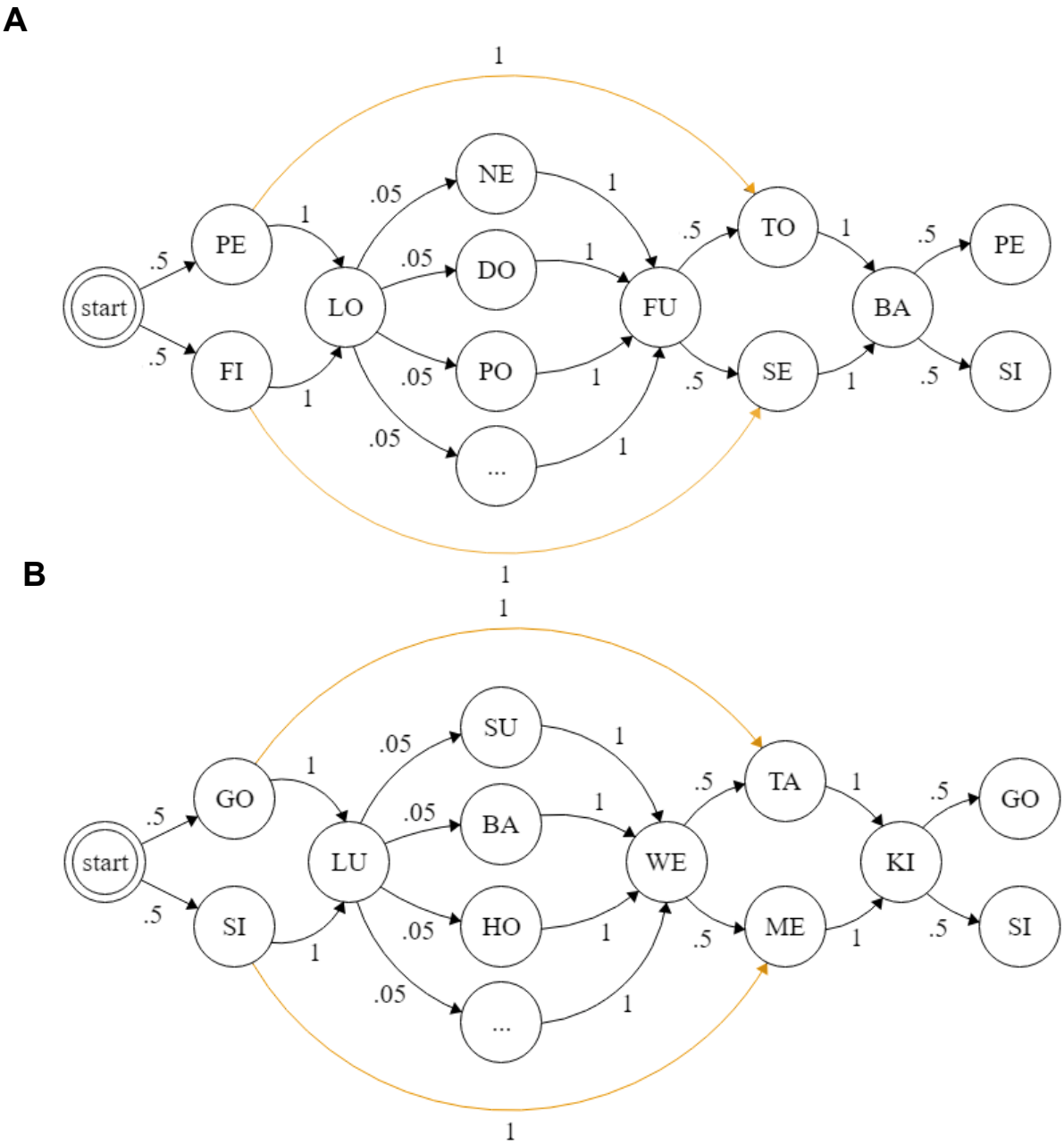


Fig. 4 Transitional probability graphs for (A) grammar 1 and (B) grammar 2

The speech streams constructed with these grammars consisted of syllables which were arranged into 6-syllable sequences according to the following grammatical rules[9]: The first two syllables in each sequence were one of two possible pairs, either *PELO* or *FILO*. These pairs always consisted of the same second syllable *LO*, with the first syllable alternating between *PE* and *FI*. Similarly to the first two syllables, the fifth and sixth syllable in each sequence, *TOBA* and *SEBA*, also alternated between two options for the fifth syllable, while the sixth syllable stayed the same across all items of each grammar. The fourth syllable, *FU*, like the second and the sixth syllable, stayed the same throughout the entire grammatical block. Only the third syllable deviates from this pattern, alternating between 20 different realisations, e.g. *PA* or *WE* or *TA* etc. The NAD is introduced by always matching one option for the first syllable with the same option for the fifth syllable, i.e. *PE* always being followed by *TO* and *FI* always being followed by *SE*. A full sequence of each grammar may then be *PELONAFUTOBA* for grammar 1 and *GOLUPOWETAKI* for grammar 2. There are two variants of each grammar, only differing in the order of the sequences in the learning phase. Two grammars are used in order to control for confounding through the specific realisation of the syllables. The syllables selected for each grammar are carefully chosen as to not systematically vary with regards to phonological features (e.g. one grammar always having a plosive as the first sound of the first syllable while the other grammar always features a fricative in the same position, thus distinguishing the two grammar types). Both sets of syllables are also chosen such that no German or English words are included which could impact participant performance.

Due to these TPs, syllable pairs one and two, three and four, and five and six each form a word. The three words combined form a phrase of the artificial grammar.

In total, for each grammar there are two different first syllables, one second syllable, 20 third syllables, one fourth syllable, two fifth syllables and one sixth syllable, totalling to 27 different syllables per grammar during the listening phase.

For the listening phase of the random block, the same sets of syllables were used which were used for the grammatical part. Instead of following the grammatical rules described above, however, the syllables were ordered pseudo-randomly (the only exceptions to randomness were disallowing German or English words and more than 2 immediate repetitions of the same syllable). As such, there was no underlying hierarchy or dependency for the participants to learn, and participants on average heard each syllable the same number of times.

For all sequences of the test phase, a set of syllables was used for the third syllable that was distinct from the set used during the listening phase. This was done in order to make sure that participants are being tested on recognizing the underlying structure of the artificial grammar,

[9] Throughout this paragraph, I will be giving examples from grammar 1 to better illustrate the rules of construction of the stimuli. The same rules also apply to grammar 2.

and not just remembering “verbatim” the sequences from the listening phase. For the test phase, 10 new syllables for the X element were introduced, with each of them occurring twice to build the total of 20 test phrases.

During the test phase, half of the sequences followed the same distributional rules of the grammar from the learning phase, and only differed from the sequences of the learning phase in the choice of the third syllable. For the other half of the sequences, the order of words A and B, i.e. syllable pairs 1+2 and 5+6, was swapped in order to introduce a grammatical violation. The advantages of using this approach for introducing the grammatical violation have been discussed in section 1.5.

2.4 EEG setup and recording

64 active electrodes were arranged in accordance with the International 10-20 System (Klem et al., 1999) and labelled following Modified Combinatorial Nomenclature labelling (Seeck et al., 2017). In addition, one electrode was fixed to the participant’s collarbone and one to their cheekbone underneath their right eye. The collarbone electrode functioned as the ground electrode, while the electrode underneath the right eye was used to record eye movement data which would later be used to remove eye movement artifacts. The Fz electrode was set as the online reference. The sampling frequency of the EEG recorder (actiChamp Plus amplifier system) was set to 500Hz, which was sampled down to 125Hz during analysis. Impedances were kept below 25kΩ. The EEG was recorded using the BrainVision Recorder recording software. After recording of the EEG, the signal was re-referenced offline to averaged mastoids.

In order to prepare the recorded EEG data for analysis, several pre-processing steps were applied. All of these operations were performed using MATLAB V. R2023a (MathWorks, 2023). The plug-ins used were EEGLAB V. 2023.1 (Delorme & Makeig, 2004) and FieldTrip Toolbox V. 20230913 (Oostenveld et al., 2011). The data were bandpass filtered to a frequency range between 0.1Hz and 30Hz, as frequencies outside this range were not of interest for the analysis. Eye artifacts were removed by independent component analysis (ICA). For the ICA, separate filtered datasets limited to the 1Hz to 30Hz range were used to aid ICA decomposition. After component selection, the resulting decomposition matrix was re-applied to the 0.1Hz – 30Hz data sets. Artifacts were removed semi-automatically using EEGLAB’s built-in artifact detection on a threshold of $\pm$175mV as well as manual selection of epochs with strong artifacts. Removed or faulty channels were then interpolated using a spherical interpolation method. For power analysis, a Steady-State Auditory Evoked Potential (SSAEP) analysis with overlapping epochs of 15 seconds length was used. A Hanning taper was applied

and the frequency range of interest limited to between 0 and 4Hz. The sampling rate of 125Hz on 15 second epochs yields a frequency resolution of 0.0667s.

2.5 Statistical methods for analysis

A cluster-based permutation test (CBPT) is used to further analyse the EEG response. The CBPT used is a non-parametric dependent-samples test. It is used to determine significant differences in the brain responses within participants in the grammatical vs ungrammatical condition. Participants who received grammar 1 were also compared to those who received grammar 2. Because this was a between rather than within participant measurement, a dependent samples variation of the CBPT was used. Due to the individual nature of the neurophysiological response, comparisons of EEG responses between participants are of limited explanatory power, and are performed in this case as a sanity check. The CBPT was conducted at the frequencies of interest, namely 0.55Hz for phrasal frequency, 1.1Hz for the 2nd harmonic of the phrasal frequency, 1.65Hz for the word frequency and 3.3Hz for the syllable frequency.

For the analysis of the main effect of the behavioural test, a one-sided one-sample t-test was used. In order to analyse the responses of the participants, the scalar responses of the participants are re-mapped to binary choices in a binary decision task. This is done because the question they are asked when choosing their response to the presented stimuli during the test phase, “How sure are you that the syllable sequence you just heard was following the same pattern as what you heard during the listening phase?”, already implies that there is only a right (= the same as the previous material) or wrong (= different from the previous material) answer, and the rating scale is only used to judge the confidence of the participants in choosing between right or wrong. The phrasing of the question thus does not allow for a scalar rating of the grammaticality of the material unlike during a standard grammaticality judgement task, and thus should not be analysed as scalar data. A one-sided one-sample t-test is used because the performance of the participants is evaluated against chance level. Since the control condition was a random scramble of syllables, there were no correct or incorrect answers in the test phase of the random block which could be used as a comparison baseline. Thus, a comparison to chance level is the only choice of comparison available for analysing the performance of the participants in the behavioural task. A one-sided test is used since it is reasonable to expect the listening phase to only have a one-sided effect of potentially increasing, but not decreasing participants’ accuracy, based both on the results from the previous literature (Ding et al., 2016; Thompson & Newport 2007; Weyers et al., 2022) as well as the experimental design. In order for the one-sided one-sample t-test to be applicable, normality of data has to be present. This will be evaluated in the results section.

For the analysis of the impact of demographic features of the participants such as level of education, bi-/multilingual background and musicality, a multivariate ANOVA will be used. The dependent variables of level of education, bi-/multilingual background and musicality are measured with categorical or ordinal data and the multivariate ANOVA allows to examine both main effects as well as possible interactions between the factors. Age and reaction time will be analysed using a linear model with interacting terms, as they are scalar data, and reaction time may at least partially be a function of age.

The impact of the different grammatical conditions (which version of each grammar the participants received, as well as whether they were exposed to the grammatical or the random block first) will also be analysed using one-way ANOVAs. In case of significance, effect of groups will be differentiated using pairwise t-tests.

The raw data for all tests are provided in the supplementary materials.

3 Results

For the results of the EEG data, one participant had to be excluded due to unusable data (irreparable strong eye artifacts throughout the entire experiment), leaving a total of 28 participants included in the EEG analysis. As there were no other issues with that participant, they were included in the behavioural analysis regardless.

3.1 EEG Results

The SSAEP analysis showed spikes at all frequencies of interest as well as integer multiples of the phrasal frequency in the grammatical condition. Spikes were strongest at the phrasal and syllabic frequencies. An outlier is present at the phrasal frequency, increasing the average response level. A power spectrum analysis of a representative electrode[10] is presented in Fig. 5.

In the random condition, no spikes are present at the 2nd harmonic of phrasal nor at the word frequency, with limited spikes for some participants at the phrasal frequency. A spike at the syllable frequency is present similarly to the grammatical condition. This is visualized in the power spectrum analysis for the random condition in Fig. 6.

[10] FCz was selected here and throughout as the representative electrode as it showed a high amplitude of activation, allowing to identify small changes in activation.

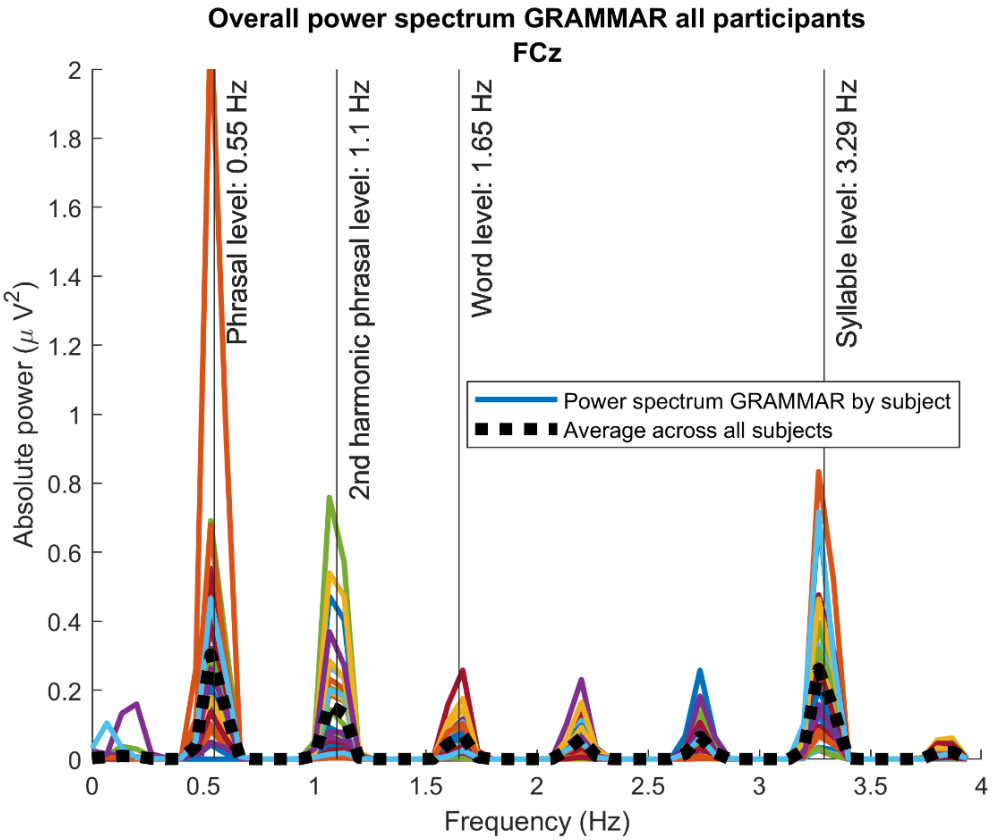


Fig. 5 Spectral analysis of overall participant response in the grammar condition

The CBPT showed a significant difference between grammatical and ungrammatical condition across all participants at the phrasal, 2nd harmonic of the phrasal, and word frequency ($p < 0.05$ in all cases). No significant difference between grammatical and ungrammatical condition was found for all participants at the syllabic frequency. Topographic plots illustrating the distributions of the clusters that differ between grammatical and ungrammatical condition across all participants at the phrasal, 2nd harmonic and word frequency are presented in Fig. 7A, B, C.[11]

[11] Note that due to the frequency resolution not perfectly aligning with the frequencies of interest, the differences presented are analysed at the closest available sampling step to each frequency of interest.

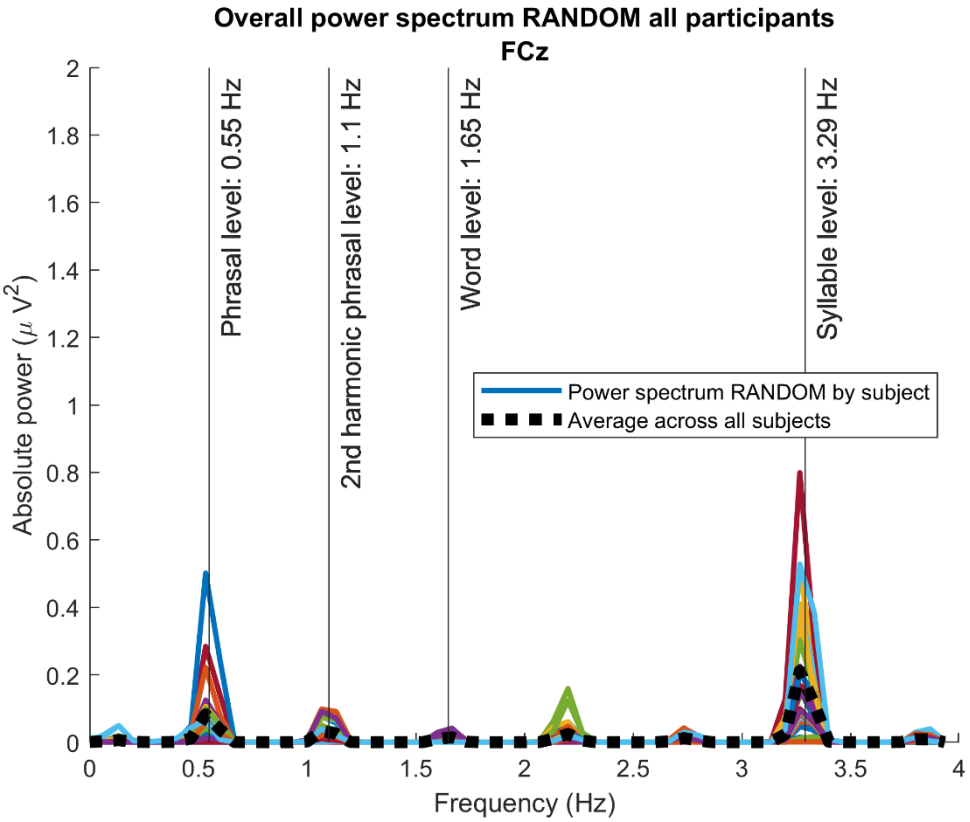


Fig. 6 Spectral analysis of overall participant response in the random condition

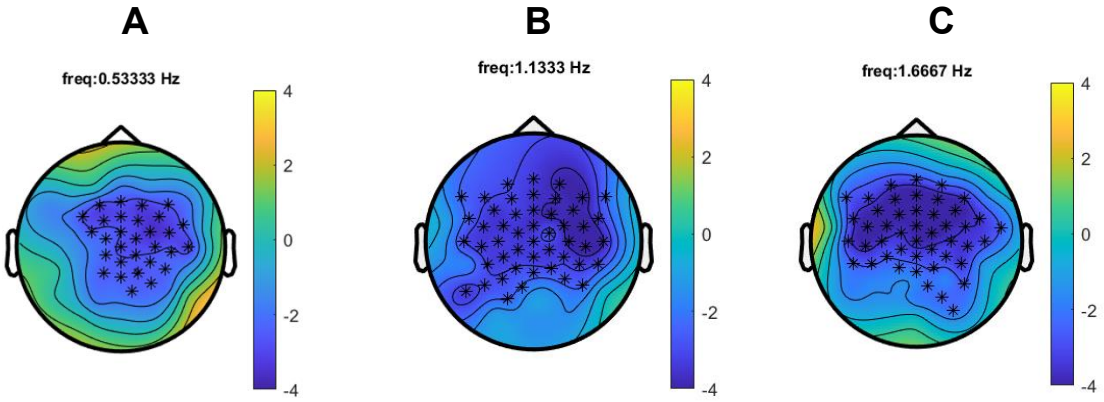
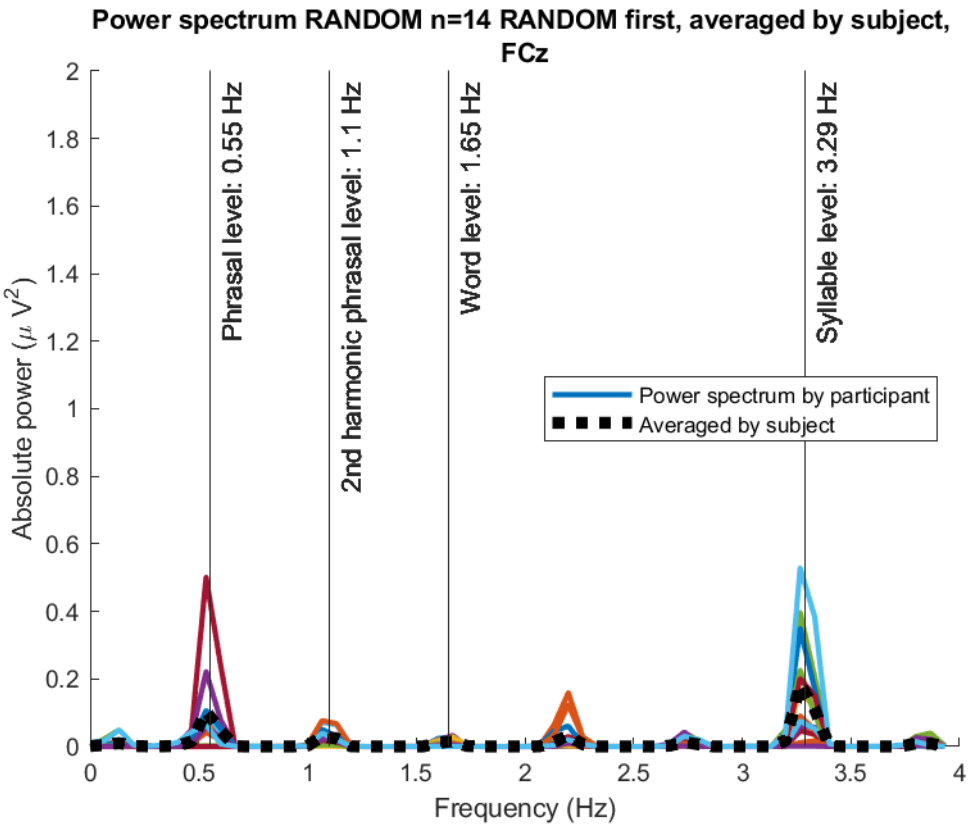


Fig. 7 Topographic plots of differentially responding clusters between random and grammar condition for overall participant group at the (A) phrasal, (B) 2nd harmonic, and (C) word frequency.

3.2 Detailed SSAEP results

In the random condition, similar levels of average power response are present at the syllabic level, but not at the word level. There is limited activation at the phrasal and 2nd harmonic level. The power spectra for the random condition in the grammar first and random first configuration are presented in Fig. 9A, B.

A



B

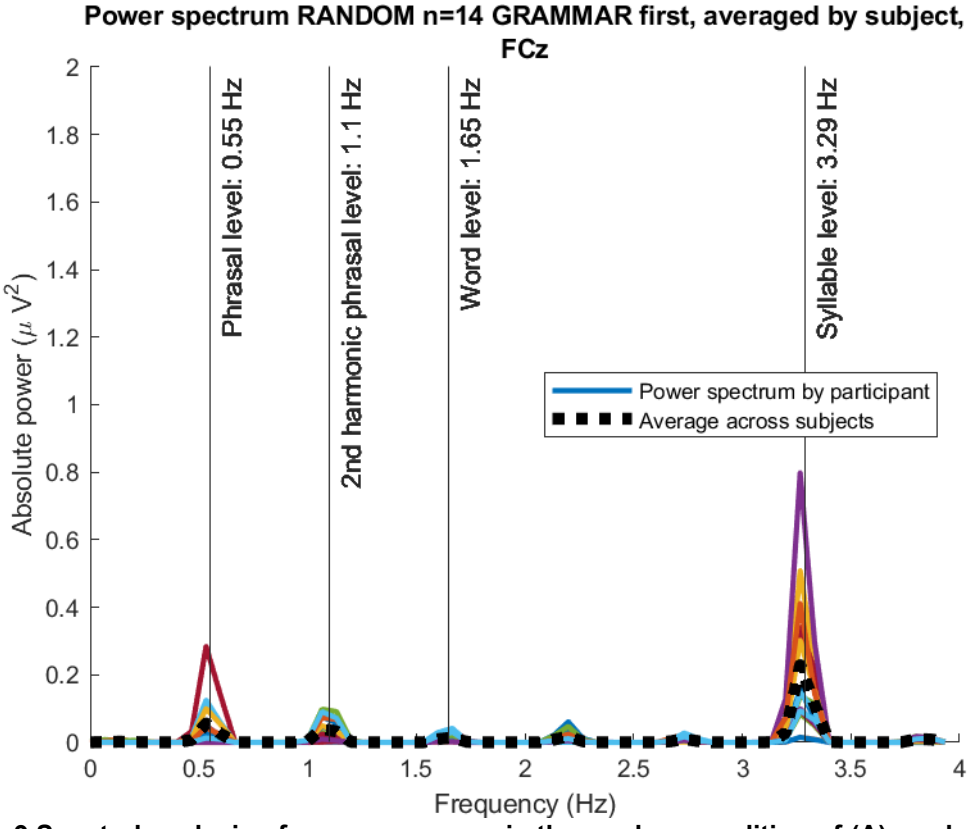


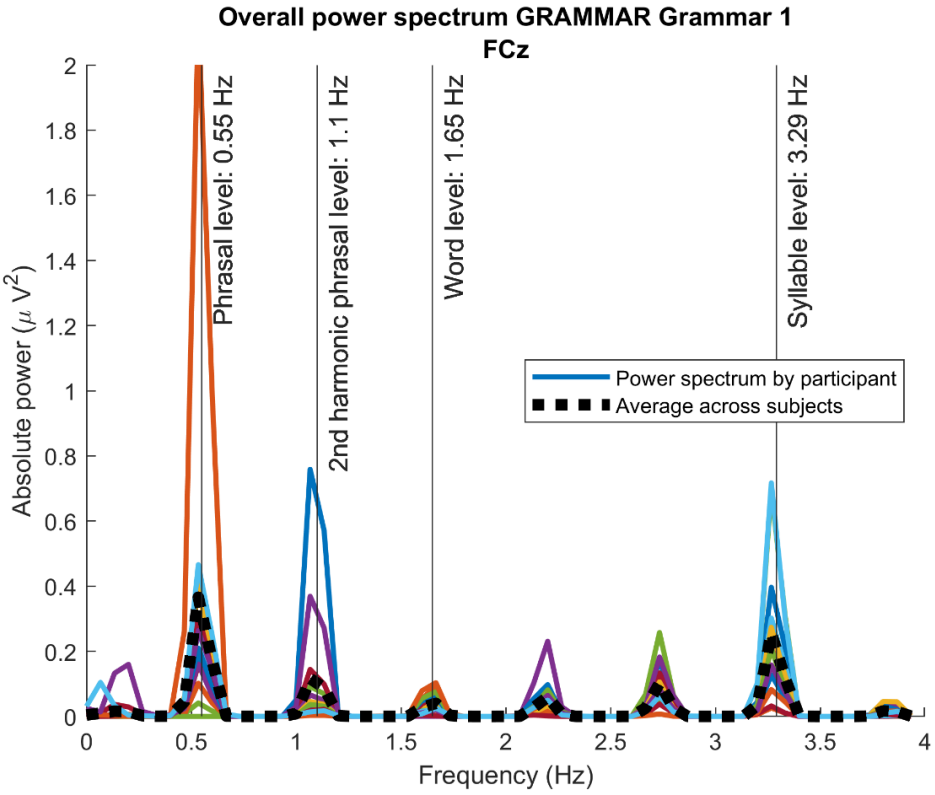
Fig. 9 Spectral analysis of power response in the random condition of (A) random first participants and (B) grammar first participants

3.2.1 Response by grammar type

Analysing the power response by grammar type shows higher activation at the phrasal frequency for grammar 1 (which includes the same outlier as in the random first grammar condition). Grammar 2 shows higher levels of average activation at the 2nd harmonic and word level. Both grammar types feature significant spikes at the syllabic frequency. The power spectra for the grammatical condition separated by grammar type are presented in Fig. 10A, B.

In the random condition, grammar 1 participants show limited spikes at the syllabic and phrasal frequencies. Grammar 2 participants show a similar response at the syllabic level to the responses of both grammar types in the grammatical condition, as well as a small spike at the phrasal frequency. An independent samples CBPT showed no significant difference in response between grammars at any of the frequencies of interest neither in the random nor the grammatical condition ($p > .05$ for all frequencies). The power spectra for the random condition separated by grammar type are presented in Fig. 11A, B.

A



B

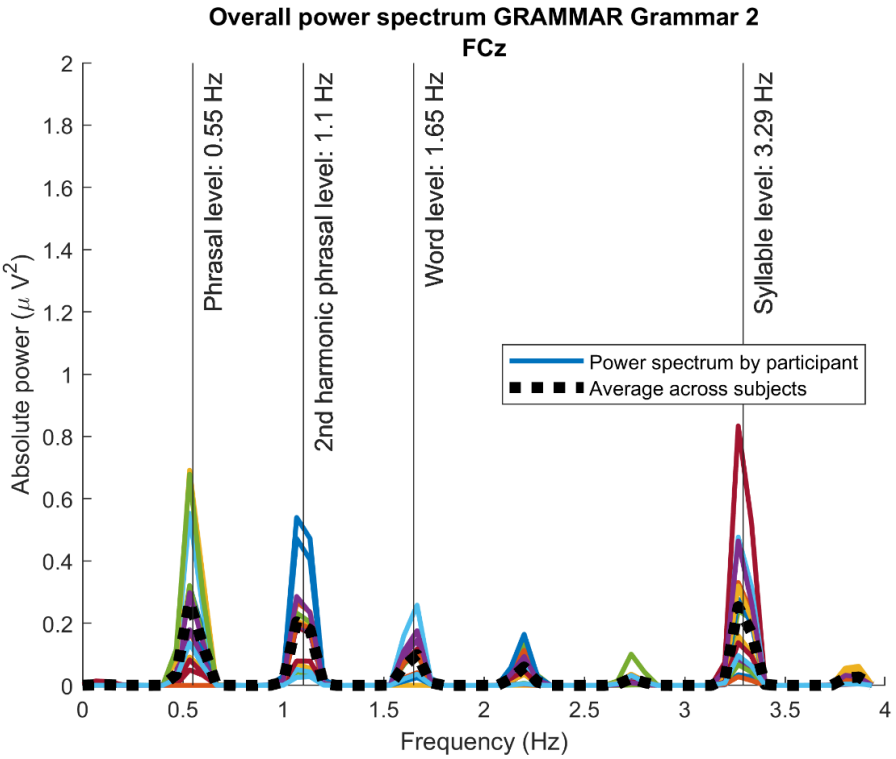
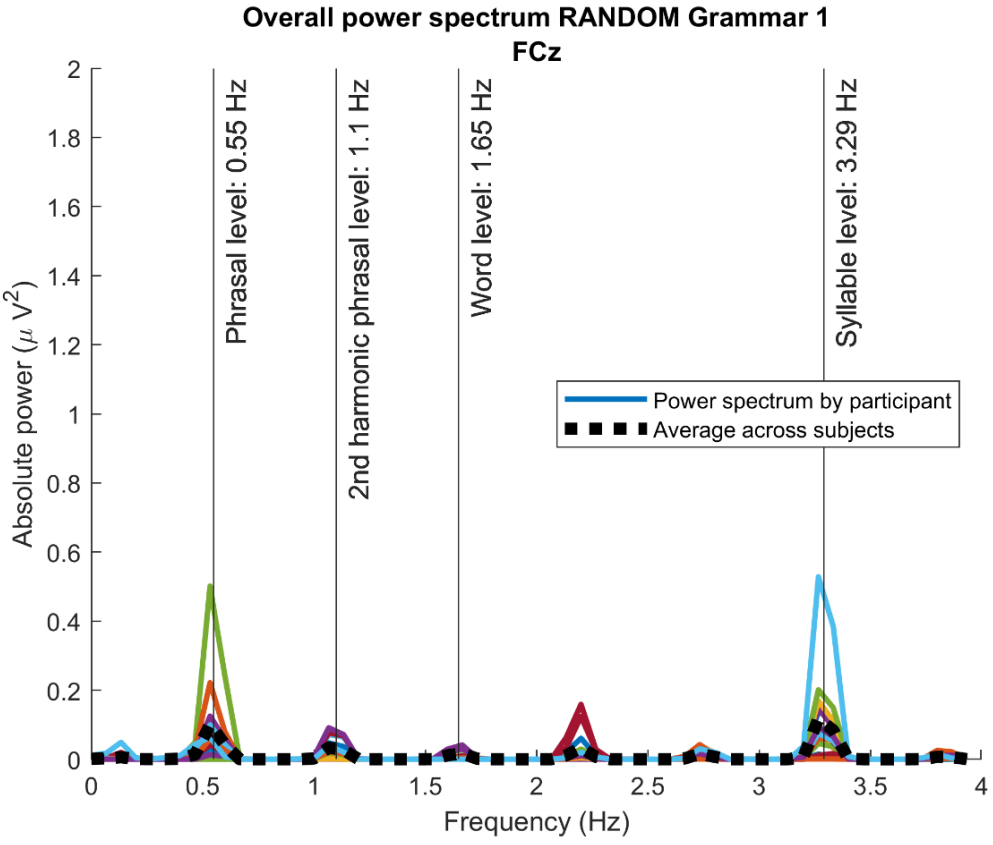


Fig. 10 Spectral analysis of power response in the grammar condition of (A) grammar 1 participants and (B) grammar 2 participants

A



B

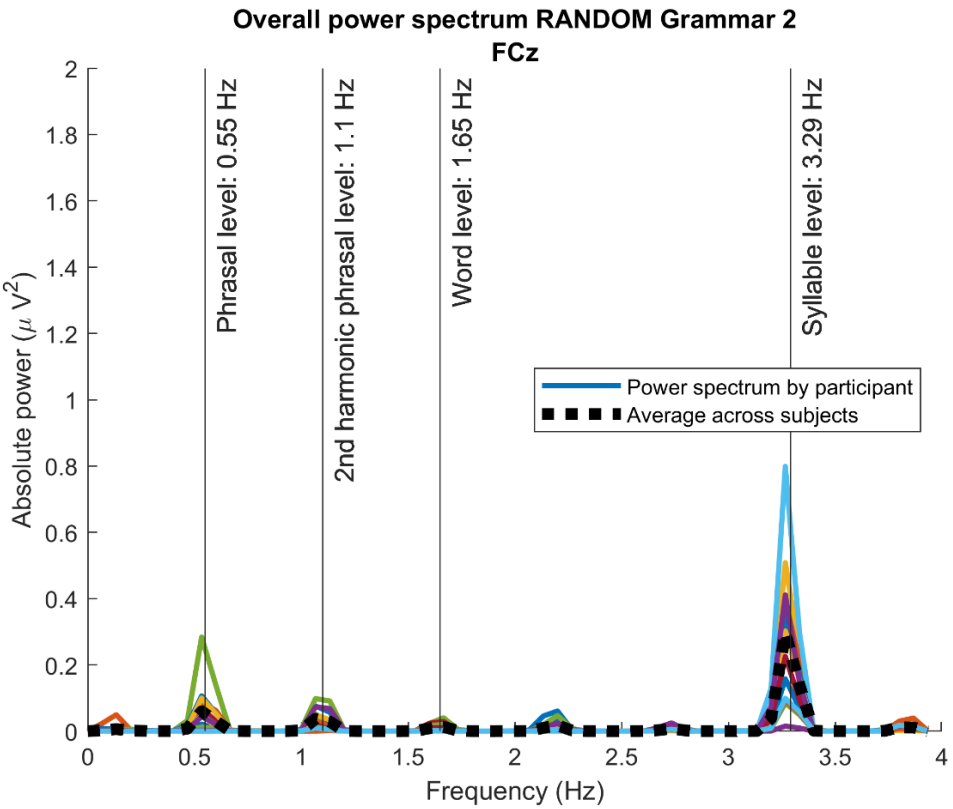


Fig. 11 Spectral analysis of power response in the random condition of (A) grammar 1 participants and (B) grammar 2 participants

3.3 Behavioural results

Average accuracy on the binary-mapped rating task across participants was 66.167% (median = 70%, SD = 24.378%). A one-sided one-sample t-test comparing participant performance to chance level returned a t-value of 3.632, compared to a t-value of 1.699 for chance level on 29 degrees of freedom at a 5% significance level. Participant performance was therefore significantly above chance level. Of the 29 participants, 6 performed below chance level, of which 4 performed more than 1 SD below chance. At a 5% significance level, the cutoff for performing significantly above chance on the 20 test sequences was at 70%, with 16 participants meeting or exceeding that level and thus being considered learners. 13 participants performed more than 1 SD above chance level. Normality of data was determined to be present based on visual inspection of histograms, which are provided in appendix 7.1. Evaluation of participant responses showed that there was no difference in the rate of false negatives compared to false positives, with a total of 101 correct sequences incorrectly rejected and 102 false sequences incorrectly approved. A boxplot of the accuracy across all participants can be seen in Fig. 12.

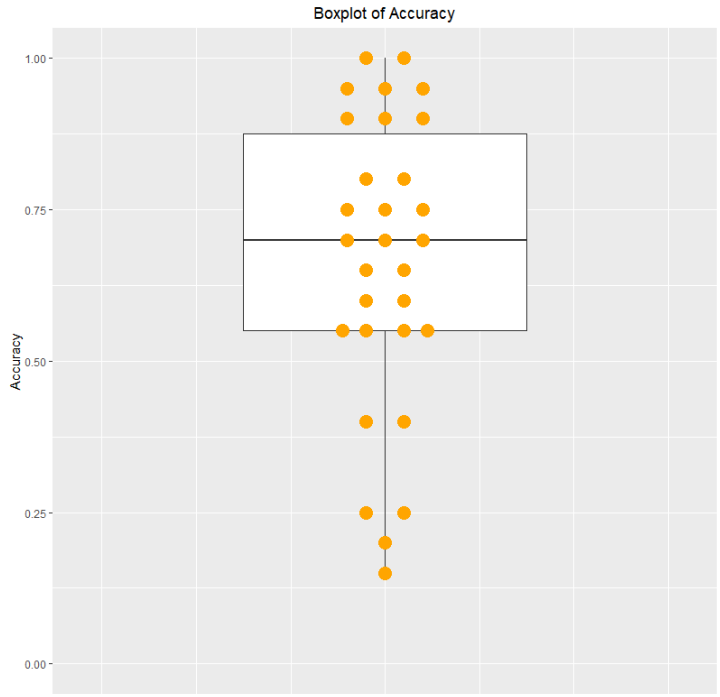


Fig. 12 Boxplot of accuracy across all participants

For the demographic analysis, ability to speak a foreign language was not included in the analysis, as every participant spoke at least one foreign language and 26 out of 31 participants spoke at least two foreign languages, not allowing for a statistically meaningful comparison to participants who spoke fewer languages. The factors included in ANOVA analysis for ordinal and categorical demographic data were level of education, multilingual background (growing up with more than one native language), and ability to play a musical instrument. A multivariate ANOVA showed no significant impact of any of these factors on accuracy ($p > .05$ for all factors).

Using a linear model with interaction terms to analyse the possible effects of age and reaction time on the accuracy of participants returned no statistically significant effect of either of the factors nor their interaction ($p > .05$ for all levels).

A Welch Two-Sample t-test revealed no significant difference in average accuracy between random first and grammar first participants ($t = -0.18$, $p > 0.05$).

A one-way ANOVA of the impact of grammar type on accuracy showed a significant difference between the level of accuracy for the different grammar types ($F = 7.34$, $p = 0.003$). Using Holm corrected pairwise t-tests by group revealed a significant difference in performance of participants with Grammar 1B compared to both Grammar 2A ($p = 0.01$) and Grammar 2B ($p = 0.017$). All other pairwise combinations of grammar types were not significantly different ($p > 0.05$). Consequently, grammars were remapped to a binary distinction between Grammar 1 and Grammar 2, summing both types of each grammar into one factor. A one-way ANOVA

revealed a significant difference in participant accuracy between grammar 1 and grammar 2, based on an F-value of 11.938 at a p-value of $p < 0.01$. A boxplot displaying the accuracy of the participants grouped by grammar type is shown in Fig. 13.

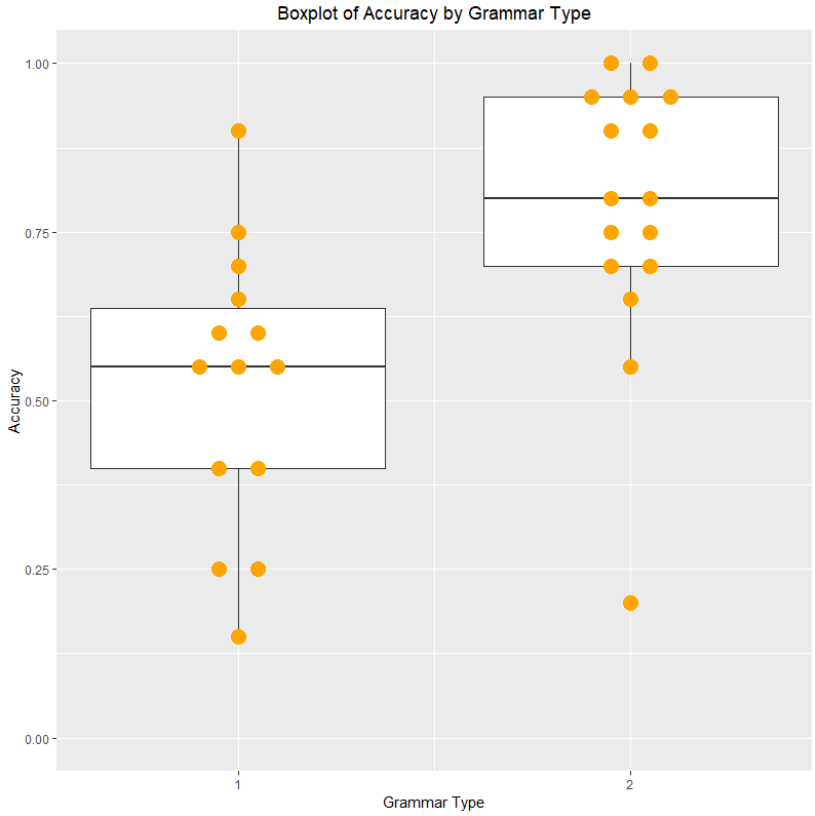


Fig. 13 Boxplot of accuracy by grammar type

4 Discussion

The present study investigated whether transitional prob abilities between syllables would be sufficient to elicit a neurophysiological and behavioural response to an NAD in a speech stream. Unlike previous studies (Buiatti et al., 2009; Ding et al., 2016; Getz et al., 2018; Müller et al., 2009), any non-statistical clues were excluded from the setup. Instead, the present study aimed to test whether a reliance on statistical learning facilitates the same entrainment effect that was found in studies with extra-statistical cues. Using the AGL paradigm from Weyers et al. (2022) and adapting it based on findings from Thompson and Newport (2007) and Getz et al. (2018), it was hypothesized that participants would show responses specific to the grammatical structure of the stimuli. Specifically, spikes in power response at the frequency of the syllables, words and phrases contained in the speech stream were expected. The adapted experimental procedure with a single, longer learning phase was also expected to improve participant accuracy in the behavioural task compared to Weyers et al. (2022). Due to conflicting results in the existing literature, no clear expectation was placed on the presence of

a possible suppression effect on the participant response to lower-level structures in the presence of higher-level hierarchical structures.

4.1 Main effect

The results show a clear difference in the response of the participants to the grammatical stimuli compared to the random stimuli. The presence of the significantly increased activation at the frequency of phrasal and word units that was expected indicates that the participants indeed may have responded to the higher-level structure in the speech stream. This supports the results from Weyers et al. (2022). Exceeding the results of that study, the present study even showed the response for the overall group of participants, with both average accuracy in the behavioural task as well as average power response at the frequencies of interest being significantly increased compared to chance/control level across all participants. The number of learners (participants whose behavioural performance was significantly above chance level) also increased, from 6 out of 24 participants in Weyers et al. (2022) to 16 out of 29 participants in the present study. This is despite the fact that the threshold for being considered a learner increased from 60.9% in Weyers et al. (2022) to 70% in the present study due to the lower number of test trials.

Both increases in performance are likely attributable to replacing the multiple shorter learning phases in Weyers et al. (2022) with a single, longer learning phase. This explanation matches the pattern shown in Getz et al. (2018), which showed a continual increase in response differentiation up until about four minutes of continual exposure to an artificial grammar stream (p. 139). The uninterrupted five minutes of learning in the present study used this fully, while the shorter, 1:28min learning phases in Weyers et al. (2022) ended well before reaching the plateau in response reported in Getz et al. (2018). Thus, the longer learning phases suggested in Getz et al. (2018) indeed seem more effective when comparing the results of the present study to the same stimuli in the shorter paradigm of Weyers et al. (2022).

Additionally, thanks to removing the intermittent test phases, the results can be more clearly attributed to a pure entrainment effect from the learning phase, as opposed to a possibly mixed effect from both learning and test phases in Weyers et al. (2022). The intermittent test phases in that study were speculated to give participants an opportunity to perform an ad-hoc adjustment to the stimuli, as during the test phases they were presented with individual phrases from the speech stream, thus revealing the structure and especially the length of the sequences contained in the stream. The participants could then have mapped this additional information onto the speech stream in the subsequent learning phases, easing segmentation of the stream and thus artificially enhancing the entrainment effect. In Bulca et al. (2019), the precursor to Weyers et al. (2022), 4 out of the 6 learners showed a significant increase in

performance in later test phases, with only 2 out of the 6 already performing above the threshold level during the first test phase (Bulca et al., 2019, p. 56). Of course, it is not clear how much of this increase over time is based on this speculative ad-hoc adaption from the intermittent test phases, and how much of it is due to the intended entrainment effect during subsequent learning phases. The fact that the participants in the present study performed significantly better and with a much higher ratio of learners while eliminating the possibility of such an ad-hoc adjustment suggests that the possible ad-hoc adjustment was not necessary to achieve high performance.

Unfortunately, the elimination of the intermittent test phases also means that it is not possible to track the development of the behavioural response over time as the amount of exposure increases, which would make a direct comparison clearer. To investigate this, one could adopt a combination of the present study with the multiple test phases from Bulca et al. (2019) and Weyers et al. (2022), where the five-minute learning phases from the present study are repeated multiple times as the shorter learning phases were in those studies. However, the already high performance in the present study suggests that a ceiling effect may occur, with performance only having limited room to increase across the multiple phases due to the fact that it is already very high. This change would also significantly increase cognitive strain on the participants, as the continual five minutes of exposure in the study were already described to be relatively taxing by some of the participants. Thus, additional learning phases may run into a counteracting fatigue effect.

A point worth discussing is how clearly attributable the responses of the participants are to the effects that we intended to measure. The response at the phrasal frequency clearly showed that the participants did segment the speech stream into 6-syllable chunks. However, the results of the EEG analysis do not offer any insight into whether they responded at this frequency because they recognized the phrase boundary between sequences, or whether they were simply tracking the fixed syllables in the stream. Because the second, fourth and sixth syllables all repeat in 6-syllable intervals, tracking these syllables would evoke a power response at the same frequency as the phrase boundary. Tracking the second and sixth syllable would also be sufficient for completing the behavioural task, as they are moved along with the first and fifth syllables in the violated condition. Because the two syllables stay constant throughout the entire learning phase, a change in those syllables may be sufficient to be registered by the participants as a violation, leading to a correct identification of the violated sequences without needing the phrasal level of analysis. This alternate strategy of tracking the fixed syllables therefore impacts the degree to which the results of the present study may be attributed to the intended causes. A possible way of identifying what the participants responded to would be to conduct an ERP component analysis during the test phase. If participants were tracking a syntactic structure, we would expect a P600 component to be triggered by the

violated sequences. If they were only tracking the syllables as syllables without a link to a hierarchical structure, we would not expect a P600 response. Unfortunately, participants may also perceive the fixed syllables as part of the syntactic structure of the sequences. This would mean that participants may show a P600 response even without analysing the sequences as phrases. An absence of a P600 response would also not necessarily indicate an absence of a hierarchical level of analysis.

In summary, the hypothesized neurophysiological response of the participants was found to be present, indicated by spikes in the power response at all frequencies of interest. As the same spikes were only present at the syllabic, but not the hierarchically higher levels of word and phrase in the random condition, these responses can be attributed to the structure of the speech stream. Due to the fixed syllable structure of the speech stream, the power response at the phrasal frequency cannot be directly linked to the presence of a phrase boundary at that frequency, as it may also be caused by simple tracking of the fixed syllables. The design of the present study does not allow to clearly differentiate between these two strategies.

As hypothesized, the performance of the participants in the behavioural task was improved significantly compared to the performance in Weyers et al. (2022), which can likely be attributed to the changes in the experimental procedure, most notably the prolonged continuous learning phase.

Looking beyond the main effect, the results of the present study invite several questions.

4.2 Performance by grammar

One of the more unexpected results found in the study is the significant difference in performance between grammar types. In the previous studies of Bulca et al. (2019) and Weyers et al. (2022), which used the same stimulus material, no such difference was discussed. As the EEG results presented in section 3.2.2 showed, the difference in power response between the two grammar types was small.

The response of grammar 2 participants at the syllabic frequency is at the same level of response as in either the random or grammatical condition of the participants in Weyers et al. (2022) and is uniform in both the grammatical and random block. The response of grammar 1 participants at the syllabic frequency in the random condition, meanwhile, is lower than the response at the syllabic frequency in any other condition in the present study, including in the grammatical condition of the same participants. It is also lower than the response shown in Weyers et al. (2022). It is not entirely clear what may cause this lower level of power response. In Weyers et al. (2022), there is no significant difference between the response at the syllabic frequency in the two different conditions. Contrary to this finding, it was expected that the participants would respond just as strongly or even more strongly (due to the absence of a

possible suppression effect) at the syllabic frequency in the random condition compared to the grammatical condition. When comparing the response of the grammar 1 group in the random condition to the responses of any condition in Bulca et al. (2019), Weyers et al. (2022) or the present study[12] including to their own response in the grammar condition, their response at the syllabic frequency is uniquely below the expected level of response. Because Weyers et al. (2022) featured the identical stimulus material (in a slightly different experimental setup) and produced a different result that aligns with the rest of the literature as well as the other conditions of the present study, it seems reasonable to assume that the grammar 1 group from the present study was underperforming. However, due to the highly individual nature of the power response and large differences in response between participants, comparisons between the power responses of different groups of participants hold only limited explanatory power. It does remain unclear why the response of the same group of participants at the syllabic frequency is higher in the grammatical condition than in the random condition, which will be discussed in more detail in section 4.4.

Besides the power response, the difference in behavioural response between the two grammar types is striking. Participants who received a variant of grammar 1 performed with a mean accuracy of 52.14% (SD = 21.01%), while participants who received a variant of grammar 2 achieved a mean accuracy of 78.44% (SD = 20.14%)[13]. A Welch Two Sample t-test confirms a significant difference between the performance of grammar 1 and grammar 2 participants ($t = 8.48$, $p < 0.001$). One-sample t-tests revealed no significant difference between grammar 1 participants and chance level performance ($t = 0.38$, $p > 0.05$) and a significant difference between grammar 2 participants and chance level performance ($t = 5.53$, $p < 0.001$). Interestingly, despite the significant difference in performance between groups, the variance was very similar between the two groups, with grammar 1 participants showing a variance of $4.41\%^2$ and grammar 2 participants $4.06\%^2$, leading to the similar standard deviations of 21.01% and 20.14% respectively as mentioned above. These differences are not statistically

[12] These two studies are the only viable direct comparisons of SSAEP response because they feature the same stimulus materials and analysis methods. Comparison to other studies is illicit because those studies employ different methods of analysis.

[13] Both reaction times and accuracy tests were conducted including the clear outlier participant in grammar 2. That participant has an accuracy of 20%, well below the next-lowest participant with 55% and the mean accuracy of 78% (for comparison, the least-accurate participant from grammar 1 had an accuracy of 15%, but there were two more participants with an accuracy of 25%, and a total of 5 out of the 14 participants with an accuracy below chance level). That participant also had an average reaction time of 0.72s, the second fastest reaction time of any participant. The low accuracy combined with the very low reaction time may be taken as an indication that the participant was performing in bad faith and simply randomly pressing a button in order to get through the experiment as fast as possible. However, achieving a 20% accuracy on a series of 20 binary guesses has a probability of 0.46%, meaning that it is implausible that one out of only 29 participants would achieve such a low accuracy by simply guessing at random. I have therefore included the results of that participant. It may be argued that they be excluded regardless, which would only increase the difference between the grammar types and push the average accuracy of grammar 2 above 80%. The decision to include or exclude the outlier therefore does not lead to any difference in analysis of the results.

significant on a Welch Two Sample t-test ($p > 0.05$). This pattern suggests that the difference in accuracy may have been caused by a baseline effect rather than an overall change in item difficulty across all items. Statistically, this baseline effect is present, with a type 3 ANOVA reporting a significant intercept ($F = 88.29$, $p < 0.001$). Such a baseline effect may have been caused by some items of grammar 2 being so easy to identify as correct that all participants always correctly identified them, leading to an overall increase in accuracy across all participants of group 2, but not of group 1. This would then lead to an overall increase in accuracy with comparable variance.

Examining the performance for each individual item during the test phase, there are no items from grammar 1 which had an accuracy rating of 100%. Conversely, there were three such items for grammar 2, two of type GOLU XXWE TAKI and one of type SILU XXWE MEKI. Additionally, one more item of type GOLU XXWE TAKI had an accuracy rating of 93.75%, with only one wrong answer out of 16. These four perfect or near-perfect items might suggest that a baseline effect is indeed partially responsible for the difference in average accuracy and the lack of a difference in variance between groups 1 and 2. However, inspecting the items that achieved these high accuracies does not suggest any linguistic reason why they would have been significantly easier to identify than any of the other items[14]. They do not include any words or part-words which were not included in other items, and the X elements which were used in these items were also used in other items. Notably, they were all correct items, meaning that participants had to accept them as correct rather than rejecting them as incorrect. They were also not consistently given as the first or last item in the test phase but distributed somewhere in the middle, excluding an ordering effect.

Another argument against treating these items as baseline outliers is the overall higher accuracy across many items in grammar 2 compared to grammar 1. The highest accuracy item from grammar 1 has an accuracy rating of 78.57%. There are ten grammar 2 items with a higher accuracy than that. 16 out of 28 grammar 2 items have an accuracy of 75+%, with only two grammar 1 items exceeding that threshold. This indicates that the overall accuracy of the participants was simply higher across the board rather than only on a few outlier items. The difference in baseline performance between the groups is therefore likely caused by an overall difference in processing difficulty rather than individual flawed items that skewed the distribution.

The syntactic structure of the stimuli was identical across both grammar types, meaning that this mismatch between grammar types must be caused by other factors than the grammatical structure of the sequences. The most obvious cause for this effect would lie in the phonological

[14] The perfect accuracy items mentioned here are items 13, 18, and 36. The near-perfect item is item no. 26. All items are provided in the supplementary materials.

features of the different syllables used in the different grammar types. Grammar type 1 featured sequences of the structure PELO XXFU TOBA and FILO XXFU SEBA, while type 2 featured the sequences GOLU XXWE TAKI and SILU XXWE MEKI. These sequences were constructed to control for phonological variation, meaning no systematic phonological features distinguish the two grammars. The variation in the X elements is arbitrary for both grammar types, and the total number of 37 syllables used per grammar type means that the effect a single syllable can have on the overall result is limited. Therefore, the phonological impact of the different X elements is unlikely to explain this effect. The items were also double-checked before starting the experiment in order to exclude sequences that would contain natural language words in German or English. Upon examining the items again post-hoc, no German words were found. In written form, some English words are included in the materials, but due to the German pronunciation of the synthesized speech, are phonetically distinct from the English words (e.g. the “go” syllable is pronounced [go] instead of [gəʊ] or [goʊ]; the syllable stream also contains the sequence “wet”, but this is pronounced [vet] instead of [wet]). If these words affected the ease of identifying correct sequences, we would expect an overall higher accuracy for items containing these words than for items which do not contain them. This is not the case, as there are no significant differences in accuracy between any of the four types of items (Canonical or reversed order for GOLUXXWETAKI and SILUXXWEMEKI) in grammar 2 ($p > 0.05$ on a two-tailed two-proportion z-test for all combinations of items). Taking a clue from the performance of the participants in the behavioural task, a possible phonological explanation is revealed. When examining the worst-performing items for Grammar 1 participants, 8 of the bottom 10 are of the FILO XXFU SEBA type, either in canonical or reversed order. Conversely, of the top 10 items by performance, only 3 are of FILO XXFU SEBA type. Excluding the varying X syllables, these items include three fricatives in the syllable onsets, compared to a maximum of one onset fricative for the other items. It is plausible that participants had increased difficulty in differentiating the different fricatives, leading to decreased performance. Even though the mean accuracy of 42.35% for the FILO XXFU SEBA sentences is lower than the 54.55% for the PELO XXFU TOBA sentences, a Welch Two Sample t-test analysing the difference in performance between sentence types 1 and 2 revealed no significant difference in accuracy between the two sentence types ($t = 1.55$, $p > 0.05$). Therefore, we cannot conclude that the increased number of fricative onsets caused the underperformance of grammar 1.

Another angle of differentiation between grammar 1 and grammar 2 is revealed by inspecting the reaction times by item. As discussed in section 3.2, there was no overall effect of reaction times on accuracy rating. However, comparing the reaction times in response to the two types of grammar using a Welch Two Sample t-test with Holm adjusted p-values reveals a significant difference in reaction times ($t = 4.44$, $p < 0.001$, 95%CI 0.58s – 1.54s). With a mean RT of 3.31s for grammar 1 compared to a mean of 2.20s for grammar 2, grammar 1 sequences were

associated with significantly slower reactions than grammar 2 sequences. An F-test confirmed that the variance in reaction time was also higher for grammar 1 sequences than grammar 2 sequences (mean RT 7.98s compared to 6.61s; $F = 0.33$, $p < 0.01$). This again is an indication of increased processing difficulty.

A different possible explanation could be demographic differences between the two participant groups of the two grammars. A Welch Two Sample t-test revealed a significant difference in age between the two groups, with group 1 being slightly younger than group 2 ($p = 0.049$, $t = -2.08$, 95%CI -5.70 – -0.02). None of the other demographic factors (Multilingual background, Ability to play a musical instrument and Educational background) differed significantly between grammar types ($p > 0.05$ for all factors).

Therefore, the demographic variance between the two participant groups also does not offer a plausible explanation for the difference in performance. On the contrary, the small but significant difference in age does support the assessment of grammar 1 being more difficult to process. This is because increased age leads to higher reaction times; the older group being faster contradicts this and is therefore more likely to indicate a difference in difficulty[15].

The large difference in accuracy combined with the significantly increased reaction times indicate that grammar 1 participants had much more difficulty evaluating the stimuli than grammar 2 participants, leading to decreased performance.

After examining these different angles of interpretation for the mismatch between the two participant groups, we can conclude that the mismatch was likely caused by an underperformance of the grammar 1 group rather than an overperformance of the grammar 2 group. While the accuracy of the grammar 2 group is quite high and higher than expected based on the results of Bulca et al. (2019) and Weyers et al. (2022), the changes to the experimental setup serve as an adequate explanation for this increase in performance. On the other hand, the seemingly increased difficulty for the grammar 1 group is not immediately attributable to a clear cause. Whether this difficulty is caused by a flaw in the stimuli or is an effect of an underperforming participant group is not ultimately clear. As there was no working memory test included in the study, there is no way of knowing whether this underperformance was simply caused by an unlikely but theoretically possible selection of a group with inherently lower level of cognitive performance. Due to the limited sample size, it is possible that by chance, participants were mismatched across groups in terms of cognitive ability. However, a few factors suggest that the problem is more likely to lie with the stimuli rather than the

[15] Admittedly, the effect of an average difference in age of around 3 years is expected to be marginal, as the average decline in reaction times is in single digit ms per decade of age difference (Hardwick et al. 2022), meaning that an average age gap of about 2.5years would be expected to lead to roughly 1ms difference in reaction time, which is negligible in comparison to the roughly 1s difference in average reaction time between the two groups.

participants. First, the performance of the participants is distributed in a random pattern around the 50% mark, which is the expected pattern when a normal group of participants takes random guesses (cf. histogram of accuracy for grammar 1, provided in appendix 7.1B). Second, the demographic analysis revealed no indication of a possible irregularity between the groups. Third, the overuse of fricatives in the onset of the syllables, while not leading to a statistically significant worsening of accuracy, may have contributed to an overall increase in the difficulty of identifying and differentiating the different syllables. It is also possible that some aspect of the stimuli used for grammar 1 was overlooked which may have made processing more difficult than expected.

Unfortunately, due to the anonymization of the data after the end of the data collection period, it is no longer possible to identify the participants and invite them to participate in the same study again with flipped group membership. A working memory or similar cognitive performance test would have helped to eliminate any doubts in the performance of the participants, and it may therefore be advisable to include such a test in future follow-up studies.

These findings significantly affect the validity of the overall results. As discussed in the previous section, overall performance across all participants in the behavioural task was significantly above chance level. Separating the participants by grammar revealed that this effect is entirely caused by the participants who were exposed to grammar 2, as performance of grammar 1 participants did not differ significantly from chance. Because there are no convincing linguistic or extra-linguistic explanations for this, the overall results, at least of the behavioural task, while not invalid, do not have the same validity as they would with a uniform performance across groups. While the reasons discussed above suggest that this mismatch is caused by an underperformance of group 1 rather than an overperformance of group 2, it should be recognized that something about the experimental setup did not function as designed. Nonetheless, the high accuracy and clear power response of the grammar 2 group show that the experimental setup works in principle, and the adjustments made to the experimental procedure based on the findings of the previous literature did lead to the predicted changes in response.

In summary, there is a significant difference in behavioural performance between grammar types 1 and 2, indicating that grammar 1 was harder to learn than grammar 2. This difference is not explained by the linguistic properties of the stimuli, nor are there other, extra-linguistic factors which offer a convincing explanation for this result other than possible but unlikely individual differences between the participant groups.

4.3 Theoretical implications

4.3.1 Suppression effect

Buiatti et al. (2009) found that the response of the participants at the syllabic frequency was inhibited in the presence of a word-level structure compared to the response to the same syllabic stimuli without the word-level structure. They postulated that this may be the result of an inhibition effect, limiting the response to lower-level structures in the presence of higher-level structures in order to be able to facilitate higher-level processing functions. A possible candidate for exhibiting this effect would be the word-level response in Weyers et al. (2022) as well as the present study. In both studies, the word-level response was significantly lower than at the phrasal level. One could assume that this was the result of a suppression from the phrasal-level structure onto the word level. However, the response at the syllabic level is unaffected by the presence of the higher-level structures, with no significant difference between the syllabic response in the random and the grammatical condition. This matches the findings of Ding et al. (2016), who found no suppression of the syllable-level response in the presence of two types of high-level hierarchical structures (phrases and sentences). Therefore, no global suppression effect onto all lower-level structures appears to exist. However, Buiatti et al. (2009) found a suppression effect using stimuli that were restricted to the syllabic and word level; their stimuli did not feature any hierarchical structures beyond the word level. Because the suppression effect seems to disappear in the presence of higher-level hierarchical structures like phrases or sentences, one could therefore conclude that syntactic operations prevent or counteract a suppression effect. This is supported by the effect found in the all-combined 5% condition of Thompson and Newport (2007), where increased variability at the syntactic level did not prevent an entrainment effect. From a linguistic perspective, embedding words into a more abstract, higher-level structure such as a sentence or even a discourse is known to lead to increased performance in word processing compared to processing words in isolation (Marslen-Wilson & Tyler, 1980, Stanovich & West, 1983). Higher-level hierarchical structures negating the suppression effect might be a plausible cause for this increase in performance. However, one of the problems with this explanation is that the directionality of the effect is unclear, i.e. whether the suppressed activation of the lower-level structure is the default case and the suppression is only suspended when the structure is embedded into a higher-level structure, or whether the unsuppressed activation of the lower-level structure is the default state and the suppression only appears in cases without the overarching linguistic structure of a sentential or phrasal embedding. One argument for assuming the latter interpretation is that natural language almost exclusively appears in context. The suppression effect could then be explained as a response to the lack of context encountered when processing the stimuli, where the absence of a higher-level linguistic

structure causes a diminished neural response at the lower level because the brain is attempting to construct a higher-level structure where there is none.

It does remain questionable why the response at the word level is so much lower than the response at the phrasal or syllabic level. A plausible ad-hoc prediction would have been to expect word-level dependencies to evoke stronger or at least comparable responses to phrasal-level dependencies as those higher-level dependencies involve more material over longer distances and feature higher-level processes than the ones at word level. The results from the present study as well as those in Weyers et al. (2022) contradict this expectation, showing instead a lower level of activation at the word frequency compared to the phrasal frequency. However, a mitigating factor could be found in the stimuli of the present study and Weyers et al. (2022), causing these relatively low responses at the word frequency. As mentioned in section 1.4, because no operations at the word level take place during the learning phase (unlike e.g. movement and optionality in Thompson & Newport, 2007), there is no reason for the participants to analyse the syllable stream as anything other than consisting of monosyllabic units. Segmenting the speech stream into an additional mid-level layer between the phrasal and syllabic levels does not yield additional information for identifying the underlying grammatical pattern in the speech stream. Because the second, fourth and sixth syllables constantly repeat in fixed intervals, segmenting the speech stream into 6-syllable chunks aids in the processing of the stream and in separating the fixed from the variable syllables. Thus, separating the stream at the phrasal frequency serves a functional purpose, even if it may not actually be indicative of the recognition of the NAD and may just indicate tracking of the fixed syllables as mentioned in section 4.1. Separating the phrases into three bisyllabic words instead of six individual syllables, however, yields no advantages in differentiating the roles of the different syllables. Even the NAD may be sufficiently entrained by simply linking the occurrence of the individual syllables rather than words. The lower response at the word frequency may then be caused simply by the word level being a less salient level of analysis than the phrasal and syllabic levels, without any relation to a hypothetical suppression effect.

Along the same lines of the findings in the present study, Getz et al. (2018) also showed that while participant responses differentiated significantly at the phrasal level over time, they did not differentiate at the word level (p. 139). At the same time, the behavioural tasks in Weyers et al. (2022) as well as the present study used the word level to introduce the syntactic violation during the test phase. Especially the present study focused on this aspect, where the violation was a pure word order violation. The success of the grammar 2 group in the behavioural task showed that the participants successfully employed the information gained during the learning phase to judge the felicity of the utterances in the test phase. However, as pointed out above, participants may have simply bypassed the word level by tracking only the fixed syllables. If

they had managed to identify the linear order of the fixed syllables during the learning phase and then heard a sequence during the test phase which violated this order, they may have recognized this as a violation of the structure of the sequence, without having to include a word level at any point of their analysis. Thus, the apparent mismatch between the high performance of the grammar 2 group in the behavioural task and the limited neurophysiological response at the word frequency should not be taken as evidence of a disjunction between the ability to perform word-level tasks and a neurophysiological response at the word frequency.

There could also be another factor in interpreting the data from the present study in relation to the suppression effect. Consider that each participant encounters each individual word with only half the frequency they encounter a phrase. Of course, each specific phrase only occurs as often as the words that it is made up of, but it seems plausible to suggest that learning a higher-level structure like a phrase is more generalizable than that of individual words, allowing the two different phrases in each grammar to be combined into a single linguistic structure more easily than the individual words. There are only two different phrases per grammar while, due to the varying X elements, there are more than a dozen different words, meaning that there is more variation between words than between phrases. This may contribute to the participants processing the phrases as similar structures and encouraging the separation of the speech stream into 6-syllable chunks, while the higher variation at the word level may inhibit the combination of syllables into 2-syllable chunks.

4.3.2 Simultaneous processing of hierarchical structures

At face value, the results of the present study appear to confirm the predictions on hierarchical processing from the existing literature. Aligning with Ding et al. (2016), the present study shows the simultaneous processing and integration of different hierarchical levels of language. On the basis of their findings, Ding et al. (2016) postulate that statistical cues are “generally not sufficient” (p. 163) for building these hierarchical structures, and instead underline the importance of tacit syntactic knowledge and capabilities of the participants (cf. p. 163). Ding et al. (2017) contrast simple N-gram prediction models as the basis for statistical learning against hierarchical linguistic knowledge, which they deem necessary in addition to statistical learning for processing complex linguistic contexts (pp. 570f). Due to the problem of tracking the fixed syllables described in the previous section, it is unfortunately not clear whether the response of the participants in the present study can be attributed to the recognition of a higher-level structure. A simple N-gram model that keeps track of every other syllable in 6-syllable intervals would be able to accurately capture the structure of the stimuli as well as differentiate grammatical from ungrammatical sequences in the test phase. While this does show that N-gram prediction is sufficient for modelling relatively complex linguistic operations like NADs, the syllable structure in the present study does not allow more insight into whether hierarchical

language-specific knowledge is needed or even involved at all in processing the speech stream. The small bit of evidence we can take from the present study as well as Weyers et al. (2022) and Bulca et al. (2019) is that participants did show power peaks at the 1-syllable, 2-syllable and 6-syllable frequencies. This does suggest that some type of hierarchical processing may have taken place, as the 2-syllable/word frequency would not have been needed to segment the structure of the syllable stream when applying purely statistical N-gram analysis, and the 6-syllable frequency response indicates that participants did analyse the syllable stream beyond the word level, tracking either the hierarchical structure or at least the re-occurrence of the fixed syllables in the next sequence. Because participants were not aware of the exact task they would be asked in the test phase but were made aware prior to the experiment that the syllable stream contained a pattern and that they should try and identify this pattern, they may have performed a broader search for hierarchical structures above the syllable level in order to find this pattern, facilitating responses at both the word and phrasal level.

4.4 Limitations

Compared to Weyers et al. (2022), the removal of the intermittent test phases, while eliminating the confound of additional training outside of the entrainment effect, also exposes a different problem of the stimulus design. When participants are simply listening to the syllable stream, only three syllable positions ever change between sequences. The systematic co-occurrence of each first syllable with the same fifth syllable is the only systematic change participants experience. Because no syntactic operations take place, participants have no incentive to analyse the stimuli at the word level. Because the syllable level is clearly the base level of analysis for the speech stream, they may be inclined to simply analyse the changing syllables at the syllabic level, “coindexing” the co-occurring syllable pairs without integrating them into a word with the second and sixth syllables. Combining the individual syllables into words is simply not necessary to build an appropriate structure to segment the speech stream, because no operation takes place at the word level that cannot be explained equivalently by analysing only at the syllable level, without having to build an entire extra layer of abstraction. The only time where a syntactic change occurs is during the test phase. In Weyers et al. (2022), this meant that the participants were first exposed to a syntactic operation in the first test phase, and subsequently went through three more learning and test phases, giving them the opportunity to adjust the learned structure and introducing a word level to the analysis. In the present study, the entirety of the learning takes place before the test phase, eliminating this possible adjustment. A remedying factor for this problem may be the inclusion of the variable X element in the third syllable position. Because that syllable changes arbitrarily and without relation to any other syllable, participants may be inclined to build word-level units in order to

more easily isolate the random variation in the third syllable from the systematic variation in the first and third syllables.

Another point worth briefly discussing is the absence of a significant difference in the CBPT results at the phrasal frequency for the participants who were given the random condition first compared to the participants who were given the grammatical condition first. The random first participants could have been expected to have an advantage in the identification of the phrasal structure of the speech stream. This is because they were already exposed to a test phase prior to being exposed to the grammatical speech stream and the corresponding test phase. While the syllable material in the test phase of the random condition was a random combination of syllables which had no structural relation to the material in the grammatical condition, the length of the phrases used in the test phase for the random condition was the same as the length of the phrases in the grammatical condition. This could have served as an aid in identifying the correct segmentation of the speech stream into sentences, which in turn would be expected to facilitate a stronger response at the phrasal frequency. The absence of a significantly increased activation at the phrasal frequency during the grammatical condition compared to the random condition means that this was not the case. This is also reflected in the behavioural performance of the participants, with no significant difference in performance between the two groups as evaluation with a Welsh Two-Sample t-test showed (mean accuracy of 65.33% vs. 67%, $p > 0.05$). It is possible that this advantage was quite small, in tune with the additional learning effect of the multiple test phases from Weyers et al. (2022) discussed section 4.1. The random first condition also gave participants additional time to adjust to the experimental setting, getting used to the sound of the synthesizer voice and the CV-syllable structure. This again could be interpreted as predicting an increased response in the participant in the random first condition. However, it is entirely plausible that these effects were completely negated by a small fatigue effect during the grammatical block caused by the previous workload during the random block. This absence of an effect is not a strong enough piece of evidence to warrant further speculation on this.

Another notable effect found was that even during the random condition, participants showed a small spike in power response at the phrasal frequency. As the random condition featured only randomly combined syllables with no underlying structure, there is no apparent reason in the stimuli why participants would respond specifically at that frequency. One possible explanation would be that the effect is caused by participants in the grammar first condition, in which they were trained to a speech stream that elicits a response at that specific frequency. They could then transfer over this segmentation strategy to the random condition, parsing the random stream of syllables by segmenting it into phrases of the same length as in the grammatical condition, even though the underlying structure no longer exists. This effect was also found in Weyers et al. (2022), where the response at the phrasal frequency was

significantly stronger in the grammatical condition but was also present to a lesser degree in the random condition despite the apparent absence of a specific stimulus at that frequency. Because that study was conducted with every participant receiving the grammatical block first, it was not possible to investigate whether such a transfer was the cause of this effect. As the results of the present study show, the spike at the phrasal frequency during the random condition is also present for those participants who were exposed to the random block first, as can be seen in the power response graph shown below in Fig. 14. Therefore, it must be caused by something other than this transfer effect.

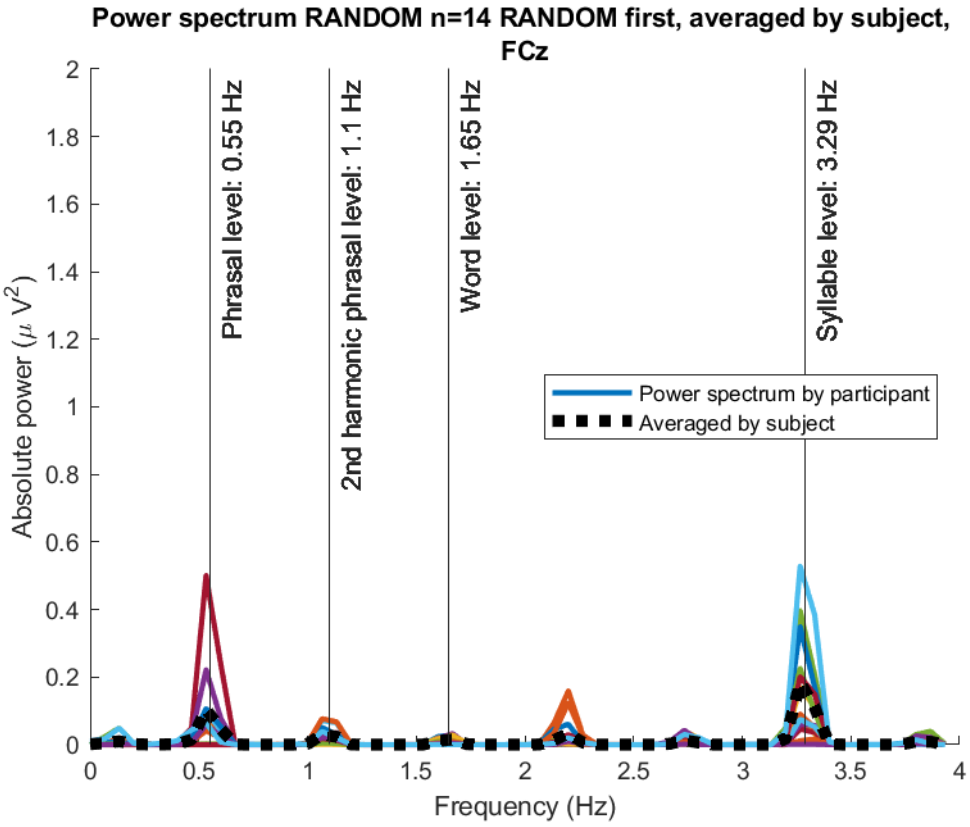


Fig. 14 Spectral analysis of power response in the random condition for participants with random block first

As the CBPT confirms, the spike at the phrasal frequency is significantly smaller than in the grammatical condition, and thus does not draw into question the main effect. It is, however, entirely unclear why this spike exists at all, as the stimulus material does not offer any cause for activation at this specific frequency. The spectral analysis also shows small responses at the frequency of the harmonics of the phrasal frequency. If the stimulus even produced harmonics, participants must have perceived some form of regularity at the phrasal frequency, even if no such regularity was intentionally introduced into the stimulus material. This also somewhat diminishes the validity of the response at the phrasal frequency in the grammatical

condition. Of course, the response in the grammatical condition was significantly stronger than in the random condition, but the presence of a small response at the phrasal frequency without any discernible basis for it in the stimulus material suggests that the response at the phrasal frequency in the grammatical condition may not be solely caused by a response to the phrasal structure. It is possible that the stimulus material contained an undiscovered feature that may provoke a response at the phrasal frequency regardless of the presence of a phrasal level at that frequency. A possible explanation would be that the selected syllable inventory overrepresented some phonological features while underrepresenting others. No such features were found even after additional post-hoc inspection of the stimuli. This fact increases the plausibility of the hypothesis that there is some hidden feature in the stimuli that causes this response, as the presence of this effect in both Weyers et al. (2022) as well as the present study eliminates the participant group as a possible cause for this.

4.5 Connections to future research and follow-up studies

An aspect of this study that warrants further investigation is the mismatch between grammar 1 and grammar 2 participant response in the behavioural task. Besides simply running the experiment again with a new group of participants, we can try to adapt the experimental setup of the present study to test for the different hypothesized causes of the underperformance of the grammar 1 group.

One of the possible factors for the underperformance of group 1 was group 1 simply being less attentive than group 2. One remedy for this factor would be to simply include a general cognition test performed prior to the experiment in the demographic analysis. More fundamentally, the role of attention in the acquisition of artificial grammars and the relation to the evoked brain responses has not been investigated for structures above the word level. Batterink and Paller (2019) investigated the impact of visual distraction on performance in a lexical decision task and a target detection task, asking participants to identify pseudowords which they heard as part of a speech stream. During the listening phase, participants were exposed to the speech stream containing the pseudowords as well as a series of kaleidoscopic images. In the Full Attention condition, participants were asked to ignore the images and focus on the speech stream, while in the Divided Attention condition, participants were asked to ignore the speech stream and were asked to perform a demanding recall task on the images (p. 60). Results showed that while the participants in the Full Attention condition showed higher performance in the target detection task, all participants were able to perform the target detection task to an “adequate” (p. 64) level. In the familiarity rating task, the groups did not significantly differ in any of the factors. Unfortunately, accuracy in the target detection task was only reported for the overall group but not by Attention group (p. 64). Nonetheless, their results show that participants are able to process a speech stream to a sufficient level for performing

rating and target detection tasks even when explicitly instructed not to pay attention to the speech stream and simultaneously distracted by a visual task. This demonstrates the robustness of implicit statistical learning from auditory stimuli even in adverse conditions. The paradigm used in the present study would be a candidate for investigating whether this ability also persists for structures beyond the word level. Alternatively, the 5% all combined condition from Thompson and Newport (2007) that was also used in Getz et al. (2018) could be well-suited for investigating the role of attention beyond the word level. During the learning phase, one could use the same visual distractors as in Batterink and Paller (2019), only swapping out the speech stream. The tasks in the test phase would have to be adapted to correspond to higher-level structures. A rating task could be constructed similarly to the word rating task in Batterink and Paller (2019), asking the participants to rate entire or part-sequences instead of words. The target detection task could also be performed on entire sequences, exposing the participants to a short speech stream consisting of a series of sequences matched in length to the sequences from the learning phase. After hearing the target sequence once, participants could then be instructed to press a button when they think the part of the speech stream they just heard matches the speech stream from the listening phase.

The main obstacle in drawing clear conclusions from the neurophysiological results of the present study is the inability to disambiguate whether the participants were indeed tracking a hierarchical structure or were simply tracking the fixed syllables in the speech stream, which would elicit the same EEG response and would enable them to perform the behavioural task without needing a higher level of analysis. For a way of investigating whether participants were actually tracking the phrasal structure or just the fixed syllables, one could adapt the stimuli of the present study so that each of the two phrases in each grammar uses a distinct set of syllables for the second, fourth and sixth syllable. This would mean that the word-internal TPs remain the same, still encoding syllable pairs 1&2, 3&4, and 5&6 as words. Unfortunately, there is no clear and easy way of introducing these changes to the fixed syllables while keeping the characteristics of the speech stream used in the present study comparable. An option would be to retain the alternating syllables in the first and fifth position and simply introduce another sequence with a different set of syllables to the speech stream. This approach would mirror the design of the present study, essentially combining both grammars into one. It does, however, introduce a change in TPs at the phrasal level. In the present study, the transition from the sixth syllable of a sequence to the first syllable of the following sequence featured a TP of .5 (excluding cases of multiple repetitions of the same variant of the sequence). Introducing another sequence with different fixed syllables but still two variations for the first and fifth syllable would mean that the TP at the phrase boundary would be reduced to .25, and the probability to encounter each fixed syllable would drop from 1 to .5 (which means tracking the fixed syllables is no longer sufficient for identifying the structure of the speech stream).

Within each sequence, however, the TPs remain identical, with the TPs within words remaining at 1 and the NAD between first and fifth syllable also remaining at 1. Fig. 15 shows a transition probability graph for such a grammar, with the two sequences constructed with syllable material from the present study.

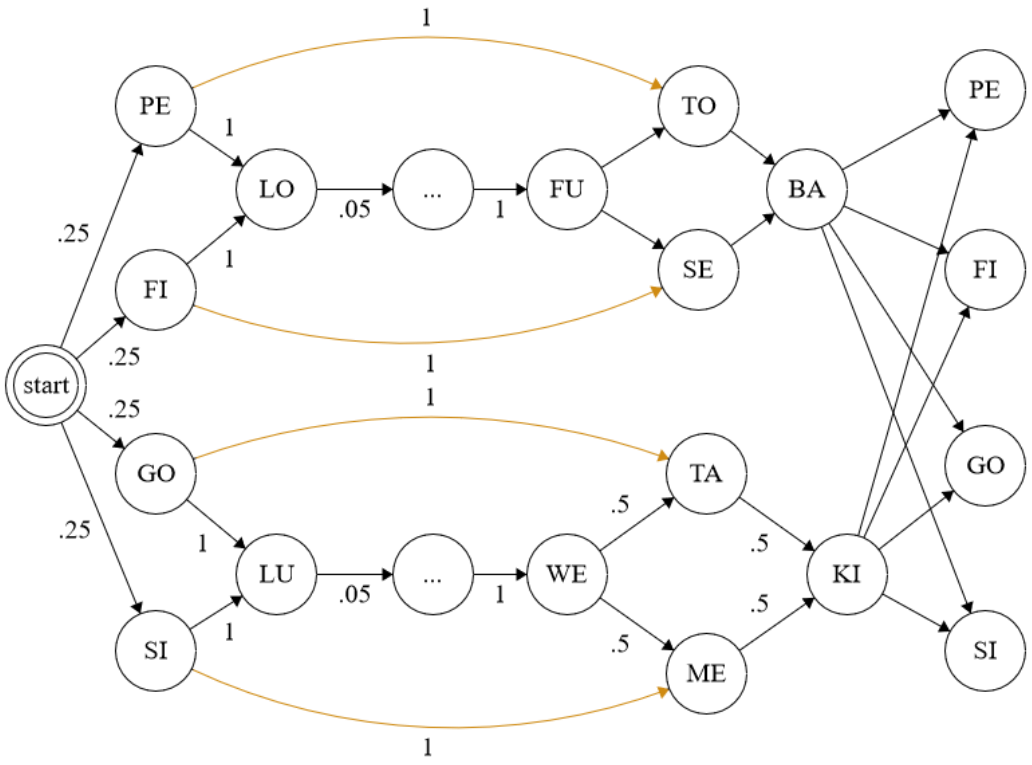


Fig. 15 Transitional probability graph for a combined grammar condition[16][17]

Because the fixed syllables no longer systematically appear at the same interval as the phrasal boundary, a neurophysiological response at that frequency would indicate tracking of either the phrasal boundary or/and the NAD. Therefore, because there are fewer stimuli repeating at that frequency, a weaker response at the phrasal frequency may be expected.

Another open question is the relevance of the word level for processing the stimuli of the present study. In order to more closely investigate this, one could run a slightly adapted version of the present study. Keeping the learning phase identical, a variation of the violation used in the test phase of the present study may provide additional insight into the question of whether

[16] TP labels for the transitions from the final syllable of the sequence to the initial syllable of the following sequence are omitted to preserve legibility, they are .25 for all transitions. Omitting the variants for the third syllables was also done for reasons of legibility, the … state is a placeholder for the 20 different variants for that syllable. NADs indicated in orange.

[17] This grammar would require even closer attention to the phonological properties of the speech stream due to the increased number of syllables used and the increased number of different transitions between syllables. Just adopting the syllable material from the present study as shown in Fig. 15 would not work; it was only chosen for demonstrative purposes.

the participants performed a word-level analysis. Instead of swapping the order of entire words like in the present study, one could instead only swap the first and fifth syllables, leaving the second and sixth syllable in position. This would ensure that participants are not simply looking for a change in the fixed syllables as discussed above. While participants may also just track the individual syllables and recognize the violated order without requiring analysis above the syllabic level, this change at the syllabic level may not be recognized as a syntactic violation. This means that an additional ERP component test during the violated sequences of the test phase could show whether participants recognized the violation of the canonical sequence at the syllabic or a higher level. If they were only tracking the syllables without a hierarchical framework, no P600 response would be expected to occur, while such a response is expected if the participants were tracking the hierarchical structure of the speech stream. However, if no P600 occurs, this is not conclusive evidence of participants not tracking the hierarchical structure of the sequences, as the absence of a P600 response is not definitive proof that no hierarchical processing has taken place. If participants do indeed exhibit a P600 response, this would indicate that they did indeed perform hierarchical processing and recognized the change in syllables as a structural violation. If the lower level of activation at the word level does correlate with a lesser entrainment of the underlying structure, we would also expect participant performance in the behavioural task to be below that in the present study. This is because the word level analysis is more relevant to the behavioural task, as the participants can no longer rely on tracking of the fixed syllables to disambiguate the canonical and violated sequences.

5 Conclusion

The present study presented an investigation of the ability of adult healthy participants to process a speech stream containing an NAD based purely on distributional statistics. In doing so, it replicated the study of Weyers et al. (2022), while making some adjustments to the experimental procedure based on findings from previous research, most notably the use of a single, uninterrupted learning phase instead of multiple shorter ones. The neurophysiological results show that participants responded at all frequencies of interest, indicating hierarchical processing of the underlying linguistic structure of the speech stream. However, due to the specific design of the stimulus material, it cannot ultimately be determined whether the response at the frequency of the higher-level structure of the speech stream is a direct indication of a response to the hierarchical structure, or whether participants were simply tracking syllables that repeated regularly at the same interval as the NAD. The behavioural results showed a significant improvement in overall performance of the participants compared to Weyers et al. (2022), with the overall group performing significantly above chance level, which was not the case in Weyers et al. (2022). While this improved performance is likely

attributable to the longer learning phase, investigation of the responses by grammar type revealed that only the grammar 2 group performed significantly above chance. The difference in performance between the two grammar types was an unexpected result, with grammar 1 appearing to be harder to process for the grammar 1 group than grammar 2 was for the grammar 2 group. While minor linguistic differences between the grammar types may have contributed to a difference in performance between the two groups, it is ultimately not clear what caused the mismatch in performance between the groups.

One of the lessons learned for future research is the difficulty of attributing the neurophysiological results to the effect that was intended to be measured. In the case of the present study, the fact that the NAD occurred at the same frequency as the fixed syllables means that, even though we see clear responses at the frequencies of interest, the conclusions that can be drawn are limited, as there is no way of telling whether the participants responded to the higher-level structure or directly to the repeating syllables. Designing an AGL paradigm that avoids this problem and enables the learning of higher-level structures purely based on statistical distribution of the syllables would provide more reliable insight into the limits of statistical language learning in adults. The study also showed that in the presence of relatively small sample sizes, admission of a cognitive ability test may be useful to eliminate a possible confound by individual differences between participants. The study also provided some evidence against a suppression effect on lower-level structures, but more research is needed to clarify whether such an effect exists and under which conditions it may occur. The fundamental question of how much of human language learning can be adequately explained by purely statistical learning and how much pre-instilled linguistic background is necessary remains open and a key motivator for future AGL research. As the present study showed, rigorous design of the experimental materials as well as their thorough analysis continue to be key prerequisites for meaningful contributions to answering that question.

6 References

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *cortex*, *90*, 31-45.

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the neural bases of implicit and statistical learning. *Topics in cognitive science*, *11*(3), 482-503.

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage*, *44*(2), 509-519.

Bulca, Ö., Nettermann, K., & Walkenhorst, B. (2019). The Investigation of Artificial Grammar Learning Using Auditory Steady-State Evoked Potentials. Study project report. Institute of Cognitive Science, University of Osnabrück.

Choi J., Cutler A., & Broersma, M. (2017) Early development of abstract language knowledge: evidence from perception–production transfer of birth-language memory. *Royal Society Open Science* 4: 160660 https://doi.org/10.1098/rsos.160660

Chomsky, N. (1957). Syntactic structures. *Mouton*.

Chomsky, N. (1965). Aspects of the theory of syntax. *MIT Press*.

Culbertson J., Koulaguina E., Gonzalez-Gomez N., Legendre G., & Nazzi T (2016). Developing knowledge

of nonadjacent dependencies. *Developmental Psychology* 52, 2174–2183.

Delorme A., & Makeig S. (2004). EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods* 134, 9-21.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience* 19(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Melloni, L., Tian, X., & Poeppel, D. (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, cognition and neuroscience*, *32*(5), 570-575.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Vrecken, O.V. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 3*, 1393-1396.

Finley S. (2015). Learning nonadjacent dependencies in phonology: transparent vowels in vowel harmony. *Language*, *91*(1), 48–72. https://doi.org/10.1353/lan.2015.0010

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829-834. https://doi.org/10.1177/0956797611409589

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science* 6 (2), 78-84. https://doi.org/10.1016/S1364-6613(00)01839-8

Getz, H., Ding, N., Newport, E. L., & Poeppel, D. (2018). Cortical tracking of constituent structure in language acquisition. *Cognition* 181, 135–140. https://doi.org/10.1016/j.cognition.2018.08.019

Gilbert, A., Lee, J. G., Coulter, K., Wolpert, M. A., Kousaie, S., Gracco, V. L., Klein, D., Titone, D., Phillips, N. A., & Baum, S. R. (2021). Spoken word segmentation in first and second language: When ERP and behavioral measures diverge. *Frontiers in Psychology* 12: 705668. https://doi.org/10.3389/fpsyg.2021.705668

Hardwick, R. M., Forrence, A. D., Costello, M. G., Zackowski, K., & Haith, A. M. (2022). Age-related increases in reaction time result from slower preparation, not delayed initiation. *Journal of neurophysiology*, *128*(3), 582-592.

Hauser M. D., Chomsky N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579.

Höhle B., Schmitz M., Santelmann L. M., & Weissenborn J. (2006). The recognition of discontinuous verbal dependencies by German 19-month-olds: Evidence for lexical and structural influences on children's early processing capacities. *Language Learning and Development* 2, 277–300.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive psychology 29*(1), 1-23.

Kleiner M, Brainard D, Pelli D, 2007, What's new in Psychtoolbox-3? Perception 36 ECVP Abstract Supplement.

Klem G.H., Lüders H.O., Jasper H.H., & Elger C. (1999). The ten-twenty electrode system of the International Federation: The International Federation of Clinical Neurophysiology. *Electroencephalography and Clinical Neurophysiology Suppl*. 52, 3-6. PMID: 10590970.

Kronrod, Y., Coppess, E. & Feldman, N.H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review* **23**, 1681–1712. https://doi.org/10.3758/s13423-016-1049-y

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, *50*(2), 93–107.

Kuhl, Patricia K. (2010). Brain mechanisms in early language acquisition. *Neuron* 67 (5), pp. 713-727. https://doi.org/10.1016/j.neuron.2010.08.038

Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. *Phonetica*, *5*(1), 5-54.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*(1), 1-71.

The MathWorks Inc. (2023). MATLAB version: 9.14.0 (2023a), Natick, Massachusetts. https://www.mathworks.com

Moser, J., Batterink, L., Hegner, Y. L., Schleger, F., Braun, C., Paller, K. A., & Preissl, H. (2021). Dynamics of nonlinguistic statistical learning: From neural entrainment to the emergence of explicit knowledge. *NeuroImage 240*, 118378.

Müller, J. L., Oberecker, R., & Friederici, A. D. (2009). Syntactic learning by mere exposure – An ERP study in adult learners. *BMC neuroscience 10*, 1-9.

Müller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences*, *109*(39), 15953-15958.

Müller, J. L., Milne, A., & Männel, C. (2018). Non-adjacent auditory sequence learning across development and primate species. *Current opinion in behavioral sciences*, *21*, 112-119.

Müller, S., Bildhauer, F., & Cook, P. (2021). German clause structure: An analysis with special consideration of so-called multiple frontings. Language Science Press.

Neville H., Nicol J.L., Barss A., Forster K.I., & Garrett M.F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of cognitive Neuroscience* 3, 151–165. PubMedID: 23972090

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data**.** *Computational Intelligence and Neuroscience* 2011: 156869. https://doi.org/10.1155/2011/156869

Picton, T. W., John, M. S., Dimitrijevic, A., & Purcell, D. (2003). Human auditory steady-state responses. *International journal of audiology*, *42*(4), 177-219.

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

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

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language* 47(1), 172-196. https://doi.org/10.1006/jmla.2001.2839

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, New Series, 274(5294), 1926–1928.

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

Seeck, M., Koessler, L., Bast, T., Leijten, F., Michel, C., Baumgartner, C., He, B., & Beniczky, S. (2017). The standardized EEG electrode array of the IFCN. *Clinical neurophysiology*, *128*(10), 2070-2077. https://doi.org/10.1016/j.clinph.2017.06.254

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General, 112*(1), 1-36.

Statistik Austria (2022). Bildungsstand der Bevölkerung. Last retrieved on 26.07.2024 from https://web.archive.org/web/20240111111346/https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bildung/bildungsstand-der-bevoelkerung

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development*, *3*(1), 1-42.

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

van Heugten, M., & Shi, R. (2010): Infants' sensitivity to non-adjacent dependencies across phonological phrase boundaries. *Journal of the Acoustical Society of America* 128, 223-228.

Weyers, I., Walkenhorst, B., Nettermann, K., Bulca, Ö, Gruber, T. & Mueller, J. L. (2022). Neural entrainment reveals simultaneous, implicit learning of adjacent word and non-adjacent phrasal structure. Poster at *14th Annual Meeting of the Society for the Neurobiology of Language*. Philadelphia, USA.

7 Appendix

7.1: Histograms of accuracy