Neural representation of speech segmentation and syntactic structure discrimination

Funding Body
This research was funded by the Max Planck Society for the Advancement of Science (www.mpg.de/en)

International Max Planck Research School (IMPRS) for Language Sciences
The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond. More information can be found at www.mpi.nl/imprs

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-41-7
Cover design and lay-out by Shuang Bi
Printed and bound by Ipskamp Drukkers, Enschede

Neural representation of speech segmentation and syntactic structure discrimination

Proefschrift

ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus, prof. dr. J.H.J.M. van Krieken,

volgens besluit van het college van de decanen

in het openbaar te verdedigen op

maandag 31 oktober 2022

om 12.30 uur precies

door

Fan Bai

geboren op 1 december 1986

te Baicheng (China)

Promotor:

Prof. dr. Antje S. Meyer

Copromotor:

Dr. Andrea E. Martin

Manuscriptcommissie:

Prof. dr. Uta Noppeney

Prof. dr. David Poeppel (ESI Frankfurt, Duitsland)

Dr. A.V.M Kösem (Centre de Recherche en Neuroscience de Lyon, Frankijk)

Neural representation of speech segmentation and syntactic structure discrimination

Dissertation

to obtain the degree of doctor

from Radboud University Nijmegen

on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,

according to the decision of the Doctorate Board

to be defended in public on

Monday, October 31, 2022

at 12.30 pm

by

Fan Bai

born on December 1, 1986

in Baicheng (China)

Supervisor:

Prof. dr. Antje S. Meyer

Co-supervisor:

Dr. Andrea E. Martin

Manuscript committee:

Prof. dr. Uta Noppeney

Prof. dr. David Poeppel (ESI Frankfurt, Germany)

Dr. A.V.M Kösem (Lyon Neuroscience Research Centre)

Contents

1 General introduction

Spoken language plays a key role in daily communication. For any type of spoken language, speech can be measured as sound waves, as the production process leads to the fluctuation of air pressure (Fry, 1979; Morse, America, & Physics, 1948; Stevens, Egan, & Miller, 1947). The temporal and spectral dynamics that are embedded in the speech stimuli include almost all the information that the speaker wants to express. However, that the information in the speech stimulus can only be extracted by speakers of the same language indicates that a language's set of rules must be applied. The process by which we extract linguistic information from speech seems automatic and effortless; however, the large body of literature on speech perception and comprehension suggests that a series of complicated operations are required. In this thesis, I focus on investigating the neural representation of two fundamental and critical processes, which are speech segmentation and syntactic structure discrimination.

1.1 Speech segmentation

A speech signal is usually continuous in time. In order to understand the information that is embedded in the speech signal, it must be segmented into basic linguistic units (Cutler, 2012; Grosjean, 1980). However, unlike written language, the boundaries between linguistic structures, such as syllables, words etc., are not explicitly marked in a spoken form. So how are listeners still able to segment speech?

Studies using traditional approaches suggest that speech segmentation benefits from the prosodic information. Cutler and Butterfield (1992) analyzed mishearing and natural slips by English listeners and found that boundaries were preferentially inserted before stressed syllables, and deletion of boundaries occurred before weak syllables. For instance, '*conduct ascends uphill*' could be reported as '*the doctor sends her bill*' (inserting boundaries before the strong final syllables in each word), or '*sons expect enlistment*' might be reported as '*sons*

expectant listen' (inserting a boundary before the stressed syllable '*-list'*, but also deleting the word boundary before weak syllable '*en-*'). The misperceptions suggested that listeners preferred to segment speech by considering strong syllables as word onsets (Cutler, 2012; Cutler & Butterfield, 1992).

In addition, using the word-spotting paradigm, where participants were asked to spot a real word in spoken nonsense strings, Cutler and Norris (1988) found that a real word (e.g. *mint*) was easier to extract from a nonsense string when it was combined with a weak syllable (e.g. *mintef*) than when it was associated with a strong syllable (e.g. *mintayf*). The authors proposed that the strong-strong structure led to a separation between the two syllables because strong syllables are often considered as onsets of words, i.e. *mintayf* separated into *min* and *tayf*. The postulation of a boundary between *min* and *tayf* made it much harder to detect the word *mint* in *mintayf* than *mintef* because one additional process of recombination of speech materials across the boundary was needed.

In brief, investigations using empirical and behavioral approaches suggest that acoustic level information, such as prosody and stress pattern, is beneficial in speech segmentation (Cutler et al., 2008; Evans, Saffran, & Robe-Torres, 2009; Fear, Cutler, & Butterfield, 1995; Hay et al., 2011; Norris et al., 2006; Pelucchi, Hay, & Saffran, 2009; Peña et al., 2002; Romberg & Saffran, 2010; Saffran, Newport, & Aslin, 1996; Smith et al., 1989).

However, speech segmentation using cues only at the phonological level, such as prosody or stress pattern, appears to omit the role of higher-level linguistic information. As such, many theories posit that linguistic structures could be extracted via an endogenous inference process (Bever & Poeppel, 2010; Brown, Tanenhaus, & Dilley, 2021; Friederici, 1995; Hagoort, 2013; Halle & Stevens, 1962; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas, 2020; Meyer, Sun, & Martin, 2020; Phillips, 2003; Poeppel & Monahan, 2011).

An MEG study conducted by Luo and Poeppel (2007) suggested that low frequency (~ 4 to 7 Hz) phase coherence reflects speech intelligibility and the extraction of basic linguistic units. In their study, participants were asked to listen to speech stimuli, for which the intelligibility was manipulated from low to high.

By decomposing neural activity into the frequency domain, the authors found that there was stronger phase coherence at the theta band (~4 to 7 Hz) when the speech stimuli were highly intelligible as compared to when the stimuli were degraded. The results indicated that the evoked neural response at low frequencies reflects the cortical analysis of speech signals. As the range of ~4 to 7 Hz approximately corresponds to the rhythm of the syllables in natural speech (Ding, Patel, et al., 2017; Doelling et al., 2014; Pellegrino, Coupé, & Marsico, 2011), the authors concluded that this low-frequency phase coherence might reflect the extraction of syllables, and therefore, syllables could be primitive units for the neural representation of speech. Similar results have been found in several other studies (Doelling et al., 2014; Howard & Poeppel, 2010; Luo & Poeppel, 2007; Peelle, Gross, & Davis, 2013).

Knowing syllables could be the primitive units for speech processing, researchers were looking for neural representations of linguistic structures that build upon syllables. An influential study conducted by Ding et al. (2016) suggested that speech segmentation can be performed via grammatical chunking or syntactic integration. More importantly, the authors found that the occurrence rate of hierarchical linguistic structures could be reflected by the intensity of the neural oscillations with the same frequencies. In their experiments, the researchers artificially synthesized three types of isochronous syllable sequences (four syllables per second, where the rhythm for syllables was 4 Hz) in Mandarin Chinese. The type-one sequences had a built-in hierarchy, where every two syllables form a phrase (the occurrence rate of phrases was 2 Hz, or two phrases per second) and every four syllables form a sentence (the occurrence rate of two-phrase sentences was 1 Hz, or one sentence per second). In contrast, the type-two and type-three sequences were control conditions involving random syllable sequences played forward and backward, respectively.

After presenting these three types of stimuli to Chinese participants, the cortical activities reflected the rhythm of linguistic structures at different levels simultaneously; i.e., there were 1 Hz, 2 Hz and 4 Hz peaks to reflect the occurrence rates of sentences, phrases and syllables for the type-one sequences, but only a 4 Hz peak to reflect the rhythm of syllables for the type-two and type-three sequences. More interestingly, when presenting these Chinese speech stimuli to English

speakers who did not understand Chinese, only a 4 Hz peak was appeared to reflect the rhythm of syllables for all three types of sequences. The contrast between various types of stimuli (three peaks for the type-one sequences vs. one peak for the control conditions) and different types of language users (three peaks for Chinese speakers vs. one peak for English speakers) suggested that the cortical response varied with the rhythm of linguistic structures and high-level linguistic knowledge (e.g., about the grammatical or syntactic relationships between units). Therefore, the authors concluded that a mechanism of grammatical chunking or syntactic integration was engaged, and the neural response that tracked the occurrence rates of hierarchical linguistic structures reflected the endogenous process of unit extraction. The neural tracking effect has been replicated and simulated in several recently conducted works (Gui, Jiang, Zang, Qi, Tan, Tanigawa, Jiang, Wen, Xu, Zhao, et al., 2020; Gwilliams & King, 2020; Jin, Lu, & Ding, 2020b; Jin et al., 2018a; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018; Martin & Doumas, 2017; Martin & Doumas, 2019; Meyer & Gumbert, 2018; Obleser & Kayser, 2019; Zhou et al., 2016).

Undoubtedly, speech segmentation benefits from high-level grammatical or syntactic knowledge. However, statistical information, such as the transitional probability between linguistic units, plays a key role in linguistic structure extraction and incorporates relationships between units at different levels. A behavioral study by Saffran, Newport, and Aslin (1996) suggested that speech segmentation can be performed using only the transitional probability (TP) between syllables (the probability that one syllable follows another, e.g., the probability of *Pi* given *Tu*). More specifically, the researchers constructed two-minute continuous syllable sequences consisting of four three-syllable nonsense words (e.g., *TuPiRo*) repeated in random order. Then, they aurally presented the two-minute speech stimuli to eight-month-old infants, after which a discrimination task followed, where the previous nonsense words as well as new three-syllable words were played to the infants. Using the infants' listening time (indexed by observing when they gazed at the researchers), the authors found that infants spent more time listening to the new words as compared to the old ones. The results are interesting as they suggest that infants could extract the three-syllable structures from the continuous speech stream using statistical information. Specifically, during the two-minute listening stage, the TP relationships between

syllables (the TP between syllables within a word was 1 and the TP between word boundaries was 1/4) constitute an implicit cue that leads to the segmentation by the infants. As no semantic or acoustic cues were available during this listening stage, and the only cue that could be used to extract the 'words' was the TP, the authors concluded that statistical information was the key for segmenting the speech stimuli.

In sum, numerous investigations in both psycholinguistics and neuroscience have suggested that acoustic-level information (e.g., stress pattern and prosody), high-level linguistic knowledge (e.g., grammatical and syntactic knowledge), and statistical regularities (the transitional probabilities between linguistic units) are all beneficial factors in speech segmentation.

1.2 Syntactic representation

Extracting basic linguistic units, such as syllables or words, from speech is not enough for language comprehension. Succesfully extracting the words *the*, *red* and *vase* from the phrase *the red vase* does not suffice to determine how these words are syntactically structured to form meaning. Therefore, in addition to extracting linguistic units, building the relationships between linguistic units is a necessary step for speech comprehension.

Behavioral studies and early analyses of natural speech have suggested the existence of abstract mental representations of syntactic structures, which are independent of lexical items. Levelt and Kelter (1982) found that shopkeepers tended to repeat the form of a question in their answers. For example, when asked *(At) what time do you close?* the shopkeepers tended to answer: *(At) five o'clock*, matching the utterance to the surface form (with or without preposition) of the question. Weiner and Labov (1983) also found that the best predictor of the occurrence of a passive structure in an interview was the recent occurrence of another passive structure. A laboratory study using syntactic priming by Bock (1986) also confirmed the facilitation effect for abstract syntactic structures. Specifically, she asked participants to repeat sentences or describe pictures during a task where some of the sentences were actually primes for picture descriptions. She found that participants tended to use the same syntactic structure as in

previously repeated sentences to describe pictures; i.e., participants were more likely to describe a picture with a passive rather than active structure after repeating a passive prime sentence.

All of the above-mentioned meta analyses and behavioral investigations indicate the existence of abstract syntactic representation, and therefore form a basis for electrophysiological research on syntactic representation. Electroencephalography (EEG) was the first neural imaging technique to visualize the process of syntactic processing in language comprehension. Osterhout and Holcomb (1992) conducted an EEG study and found that the neural response was sensitive to syntactic anomaly in a phase-locked manner. In their experiment, two types of sentences, e.g., *the broker hoped to sell the stock was sent to jail* (type-one sentence) and *the broker persuaded to sell the stock was sent to jail* (type-two sentence), were visually presented to participants word by word. Both types of sentences contain a clausal complement (e.g., *to sell the stock*). The intransitive verb *hoped* in the type-one sentence allows the clausal complement to be easily attached to the main clause; however, the transitive verb *persuaded* in the type-two sentence is most commonly combined with a direct object. When it is followed by the phrasal complement *to sell*, the word *persuaded* needs to be reanalyzed as a past participle rather than a past-tense finite verb form, e.g., *the broker (who was) persuaded to sell the stock was sent to jail*. Therefore, the type-two sentence is syntactically anamolous compared to the type-one sentence. Using these types of manipulated stimuli, the authors examined the syntactic anomaly recognition by measuring event-related potentials (ERPs). As expected, a positive phase-locked component (right hemisphere dominant) was found for the type-two sentences with a peak at approximately 600 ms after the infinitive *to* when compared to type-one sentences. The authors defined this component as 'P600' and considered it as a reflection of syntactic anomaly. This result is important, as it is the first direct evidence to show that syntactic processing can be reflected in phase-locked (evoked) neural responses. Building on this seminal research, a large number of ERP studies have confirmed the robustness of this effect (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Hagoort, Brown, & Groothusen, 1993; Osterhout & Mobley, 1995; Patel et al., 1998).

To form compositional meaning, individual words need to be chunked into syntactic structures (Baddeley, Hitch, & Allen, 2009; Bonhage et al., 2017; Chomsky, 2002; Cowan, 2016; Meyer et al., 2017). The process of forming syntactic relationships using extracted linguistic units depends on the encoding and retrieving of these units (Futrell, Mahowald, & Gibson, 2015; Gibson et al., 2000; King & Just, 1991; Lewis, 1996; Lewis, Vasishth, & Van Dyke, 2006; Meyer et al., 2015; Meyer, Obleser, & Friederici, 2013; Meyer et al., 2012; Nicol, Fodor, & Swinney, 1994). As syntactic processing builds on successful speech segmentation, at least two linguistic units are required to form a compositional structure. The syntactic processing might happen after the segmentation of necessary linguistic units or it could overlap with the stage of segmenting linguistic units. For instance, a syntactic phrase *the red vase* is formed by the brain after or possibly at the same time as the extraction of the lexical components (*the*, *red* and *vase*). As mentioned earlier, in the study of Osterhout and Holcomb (1992), the P600 also appeared after the extraction of the word of interest *to*, and the ongoing positive deflection overlapped with the processing of subsequent words. The dependency of the syntactic integration on the segmentation of linguistic units, and the potential overlap in time with non-syntactic processing stages, makes isolating syntactic processing challenging.

Meyer et al. (2017) conducted an EEG study in which the researchers presented their participants with syntactically ambiguous German sentences. The sentences were disambiguated by prosodic cues that marked either the ending or the continuation of a phrase. For instance, the authors inserted a long pause at the end of the word *verklagte* ('sued', type-one) or the end of the word *Mörder* ('murderer', type-two) in the sentence *Der Klient verklagte /den Mörder/ mit dem korrupten Anwalt* ('The client sued the murderer with the corrupt lawyer'). After listening, the participants answered questions such as *wer hatte den korrupten Anwalt?* ('Who had the corrupt lawyer?'), probing the attachment of the prepositional phrase. Because of the different positions (prosodic cues) of the inserted long pause, the type-one sentences would form a one-phrase sentence. For example, the prepositional phrase *mit dem korrupten Anwalt* ('with the corrupt lawyer') and the preceding object phrase *den Mörder* ('the murderer') form the joint phrase *den Mörder mit dem korrupten Anwalt* ('the murderer with the

corrupt lawyer'). Therefore, the answer to the probing question should be 'the murderer had the corrupt lawyer'. However, the type-two sentences would form two phrases, e.g., the prepositional phrase *mit dem korrupten anwalt* ('with the corrupt lawyer') forms a separate phrase, interpreted as linking to the subject phrase *der Klient* ('the client'). Therefore, the answer to the probing question should be 'the client had the corrupt lawyer'. The behavioral results showed the participants' sensitivity to the prosodic cues. They made significantly more two-phrase choices for type-two sentences as compared to type-one sentences. More importantly, the delta band (< ~ 4 Hz) oscillatory phase robustly separated the two grouping choices; i.e., the delta band oscillations separated the one-phrase grouping from the two-phrase grouping. These results suggest that delta band oscillations play a key role in the mental representation of syntactic structures. In line with Meyer et al. (2017), Bonhage et al. (2017) also found that the intensity of delta band oscillations was increased when participants encoded sentence fragments (for which syntactic structures can be formed) as compared to random word lists (for which syntactic structures cannot be formed).

As mentioned earlier in this section, encoding and retrieving extracted linguistic units is crucial for syntactic structure construction, and research into this has suggested the involvement of alpha band oscillations in verbal working memory (Haegens et al., 2010; Obleser et al., 2012; Ten Oever, De Weerd, & Sack, 2020; Wilsch & Obleser, 2016). Alpha band oscillations were also shown to be related to auditory attention (Strauß, Wöstmann, & Obleser, 2014; Wöstmann et al., 2016; Wöstmann et al., 2015; Wöstmann, Lim, & Obleser, 2017). Furthermore, a neural physiology model of speech perception considered alpha band neural oscillations as a 'top-down' perceptual gating control (Ghitza, Giraud, & Poeppel, 2013; Giraud & Poeppel, 2012). In short, the above-mentioned investigations support the role of alpha band oscillations in generalized auditory processing, which implies that alpha band oscillations may not be specific to speech processing. However, studies have also shown that neural activities at the alpha band reflect speech intelligibility (Becker et al., 2013; Dimitrijevic et al., 2017; Meyer et al., 2012), which indicates the engagement of alpha band oscillations in syntactic or semantic processing. Overall, the question of whether or not alpha band oscillations are related to high-level speech processing, e.g. syntactic representation, is still debatable.

Lastly, work using modeling to investigate the encoding of acoustic features has added new information and broadened the scope for exploring syntactic representation. Studies using the spectral-temporal response function (STRF) have shown that acoustic features in speech can be reliably reflected in the low-frequency neural response via a phase-locked manner (Ding & Simon, 2012a, 2012b, 2013b) and phonemic-level processing can also be shown in the low-frequency entrainment to speech (Di Liberto, O'Sullivan, & Lalor, 2015; Donhauser & Baillet, 2020; Weissbart, Kandylaki, & Reichenbach, 2020). The results of these modeling works revealed that low-frequency neural responses have a role in representing acoustic features. However, the extent to which syntactic-level information can be reflected by the encoding of acoustic features remains an open question.

1.3 The current thesis

Building on the above literature review, in this thesis, I present an investigation into the neural representation of speech segmentation and syntactic structure discrimination.

In Chapter 2, I focus on the role of statistical information, i.e., the transitional probability, in speech segmentation. The effect of cortical activity that tracks the rhythm of hierarchical linguistic structures is an intriguing phenomenon (Ding et al., 2016) as it might suggest that the brain uses grammatical or syntactic knowledge to construct linguistic structures at different levels via an endogenous inference approach. However, as Saffran, Aslin, and Newport (1996) suggested, the relationships between linguistic units could also be reflected by statistical information, i.e., the transitional probability. Specifically, by comparing the studies of Saffran, Aslin, and Newport (1996) and Ding et al. (2016), we found that the high-level linguistic information (e.g., grammatical or syntactic relationships) coexists with the statistical regularities (e.g., the transitional probability) in syllable sequences with a built-in hierarchy, where syllables, words and sentences appear with a fixed rhythm. Therefore, the cortical tracking effect might be driven by both of these factors. Figuring out whether the neural activity that tracks the rhythm of linguistic structures could be introduced solely by statistical information is important, as structuring units via grammatical chunking or syntactic integration

requires high-level linguistic knowledge, whereas linguistic knowledge is not necessarily involved when structure extraction is conducted using statistical information. In addition, if unit extraction could be performed using statistical information, would it be reflected by the cortical tracking effect? In the study described in Chapter 2, we constructed several types of syllable sequences which were controlled for both linguistic and statistical properties, and used in six MEG experiments with Dutch participants. We focused on determining the role of statistical information in the speech segmentation, and checking whether the units' extraction via statistical information could be reflected by the cortical tracking effect.

In Chapter 3, a generalization question is considered. Specifically, I report another set of MEG experiments with Chinese participants. The hypothesis is that if speech segmentation can be conducted using statistical information and the segmentation process is reflected by the cortical tracking effect, then having users of a different language perform the same experiments as in Chapter 2 will not change the effect, because speech segmentation using a non-language-specific factor, i.e. transitional probability, is a generalized perceptual process. Therefore, I investigate whether speech segmentation via statistical information can be accomplished by a different type of language user, and more importantly, if the neural representation of this segmentation process varies between different types of language users.

In Chapter 4, I describe an EEG study exploring how the phase-related neural activity reflects the discrimination between two types of syntactic structures (i.e., phrases and sentences). In each trial of the experiment, Dutch participants listened to a speech stimulus, which was either a phrase, e.g. *de rode vaas* ('the red vase'), or a sentence, e.g. *de vaas is rood* ('the vase is red'). To extract the optimized effect driven by syntactic structure discrimination, we matched the physical and semantic properties across the two types of stimuli (for details see Chapter 4). As this is an exploratory study, Chapter 4 addresses three questions. First, as previous studies have shown that low-frequency phase measures play a fundamental role in syntactic integration (Bonhage et al., 2017; Meyer et al., 2015; Meyer et al., 2017), we asked whether the low-frequency (< 8 Hz) phase coherence would reflect the discrimination between phrases and sentences. Second, if the brain could separate

the two types of syntactic structures, would this be reflected by the functional connectivity via temporal synchronization (phase connectivity)? Lastly, referencing the generalized speech perception model proposed by Giraud and Poeppel (2012), we asked if the low-frequency phase entrains with the high-frequency amplitude that exists during spoken language comprehension, and if so, whether this coupling mechanism separates the two different types of syntactic structures.

Chapter 5 presents a study that used the same dataset as in Chapter 4. We performed additional sets of analyses to show how syntactic structure discrimination would be reflected in the intensity-related brain measures. Furthermore, I describe modeling work using the STRF on the encoding of acoustic features. More concretely, as I mentioned before in the literature review on syntactic representation, there is still an ongoing debate about whether the alpha band neural oscillations are related to high-level (e.g. syntactic) speech processing. By considering this unsolved issue, we first asked whether the induced neural activity would reflect syntactic structure discrimination, and in particular, whether alpha band neural oscillations are involved in syntactic structure discrimination. Second, functional connectivity studies often paid attention on temporal synchronization (phase connectivity), so in this study, we further explored whether the neural networks that are constructed by the intensity of the neural activities (intensity connectivity) would separate the phrases from the sentences. Finally, as recently conducted works have shown that the acoustic features can be selectively represented in the neural response (Di Liberto & Lalor, 2017; Ding & Simon, 2012a, 2012b, 2013b), we use a computational modeling approach, the STRF, to explore how acoustic features are phase-locked encoded in both temporal and spectral dimension to separate the phrases from the sentences.

In the concluding section, Chapter 6, I summarize the core findings on the neural representation of speech segmentation and syntactic structure discrimination. This is followed by a discussion of the relationships between these findings and those of previous studies, to uncover the implications and contributions. Finally, I discuss the research questions that have arisen from our studies and might be tackled in future investigations.

Note that a certain amount of overlap exists across the chapters, as the same sets of MEG experiments were conducted with Dutch participants in Chapter 2 and Chinese participants in Chapter 3, and the same EEG dataset was analyzed from different directions in Chapter 4 and Chapter 5. Thus, some of the critical information in the general descriptions is reiterated in each chapter.

2 | Neural representation of speech segmentation via statistical inference

Abstract

A fundamental question in speech comprehension is how continuous speech signals are segmented by the brain. A study by Ding et al. (2016) suggested that the rhythm of hierarchical structures in speech can be reflected in the neural activities at the corresponding frequencies. They explained the phenomenon as the cortical tracking of linguistic units by endogenous grammatical chunking. However, online chunking clearly benefits from the statistical information (e.g., transitional probability) between linguistic units. Saffran, Aslin, and Newport (1996) found that infants could extract 'words' from fluent 'speech' after a two-minute exposure, in which transitional probabilities were considered as a key in the speech segmentation. A natural question raised by comparing the conclusions of these two studies is whether the cortical tracking effect could be driven by the transitional probabilities. We conducted a six-session magnetoencephalography experiment with Dutch native speakers to investigate the role of transitional probabilities (TPs) in the cortical tracking effect. Using the discrete Fourier transform with generalized eigen-decomposition, our analysis indicated that cortical tracking could be introduced solely by TPs, and the effect might not be a pure reflection of unit-chunking using high-level linguistic knowledge.

2.1 Introduction

Speech contains an abundance of acoustic features in both the time and frequency domains (Shannon et al., 1995; Smith, Delgutte, & Oxenham, 2002; Zeng et al., 2005). While these features are crucial for speech comprehension, they do not themselves signpost the linguistic units and structures that give rise to meaning. Spoken language comprehension therefore relies on listeners to go beyond the information given and infer the presence of linguistic structure based on their knowledge of language. As such, many theories posit that linguistic structures – ranging from syllables to words to syntactic structures – are constructed via an endogenous inference process (Bever & Poeppel, 2010; Brown, Tanenhaus, & Dilley, 2021; Friederici, 1995; Hagoort, 2013; Halle & Stevens, 1962; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas, 2020; Meyer, Sun, & Martin, 2020; Phillips, 2003; Poeppel & Monahan, 2011). Recent studies have begun to investigate the neural activity that corresponds to the emergence of linguistic structure (Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018; Martin & Doumas, 2017; Meyer & Gumbert, 2018), in particular in terms of the temporal and spatial dynamics of brain rhythms.

An influential study by Ding et al. (2016) found that the occurrence rate of linguistic structures (syllables, phrases and sentences) in speech can be reflected in neural activities at the corresponding frequencies. They consider this effect to be a neural representation of the construction of linguistic structures. In one experiment of their study, the speech stimuli were isochronous syllable sequences, in which syllables were aurally presented four times per second. The authors manipulated the relationship between the syllables so that every two syllables formed a phrase and every four syllables formed a sentence. This yielded rates of syllables, phrases, and sentences of 4 Hz (four syllables in one second), 2 Hz (two phrases in one second), and 1 Hz (one sentence per second), respectively. When participants who knew the language of the stimuli listened to the sequences, neural activities with the same frequencies as these hierarchical linguistic units were observed (4 Hz, 2 Hz and 1 Hz). However, in participants who did not know the language, the neural response only showed a peak at 4 Hz, which corresponded to the occurrence rate of syllables. It is reasonable to explain the phenomenon as

cortical tracking of linguistic structures by endogenous syntactic integration because the additional peaks in the neural responses at the occurrence rate of phrases (2 Hz) and sentences (1 Hz) in participants who knew the language could reflect processes that chunk syllables into higher-level linguistic structures. Similar findings have been shown in many experimental and computational studies (Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018; Martin & Doumas, 2017; Martin & Doumas, 2019).

However, online chunking clearly benefits from variation in transitional probability (TP) between linguistic units. Saffran, Aslin, and Newport (1996) found that eight-month-old infants could extract artificial words from fluent speech after a two-minute training. Specifically, the researchers constructed a continuous speech stream consisting of four three-syllable nonsense words, e.g., *TuPiRo*, repeated in a random order. After two minutes of exposure to the sequence, a discrimination task followed, in which the trained 'words' and new three-syllable 'words' were played one by one to the infants. Using the infants' listening time (indicated by gazes to the speaker) as an index, the researchers found that the infants listened longer to new words as compared to the trained words. The results implied that infants could discriminate between the two. Since no semantic or acoustic boundaries were available during the training session, and the only cue to extract the 'words' from the continuous syllable sequence was the transitional probability between syllables (TP within words was 1, TP between words was 1/3). The authors thus suggested that the infants used a computational mechanism based on the statistical properties of the stimuli (Saffran, Aslin, & Newport, 1996).

A question raised by comparing the conclusions of these two studies is whether the cortical tracking in Ding et al. (2016) was driven by the syntactic integration of linguistic units, or by the TP between syllables in the speech stimuli. In this study, syntactic integration and TP are confounded because the construction of higher-level structures, i.e., phrases and sentences, can be explained by both syntactic integration and TP. A speaker of the language of the stimuli is familiar both with the syntactic structure and the TP relationships between linguistic units. Neither source of information is available to a listener who is not familiar with the language. Separating the effects of syntactic integration from those of TP is important because the former is based on high-level language processing, involving syntactic

and/or semantic knowledge. In contrast, structuring by statistical information (TP) is a low-level perceptual process, which means that the chunking of low-level linguistic units (e.g., syllables) into high-level structures (e.g., phrases) reflects the statistical properties of the stimuli. Linguistic knowledge such as syntactic or semantic information is not necessarily involved in this process. In the present magnetoencephalography (MEG) project, we asked whether TP information contributed to linguistic structure extraction and whether would be reflected in the cortical tracking effect.

To answer these questions, we designed a series of experiments using a syllable recognition task (for details see section 2.2, Methods) to investigate the role of TPs in the extraction of hierarchical linguistic structures. In ***Experiment 1***, we asked Dutch participants listen to Dutch syllable sequences. As in Ding et al. (2016), we constructed three types of isochronous syllable sequences with the occurrence rate of 4 Hz (four syllables in one second). In type-one sequences, syllable pairs formed existing singular nouns; e.g., the syllable *tij* and syllable *ger* formed the Dutch singular noun *tijger* ('tiger'). Therefore, the occurrence rate of words was controlled to be 2 Hz (two nouns per second). In type-two sequences, syllables were randomly presented with an occurrence rate of 4 Hz (four syllables per second), in which successive syllables did not form Dutch words. Type-three sequences were random syllable sequences played backward. The reason for conducting ***Experiment 1*** was to obtain a baseline for our following experiments, and to make sure that the effect of cortical tracking of hierarchical linguistic structures could be replicated in Dutch. Therefore, we expected that the same effect would appear as in the study by Ding and colleagues (2016). Our hypothesis for ***Experiment 1*** was that we would see a 2 Hz and 4 Hz peak at the neural response when participants listened to the type-one sequences, and only a 4 Hz peak when they listened to type-two and type-three sequences.

To examine the role of TP in the tracking effect, we conducted ***Experiment 2***, where Dutch participants listened to the same three types of syllable sequences as in ***Experiment 1***, but in Mandarin Chinese, which they did not understand. We set the TPs to be 1/10 between every two consecutive syllables to serve as a cue indicating a grouped structure (two-syllable words) in the type-one sequence. The aim was to remove high-level, language-related cues, such as syntactic and

semantic information from the stimuli, and to assess whether a frequency response could be introduced to reflect the occurrence rate of words (2 Hz) using only the TPs. Based on the findings by Saffran, Aslin, and Newport (1996), we predicted that there would be two peaks in neural activities corresponding to the occurrence rates of words (2 Hz, two nouns per second) and syllables (4 Hz, four syllables per second) when participants listened to the type-one sequence. In contrast, there should be only a 4 Hz peak in the neural response for the type-two and type-three sequences because the TPs in these types were not controlled. As the Dutch participants did not understand Chinese, the effects could not reflect syntactic or semantic integration. Finding results that are consistent with these predictions would therefore show that TPs are a key component of the cortical tracking effect.

In Ding et al. (2016), cortical activity was found to track linguistic structures at different levels, i.e. syllables (4 Hz), phrases (2 Hz) and sentences (1 Hz), simultaneously. To match the hierarchy of our stimuli with the structure of their syllable sequences, we conducted ***Experiment 3***, in which we trained Dutch participants to learn four-syllable (one-second) novel compounds (i.e. compounds that do not exist in Dutch), such as one made up of the Dutch words meaning 'tiger' and 'noise': *tij-ger-la-waai*. Specifically, we aurally presented a syllable sequence in each trial with one, two or three such compounds while holding the TP between these four-syllable structures at 1/25 using a Markov chain (for details see Methods. By doing so, participants would learn to segment syllable sequences using the statistical cue because the TP between these compounds was controlled (1/25). They would not be able to use syntactic integration since these compounds that concatenate two singular nouns do not exist in Dutch (see section 2.4, Discussion), and furthermore were produced with list intonation, i.e., were not prosodically marked as units.

After this training stage, ***Experiment 4*** was conducted to assess the cortical tracking effect using the trained stimuli. Specifically, we constructed the same three types of sequence as in Experiments 1 and 2, which are noun sequences (type-one), and random syllable sequences played forward (type-two) and backward (type-three). However, as the participants had been trained to extract novel compounds in Experiment 3, we expected that there would be peaks in neural activity to reflect the occurrence rates of syllables (4 Hz), words (2 Hz) and novel

compounds (1 Hz) for the type-one sequences. There should only be a 4 Hz peak in the type-two and type-three sequences to reflect the rhythm of syllables.

One important factor still needs to be eliminated in order to say that the cortical tracking effect can be solely driven by statistical information. In Experiments 3 and 4, participants listened to their own language, and hence they could construct compounds using semantic information. For instance, the sequence *tij-ger-la-waai* was not an existing compound, but because it has components that are meaningful lexical units, participants could construct a semantic relationship between them, e.g., 'the tiger is making noise'. To address this concern, we designed ***Experiments 5 and 6***, which used the same procedures as Experiments 3 and 4, but the stimuli were in Mandarin Chinese. Now semantic and syntactic integration processes were ruled out because the participants did not understand Chinese. This meant that if we still found a frequency response corresponding to the rhythm of the trained compounds (1 Hz) in the type-one sequences, we would be able to say that the cortical tracking effect could be solely driven by TPs. Our hypothesis for ***Experiment 6*** was that there would be frequency peaks in the brain to reflect the rates of syllables (4 Hz), words (2 Hz), and the trained compounds (1 Hz) for the type-one sequences. As before, for the type-two and type-three sequences, we expected only a 4 Hz peak to reflect the rhythm of syllables.

2.2 Methods

Participants

Fourteen Dutch native speakers (8 females and 6 males), aged 20 to 35, participated in the study. All of them were undergraduate or graduate students and were right-handed. They reported no history of hearing impairment or neurological disorder. The experimental procedure was approved by the Ethics Committee of the Social Sciences Department at Radboud University. Written informed consent was obtained from each participant before the experiment, and they were paid for their participation.

Acoustic manipulations

To create the Dutch materials, 20 bi-syllabic singular nouns were synthesized by the ReadSpeaker synthesizer (https://www.readspeaker.com/, the male voice, Guus), and then 40 syllables were extracted manually without missing any meaningful dynamics (***Table 1***).

Using the same method, 20 nouns from Mandarin Chinese (which has no singular vs. plural distinction) were synthesized by ReadSpeaker (the male voice, Liang), following which 40 syllables were extracted (***Table 2***).

In both languages, syllables were 153 to 302 ms (mean 230 ms) in duration. To normalize the stimuli, each syllable was first resampled to 44.1 kHz, then adjusted to 250 ms by truncation or zero padding evenly at both ends. Five percent of both ends of each syllable was ramped by a cosine wave. The root-mean-square value of each syllable was normalized to -16 dB.

For all experiments, auditory stimuli were isochronous syllabic sequences with, and the length varied depending on the particular experiment. No existing compounds could be constructed from any two of the 20 nouns (two-syllable singular nouns).

Dutch items

The 1st syllables	***The 2nd syllables***	***Words***	***In English***
tij	ger	tijger	tiger
ta	fel	tafel	table
la	waai	lawaai	noise
var	ken	varken	pig
be	zem	bezem	broom
tar	we	tarwe	wheat
hal	te	halte	station
ba	naan	banaan	banana
ri	vier	rivier	river
wei	de	weide	pasture
gor	dijn	gordijn	curtain
ze	nuw	zenuw	nerve
sei	zoen	seizoen	season
sui	ker	suiker	sugar
bo	ter	boter	butter
li	moen	limoen	lemon
ko	ning	koning	king
ha	mer	hamer	hammer
le	pel	lepel	spoon
wor	tel	wortel	carrot

Table 1. Dutch materials used in the experiments. *The first and second columns represent the first and second syllables of the Dutch bi-syllabic singular nouns. The third and fourth columns show each full Dutch word and its translation in English, respectively.*

Chinese items			
The 1st syllables	***The 2nd syllables***	***Words***	***In English***
怀 (huái)	表 (biǎo)	怀表 (huái biǎo)	pocket watch
键 (jiàn)	盘 (pán)	键盘 (jiàn pán)	keyboard
相 (xiāng)	机 (jī)	相机 (xiàng jī)	camera
电 (diàn)	视 (shì)	电视 (diàn shì)	televison
熨 (yùn)	斗 (dǒu)	熨斗 (yùn dǒu)	iron
衣 (yī)	柜 (guì)	衣柜 (yī guì)	wardrobe
冰 (bīng)	箱 (xiāng)	冰箱 (bīng xiāng)	refrigerator
吉 (jí)	他 (tā)	吉他 (jí tā)	guitar
沙 (shā)	发 (fā)	沙发 (shā fā)	sofa
帐 (zhàng)	篷 (péng)	帐篷 (zhàng péng)	tent
腰 (yāo)	带 (dài)	腰带 (yāo dài)	belt
牙 (yá)	膏 (gāo)	牙膏 (yá gāo)	toothpaste
钢 (gāng)	笔 (bǐ)	钢笔 (gāng bǐ)	pen
篮 (lán)	球 (qiú)	篮球 (lán qiú)	basketball
汽 (qì)	车 (chē)	汽车 (qì chē)	car
围 (wéi)	巾 (jīn)	围巾 (wéi jīn)	scarf
台 (tái)	灯 (dēng)	台灯 (tái dēng)	table lamp
钱 (qián)	包 (bāo)	钱包 (qián bāo)	wallet
耳 (ěr)	环 (huán)	耳环 (ěr huán)	earring
皮 (pí)	鞋 (xié)	皮鞋 (pí xié)	leather shoe

Table 2. Chinese items used in the experiments. *The first two columns represent the first and second syllables of the Chinese bi-syllabic words. The third and fourth columns show each Chinese word and its translation in English, respectively. Note that the content in the brackets are phonetic labels for the corresponding items.*

Acoustic analysis

The Hilbert transform was first applied on the half-wave rectified speech signal to extract the temporal envelopes, and then the discrete Fourier transform of the down-sampled (200 Hz) temporal envelope was calculated to reflect the frequency characteristics of the stimuli.

Experimental procedure

The study consisted of six experiments (including two training experiments) using a syllable recognition paradigm. On each trial, participants first listened to an isochronous syllabic sequence, and after two or three seconds of silence, a syllable target would be presented. Their task was to indicate by pressing a button (using the right hand), whether or not the syllable target had appeared in the preceding sequence. The next trial started between 2000 and 2800 ms (random jitter) after participants gave their response.

In order to prevent the participants from transposing the higher-level structures from the Dutch to the Chinese stimuli, the experiments using Chinese stimuli preceded those using Dutch. Thus, the order of the experiments was as follows: ***Experiment 2*** (Dutch listen to Chinese), ***Experiment 1*** (Dutch listen to Dutch), ***Experiment 5*** (Dutch receive training on Chinese compounds), ***Experiment 6*** (Dutch listen to trained Chinese stimuli), ***Experiment 3*** (Dutch receive training on Dutch compounds), and ***Experiment 4*** (Dutch listen to trained Dutch stimuli).

Experiment 1. The Dutch participants listened to Dutch syllable sequences in this experiment. We first randomly selected 10 singular nouns (20 syllables) from a pool of 20 words. Then using these selected words, five on each set, to stochastically concatenate a set of 100 noun sequences (type-one sequences, four seconds long, including eight singular nouns or 16 syllables, with a TP between nouns of 1/10). Then by shuffling all the selected syllables, a set of 80 random syllable sequences was constructed. We then randomly selected forty sequences from these 80 (type-two sequences, 16 syllables). The remaining 40 sequences were played backward and used as the last type of stimuli (type-three sequences, 16 syllables). All of these sequences were pseudo-randomly arranged in six blocks with 30 sequences in each block. During the syllable detection task, the silent gap between the sequence and the target syllable was three seconds. To balance the response, for each type of sequence, the syllable target of half of the trials was selected from the preceding syllable sequence, and the syllable target of the other half of the trials was selected from the unused 20 syllables.

Experiment 2. The Dutch participants listened to Chinese syllable sequences. The same arrangement and three types of sequences were used as in Experiment 1, except that the stimuli were Chinese rather than Dutch.

Experiment 3. The Dutch participants were trained on Dutch novel compounds. In this experiment, we first randomly selected 10 words from the 20-word pool. Then arranging five words on one set with the remaining words on the other set, the full combination (5*5) of these words generated 25 four-syllable novel compounds. Using a Markov chain, we generated a series of syllable sequences containing either one such compound (four syllables, one second long), or two (eight syllables, two seconds), or three (12 syllables, three seconds). On each trial of this training session, participants listened to one of the syllable sequences, then performed a syllable recognition task with a silent interval of two seconds between the syllable sequence and syllable target. Note that the TP between each structural level was controlled to serve as cues for participants to segment the syllable sequence. The TP between syllables in a word was 1, between words in a compound it was 1/5, and between compounds it was 1/25.

Experiment 4. The Dutch participants listened to Dutch syllable sequences with the trained stimuli (the 10 singular nouns from Experiment 3). As in Experiments 1 and 2, we constructed three types of syllable sequences: noun sequences (type-one), and random syllable sequences played forward (type-two) and backward (type-three). Note that Experiment 4 was conducted 15 to 30 minutes after Experiment 3.

Experiment 5. The Dutch participants were trained on Chinese novel compounds. The same procedure and arrangement were used as in Experiment 3, except the stimuli were in Chinese rather than Dutch. This is because we wanted to eliminate high-level language processing, e.g. of grammar, syntax and semantics, from the structuring processes.

Experiment 6. The same procedure was applied as for Experiment 4, apart from the fact that the trained items were from Experiment 5. In this experiment, all the effects we observed reflected sequence segmentation by statistical information (TP) because opportunities for high-level language processing, such as chunking according to syntactic, grammatical and semantic information, are removed by using a language that participants do not understand.

Localizer task. A localizer task was performed as well, in which a 'beep' tone (1 kHz, 50 ms in duration) was played 100 times (jitter 2 to 3 seconds) to localize the auditory cortex by using the canonical M100 auditory response.

Scalp surface scanning. Each subject's head shape was digitized using a Polhemus Fastrak three-dimensional digitizer (Polhemus, VT, USA).

Anatomical MRI scanning. Anatomical magnetic resonance images (MRIs) of each participant's brain were acquired using a 1.5 T Siemens Magnetom Sonata system.

Neural recordings

Neural responses were recorded using a 275-channel axial gradiometer MEG system (CTF, Canada), with a sampling rate of 1.2 kHz, in a magnetically shielded room. An infrared eye tracker (EyeLink, Canada) was used to monitor eye activity. In addition, online head position was recorded with three fiducial sensors referencing three anatomical landmarks (Nasion, left and right ear canals).

Speech stimuli were presented using MATLAB 2019a (The MathWorks, Natick, MA) with Psychtoolbox-3 (Brainard, 1997). Auditory stimuli were played at 65 dB SPL and delivered through air-tube earplugs (Etymotic ER-3C, Etymotic Research, Inc.). Event markers were sent via serial port for tagging the onset of the events under investigation (i.e., speech onset, task index onset, etc.).

MEG data preprocessing

MEG data was preprocessed via MATLAB using FieldTrip (Oostenveld et al., 2011), EEGLAB (Delorme & Makeig, 2004), and customized scripts. We first down-sampled the data to 200 Hz, and then high-pass filtered it at 0.5 Hz (finite impulse response filter, FIR; zero-phase lag), and cleaned it using the time-sliding PCA (Chang et al., 2018; Kothe & Jung, 2016).

Following the above steps, we extracted epochs of two seconds preceding and 10 seconds after the auditory stimulus onset. We eliminated bad trials and artifacts in the following two steps. First, we used the short-time Fourier transform to calculate the power spectrum in every one-second window, in which we extracted a value that was calculated by the power summation between 15 and 50 Hz (instantaneous muscle artifacts). Then all the extracted values, one value per

window, formed a distribution for each sensor. From this distribution, we transformed all the extracted values into z-scores. The epochs with values outside the standard deviation range of plus or minus three were deleted. Second, ICA was conducted on the trial-rejected data for the elimination of heartbeat and eye-related artifacts and sustained muscular activities.

MEG data analysis

Frequency tagging analysis. To eliminate the transient evoked neural (e.g. M100) response, each trial was initially epoched from two to four seconds (the neural response that corresponds to the first four syllables was removed) after the speech signal onset. Then a ramping taper (a cosine wave), smoothing 5% of each end, was applied to attenuate frequency leakage. We applied a bootstrapping approach to balance the number of trials across different conditions. More concretely, we generated 50 trials that each lasted 12 seconds by randomly concatenating four extracted epochs (of three seconds each) for every condition. The trial manipulations resulted in a frequency resolution of 1/12 Hz (~ 0.08 Hz). To optimize the frequency response, we performed the following three steps. First, we conducted a narrow band filtering via Gaussian frequency where the full-width-half-maximum (FWHM) value equaled 0.1 Hz for each frequency bin. Then two covariance matrices, one for the filtered data and the other for the original data, were calculated for constructing a spatial filter using generalized eigen-decomposition (GED). The spatial filter was defined as the generalized eigenvector corresponding to the biggest eigenvalue. Finally, after filtering the data, the discrete Fourier transform was applied to extract the specific frequency response. The harmonics of the fundamental frequency (1/3 Hz), which are introduced by the epochs' concatenation, were regressed out by minus the average amplitude of its previous and post harmonic bins.

Statistical analysis

For spectral peaks of interest (1 Hz, 2 Hz and 4 Hz), a one-tailed paired sample t-test with the Bonferroni correction was conducted to test whether the peak activity at one frequency bin was significantly higher than the average of the neighboring four bins around it (two bins on each side).

2.3 Results

The cortical tracking effect occurs when Dutch participants listen to Dutch hierarchical syllable sequences

Experiment 1 served as an experimental baseline, in which we asked Dutch participants to listen to three types of Dutch syllable sequences (see ***Figure 1a***). For the noun sequences (type-one), Dutch bi-syllabic words occurred at the rate of two times per second (2 Hz), while the syllables occurred four times per second (4 Hz). As expected, the neural activity showed peaks corresponding to both words (2 Hz, $t(13) = 10.13$, $p<7.72e-8$, Bonferroni-corrected) and syllables (4 Hz, $t(13) = 5.12$, $p<8.62e-5$, Bonferroni-corrected). In contrast, the neural activity only showed a peak at the syllable rate for the random syllable sequences played forward (4 Hz, $t(13) = 6.44$, $p<1.09e-5$, Bonferroni-corrected) and backward (4 Hz, $t(13) = 5.88$, $p<2.68e-5$, Bonferroni-corrected). The results are shown in ***Figure 1b***, in which the red line represents the frequency response of participants listening to the noun sequences (type-one), and the dark blue and light blue lines represent participants listening to the random syllable sequences played forward (type-two) and backward (type-three), respectively. The shaded areas represent two standard errors of the mean. The topographical distributions show the absolute values of the GED weights (for details see Methods) for the frequencies of interest (2 Hz and 4 Hz), in which the size of the red circles indicates the weight of the sensors. Note that only the weight distribution of the type-one sequence is shown here, as in Ding et al. (2016).

Our analysis supports the existence of cortical tracking of hierarchical linguistic structures in Dutch. However, the underlying mechanism can be explained by different accounts. Specifically, the effect could be the result of grammatical chunking since it might be a reflection of syllable grouping using high-level linguistic knowledge (Ding et al., 2016). Alternatively, it could arise from syllable sequence segmentation using statistical information (Saffran, Aslin, & Newport, 1996) because participants had heard the grouped syllables (two-syllable nouns) repeatedly in their daily life, which means the TP relationships (the TP is higher between syllables within a word than across word boundaries) had been trained. From the results of this experiment, we are not able to determine whether speech segmentation can be performed using only statistical information, but the

occurrence of the cortical tracking effect in Dutch indicates that our experimental manipulations were effective.

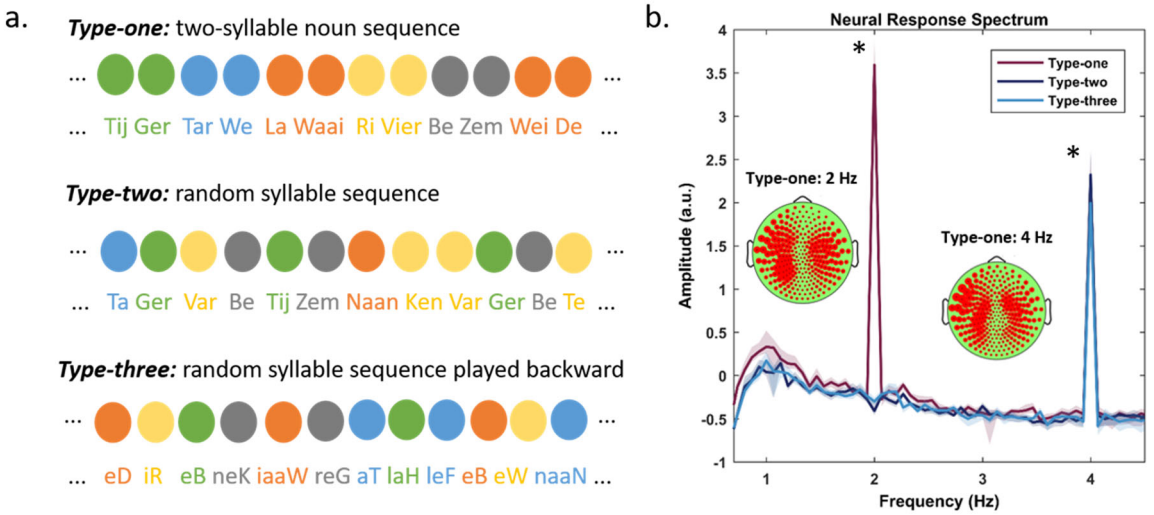

Figure 1. The replication of the cortical tracking effect in Dutch. (a) *The structure of three types of syllable sequences, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively. In the type-one sequence, except for syllables which occurred at the rate of 4 Hz, Dutch singular nouns occurred at the rate of 2 Hz. In the type-two and type-three sequences, syllables occurred at the rate of 4 Hz, and no higher-level structures could be constructed either linguistically or statistically.* ***(b)*** *The neural response spectrum for each type of sequence, in which the 2 Hz peak was significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The shaded areas for each line represent two SEM. The topographical distributions represent the GED weights for the peaks of interest.*

The cortical tracking effect can be introduced by statistically defined structures

In ***Experiment 2***, we assessed the account of sequence segmentation by statistical information (Saffran, Aslin, & Newport, 1996). More concretely, we constructed the same three types of syllable sequences as in Experiment 1, but the stimuli were in Mandarin Chinese rather than Dutch (see ***Figure 2a***). Using the same frequency domain analysis as in Experiment 1, we found that there were two peaks corresponding to the occurrence rates of words (2 Hz, $t(13) = 5.35$, $p < 6.5e-$

5, Bonferroni-corrected) and syllables (4 Hz, t (13) = 8.90, p<3.39e-7, Bonferroni-corrected) for the type-one sequences, and only one peak of activity indicating the rate of syllables for the type-two (4 Hz, t (13) = 8.12, p<9.42e-7, Bonferroni-corrected) and type-three sequences (4 Hz, t (13) = 5.98, p<2.28e-5, Bonferroni-corrected). The results are shown in ***Figure 2b***, in which the red line represents the frequency response corresponding to participants listening to the noun sequences (type-one), while the dark blue and light blue lines represent participants listening to the random syllable sequences played forward (type-two) and backward (type-three), respectively. The shaded area covers two SEM. The topographical distributions represent the GED weights of the frequencies of interest (2 and 4 Hz for the type-one sequences), where the bigger the red circle, the higher the weight of that sensor.

The results are interesting because the peak response at 2 Hz that corresponds to Dutch participants listening to Chinese type-one sequences (noun sequences) has to reflect structural chunking (of bi-syllabic words) via statistical information (TP). The reason is that high-level language processing, such as grammatical, syntactic or semantic processing, is not available for Dutch participants who do not know Chinese. Consequently, the only cue to indicate the sequence structure was the TP information: the TP between words was 1/10 and the TP between syllables in a word was 1. Therefore, our results suggest that the cortical tracking effect can be introduced solely by the TP information even when lexical, grammatical and syntactic knowledge is not available.

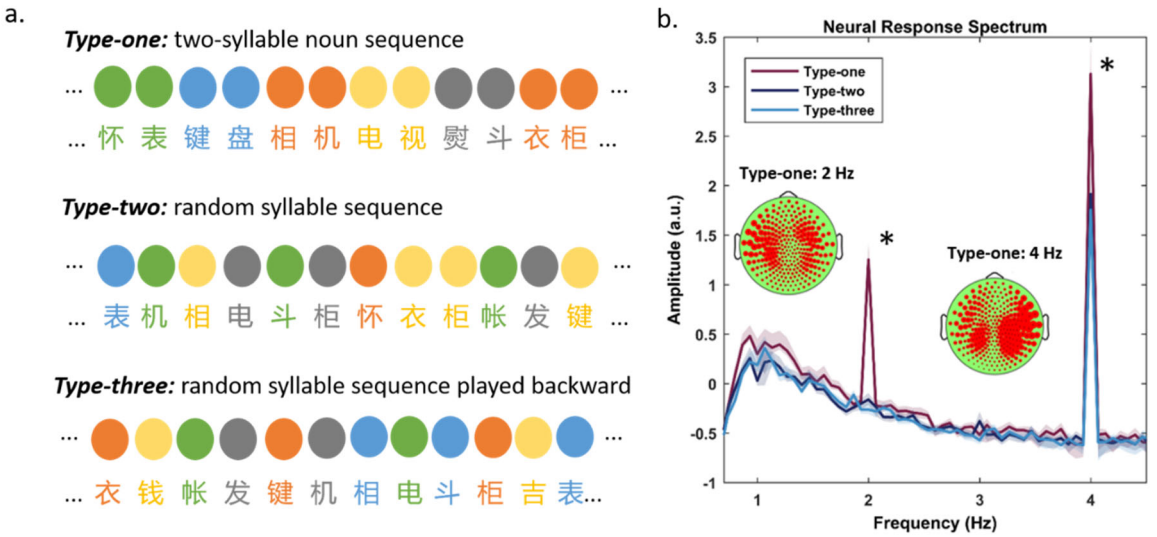

Figure 2. The cortical tracking effect occurred when Dutch participants listened to Chinese syllable sequences. (a) *The structure of three types of syllable sequences, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively. In the type-one sequence, except for syllables which occurred at the rate of 4 Hz, Chinese singular nouns occurred at the rate of 2 Hz. In the type-two and type-three sequences, syllables occurred at the rate of 4 Hz, and no higher-level structures could be constructed either linguistically or statistically.* ***(b)*** *The neural response spectrum for each type of sequence, in which the 2 Hz peak was significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The shaded areas for each line represent two SEM. The topographical distributions represent the GED weights for the peaks of interest.*

The frequency response simultaneously tracks the rhythm of TP-defined structures and lower-level structures

The results of ***Experiment 2*** suggest that the cortical tracking effect could be introduced solely by statistical information at one TP-manipulated rate (2 Hz). However, in the original study (Ding et al., 2016), frequency tagging was found in the simultaneous tracking of different levels of linguistic structure, i.e., syllables (4 Hz), phrases (2 Hz), and sentences (1 Hz). To match the hierarchy of our syllable sequences with the structure of the stimuli in the original study, we first trained Dutch participants on TP-organized, four-syllable (one-second) structures such as *tij-ger-la-waai* (***Experiment 3***). Then, we recorded the neural response when

participants listened to the syllable sequences that were constructed using the trained items (***Experiment 4***).

In ***Experiment 4***, the three types of syllable sequence and manipulations from Experiments 1 and 2 were used, but one additional TP cue which reflects the trained four-syllable compounds was fitted into the type-one sequences**.** As expected, we found that the neural activity showed three peaks that corresponded to the occurrence rates of syllables (4 Hz, t (13) = 5.27, p<7.55e-5, Bonferroni-corrected), words (2 Hz, t (13) = 13.00, p<3.97e-9, Bonferroni-corrected), and the novel compounds (1 Hz, t (13) = 6.02, p<2.12e-5, Bonferroni-corrected) for the type-one sequences. However, there was only a peak at 4 Hz corresponding to the rate of syllables for the type-two sequences (4 Hz, t (13) = 6.19, p<1.63e-5, Bonferroni-corrected) and type-three sequences (4 Hz, t (13) = 4.03, p<7.14e-4, Bonferroni-corrected). ***Figure 3a*** shows the statistical framework for constructing the 25 Dutch novel compounds and the probabilistic relationship between them, in which the TPs between syllables in a word, between words in a compound, and between compounds were 1, 1/5 and 1/25, respectively. ***Figure 3b*** shows the sample syllable sequences that were used during the training stage (Experiment 3), which were constructed using a Markov chain to make sure the statistical relationships between units hold constant. Each style of red outline around the syllable units (i.e. solid, dashed and dotted) represents a statistically defined novel compound. ***Figure 3c*** depicts the stimuli's structure for each type of sequence. The three types were the same as in Experiments 1 and 2, except that this time the syllables had been trained. The results are shown in ***Figure 3d***, in which the red, dark blue line and light blue line represent participants listening to the type-one, type-two and type-three sequences, respectively. The shaded area covers two SEM. The topographical distributions indicate the GED weights for the frequency peaks of interest (i.e., 1 Hz, 2 Hz and 4 Hz for the type-one sequences), in which the larger the red circle, the higher the weight of that sensor.

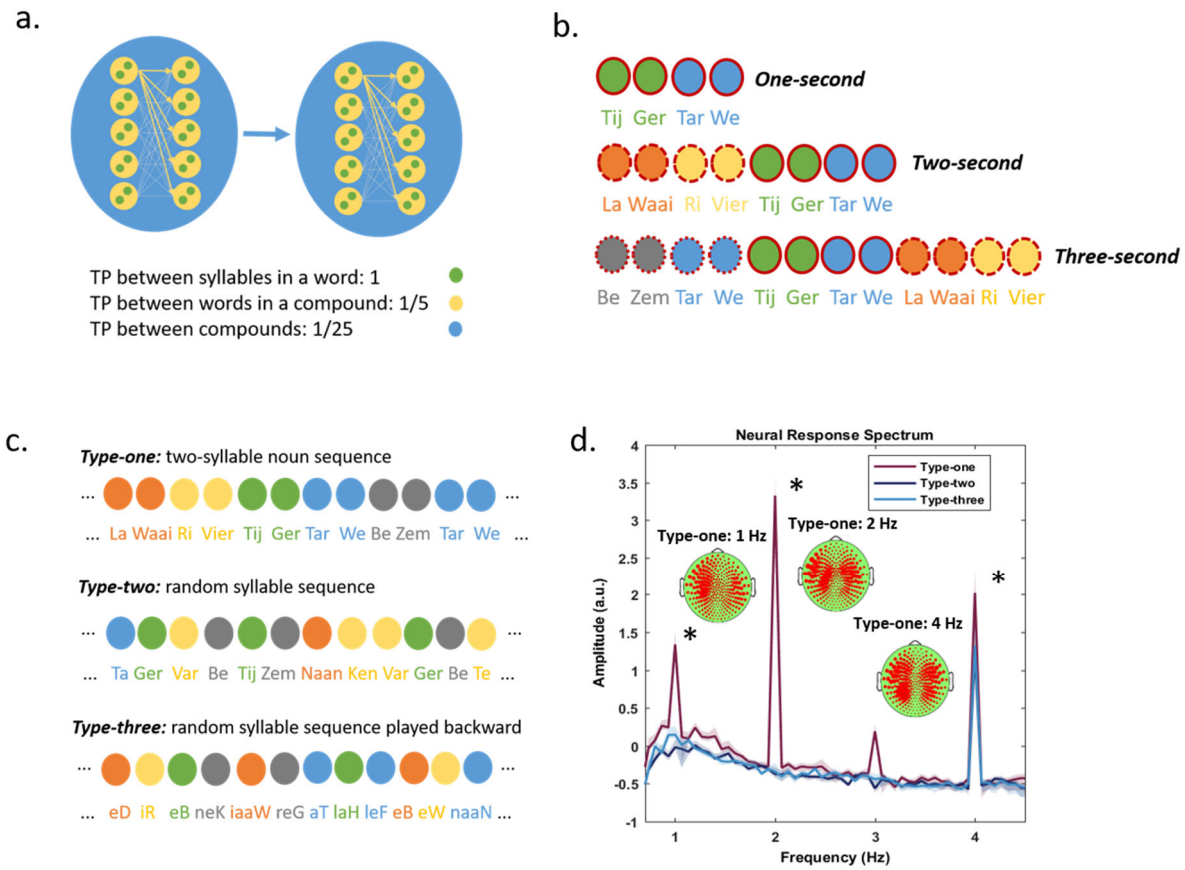

Figure 3. Neural activities track trained novel compounds together with the units that comprise them in Dutch. ***(a)*** *The statistical framework for constructing syllable sequences in Dutch, in which the TPs between syllables in a word, between words in a compound, and between compounds were 1, 1/5 and 1/25, respectively.* ***(b)*** *Sample sequences that were presented during the training experiment. The sequences were generated using a Markov chain to stabilize the statistical relationships between units at different levels. To make sure participants could extract the statistically defined compounds, sequences were manipulated so that they were one, two or three seconds in length.* ***(c)*** *The sequence structure used in Experiment 4, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively.* ***(d)*** *The neural response spectrum for each type of sequence, in which the 1 and 2 Hz peaks were significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The topographical distributions represent the GED weights for the peaks of interest.*

The results in this section are informative because, first, we constructed a type of sequences with structures at three different levels and found that the brain could track these hierarchies at the same time. This suggests that the brain can handle the boundaries at different levels simultaneously. Secondly and most importantly, the highest-level structures in our stimuli (one-second, four-syllable novel compounds) were tracked not in Experiment 1, but after training. This implies that these compounds were not linguistically marked as units as per the accounts based on syntactic integration. Therefore, the frequency tagging to the highest level cannot be explained by syntactic integration; instead, it can be explained by sequence segmentation using statistical information. Our manipulation of the highest-level structures forms a contrast between the accounts of syntactic integration and those of sequence segmentation by TP, and supports the notion that the cortical tracking effect can be manipulated using statistical information.

The frequency-tagging effect occurs when high-level linguistic information is removed from the stimuli

In Experiments 3 and 4, we showed that the frequency-tagging effect could be introduced by the TP-defined compounds. However, during these experiments, the Dutch participants were listening to the stimuli in their own language, which might raise concerns about the semantic combination of two singular nouns. It is possible that the brain tracks semantically associated structures that are formed by two singular nouns, e.g., *tij-ger-la-waai* (made up of the Dutch words meaning 'tiger' and 'noise'), although the semantic content might not be consistent across participants.

To address this concern, we conducted ***Experiments 5 and 6***, in which the same procedures as in Experiment 3 and 4 were used, except that all stimuli were in Chinese, which the participants did not know. By doing so, we believe that the high-level, language-related cues, such as grammatical, syntactic and semantic information, are removed from participants' processing of the stimuli.

The results confirm our predictions: we found three peaks in the neural response to reflect the occurrence rates of syllables (4 Hz, $t(13) = 9.39$, $p<1.83e-7$, Bonferroni-corrected), words (2 Hz, $t(13) = 3.02$, $p<4.9e-3$, Bonferroni-corrected), and the four-syllable novel compounds (1 Hz, $t(13) = 6.64$, $p<8.05e-6$, Bonferroni-

corrected) for the type-one sequences. As expected, there was only a 4 Hz peak to reflect the rate of syllables for the type-two sequences (4 Hz, t (13) = 10.58, p<4.63e-8, Bonferroni-corrected) and the type-three sequences (4 Hz, t (13) = 7.16, p<3.67e-6, Bonferroni-corrected). ***Figure 4a*** shows the statistical framework for constructing the 25 Chinese novel compounds and the probabilistic relationship between compounds, in which the TPs between syllables in a word, between words in a compound, and between compounds were 1, 1/5 and 1/25, respectively. ***Figure 4b*** shows the sample syllable sequences that were used during the training stage (Experiment 3), which were constructed using a Markov chain to make sure the statistical relationships between units at different levels hold constant. Each style of red outline around the syllable units (i.e. solid, dashed and dotted) represents a statistically defined novel compound. ***Figure 4c*** depicts the stimuli's structure for each type of sequence. The three types were the same as in Experiments 1 and 2, except that this time the syllables had been trained. The results are shown in ***Figure 4d***. The frequency responses corresponding to the type-one, type-two and type-three sequences are indicated by the red, dark blue and light blue lines, respectively. The shaded area covers two SEM. The topographical distributions represent the GED weights for the frequency peaks of interest, where the bigger the red circle, the higher the weight of the sensor.

These results suggest that the brain can simultaneously track different levels of TP-defined structures, i.e., syllables (4 Hz), words (2 Hz), and novel compounds (1 Hz). In addition, our results are at odds with the account postulating that the cortical tracking of hierarchical linguistic structures is purely a reflection of grammatical chunking or syntactic integration (Ding et al., 2016). Instead, by connecting the TP with the cortical tracking effect and removing the high-level language information, we find that the frequency activities tagging the occurrence rate of hierarchical structures can be solely driven by statistical information (TP).

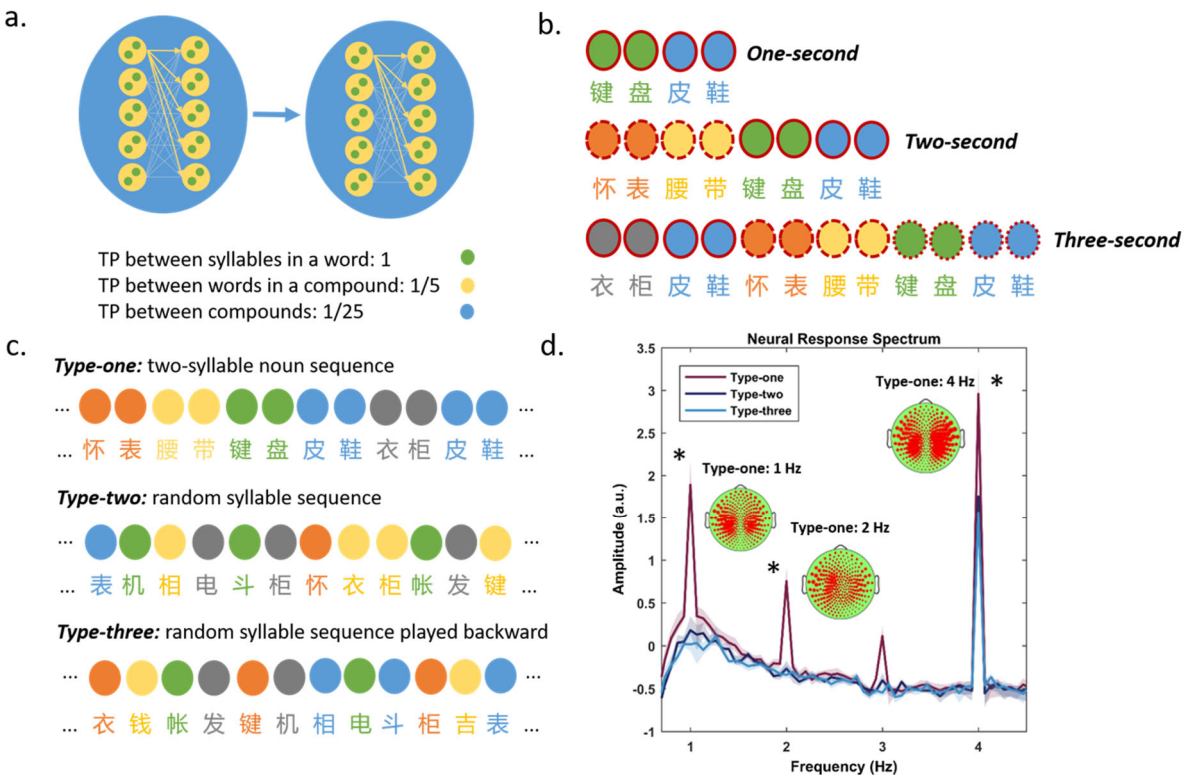

Figure 4. Neural activities of Dutch participants track statistically defined units in Chinese. ***(a)*** *The statistical framework for constructing syllable sequences in Chinese, in which the TPs between syllables in a word, between words in a compound, and between compounds were 1, 1/5 and 1/25, respectively.* ***(b)*** *Sample sequences that were presented in each trial during the training experiment. The sequences were generated using a Markov chain to stabilize the statistical relationships between different levels' units. To make sure participants could extract the statistically defined compounds, sequences were manipulated to be one, two or three seconds in length.* ***(c)*** *The sequence structure used in Experiment 6, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively.* ***(d)*** *The neural response spectrum for each type of sequence, in which the 1 Hz and 2 Hz peaks were significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The topographical distributions represent the GED weights for the peaks of interest.*

2.4 Discussion

This chapter has reported the results of a series of MEG experiments investigating the role of statistical information on speech segmentation. By connecting the effect of cortical tracking to hierarchical linguistic structures with a statistical learning approach, we found that the frequency response in the brain that tags different levels of structure can be introduced solely by statistical cues (TP). Our results support the argument that speech segmentation (indexed by extracting structures) can be conducted without using high-level language knowledge such as grammatical, syntactic and semantic information.

In Experiment 1, we replicated the cortical tracking effect, that was found by Ding et al. (2016), in Dutch. The results suggest that this effect might be a language independent phenomenon; at least, it can be introduced when Dutch participants listen to their own language. We found the occurrence rate of words (2 Hz) and syllables (4 Hz) to be reflected in the brain by using the same experimental paradigm as Ding et al. (2016). However, in Experiment 1, language-related cues such as grammatical, syntactic and semantic knowledge coexisted with the statistical information. Therefore, the observed effect, i.e. the neural activity tracking the rhythm of hierarchical linguistic structures, could be explained by either of these two types of cues.

To remove the impact of linguistic knowledge, we constructed the same three types of syllable sequences as in Experiment 1, but in Mandarin Chinese which the participants did not know. Using the same experimental procedure and analysis methods, we found that frequency activity in the brain still tracked the occurrence rates of words (2 Hz) and syllables (4 Hz). The results can only be explained by structure chunking using statistical cues, as language-related information was not available during listening.

In the study by Ding et al. (2016), the frequency response was found to track multiple levels of linguistic structure, namely syllables (4 Hz, one-syllable structure), phrases (2 Hz, two-syllable structure), and sentences (1 Hz, four-syllable structure), simultaneously. However, in Experiments 1 and 2, we only showed that the brain tracked units at two levels (syllables and words). To match the gap between our experiments (two levels of tracking) with the original study

(three levels of tracking) and test if we could introduce the cortical tracking effect that tracks multiple levels' units, we first need a type of structure that is built on top of words. More importantly, if we want to introduce an additional peak to reflect structure chunking by statistical information, this type of structure needs to be statistically but not syntactically defined. In addition, to satisfy the rule of thumb, which is using the same stimuli to introduce different neural responses, the characteristics of the stimuli in the new experiment should be the same as the old one (Experiment 1, Dutch participants listen to Dutch). To satisfy these criteria, we constructed four-syllable (one-second-long) novel compounds using singular nouns from Experiment 1, and then trained participants to learn these compounds in Markov chain manipulated sequences (Experiment 3, for details see section 2.2, Methods). After this learning took place, we conducted Experiment 4, in which the same three types of sequences, namely the noun sequences (type-one) and random syllable sequences played forward (type-two) and backward (type-three), were constructed using the trained items (10 bi-syllabic nouns in Experiment 3). By doing so, we found that there were three peaks in the brain's frequency response to reflect the occurrence rates of syllables (4 Hz), words (2 Hz), and novel compounds (1 Hz) when Dutch participants listened to Dutch noun sequences (type-one sequences). The results are quite interesting. First of all, we found that the occurrence rate of different levels of structure were reflected in the neural response, which indicates that the brain can handle the boundaries from different levels' structures simultaneously. Secondly, the additional peak corresponding to the rhythm of novel compounds (1 Hz) reflected statistical chunking (and could reflect semantic association) because this 1 Hz peak did not occur in Experiment 1, in which the same noun sequences only introduced frequency responses corresponding to the rates of words (2 Hz) and syllables (4 Hz). For the same reason, we argue that this additional 1 Hz peak was not a reflection of syntactic integration, because if the singular nouns can be chunked into a higher-level structure syntactically, such as compounds, there should be a 1 Hz neural activity to reflect this process in Experiment 1. Lastly, by comparing Experiment 1 and Experiment 4, we can say that the cortical tracking effect reflects an endogenous inference process that sensitive to statistical information as the same stimuli can introduce different neural responses, i.e., two peaks (2 and 4 Hz) in Experiment 1 and three peaks (1, 2 and 4 Hz) in Experiment 4.

However, in Experiments 3 and 4, Dutch participants listened to their own language, which means we cannot construct a type of purely TP-defined structure, as the semantic association has always existed. For instance, *tij-ger-la-waai* is not a compound in Dutch, but participants could semantically associate them together (even though the semantic content might not be consistent across participants). Therefore, one might hold concerns that the additional 1 Hz response could be a reflection of semantic association. To address these concerns, we conducted Experiments 5 and 6, in which the same experimental procedure and parameters as Experiments 3 and 4 were used, but with the stimuli in a language unknown to the participants (Mandarin Chinese). This allowed us to remove all higher-level language related cues that lead to structure chunking from the processing of the syllable sequence. As expected, we still identified three peaks corresponding to the rhythm of syllables (4 Hz), words (2 Hz), and novel compounds (1 Hz) when participants listened to the Chinese noun sequences. In addition to the conclusions drawn from Experiments 3 and 4, the results at this stage could be evidence that the cortical activity tracking multiple levels' structures in speech may be introduced solely by statistical information (TP), which is at odds with the account that the effect is purely a reflection of syntactic integration.

In sum, across all the results explored in this chapter, we showed that speech segmentation can be performed solely using statistical information. More importantly, this inference segmentation process using statistical information could be reflected by the cortical tracking effect which was originally considered as a pure reflection of syntactic integration or grammatical chunking. Demonstrating that the cortical tracking effect can be driven by statistical information alone is important. One reason is that this finding helps to establish how speech segmentation can be performed prior to acquiring high-level language knowledge. From this perspective, our results are in accordance with those of Saffran, Aslin, and Newport (1996) suggesting that speech segmentation using statistical information could be an initial inference mechanism in language acquisition. In addition, we showed that the cortical response which simultaneously tracks the TP-defined boundaries could be the neural representation of this endogenous process. Furthermore, in contrast to the account of grammatical chunking or syntactic integration, our experiments provide evidence that the cortical tracking effect does

not necessarily constitute a pure reflection of the generation of hierarchical structures; instead, the fact that the effect can be introduced without the involvement of high-level linguistic knowledge suggests it might be a generalized perceptual process. It is undeniable that higher-level linguistic knowledge is helpful for speech segmentation, and that successful speech comprehension needs grammatical and syntactic information. However, the neural representation of higher-level linguistic information, e.g., syntactic structure, is not necessarily the only information reflected in the cortical tracking effect.

3 | Generalization of the cortical tracking effect based on statistical inference

Abstract

In Chapter 2, we showed that speech segmentation could be performed via statistical inference without understanding the speech stimuli, and the endogenous inference process was robustly reflected by the cortical activity tracking the rhythm of multiple layers of units. However, all six magnetoencephalography (MEG) experiments in Chapter 2 were conducted with Dutch participants, and it is possible that all the effects we reported were driven by participants' specific linguistic knowledge (e.g., the Dutch participants had Dutch linguistic knowledge). To test whether the cortical tracking effect driven by statistical information could be generalized to different types of language users, and especially to see if the cortical response tracking the rhythm of hierarchical units reflects a general perceptual processing, we conducted the same sets of experiments with Chinese participants. The findings, which are explored in this chapter, were that linguistic knowledge itself did not affect the statistical inference process. In other words, each experiment discussed in this chapter obtained the same pattern of results as its counterpart in the last chapter. Our results support the idea that the cortical tracking effect can be solely driven by statistical information and is independent of the linguistic knowledge that participants have.

3.1 Introduction

In Chapter 2, using six MEG experiments with Dutch native speakers, we showed that statistical information (i.e., transitional probability) plays an important role in speech segmentation, and the extraction of units via statistical inference can be reflected by the cortical tracking effect. The results in Chapter 2 revealed that neural oscillations are robust indices that varied according to the rhythm of the statistically defined units. Based on these findings, we argue that the cortical tracking effect is not a pure reflection of linguistic units' extraction via grammatical chunking or syntactic integration; rather, it also reflects an endogenous inference that can be evoked solely by statistical information. However, we have only demonstrated this in the case of Dutch native speakers, which means the conclusions are only applied to Dutch speakers. A straightforward question is whether the statistical cue-based segmentation process would be represented differently in the brain when a different type of speaker, i.e. Chinese native speakers, perform the same sets of experiments. The concern is important for two reasons. First, in Chapter 2, the conclusion was drawn by removing the availability of high-level linguistic information (i.e., contrasting Dutch participants listening to Dutch speech stimuli with Dutch participants listening to Chinese syllable sequences), but whether the language knowledge itself that participants had (Dutch vs. Chinese) would affect the cortical tracking effect is not known. It is possible that all the effects we presented in Chapter 2 were driven by participants' specific linguistic knowledge of Dutch. Second, referencing the previous studies on speech segmentation via statistical inference (Henin et al., 2021; Saffran, Aslin, & Newport, 1996) and the results in Chapter 2, we hypothesized that the phenomenon of neural activity tracking the rhythm of linguistic structures could reflect a generalized perceptual processing (statistical inference). If this is the case, then the effect should be independent of participants' linguistic knowledge, which means we would get the same pattern of results when experiments are conducted with users of a different language (i.e. Chinese). In contrast, if our hypothesis was not supported, the critical question would be how the statistical inference process is represented in the brain when users of a different language are doing the experiments. In short, the experiment reported in Chapter 2 served to assess whether the cortical tracking effect could be introduced after removing the availability of high-level linguistic knowledge. To draw a full picture, the present

chapter discusses six MEG experiments with Chinese native speakers designed to check the weight of participants' linguistic knowledge in the frequency-tagging effect and ascertain whether the endogenous statistical inference has a consistent neural representation across different types of language users.

Specifically, in ***Experiment 1***, we asked Chinese participants to listen to Chinese syllable sequences. The same three types of syllable sequences as in Chapter 2 were constructed, i.e., noun sequences (type-one) and random syllable sequences played forward (type-two) and backward (type-three). The reason for conducting ***Experiment 1*** was the same as for its counterpart in Chapter 2: to obtain an experimental baseline and make sure that the manipulations are effective enough to introduce the cortical tracking effect in Chinese. Therefore, we expected that we would see a 2 Hz and 4 Hz peak at the neural response when participants listened to the noun sequences, and only a 4 Hz peak when they listened to type-two and type-three sequences.

In ***Experiment 2***, the Chinese participants listened to Dutch syllable sequences, in which the same three types of sequences as in Experiment 1 were used. We set the TP (at 1/10 between two-syllable nouns) to serve as a statistical cue to indicate grouped structures in the type-one sequence (for details, see section 3.2, Methods). The aim was the same as its counterpart in Chapter 2 (Experiment 2), which was to remove the availability of high-level language-related information. In addition, by involving users of a different language, we wanted to check whether the particular linguistic knowledge itself that participants had would affect the inference process. Based on the findings of Saffran, Aslin, and Newport (1996) and the results of the corresponding experiment in Chapter 2, we predicted that we would get the same pattern of results, i.e. two peaks in the brain to reflect the rhythm of words (2 Hz) and syllables (4 Hz) when Chinese participants listened to Dutch type-one sequences. In contrast, there would be only a 4 Hz peak in the brain for the remaining two control conditions.

Experiments 3 and 4 were conducted in order to match the hierarchy of our syllable sequences with the structure of stimuli in the original study (Ding et al., 2016) and check whether multiple levels of units can be handled by the brain simultaneously when Chinese participants listen to Chinese speech stimuli. These experiments were expected to give us the data necessary to compare the neural

representation of statistical inference between the two different types of language users (Dutch speakers vs. Chinese speakers). In short, we first trained the Chinese participants in ***Experiment 3*** with the same procedures as its counterpart in Chapter 2 to learn four-syllable, one-second novel compounds (i.e., compounds that do not exist in Chinese), such as *jí-tā-bīng-xiāng* (made up of words meaning 'guitar' and 'refrigerator'). Then in ***Experiment 4***, using the trained stimuli we assessed the cortical tracking effect after training. Like before, three types of sequences were constructed, namely, the noun sequences (type-one) and random syllable sequences played forward (type-two) and backward (type-three). As participants had learned to extract the novel compounds in ***Experiment 3***, the statistical information indicating the trained novel compounds was fitted into the type-one sequences in ***Experiment 4***. Therefore, we expected that there would be peaks in neural activity to reflect the occurrence rate of syllables (4 Hz), words (2 Hz), and the novel compounds (1 Hz) for the type-one sequence. There should only be a 4 Hz peak in the type-two and type-three sequences to reflect the rhythm of syllables.

To address the concern about semantic association (for details see Chapter 2, section 2.4), we performed ***Experiments 5 and 6***, in which the same procedure as in Experiments 3 and 4 was used, but the stimuli were in Dutch. By doing so, we were able to rule out semantic and syntactic integration processes, as the participants did not understand Dutch. Furthermore, the experiments were conducted to draw a comparison between how the two types of language speakers (Dutch speakers in Chapter 2 vs. Chinese speakers in Chapter 3) chunk units from speech input in an unknown language. If we still get a frequency response corresponding to the rhythm of the trained compounds (1 Hz), we would say that the cortical tracking effect could be solely driven by statistical information (i.e., the TP) and any language-specific knowledge that the participants had is an unrelated factor. Our hypothesis for ***Experiment 6*** was that there would be frequency peaks in the brain to reflect the rate of syllables (4 Hz), words (2 Hz), and the trained compounds (1 Hz) for the type-one sequences. As before, for the type-two and type-three sequences, we expected only a 4 Hz peak to reflect the rhythm of syllables.

3.2 Methods

Participants

Fourteen Chinese native speakers (12 females and 2 males), aged 20 to 35, participated in all six experiments. All of them were undergraduate or graduate students and were right-handed. They reported no history of hearing impairment or neurological disorder. The experimental procedure was approved by the Ethics Committee of the Social Sciences Department at Radboud University. Written informed consent was obtained from each participant before the experiment, and they were paid for their participation.

Acoustic manipulations

As noted in Chapter 2, to create the original Dutch materials, 20 Dutch bi-syllabic singular nouns were synthesized by the ReadSpeaker synthesizer (https://www.readspeaker.com/, the male voice, Guus), and then 40 syllables were extracted manually without missing any meaningful dynamics (see ***Table 1 in Chapter 2***). Using the same method that we used to construct the Dutch materials, 20 Mandarin Chinese nouns (not marked for number, as there is no lexical distinction between singular and plural in Chinese) were synthesized by ReadSpeaker (the male voice, Liang), following which 40 syllables were extracted (see ***Table 2 in Chapter 2***).

In both languages, the syllables were 153 to 302 ms (mean 230 ms) in duration. To normalize the stimuli, each syllable was first resampled to 44.1 kHz, and then adjusted to 250 ms by truncation or zero padding evenly at both ends. Five percent of both ends of each syllable was ramped by a cosine wave. The root-mean-square value of each syllable was normalized to -16 dB.

For all experiments, the auditory stimuli were isochronous syllabic sequences, and the length varied depending on the particular experiment. No existing compounds could be constructed from any two of the 20 nouns (two-syllable nouns).

Acoustic analysis

The Hilbert transform was first applied on the half-wave rectified speech signal to extract the temporal envelopes, and then the discrete Fourier transform

of the down-sampled (200 Hz) temporal envelope was calculated to reflect the frequency characteristics of the stimuli.

Experimental procedure

The experimental procedure is the same as in Chapter 2. The study consisted of six experiments (including two training experiments) using a syllable recognition paradigm. On each trial, participants first listened to an isochronous syllabic sequence, and after two or three seconds of silence, a syllable target would be presented. Their task was to indicate by pressing a button (using the right hand), whether or not the syllable target had appeared in the preceding sequence. The next trial started between 2000 and 2800 ms (random jitter) after participants gave their response.

In order to prevent the participants from transposing the higher-level structures from the Chinese to the Dutch stimuli, the experiments using Dutch stimuli preceded those using Chinese. Thus, the order of the experiments was as follows: ***Experiment 2*** (Chinese listen to Dutch), ***Experiment 1*** (Chinese listen to Chinese), ***Experiment 5*** (Chinese receive training on Dutch compounds), ***Experiment 6*** (Chinese listen to trained Dutch stimuli), ***Experiment 3*** (Chinese receive training on Chinese compounds), and ***Experiment 4*** (Chinese listen to trained Chinese stimuli).

Experiment 1. The Chinese participants listened to Chinese syllable sequences in this experiment. We first randomly selected 10 singular nouns (20 syllables) from a pool of 20 words. Then using these selected words, five on each set, to stochastically concatenate a set of 100 noun sequences (type-one sequences, four seconds long, including eight singular nouns or 16 syllables, with a TP between nouns of 1/5). Then by shuffling all the selected syllables, a set of eighty random syllable sequences was constructed. We then randomly selected 40 sequences from these 80 (type-two sequences, 16 syllables). The remaining 40 sequences were played backward and used as the last type of stimuli (type-three sequences, 16 syllables). All of these sequences were pseudo-randomly arranged in six blocks with 30 sequences in each block. During the syllable detection task, the silent gap between the sequence and the target syllable was three seconds. For each type of sequence, the syllable target of half of the trials had appeared in the preceding

syllable sequence, and the syllable target of the other half of the trials was selected from the unused 20 syllables (unused 10 words).

Experiment 2. The Chinese participants listened to Dutch syllable sequences. The same arrangement and three types of sequences were used as in Experiment 1, except that Dutch rather than Chinese stimuli were used.

Experiment 3. The Chinese participants were trained on Chinese novel compounds. In this experiment, we first randomly selected 10 words from the 20-word pool. Then arranging five words on one set with the remaining words on the other set, the full combination (5*5) of these words generated 25 novel compounds (four syllables each). Using a Markov chain, we generated a series of syllable sequences containing either one such compound (four syllables, one second long), or two (eight syllables, two seconds), or three (12 syllables, three seconds). On each trial of the training session, participants listened to one of the syllable sequences, then performed a syllable recognition task with a silent interval of two seconds between the syllable sequence and syllable target. Note that the TP between each structural level was controlled to serve as cues for participants to segment the syllable sequence. The TP between syllables in a word was 1, between words in a compound it was 1/5, and between compounds it was 1/25.

Experiment 4. The Chinese participants listened to Chinese syllable sequences with the trained stimuli (the 10 singular nouns from Experiment 3). Like in Experiments 1 and 2, we constructed three types of syllable sequences, noun sequences (type-one) and random syllable sequences played forward (type-two) and backward (type-three). Note that Experiment 4 was conducted 15 to 30 minutes after Experiment 3.

Experiment 5. The Chinese participants were trained on Dutch novel compounds. The same procedure and arrangement were used as in Experiment 3, except the stimuli were in Dutch rather than Chinese. This is because we wanted to remove high-level grammatical, syntactic, and semantic processing from perceptual processing.

Experiment 6. The same procedure was applied as in Experiment 4, apart from the fact that the trained items were from Experiment 5. In this experiment, all the effects we observed reflected sequence segmentation by statistical information (TP) because opportunities for high-level language processing, such as

chunking according to syntactic, grammatical, and semantic information, are removed by using a language that participants do not understand.

Localizer task. A localizer task was performed as well, in which a 'beep' tone (1 kHz, 50 ms in duration) was played 100 times (jitter 2 to 3 seconds) to localize the auditory cortex by using the canonical M100 auditory response.

Scalp surface scanning. Each subject's head shape was digitized using a Polhemus Fastrak three-dimensional digitizer (Polhemus, VT, USA).

Anatomical MRI scanning. Anatomical magnetic resonance images (MRIs) of each participant's brain were acquired using a 1.5 T Siemens Magnetom Sonata system.

Neural recordings

Neural activity was recorded using a 275-channel axial gradiometer MEG system (CTF, Canada), with a sampling rate of 1.2 kHz, in a magnetically shielded room. An infrared eye tracker (EyeLink, Canada) was used to monitor eye activity. In addition, online head position was recorded with three fiducial sensors referencing three anatomical landmarks (Nasion, left and right ear canals). Speech stimuli were presented using MATLAB 2019a (The MathWorks, Natick, MA) with Psychtoolbox-3 (Brainard, 1997). Auditory stimuli were played at 65 dB SPL and delivered through air-tube earplugs (Etymotic ER-3C, Etymotic Research, Inc.). Event markers were sent via serial port for tagging the onset of the events under investigation (i.e., speech onset, task index onset, etc.).

MEG data preprocessing

MEG data was preprocessed via MATLAB using FieldTrip (Oostenveld et al., 2011), EEGLAB (Delorme & Makeig, 2004), and customized scripts. We first down-sampled the data to 200 Hz, and then high-pass filtered it at 0.5 Hz (finite impulse response filter, FIR; zero-phase lag), and cleaned it using the time-sliding PCA (Chang et al., 2018; Kothe & Jung, 2016). Following the above steps, we extracted epochs of two seconds preceding and 10 seconds after the auditory stimulus onset. We eliminated bad trials and artifacts in the following two steps. First, we used the short-time Fourier transform to calculate the power spectrum in every one-second window, in which we extracted a value that was calculated by the power summation between 15 and 50 Hz. Then all the extracted values, one value per window, formed

a distribution for each sensor. From this distribution, we transformed all the extracted values into z-scores. The epochs with values outside the standard deviation range of plus or minus three were deleted. Second, ICA was conducted on the trial-rejected data for the elimination of heartbeat and eye-related artifacts and sustained muscular activities.

MEG data analysis

Frequency tagging analysis

To eliminate the transient evoked neural (e.g. M100) response, each trial was initially epoched from two to four seconds (the neural response that corresponds to the first four syllables was removed) after the speech signal onset. Then a ramping taper (a cosine wave), smoothing 5% of each end, was applied to attenuate frequency leakage. We applied a bootstrapping approach to balance the number of trials across different conditions. More concretely, we generated 50 trials that each lasted 15 seconds by randomly concatenating four extracted epochs (of three seconds each) for every condition. The trial manipulations resulted in a frequency resolution of 1/15 Hz (~ 0.07 Hz). To optimize the frequency response, we performed the following three steps. First, we conducted a narrow band filtering via Gaussian frequency where the full-width-half-maximum value equaled 0.1 Hz for each frequency bin, e.g., 0.46 Hz. Then two covariance matrices, one for the filtered data and the other for the original data, were calculated for constructing a spatial filter using the generalized eigen-decomposition (GED). The spatial filter was defined as the generalized eigenvector corresponding to the biggest eigenvalue. Finally, after filtering the data, the discrete Fourier transform was applied to extract the specific frequency response. The harmonics of the fundamental frequency (1/3 Hz), which are introduced by the epochs' concatenation, were regressed out by minus the average amplitude of its pre- and post-harmonic bins.

Statistical analysis

For spectral peaks of interest (1 Hz, 2 Hz and 4 Hz), a one-tailed paired sample t-test with the Bonferroni correction was conducted to test whether the peak activity at one frequency bin was significantly higher than the average of the neighboring four bins around it (two bins on each side).

3.3 Results

The cortical tracking effect occurs when Chinese participants listen to Chinese hierarchical syllable sequences

In ***Experiment 1***, we asked Chinese participants to listen to three types of Chinese syllable sequences. For the noun sequences (type-one), Chinese bi-syllabic words occurred at the rate of two times per second (2 Hz), while the syllables occurred four times per second (4 Hz). In contrast, for the remaining two control conditions, i.e. the random syllable sequences played forward and backward, only the rhythm of syllables was controlled to be 4 Hz. The sample structures for each type of sequence are shown in ***Figure 1a***. As expected, the neural activity showed peaks corresponding to both words (2 Hz, $t(13) = 8.81$, $p<3.86e-7$, Bonferroni-corrected) and syllables (4 Hz, $t(13) = 10.52$, $p<4.94e-8$, Bonferroni-corrected) for the noun sequences (type-one). However, the neural activity only showed a peak at the rate of syllables for the random sequences played forward (type-two, 4 Hz, $t(13) = 8.52$, $p<5.50e-7$, Bonferroni-corrected) and backward (type-three, 4 Hz, $t(13) = 6.61$, $p<8.47e-6$, Bonferroni-corrected). The results are shown in ***Figure 1b***, in which the red line represents the frequency response of participants listening to the noun sequences (type-one), and the dark blue and light blue lines represent participants listening to the random syllable sequences played forward (type-two) and backward (type-three), respectively. The shaded areas represent two SEM. The topographical distributions show the absolute values of the GED weights (for details see section 3.2, Methods) for the frequencies of interest (2 and 4 Hz), in which the size of the red circles indicates the weight of the sensors.

The results replicated the cortical tracking effect found by Ding et al. (2016) in Chinese. By considering the results of this experiment with its counterpart in Chapter 2 (Experiment 1, wherein Dutch participants listened to Dutch speech stimuli), our analysis indicated that neural oscillations tracking the rhythm of hierarchical linguistic units could be independent of the linguistic knowledge that participants had. The same inference mechanism was engaged across different types of language users (Dutch speakers vs. Chinese speakers). However, two comparable accounts, which are the grammatical chunking and statistical structuring, both explain the effect.

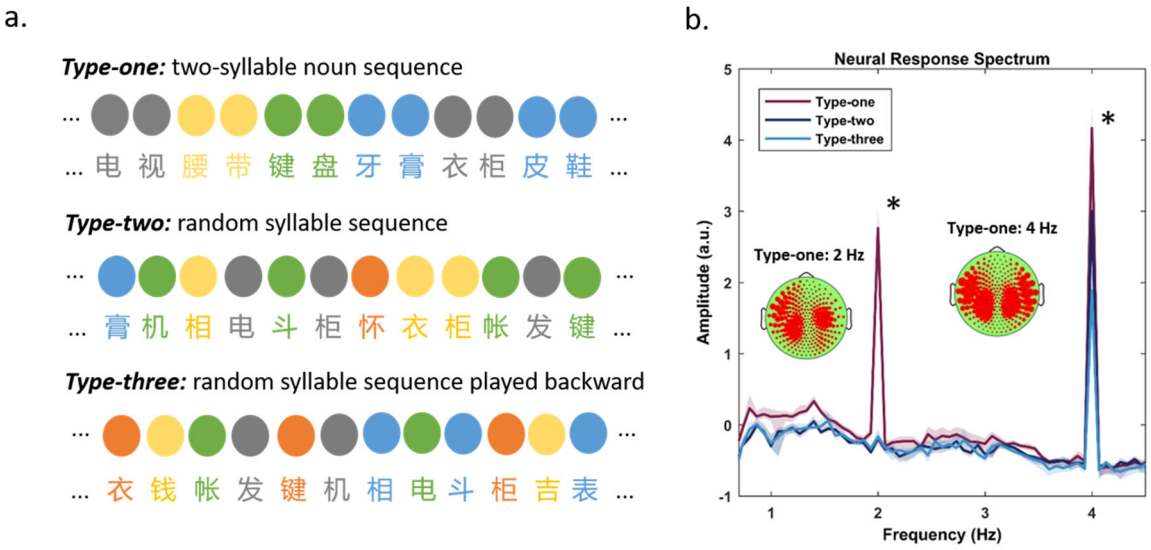

Figure 1. The cortical tracking effect occurred when Chinese participants listened to Chinese hierarchical syllable sequences. (a) *The structure of three types of syllable sequences, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively. In the type-one sequence, except for syllables which occurred at the rate of 4 Hz, Chinese nouns occurred at the rate of 2 Hz. In the type-two and type-three sequences, syllables occurred at the rate of 4 Hz, and no higher-level structures could be constructed either linguistically or statistically.* ***(b)*** *The neural response spectrum for each type of sequence, in which the 2 Hz peak was significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The shaded areas for each line represent two SEM. The topographical distributions represent the GED weights for the peaks of interest.*

The cortical tracking effect reflects speech segmentation via statistical inference

In ***Experiment 2***, we constructed the same three types of syllable sequences as in Experiment 1, but the stimuli were in Dutch rather than Chinese (as shown in ***Figure 2a***). By doing so, we wanted to test the effect after removing the availability of higher-level linguistic information and check whether speech segmentation via statistical inference could be introduced by users of a different language (Chinese participants). Using the same frequency decomposition as in Experiment 1, we found that there were two peaks corresponding to the occurrence rates of words (2 Hz, t (13) = 16.79, p < 1.70e-10, Bonferroni-corrected) and

syllables (4 Hz, t (13) = 10.63, p<4.38e-8, Bonferroni-corrected) for the type-one sequences, and only one peak of activity indicating the rate of syllables for the type-two (4 Hz, t (13) = 6.49, p<1.00e-5, Bonferroni-corrected) and type-three sequences (4 Hz, t (13) = 7.53, p<2.16e-6, Bonferroni-corrected). The results are shown in ***Figure 2b***, in which the red line represents the frequency response corresponding to participants listening to the noun sequences (type-one), while the dark blue and light blue lines represent participants listening to the random syllable sequences played forward (type-two) and backward (type-three), respectively. The shaded area covers two SEM. The topographical distributions show the GED weight of the frequencies of interest, where the bigger the red circle, the higher the weight of that sensor.

The results confirmed our hypothesis which is that the neural response tracking the rhythm of units at different levels is not a pure reflection of unit structuring using high-level linguistic knowledge. Instead, by comparing the results of this experiment with those of its counterpart in Chapter 2 (Experiment 2, where Dutch participants listened to Chinese stimuli), we found that the linguistic knowledge that participants had did not affect the tracking regime. In addition, as high-level linguistic information was not available when participants listened to the speech in an unfamiliar language, the effect has to reflect statistical inference.

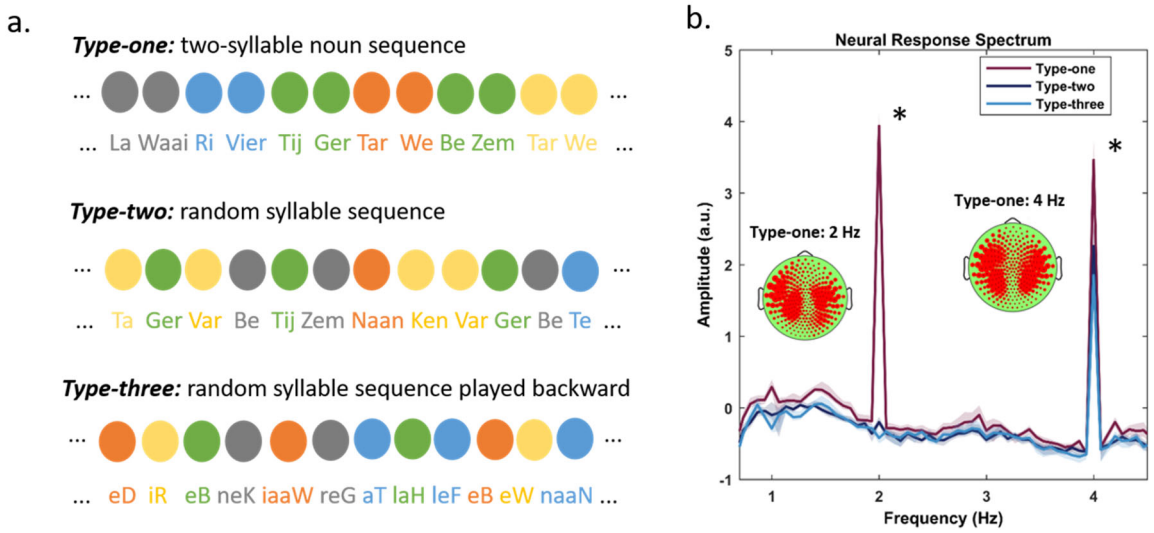

Figure 2. Cortical tracking effect reflects speech segmentation via statistical inference. (a) *The structure of three types of syllable sequences, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three*

sequences, respectively. In the type-one sequence, except for syllables which occurred at the rate of 4 Hz, Dutch singular nouns occurred at the rate of 2 Hz. In the type-two and type-three sequences, syllables occurred at the rate of 4 Hz, and no higher-level structures could be constructed either linguistically or statistically. ***(b)*** *The neural response spectrum for each type of sequence, in which the 2 Hz peak was significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The shaded areas for each line represent two SEM. The topographical distributions represent the GED weights for the peaks of interest.*

The cortical response tracks the rate of multiple levels' units regardless of what linguistic knowledge the participants had

The results of Experiment 2 in both Chapter 2 and Chapter 3 suggest that the cortical tracking effect could be solely introduced via statistical inference at one TP-manipulated rate (2 Hz, nouns). However, as stated in Chapter 2, frequency tagging was found in the tracking of different levels of linguistic structure, i.e., syllables (4 Hz), phrases (2 Hz), and sentences (1 Hz), simultaneously in the original study (Ding et al., 2016). To match the hierarchy of our syllable sequences with the structure of the stimuli in the original study and check the role of participants' linguistic knowledge in the tracking phenomenon, we first trained Chinese participants on Chinese TP-organized novel compounds (***Experiment 3***), such as *jí-tā-bīng-xiāng* (made up of words meaning 'guitar' and 'refrigerator'). The statistical framework for constructing these novel compounds is shown in ***Figure 3a***, in which the green, yellow and blue circles represent syllables, words and novel compounds, respectively. The TPs between syllables in a word, between words in a novel compound, and between novel compounds were controlled to be 1, 1/5 and 1/25, respectively. To make sure the statistical cues (the TP between units at different levels) were held constant, we extracted syllable sequences from the framework using a Markov chain. The extracted sequences could be one, two or three seconds in length. Sample sequences that were used in the training stage of Experiment 3 are shown in ***Figure 3b***, in which each style of red outline indicates one novel compound.

In ***Experiment 4***, the same three types of syllable sequences and manipulations were used as before (in Experiments 1 and 2), except that one

additional TP cue which reflects how to extract the trained compounds was fitted into the type-one sequences (***Figure 3c***). As expected, we found that the neural activity showed three peaks that corresponded to the occurrence rates of syllables (4 Hz, t (13) = 10.38, p<5.79e-8, Bonferroni-corrected), words (2 Hz, t (13) = 8.16, p<9.00e-7, Bonferroni-corrected) and the TP-organized compounds (1 Hz, t (13) = 4.49, p<2.98e-4, Bonferroni-corrected) for the type-one sequences. However, there was only a peak at 4 Hz corresponding to the rate of syllables for the type-two sequences (4 Hz, t (13) = 7.36, p<2.72e-6, Bonferroni-corrected) and the type-three sequences (4 Hz, t (13) = 9.25, p<2.18e-7, Bonferroni-corrected). The results are shown in ***Figure 3d***, in which the red, dark blue and light blue lines represent participants listening to the type-one, type-two and type-three sequences, respectively. The shaded area covers two SEM. The topographical distributions show the GED weights of the peaks of interest, in which the bigger the red circle, the higher the weight of that sensor.

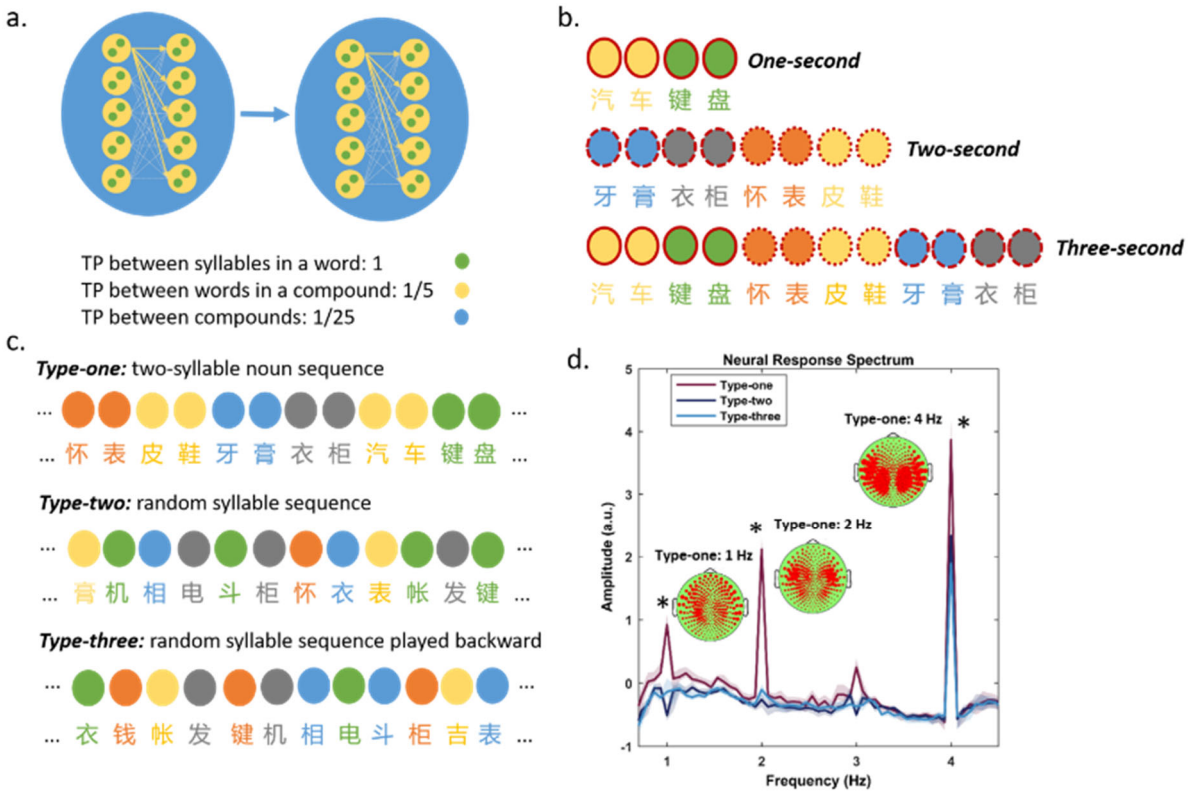

Figure 3. Neural activities track trained novel compounds together with the units that comprise them. (a) *The statistical framework for constructing syllable sequences in Chinese, in which the TPs between syllables in a word, between words in a compound and between compounds were 1, 1/5 and 1/25, respectively.* ***(b)*** *Sample*

sequences that were presented during the training experiment. The sequences were generated using a Markov chain to stabilize the statistical relationships between units at different levels. To make sure participants could extract the statistically defined compounds, sequences were manipulated so that they were one, two or three seconds in length. ***(c)*** *The sequence structure used in Experiment 4, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively.* ***(d)*** *The neural response spectrum for each type of sequence, in which the 1 and 2 Hz peaks were significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The topographical distributions represent the GED weights for the peak of interest.*

The results of this experiment were a perfect replication of its counterpart in Chapter 2. First, it indicates that the brain can handle several boundaries between different levels' units simultaneously. Second, it suggests that the occurrence of frequency peaks can be manipulated via statistical learning (for details see Chapter 2). Lastly and most importantly, we found that the linguistic knowledge that participants had did not change the pattern of results across different types of language users, which indicates that the frequency response tracking the rhythm of different levels' units could reflect a generalized perceptual inference (i.e., statistical inference).

Neural oscillations tracking the rhythm of units at multiple levels reflect statistical inference

In Experiment 3 and 4, when Chinese participants were listening to the stimuli in their own language, the same concerns about semantic association were held as for their counterparts in the previous chapter (see Experiments 5 and 6 in Chapter 2). To address this issue, we conducted ***Experiments 5 and 6***, where the same procedures as in Experiment 3 and 4 were used, except that all stimuli were in Dutch. By doing so, we removed the availability of higher-level linguistic knowledge, such as grammatical, syntactic and semantic information. In addition, we could check whether the language-specific knowledge that participants had (about Dutch vs. Chinese) would alter the frequency tracking effect. ***Figure 4a*** and ***Figure 4b*** show the statistical framework for generating novel compounds and the statistically controlled sample sequences (from Experiments 3 and 4) that

were used in ***Experiment 5***. The structures of the three types of syllable sequences that were used in ***Experiment 6*** are represented in ***Figure 4c***.

The results of ***Experiment 6*** confirmed our predictions, as we still found three peaks in the neural response to reflect the occurrence rates of syllables (4 Hz, t (13) = 9.30, p<1.6e-3, Bonferroni-corrected), words (2 Hz, t (13) = 17.62, p<9.33e-11, Bonferroni-corrected), and the four-syllable artificial structures (1 Hz, t (13) = 3.60, p<2.07e-7, Bonferroni-corrected) for the type-one sequences. As expected, there was only a 4 Hz peak to reflect the rate of syllables for the type-two sequences (4 Hz, t (13) = 8.03, p<1.06e-6, Bonferroni-corrected) and the type-three sequences (4 Hz, t (13) = 6.08, p<1.92e-5, Bonferroni-corrected). The results are shown in ***Figure 4d***. The frequency responses corresponding to the type-one, type-two and type-three sequences are signified by the red, dark blue and light blue lines, respectively. The shaded area covers two SEM. The topographical distributions show the GED weights for the peaks of interest, where the larger the red circle, the higher the weight of the sensor.

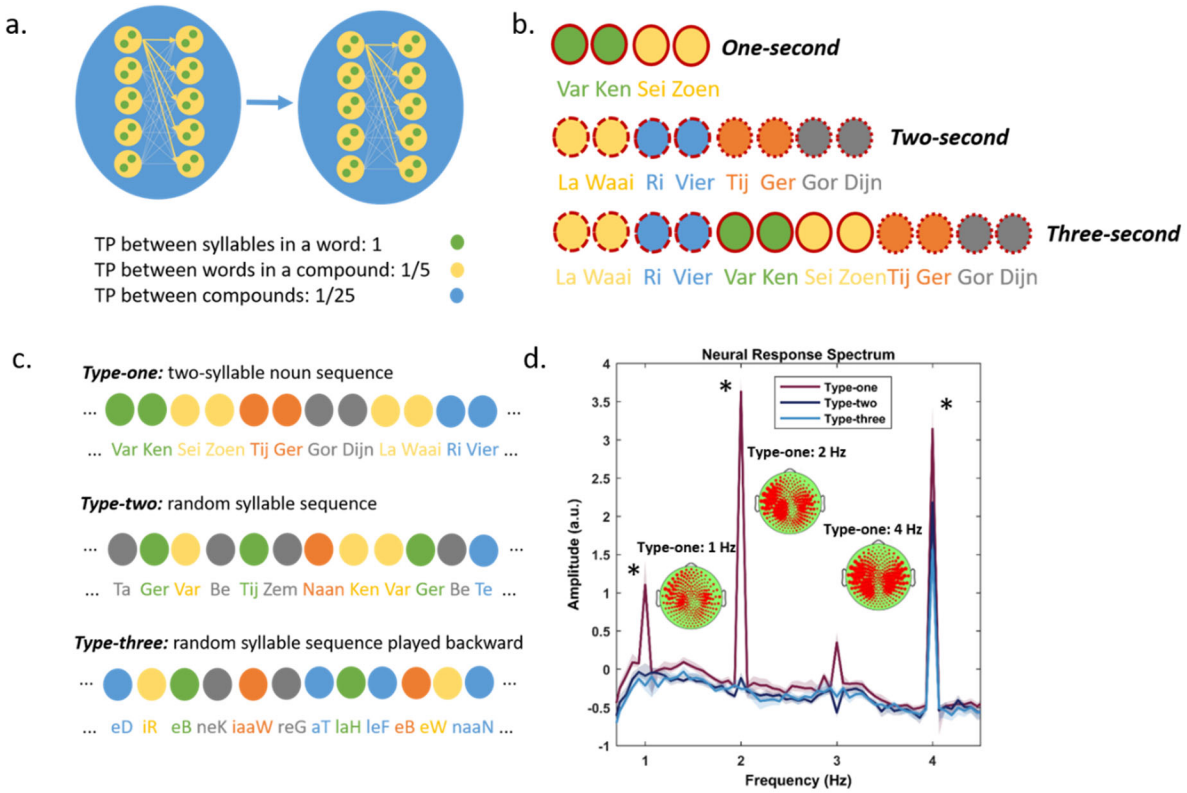

Figure 4. Neural activities of Chinese participants track statistically defined units in Dutch. (a) *The statistical framework for constructing syllable sequences in Dutch, in which the TPs between syllables in a word, between words in a compound, and*

between compounds were 1, 1/5 and 1/25, respectively. ***(b)*** *Sample sequences that were presented in each trial during the training experiment. The sequences were generated using a Markov chain to holding the statistical relationships between different levels' units. To make sure participants could extract the statistically defined compounds, sequences were manipulated to be one, two or three seconds in length.* ***(c)*** *The sequence structure used in Experiment 6, where the upper, middle and lower panels represent the structure of type-one, type-two and type-three sequences, respectively.* ***(d)*** *The neural response spectrum for each type of sequence, in which the 1 and 2 Hz peaks were significant for only type-one sequences, whereas the 4 Hz peak was significant for all three types. The topographical distributions represent the GED weights for the peaks of interest.*

Therefore, the data from Experiment 6 showed the same pattern as its counterpart in Chapter 2, suggesting that neural oscillations can simultaneously tag statistically defined units at different levels. In addition, the linguistic knowledge that participants had did not vary the pattern of results across users of different languages (Dutch vs. Chinese). More importantly, our results are at odds with the account of frequency tagging as an exclusive readout of unit extraction via inference using high-level linguistic knowledge, in that the neural readout cannot be said to be purely a reflection of grammatical processing or syntactic integration (Ding et al., 2016). Instead, by fitting statistical information into the cortical tracking effect, removing the availability of high-level language information and conducting experiments with different types of language users, we found that the frequency activities tagging the occurrence rate of hierarchical structures could be solely driven by statistical information. Moreover, by comparing the results in Chapters 2 and 3, we are able to argue that the cortical tracking effect (can also) reflect a generalized perceptual process – statistical inference.

3.4 Discussion

Both Chapter 2 and Chapter 3 have reported the results of a series of MEG experiments that investigate the fundamental question of how the brain segments speech into unit representations. By connecting the effect of the cortical tracking of hierarchical linguistic structures with statistical inference, we found that the

frequency-tagging response in the brain can be induced at different levels of structure solely by manipulating statistical cues (TP). Our results support the notion that speech segmentation (indexed by extracting statistically defined structures) can be conducted without using high-level language knowledge such as grammatical, syntactic and semantic information. This finding suggests that speech tracking or frequency tagging could be a general perceptual processing readout.

In the Experiment 1 of both chapters, we replicated the cortical tracking effect that was found by Ding et al. (2016), in both Dutch and Chinese. The results support the argument that the effect could be a language-independent phenomenon, and shows that it can at least be introduced when Dutch and Chinese participants listen to their own language. The occurrence rates of words (2 Hz) and syllables (4 Hz) were robustly reflected in the brain by using the same experimental paradigm as Ding et al. (2016). However, language-related cues such as grammatical, syntactic and semantic knowledge coexisted with the statistical information. Therefore, the frequency response tagging the rhythm of words and syllables could be explained by either of these two types of cues.

In Experiment 2 of both chapters, we removed the impact of linguistic knowledge by constructing three types of syllable sequences in a language that the participants did not know. Using the same experimental procedure and analysis methods, we found that frequency activity in the brain still tracked the occurrence rates of words (2 Hz) and syllables (4 Hz). The results can only be explained by structure chunking using statistical cues, as language-related information was not available to the participants. In addition, to determine whether the tracking mechanism could be independent of the linguistic knowledge that participants had, we conducted the experiment with both Dutch and Chinese participants. Our results pointed to a consistent conclusion which is that the effect is not a pure reflection of unit chunking using high-level linguistic knowledge. Instead, the neural oscillations which track the rhythm of linguistic structures could also be reflected by unit structuring via statistical inference.

However, in Experiments 1 and 2, we only showed that the brain tracked units at two levels (syllables and words). To fill the gap between our findings (two levels of tracking) and the original study (three levels of tracking) and see if we could

introduce the cortical tracking effect that tracks units at multiple levels, we needed a type of structure that is built on top of words. More importantly, if we wanted to introduce an additional peak to reflect structure chunking by statistical information, this type of structure needed to be statistically but not syntactically defined. In addition, to satisfy the rule of thumb, which is using the same stimuli to introduce a different neural response, the characteristics of the stimuli in the new experiment should be the same as the old one (Experiment 1). To satisfy these criteria, we constructed a type of four-syllable (one-second) novel compound using words from Experiment 1, and then training participants to extract these compounds in Markov-chain-manipulated sequences (Experiment 3, for details see section 3.2, Methods). After training, we conducted Experiment 4, in which the same three types of sequences, namely the noun sequences (type-one) and random syllable sequences played forward (type-two) and backward (type-three), were constructed using the trained items (10 bi-syllabic nouns from Experiment 3). By doing so, we found that there were three peaks in the brain's frequency response to reflect the occurrence rates of syllables (4 Hz), words (2 Hz), and novel compounds (1 Hz) when participants listen to the noun sequences (type-one sequences) in their own language.

The results are quite compelling. First, we found that the occurrence rates of different levels of structure were reflected in the neural response, which indicates that the brain can handle the boundaries from different structural levels simultaneously. Second, the additional peak corresponding to the rhythm of novel compounds (1 Hz) reflected statistical chunking (and could reflect semantic association, as noted below) because this 1 Hz peak did not occur in Experiment 1, in which the same noun sequences only introduced frequency responses corresponding to the rates of words (2 Hz) and syllables (4 Hz). For the same reason, we argue that this additional 1 Hz peak was not a reflection of unit chunking using high-level linguistic knowledge, because if the singular nouns can be chunked into a higher-level structure syntactically, there should be a 1 Hz neural activity to reflect this process in Experiment 1. Moreover, by comparing Experiment 1 and Experiment 4, it appears that the cortical tracking effect reflects an endogenous perceptual mechanism for segmenting speech into chunked structures, because the same stimuli can introduce different types of neural responses, i.e., two peaks (2 and 4 Hz) in Experiment 1 and three peaks (1, 2 and 4 Hz) in Experiment 4. Lastly,

we conducted the experiments with both Dutch and Chinese participants, and the consistent pattern across different types of participants indicated that linguistic knowledge itself was not sensitive to the tagging effect, which implicitly suggested that the effect might be associated with a general inference process.

Showing cortical activities tracking the statistically defined compounds is necessary to prove that the effect could be introduced by statistical information; however, in Experiments 3 and 4 of both chapters, the participants still listened to their own language, which means semantic association will inevitably occur. Therefore, one might hold a concern that the additional 1 Hz response corresponding to the compounds could reflect semantic association. To address this issue, we conducted Experiments 5 and 6, in which the same experimental procedures and parameters as Experiments 3 and 4 were used, but with the stimuli in an unfamiliar language. This enabled all higher-level, language-related cues that lead to structure chunking to be removed from the processing of the syllable sequence. As expected, we still found three peaks that correspond to the rhythm of syllables (4 Hz), words (2 Hz), and novel compounds (1 Hz) when participants listened to noun sequences in a language they did not know. In addition to the conclusions drawn from Experiments 3 and 4, the results at this stage could be evidence that the cortical activity tracking multiple levels' structures in speech can be introduced by statistical information alone (TP), which is at odds with the account that the effect is purely a reflection of unit chunking using high-level linguistic knowledge. The fact that the data from users of different languages showed the same pattern also suggests that the cortical tracking effect can be introduced by perceptual statistical inference.

Demonstrating that the cortical tracking effect can be solely driven by statistical information is important. One reason is that it enables us to establish whether speech segmentation can be performed prior to acquiring high-level language-related knowledge. For instance, as humans, our language acquisition starts with exposure to a language environment that is not yet meaningful to us; when we listen to speech as infants, the mechanism of segmenting speech into analyzable units might be our initial steps toward acquiring a language. From this perspective, the results are consistent with Saffran, Aslin, and Newport (1996), suggesting that speech segmentation could be conducted via statistical inference. In addition, our results show that different levels of TP-defined boundaries can be

tracked by the brain simultaneously. Another motivation behind the study is to check whether this effect is a purely reflection of inference using high-level linguistic knowledge. Apparently, our results do not support this account; instead, our experiments provide evidence that the neural activity tagging the rhythm of linguistic structures can be introduced without the presence of high-level linguistic knowledge. By conducting all of the experiments with both Dutch and Chinese participants, we also found that the type of linguistic knowledge that participants had was not a related factor. Therefore, based on our findings, we argue that the effect cannot be purely a reflection of an inference process using high-level linguistic knowledge. Instead, it can also be introduced via statistical inference (without language knowledge).

4 | Phase consistency as a window onto syntactic structure representation[1]

Abstract

Speech stands out in the natural world as a biological signal that communicates formally-specifiable complex meanings. However, the acoustic and physical dynamics of speech do not injectively mark the linguistic structure and meaning that we perceive. Linguistic structure must therefore be inferred through the human brain's endogenous mechanisms, which remain poorly understood. Using electroencephalography (EEG), we investigated the neural responses to synthesized spoken phrases and sentences that were closely matched physically but differed in syntactic structure. Differences in syntactic structure were well-captured in theta band (~ 2 to 7 Hz) phase coherence, with phase synchronization at low frequencies (< ~ 2 Hz). Theta-gamma phase-amplitude coupling was found when participants listened to speech, but it did not discriminate between syntactic structures. Our findings provide a comprehensive description of how the brain separates linguistic structures in the dynamics of neural responses, and imply that phase synchronization and connectivity strength can be used as readouts for constituent structure, providing a novel basis for future neurophysiological research on linguistic structure representation in the brain.

[1] Adapted from Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. PLoS Biology, 20(7), e3001713.

4.1 Introduction

To successfully understand speech, syntactic representations have to be formed via an inferential (top-down) process where grammatical relationships between hierarchical linguistic structures are constructed (Berwick et al., 2013; Chomsky, 2009; Phillips, 2003). Previous research has shown that neural activity synchronizes with the presence of linguistic structures in speech, which suggests the temporal properties of the neural oscillations (e.g., phase coherence) could reflect the building of linguistic structures by the brain (Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018). However, the construction of syntactic structures is often hard to isolate as both perceptual-level (e.g. evoked auditory) and linguistic-level (e.g. speech chunking via grammatical knowledge) processes are paralleled with it. Therefore, how the brain represents and discriminates syntactic structures is largely unknown. In this chapter, we investigate how the discrimination between two types of syntactic structures (i.e., a phrase vs. a sentence which has highly similar temporal-spectral features) is represented in the temporal synchronization of neural oscillations.

Low-frequency phase coherence (< 8 Hz) was heavily weighted as a critical neural readout for speech comprehension (Doelling et al., 2014; Howard & Poeppel, 2010; Luo & Poeppel, 2007; Peelle, Gross, & Davis, 2013). In the MEG study by Luo and Poeppel (2007), participants listened to sentences with systematically varied intelligibility from low to high. By fitting the low-frequency (theta band, 4 to 7 Hz) phase coherence as a function of the degree of intelligibility, the researchers found that the consistency of the theta band phase reliably reflected the intelligibility of the speech stimuli, in which higher-level phase coherence was evoked by highly intelligible sentences compared to degraded ones. This stimulus-driven temporal synchronization of the neural activities provided initial evidence that low-frequency phase tracking is an important component leading to comprehension. As such, the authors concluded that theta-phase entrained with the rhythm of syllables (often represented by the temporal envelope of speech stimuli) was a necessary condition for spoken language comprehension. Therefore, low-frequency phase coherence could reflect the neural representation of syllables.

Following the investigation by Luo and Poeppel (2007), an MEG study by Ding et al. (2016) further showed that neural oscillations simultaneously track the

occurrence rates of hierarchical linguistic structures. Specifically, the authors artificially synthesized isochronous syllable sequences with a built-in hierarchy and found that the frequency of cortical activity robustly reflected the occurrence rates of linguistic structures at different levels. The phenomenon of the frequency of neural activity tracking the rhythm of linguistic structures has been replicated in many other studies (Ding, Melloni, et al., 2017; Gui, Jiang, Zang, Qi, Tan, Tanigawa, Jiang, Wen, Xu, Zhao, et al., 2020; Jin, Lu, & Ding, 2020b; Jin et al., 2018a; Zhou et al., 2016).

Low-frequency neural oscillations may be especially important for speech processing because they occur roughly at the average syllable rate across various human languages (Ding, Patel, et al., 2017; Pellegrino, Coupé, & Marsico, 2011; Varnet et al., 2017). The brain may use syllables, which are abstract linguistic units, as the primitive units to analyze spoken language (Giraud & Poeppel, 2012; Luo & Poeppel, 2007; Poeppel & Assaneo, 2020). Indeed, a view has emerged wherein the brain employs an inherent cortical rhythm at a syllabic rate that can be altered by manipulations of linguistic structure or intelligibility. One possible synthesis of previous results is that low-frequency power reflects the construction of linguistic structures (Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018), whereas low-frequency phase coherence reflects the parsing and segmenting of speech signals (Doelling et al., 2014; Howard & Poeppel, 2010; Luo & Poeppel, 2007; Peelle, Gross, & Davis, 2013). Inspired by these hypotheses and empirical works, Martin and Doumas (2017) provided a theoretical, computationally explicit framework for understanding the role of low-frequency neural oscillations in generating linguistic structure. They reproduced the frequency-tagging results reported by Ding et al. (2016) in an artificial neural network model that uses time (unit firing asynchrony) to encode structural relations between words (Martin & Doumas, 2019). Based on their model, Martin and Doumas hypothesized that low frequency power and temporal synchronization should depend on the number of constituents that are represented at a given time step. In their model's coding scheme, constituents are represented as (localist) relations between distributed representations in time. Thus, the ongoing dynamics of the neural ensembles involved in coding linguistic units and their structural relations are what constitute 'linguistic structure' in such a neural system (Martin, 2016, 2020).

Indeed, we saw that the inherent cortical rhythm at the syllabic rate can be altered by both syntactic structures and semantic manipulations. As such, extracting the isolated neural readout for syntactic representation is helpful for better understanding speech perception and language comprehension. As an exploratory study, the current chapter investigated the role of the temporal synchronization of neural oscillations in syntactic structure discrimination and tested the hypothesis proposed by Martin and Doumas (Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas, 2020). In order to increase the likelihood that any observed patterns are due to representing and processing syntactic structures, we strictly controlled the physical and semantic features of our materials. We extend the work of Ding et al. (2016) and others to ask whether the 1 Hz neural response can be decomposed to reflect the discrimination of syntactic structures (phrases versus sentences). To assess this, we used two types of natural speech stimuli in Dutch, namely determiner phrases such as *de rode vaas* ('the red vase') and sentences such as *de vaas is rood* ('the vase is red'), which combine a subject with a verb into a proposition. These phrases and sentences were given matching properties in both physical and semantic dimensions, such as the number of syllables (four), the semantic components (same color and object), the duration in time (one second), the sampling rate (44.1 k Hz), and the overall energy (root-mean-square value equals -16 dB).

We formulated a general hypothesis that low-frequency neural oscillations would be sensitive to the difference in syntactic structure of the phrases and sentences. However, we did not limit our analysis to low-frequency phase coherence, as previous researchers had done (Brennan & Martin, 2020; Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018). We hypothesized that the neural response difference between phrases and sentences may manifest itself in a number of dimensions that are outside of the view of typical analyses of low-frequency phase coherence. We therefore employed additional methods to decompose the neural response to phrases and sentences, in order to address the following three questions:

Question 1. As previous studies have demonstrated the fundamental role of theta-band phase coherence in speech perception, our first concern was to test whether low-frequency (< 8 Hz) phase coherence could separate speech stimuli with different syntactic structures. To assess this, two types of speech stimuli were

constructed, such as *de rode vaas* (a phrase, 'the red vase') and *de vaas is rood* (a sentence, 'the vase is red'), with strict controls on the physical and semantic properties (for details see section 4.2, Methods). The number of constituents for sentences (a noun phrase and a verb phrase) is higher than and the number of units in phrases (a noun phrase), and the syntactic complexity is higher for sentences than phrases (Chomsky, 2009). Therefore, based on the theoretical model proposed by Martin and Doumas, we expected to see a higher-level phase coherence for sentences than phrases.

Question 2. We wondered whether phrases and sentences have different effects on brain dynamics as reflected at the functional neural network level (viz., functional connectivity). In the field of neuroscience, there is a rapidly growing interest in investigating functional connectivity to study whole-brain dynamics in sensor space (Cabral, Kringelbach, & Deco, 2014; Cohen, 2014; Cohen, 2015; Hutchison et al., 2013; Sporns, 2010), which can reveal temporal synchronization (viz., phase coherence) between brain regions. Neurophysiological techniques such as EEG and MEG have a high temporal resolution and are suitable for calculating synchronization across frequency bands in functional brain networks (Stam, Nolte, & Daffertshofer, 2007). Describing the temporal synchronization of the neural activity over the whole brain is the first step in decomposing neural responses to high-level variables like syntactic structure. We therefore investigated whether phrases and sentences have different effects on inter-site phase coherence (ISPC), which are considered to reflect the temporal synchronization of neural activity across different brain regions (Cohen, 2014; Lachaux et al., 2000; Mormann et al., 2000).

Question 3. We asked whether phrases and sentences have different effects on the coupling between the lower frequency phase and high frequency intensity. This question is related to the theoretical model proposed by Giraud and Poeppel (2012) on a generalized neural mechanism for speech perception. The model suggests that presentation of the speech stimulus first entrains an inherent neural response at low frequencies (< ~ 8 Hz) in order to track to the speech envelope, from which the neural representation of syllables is then constructed. Then, this low frequency response evokes a neural response at a higher frequency (~25 to ~35 Hz), which reflects the brain's analysis of phonemic-level information. The model proposes that the coupling of the low- and high-frequency neural responses (theta

and gamma, respectively) is the fundamental neural mechanism for speech perception up to the syllable. We therefore investigated whether theta-gamma frequency coupling may also differentiate higher-level linguistic structures, namely phrases and sentences.

In sum, by performing an electroencephalography (EEG) experiment, we explored how the discrimination between two types of (normalized) syntactic structures would be reflected in the temporal characteristics of the neural response. As set out in this chapter, our investigations may serve as a trail marker on the path towards a theory of the neural computations underlying the formation of syntactic structure.

4.2 Methods

Participants

Fifteen Dutch native speakers (8 females and 7 males), aged 22 to 35, participated in the study. All of them were undergraduate or graduate students and were right-handed. They reported no history of hearing impairment or neurological disorder. The experimental procedure was approved by the Ethics Committee of the Social Sciences Department at Radboud University. Written informed consent was obtained from each participant before the experiment, and they were paid for their participation.

Stimuli

Fifty line-drawings of common objects were selected from a standardized corpus (Snodgrass & Vanderwart, 1980). The Dutch names of all the objects were mono-syllabic and had non-neuter gender. In our experiment, the objects appeared as colored line-drawings on a grey background. More specifically, we presented each line-drawing in five colors: blue (*blauw*), red (*rood*) yellow (*geel*), green (*groen*), and purple (*paars*). In total, this yielded 250 pictures. The line-drawings were sized to fit into a virtual frame of 4 by 4 cm, corresponding to a 2.29° visual angle for the participants.

We then selected 50 drawings of different objects, 10 for each color, to create the speech stimuli. For each selected line-drawing, a four-syllable phrase-sentence pair was created, e.g. *de rode vaas* (‘the red vase’) and *de vaas is rood* (‘the vase is

red'). This means that in total, we had 100 speech stimuli (50 phrases and 50 sentences; see Appendix 1). All stimuli were synthesized by an online synthesizer (www.neospeech.com) using a Dutch male voice, Guus. All stimuli were 733 to 1125 ms in duration (mean = 839 ms, SD = 65 ms). To normalize the synthesized auditory stimuli, they were first resampled to 44.1 kHz. Then each speech stimulus was cut or zero-padded to fit in a 1000 ms window without missing any meaningful dynamics. Ten percent at both ends of each stimulus was smoothed by a linear ramp (a cosine wave) to remove the abrupt sound burst. To normalize the intensity of the stimuli, the root-mean-square value of each stimulus was controlled to be -16 dB (see ***Figure 3***).

Experimental procedure

Each trial started with a fixation cross being visible at the center of the screen (for 500 ms). Participants were asked to look at the screen. Immediately after the fixation cross had disappeared, the participants heard a 1000 ms spoken stimulus, either a phrase or a sentence, followed by a three-second silence; then the participants were asked to perform one of three types of task, indicated to them by an index number (1, 2 or 3 showing at the center of the screen, for 500 ms in duration). If the index number was '1', they did a linguistic structure discrimination task (type one), in which they had to judge whether the spoken stimulus was a phrase or a sentence. If the index number was '2', there was a 1000 ms pause and then a picture was shown for 200 ms. Then participants would do a color-matching task (type two), in which they had to judge whether the color described in the spoken stimulus matched the color shown in the picture. If the index was '3', they would experience the same procedure as the type-two task, except instead of matching colors, they would do an object-matching task (type three), judging whether the objects in the spoken stimulus were the same as in the picture. All responses were recorded via a parallel port response box, in which the two buttons were labeled as 'phrase/match' and 'sentence/mismatch'. Each response was followed by a silent interval of 3 to 5.2 seconds (jittered).

The data collection was broken down into five blocks, with 48 trials in each block. Before the core data collection, several practice trials were conducted for each participant to make sure they understood the task. Trials in each block were fully matched in across linguistic structure (phrase or sentence) and task type (1, 2

or 3). For instance, half of the spoken stimuli were phrases and half were sentences (24 of each type), and six combinations (eight trials for each type) were evenly distributed in each block (eight trials times two linguistic structures times three task types). The order of the trials was pseudo-random throughout the whole experiment. The behavioral results indicated that the task was relatively easy and no difference was found between the phrases and the sentences. For all tasks combined, the accuracy rates for phrases and sentences were 97.9 ± 3% and 97.3 ± 3% (p = 0.30), respectively. The experimental procedures are shown in ***Figure 1***.

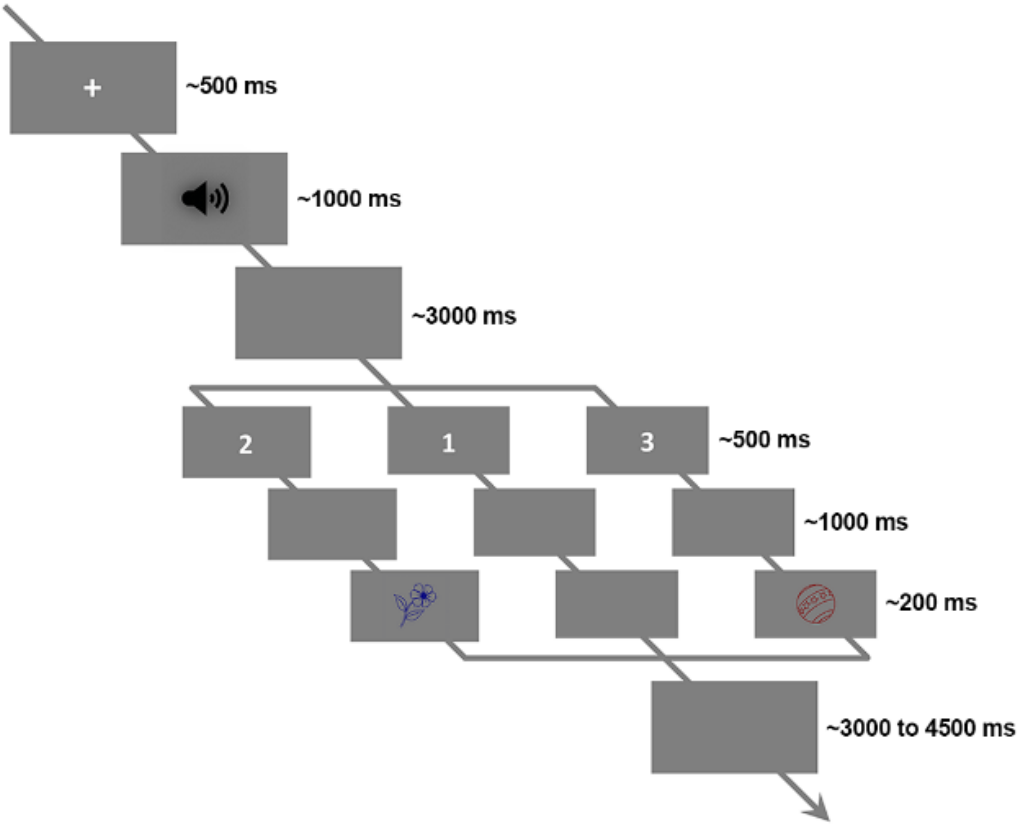

Figure 1. An illustration of the experimental procedure. *Participants were asked to look at the center of the screen, and after hearing the speech stimulus they would do a task indexed by a number that appeared on the screen. If the number was '1', they would discriminate whether the stimulus they heard was a phrase or a sentence. If the number was '2', they would see a picture and then judge whether the color in the picture was the same as the color described in the speech stimulus. If the number was '3', they would judge whether the object in the picture was the same as the object described in the speech stimulus. Trial type was pseudo-randomly assigned throughout the whole experiment.*

After the main experiment, a localizer task was performed, in which a 'beep' tone (1 kHz, 50 ms in duration) was played 100 times (jitter 2 to 3 seconds) for each participant, in order to localize the canonical auditory response (N1-P2 complex). The topographies for N1 and P2 are shown in ***Figure 2.*** The upper panel shows the averaged N1-P2 complex of all participants over the time bin from 90 to 110 ms for N1 and 190 to 210 ms for P2. The lower panel shows the N1-P2 complex after applying surface Laplacian, in which the effect of the volume conduction was attenuated. The topographies indicated that all participants had the canonical auditory response.

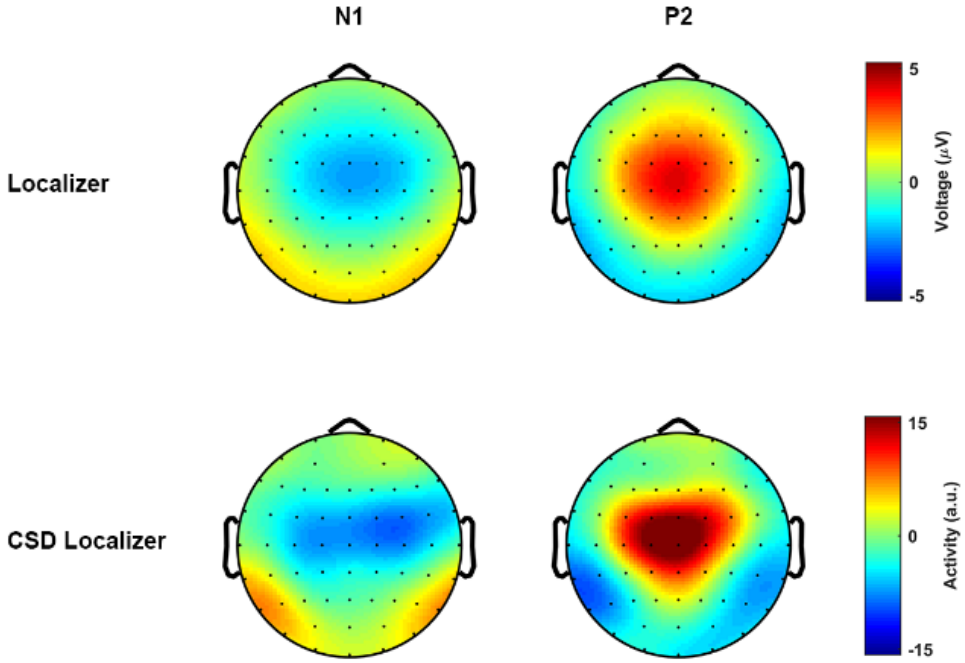

Figure 2. Effect of volume conduction attenuation. *The topographical distribution of the canonical auditory N1-P2 complex evoked by the localizer task. The upper panel shows the N1-P2 complex that was averaged across all participants (N=15). The lower panel shows the N1-P2 complex after applying surface Laplacian (current source density), in which the effect of volume conduction was attenuated.*

EEG recording

EEG data was recorded using a 64-channel active sensor system from Brain Products (GmbH) in a sound-dampened, electrically shielded room. Signals were digitized online at 1000 Hz, with high-pass and low-pass at 0.01 Hz and 249 Hz,

respectively. Two electrodes, AFz and FCz, were used as ground and reference. All electrodes were placed on the scalp based on the international 10-20 system and the impedance of each one was kept below 5 kΩ.

The experimental procedure was controlled by MATLAB 2019a (The MathWorks, Natick, MA) with Psychtoolbox-3 (Brainard, 1997). Auditory stimuli were played at 65 dB SPL and delivered through air-tubes earplugs (Etymotic ER-3C, Etymotic Research, Inc.). Event markers were sent via a parallel port for tagging the onset of the events under investigation (i.e., speech onset, task index onset, etc.).

EEG data preprocessing

The EEG data preprocessing was conducted via MATLAB using the EEGLAB toolbox (Delorme & Makeig, 2004) and customized scripts. The data were first down-sampled to 256Hz then high-pass filtered at 0.5 Hz (finite impulse response filter, FIR; zero-phase lag). The raw data were first cleaned by the time-sliding PCA (Chang et al., 2018; Kothe & Jung, 2016). Then all bad channels were interpolated with spherical interpolation. After transferring the data to average reference, the online reference FCz was recovered and the line noise, 50 Hz and its harmonics, was removed.

Following the above steps, epochs of two seconds preceding and nine seconds following the auditory stimulus onset were extracted. The deletion of bad trials and removal of artifacts were conducted in two steps. First, independent component analysis (ICA) was used for decomposing the data into the component space (number of components equals data rank). Then for each independent component, we used the short-time Fourier transform to convert each trial into the power spectrum, in which we extracted a value that was calculated by the power summation between 15 and 50 Hz. Then all the extracted values in each component formed a distribution. From this distribution, we transformed all the extracted values into z-scores, and the epochs with values outside the range of plus or minus three standard deviation were deleted. Second, ICA was conducted again on the trial-rejected data for eye-related artifact removal and muscle activity elimination. Artifact components were identified and removed using an automatic classification algorithm (Winkler, Haufe, & Tangermann, 2011). All the preprocessing steps resulted in the removal of, on average, 7 components (range 4 to 11) and 22 trials

(including incorrect trials and trials with excessively slow responses, range 10 to 30, 4% to 12.5%) per participant. Finally, volume conduction was attenuated by applying surface Laplacian (Cohen, 2014; Srinivasan et al., 2007; Winter et al., 2007).

EEG data analysis

Time frequency decomposition

To perform time-frequency decomposition, the single-trial time series were convolved with a family of complex wavelets (1 to 50 Hz in 70 logarithmically spaced steps). Temporal and spectral resolution were optimized by changing the cycle from 3 to 30 in logarithmical steps. Phase coherence was calculated by inter-trial phase clustering (ITPC) (Cohen, 2014; Lachaux et al., 1999). At each time-frequency bin, the wavelet coefficients of all trials were divided by their corresponding magnitude and averaged across trials. The magnitude of the averaged complex output was represented as phase coherence (ITPC).

Phase connectivity

Trials in each condition first experienced the wavelet convolution (with the same parameters as the time-frequency decomposition). Then the cross-spectral density (CSD) was calculated for each sensor pair at each frequency-time-trial bin. Phase connectivity over the sensor space was calculated by inter-site phase coherence (ISPC) (Cohen, 2014; Lachaux et al., 2000; Mormann et al., 2000), in which we divided the complex coefficients from the CSD output by the corresponding amplitude at each frequency-time-trial bin. Then averaging was conducted across all trials. The amplitude of the averaged output was represented as phase connectivity between sensors (ISPC). After the above steps, the phase connectivity at each time-frequency bin had a matrix representation where the connectivity between all sensor pairs was represented as an all-channels-by-all-channels matrix (graph), in our case 65 channels * 65 channels.

For comparing the phase connectivity level between the phrases and sentences in the time-frequency space, a statistical threshold method was deployed. More specifically, at each time-frequency bin, we formed a distribution by pooling together all the connectivity values from both conditions, and then defined the threshold as the value at half the standard deviation above the median. We then

binarized the connectivity graph for both conditions by using this threshold at each bin. The connectivity level at each time-frequency bin was represented as the total number of connectivity values that were above this threshold. Finally, we transferred the connectivity level at each time-frequency bin as the percentage change relative to the baseline, which was calculated as the average connectivity level at 800 to 200 ms before the audio onset.

Phase-amplitude coupling (PAC)

Since the low-frequency phase and high-frequency amplitude were supposed to show coupling when the speech stimuli were processed, we initially defined the frequency range for the phase series as 1 to 16 Hz in a linear step of 1.5 Hz, and the frequency range for the amplitude series as 8 to 50 Hz in 12 logarithmic steps. Then, the wavelet convolution was performed to extract the analytic signal, in which the phase time series and amplitude time series were extracted at the specified frequency ranges from 50 ms before to 1500 ms after the audio onset. At each phase-amplitude bin, a complex time series composed of the phase angle of the phase time series and the magnitude of the amplitude time series was constructed. The PAC at each bin was calculated by extracting the magnitude of the average of all the vectors in the complex time series (Canolty et al., 2006; Cohen, 2014). Since the variation of the amplitude response, a z-score normalization was also performed for each phase-amplitude bin. More specifically, we first calculated the real PAC value by using the raw complex time series. Then the random PAC value was computed 1000 times by using the constructed complex time series. These constructed series were built by temporally shifting the power time series with a random temporal offset without changing the phase time series. These 1000 random PAC values formed a reference distribution for each phase-amplitude bin. Then the z-score of the real PAC value in this distribution was represented as the index of the phase-amplitude coupling, PAC-Z.

Statistical analysis

In addition to using parametric statistical methods to check whether the difference between phrases and sentences was significant, a cluster-based non-parametric permutation test was applied. This method deals with the multiple-comparisons problem and at the same time takes the data's dependency (temporal, spatial and spectral adjacency) into account. For all types of analysis that followed

this inference method, the subject-level data were initially averaged over trials and for each single sample, i.e. a time-frequency-channel point, a dependent t-test was performed. We selected all samples for which the t-value exceeded an a priori threshold, $p<0.025$, and these samples were subsequently clustered based on spatial and temporal-spectral adjacency. The sum of the t-values within a cluster was used as a cluster-level statistic. The cluster with the maximum sum was subsequently used as test statistic. By randomizing the data across the two conditions and recalculating the test statistic 1000 times, we obtained a reference distribution of the maximum cluster t-values. This distribution was used to evaluate the statistics of the actual data. This statistical method was carried out using the FieldTrip toolbox (Maris & Oostenveld, 2007; Oostenveld et al., 2011).

Acoustic normalization and analyses

In order to normalize the synthesized auditory stimuli, they were first resampled to 44.1 kHz. Then all speech stimuli were adjusted by truncation or zero padding at both ends to 1000 ms without missing any meaningful dynamics. Then 10% at both ends of each stimulus was smoothed by a linear ramp (a cosine wave) for removing the abrupt sound burst. Finally, to control the intensity of the speech stimuli, the root-mean-square value of each stimulus was normalized to -16 dB.

The intensity fluctuation of each speech stimulus was characterized by the corresponding temporal envelope, which was extracted by the Hilbert transform of the half-wave rectified speech signal. Then each extracted temporal envelope was down-sampled to 400 Hz. For checking the acoustic properties in the frequency domain, the discrete Fourier transform was performed to extract the spectrum of the temporal envelope. Decibel transformation for the spectrum of each speech stimulus was performed by using the highest frequency response in the corresponding phrase-sentence pair as the reference.

Figures 3a and **3b** show the syntactic representation of the phrases and sentences. Since all the phrases, as well as all the sentences, have the same syntactic structure, we selected a sample pair, *de rode vaas* ('the red vase') and *de vaas is rood* ('the vase is red'), to show the syntactic decomposition. Four syllables were strictly controlled to be the physical input for both conditions. Syntactic integration, i.e. the way in which the physical input is combined into a linguistically logical structure, is different. The syntactic structure for the sentences is more

complicated than that of the phrases with respect to numbers and types of constituents. **Figure 3c** and **3d** show the spectrogram of a sample phrase-sentence pair. The comparison suggests a similar temporal-spectral pattern in this sampled pair. **Figure 3e** shows the temporal envelopes of this sample pair, the blue line for the phrase and the red line for the sentence, respectively. The comparison suggests a highly similar energy fluctuation between the phrase and the sentence. **Figure 3f** shows the intensity relationship of this sample pair in each frequency bin. The Pearson correlation was calculated to reveal the similarity between the spectrum of this sample pair ($r = 0.94$, $p < 1e\text{-}4$ ***). The comparisons indicated that they are highly similar in acoustic features. In this figure, the darker the dots, the lower the frequency of the spectrum.

Figure 3g shows the temporal envelope averaged across all the stimuli (N=50), with the blue and red lines representing the phrases and sentences, respectively. The shaded areas cover two SEM. To check the similarity of the instantaneous intensity on the temporal envelopes between the phrases and sentences, we first calculated the cosine similarity. For each time bin (400 bins in total), the similarity measure simultaneously treats the activity of all stimuli as one vector while considering each stimulus as one dimension (50 dimensions in total). To add the signal-to-noise ratio, the energy fluctuation was averaged using a 50 ms window centered on each bin. Statistical significance was evaluated via a permutation approach. Specifically, we generated a reference distribution with 1000 similarity values, each of which was selected as the largest value of the cosine similarities that were calculated using the raw phrase envelopes with the time-shuffled sentence envelopes. Our simulations suggested a threshold of 0.884 corresponding to the p-value of 0.05, as shown on the right-hand vertical axis. The statistical analysis indicated a high similarity between the phrases and sentences on the temporal dimension of the energy profile.

Figure 3h shows the comparison between the averaged spectrums of all phrases and all sentences. These spectrums were considered to reflect the prosodic information of the speech stimulus (Ding et al., 2016; Gui, Jiang, Zang, Qi, Tan, Tanigawa, Jiang, Wen, Xu, & Zhao, 2020; Henin et al., 2021; Jin, Lu, & Ding, 2020a; Jin et al., 2018b). In this figure, the shaded area covers two SEM across the stimuli. A statistical comparison using a paired sample t-test was conducted at each frequency bin, in which no evidence was found to indicate a significant physical

difference between the phrases and sentences. In addition, to show a statistically similar frequency response on the energy profiles between the phrases and sentences, a robust Bayesian inference on all frequency bins above 1 Hz was conducted. Specifically, for each frequency bin, we first combined the instantaneous intensities across conditions into one pool. Then, a prior gamma-distribution for the mean of each condition was generated, where the mean equals the average value of the pool and the standard deviation equals five times the standard deviation of the pool. The normality for both conditions was governed by a constant value of 30. The posterior distribution was recurrently updated using a Markov chain Monte Carlo (MCMC), and the statistical significance was determined according to whether zero was located in the 95% highest density interval (HDI) of the posterior distribution for the difference of the means. A robust Bayesian estimation allows us to accept the null hypothesis when the 95% HDI is entirely located within the empirical range (-0.1 to 0.1) for the region of practical equivalence (ROPE) (Carlin & Louis, 2009; Freedman, Lowe, & Macaskill, 1984; Freeman, Spiegelhalter, & Parmar, 1994; Hobbs & Carlin, 2007; Kruschke, 2014; Kruschke, 2011, 2013, 2018). Our analysis on all frequency bins suggested that there is no difference in the spectral dimension of the envelopes (the 95% HDI located in the ROPE range from -0.1 to 0.1) between the phrases and sentences. **Figure 3i** shows the simulation results using TRF. The reason for doing this is to demonstrate that any effect observed in this study is not driven by acoustic differences, and that the acoustic features are statistically matched in the temporal dimension. The underlying assumption is that if the physical-acoustical properties of the phrases and sentences are similar, then fitting a kernel (TRF) using these speech stimuli with the same signal would give similar results. By testing this hypothesis, we fitted two TRFs for each condition 15 times (to imitate the number of participants), each time with 100 simulated acoustic-response pairs. The acoustic input was constructed by randomly selecting 15 speech stimuli in the corresponding condition, and the simulated response was sampled from the standard Gaussian distribution. Optimization was performed using ridge regression and leave-one-out cross validation. After fitting the kernels, a paired sample t-test on each time point was conducted, and the comparison suggested no difference between the TRFs. Therefore, the simulation results also indicate that

the phrases and sentences had statistically similar acoustic properties with regard to time.

As one might be interested in the acoustic comparison not only of the targeted pairs (e.g., the pairs with the same semantic components), but also the pairs within conditions (e.g., a comparison of two phrases), we performed a similarity analysis on all the possible pairs in our stimuli. To do so, we first calculated the cosine similarity on the energy profile between any two of our stimuli, then depicted the results in a representational similarity matrix (RSM). Our analysis suggested a high similarity pattern, in which the mean similarity value was 0.74 (maximum 1, ranging from 0.43 to 0.94 when omitting pairs with the same stimuli which equals 1). To test the statistical significance, we performed a permutation test 1000 times to form a null distribution. In each iteration, we calculated the cosine similarity between a randomly selected real envelope and another randomly selected envelope that was shuffled in time. Our manipulations suggested a threshold of 0.496 corresponding to a p-value of 0.05. The results indicated that 98.91% of all pairs were statistically similar. Note that only one targeted pair (i.e., the pair with controlled semantic components) did not reach the threshold. The results are shown in **Figure 3j**, in which the pairs with the similarity values lower than the threshold are represented by dark blue squares. (Note that the dark blue cross that separates the RSM into four regions serves merely as a reference grid, not data points.)

In order to check whether syllables were the initial processing units, and also whether the syntactic integration would be reflected at the one-second interval we conducted a frequency-tagging analysis. In doing so, we constructed 40 trials of 15 seconds for each participant by randomly selecting the neural responses corresponding to the phrase condition and the sentence condition. Then the discrete Fourier transform was performed to extract the frequency neural response. Decibel transformation was conducted based on the neural response at the baseline stage. A grand average was calculated to check the frequency domain characteristics. **Figure 3k** shows that there was a 1 Hz peak for both conditions. To check the statistical significance, a paired sample t-test, for both conditions, was conducted between the 1 Hz peak and the averaged frequency response around it, with a window of five bins on each side. The 1 Hz peak was statistically significant for both the phrase condition ($t(14) = 8.72$, $p < 4.9e\text{-}7$ ***) and the sentence

condition (t (14) = 8.46, p < 7.1e-7 ***). The results indicate that syntactic integration (Ding et al., 2016) happened at the one-second interval and our one-second-duration normalization was effective. However, we can see that using the frequency-tagging approach makes it difficult to separate the two types of syntactic structures (t (14) = 0.63, p=0.53).

Figure 3l shows the response spectrum around 4 Hz. A paired sample t-test suggests that there was a strong 4 Hz response for both the phrases (t (14) = 7.79, p < 1.8e-6 ***) and the sentences (t (14) = 9.43, p < 1.9e-7 ***). The results suggest that syllables were the initial processing units for both phrases and sentences (Ding et al., 2016).

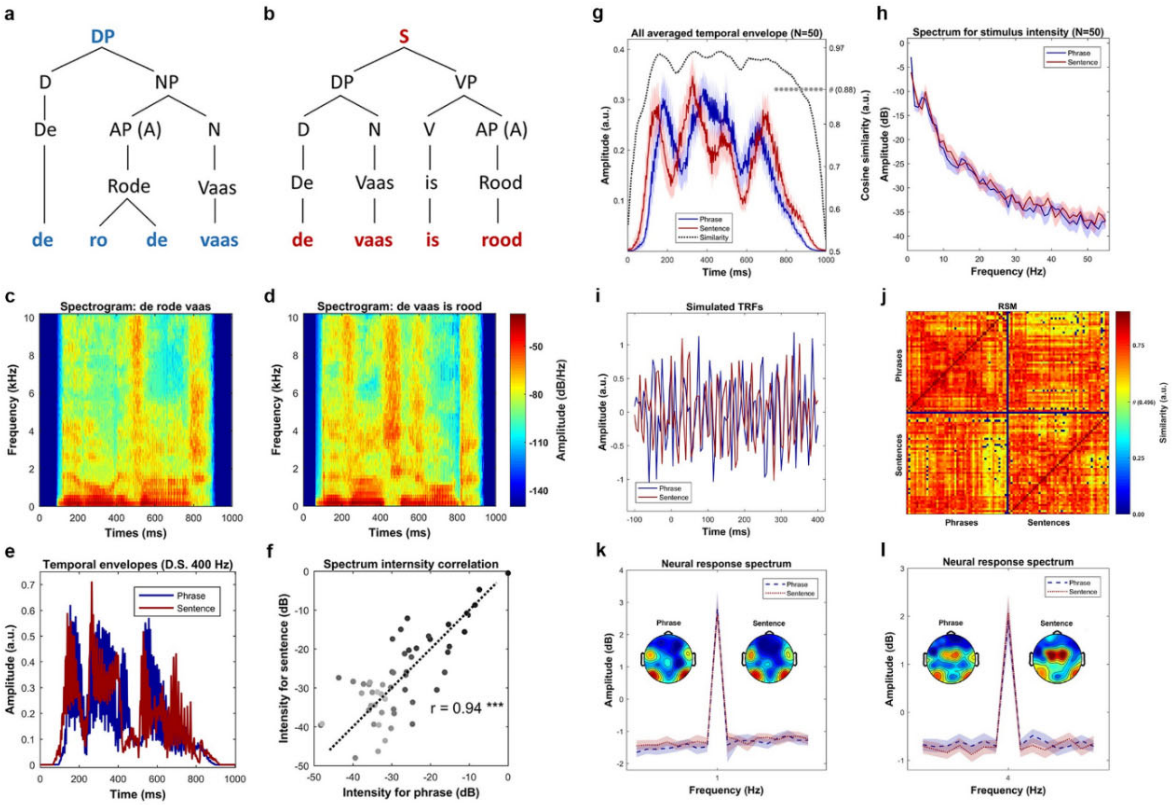

Figure 3. Stimulus comparison between phrases and sentences. (a) *The syntactic structure of the phrases, which is represented by a sample phrase,* de rode vaas *('the red vase'). The determiner phase (DP) can first be decomposed into a determiner (D) and a noun phrase (NP), in which the NP can be separated into an adjective phrase (AP), which constitutes an adjective (A) and a noun (N). Finally, these words can be decomposed into four syllables.* ***(b)*** *The syntactic structure of sentences, which is represented by a sample sentence,* de vaas is rood *('the vase is red'). The sentence can be decomposed into two parts, which are a determiner phrase (DP) and a verb phrase (VP).*

The DP can then be separated into a determiner (D) and a noun (N), and the VP can be separated into a verb (V) and an adjective (A). All these words are finally decomposed into four syllables. ***(c)*** *and* ***(d)*** *The spectrograms of the sample phrase and sample sentence. The comparison between the spectrograms indicates a similar pattern between these two types of stimulus.* ***(e)*** *The comparison of the temporal envelopes of the sample phrase-sentence pair, i.e.,* de rode vaas *('the red vase') vs.* de vaas is rood *('the vase is red'), which were down-sampled to 400 Hz. The comparison suggests a similar energy profile across the sample pair.* ***(f)*** *The spectrum for the sample phrase-sentence pair, in which the horizontal and vertical axes indicate the frequency response of the temporal envelope of the phrase and sentence, respectively. The darker the dot indicates the higher the frequency. The Pearson correlation suggested that the spectrum of the sample phrase and sample sentence are highly similar (r = 0.94, p < 1e-5 ***).* ***(g)*** *The averaged temporal envelope of these two types of stimuli, blue for phrases and red for sentences. The black dotted line indicates highly similar physical properties between them in the time domain by cosine similarity. The statistical analysis on the similarity measure using a permutation test indicated an inseparable pattern.* ***(h)*** *Spectrum of the averaged envelopes for the two types of speech stimuli. The shaded area for each condition covers two SEM (N=50). Statistical analysis using Bayesian inference suggested a highly similar frequency response.* ***(i)*** *The results of the simulations using the temporal response function. Statistical analysis using a pairwise t-test indicated no difference between the two types of stimuli in any time point, which suggests similar acoustic properties between the two types of stimuli.* ***(j)*** *Similarity comparison for all possible stimuli pairs. As shown in the RSMs, the upper-left and lower-right panels show the comparison of all phrase pairs and all sentence pairs, respectively. The upper-right and lower-left matrices show the comparison of all possible phrase-sentence pairs. Statistical comparison using a permutation test indicated highly similar acoustic properties across all possible pairs.* ***(k)*** *The frequency-tagging effect at 1 Hz. The figure shows a strong peak at 1 Hz for the phrases (t(14) = 8.72, p < 4.9e-7 ***) and the sentences (t(14)=8.46, p < 7.1e-7 ***). It reflects that syntactic integration happened at 1 Hz, and our duration normalization (at one second) was effective. However, no difference in the 1 Hz activity was found between the two conditions, which points to the difficulty of separating the two types of syntactic structures (t(14) = 0.63, p=0.53) using the frequency-tagging approach.* ***(l)*** *The frequency-tagging effect at 4 Hz. The strong 4 Hz peak for the phrases*

*(t(14) = 7.79, p < 1.8e-6 ***) and the sentences (t(14) = 9.43, p < 1.9e-7 ***) suggests that syllables were the initial processing units for syntactic integration.*

4.3 Results

Low-frequency phase coherence distinguishes phrases from sentences

To answer our first question, whether the low frequency neural oscillations distinguish phrases and sentences, we calculated the inter-trial phase coherence. We then performed non-parametric cluster-based permutation tests (1000 permutations) on a time window of 1200 ms starting at the audio onset and over the frequencies from 1 Hz to 8 Hz. The comparison indicated that phase coherence was significantly higher for sentences than phrases (p < 1e-4 ***, two-tailed). In the selected latency and frequency range, the effect was most pronounced at central electrodes. ***Figure 4a*** shows the temporal evolution, in steps of 50 ms, of the effect which is computed as the phase coherence of the phrase condition minus the phrase coherence of the sentence condition. ***Figure 4b*** shows the time-frequency plot using all the sensors in this cluster, in which the upper and lower panels are the plots for the phrase condition and sentence condition, respectively.

The results indicate that the low-frequency phase coherence can reliably distinguish phrases and sentences, consistent with the hypothesis that low-frequency phase coherence represents cortical computations over speech stimuli (Brennan & Martin, 2020; Doelling et al., 2014; Howard & Poeppel, 2010; Kaufeld et al., 2020; Luo & Poeppel, 2007; Martin, 2016, 2020; Meyer & Gumbert, 2018; Peelle, Gross, & Davis, 2013; Rimmele et al., 2018). Our findings therefore suggest that low-frequency phase coherence is involved in speech comprehension at the level of syntactic inference.

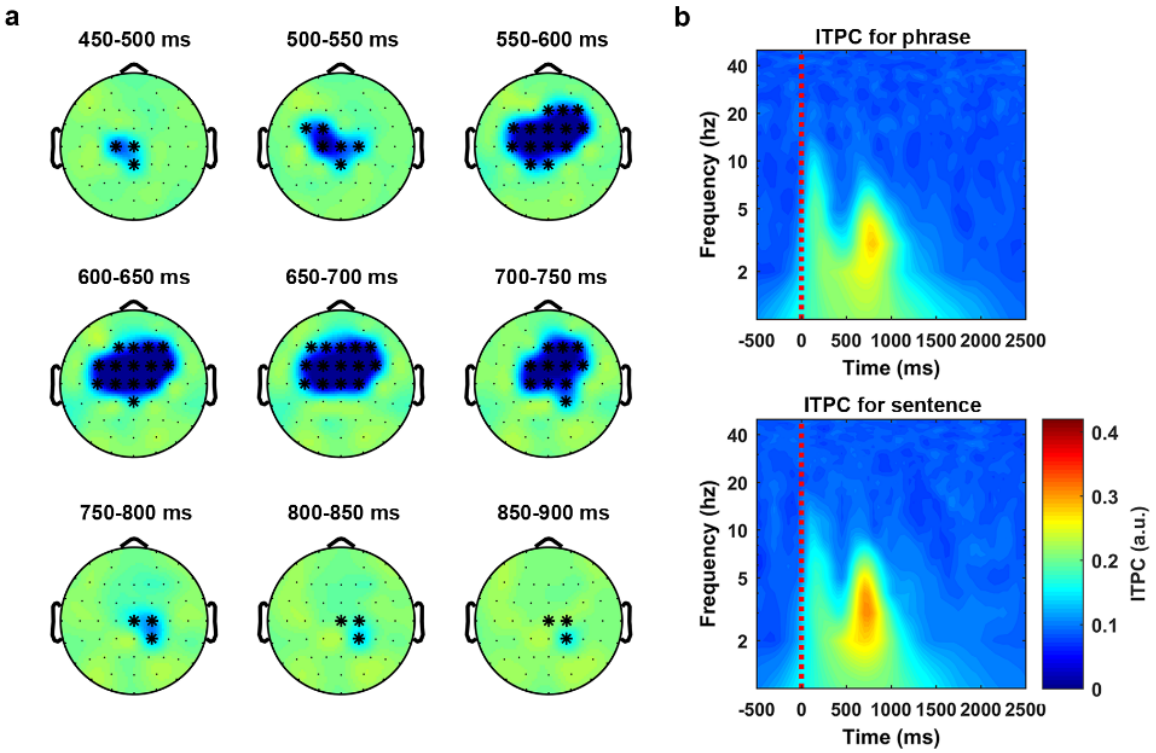

Figure 4. Phase coherence separates phrases from sentences at the theta band. *Statistical analysis on the phase coherence (ITPC) was conducted by using the non-parametric cluster-based permutation test (1000 times) on a time window of 1200 ms, which started at the audio onset and over the frequencies from 2 to 8 Hz. The results indicated that the phase coherence was higher for the sentences than the phrases (p < 1e-4 ***, two-tailed).* ***(a)*** *The temporal evolution of the cluster that corresponds to the separation effect. The activity was drawn by using the ITPC of the phrases minus the ITPC of the sentences. The topographies were plotted in steps of 50 ms.* ***(b)*** *The ITPC averaged over all the sensors in the cluster. The upper and lower panels show the ITPC of the phrase condition and the sentence condition, respectively.*

The degree of low-frequency (< ~ 2 Hz) phase connectivity separates phrases and sentences

We initially calculated the phase connectivity over the sensor space by ISPC at each time-frequency bin (for details see section 4.2, Methods). We then used a statistical threshold method to transform each connectivity representation into a super-threshold count at each bin. After baseline correction, we conducted a cluster-based permutation test 1000 times on a time window of 3500 ms starting at the audio onset and over the frequencies from 1 to 8 Hz, to compare the degree

of the phase connectivity between the phrases and sentences. The two structures showed a significant difference in connectivity ($p < 0.01$ **, two-tailed). The effect corresponded to a cluster extended from ~1800 ms to ~2600 ms after the speech stimulus onset, and was mainly located at a very low frequency range (< ~2 Hz). In the selected latency and frequency range, the effect was most pronounced at the right posterior region. ***Figure 5a*** shows the temporal evolution of this separation effect, which is represented by the degree of connectivity of the phrase condition minus that of the sentence condition (in steps of 100 ms). ***Figure 5b*** shows the time-frequency plot of the degree of phase connectivity, which is averaged across all sensors in this cluster. The left and right panels are the time-frequency plots for the phrase condition and the sentence condition, respectively.

Since the statistical analysis indicated a difference between phrases and sentences in the degree of phase connectivity, we assessed how this effect was distributed in the sensor space. To do so, we extracted all binarized connectivity matrices that corresponded to the time and frequency range of the cluster and averaged all the matrices in this range for both conditions. **Figure 5c** shows the averaged matrix representation of the sentence condition minus that of the phrase condition. This result suggests that the connectivity difference was mainly localized at the frontal-central area. After extracting the matrix representation, we used all sensors of this cluster as seeds to plot connectivity topographies for both conditions. **Figure 5d** shows the pattern of the thresholded phase connectivity. The black triangles represent the seed sensors. The upper and lower panels represent the phrase and sentence condition, respectively. The figure shows how the phase connectivity (synchronization) is distributed on the scalp in each condition. From this figure we can see that the overall degree of the phase connectivity was stronger for the sentence condition than the phrase condition.

The analysis indicated that the degree of phase connectivity over the sensor space at the low frequency range (< ~ 2 Hz) could reliably separate the two syntactically different stimuli and that the effect was most prominent at the right posterior region.

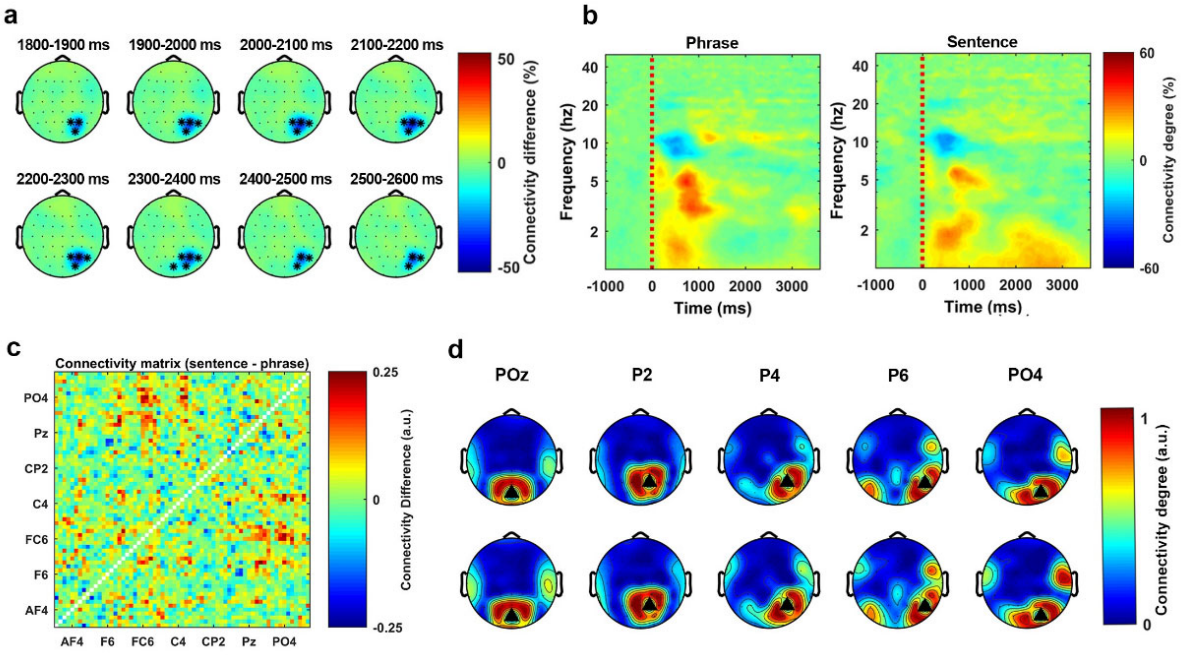

Figure 5. Low-frequency phase connectivity separates phrases and sentences. *Statistical analysis on the phase connectivity degree was conducted by using the non-parametric, cluster-based permutation test (1000 times) on a time window of 3500 ms, which started at the audio onset and over the frequencies from 1 to 8 Hz. The results indicated that the degree of phase connectivity was higher for the sentences than the phrases (p < 0.01**, two-tailed).* ***(a)*** *The temporal evolution of this cluster. The activity was drawn by using the degree of connectivity of the phrase condition minus that of the sentence condition. The topographies were plotted in steps of 100 ms.* ***(b)*** *The time-frequency plot of the degree of connectivity, which was averaged over all the sensors in this cluster. The left and right panels show the degree of connectivity of the phrase condition and sentence condition, respectively.* ***(c)*** *The matrix representation of the difference in phase connectivity between phrases and sentences. The figure was drawn by using the averaged connectivity matrix of the phrases minus that of the sentences.* ***(d)*** *All the sensors in this cluster were used as the seed sensors to plot the topographical representation of the phase connectivity. The upper and lower panels show the phase connectivity of the phrases and sentences, respectively.*

Phase-amplitude coupling (PAC) as a generalized neural mechanism for speech perception

To assess whether PAC distinguished phrases from sentences, we calculated the PAC value at each phase-amplitude bin for each condition, and then transformed it into the PAC-Z (for details see section 4.2, Methods). The grand average of the PAC-Z (average across sensors, conditions and participants) showed a strong activation over a region from 4 to 10 Hz for the frequency of phase and from 15 to 40 Hz for the frequency of amplitude. We therefore used the average PAC-Z value in this region of interest (ROI) for sensor clustering. For each participant, we first selected eight sensors that had the highest PAC-Z (conditions averaged) at each hemisphere. Averaging over sensors was conducted separately for the two conditions (phrase and sentence) and two hemispheres (see ***Figure 6a***). The Bonferroni correction was performed to address the multiple comparison problem. This resulted in a z-score of 3.73 for statistical significance (p=0.05; the z-score corresponded to the p-value of 0.05 ÷ 11 (the number of phase bins) * 12 (the number of amplitude bins) * 4 (the number of conditions)). From the results, we can see that there was a strong low-frequency phase response (4 to 10 Hz) entrained to high frequency amplitude (15 to 40 Hz). The results indicate that the PAC was introduced when participants listened to the speech stimuli.

Figure 6b shows how the sensors were selected. The larger the red circle, the more often the sensor was selected across participants. **Figure 6c** shows the topographical representation of the PAC-Z. The results indicate that when the participants listened to the speech stimuli, PAC was introduced symmetrically at both hemispheres over the central area. This could be evidence for the existence of PAC when speech stimuli are being processed. However, both the parametric and non-parametric statistical analyses failed to show a significant difference in the PAC-Z of phrases vs. sentences, which means we do not have evidence to show that the PAC was related to syntactic information processing. Therefore, our results suggest that PAC could be a generalized neural mechanism for speech perception, rather than a mechanism specifically recruited during the processing of higher-level linguistic structures.

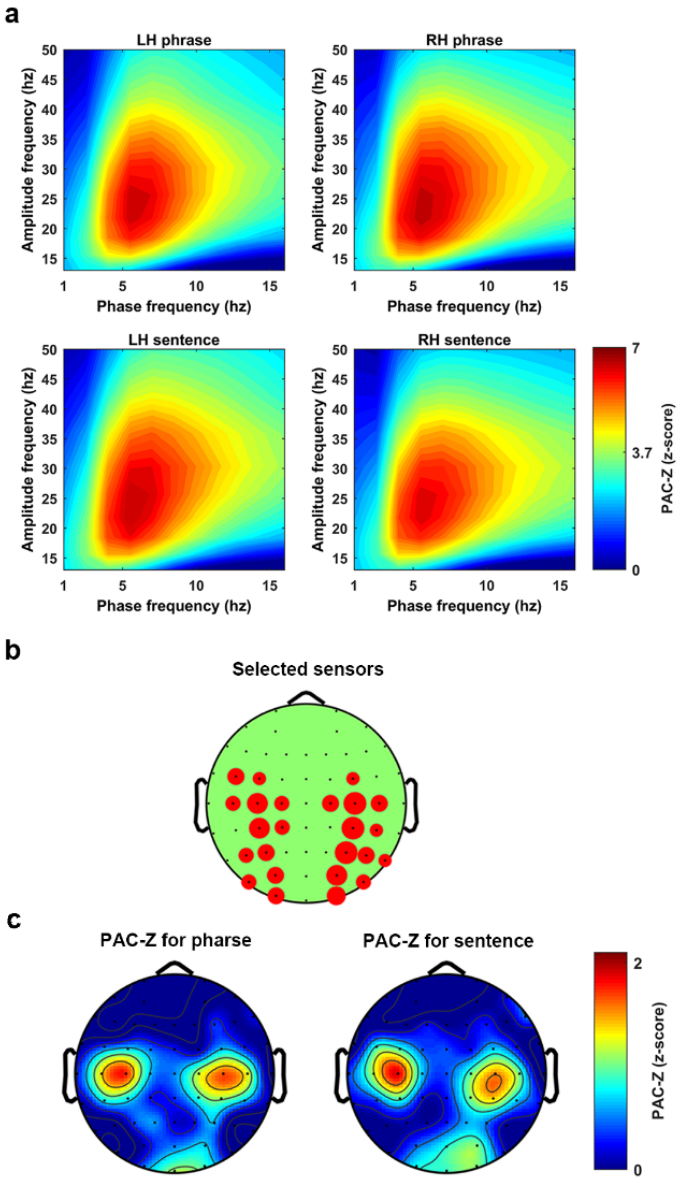

Figure 6. Phase-amplitude coupling as a general mechanism for speech perception. *The figure shows PAC transformed into a z-score, PAC-Z.* ***(a)*** *The PAC-Z for the phrases and the sentences at each hemisphere. Each figure was created by averaging the eight sensors which showed the biggest PAC-Z over the ROI. A z-score transformation with Bonferroni correction was conducted to test the significance, which led to a threshold of 3.73, corresponding to p=0.05.* ***(b)*** *How sensors were selected at each hemisphere. The bigger the red circle, the more times the sensor was selected.* ***(c)*** *The topographical distribution of the PAC-Z, which indicates that PAC was largely localized at the bilateral central areas.*

4.4 Discussion

In this chapter, we reported on an investigation into how the temporal dimension of neural oscillations distinguishes the linguistic structure of phrases from that of sentences. Using a series of analytical techniques, we gave a comprehensive description of the dimensions of neural readouts that were sensitive to the syntactic structure discrimination. We found that phrases and sentences have different effects on phase coherence; i.e., sentences show more phase coherence compared to phrases. In addition, we demonstrated that while phrases and sentences recruit similar functional networks that are constructed by temporal synchronization, the engagement of those networks is scaled according to the syntactic information of the linguistic structures. The connectivity pattern suggests that phrases and sentences have different impacts on the distribution and intensity of the neural networks involved in speech comprehension. Furthermore, we found that phase-amplitude coupling between theta and gamma, which has been implicated in speech processing, is not sensitive to syntactic structure differences in speech. In the following sections we give more detail about our findings and discuss potential interpretations of them.

Consistent with previous studies that showed the role of low-frequency phase coherence in the neural representation of speech (Doelling et al., 2014; Luo & Poeppel, 2007; Peelle & Davis, 2012; Peelle, Gross, & Davis, 2013), our analysis indicated that the degree of low-frequency phase coherence differed between phrases and sentences. In addition, at the selected range of interest, the statistical comparison of the phase coherence between phrases and sentences indicated a discrimination effect, which corresponded to a cluster that was expanded from 450 ms to 900 ms at a low frequency range (~2 to ~8 Hz) and was most pronounced over the central sensors. Our results reinforce the role of low frequency synchronization in speech representation, and more importantly, show the engagement of low-frequency neural oscillations in syntactic structure discrimination. As we have matched the physical intensity in both the temporal and spectral dimensions between the phrases and the sentences, the effect of the phase coherence separating the two conditions should not reflect acoustic-level differences. Instead, as the effect occurred during the listening stage, we consider that it reflects the online extraction of critical information, such as a verb in a

sentence, for further syntactic representation. In this view, our results are consistent with the notion of 'phase sets' in computational models of structured representations that exploit oscillatory dynamics. Phase sets are representational groupings that are formed by treating distributed patterns of activation as a set when units are in (or out) of phase with one another across the network (Doumas & Martin, 2018; Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2020; Martin & Doumas, 2020).

Phase connectivity was also a robust neural readout for discriminating between phrases and sentences. In the predefined time and frequency range of interest, the statistical comparison indicated a difference corresponding to a cluster approximately 800 to 1600 ms after the audio offset, occurring at the very low frequency range (< ~2 Hz) that was most pronounced over the right posterior region. Phrases and sentences thus differentially impact the temporal synchronization of neural responses.

Several aspects of the results are noteworthy. First of all, the relatively late latency suggests that the effect on temporal synchronization occurs after the initial presentation the speech stimulus. In our experiment, participants were randomly presented with a prompt for one of three possible tasks (color discrimination, object discrimination, and structure discrimination), which asked them to identify either 'semantic' information (object or color) or 'syntactic' information (whether the stimulus was a phrase or sentence) from the speech stimulus. Because of the random order of the task trials, participants had to pay close attention to the stimuli and continue to represent each stimulus after hearing it, namely until they received the task prompt. The tasks also ensured that participants could not select only one dimension of the stimulus for processing. Similarly, because we used an object and a color task, participants had to distribute their attention evenly across the adjectives and nouns, mitigating word-order differences between structures. In light of these controls and task demands, we consider it unlikely that the observed phase connectivity effects reflect mere differences in attention to phrases or sentences. Rather, we attribute the observed effects to differences in syntactic structure representation.

Secondly, the low frequency range (< 2 Hz) of the observed effect is consistent with previous research (Brennan & Martin, 2020; Ding et al., 2016; Kaufeld et al.,

2020; Keitel, Gross, & Kayser, 2018; Meyer et al., 2017). In Ding et al. (2016), the cortical response was modulated by the timing of the occurrence of linguistic structure; low-frequency neural responses (1-2 Hz) were found to track the highest-level linguistic structures (phrases and sentences) in their stimuli. Here we extended their work to ask whether the 1 Hz response could be decomposed to reflect separate syntactic structures (phrases vs. sentences), and we identified the role of the phase in discriminating between these structures. In our study, all speech stimuli lasted one second, and except for the presence of syntactic structure, the stimuli were normalized to be highly similar. Our pattern of results therefore suggests that functional connectivity, as reflected in the temporal synchronization of the induced neural response, distinguishes between phrases and sentences.

Lastly, phrases and sentences differed most strongly over the right posterior region of the brain, which is broadly consistent with previous research on speech comprehension. Functional magnetic resonance imaging (fMRI) studies implicate the posterior right hemisphere in processing syntactic structure (de Bode et al., 2015; Grodzinsky, 2000; Grodzinsky & Friederici, 2006; Maess et al., 2001). Neurophysiological research also suggests the involvement of the right hemisphere in the extraction of slow timescale information (Abrams et al., 2008; Giraud et al., 2007; Morillon et al., 2012; Poeppel, 2003). In addition, the P600, a positive ERP component often associated with syntactic processing, has a robust right-posterior topographical dominance (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Patel et al., 1998). In light of the existing literature, therefore, the right posterior distribution of the phase connectivity effects is consistent with the processing of syntactic structures, although we refrain from claims about underlying neural sources based on our EEG data.

We observed PAC during speech comprehension, as the low frequency phase (~ 4 to 10 Hz) strongly entrained with high frequency amplitude (~ 15 to 40 Hz). This effect appeared largely over the bilateral central area. The bilateral central topographical distribution has been repeatedly shown to reflect sensory-motor integration (Babiloni et al., 2011; Klimesch, Sauseng, & Hanslmayr, 2007; Neuper, Wörtz, & Pfurtscheller, 2006; Pfurtscheller et al., 2006; Pfurtscheller et al., 1998; Schlögl et al., 2005; Suffczynski et al., 2001), which is consistent with the proposal from Giraud and Poeppel (2012) that PAC reflects an early step in speech encoding

involving sensory-motor alignment between the auditory and articulatory systems. Crucially, however, this effect did not distinguish phrases from sentences. Despite this null result, the pattern is compatible with the generalized model for speech perception (Giraud & Poeppel, 2012). This early step is presumably similar for phrases and sentences, and perhaps for any type of structure above the syllable level.

In this chapter, we mainly focused on investigating the role of the temporal properties of neural oscillations in representing syntactic structure discrimination (phrases vs. sentences). Our results indicated a strong involvement of low-frequency phase coherence and phase connectivity in syntactic representation. In addition, consistent with Giraud and Poeppel (2012), we showed that PAC was present when participants listened to speech, although it did not reach the level of syntactic structure discrimination. Our investigations provided a comprehensive picture on how the phase-related measures reflected syntactic structure discrimination. However, to draw a full picture, we need to explore the role of the intensity of the neural oscillations in syntactic structure representation and model how the acoustic features are encoded when differences in syntactic structure are being represented.

Appendix 1: All the phrase-sentence pairs used in the experiment

1. de blauwe bal (‘the blue ball’) / de bal is blauw (‘the ball is blue’)
2. de blauwe knoop (‘the blue button’) / de knoop is blauw (‘the button is blue’)
3. de blauwe sok (‘the blue sock’) / de sok is blauw (‘the sock is blue’)
4. de blauwe strik (‘the blue bow’) / de strik is blauw (‘the bow is blue’)
5. de blauwe stoel (‘the blue chair’) / de stoel is blauw (‘the chair is blue’)
6. de blauwe pijl (‘the blue arrow’) / de pijl is blauw (‘the arrow is blue’)
7. de blauwe bel (‘the blue bell’) / de bel is blauw (‘the bell is blue’)
8. de blauwe helm (‘the blue helmet’) / de helm is blauw (‘the helmet is blue’)
9. de blauwe boot (‘the blue boat’) / de boot is blauw (‘the boat is blue’)

10. de blauwe mok (‘the blue mug’) / de mok is blauw (‘the mug is blue’)
11. de groene tas (‘the green purse’) / de tas is groen (‘the purse is green’)
12. de groene laars (‘the green boot’) / de laars is groen (‘the boot is green’)
13. de groene bijl (‘the green ax’) / de bijl is groen (‘the ax is green’)
14. de groene schoen (‘the green shoe’) / de schoen is groen (‘the shoe is green’)
15. de groene ster (‘the green star’) / de ster is groen (‘the star is green’)
16. de groene kom (‘the green bowl’) / de kom is groen (‘the bowl is green’)
17. de groene deur (‘the green door’) / de deur is groen (‘the door is green’)
18. de groene kam (‘the green comb’) / de kam is groen (‘the comb is green’)
19. de groene pen (‘the green pen’) / de pen is groen (‘the pen is green’)
20. de groene brug (‘the green bridge’) / de brug is groen (‘the bridge is green’)
21. de rode trui (‘the red sweater’) / de trui is rood (‘the sweater is red’)
22. de rode doos (‘the red box’) / de doos is rood (‘the box is red’)
23. de rode vaas (‘the red vase’) / de vaas is rood (‘the vase is red’)
24. de rode vis (‘the red fish’) / de vis is rood (‘the fish is red’)
25. de rode bank (‘the red couch’) / de bank is rood (‘the couch is red’)
26. de rode mier (‘the red ant’) / de mier is rood (‘the ant is red’)
27. de rode bus (‘the red bus’) / de bus is rood (‘the bus is red’)
28. de rode tent (‘the red tent’) / de tent is rood (‘the tent is red’)
29. de rode bloem (‘the red flower’) / de bloem is rood (‘the flower is red’)
30. de rode draak (‘the red dragon’) / de draak is rood (‘the dragon is red’)
31. de gele hoed (‘the yellow hat’) / de hoed is geel (‘the hat is yellow’)
32. de gele riem (‘the yellow belt’) / de riem is geel (‘the belt is yellow’)
33. de gele sjaal (‘the yellow scarf’) / de sjaal is geel (‘the scarf is yellow’)
34. de gele broek (‘the yellow pants’) / de broek is geel (‘the pants is yellow’)
35. de gele kan (‘the yellow pitcher’) / de kan is geel (‘the pitcher is yellow’)
36. de gele lamp (‘the yellow lamp’) / de lamp is geel (‘the lamp is yellow’)

37. de gele eend (‘the yellow duck’) / de eend is geel (‘the duck is yellow’)
38. de gele jas (‘the yellow jacket’) / de jas is geel (‘the jacket is yellow’)
39. de gele fles (‘the yellow bottle’) / de fles is geel (‘the bottle is yellow’)
40. de gele fiets (‘the yellow bicycle’) / de fiets is geel (‘the bicycle is yellow’)
41. de paarse vlag (‘the purple flag’) / de vlag is paars (‘the flag is purple’)
42. de paarse tas (‘the purple bag’) / de tas is paars (‘the bag is purple’)
43. de paarse mand (‘the purple basket’) / de mand is paars (‘the basket is purple’)
44. de paarse jurk (‘the purple dress’) / de jurk is paars (‘the dress is purple’)
45. de paarse bril (‘the purple glasses’) / de bril is paars (‘the glasses is purple’)
46. de paarse kerk (‘the purple church’) / de kerk is paars (‘the church is purple’)
47. de paarse pop (‘the purple doll’) / de pop is paars (‘the doll is purple’)
48. de paarse muur (‘the purple wall’) / de muur is paars (‘the wall is purple’)
49. de paarse veer (‘the purple feather’) / de veer is paars (‘the feather is purple’)
50. de paarse kast (‘the purple closet’) / de kast is paars (‘the closet is purple’)

5 | Representing syntactic structure discrimination in the intensity of neural oscillations[2]

Abstract

Using the same dataset as in Chapter 4, we extended our investigation into the neural representation of syntactic structure discrimination. Unlike the analysis in the last chapter, where temporal synchronization was heavily weighted, this chapter explores how syntactic structure discrimination is reflected in the intensity of neural oscillations, and how acoustic features are differently encoded to separate phrases from sentences. We found that syntactic structure discrimination was well captured in both the intensity and degree of power connectivity of induced neural responses in the alpha band (~ 7.5 to 13.5 Hz). In addition, our modeling suggested that there were different encoding states in both the temporal and spectral dimensions as a function of the quantities and types of linguistic structures perceived, over and above the acoustically driven neural response. Complementing the findings in the last chapter, our results in this chapter provide new insights into the neural readout for syntactic structure discrimination.

[2] Adapted from Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. PLoS Biology, 20(7), e3001713.

5.1 Introduction

In the last chapter, we explored how the temporal synchronization of the neural oscillations separates the syntactically different linguistic structures. In our phase-related analysis, we found that the low-frequency (< 8 Hz) phase measures contributed to separating the phrases from the sentences even when the two types of stimuli had highly similar temporal-spectral properties. However, the intensity of the neural oscillations might also be involved in discriminating the two types of speech stimuli (phrases vs. sentences). Building upon the work in Chapter 4 and using the same dataset, we now address the following three questions. The first is how the intensity (e.g. power) of the induced neural response contributes to the differentiation between the phrases and sentences. The second is whether the structural differences between the phrases and sentences could be reflected by the intensity connectivity of neural oscillations. The third and final question is how the acoustic features of the phrases and sentences are represented in the brain to reflect the discrimination of syntactic structures.

Research into speech perception and comprehension has shown evidence that the intensity of neural oscillations from almost all canonical bands were correlated. As for neural oscillations at slow (< ~8 Hz) and fast (> ~25 Hz) timescales, researchers are largely in agreement regarding their role in speech processing. For instance, the intensity of low-frequency (< 8 Hz) neural oscillations has proved to be related to the representation of hierarchical linguistic structures (Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018) and speech intelligibility (Brennan & Martin, 2020; Doelling et al., 2014; Luo & Poeppel, 2007; Peelle & Davis, 2012; Peelle, Gross, & Davis, 2013). The high frequency (> 25 Hz) amplitude has also revealed its engagement in encoding the phonemic-level units (Gross et al., 2013; Kerlin, Shahin, & Miller, 2010; Morillon et al., 2012; Palva et al., 2002; Peña & Melloni, 2012; Shahin, Picton, & Miller, 2009). However, whether the induced power of alpha band (~ 8 to ~13 Hz) oscillations is a reflection of auditory processing or language processing is still under debate. Studies have shown that alpha band oscillations correlate with verbal working memory (Obleser et al., 2012; Wilsch & Obleser, 2016) and auditory attention (Strauß, Wöstmann, & Obleser, 2014; Wöstmann et al., 2016; Wöstmann et al., 2015; Wöstmann, Lim, & Obleser, 2017). A neural physiology model of speech perception also considered the induced

neural response at alpha band as an endogenous (top-down) gating control (Ghitza, Giraud, & Poeppel, 2013; Giraud & Poeppel, 2012). However, studies have further shown that alpha band oscillations reflect speech intelligibility (Becker et al., 2013; Dimitrijevic et al., 2017; Obleser & Weisz, 2012).

Due to these inconsistencies, we wanted to check how the intensity of neural oscillations would reflect the discrimination between the two types of syntactic structures (phrases vs. sentences). If the induced neural activities reflect the syntactic differences between the phrases and sentences, which frequencies are involved in this separation? Could we find evidence that alpha band oscillations are engaged in representing the differences between phrases and sentences?

Successful speech comprehension involves inter-communication among different brain regions, and one impacted neural physiological model (Hickok & Poeppel, 2000, 2004, 2007) emphasized the critical role of functional connectivity via neural oscillations. Unlike the phase connectivity, which indicates the temporal synchronization across brain regions, the neural networks that are generated by intensity describe an energy-organized network reflecting the encoding and representing of the information extracted from the sensory input. As far as we know, studies of high-level linguistic processing, e.g. syntactic structure discrimination, rarely use this approach.

As this is an exploratory study, we wanted to know whether the functional connectivity via the intensity of neural oscillations would reflect the separation between the phrases and the sentences. If so, which frequency bands would be involved in discriminating different types of syntactic structures?

Analyzing the neural activities that correspond to participants doing a cognitive task is sort of indirect in terms of showing how the stimuli were encoded. A more straightforward approach is to model how the pertinent features of the stimuli are being represented (encoded) in the brain. Previous research using the spectral-temporal response function (STRF) has shown that low-frequency neural activities reflect the encoding of acoustic features in speech (Ding & Simon, 2012a, 2012b, 2013b), and phonemic-level information can be reflected in the low-frequency neural response entrained to speech (Di Liberto, O'Sullivan, & Lalor, 2015). Inspired by studies using the STRF, we aimed to find out whether the syntactic structure discrimination would be reflected in the encoding of acoustic

features, which is especially interesting as the physical properties of our speech stimuli are fully matched. We expected to see that a different encoding regime would be employed in order to represent the two different syntactic structures.

In sum, using the same dataset as Chapter 4, in this chapter we convey additional information on how the intensity of neural oscillations works when discriminating between phrases and sentences, as well as whether the acoustic features are encoded differently to represent the two syntactically different structures.

5.2 Methods

Note that the experiments presented in this chapter used the same dataset as those in Chapter 4; therefore, some parts are reiterated here for easy reference.

Participants

Fifteen Dutch native speakers (8 females and 7 males), aged 22 to 35, participated in the study. All participants were undergraduate or graduate students, and were right-handed. They reported no history of hearing impairment or neurological disorder. The experimental procedure was approved by the Ethics Committee of the Social Sciences Department at Radboud University. Written informed consent was obtained from each participant before the experiment, and they were paid for their participation.

Stimuli

Fifty line-drawings of common objects were selected from a standardized corpus (Snodgrass & Vanderwart, 1980). The Dutch names of all the objects were mono-syllabic and had non-neuter gender. The objects appeared as colored line-drawings on a grey background. We presented each line-drawing in five colors: blue (*blauw*), red (*rood*) yellow (*geel*), green (*groen*), and purple (*paars*). In total, this yielded 250 pictures. The line-drawings were sized to fit into a virtual frame of 4 by 4 cm, corresponding to a 2.29° of visual angle for the participants.

We then selected 50 figures (50 different objects in five colors) to create the speech stimuli. For each selected line-drawing, a four-syllable phrase-sentence pair was created, e.g. *de rode vaas* ('the red vase') and *de vaas is rood* ('the vase is red').

This means that in total, we had 100 speech stimuli (50 phrases and 50 sentences). All stimuli were synthesized by an online synthesizer (www.neospeech.com), using a Dutch male voice, Guus. All speech stimuli were 733 to 1125 ms in duration (mean = 839 ms, SD = 65 ms). To normalize the synthesized auditory stimuli, they were first resampled to 44.1 kHz. Then any speech stimuli that were longer than 1000 ms were cut at both sides to shorting them to less than 1000 ms without missing any meaningful dynamics. The 10% at both ends of each stimulus was smoothed by a linear ramp (a cosine wave) to remove the abrupt sound burst. All stimuli were fitted into 1000 ms with symmetric zero paddings. Finally, to normalize the intensity of the stimuli, the root-mean-square value of each stimulus was normalized to -16 dB.

Experimental procedure

Each trial started with a fixation cross being visible at the center of the screen (for 500 ms in duration). Participants were asked to look at the screen. Immediately after the fixation cross had disappeared, the participants heard a 1000 ms spoken stimulus, either a phrase or a sentence, followed by a three-second silence; then the participants were asked to perform one of three discrimination tasks, indicated to them by an index number (1, 2 or 3 showing at the center of the screen for 500 ms). All responses were recorded via a parallel port response box, in which the two buttons were labeled as 'phrase/match' and 'sentence/mismatch'. Each response was followed by a silent interval of 3 to 5.2 seconds.

The data collection was broken down into five blocks, with 48 trials in each block. Before the core data collection, several practice trials were conducted for each participant to make sure they understood the task. Trials in each block were fully matched in across linguistic structure (phrase or sentence) and task type (1, 2 or 3). For instance, half the spoken stimuli were phrases and half were sentences (24 of each structure), and six combinations (eight trials for each type) were evenly distributed in each block (eight trials times two linguistic structures times three task types). The order of the trials was pseudo-random throughout the whole experiment. The behavioral results indicated that the task was relatively easy and no differences were found between the phrases and the sentences. For all tasks combined, the accuracy rates for phrases and sentences were 97.9 ± 3% and 97.3 ± 3% ($p = 0.30$), respectively.

After the main experiment, a localizer task was performed, in which a 'beep' tone (1 kHz, 50 ms in duration) was played 100 times (jitter 2 to 3 seconds) for each participant, in order to localize the canonical auditory response (N1-P2 complex).

EEG recording

EEG data was recorded using a 64-channel active sensor system from Brain Products (GmbH) in a sound-dampened, electrically shielded room. Signals were digitized online at 1000 Hz, with high-pass and low-pass at 0.01 Hz and 249 Hz, respectively. Two electrodes, AFz and FCz, were used as ground and reference. All electrodes were placed on the scalp based on the international 10-20 system and the impedance of each one was kept below 5 kΩ. The experimental procedure was controlled by MATLAB 2019a (The MathWorks, Natick, MA) with Psychtoolbox-3 (Brainard, 1997). Auditory stimuli were played at 65 dB SPL and delivered through air-tube earplugs (Etymotic ER-3C, Etymotic Research, Inc.). Event markers were sent via a parallel port for tagging the onset of the events under investigation (i.e., speech onset, task index onset, etc.).

EEG data preprocessing

The EEG data preprocessing was conducted via MATLAB using the EEGLAB toolbox (Delorme & Makeig, 2004) and customized scripts. The data were first down-sampled to 256Hz then high-pass filtered at 0.5 Hz (finite impulse response filter, FIR; zero-phase lag). The raw data were first cleaned by the time-sliding PCA (Chang et al., 2018; Kothe & Jung, 2016). Then all detected bad channels were interpolated with spherical interpolation. After transferring the data to average reference, the online reference FCz was recovered and the line noise, 50 Hz and its harmonics, was removed.

Following the above steps, epochs of two seconds preceding and nine seconds following the auditory stimulus onset were extracted. The deletion of bad trials and removal of artifacts were conducted in two steps. First, independent component analysis (ICA) was used for decomposing the data into the component space (number of components equals data rank). Then for each independent component, we used the short-time Fourier transform to convert each trial into the power spectrum, in which we extracted a value that was calculated by the power summation between 15 and 50 Hz. Then all the extracted values in each component

formed a distribution. From this distribution, we transformed all the extracted values to z-scores, and the epochs with values outside the range of plus or minus three standard deviation were deleted. Second, ICA was conducted again on the trial-rejected data for eye-related artifact removal and muscle activity elimination. Artifact components were identified and removed using an automatic classification algorithm (Winkler, Haufe, & Tangermann, 2011). All the preprocessing steps resulted in the removal of, on average, 7 components (range 4 to 11) and 22 trials (including incorrect trials and trials with excessively slow responses, range 10 to 30, 4% to 12.5%) per participant. Finally, volume conduction was attenuated by applying surface Laplacian (Cohen, 2014; Srinivasan et al., 2007; Winter et al., 2007).

EEG data analysis

Time frequency decomposition

To perform time-frequency decomposition, the single-trial time series were convolved with a family of complex wavelets (1 to 50 Hz in 70 logarithmically spaced steps). Temporal and spectral resolution were optimized by changing the cycle from 3 to 30 in logarithmic steps. The induced response (power) was then extracted from the analytical output at each time-frequency bin by taking the summation of the squared wavelet coefficients. Decibel transformation was performed at each frequency, in which the average power at the duration from 800 to 200 ms before the audio onset was used as the baseline.

Power connectivity

After time-frequency decomposition, the induced power at each channel-time-frequency-trial bin was extracted. For each condition, the power connectivity between each sensor pair at each time-frequency bin was calculated as the rank correlation between the power response of all trials in one sensor and the power response of all trials in the other sensor. The power connectivity calculation resulted in an all-sensors-to-all-sensors (65*65 in our data) representation at each time-frequency bin for each condition.

To compare the power connectivity levels between the phrases and sentences in the time-frequency space, a statistical threshold method was used. More specifically, at each time-frequency bin, we formed a distribution by pooling

together all the connectivity values from both conditions, and then defined the threshold as the value at half the standard deviation above the median. We then binarized the connectivity matrix at each bin for each condition by applying the corresponding threshold. The connectivity level at each time-frequency bin was represented as the total number of connectivity values that were above this threshold. Finally, we transferred the connectivity level at each time-frequency bin as the percentage change relative to the connectivity level of the baseline, which was calculated as the average connectivity level in the duration from 800 to 200 ms before the audio onset.

Spectral-temporal response function (STRF)

The STRF is a linear kernel which convolves with the specified features of the speech signal to estimate the neural response in time. It can be interpreted as a linear filter which transforms the stimulus feature into the neural response (Crosse et al., 2016; Di Liberto, O'Sullivan, & Lalor, 2015).

In our study, the stimulus features were defined as the narrow-band temporal envelopes, which were obtained by first filtering the speech stimulus into 16 logarithmically spaced frequency bands between 0.05 and 8 kHz to simulate the frequency decomposition by the brain (Greenwood, 1990). These were then extracted by the Hilbert transform.

To construct the stimulus-response pairs, we first applied a linear ramp to both sides of each neural response corresponding to a trial (10% at each side) to attenuate the abrupt onset and offset. Then, we matched each one-second neural response with the corresponding stimulus features.

Since each trial was one second long, to optimize the estimation of the STRF, a randomization procedure was applied to create a new data structure. We first randomly selected 80% of all unique speech stimuli, and then the stimulus-response pairs that corresponded to the selected speech stimuli were extracted as the seed data to construct the dataset for performing the cross validation. We constructed 35 stimulus-response pairs lasting 10 seconds each, which were all concatenations of the 10 randomly selected one-second stimulus-response pairs (bootstrapping).

The STRF was estimated using the ridge regression with the leave-one-out cross validation. Since the ridge regression weights the diagonal elements of the

covariance matrix of the stimulus features with a lambda parameter (Crosse et al., 2016; Tikhonov & Arsenin, 1977), we predefined the range of lambda as 10 values from 6 to 100 in linear steps before the cross validation. We used the extracted dataset (35 stimulus-response pairs lasting 10 seconds each) to conduct the cross validation for optimizing the STRF. The Pearson correlation between each real neural response and each predicted response was calculated. The average of all the coefficients of the Pearson correlation (across all sensors and all trials) was defined as the performance of the STRF. The model with the lambda parameter which gave the best performance was used as the optimized STRF.

Since previous research has shown that a slow (low-frequency) neural response reliably reflects the neural representation of the acoustic features in speech (Ding & Simon, 2012a, 2012b), we initially checked whether the STRFs for different frequency bands faithfully reflect the encoding of the acoustic features. To do so, we first filtered the neural responses corresponding to each trial into five canonical frequency bands, which were delta (< 4 Hz), theta (4 to 7 Hz), alpha (8 to 13 Hz), beta (14 to 30 Hz), and low-gamma (31 to 50 Hz). Then, the STRF for each condition at each frequency band was estimated using the procedure mentioned above. In order to check the performance of each estimated STRF, the stimulus-response pairs (the unseen pairs for each STRF) that corresponded to the remaining 20% of the speech stimuli were extracted as the seed data for constructing the testing dataset. We extracted five stimulus-response pairs of four seconds each from the testing data for each STRF. All of them were concatenations of four randomly selected pairs lasting 1 seconds each.

The real performance of each STRF was calculated by using the frequency- and condition-matched stimulus-response pairs. The random performance was calculated 1000 times by using randomly selected stimulus-response pairs.

For fitting the low frequency STRF (< 9 Hz), the same procedure was followed. The performance of the averaged STRF (averaged across participants) in each condition was computed using the average of the Pearson correlations for the real neural responses and the predicted responses across sensors.

The temporal response function (TRF) and the spectral response function (SRF) for each participant in each condition were extracted by averaging the STRF over the frequencies from 0.1 kHz to 800 kHz and by averaging the STRF over the

time from 0 to 400 ms, respectively. All the calculations in this section were conducted using customized scripts, the scripts from the EEGLAB toolbox (Delorme & Makeig, 2004) and the Multivariate Temporal Response Function Toolbox (Crosse et al., 2016).

Statistical analysis

In addition to using parametric statistical methods to check whether the difference between phrases and sentences was significant, a cluster-based non-parametric permutation test was applied. This method deals with the multiple-comparisons problem and at the same time takes the data's dependency (temporal, spatial and spectral adjacency) into account. For all types of analysis that followed this inference method, the subject-level data were initially averaged across trials and for each single sample, i.e. a time-frequency-channel point, a dependent t-test was performed. We selected all samples for which the t-value exceeded an a priori threshold, $p<0.05$, and these were subsequently clustered on the basis of spatial and temporal-spectral adjacency. The sum of the t-values within a cluster was used as a cluster-level statistic. The cluster with the maximum sum was subsequently used as test statistic. By randomizing the data across the two conditions and recalculating the test statistic 1000 times, we obtained a reference distribution of the maximum cluster t-values. This distribution was used to evaluate the statistics of the actual data. This method was carried out using the FieldTrip toolbox (Maris & Oostenveld, 2007; Oostenveld et al., 2011).

Acoustic normalization and analyses

See the corresponding section in ***Chapter 4***.

5.3 Results

Alpha band inhibition reflects discrimination between phrases and sentences

To query whether neural oscillations at the alpha band reflect the processing of syntactic structure, we calculated the induced power. The grand average (over all participants and all conditions) of the induced power showed a strong inhibition at the alpha band (~7.5 to 13.5 Hz). Therefore, we checked whether this alpha band

inhibition could separate the two types of linguistic structures. A statistical analysis was conducted using the non-parametric cluster-based permutation test (1000 times) over the frequencies of the alpha band with a 1000 ms time window that started at the audio onset. The results indicated that the alpha band inhibition was stronger for the phrase condition than the sentence condition ($p < 0.01$ **, two-tailed). In the selected time and frequency range, this effect corresponded to a cluster that lasted from ~350 to ~1000 ms after the audio onset and was largely localized at the left hemisphere, though the right frontal-central sensors were also involved during the temporal evolution of this cluster. ***Figure 1a*** shows the temporal evolution of this cluster in steps of 50 ms using the induced power of the phrase condition minus the induced power of the sentence condition. ***Figure 1b*** shows the time-frequency plot of the induced power using the average of all the sensors in this cluster. The upper and lower panels show the phrase and sentence condition, respectively. From these figures, we can see that the alpha band inhibition was stronger for the phrase condition than the sentence condition. These results show that the processing of phrases and sentences is reflected in the intensity of the induced neural response in the alpha band.

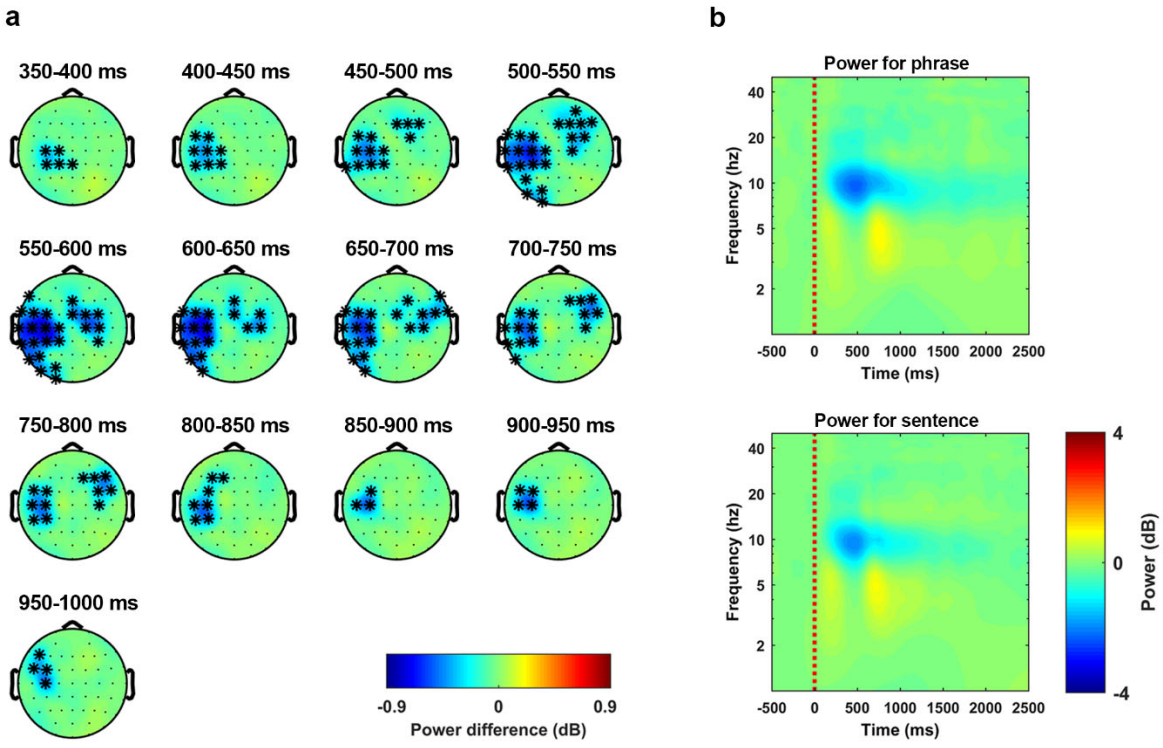

Figure 1. Alpha band inhibition suggests a separation between phrases and sentences. *Statistical analysis on the induced neural response was conducted by using*

*the non-parametric cluster-based permutation test (1000 times) on a time window of 1000 ms, which started at the audio onset and over the frequencies from 7.5 to 13.5 Hz. The results indicated that the power was significantly higher for sentences than phrases (p < 0.01 **, two-tailed).* ***(a)*** *The temporal evolution of the cluster that corresponds to this separation effect. The activity was drawn by using the induced power of the phrase condition minus that of the sentence condition. The topographies were plotted in steps of 50 ms.* ***(b)*** *The induced power averaged over all the sensors in this cluster. The upper and lower panels show the induced power of the phrases and sentences, respectively.*

The power connectivity in the alpha band indicates a network-level separation between phrases and sentences

We calculated power connectivity in each sensor-pair at each time-frequency bin using a rank correlation (for details see section 5.2, Methods). The grand average of the power connectivity level (over all participants and all conditions) showed a strong inhibition at the alpha band from 100 to 2200 ms after the audio onset. Because it revealed a strong power connectivity inhibition, this region was defined as the ROI. For each participant, we selected eight sensors at each hemisphere that indicated the greatest inhibition on the condition-averaged power connectivity. This was followed by averaging across all the selected sensors, which resulted in four conditions for each participant (left-phrase, left-sentence, right-phrase, and right-sentence).

Figure 2a shows the degree of power connectivity, which was averaged over all participants for each condition. To check whether this metric could separate the phrases and the sentences, a Stimulus-Type*Hemisphere two-way repeated-measure ANOVA was conducted. The comparison revealed that the main effect was derived from Stimulus-Type ($F(1, 14) = 5.28$, $p = 0.033$ *). A planned post-hoc comparison using paired sample t-tests on the main effect of the Stimulus-Type showed that the power connectivity inhibition was stronger for the phrases than the sentences ($t(29) = 2.82$, $p = 0.0085$ ***). **Figure 2b** shows the power connectivity degree for each extracted condition. **Figure 2c** illustrates what sensors were used. The larger the red circle, the more times the sensor was selected.

Since the degree of the power connectivity over the alpha band indicated a separation between the phrases and sentences, we also checked how this difference

was distributed in the sensor space. To do so, we extracted the binarized power connectivity representations (matrices) that are located in the ROI, and then averaging was performed for each condition across all connectivity matrices. **Figure 2d** shows the difference in the degree of power connectivity over the sensor space using the average of the binarized sentence connectivity matrix minus the average of the binarized phrase connectivity matrix. The results indicate that the inhibition of the power connectivity was stronger for phrases than for sentences. In other words, the overall level of the power connectivity was higher for sentences than phrases. **Figure 2e** is the topographical representation of this, which was plotted using the binarized power connectivity of the selected sensors. The upper and lower panels indicate the phrase condition and the sentence condition, respectively. From this figure, we can see that the difference was largely localized at the bilateral central area, and more strongly present in the left than the right hemisphere. These results reflect that the neural network which was organized by the intensity of the induced power at the alpha band was different for the two types of syntactic structure.

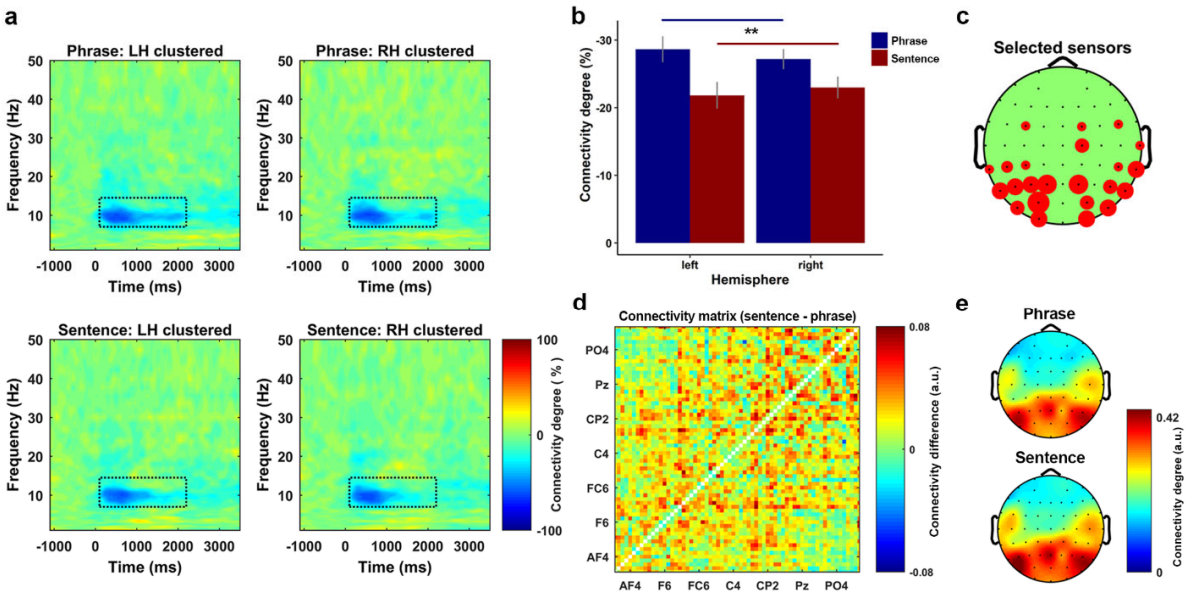

Figure 2. Power connectivity in the alpha band suggests a separation between phrases and sentences. (a) *The degree of power connectivity for all conditions. Each plot was clustered by the sensors at each hemisphere that showed the biggest inhibition of the power connectivity (as a grand average).* ***(b)*** *The results of a two-way repeated-measure ANOVA for the power connectivity on the factors of stimulus-type (phrase or sentence) and hemisphere (left or right). The results indicate a*

*significant main effect of Stimulus-Type, and a post-hoc comparison on the main effect indicated that the overall inhibition of the power connectivity was stronger for the phrases than the sentences (t(29) = 2.82, p = 0.0085 ***, two-sided).* ***(c)*** *The selection of sensors for the clustering. The bigger the red circle, the more times the sensor was selected across participants.* ***(d)*** *The connectivity differences between the phrases and sentences on all sensor pairs. The figure was drawn using the average of the binarized connectivity matrix of the sentence condition minus that of the phrase condition. The results indicate that the degree of connectivity over the sensor space for the sentence condition was higher than for the phrase condition.* ***(e)*** *Topographical plot of the binarized connectivity that was clustered by the sensors showing the most inhibition of power connectivity. The upper and lower panels show the phrase and sentence condition, respectively.*

Different encoding states for phrases versus sentences in both temporal and spectral dimensions

Previous research has shown that the low-frequency neural response reliably reflects the phase-locked encoding of the acoustic features of speech (Ding & Simon, 2012a, 2012b). Therefore, we initially tested whether the neural response from all canonical frequency bands could equally reflect the encoding of the acoustic features. To do so, we fitted the STRF for each condition at all frequency bands, which are delta (< 4 Hz), theta (4 to 7 Hz), alpha (8 to 13 Hz), beta (14 to 30 Hz), and low-gamma (31 to 50 Hz). Then we compared the real performance of the STRFs to their random performance (for details see section 5.2, Methods). **Figure 3a** shows the results of this comparison. The blue and red dots represent the real performance of the STRFs, and the error bar indicates one SEM on each side. The small gray dots represent the random performance (1000 times in each frequency band per condition). The upper border delineated by these gray dots represents the percentile of 97.5 for the random performance. The performance of the STRFs was above chance level only at the low frequency (delta and theta) bands, which is consistent with previous research (Ding & Simon, 2012a, 2012b). Our results verified that the low frequency STRF reliably reflected the relationship between the acoustic features of speech and the neural response at low frequencies.

Since only low-frequency neural responses robustly reflected the encoding of the speech stimuli, we fitted the STRF for both conditions using the neural response that was low-passed at 9 Hz. Leave-one-out cross validation was used to maximize the performance of the STRFs. **Figure 3b** shows the performance of the STRF for each condition. The light dots, blue for phrases and red for sentences, represent the model's performance on each testing trial. The solid dots represent the model's performance that was averaged over all trials, and the error bars represent one SEM on each side. A paired sample t-test was used to compare the performance between the phrase condition and the sentence condition. No evidence was found to indicate a difference in performance between these two conditions ($t(74) = 1.25$, $p = 0.21$). The results indicate that the STRFs fitted equally well for phrases and sentences. Thus, any difference in temporal-spectral features between the STRFs of phrases vs. sentences cannot be driven by the model's performance. **Figure 3c** shows the comparison between the real neural response and the response predicted by the model at a sample sensor, Cz. The upper and lower panels show the performance of the STRF for phrases ($r = 0.47$, $N=1024$, $p < 1e\text{-}5$ ***) and sentences ($r = 0.41$, $N=1024$, $p < 1e\text{-}5$ ***), respectively.

The grand average of the STRFs was negative from 0 to 400 ms in the time dimension and from 100 to 1000 Hz in the frequency dimension, and the sensor clustering of the STRF was conducted based on the average activation in this ROI. More concretely, we selected eight sensors at each hemisphere for each participant, which showed the strongest average magnitude (negative) in this region. **Figure 3d** shows the clustered STRFs that were averaged across all participants. **Figure 3e** depicts the sensors that were selected across the participants: the bigger the red circle, the more often the given sensor was selected.

To compare the differences in the kernel (STRF) for both the temporal and spectral dimensions, the temporal response function (TRF) and the spectral response function (SRF) were extracted separately for each condition. **Figure 3f** shows the TRFs that were averaged across all participants. The grand average of all TRFs showed two peaks at ~100 ms and ~300 ms. We therefore defined the first temporal window as 50 to 150 ms (center at 100 ms) and the second temporal window as 250 to 350 (center at 300 ms) to search for the magnitude and latency of these two peaks. The latency of each peak was defined as the time when it appeared, and the magnitude was defined as the average strength over a 5 ms

window on both sides around each peak. After extracting these measurements, a Stimulus-type*Peak-type*Hemisphere three-way repeated-measure ANOVA was conducted on both the magnitude and the latency.

For the magnitude of the TRF (**Figure 3g**), the statistical comparison showed that there was a significant main effect from the Stimulus-Type ($F(1, 14) = 13.58$, $P = 0.002$ ***) and a significant three-way interaction involving Stimulus-type*Peak-type*Hemisphere ($F(1, 14) = 15.25$, $P = 0.001$ ***).

The post-hoc comparison on the main effect of Stimulus-Type using paired sample t-tests showed that the magnitude for phrases was significantly stronger than the magnitude for sentences ($t(59) = 4.55$, $P < 2e\text{-}5$ ***). The results suggest that the instantaneous neural activity in response to phrases had a stronger phase-locked dependency on the acoustic features than in response to sentences.

To investigate the three-way interaction of Stimulus-Type*Peak-Type*Hemisphere, two-way repeated-measure ANOVAs with the Bonferroni correction were conducted on the factors of Hemisphere and Stimulus-Type at each level of the Peak-Type. The results indicated a main effect of Stimulus-Type at the first peak ($F(1, 14) = 8.19$, $p = 0.012$ *) and a two-way Hemisphere*Stimulus-Type interaction at the second peak ($F(1, 14) = 6.42$, $p = 0.023$ *).

At the first peak, we conducted a post-hoc comparison on the main effect of Stimulus-Type using paired sample t-tests, which showed that the magnitude of the phrase condition was higher than the magnitude of the sentence condition ($t(29) = 3.49$, $p = 0.001$ ***). The results indicate that the instantaneous neural activity was more strongly driven by the acoustic features that were presented ~100 ms ago when phrases than when sentences were presented.

For the two-way Hemisphere*Stimulus-Type interaction at the second peak, the paired sample t-tests with Bonferroni correction were conducted to compare the difference in the magnitude between phrases and sentences at each hemisphere. The results indicate that the magnitude at the second peak was stronger for phrases than sentences in the right hemisphere ($t(14) = 3.21$, $p = 0.006$ **), but not the left hemisphere ($t(14) = 0.86$, $p = 0.40$). The findings suggest that, at the right hemisphere, the instantaneous neural activity of the phrases was more strongly driven by the acoustic features that were present approximately 300 ms than it was under sentences.

For the latency of the TRF (**Figure 3h**), the comparison showed a main effect of the Peak-type ($F(1, 14) = 1e+3$, $p<1e-14$ ***) and a three-way interaction of Stimulus-Type*Peak-Type*Hemisphere ($F(1, 14) = 8.04$, $p = 0.013$ *). The post-hoc comparison for the main effect of the Peak-Type with paired sample t-tests showed, as expected, that the latency of the first peak was significantly shorter than the second one ($t(59) = 38.89$, $p < 2e-16$ ***). The result is unsurprising since regardless of the method for searching the time windows, the latency of the first one will always be shorter than the second.

To investigate the three-way Stimulus-type*Peak-type*Hemisphere interaction, two-way repeated-measure ANOVA with the Bonferroni correction were conducted on the factors of Hemisphere and Stimulus-Type for each level of the Peak-Type. The comparison suggested a two-way Hemisphere*Stimulus-Type interaction at the first peak ($F(1, 14) = 12.83$, $p = 0.002$ **). The post-hoc comparison on this two-way interaction using paired sample t-tests with the Bonferroni correction indicated that the latency at the first peak was significantly longer for sentences than for phrases at the right hemisphere ($t(14) = 3.55$, $p = 0.003$ **), but not the left ($t(14) = 0.58$, $p = 0.56$). The results suggest that, within the first temporal window (~50 to 150 ms), and only at the right hemisphere, the neural response to the sentences was predominantly driven by the acoustic features earlier in time than the response to the phrases.

Figure 3i shows the SRFs that were averaged across all participants. The grand average of the STRFs suggested that the activation of the kernel was most prominent in the frequency range from 0.1 to 0.8 kHz. To compare the differences in the encoding of acoustic features in the spectral dimension, we separated the SRF into three frequency bands: lower than 0.1 kHz; 0.1 to 0.8 kHz; and higher than 0.8 kHz. We then averaged the response in each extracted frequency band for each condition. The statistical comparison was conducted using a three-way repeated-measure ANOVA on the factors of Hemisphere, Stimulus-Type and Band-Type. The results (**Figure 3j**) indicated a main effect of Band-Type ($F(2, 28) = 119.67$, $p < 2e-14$ ***) and a two-way interaction of Band-Type*Stimulus-Type ($F(2, 28) = 27.61$, $p < 3e-7$ ***).

It was revealed by the post-hoc comparison on the main effect of Band-Type using paired sample t-tests with the Bonferroni correction that the magnitude of

the middle frequency band was stronger than that of the low frequency band ($t(59) = 17.9$, $p < 4e\text{-}25$ ***) and high frequency band ($t(59) = 18.7$, $p < 5e\text{-}26$ ***). The results indicate that the acoustic features from different frequency bands contributed differently to the evoked neural response. In other words, for both conditions, the neural response was predominantly driven by the encoding of the acoustic features from 0.1 to 0.8 kHz, which are considered as the spectral-temporal features at the range of the first formant (Catford, 1988; Jeans, 1968; Titze et al., 2015; Titze & Martin, 1998).

The post-hoc comparison using paired sample t-tests with the Bonferroni correction on the Band-Type*Stimulus-Type interaction showed that the amplitude of the SRF was stronger for the phrase condition than the sentence condition only at the middle frequency band ($t (29) = 4.67$, $p < 6e\text{-}5$ ***). The results signify that at the middle frequency range, the neural response of phrases was more strongly predicted solely by modeling the encoding of the acoustic features than it was in the sentence condition. This pattern of results suggests that the neural representation of sentences is more abstracted away from the neural response that is driven by the physicality of the stimulus.

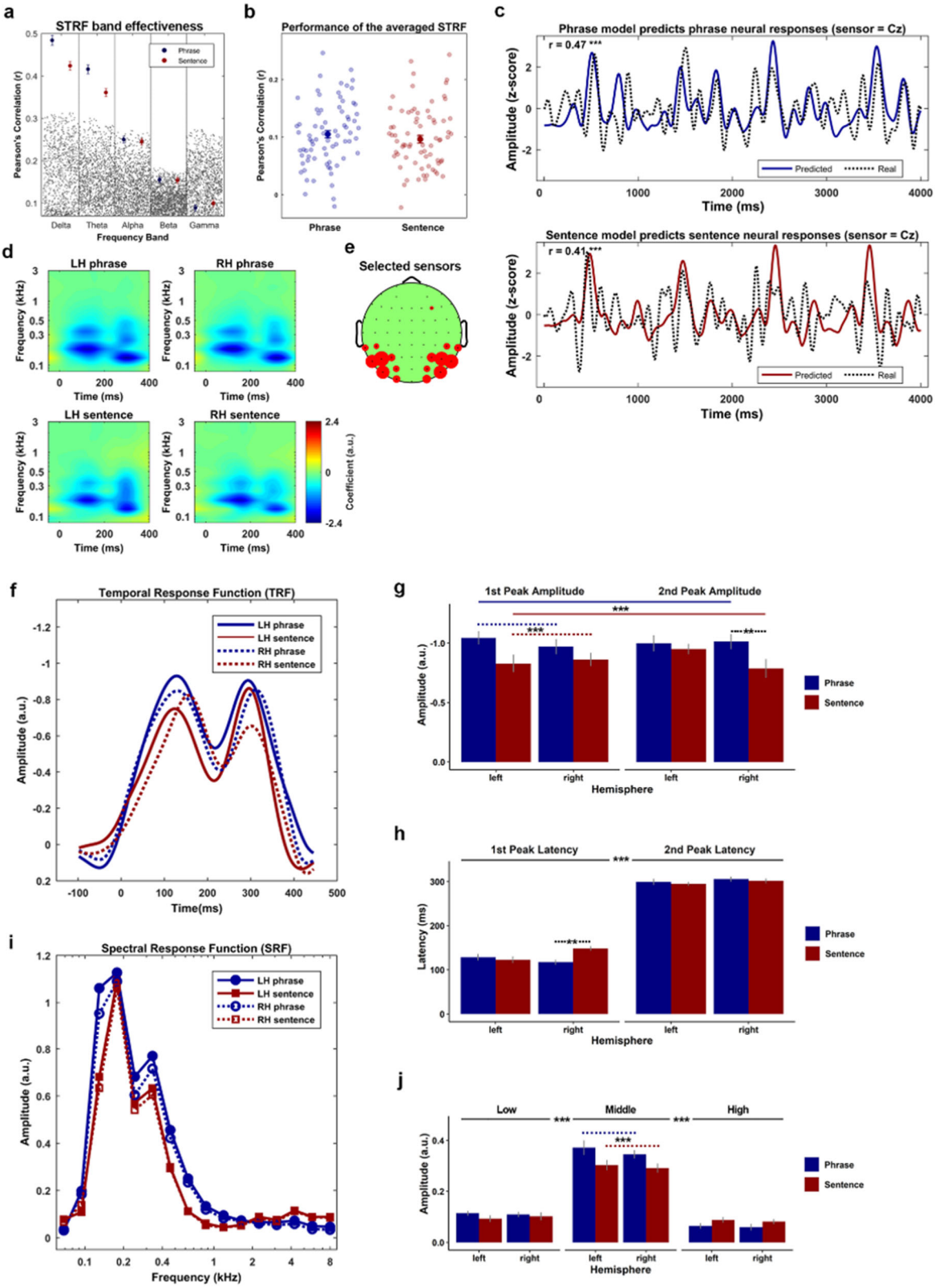

Figure 3. Acoustic features are encoded differently for phrases vs. sentences in a phase locked manner. ***(a)*** *Comparison between the real and random performance of the STRF in each frequency band. The results suggested that only the performance of the STRF in the delta band (< 4 Hz) and theta band (4-8 Hz) was better*

than the random performance. The blue and red dots represent the real performance of the STRFs for the phrases and sentences, respectively. The gray dots represent the random performance, and the error bar represents two SEM. ***(b)*** *The performance of the low frequency (< 8 Hz) STRF that was averaged across all participants. The solid blue and red dots represent the average performance across all the testing trials. The error bar represents two SEM. The light blue and red dots represent the model's performance on each testing trial for the phrase condition and the sentence condition, respectively. The results indicate no difference in the kernel between the phrase and sentence condition.* ***(c)*** *The comparison between the real neural responses (dashed lines) and the average responses predicted by the model (solid blue for the phrases, solid red for the sentences) at a sample sensor, Cz. The results suggest that the STRFs performed equally well for the phrases (r = 0.47 ***, n=1024) and sentences (r=0.41 ***, n=1024).* ***(d)*** *The STRF clustered according to the selected sensors that showed the biggest responses (negative) on the ROI. The figures on the left and right sides of the upper panel represent the clustered STRF for the phrases at the left and right hemisphere, respectively. The corresponding positions on the lower panel represent the clustering for the sentence condition.* ***(e)*** *The selection of sensors. The bigger the red circle, the more times the sensor was selected across all participants.* ***(f)*** *The TRFs that were decomposed from the STRFs. The blue and red lines represent the phrase condition and sentence condition, respectively. The solid and dashed lines respectively indicate the left and right hemisphere.*

(g) *The comparison of the magnitude of the TRFs. The blue and red bars represent the condition and sentence condition, respectively. The error bar shows one SEM on each side. A three-way repeated-measure ANOVA of the peak magnitude was conducted on the factors of Stimulus-Type (phrase or sentence), Hemisphere (left or right) and Peak-Type (~100 ms or ~300 ms). The results indicated a main effect of Stimulus-Type and a three-way interaction. The post-hoc comparison on the former suggested that the amplitude (negative) was stronger for the phrase condition than the sentence condition (t (59) = 4.55, P < 2e-5 ***). To investigate the three-way interaction of Stimulus-Type*Peak-Type*Hemisphere, two-way repeated-measure ANOVA with the Bonferroni correction were conducted on the factors of Hemisphere and Audio-Type at each level of the Peak-Type. The results indicated a main effect of Stimulus-Type at the first peak (F (1, 14) = 8.19, p = 0.012 *) and a two-way Hemisphere*Stimulus-Type interaction at the second peak (F (1, 14) = 6.42, p = 0.023 *). At the first peak, a post-hoc comparison on the*

*main effect of Stimulus-Type was conducted using paired sample t-tests, and the results showed that the magnitude of the phrase condition was higher than the magnitude of the sentence condition (t(29) = 3.49, p = 0.001 ***). For the two-way Hemisphere*Stimulus-Type interaction at the second peak, paired sample t-tests with the Bonferroni correction were conducted to compare the difference in the magnitude between the phrase and sentence condition at each hemisphere. The results revealed that the magnitude at the second peak was stronger for the phrase condition than the sentence condition in the right hemisphere (t (14) = 3.21, p = 0.006 **), but not the left hemisphere (t (14) = 0.86, p = 0.40).* ***(h)*** *The comparison of the peak latency of the TRFs, with the blue and red bars representing the phrase condition and sentence condition, respectively. The error bar shows one SEM on each side. A three-way repeated-measure ANOVA of the peak latency was conducted on the factors of Stimulus-Type (phrase or sentence), Hemisphere (left or right), and Peak-Type (~100 ms or ~300 ms). The results pointed to a main effect of Peak-Type and a three-way interaction. The post-hoc comparison on the former suggested that the latency of the first peak was significantly faster than the second peak (t (59) = 38.89, p < 2e-16 ***). The post-hoc comparison on the latter with the Bonferroni correction on the factors of Hemisphere and Stimulus-Type for each level of the Peak-Type suggested a two-way interaction between them at the first peak (F (1, 14) = 12.83, p = 0.002***). The results of the post-hoc comparison on this two-way interaction using paired sample t-tests with the Bonferroni correction showed that the latency at the first peak was significantly longer for the sentences than the phrases at the right hemisphere (t(14) = 3.55, p = 0.003 ***), but not the left hemisphere (t(14) = 0.58, p = 0.56).* ***(i)*** *The SRFs which were decomposed from the STRFs. The red and blue lines signify the phrase condition and the sentence condition, respectively, and the solid and dashed lines represent the left and right hemisphere.* ***(j)*** *The comparison of the amplitude of the SRFs. The SRF was first separated into three bands, low (< 0.1 kHz), middle (0.1 to 0.8 kHz) and high (> 0.8 kHz) based on the averaged frequency response of the STRF. Then a three-way repeated-measure ANOVA of the amplitude was conducted on the factors of Stimulus-Type (phrase or sentence), Hemisphere (left or right), and Frequency-Band (low, middle or high). The results indicated a main effect of Band-Type (F (2, 28) = 119.67, p < 2e-14 ***) and a two-way Band-Type*Stimulus-Type interaction (F (2, 28) = 27.61, p < 3e-7 ***). The post-hoc comparison on the former using paired sample t-tests with the Bonferroni correction showed that the magnitude of the middle frequency band was stronger than the low frequency band (t(59) = 17.9, p < 4e-25 ***) and high frequency*

*band (t(59) = 18.7, p < 5e-26 ***). The post-hoc comparison on the Band-Type*Stimulus-Type interaction using paired sample t-tests with the Bonferroni correction showed that the amplitude of the SRF was stronger for the phrases than the sentences only at the middle frequency band (t (29) = 4.67, p < 6e-5 ***).*

5.4 Discussion

Using the same dataset as in Chapter 4, our analysis in this chapter indicated that the intensity of the induced power at the alpha band represented the differentiation between the phrases and sentences, which could be evidence that alpha band neural oscillations are involved in syntactic structure discrimination. Moreover, the functional connectivity via induced intensity suggested that the degree of alpha band connectivity was stronger for the sentences than the phrases. As the degree of connectivity was reflected by the inhibition level, the results indicated that the inhibition of the alpha band connectivity was stronger for the phrases than the sentences. The results revealed that the intensity-organized neural network was a robust readout for syntactic structure discrimination between the phrases and the sentences. Lastly, using the STRF we showed that the brain exploited different encoding mechanisms for representing the syntactically different linguistic structures. In both the temporal and spectral dimensions, the STRF showed different encoding characteristics between the phrases and the sentences, which suggests that similar acoustic features can be represented differently via an endogenous phase-locked encoding.

As I mentioned in the introduction to this chapter, the core debate about the role of alpha band oscillations is whether they reach to high-level linguistic processing (e.g., syntactic representation). In the work described in this chapter, using two types of normalized speech stimuli, we found evidence showing the involvement of the induced alpha power in syntactic structure discrimination. Previous studies have shown that different physical loads for attention (Strauß, Wöstmann, & Obleser, 2014; Wöstmann et al., 2016; Wöstmann et al., 2015; Wöstmann, Lim, & Obleser, 2017) or working memory (Obleser et al., 2012; Wilsch & Obleser, 2016) across conditions can be represented in the alpha band activities. Therefore, researchers have agreed on the role of alpha band oscillations in

perceptual-level processing (e.g. auditory processing). However, this perceptual-level account does not fit well with our results. We matched both the physical and semantic features between the phrases and sentences (see **Figure 3** in Chapter 4), so no physical differences existed in the stimuli itself across conditions. Moreover, our experimental task, which asked participants to evenly distribute their attention to three types of properties of the stimuli (color, object and syntactic structure), had made sure that participants listened to the speech stimuli and the perceptual-level load was the same across the two conditions. In other words, there were no differences in the load placed on working memory or attention.

Therefore, the finding that the alpha-band-induced power separated sentences and phrases pointed to the role of alpha band neural oscillations in syntactic structure representation. In addition, the topographical distribution of this representation was largely localized at the left hemisphere, which suggests high-level language processing (Cutting, 1974; Hickok & Poeppel, 2000; Kimura, 1961; Strauss & Wada, 1983). Furthermore, the temporal evolution of this discrimination effect was largely consistent with the study showing that alpha band oscillations reflect speech intelligibility (Obleser & Weisz, 2012). From this perspective, our results suggest that the intelligibility effect found by Obleser and Weisz (2012) could be an indication of syntactic structure integration.

It needs to be noted that neural oscillations reflect participants' mental processes when facing a perceptual task. This stems from the fact that the neural activity varies as a function of the task (Hickok & Poeppel, 2000); it is not a necessary condition that the response at one band (e.g. alpha) has to reflect one specific mental process. The fact that alpha band activity reflects low-level perceptual gating does not rule out its role in high-level linguistic processing. Therefore, the inherent alpha band oscillations could be altered by both perceptual and language-level tasks.

The power connectivity analysis also suggested a separation between the phrases and sentences. The overall inhibition degree of the power connectivity at the alpha band (~ 7.5 to ~ 13.5 Hz) was stronger for the phrases than the sentences. Several aspects are worth mentioning, starting with the duration of the connectivity effect. In the results relating to induced power, we saw that the syntactic structure discrimination was well captured in the alpha band, as observed

within the listening stage, from ~350 to ~1000 ms. In contrast, the separation effect reflected by the power connectivity lasted from ~100 to ~2200 ms after the stimulus onset. The long-lasting effect of the power connectivity, which extended beyond the listening stage, suggests that the brain needed more time to construct a distributional syntactic structure representation. Apparently, to perform the experimental task, participants would first extract the necessary components, then consecutively integrate and represent the extracted units. It makes sense that the integration and representation of the extracted components lasted beyond the stage of listening. In this view, the timing effect of the power connectivity could reflect syntactic structure integration. Second, previous research using P600 has consistently found that the posterior region is involved in syntactic integration (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Frisch et al., 2002; Hagoort, Brown, & Groothusen, 1993; Kaan et al., 2000; Neville et al., 1991; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Patel et al., 1998; Van Herten, Kolk, & Chwilla, 2005). In accordance with this body of work, our analysis also found that the largest connectivity inhibition was located at the posterior region. The spatial consistency between our findings and previous results indicates that the readout reflected syntactic structure representation. Third, after determining the power connectivity differences between the phrases and sentences, we extracted the connectivity distribution using the ROI (which was 100 to 2200 ms in time and 7.5 to 13.5 Hz in frequency). The connectivity pattern showed that the overall degree of connectivity was stronger for the sentences than the phrases, which suggests that the phrases and sentences are represented by a similar connectivity network, but the inter-regional connectivity was stronger for the sentences than the phrases. The results are consistent with the prediction of the computational model proposed by Martin and Doumas (Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas, 2020), who hypothesized that higher-level connectivity is required to represent a more abstract syntactic structure (the sentences have a more complicated syntactic decomposition than the phrases; see **Figure 3 in Chapter 4**). Lastly, our analysis showed that the inhibition of the power connectivity was stronger for the phrases than the sentences. A stronger inhibition of connectivity indicates a weaker connection. Because the inter-regional connectivity inhibition was stronger for the phrases than the sentences, this implied that a higher-level connectivity was

constructed for the sentences to separate it from the phrases. In sum, our analysis suggests that the syntactic structure discrimination between the phrases and sentences can be represented by the degree and the pattern of the power connectivity. The two types of syntactic structures had a similar distributional representation; however, the intensity of the connectivity was higher for the sentences than the phrases.

The STRF analysis highlighted that acoustic features for the phrases and the sentences are encoded differently by the brain in both the temporal and the spatial dimension. In the following section, I discuss the implications of this finding.

First, consistent with previous research (Ding & Simon, 2012a, 2012b), for both phrases and sentences we found that only the low-frequency (< 8 Hz) neural activity robustly reflected the representation of the acoustic features. This suggests that the acoustic features of speech are represented in a relatively slow neural response via phase-locked encoding.

Second, by fitting STRFs using low frequency (< 8 Hz) activities with the narrow band envelopes, we demonstrated that the slow temporal modulations of speech (low-frequency acoustic features, < 1 kHz) were represented bilaterally in the brain. Moreover, the sensors that prominently reflected the phase-locked relationship between the acoustic features and the low frequency activities were localized at the posterior region of both hemispheres. Neurophysiological studies have repeatedly shown the relatedness of the posterior regions in syntactic integration (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Frisch et al., 2002; Hagoort, Brown, & Groothusen, 1993; Kaan et al., 2000; Neville et al., 1991; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Patel et al., 1998; Van Herten, Kolk, & Chwilla, 2005). Associating our findings with previous studies on syntactic integration, we found that the encoding of the acoustic features, which were represented by narrow-band temporal envelopes, indicates the neural representation of syntactic structures.

Third, to explore how the acoustic features are encoded in both the temporal and spectral dimensions, we decomposed the STRF into the TRF and SRF. The results from both suggested that acoustics for phrases and the sentences are encoded differently. More specifically, from the results of the TRF, we know that the brain represents acoustic features in the low-frequency neural response with a

two-peak temporal dependency (~100 and ~300 ms), which reflects that the instantaneous low-frequency neural activity was predominantly driven by the encoding of acoustic features that were presented ~100 and ~300 ms ago.

When we only consider intensity (approximately 100-ms time window), sentences depended on acoustic features less strongly than phrases. This result is consistent with the idea that sentence representations are more abstracted away from the physical input because they contain more linguistic structural units (i.e., constituents) that are not vertically present in the physical or sensory stimulus. Consistent with previous research, we found that the instantaneous neural response was strongly driven by the encoding of the acoustic features presented approximately 100-ms ago (Brodbeck, Hong, & Simon, 2018; Crosse & Lalor, 2014; Di Liberto, O'Sullivan, & Lalor, 2015; Ding & Simon, 2012a, 2012b, 2013b; Golumbic et al., 2013; Puvvada & Simon, 2017; Wang et al., 2019).

When we only consider the latency (approximately 100-ms time window), and only the right hemisphere, the low-frequency neural response to sentences was predominantly driven by the acoustic features that appeared earlier in time than the acoustic features that drove the neural response to phrases. Our results imply that the brain distinguishes syntactically different linguistic structures by how its responses are driven by the acoustic features that appeared approximately 100-ms ago. More importantly, over the right hemisphere, our findings suggest that the low-frequency neural response to sentences reflected the encoding of the acoustic features that appeared earlier in time than the acoustic features that drove the neural response to phrases. This could be evidence that the right hemisphere is dominant in extracting the slow timescale information of speech that is relevant for, or even shapes, higher-level linguistic structure processing, e.g., syntactic structure building (Ding & Simon, 2012a, 2012b; Poeppel, 2003). It is noteworthy to see that the distribution in time and space of these patterns is consistent with the idea that the brain is extracting information from the sensory input at different timescales and that this process is, in turn, is reflected in the degree of departure (in terms of informational similarity) of the neural response from physical features of the sensory input.

At ~300 ms, when we only consider the intensity of the acoustic features, the low-frequency neural response to the phrases more strongly depended on the

acoustic features than the response to the sentences did. However, the TRF comparison indicated that the brain exploited a different encoding mechanism across the hemispheres to discriminate between the phrases and sentences. Specifically, our analysis shows that the low-frequency neural response to the phrases had a stronger dependency on the intensity of the acoustic features than the sentence condition did at the right hemisphere, but not at the left hemisphere. These results first imply that the instantaneous low-frequency neural response reflects the encoding of the acoustic features that were present ~300 ms ago for both conditions. Moreover, only at the right hemisphere, the low-frequency neural response of the phrases more strongly depends on the acoustic features from ~300 ms ago when compared with the response to the sentences. The findings remind us of the results of the phase connectivity study reported in Chapter 4, in which the degree of phase connectivity also showed a different pattern between the phrases and sentences at the right posterior region. Consistent with previous works that point to the involvement of the right hemisphere in processing the slow modulations (Abrams et al., 2008; Ding & Simon, 2012a, 2012b; Giraud et al., 2007; Kerlin, Shahin, & Miller, 2010; Luo & Poeppel, 2007; Poeppel, 2003), our results further underline that the brain can discriminate these two types of structures by differently representing the acoustic features that appeared ~300 ms ago at the right hemisphere. That sentence representations were more abstract and less driven by the acoustics in the left hemisphere is consistent with contemporary neurobiological models of sentence processing (Friederici, 1995; Hagoort, 2013).

The results of the SRF indicated that the brain can discriminate between phrases and sentences by representing acoustic features in the first formant (Catford, 1988; Jeans, 1968; Titze et al., 2015; Titze & Martin, 1998). More specifically, within the range of the first formant, the low-frequency neural response to the phrases reflected a stronger dependency on the acoustic features than the response to the sentences did. Unlike consonants, the intensity of vowels can be well reflected at the first formant (<1 kHz) (Catford, 1988; Jeans, 1968; Titze et al., 2015; Titze & Martin, 1998). Given that the stimuli were not physically different, this pattern of results convincingly suggests that the brain is 'adding' information, for example by actively selecting and representing linguistic structures that are cued by the physical input and its sensory correlate. This is consistent with the finding that low-frequency cortical entrainment to speech

reflects phoneme-level processing (Di Liberto & Lalor, 2017; Di Liberto, O'Sullivan, & Lalor, 2015; Keitel, Gross, & Kayser, 2018; Khalighinejad, da Silva, & Mesgarani, 2017). In addition, our results imply that this phonemic-level representation of acoustic features can reflect the syntactic differences between linguistic structures.

In complement to Chapter 4, in this chapter, we found that syntactic structure differences can be represented in the intensity and connectivity of the induced neural response. More interestingly, by modeling the encoding of acoustic features, we showed that the brain can represent the syntactic differences between phrases and sentences by conveying the phonemic-level acoustic features in the low-frequency neural response. On the whole, combining the results of Chapters 4 and 5, we have provided a comprehensive readout on how the syntactic differences between phrases and sentences are reflected in the brain.

6 | General discussion

Speech segmentation and syntactic representation are crucial steps leading to language comprehension. In the previous chapters, I reported a series of MEG experiments, first with Dutch native speakers to show how the brain segments speech stimuli into chunked units at different levels of linguistic representation, and especially how statistical information is used to perform cue-based structure extraction (Chapter 2). Then, a parallel series of MEG experiments with Chinese native speakers showed the stability and consistency of the involvement of this statistical information in the inference process (Chapter 3). Finally, I reported findings of an EEG experiment assessing differences between two types of syntactic structures, phrases and sentences, in various dimensions of the neural response (Chapters 4 and 5). In the following parts of this section, I first provide a brief summary of the findings concerning speech segmentation and syntactic representation during language comprehension. Then I discuss the implication of the findings in a broader context and the progression of our investigation into speech segmentation and syntactic representation. Finally, I outline some questions arising from our research and possible directions for further studies.

6.1 Summary of core findings

In ***Chapter 2****,* the results of six MEG experiments with Dutch native speakers were reported to explore the role of statistical information, i.e., transitional probability, in speech segmentation. Using a revised paradigm from Saffran, Aslin, and Newport (1996) with isochronous syllable sequences (four syllables per second, 4 Hz) in a language which was either known or unknown to participants, we determined the role of transitional probability in the cortical tracking of hierarchical linguistic structures (Ding et al., 2016). We found that neural oscillations indicating the occurrence rate of linguistic structures at multiple levels

could be solely introduced by statistical information, namely, transitional probability (TP).

More specifically, in ***Experiment 1***, using three types of Dutch syllable sequences, i.e. noun sequences and random syllable sequences played forward and backward, we first showed that the cortical tracking effect exists in Dutch for native speakers. Then, in ***Experiment 2***, by changing the three types of speech stimuli to Mandarin Chinese (which the Dutch participants did not know), we removed high-level language information, such as semantic and grammatical knowledge, and showed that the neural activities tracking the rhythm of linguistic structures could be introduced solely by statistical information. Moreover, to check whether the cortical tracking effect still exists when multiple layers of units are fitted in and to match the structure of our syllable sequences with the hierarchy of the stimuli in Ding et al. (2016), we conducted Experiments 3 and 4. In ***Experiment 3***, we first trained participants to statistically combine noun pairs into novel compounds that do not exist in Dutch. Then in ***Experiment 4***, we scanned participants when they listened to the materials trained before, i.e. in Experiment 3. As expected, we found three peaks reflecting the rhythm of novel compounds (1 Hz), words (2 Hz), and syllables (4 Hz) for the noun sequences. The results show that neural activities can track multiple levels of structures simultaneously. More importantly, the neural activity that appeared at the frequency of the highest-level structures (novel compounds) was not found in Experiment 1, which underlines the effectiveness of training and indicates that the cortical response to the rhythm of linguistic structures can be manipulated by using statistical information.

However, in Experiment 3, participants were trained on stimuli in their own language. Except for the statistical information (TP) that was learned during the training stage, participants could also semantically associate word pairs. To assess the concerns about semantic association, we conducted ***Experiments 5 and 6***, in which the same procedures and manipulations as Experiments 3 and 4 were used, except that the stimuli were in Mandarin Chinese. This meant that higher-level linguistic information was removed from the processing, as the Dutch participants did not understand the stimuli. Using the same analytical methods, as expected, we got the same results: there were still frequency peaks to reflect the rhythm of compounds (1 Hz), words (2 Hz), and syllables (4 Hz) for the noun sequences. We concluded that the cortical tracking effect can be solely introduced

by statistical information, and the brain can track the statistically defined structures at different levels simultaneously. Note that this does not preclude the possibility of cortical tracking signatures for native language processing in the absence of statistical learning. In other words, the fact that cortical tracking occurs in response to statistically defined stimuli on different timescales does not mean that linguistic representations, recognized from speech input based on endogenous linguistic knowledge of a native language, are not also subject to cortical tracking.

In ***Chapter 3***, I reported the results of the same six MEG experiments as in Chapter 2, but with Chinese participants. The reason for doing so is that we hypothesized that structure chunking via statistical information could be a generalized perceptual mechanism, which means the cortical tracking effect that reflects this endogenous process could be independent of language comprehension in some circumstances. Therefore, we predicted that we would obtain the same results if we conducted these experiments with users of a different language; i.e., the findings would be the same regardless of whether the speaker could comprehend the language of the speech input. Our results fully confirm this prediction: the results of Experiments 1 to 6 in Chapter 3 closely correspond to the results of the analogous experiments reported in Chapter 2. Finding the same pattern of results across different language users suggests that the frequency-tagging effect reflects generalized perceptual processing, and higher-level linguistic knowledge is not necessary to introduce the effect.

In ***Chapter 4***, we investigated how two types of syntactic structures, i.e., phrases and sentences, are differently represented in phase-related measures of neural readouts. First, we calculated the inter-trial phase coherence (ITPC). The cluster-based permutation tests suggested that the phase coherence for the sentences was significantly higher than for the phrases. The differences were found at the theta band (~2 to 7 Hz) during listening (~ 450 to 900 ms) over the central electrodes. The results reflect the processes by which the brain can extract a key component, e.g. a verb in a sentence, from physically inseparable stimuli via temporal synchronization to separate the two types of syntactic structure. Second, we calculated phase connectivity, indexed by inter-site phase coherence (ISPC), across the sensor space. Cluster-based permutation tests suggested that the degree of phase connectivity was significantly higher for the sentence condition than the phrase condition. The differences were found between ~1800 and ~2600 ms after

audio onset and were largely localized in the right posterior region with frequencies falling within a very low range (<~2 Hz). The long latency of the effect suggested a long-lasting process of building spatial connectivity distributions to distinguish the two types of structures. Third, low-frequency (< ~ 2 Hz) phase connectivity indicates a slow temporal synchronization over the sensor space, which could reflect the act of integrating information extracted during listening. Lastly, the differences in the degree of connectivity that were localized in the right posterior region emphasize that the right hemisphere is strongly engaged in syntactic representation. Our last concern was the role of phase-amplitude coupling in representing syntactic differences, as the low frequency phase entrained with high frequency amplitude was considered as a generalized neural mechanism for speech perception (Giraud & Poeppel, 2012). To examine this, we calculated the normalized phase-amplitude coupling (PAC-Z) values and found that there was a strong coupling between the low frequency phase (~4 to 10 Hz) and high frequency amplitude (~15 to 40 Hz) during the listening stage. However, there was no evidence suggesting a difference between the phases and sentences. Consistent with Giraud and Poeppel (2012), we considered PAC to be a generalized mechanism for speech perception, and our analysis suggested that it might not reflect processing at the syntactic level.

In ***Chapter 5***, using the same data as in Chapter 4, we investigated how the intensity of the neural response would reflect the differences between the two types of syntactic structure. In addition, STRF (spectral-temporal response function) modeling was conducted to show how acoustic features are encoded to separate phrases from sentences in a phase-locked manner. First, the comparison of the induced power suggested a difference between the phrases and sentences over the alpha band (7.5 to 13.5 Hz). The effect was largely localized at the left hemisphere and extended from ~350 to 1000 ms after the audio onset. The results suggested that alpha band inhibition could reflect syntactic structure discrimination between phrases and sentences.

Then, an intensity (power) connectivity analysis showed that alpha band connectivity was higher for sentences than phrases. The analysis indicated that alpha band connectivity could effectively separate phrases and sentences without hemisphere dominance.

Lastly, a phase-locked encoding model using STRFs was conducted to show the differences between the phrases and sentences. We first demonstrated that only low-frequency (< 9 Hz) neural responses effectively reflect the encoding of acoustic features. Then, using the neural activity below 9 Hz, we fitted STRFs for both conditions. As the temporal and spectral dimensions might feature different encoding characteristics, we extracted the temporal response function (TRF) and spectral response function (SRF) separately.

For the magnitude of the TRF, statistical comparisons suggested that the instantaneous neural activity in response to phrases had a stronger phase-locked dependency on the acoustic features than in response to sentences. In addition, the overall dependency on acoustic features was statistically higher for the phrase condition than the sentence condition within a temporal window of ~50 to ~150 ms. However, in the window from ~250 to ~350 ms, the magnitude was stronger for phrases than sentences only in the right hemisphere.

As for the latency of the TRF, statistical comparisons indicated that the latency at the first peak was significantly longer for sentences than for phrases only at the right hemisphere. The results highlight that within the first temporal window (~50 to 150 ms), only at the right hemisphere, the low-frequency neural response to sentences reflected the encoding of the acoustic features that appeared earlier in time than the acoustic features that drove the neural response to phrases.

For the SRF, the statistical analysis indicated that the neural response was predominantly driven by the encoding of the acoustic features from 0.1 to 0.8 kHz corresponding to the range of the first formant (Catford, 1988; Jeans, 1968; Titze et al., 2015; Titze & Martin, 1998) for both conditions. Additionally, in this range, the neural response to phrases was more strongly predicted solely by modeling the encoding of the acoustic features than it was in the sentence condition.

6.2 Speech segmentation using statistical information

Investigations have showed that information in both the temporal and spectral dimensions of speech contributes to structure extraction (Shannon et al., 1995; Smith, Delgutte, & Oxenham, 2002; Zeng et al., 2005). For instance, speech segmentation can be performed using cues that naturally reside in acoustics such as prosody (Cutler, Dahan, & Van Donselaar, 1997; Peña et al., 2002; Steinhauer,

Alter, & Friederici, 1999) and stress pattern (Cutler, 2012; Cutler & Butterfield, 1992). Although these acoustic-level features are crucial for structure extraction, segmenting speech using only phonological characteristics seems to omit the grammatical and semantic relationships between hierarchical linguistic structures. As such, many theories posit that linguistic structures can be extracted via an endogenous inference process (Bever & Poeppel, 2010; Brown, Tanenhaus, & Dilley, 2021; Friederici, 1995; Hagoort, 2013; Halle & Stevens, 1962; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas, 2020; Meyer, Sun, & Martin, 2020; Phillips, 2003; Poeppel & Monahan, 2011). In accordance with this inference argument, several studies have suggested that higher-level linguistic knowledge, such as grammatical or syntactic information, is crucial in structure extraction (Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018; Meyer & Gumbert, 2018). The neural activities tracking the occurrence rate of hierarchical linguistic structures were considered to be a reflection of speech segmentation using higher-level linguistic knowledge. Under this account, grammatical or syntactic information appears to be the key factor driving the cortical tracking effect. However, the endogenous cue-based segmentation also benefits from statistical information, i.e., the transitional probability between multiple levels of linguistic structures (Saffran, Aslin, & Newport, 1996).

The TP often reflects the relationships between upper- and lower-level linguistic units as well. For instance, the TP between syllables in a word is higher than the TP between syllables at word boundaries, and the TP between words in a phrase is also higher than the TP between two words that cannot form a phrase. By comparing the two accounts, i.e. grammatical chunking vs. structuring via TP in speech segmentation, we designed an investigation into whether the cortical tracking effect could be solely introduced by statistical information. Building on previous studies (Ding et al., 2016; Saffran, Aslin, & Newport, 1996), we connected the effect of cortical tracking to linguistic structures (Ding et al., 2016) with the account of structure extraction via statistical information (Saffran, Aslin, & Newport, 1996). Using artificially synthesized speech stimuli, we first eliminated the cues at the acoustic level (e.g. prosody and stress pattern) from the structure chunking by presenting the syllables isochronously, i.e. four times per second (4

Hz). Furthermore, by holding the statistical information consistent between experiments, we were able to compare the effect introduced by stimuli in participants' own language (linguistic knowledge involved) with the effect evoked by stimuli in an unfamiliar language (linguistic knowledge removed). In addition, to show whether the statistically defined structures could be tracked simultaneously at different levels, we also trained participants to extract statistically defined novel compounds. Finally, to check whether the statistically driven effect is independent of language, we repeated all the experiments with Chinese participants. All these operations and manipulations led to one conclusion, which is that the cortical tracking effect can be introduced without higher-level linguistic knowledge.

Our findings are informative as to the flexibility of cortical tracking. Firstly, we showed that the tracking effect could be introduced solely by statistical information, which is at odds with the account that the effect is purely a reflection of syntactic integration or grammatical chunking via higher-level linguistic knowledge. In natural language, one important source of information that defines boundaries between linguistic units is the transitional probability. Within a corpus of speech, this kind of statistical information is available in measurable quantities which can be used to extract sound sequences that comprise linguistic structures. The transitional probability from one sound to the next will generally be higher when the two sounds follow one another within a unit than when they are located at a boundary between units (Chomsky, 2014; Saffran, Aslin, & Newport, 1996). Behavioral studies have also shown that the measurable transitional probabilities can be used to extract structures from continuous speech stimuli (Pelucchi, Hay, & Saffran, 2009; Saffran, 2003; Saffran et al., 1999; Saffran, Newport, & Aslin, 1996). However, to examine the role of statistical information in structure chunking, artificial speech stimuli have often been the first choice. Using such stimuli with unnatural language would potentially disrupt the language-specific characteristics and limit the applicability of the conclusions. In contrast, our experimental manipulations used natural speech stimuli while holding the statistical information (TP) between the hierarchical units constant, effectively separating the role of statistical information from the role of higher-level linguistic information that coexists with statistical information in structure extraction. This enabled us to demonstrate that structure chunking via statistical information can introduce the

cortical tracking effect, which shed light on the neural representation of how statistical information is utilized to segment speech.

Secondly, consistent with Saffran, Aslin, and Newport (1996), our results suggested that the endogenous process of extracting structures from continuous speech via statistical information could play a fundamental role in language acquisition, which implies that grammatical analysis and syntactic integration can be performed after unit extraction. A generalized speech-perception model proposed by Giraud and Poeppel (2012) also considered that extracting syllables via low-frequency phase alignment is the first step toward linguistic analysis via higher-frequency neural activities. This might be obvious when considering how language learners try to understand a piece of speech. For instance, an English learner may have a perfect understanding of a sentence in its written form, but fail to extract meaning from it in its spoken form. The difficulty of understanding the sentence in the latter form could be due to the learner failing to perform a linguistic analysis, as this requires successful speech segmentation. It is undeniable that higher-level linguistic knowledge is helpful for speech segmentation, and successful comprehension requires grammatical and syntactic information. However, the neural representation of higher-level linguistic information, e.g. syntactic structure, is not necessarily the only data reflected in the cortical tracking effect.

Lastly, our experiments were conducted with both Dutch and Chinese speakers. The consistent results across languages suggest that structure extraction via statistical information, which was reflected by the cortical tracking effect, could be a generalized perceptual mechanism. Despite the differences between Dutch and Chinese ranging from physical characteristics (e.g., Dutch is a stress-based language and Chinese is a tone-based language) to high-level linguistic regularities (e.g., there is a difference between singular and plural nouns in Dutch, but not in Chinese), consistent results were always acquired no matter what language the participants listened to. The results indicated that linguistic properties did not vary the frequency response tagging the rhythm of hierarchical linguistic structures, which could be a sign that frequency peaks in the brain are not a function of linguistic properties. In addition, experiments conducted using artificial speech sequences and visual stimuli with manipulated transitional probabilities have shown the same cortical tracking effect (Henin et al., 2021). The existence of this

effect in visual-perceptual tasks further strengthens the argument that cortical activity tracking the rhythm of stimuli might not be a language-specific phenomenon.

In sum, we pinpointed the role of statistical information, i.e. transitional probabilities, in speech segmentation by conducting a number of MEG experiments. As the frequency response tagging the rhythm of units was sensitive to the manipulation of statistical properties, but insensitive to the physical and language-specific properties, we argue that the cortical tracking effect could reflect a generalized perceptual mechanism for structure extraction. Our investigations provide new evidence and readouts on the neural representation of speech segmentation via statistical information. However, since our exploration focused on speech segmentation via TP and was conducted using isochronous syllable sequences, future studies might aim to examine the neural representation of grammatical chunking or syntactic integration, prioritizing the use of stimuli that are more natural. An especially fruitful pursuit would be to compare how the acoustic features that are encoded when segmenting speech differ depending on whether or not the individual understands the given language.

6.3 Neural representation of syntactic structure discrimination

In ***Chapters 2 and 3***, we explored the role of statistical information in speech segmentation; we showed that the frequency responses in the brain also reflect the statistical properties of the stimuli. If the cortical response tracking linguistic structures at different levels reflects structure chunking via statistical information, how then does the brain represent the syntactic structure as it extracts it from the speech input? Which dimensions of the neural response could reflect syntactic structures? To address these questions, we investigated various neural readouts to look for links to syntactic representation, as reported in ***Chapters 4 and 5***. We increased our chances of finding such links by maximizing the physical and semantic similarities between the linguistic structures. More specifically, we investigated which dimensions of neural activity distinguish the linguistic structure of phrases and sentences and used a series of analytical techniques to better describe the dimensions of neural readouts that were sensitive to the distinctions.

We asked whether phrases and sentences have different effects on functional connectivity, and found, first, that while phrases and sentences recruit similar functional networks, the intensity of those networks was scaled with different types of linguistic structure. Sentences showed more phase coherence and power connectivity compared to phrases. This pattern suggests that phrases and sentences differently impact the distribution and intensity of the neural networks involved in speech comprehension. Second, we found that phase-amplitude coupling between theta and gamma, which has been implicated in speech processing, is not sensitive to structural differences in spoken language. Third, we found that activity in the alpha band was sensitive to linguistic structure. Lastly, by modeling acoustic fluctuations in the stimulus and brain response with STRFs, we found that during perception of phrases and sentences, the neural readout differentially relied on the encoding of acoustic features in the brain, and that sentences were more abstracted away from acoustic dynamics in the brain response. In the following three sections I give more details about these findings on phase coherence, phase connectivity, phase-amplitude coupling, induced alpha power, and power connectivity, and discuss potential interpretations of them.

Phase coherence. Consistent with previous research (Doelling et al., 2014; Luo & Poeppel, 2007; Peelle & Davis, 2012; Peelle, Gross, & Davis, 2013), our phase synchronization analysis detected low-frequency phase coherence during speech comprehension. Moreover, phase coherence distinguished between phrases and sentences, yielding a cluster between ~450 and ~900 ms after the audio onset and with frequencies from ~2 Hz to ~8 Hz, which was most pronounced over the central electrodes. These results therefore suggest that syntactic structure may be encoded by low-frequency phase coherence, through the systematic organization of activity in neural networks, in particular their temporal dynamics. Our results are consistent with the notion of 'phase sets' in computational models of structured representations that exploit oscillatory dynamics. Phase sets are representational groupings that are formed by treating distributed patterns of activation as a set when units are in (or out) of phase with one another across the network (Doumas & Martin, 2018; Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2020; Martin & Doumas, 2020). They are key to the representation of structure in artificial neural network models.

Phase connectivity. Phrases and sentences also yielded differences in phase connectivity. In the predefined time and frequency range of interest, the statistical comparison indicated a difference corresponding to a cluster from ~800 to ~1600 ms after the audio offset, occurring at the very low frequency range (< ~2 Hz) that was most pronounced over the right posterior region. Phrases and sentences thus differentially impact the temporal synchronization of neural responses.

Several aspects of the results are noteworthy. First, the relatively late effect suggests that the impact on temporal synchronization occurs after the initial presentation of the speech stimulus. In our experiment, participants were randomly presented with a prompt for three possible tasks (color discrimination, object discrimination, and linguistic structure discrimination), which asked them to identify either 'semantic' information (object or color) or 'syntactic' information (whether the stimulus was a phrase or sentence) from the speech stimulus. Because of the random order of the task trials, participants had to pay close attention to the stimuli and maintain each stimulus in working memory until they received the task prompt. The tasks also ensured that participants could not select just one dimension of the stimulus for processing. Similarly, because we used an object and a color task, participants had to distribute their attention evenly across the adjectives and nouns. Due to these controls and task demands, it was unlikely that the observed connectivity effects reflected mere differences in attention to phrases or sentences. Rather, we were able to attribute the observed effects to the syntactic differences between them.

Second, the low frequency range (< 2 Hz) of the observed effect is consistent with previous research (Brennan & Martin, 2020; Ding et al., 2016; Kaufeld et al., 2020; Keitel, Gross, & Kayser, 2018; Meyer et al., 2017). In Ding et al. (2016), the cortical response was modulated by the timing of the linguistic structure's occurrence; low-frequency neural responses (1 to 2 Hz) were found to track the highest-level linguistic structures (phrases and sentences) in their stimuli. Here we extended their work to ask whether the 1 Hz response could be decomposed to reflect separate syntactic structures (phrases vs. sentences), and we identified the role of phases in discriminating between these structures. In our study, all speech stimuli lasted for one second, and except for the presence of syntactic structure, the stimuli were normalized to be highly similar. Our pattern of results therefore

suggests that functional connectivity, as reflected in the temporal synchronization of the induced neural response, distinguishes between phrases and sentences.

Lastly, phrases and sentences differed most strongly over the right posterior region of the brain, which is broadly consistent with previous research on speech comprehension. Functional magnetic resonance imaging (fMRI) studies implicate the posterior right hemisphere in processing syntactic structure (de Bode et al., 2015; Grodzinsky, 2000; Grodzinsky & Friederici, 2006; Maess et al., 2001). Neurophysiological research also suggests the involvement of the right hemisphere in slow-timescale information extraction (Abrams et al., 2008; Giraud et al., 2007; Morillon et al., 2012; Poeppel, 2003). Additionally, the P600, a positive ERP component often associated with syntactic processing, has a robust right-posterior topographical dominance (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Patel et al., 1998). In light of the existing literature, therefore, the right posterior distribution of the phase connectivity effects is consistent with the processing of syntactic structures, although we refrain from claims about underlying neural sources based on our EEG data.

Phase-amplitude coupling (PAC). We observed PAC during speech comprehension, as a low-frequency phase response (~ 4 to 10 Hz) entrained with high frequency amplitude (~ 15 to 40 Hz). This effect appeared largely over the bilateral central area. The bilateriality of the topographical distribution has been observed in sensory-motor integration (Babiloni et al., 2011; Klimesch, Sauseng, & Hanslmayr, 2007; Neuper, Wörtz, & Pfurtscheller, 2006; Pfurtscheller et al., 2006; Pfurtscheller et al., 1998; Schlögl et al., 2005; Suffczynski et al., 2001), which is consistent with the proposal from Giraud and Poeppel (2012) that PAC reflects an early step in speech encoding involving sensory-motor alignment between the auditory and articulatory systems. Crucially however, this effect did not differ between phrases and sentences. Although this is a null result, the pattern is compatible with the generalized model for speech perception (Giraud & Poeppel, 2012). This early step is presumably similar for phrases and sentences, and perhaps for any type of structure above the syllable level.

Induced alpha power. Induced alpha-band power was also found to distinguish phrases from sentences, and this effect was most pronounced at the left

hemisphere. This pattern implies the involvement of alpha band oscillations in syntactic structure processing. Although alpha band activity is often associated with processing related to attention or working memory (Haegens et al., 2010; Obleser et al., 2012; Strauß, Wöstmann, & Obleser, 2014; Ten Oever, De Weerd, & Sack, 2020; Wilsch & Obleser, 2016; Wöstmann et al., 2016; Wöstmann et al., 2015; Wöstmann, Lim, & Obleser, 2017), we do not consider this a very plausible alternative explanation for our results, due to the following reasons. First, it is not clear why phrases and sentences would differ in their attentional demands. Second, we employed an experimental task to ensure that participants had to pay similar attention to phrases and sentences, and these two structures were associated with similar behavioral performance in each task (with a caveat that performance was at ceiling and therefore may not have detected small differences between conditions). Thus, we do not claim that that all speech-elicited alpha band effects reflect syntactic processing. Some observed effects may well reflect perceptual processing during speech comprehension (e.g. Obleser & Weisz, 2012), especially in experiments designed to manipulate perceptual processing, such as speech-in-noise manipulations. Neural responses in a given band, e.g. the alpha band, need not reflect only one particular perceptual process. Likewise, the fact that the alpha-band neural response could reflect lower-level perceptual processes or working memory load in certain contexts does not necessarily rule out its role in the representation of higher-level linguistic information such as syntax.

Power connectivity. Phrases and sentences elicit differences in induced power connectivity in alpha band activity (7.5 to 13.5 Hz). Phrases showed more inhibition in power connectivity than sentences; in other words, sentences showed a stronger degree of connectivity over the sensor space in the alpha band than phrases did. Several aspects of these results are notable. First, we observed this effect from ~100 to ~2200 ms after the stimulus onset, which suggests that the functional connectivity persisted and outlasted the observed effects in induced alpha power, which we observed from ~350 to ~1000 ms after the audio onset (during the listening stage). The extended nature of the functional connectivity effect could reflect the continuing integration and representation of syntactic and semantic components. Second, alongside differences between phrases and sentences in power connectivity, we also extracted the sensor connectivity pattern (over an ROI ranging from 100 to 2200 ms in time and 7.5 to 13.5 Hz in frequency).

Whereas phrases and sentences showed similar functional connectivity in the intensity of the neural response, sentences showed stronger inter-regional (sensor) connectivity than phrases. By design, in our stimuli the sentences had more constituents than the phrases did. If local network activity is more organized or coherent as a function of linguistic structure, then the difference observed here could reflect the encoding of additional constituents in sentences as compared to phrases. Lastly, phrases elicited a stronger inhibition of induced power connectivity than sentences did. This indicates weaker cooperation between brain regions. In contrast, inter-regional connectivity was stronger for the sentence condition than the phrase condition, which suggested a higher intensity of connectivity between the brain regions for the sentences in order to separate them from the phrases.

In sum, phrases and sentences elicited robust differences in induced power connectivity. A similar functional connectivity pattern was deployed for representing phrases and sentences, but the intensity of the connectivity was stronger for sentences than phrases. This finding is consistent with the prediction that low frequency power and network organization should increase as linguistic structure increases. Our stimuli were designed to allow the measurement of differences in neural dynamics between phrases and sentences, and as such differed in the number and type of linguistic constituents that were perceived. Beyond the number and type of constituents, the phrase and sentence structures also differed in terms of the relations between their constituents, or with respect to the linguistic notion of hierarchy.

Spectral-temporal response function (STRF). We performed an STRF analysis to investigate whether phrases and sentences are encoded differently. Firstly, consistent with previous research (Ding & Simon, 2012a, 2012b), only low-frequency (< 9 Hz) neural responses robustly reflected the phase-locked encoding of the acoustic features. Secondly, we observed a bilateral representation of the slow temporal modulations of speech, in particular at the posterior sensors. The posterior region has consistently been found to be involved in syntactic integration (Coulson, King, & Kutas, 1998; Friederici, Pfeifer, & Hahne, 1993; Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995; Patel et al., 1998). The low-frequency neural response that models the phase-locked encoding of acoustic features can capture structural differences between phrases

and sentences, even without explicitly using a hand-coded annotation on the syntactic level to reconstruct the data. Thirdly and most importantly, we explored these patterns further in both the temporal and spectral dimensions, by decomposing the STRF into the TRF and SRF. Both functions pointed to a different encoding mechanism for phrases vs. sentences. More specifically, the TRF results showed that the brain transduces the speech stimulus into the low-frequency neural response via an encoding mechanism with two peaks in time (at ~100 and ~300 ms). The two peaks reflect the instantaneous low frequency response that is predominantly driven by the encoding of acoustic features that were presented ~100 and ~300 ms ago. In the two windows that centered at ~100 and ~300 ms, phrases and sentences showed a different dependency on the acoustic features in terms of both latency and intensity.

When we only consider intensity (at ~100 ms time window), sentences depend on acoustic features less strongly than phrases. This result is consistent with the idea that representations of sentences are more abstracted away from the physical input because they contain more structural linguistic units (i.e., constituents) that are not present in the physical or sensory stimulus. In accordance with previous research, we found that the instantaneous neural response was strongly driven by the encoding of the acoustic features presented ~100 ms ago (Brodbeck, Hong, & Simon, 2018; Crosse & Lalor, 2014; Di Liberto, O'Sullivan, & Lalor, 2015; Ding & Simon, 2012a, 2012b, 2013a; Golumbic et al., 2013; Puvvada & Simon, 2017; Wang et al., 2019).

When we only consider the latency (again using a ~100 ms time window), and only the right hemisphere, the low-frequency neural response of the sentences was predominantly driven by the acoustic features that appeared earlier in time than the features that drove the response to the phrases. Our results imply that the brain distinguishes syntactically different linguistic structures according to how its responses are driven by the acoustic features that appeared ~100 ms ago. More importantly, at the right hemisphere, the findings suggest that the low-frequency neural response of the sentences reflects the encoding of acoustic features that appeared earlier than the features which triggered the response to the phrases. This could be evidence that the right hemisphere is dominant in extracting the slow-timescale information of speech that is relevant for, or even shapes, higher-level

linguistic processing such as the building of syntactic structures (Ding & Simon, 2012a, 2012b; Poeppel, 2003).

Turning to the SRF results, these indicated that the brain can begin to separate phrases and sentences via a differential reliance on the encoding of the acoustic features from roughly the first formant, and in a phase-locked manner. More particularly, in the range of the first formant, the low-frequency neural response reflected a stronger dependency on the acoustic features in the phrases than in the sentences. Unlike consonants, the intensity of vowels is well reflected at the first formant (<1 kHz) (Catford, 1988; Jeans, 1968; Titze et al., 2015; Titze & Martin, 1998). Although the overall physical intensity of the speech stimulus of the phrases was not different from the sentences, the neural response contains information that discriminates between these syntactic structures. Given that the stimuli were not physically different, this pattern of results strongly suggests that the brain is 'adding' information, for example by actively selecting and representing linguistic structures that are cued by the physical input and its sensory correlate. For example, the brain could be adding phonemic-level information such as vowels via a top-down mechanism; in certain situations and languages, even a single vowel can cue a differential syntactic structure. In fact, in our stimuli, the schwa carries agreement information that indicates the phrasal relationship between *roode* ('red') and *vaas* ('vase') in the phrase *de roode vaas*. Our results, which feature both dependence on, but also departure from, the acoustic signal, are consistent with previous findings that have demonstrated low-frequency cortical entrainment to speech and argued that it can reflect phonemic-level processing (Di Liberto & Lalor, 2017; Di Liberto, O'Sullivan, & Lalor, 2015; Keitel, Gross, & Kayser, 2018; Khalighinejad, da Silva, & Mesgarani, 2017). We extend these findings by showing that when lower-level variables in the stimuli are modeled, the brain can discriminate between syntactic structures even without the addition of higher-level linguistic annotations.

6.4 Future research directions

In the thesis, I first reported our exploration of how speech segmentation using statistical information is reflected in the brain (Chapters 2 and 3), and then presented our investigation into which dimensions of the neural response could

reflect the discrimination between two types of syntactic structures, i.e. phrases and sentences (Chapters 4 and 5). Several extended questions for further investigation are discussed in this final section.

In ***Chapters 2 and 3*** our results suggested that statistical information can be used as an effective cue to segment speech, and the statistical cue-based endogenous process could be reflected in the neural activities tracking the rhythm of hierarchical linguistic structures. Consistent effects across different types of language users were observed no matter whether the language of the stimuli was known or unknown to participants. However, several aspects still need to be investigated. First, the mental processes will be different when same stimuli are presented as a function of whether or not the participants can understand the stimulus language. For instance, the neural response of Dutch participants listening to Dutch stimuli has to be different from Chinese participants listening to the same stimuli, because automatic high-level linguistic processing, e.g. semantic and grammatical processing, is involved in the former situation. Although it is not necessary for the processing differences to be reflected by the frequency-tagging effect, other dimensions of neural activity could still reflect this linguistic-level difference. More extensive documentation of neural readouts that reflect this difference would provide valuable information on the neural representation of high-level language processing. Second, the encoding of acoustic features might vary when users of different languages listen to the same type of speech stimuli. An analytical approach such as STRF or mutual information could be harnessed to show how acoustic features in both the temporal and spectral dimension are utilized differently to discriminate the mental activities that involve high-level language processing (e.g., Dutch participants listening to Dutch) from the processes that do not (e.g., Chinese participants listening to Dutch). Third, using GED, we extracted the most optimized source-level frequency response for each frequency bin; however, this decomposition method was limited in that we could not obtain information on where the neural activities had originated. A source-level analytical method such as the beamforming approach could be deployed to find the answer to this question. Fourth, as outlined in ***Chapters 4 and 5***, a functional connectivity analysis could be carried out to show the network-level differences in comparisons within or between different types of stimuli or participants. Lastly, our investigation was conducted only with Chinese and Dutch participants, so to

further assess the extent to which the cortical tracking effect reflects a generalized perceptual mechanism, studies using other types of language users such as German speakers could be performed.

In ***Chapters 4 and 5*** we found a neural differentiation between spoken phrases and sentences that were physically and semantically similar. Moreover, we found that this differentiation was captured in several readouts, or dimensions of brain activity. In addition, by modeling the phase-locked encoding of the acoustic features, we further showed that the brain can represent the syntactic difference between phrases and sentences in the low-frequency neural response, but that the more structured a stimulus is, the more it departs from the acoustically-driven neural response to the stimulus, even when the physicality of the stimulus is held constant. Across all our results, we provide a comprehensive picture of how the brain separates two different types of syntactic structures. However, further research is still needed to explore the relationship between these different neural readouts that index syntactic differences, e.g., how the induced neural response at the alpha band interacts with phase coherence in the low frequency range (< 8 Hz), and how these separation effects are represented at the neural source level. Finally, more extensive comparisons between additional types of syntactic structures could also be conducted in future studies.

References

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience, 28*(15), 3958-3965.

Babiloni, C., Infarinato, F., Marzano, N., Iacoboni, M., Dassù, F., Soricelli, A., Rossini, P. M., Limatola, C., & Del Percio, C. (2011). Intra-hemispheric functional coupling of alpha rhythms is related to golfer's performance: A coherence EEG study. *International Journal of Psychophysiology, 82*(3), 260-268.

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language, 61*(3), 438-456.

Becker, R., Pefkou, M., Michel, C. M., & Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Frontiers in Systems Neuroscience, 7*, 121.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences, 17*(2), 89-98. doi:10.1016/j.tics.2012.12.002

Bever, T. G., & Poeppel, D. (2010). Analysis by synthesis: a (re-) emerging program of research for language and vision. *Biolinguistics, 4*(2-3), 174-200.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355-387.

Bonhage, C. E., Meyer, L., Gruber, T., Friederici, A. D., & Mueller, J. L. (2017). Oscillatory EEG dynamics underlying automatic chunking during sentence processing. *NeuroImage, 152*, 647-657.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436.

Brennan, J. R., & Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B, 375*(1791), 20190305.

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology, 28*(24), 3976-3983. e3975.

Brown, M., Tanenhaus, M. K., & Dilley, L. (2021). Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science, 13*(2), 351-398.

Cabral, J., Kringelbach, M. L., & Deco, G. (2014). Exploring the network dynamics underlying brain activity during rest. *Progress in Neurobiology, 114*, 102-131.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science, 313*(5793), 1626-1628.

Carlin, B., & Louis, T. (2009). Bayesian Methods for Analysis.

Catford, J. C. (1988). *A Practical Introduction to Phonetics*: Clarendon Press Oxford, UK.

Chang, C.-Y., Hsu, S.-H., Pion-Tonachini, L., & Jung, T.-P. (2018). *Evaluation of artifact subspace reconstruction for automatic EEG artifact removal*. Paper presented at the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Chomsky, N. (2002). *Syntactic Structures*: Walter de Gruyter.

Chomsky, N. (2009) *Syntactic Structures*: De Gruyter Mouton.

Chomsky, N. (2014). *Aspects of the Theory of Syntax* (Vol. 11): MIT press.

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*: MIT press.

Cohen, M. X. (2015). Effects of time lag and frequency matching on phase-based connectivity. *Journal of Neuroscience Methods, 250*, 137-146.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes, 13*(1), 21-58.

Cowan, N. (2016). *Working Memory Capacity: Classic Edition*: Psychology press.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience, 10*, 604.

Crosse, M. J., & Lalor, E. C. (2014). The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech. *Journal of Neurophysiology, 111*(7), 1400-1408.

Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*: Mit Press.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*(2), 218-236.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141-201.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boudaries. *In Proceedings of Interspeech*, 2056-2056.

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

Cutting, J. E. (1974). Two left-hemisphere mechanisms in speech perception. *Perception & Psychophysics, 16*(3), 601-612.

de Bode, S., Smets, L., Mathern, G. W., & Dubinsky, S. (2015). Complex syntax in the isolated right hemisphere: Receptive grammatical abilities after cerebral hemispherectomy. *Epilepsy & Behavior, 51*, 33-39.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Di Liberto, G. M., & Lalor, E. C. (2017). Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. *Hearing Research, 348*, 70-77.

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology, 25*(19), 2457-2465.

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2017). Cortical alpha oscillations predict speech intelligibility. *Frontiers in Human Neuroscience, 11*, 88.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience, 11*, 481.

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.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews, 81*, 181-187.

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109*(29), 11854-11859.

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology, 107*(1), 78-89.

Ding, N., & Simon, J. Z. (2013a). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience, 33*(13), 5728-5735.

Ding, N., & Simon, J. Z. (2013b). Robust cortical encoding of slow temporal modulations of speech *Basic Aspects of Hearing* (pp. 373-381): Springer.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage, 85*, 761-768.

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron, 105*(2), 385-393. e389.

Doumas, L. A., & Martin, A. E. (2018). Learning structured representations from experience *Psychology of Learning and Motivation* (Vol. 69, pp. 165-203): Elsevier.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*(2), 321-335.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *The Journal of the Acoustical Society of America, 97*(3), 1893-1904.

Freedman, L. S., Lowe, D., & Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics*, 575-586.

Freeman, L., Spiegelhalter, D., & Parmar, M. (1994). The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine, 13*, 1371-1371.

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and Language, 50*(3), 259-281.

Friederici, A. D., Pfeifer, E., & Hahne, A. (1993). Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive Brain Research, 1*(3), 183-192.

Frisch, S., Schlesewsky, M., Saddy, D., & Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. *Cognition, 85*(3), B83-B92.

Fry, D. B. (1979). *The Physics of Speech*: Cambridge University Press.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336-10341.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience, 6*, 340.

Gibson, E., Marantz, A., Miyashita, Y., & O'Neil, W. (2000). Image, language, brain.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron, 56*(6), 1127-1134.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., & Simon, J. Z. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron, 77*(5), 980-991.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America, 87*(6), 2592-2605.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences, 23*(1), 1-21.

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology, 16*(2), 240-246.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics, 28*(4), 267-283.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology, 11*(12), e1001752.

Gui, P., Jiang, Y., Zang, D., Qi, Z., Tan, J., Tanigawa, H., Jiang, J., Wen, Y., Xu, L., & Zhao, J. (2020). Assessing the depth of language processing in patients with disorders of consciousness. *Nature neuroscience*, 1-10.

Gui, P., Jiang, Y., Zang, D., Qi, Z., Tan, J., Tanigawa, H., Jiang, J., Wen, Y., Xu, L., Zhao, J., Mao, Y., Poo, M. M., Ding, N., Dehaene, S., Wu, X., & Wang, L. (2020). Assessing the depth of language processing in patients with disorders of consciousness. *Nature Neuroscience, 23*(6), 761-770. doi:10.1038/s41593-020-0639-1

Gwilliams, L., & King, J.-R. (2020). Recurrent processes support a cascade of hierarchical decisions. *Elife, 9*, e56603.

Haegens, S., Osipova, D., Oostenveld, R., & Jensen, O. (2010). Somatosensory working memory performance in humans depends on both engagement and disengagement of regions in a distributed network. *Human Brain Mapping, 31*(1), 26-35.

Hagoort, P. (2013). MUC (memory, unification, control) and beyond. *Frontiers in Psychology, 4*, 416.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8*(4), 439-483.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory, 8*(2), 155-159.

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

Henin, S., Turk-Browne, N. B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., & Melloni, L. (2021). Learning hierarchical sequence representations across human cortex and hippocampus. *Science Advances, 7*(8), eabc4530.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(4), 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Hobbs, B. P., & Carlin, B. P. (2007). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics, 18*(1), 54-80.

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology, 104*(5), 2500-2511.

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., & Gonzalez-Castillo, J. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage, 80*, 360-378.

Jeans, J. H. (1968). *Science & Music*: Courier Corporation.

Jin, P., Lu, Y., & Ding, N. (2020a). Low-frequency neural activity reflects rule-based chunking during speech listening. *Elife, 9*, e55613.

Jin, P., Lu, Y., & Ding, N. (2020b). Low-frequency neural activity reflects rule-based chunking during speech listening. *Elife, 9*. doi:10.7554/eLife.55613

Jin, P., Zou, J., Zhou, T., & Ding, N. (2018a). Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature Communications, 9*(1), 5374. doi:10.1038/s41467-018-07773-y

Jin, P., Zou, J., Zhou, T., & Ding, N. (2018b). Eye activity tracks task-relevant structures during speech and auditory sequence perception. *Nature communications, 9*(1), 1-15.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15*(2), 159-201.

Kaufeld, G., Bosker, H. R., Ten Oever, S., Alday, P. M., Meyer, A. S., & Martin, A. E. (2020). Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *Journal of Neuroscience, 40*(49), 9467-9475.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS Biology, 16*(3), e2004473.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience, 30*(2), 620-628.

Khalighinejad, B., da Silva, G. C., & Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience, 37*(8), 2176-2185.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology/Revue canadienne de psychologie, 15*(3), 166.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*(5), 580-602.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: the inhibition–timing hypothesis. *Brain Research Reviews, 53*(1), 63-88.

Kothe, C. A. E., & Jung, T.-p. (2016). Artifact removal techniques with signal reconstruction: Google Patents.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*: Academic Press/Elsevier.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299-312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270-280.

Lachaux, J.-P., Rodriguez, E., Le Van Quyen, M., Lutz, A., Martinerie, J., & Varela, F. J. (2000). Studying single-trials of phase synchronous activity in

the brain. *International Journal of Bifurcation and Chaos, 10*(10), 2429-2439.

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping, 8*(4), 194-208.

Levelt, W. J., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*(1), 78-106.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93-115.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*(10), 447-454.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010. doi:10.1016/j.neuron.2007.06.004

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature Neuroscience, 4*(5), 540-545.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190.

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1-2), 71-102.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*(1), 29-63.

Martin, & Doumas, L. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology, 15*(3), e2000663. doi:10.1371/journal.pbio.2000663

Martin, & Doumas, L. A. (2019). Predicate learning in neural systems: using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences, 29*, 77-83.

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology, 7*, 120.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience, 32*(8), 1407-1427.

Martin, A. E., & Doumas, L. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B, 375*(1791), 20190306.

Meyer, L., Grigutsch, M., Schmuck, N., Gaston, P., & Friederici, A. D. (2015). Frontal–posterior theta oscillations reflect memory retrieval during sentence comprehension. *Cortex, 71*, 205-218.

Meyer, L., & Gumbert, M. (2018). Synchronization of electrophysiological responses with speech benefits syntactic information processing. *Journal of Cognitive Neuroscience, 30*(8), 1066-1074.

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex, 27*(9), 4293-4302.

Meyer, L., Obleser, J., & Friederici, A. D. (2013). Left parietal alpha enhancement during working memory-intensive sentence processing. *Cortex, 49*(3), 711-721.

Meyer, L., Obleser, J., Kiebel, S. J., & Friederici, A. D. (2012). Spatiotemporal dynamics of argument retrieval and reordering: an fMRI and EEG study on sentence processing. *Frontiers in Psychology, 3*, 523.

Meyer, L., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience, 35*(9), 1089-1099.

Morillon, B., Liégeois-Chauvel, C., Arnal, L. H., Bénar, C. G., & Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing: an intra-cortical study. *Frontiers in Psychology, 3*, 248.

Mormann, F., Lehnertz, K., David, P., & Elger, C. E. (2000). Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D: Nonlinear Phenomena, 144*(3-4), 358-369.

Morse, P. M., America, A. S. o., & Physics, A. I. o. (1948). *Vibration and Sound* (Vol. 2): McGraw-Hill New York.

Neuper, C., Wörtz, M., & Pfurtscheller, G. (2006). ERD/ERS patterns reflecting sensorimotor activation and deactivation. *Progress in Brain Research, 159*, 211-222.

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*(2), 151-165.

Nicol, J. L., Fodor, J. D., & Swinney, D. (1994). Using cross-modal lexical decision tasks to investigate sentence processing.

Norris, D., Cutler, A., McQueen, J. M., & Butterfield, S. (2006). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology, 53*(2), 146-193.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences, 23*(11), 913-926.

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex, 22*(11), 2466-2477.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *Journal of Neuroscience, 32*(36), 12376-12383.

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*.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*(6), 785-806.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and language, 34*(6), 739-773.

Palva, S., Palva, J. M., Shtyrov, Y., Kujala, T., Ilmoniemi, R. J., Kaila, K., & Näätänen, R. (2002). Distinct gamma-band evoked responses to speech and non-speech sounds in humans. *Journal of Neuroscience, 22*(4), RC211-RC211.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience, 10*(6), 717-733.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*, 320.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex, 23*(6), 1378-1387.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 539-558.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*(2), 244-247.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*(5593), 604-607.

Peña, M., & Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience, 24*(5), 1149-1164.

Pfurtscheller, G., Brunner, C., Schlögl, A., & Da Silva, F. L. (2006). Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage, 31*(1), 153-159.

Pfurtscheller, G., Neuper, C., Schlogl, A., & Lugger, K. (1998). Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters. *IEEE transactions on Rehabilitation Engineering, 6*(3), 316-325.

Phillips, C. (2003). Linear order and constituency. *Linguistic Inquiry, 34*(1), 37-90.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*(1), 245-255.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience, 21*(6), 322-334.

Poeppel, D., & Monahan, P. J. (2011). Feedforward and feedback in speech perception: Revisiting analysis by synthesis. *Language and Cognitive Processes, 26*(7), 935-951.

Puvvada, K. C., & Simon, J. Z. (2017). Cortical representations of speech in a multitalker auditory scene. *Journal of Neuroscience, 37*(38), 9189-9196.

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences, 22*(10), 870-882.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906-914.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*(4), 110-114.

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

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

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.

Schlögl, A., Lee, F., Bischof, H., & Pfurtscheller, G. (2005). Characterization of four-class motor imagery EEG data for the BCI-competition 2005. *Journal of Neural Engineering, 2*(4), L14.

Shahin, A. J., Picton, T. W., & Miller, L. M. (2009). Brain oscillations during semantic evaluation of speech. *Brain and Cognition, 70*(3), 259-266.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304.

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

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416*(6876), 87-90.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174.

Sporns, O. (2010). *Networks of the Brain*: MIT press.

Srinivasan, R., Winter, W. R., Ding, J., & Nunez, P. L. (2007). EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of Neuroscience Methods, 166*(1), 41-52.

Stam, C. J., Nolte, G., & Daffertshofer, A. (2007). Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. *Human Brain Mapping, 28*(11), 1178-1193.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*(2), 191-196.

Stevens, S., Egan, J., & Miller, G. (1947). Methods of measuring speech spectra. *The Journal of the Acoustical Society of America, 19*(5), 771-780.

Strauß, A., Wöstmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience, 8*, 350.

Strauss, E., & Wada, J. (1983). Lateral preferences and cerebral speech dominance. *Cortex, 19*(2), 165-177.

Suffczynski, P., Kalitzin, S., Pfurtscheller, G., & Da Silva, F. L. (2001). Computational model of thalamo-cortical networks: dynamical control of alpha rhythms in relation to focal attention. *International Journal of Psychophysiology, 43*(1), 25-40.

Ten Oever, S., De Weerd, P., & Sack, A. T. (2020). Phase-dependent amplification of working memory content and performance. *Nature Communications, 11*(1), 1-8.

Tikhonov, A. N., & Arsenin, V. Y. (1977). Solutions of ill-posed problems. *New York*, 1-30.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., & Kent, R. D. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America, 137*(5), 3005-3007.

Titze, I. R., & Martin, D. W. (1998). Principles of voice production: Acoustical Society of America.

Van Herten, M., Kolk, H. H., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research, 22*(2), 241-255.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America, 142*(4), 1976-1989.

Wang, Y., Zhang, J., Zou, J., Luo, H., & Ding, N. (2019). Prior knowledge guides speech segregation in human auditory cortex. *Cerebral Cortex, 29*(4), 1561-1571.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, 29-58.

Weissbart, H., Kandylaki, K. D., & Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of Cognitive Neuroscience, 32*(1), 155-166.

Wilsch, A., & Obleser, J. (2016). What works in auditory working memory? A neural oscillations perspective. *Brain Research, 1640*, 193-207.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions, 7*(1), 30.

Winter, W. R., Nunez, P. L., Ding, J., & Srinivasan, R. (2007). Comparison of the effect of volume conduction on EEG coherence with the effect of field spread on MEG coherence. *Statistics in Medicine, 26*(21), 3946-3957.

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences, 113*(14), 3873-3878.

Wöstmann, M., Herrmann, B., Wilsch, A., & Obleser, J. (2015). Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits. *Journal of Neuroscience, 35*(4), 1458-1467.

Wöstmann, M., Lim, S.-J., & Obleser, J. (2017). The human neural alpha response to speech is a proxy of attentional control. *Cerebral Cortex, 27*(6), 3307-3317.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., & Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences, 102*(7), 2293-2298.

Zhou, H., Melloni, L., Poeppel, D., & Ding, N. (2016). Interpretations of frequency domain analyses of neural entrainment: periodicity, fundamental frequency, and harmonics. *Frontiers in Human Neuroscience, 10*, 274.

English summary

Speech segmentation and syntactic representation are intriguing fields in neuroscience and psycholinguistics as they are critical and necessary steps leading to spoken language comprehension. In my doctoral thesis, I investigate the neural representation of speech segmentation via statistical inference and explore how syntactic structure discrimination could be reflected in various dimensions of neural activities.

Specifically, in ***Chapter 2***, I report on six MEG experiments with Dutch native speakers. By varying the speech stimuli so that it they are either understandable or not to participants, and by manipulating the statistical information between multiple layers of units, we found that speech segmentation or linguistic unit extraction can be conducted via statistical inference without comprehending the stimuli. More importantly, the segmentation process can be represented by cortical activity that tracks the rhythm of linguistic structures. The results provide new insights into the phenomenon of cortical tracking of hierarchical linguistic structures, and are at odds with previous views that considered the tracking effect only as a reflection of chunking using high-level linguistic knowledge. Furthermore, our findings shed light on the role of statistical learning in language acquisition. Consistent with Saffran, Aslin, and Newport (1996), our results reveal that speech segmentation via statistical inference may be an initial step in speech perception and a necessary building block in spoken language comprehension.

In ***Chapter 3***, we mainly focus on expanding our findings into a broader context and eliminating the possibility that the effects we reported in Chapter 2 were driven by the specific type of linguistic knowledge that participants had. To do so, we conducted the same sets of MEG experiments as in Chapter 2 but with Chinese participants. Our hypothesis was that if the cortical activity tracking the rhythm of multiple layers of units can reflect speech segmentation via statistical inference, then the process should be independent of participants' linguistic knowledge. As expected, the same pattern of results as in Chapter 2 was obtained, indicating that the linguistic knowledge itself did not vary the cortical tracking effect. This first supports the consistency and stability of the neural representation of speech segmentation via statistical inference across different types of language

users and indicates the insensitivity of the effect to participants' linguistic knowledge. Second, the consistent results across experiments reveals that the cortical tracking effect could reflect a generalized perceptual process, i.e., speech segmentation using statistical information, as regardless of what type of linguistic knowledge the participants had and whether they understood the stimuli, the tracking regime did not change. In sum, combining the results of Chapters 2 and 3, our findings demonstrate that speech segmentation can be performed via statistical inference, and the perceptual-level endogenous process can be reflected by cortical activities that simultaneously track the rhythm of multiple layers of linguistic structures.

Building syntactic relationships from extracted units is a necessary step in the acquisition of semantics. In complement to Chapters 2 and 3, which explored the neural representation of speech segmentation, ***Chapters 4 and 5*** report our exploratory investigations into the neural representation of syntactic structure discrimination. Both chapters used the same dataset from an EEG experiment, in which we asked participants to listen to two types of artificially synthesized speech stimuli, i.e. phrases and sentences. The stimuli were normalized in terms of both physical and semantic properties across conditions in order to highlight the differences between syntactic structures. Concretely, in ***Chapter 4***, we put weights on the temporal synchronization of the neural response in representing syntactic structure discrimination. Building on the previous studies that showed the role of low-frequency neural oscillations in speech perception and comprehension, our results indicate that syntactic structure extraction may be a feedback process in which acoustic features are encoded in an endogenous approach that is guided by high-level linguistic knowledge. In addition, our analysis using functional connectivity via phase coherence suggests that syntactic structure differences can be represented in a distributional approach. These differences were represented by the temporal synchronization of neural oscillations at a very low frequency range (< ~2 Hz). Our results are highly consistent with physiological works showing that the neural indices of syntactic structure representation have the characteristics of late latency (e.g., P600, 600 ms after the onset of the target) and right hemisphere dominance. Moreover, the results are consistent with the hypothesis proposed by Martin and Doumas (Martin & Doumas, 2017; Martin & Doumas, 2019; Martin, 2016, 2020; Martin & Doumas,

2020), which predicted that the more complicated the syntactic structure is, the more phase coherence will be evoked in the neural response. The last test described Chapter 4 was designed to check whether low frequency phase entrained with high frequency amplitude (Giraud & Poeppel, 2012) would be introduced when participants listened to the speech stimuli, and if so, whether the coupling readout would reflect syntactic structure discrimination. Consistent with the hypothesis of Giraud and Poeppel (2012), we found a strong phase-amplitude coupling (4 to 10 Hz for phase, 15 to 40 Hz for amplitude) when speech stimuli were presented for both conditions. However, no evidence was shown to indicate whether the entrainment reaches the syntactic level. The results suggest that low frequency phase entrained with high frequency amplitude could reflect generalized aspects of speech perception, such as semantic analyses of extracted units.

In ***Chapter 5***, we expand our exploration into how syntactic structure discrimination might be reflected in the intensity of neural oscillations. We modeled how acoustic features are encoded to represent syntactic structure differences. By doing so, we first found that alpha-band-induced (~8 to 13 Hz) power and connectivity could robustly reflect the syntactic discrimination between phrases and sentences. The findings could be evidence of the involvement of alpha band oscillations in the representation of syntactic structures. Furthermore, our modeling work using the STRF showed that acoustic features in both the temporal and spectral dimensions could be selectively encoded to represent the differences between syntactic structures.

In the last chapter, ***Chapter 6***, I summarize the implications of all our findings, connect the results to previous studies, and point out the directions for further research. In sum, the chapter concludes that our work has provided a comprehensive picture of the neural representation of speech segmentation via statistical inference and syntactic structure discrimination. The results in this doctoral thesis also point to the value of studying neural oscillations as an effective approach for uncovering issues in speech perception and spoken language comprehension.

Nederlandse samenvatting

Spraaksegmentatie en syntactische representaties zijn intrigerende gebieden in de neurowetenschappen en psycholinguïstiek, omdat dit cruciale en noodzakelijke stappen zijn die leiden tot het begrip van gesproken taal. In mijn proefschrift heb ik de neurale representatie van spraaksegmentatie via statistische inferentie onderzocht, en ik heb onderzocht hoe het discrimineren van syntactische structuur weerspiegeld kan worden in verschillende dimensies van neurale activiteit.

In hoofdstuk 2 rapporteer ik zes MEG-experimenten met moedertaalsprekers van het Nederlands. Door de gesproken stimuli in deze experimenten zo te variëren dat ze wel of niet te begrijpen zijn voor de proefpersonen, en door de statistische informatie tussen meerdere lagen van linguïstische eenheden te manipuleren, ontdekten we dat spraaksegmentatie, ofwel de extractie van linguïstische eenheden, uitgevoerd kan worden via statistische inferentie zonder dat de stimuli daadwerkelijk begrepen worden. Wat nog belangrijker is, is dat het segmentatieproces gerepresenteerd kan worden door corticale hersenactiviteit die het ritme van de linguïstische structuren volgt ("trackt"). De resultaten bieden nieuwe inzichten in het fenomeen van corticale tracking van hiërarchische linguïstische structuren, en zijn in strijd met eerdere opvattingen die het tracking-effect beschouwen als een weerspiegeling van een "chunking" proces dat gebruik maakt van linguïstische kennis van hogere orde. Daarnaast schijnen onze resultaten licht op de rol van statistisch leren in taalverwerving. In overeenstemming met Saffran, Aslin, en Newport (1996) laten onze resultaten zien dat spraaksegmentatie via statistische inferentie een eerste stap zou kunnen zijn in spraakperceptie, en dat het een noodzakelijke bouwsteen zou kunnen zijn in het begrijpen van gesproken taal.

In hoofdstuk 3 hebben we ons voornamelijk gericht op het uitbreiden van onze bevindingen naar een bredere context en op het uitsluiten van de mogelijkheid dat de effecten uit hoofdstuk 2 gedreven werden door een specifiek type linguïstische kennis dat de proefpersonen hadden. Hiervoor hebben we dezelfde set MEG-experimenten uitgevoerd, maar dan met Chinese proefpersonen. Onze hypothese was dat als de corticale activiteit, die het ritme van meerdere lagen van eenheden volgt, spraaksegmentatie via statistische inferentie weerspiegelt, dan zou het

proces onafhankelijk moeten zijn van de linguïstische kennis die de proefpersonen hebben. Zoals verwacht vonden we hetzelfde patroon aan resultaten als in hoofdstuk 2. Dit geeft aan dat de linguïstische kennis die de proefpersonen hadden het corticale tracking-effect niet beïnvloedde. In de eerste plaats liet het de consistentie en stabiliteit zien van de neurale representatie van spraaksegmentatie via statistische inferentie over verschillende soorten taalgebruikers, en wees het op de ongevoeligheid voor het effect van de linguïstische kennis die deelnemers hadden. Ten tweede geven de consistente resultaten in de experimenten en hoofdstukken aan dat het corticale tracking-effect een algemeen perceptueel proces zou kunnen weerspiegelen, namelijk spraaksegmentatie met behulp van statistische informatie, ongeacht of de stimuli worden begrepen en ongeacht de soorten linguïstische kennis die de deelnemers hebben. Kortom, als we de resultaten van hoofdstuk 2 en 3 combineren, wijzen onze bevindingen erop dat spraaksegmentatie zou kunnen worden uitgevoerd via statistische inferentie en dat het endogene proces op perceptueel niveau kan worden weerspiegeld door corticale activiteit die tegelijkertijd het ritme van meerdere lagen van linguïstische structuren volgt.

Het opbouwen van syntactische relaties van geëxtraheerde eenheden is een noodzakelijke stap bij het verwerven van betekenis. Als aanvulling op hoofdstuk 2 en 3, waarin de neurale representatie van spraaksegmentatie werd onderzocht, rapporteer ik in hoofdstuk 4 en 5 de bevindingen van onze verkennende onderzoeken naar de neurale representatie van het discrimineren van syntactische structuur. Beide hoofdstukken gebruikten dezelfde dataset van een EEG-experiment waarin we proefpersonen vroegen te luisteren naar twee soorten kunstmatig geproduceerde spraakstimuli, namelijk constituenten en zinnen. Om de verschillen in syntactische structuren naar voren te brengen werden de stimuli in verschillende condities op elkaar afgestemd in zowel fysieke als semantische eigenschappen. In hoofdstuk 4 benadrukten we de temporele synchronisatie van de neurale respons bij het representeren van syntactische structuurdiscriminatie. Voortbouwend op de eerdere studies die de rol van laagfrequente neurale oscillaties aantoonden in spraakperceptie en spraakbegrip, lieten onze resultaten allereerst zien dat fasecoherentie in de theta band (~ 2 tot 7 Hz) een belangrijke component is, bijvoorbeeld een lettergreep 'is' voor een zin, extractie bij het representeren van syntactische structuren gezien de fysieke gelijkenis van de

stimuli in verschillende condities. De resultaten geven aan dat extractie van syntactische structuren een feedbackproces kan zijn, waarbij akoestische kenmerken worden gecodeerd in een endogene benadering die wordt geleid door linguïstische kennis van hogere orde. Daarnaast suggereert onze analyse met behulp van functionele connectiviteit via fasecoherentie dat verschillen in syntactische structuur weergegeven zouden kunnen worden in een distributieve benadering. Verschillen in syntactische structuur werden gerepresenteerd door de temporele synchronisatie van neurale oscillaties op een zeer lage frequentie (< ~2 Hz). Onze resultaten zijn in lijn met fysiologisch onderzoek dat heeft aangetoond dat de neurale eigenschappen van de representatie van syntactische structuur de kenmerken hebben van een late respons (bijvoorbeeld de P600, die 600 ms na een target begint) en dominantie van de rechterhersenhelft. Bovendien komen de resultaten overeen met de hypothese van Martin en Doumas (2017; 2019, 2020;; Martin, 2016, 2020;, die stelt dat hoe ingewikkelder de syntactische structuur is, hoe meer fasecoherentie wordt opgeroepen in de neurale respons. Het laatste dat we in dit hoofdstuk onderzochten, was of er een koppeling ("entrainment") tussen de fase van laagfrequente activiteit en de amplitude van hoogfrequente activiteit (Giraud & Poeppel, 2012) zou ontstaan wanneer de proefpersonen naar de spraakstimuli luisterden. Zo ja, dan zou dat betekenen dat het het discrimineren van syntactische structuur weerspiegelt. In overeenstemming met de hypothese van Giraud en Poeppel (2012) vonden we een sterke fase-amplitudekoppeling (4 tot 10 Hz voor fase, 15 tot 40 Hz voor amplitude) wanneer spraakstimuli werden gepresenteerd in beide condities. Er werd echter geen evidentie gevonden die aangaf dat de koppeling een syntactisch niveau bereikt. Deze resultaten suggereren dat de koppeling tussen laagfrequente fase en hoogfrequente amplitude gegeneraliseerde aspecten van spraakperceptie zou kunnen weerspiegelen, zoals bijvoorbeeld semantische analyse van geëxtraheerde eenheden.

In hoofdstuk 5 hebben we verder onderzocht hoe het discrimineren van syntactische structuur weerspiegeld zou kunnen zijn in de intensiteit van neurale oscillaties. We hebben gemodelleerd hoe akoestische kenmerken worden gecodeerd om verschillen in syntactische structuur weer te geven. Allereerst ontdekten we dat geïnduceerde "power" en connectiviteit in de alfaband (~8 tot 13 Hz) de syntactische structuurdiscriminatie tussen constituenten en zinnen kan weerspiegelen. De bevindingen zouden kunnen wijzen op de betrokkenheid van

alfa-bandoscillaties bij de representatie van syntactische structuren. Bovendien toonde ons modelleringswerk met STRFs aan dat akoestische kenmerken in zowel de temporele als de spectrale dimensie selectief kunnen worden gecodeerd om syntactische structuurverschillen weer te geven.

In het laatste hoofdstuk, hoofdstuk 6, vatte ik de implicaties van al onze bevindingen samen, verbond ik onze resultaten met eerdere studies en gaf ik de richting aan voor verder onderzoek. Onze resultaten geven een uitgebreid beeld van de neurale representatie van spraaksegmentatie via statistische inferentie en syntactische structuurdiscriminatie. De resultaten in dit proefschrift tonen ook aan dat neurale oscillaties een effectieve benadering kunnen zijn om vraagstukken over spraakperceptie en gesproken taalbegrip te behandelen.

Acknowledgements

The furthest distance in this world is not from life to death, is from ‘making a thesis title’ to ‘drawing a conclusion’ in English. With the help of many people, I reached the last session before defending my dissertation. I would like to thank the contributors of the thesis because without you I couldn’t have finished it (so late, for some of my best friends).

Andrea Martin, my beautiful and smart daily supervisor, I would like to express my thanks to your daily guidance and help in both academics and life. I am proud of myself to be one of your PhD students, you gave me the opportunity to be a PhD student (and a postdoctoral researcher) at the Max Planck Institute and always support me whenever I need it. Your guidance in conducting research is one of my strong motivations to survive from the whole PhD stage. You were always the target of my endogenous search whenever I reached the session of answering questions in a presentation, I appreciate your super power to reset my alpha phase response during talks. Many thanks to you for answering most of the questions in our presentations. Throughout 5 years of our weekly meetings, I truly believe that you are qualified to teach people in understanding the combination of sign and Chinglish language in addition to neural oscillations and computational modelling (you are welcome). Congratulations to you for having a new cute member in your family, I am fully prepared to make an adventure to search black chicken at any time in Europe whenever Mante is given a difficult job.

Antje Meyer, my beautiful and smart promoter, thank you so much for your guidance on the thesis writing and discussing the projects with me and Andrea. It is reasonable to say that you and I are the co-first authors of the dissertation. I appreciate your huge amount of patience in correcting my thesis’ writing. Please forgive me that I sent so many revising tasks to you in the weekends and holidays. You never delayed in giving feedback and always tolerated my stupidity. I am so grateful that you would like to teach me how to become a qualified frog eater and sorry for haven’t reached that point yet. Andrea once said during the pandemic ‘… no worries, go collect your data, me and Antje will always support you …’, I actually imagined a situation where I could not find a single participant and you supported me in a way that asked Peter Hagoort to be a participant in our experiment. Antje, thank you so much for your guidance and help throughout the whole PhD stage. I’d

remember your rigorous attitudes toward research, the city center exploration, the golf playing, the delicious dinners, and especially, the early notice of the bad news (the campus would not allow smoking from October 2020 onwards) .

I would also like to appreciate the members of my thesis reading committee, Prof. Uta Noppeney, Prof. David Poeppel and Dr. A.V.M Kösem. Thanks so much for taking your time to read and evaluate my thesis.

I am really fortunate to have been part of the amazing scientific community at MPI and DCCN. Thanks to all my colleagues and friends for your positive attitude and innovative ideas. It is a colorful and precious experience working together with you. Rong Ding and Filiz Tezcan Semerci, it is my honor to have you two as my paranymphs, I am so glad to stand here together with you. I enjoy and appreciate the time talking to you and always impressed by your true personalities and unique experiences. Kevin Lam, you are a great coordinator and supervisor, I really appreciate your effort in organizing the IMPRS events. Special thanks to your tolerance and forgiveness when I missed the deadline of Checkpoints. Thanks to you for making the suggestions on selecting courses and being a nice supervisor during my journey to defense. Mante Nieuwland and Sanne ten Oever, thanks so much for reading and providing comments on our EEG paper and always nice to answer questions as senior level researchers. Rowan Sommers, you are so nice to let me join your MEG data collection and teach me how to collect data. I appreciate so much that you would like to be my participant when I failed to find anyone. I'd remember the time we collected data together and talked in person. Cas Coopmans, thanks so much for teaching me how to conduct syntactic structure decomposition and translate my thesis summary into Dutch. I am impressed by your passion and speed in conducting research. Hugo Weissbart, I appreciate your suggestions on showing acoustic similarity using TRF and being so nice when I collected MEG data. Many thanks to your suggestions on improving my English speaking. Karthikeya Kaushik, thanks for your introduction and explanation on the DORA model and your recommendation in reading papers. Ashley Lewis, thanks a lot for providing your experimental stimuli and sharing your experience for finding MEG participants. Your detailed emails were really helpful. Laurel Brehm and Sophie Slaats, I appreciate your encouragement during my presentation and always nicely expressed your suggestions on my experiments. Birgit Knudsen, I appreciate your help in making the Dutch stimuli for our MEG project and sharing your experience

in EEG data collection. I really enjoy the time talking to you in the EEG lab and group events. Annelies van Wijngaarden, thank you so much for your help in making the speech stimuli and finding student assistant. Jenny Webster, I appreciate a lot for your proofreading. Your careful and detailed checking of the thesis made me moved. I would also like to acknowledge the Max Planck Society and Radboud University for providing financial support during the whole PhD stage.

For my dear Chinese friends and family, 洁莹，真高兴在读博期间认识你，你来了以后咱几乎天天一起吃午饭，聊天，谢谢你跟我分享各种生活和学习上遇到的事，师兄是个不合格的师兄，但师妹绝对是个好师妹，在我做 MEG 实验的特殊时段，你一点不犹豫的帮我，又一起测试程序又当被试。你总能在我需要人帮忙的时候出现，师兄也没帮不上你啥忙，对不起啊。师兄支持你做的选择，你肯定能实现自己的想法，祝你之后一路顺利，遇到的困难都逐一克服。感谢周末一起吃饭聊天的好朋友们, 有你们的陪伴让我和老婆在这期间的生活丰富多彩。感谢家人的支持，尤其老婆的支持。

Curriculum vitae

Fan Bai was born in Baicheng, China, in 1986. He obtained his master's degree in fundamental psychology from Northeast Normal University in China. In September 2017, Fan started his PhD project funded by IMPRS at the Max Planck Institute for Psycholinguistics. He is now a postdoctoral researcher at Donders Institute for Brain, Cognition and Behaviour.

Publications

Tian, X., Ding, N., Teng, X., Bai, F., & Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. Nature Human Behaviour, 2(3), 225-234.

Bai, F., Meyer, A. S., & Martin, A. E. (2022). Neural dynamics differentially encode phrases and sentences during spoken language comprehension. PLoS Biology, 20(7), e3001713.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. Miranda I. van Turennout

2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. Niels O. Schiller

3. Lexical access in the production of ellipsis and pronouns. Bernadette M. Schmitt

4. The open-/closed class distinction in spoken-word recognition. Alette Petra Haveman

5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. Kay Behnke

6. Gesture and speech production. Jan-Peter de Ruiter

7. Comparative intonational phonology: English and German. Esther Grabe

8. Finiteness in adult and child German. Ingeborg Lasser

9. Language input for word discovery. Joost van de Weijer

10. Inherent complement verbs revisited: Towards an understanding of argu-ment structure in Ewe. James Essegbey

11. Producing past and plural inflections. Dirk J.Janssen

12. Valence and transitivity in Saliba: An Oceanic language of Papua New Gui-nea. Anna Margetts

13. From speech to words. Arie H. van der Lugt

14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. Eva Schultze-Berndt

15. Interpreting indefinites: An experimental study of children's language com-prehension. Irene Krämer

16. Language-specific listening: The case of phonetic sequences. Andrea Chri-stine Weber

17. Moving eyes and naming objects. Femke Frederike van der Meulen

18. Analogy in morphology: The selection of linking elements in Dutch compounds. Andrea Krott

19. Morphology in speech comprehension. Kerstin Mauth

20. Morphological families in the mental lexicon. Nivja Helena de Jong

21. Fixed expressions and the production of idioms. Simone Annegret Sprenger

22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria. Birgit Hellwig

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. Fermín Moscoso del Prado Martín

24. Contextual influences on spoken-word processing: An electrophysiological approach. Danielle van den Brink

25. Perceptual relevance of prevoicing in Dutch. Petra Martine van Alphen

26. Syllables in speech production: Effects of syllable preparation and syllable frequency. Joana Cholin

27. Producing complex spoken numerals for time and space. Marjolein Hen-riëtte Wilhelmina Meeuwissen

28. Morphology in auditory lexical processing: Sensitivity to fine phonetic de-tail and insensitivity to suffix reduction. Rachèl Jenny Judith Karin Kemps

29. At the same time. . . : The expression of simultaneity in learner varieties. Barbara Schmiedtová

30. A grammar of Jalonke argument structure. Friederike Lüpke

31. Agrammatic comprehension: An electrophysiological approach. Marijtje Elizabeth Debora Wassenaar

32. The structure and use of shape-based noun classes in Miraña (North West Amazon). Frank Seifart

33. Prosodically-conditioned detail in the recognition of spoken words. Anne Pier Salverda

34. Phonetic and lexical processing in a second language. Mirjam Elisabeth Broersma

35. Retrieving semantic and syntactic word properties: ERP studies on the time course in language comprehension. Oliver Müller

36. Lexically-guided perceptual learning in speech processing. Frank Eisner

37. Sensitivity to detailed acoustic information in word recognition. Keren Ba-tya Shatzman

38. The relationship between spoken word production and comprehension. Rebecca Özdemir

39. Disfluency: Interrupting speech and gesture. Mandana Seyfeddinipur

40. The acquisition of phonological structure: Distinguishing contrastive from non-constrative variation. Christiane Dietrich

41. Cognitive cladistics and the relativity of spatial cognition. Daniel Haun

42. The acquisition of auditory categories. Martijn Bastiaan Goudbeek

43. Affix reduction in spoken Dutch: Probabilistic effects in production and perception. Mark Pluymaekers

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. Valesca Madalla Kooijman

45. Space and iconicity in German sign language (DGS). Pamela M. Perniss

46. On the production of morphologically complex words with special attenti-on to effects of frequency. Heidrun Bien

47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. Amanda Brown

48. The acquisition of verb compounding in Mandarin Chinese. Jidong Chen

49. Phoneme inventories and patterns of speech sound perception. Anita Eva Wagner

50. Lexical processing of morphologically complex words: An information-theoretical perspective. Victor Kuperman

51. A grammar of Savosavo: A Papuan language of the Solomon Islands. Clau-dia Ursula Wegener

52. Prosodic structure in speech production and perception. Claudia Kuzla

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. Sarah Schimke

54. Studies on intonation and information structure in child and adult Ger-man. Laura de Ruiter

55. Processing the fine temporal structure of spoken words. Eva Reinisch

56. Semantics and (ir)regular inflection in morphological processing. Wieke Tabak

57. Processing strongly reduced forms in casual speech. Susanne Brouwer

58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. Miriam Ellert

59. Lexical interactions in non-native speech comprehension: Evidence from electroencephalography, eye-tracking, and functional magnetic resonance imaging. Ian FitzPatrick

60. Processing casual speech in native and non-native language. Annelie Tuin-man

61. Split intransitivity in Rotokas, a Papuan language of Bougainville. Stuart Payton Robinson

62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. Sonja Gipper

63. The influence of information structure on language comprehension: A neurocognitive perspective. Lin Wang

64. The meaning and use of ideophones in Siwu. Mark Dingemanse

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. Marco van de Ven

66. Speech reduction in spontaneous French and Spanish. Francisco Torreira

67. The relevance of early word recognition: Insights from the infant brain. Caroline Mary Magteld Junge

68. Adjusting to different speakers: Extrinsic normalization in vowel percepti-on. Matthias Johannes Sjerps

69. Structuring language: Contributions to the neurocognition of syntax. Ka-trien Rachel Segaert

70. Infants' appreciation of others' mental states in prelinguistic communica-tion: A second person approach to mindreading. Birgit Knudsen

71. Gaze behavior in face-to-face interaction. Federico Rossano

72. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. Connie de Vos

73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. Attila Andics

74. Lexical processing of foreign-accented speech: Rapid and flexible adapta-tion. Marijt Witteman

75. The use of deictic versus representational gestures in infancy. Daniel Puc-cini

76. Territories of knowledge in Japanese conversation. Kaoru Hayano

77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. Kimberley Mulder

78. Contributions of executive control to individual differences in word pro-duction. Zeshu Shao

79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. Patrick van der Zande

80. High pitches and thick voices: The role of language in space-pitch associa-tions. Sarah Dolscheid

81. Seeing what's next: Processing and anticipating language referring to ob-jects. Joost Rommers

82. Mental representation and processing of reduced words in casual speech. Iris Hanique

83. The many ways listeners adapt to reductions in casual speech. Katja Pöll-mann

84. Contrasting opposite polarity in Germanic and Romance languages: Ver-um Focus and affirmative particles in native speakers and advanced L2 learners. Giuseppina Turco

85. Morphological processing in younger and older people: Evidence for flexi-ble dual-route access. Jana Reifegerste

86. Semantic and syntactic constraints on the production of subject-verb agree-ment. Alma Veenstra

87. The acquisition of morphophonological alternations across languages. He-len Buckler

88. The evolutionary dynamics of motion event encoding. Annemarie Verkerk

89. Rediscovering a forgotten language. Jiyoun Choi

90. The road to native listening: Language-general perception, language-specific input. Sho Tsuji

91. Infants'understanding of communication as participants and observers. Gud-mundur Bjarki Thorgrímsson

92. Information structure in Avatime. Saskia van Putten

93. Switch reference in Whitesands. Jeremy Hammond

94. Machine learning for gesture recognition from videos. Binyam Gebrekidan Gebre

95. Acquisition of spatial language by signing and speaking children: A com-parison of Turkish sign language (TID) and Turkish. Beyza Sumer

96. An ear for pitch: On the effects of experience and aptitude in processing pitch in language and music. Salomi Savvatia Asaridou

97. lncrementality and Flexibility in Sentence Production. Maartje van de Velde

98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) ani-mal culture. Edwin van Leeuwen

99. The request system in Italian interaction. Giovanni Rossi

100. Timing turns in conversation: A temporal preparation account. Lilla Magya-ri

101. Assessing birth language memory in young adoptees. Wencui Zhou

102. A social and neurobiological approach to pointing in speech and gesture. David Peeters

103. Investigating the genetic basis of reading and language skills. Alessandro Gialluisi

104. Conversation electrified: The electrophysiology of spoken speech act recognition. Rósa Signý Gisladottir

105. Modelling multimodal language processing. Alastair Charles Smith

106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. Florian Hintz

107. Situational variation in non-native communication. Huib Kouwenhoven

108. Sustained attention in language production. Suzanne Jongman

109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntactic, and orthographic effects. Malte Viebahn

110. Nativeness, dominance, and the flexibility of listening to spoken language. Laurence Bruggeman

111. Semantic specificity of perception verbs in Maniq. Ewelina Wnuk

112. On the identification of FOXP2 gene enhancers and their role in brain development. Martin Becker

113. Events in language and thought: The case of serial verb constructions in Avatime. Rebecca Defina

114. Deciphering common and rare genetic effects on reading ability. Amaia Carrión Castillo

115. Music and language comprehension in the brain. Richard Kunert

116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. Nietzsche H.L. Lam

117. The biology of variation in anatomical brain asymmetries. Tulio Guadalupe

118. Language processing in a conversation context. Lotte Schoot

119. Achieving mutual understanding in Argentine Sign Language. Elizabeth Manrique

120. Talking sense: the behavioural and neural correlates of sound symbolism. Gwilym Lockwood

121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. Franziska Hartung

122. Sensorimotor experience in speech perception. Will Schuerman

123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. Ashley Lewis

124. Influences on the magnitude of syntactic priming. Evelien Heyselaar

125. Lapse organization in interaction. Elliott Hoey

126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. Sophie Brand

127. The neighbors will tell you what to expect: effects of aging and predictability on language processing. Cornelia Moers

128. The role of voice and word order in incremental sentence processing. Studies on sentence production and comprehension in Tagalog and German. Sebastian Sauppe

129. Learning from the (un)expected: age and individual differences in statistical learning and perceptual learning in speech. Thordis Neger

130. Mental representations of Dutch regular morphologically complex neologisms. Laura de Vaan

131. Speech production, perception, and input of simultaneous bilingual pre-schoolers: Evidence from voice onset time. Antje Stoehr

132. A holistic approach to understanding pre-history. Vishnupriya Kolipakam

133. Characterization of transcription factors in monogenic disorders of speech and language. Sara Busquets Estruch

134. Indirect request comprehension in different contexts. Johanne Tromp

135. Envisioning language: An exploration of perceptual processes in language comprehension. Markus Ostarek

136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. Juliane Kirsch

137. Let the agents do the talking: On the influence of vocal tract anatomy on speech during ontogeny and glossogeny. Rick Janssen

138. Age and hearing loss effects on speech processing. Xaver Koch

139. Vocabulary knowledge and learning: Individual differences in adult native speakers. Nina Mainz

140. The face in face-to-face communication: Signals of understanding and non-understanding. Paul Hömke

141. Person reference and interaction in Umpila/Kuuku Ya'u narrative. Clair Hill

142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. Jana Bašnáková

143. From Kawapanan to Shawi: Topics in language variation and change. Luis Miguel Rojas Berscia

144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. Linda Drijvers

145. Linguistic dual-tasking: Understanding temporal overlap between production and comprehension. Amie Fairs

146. The role of exemplars in speech comprehension. Annika Nijveld

147. A network of interacting proteins disrupted in language-related disorders. Elliot Sollis

148. Fast speech can sound slow: Effects of contextual speech rate on word recognition. Merel Maslowski

149. Reasons for every-day activities. Julija Baranova

150. Speech planning in dialogue - Psycholinguistic studies of the timing of turn taking. Mathias Barthel

151. The role of neural feedback in language unification: How awareness affects combinatorial processing. Valeria Mongelli

152. Exploring social biases in language processing. Sara Iacozza

153. Vocal learning in the pale spear-nosed bat, Phyllostomus discolor. Ella Lattenkamp

154. Effect of language contact on speech and gesture: The case of Turkish-Dutch bilinguals in the Netherlands. Elif Zeynep Azar

155. Language and society: How social pressures shape grammatical structure Limor Raviv

156. The moment in between: Planning speech while listening. Svetlana-Lito Gerakaki

157. How speaking fast is like running: Modelling control of speaking rate. Joe Rodd

158. The power of context: How linguistic contextual information shapes brain dynamics during sentence processing. René Terporten

159. Neurobiological models of sentence processing. Marvin Uhlmann

160. Individual differences in syntactic knowledge and processing: The role of literacy experience. Saoradh Favier

161. Memory for speaking and listening. Eirini Zormpa

162. Masculine generic pronouns: Investigating the processing of an uninten-ded gender cue. Theresa Redl

163. Properties, structures and operations: Studies on language processing in the brain using computational linguistics and naturalistic stimuli. Alessan-dro Lopopolo

164. Investigating spoken language comprehension as perceptual inference. Greta Kaufeld

165. What was that Spanish word again? Investigations into the cognitive mechanisms underlying foreign language attrition. Anne Mickan

166. A tale of two modalities: How modality shapes language production and visual attention. Francie Manhardt

167. Why do we change how we speak? Multivariate genetic analyses of language and related traits across development and disorder. Ellen Verhoef

168. Variation in form and meaning across the Japonic language family with a focus on the Ryukyuan languages. John Huisman

169. Bilingual sentence production and code-switching: Neural network simulations. Chara Tsoukala

170. Effects of aging and cognitive abilities on multimodal language production and comprehension in context. Louise Schubotz

171. Speaking while listening: Language processing in speech shadowing and translation. Jeroen van Paridon

172. Left-right asymmetry of the human brain: Associations with neurodevelopmental disorders and genetic factors. Merel Postema

173. Abstract neural representations of language during sentence comprehension: Evidence from MEG and Behaviour. Sophie Arana

174. Infants' perception of sound patterns in oral language play. Laura Hahn

175. Spoken and written word processing: Effects of presentation modality and individual differences in experience to written language. Merel Wolf

176. Development of Spatial Language and Memory: Effects of Language Modality and Late Sign Language Exposure. Dilay Karadoller

177. Kata Kolok - Variation and acquisition. Hannah Lutzenberger

178. Individual differences in speech production and maximum speech performance. Chen Shen

179. Non-native phonetic accommodation in interactions with humans and with computers. Aurora Troncoso Ruiz

180. About time: Exploring the role of grammatical aspect in event cognition. Julia Misersky

181. Non-native Lombard speech: The acoustics, perception, and comprehension of English Lombard speech by Dutch natives. Katherine Marcoux

182. The enlanguaged brain: Cognitive and neural mechanisms of linguistic influence on perception. Ksenija Slivac

183. Discovering the units in language cognition: From empirical evidence to a computational model. Jinbiao Yang

184. Neural representation of speech segmentation and syntactic structure discrimination. Fan Bai