Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Registered Report

Frequency tagging of syntactic structure or lexical properties; a registered MEG study

Evgenii Kalenkovich [a,*], Anna Shestakova [c] and Nina Kazanina [b,c]

[a] HSE University, Centre for Cognition and Decision Making, Institute for Cognitive Neuroscience, National Research University Higher School of Economics, Russian Federation
[b] University of Bristol, School of Psychological Science, Bristol, UK
[c] International Laboratory of Social Neurobiology, Institute of Cognitive Neuroscience, National Research University Higher School of Economics, Moscow, Russia

ARTICLE INFO

Article history:
Received 3 July 2021
Reviewed 9 August 2021
Revised 22 September 2021
Accepted 30 September 2021
Action editor Pia Rotshtein
Published online 20 October 2021

Keywords:
Sentence comprehension
Syntactic processing
Hierarchical sentence structure
MEG
Frequency tagging
Distributional semantic vectors

ABSTRACT

A traditional view on sentence comprehension holds that the listener parses linguistic input using hierarchical syntactic rules. Recently, physiological evidence for such a claim has been provided by Ding et al.'s (2016) MEG study that demonstrated, using a frequency-tagging paradigm, that regularly occurring syntactic constituents were spontaneously tracked by listeners. Even more recently, this study's results have been challenged as artifactual by Frank and Yang (2018) who successfully re-created Ding's results using a distributional semantic vector model that relied exclusively on lexical information and did not appeal to any hierarchical syntactic representations. The current MEG study was designed to dissociate the two interpretations of Ding et al.'s results. Taking advantage of the morphological richness of Russian, we constructed two types of sentences of different syntactic structure; critically, this was achieved by manipulating a single affix on one of the words while all other lexical roots and affixes in the sentence were kept the same. In Experiment 1, we successfully verified the intuition that due to almost complete lexical overlap the two types of sentences should yield the same activity pattern according to Frank and Yang's (2018) lexico-semantic model. In Experiment 2, we recorded Russian listeners' MEG activity while they listened to the two types of sentences. Contradicting the hierarchical syntactic account and consistent with the lexico-semantic one, we observed no difference across the conditions in the way participants tracked the stimuli properties. Corroborated by other recent evidence, our findings show that peaks interpreted by Ding et al. as reflecting higher-level syntactic constituency may stem from non-syntactic factors.

© 2021 Elsevier Ltd. All rights reserved.

* *Corresponding author*. 3, Krivokolenny per., Moscow, Russia.
E-mail address: ekalenkovich@hse.ru (E. Kalenkovich).

https://doi.org/10.1016/j.cortex.2021.09.012
0010-9452/© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

Language is a primary means of communicating information among humans. A key reason for the success of language in this task lies in its productive power, i.e. the speaker's ability to generate new sentences that express novel ideas combined with the listeners' ability to comprehend them. What makes it possible to understand novel, previously unheard sentences? A dominant linguistic theory (Berwick & Weinberg, 1984, p. 325; Chomsky, 2002, p. 117; Everaert, Huybregts, Chomsky, Berwick, & Bolhuis, 2015) proposes that this ability is underpinned by the hierarchical syntactic rules shared by the speaker and the listener. On such theory, combining words into units known as “syntactic constituents” is a critical step in sentence comprehension.

In a recent study, Ding, Melloni, Zhang, Tian, and Poeppel (2016) presented physiological evidence that listeners automatically “extract” hierarchical syntactic structure. In a magnetoencephalographic (MEG) study using a frequency-tagging paradigm, participants listened to sequences of words. These sequences could be either representing a series of individual monosyllabic words (e.g., *black went must from* ...), or parsed into larger syntactic constituents such as 2-syllable phrases (e.g., *new plans big box* ...; “phrase condition”) or sentences consisting of two 2-syllable phrases (e.g. *new plans give hope, big fish escaped* ...; “sentence condition”). Across conditions, the syllables were presented isochronously (e.g., each syllable lasted exactly 250 ms); all prosodic cues were controlled for so that the stream of words had no prosodic cues to phrase or sentence boundaries. All conditions, expectedly, yielded a 4 Hz peak in the participants' MEG power spectrum, corresponding to the regular rate of syllable presentation (Fig. 1a). Critically, the phrase condition featured an additional peak at 2 Hz corresponding to the phrase rate (Fig. 1b); the sentence condition yielded a 4 Hz syllable peak, a 2 Hz phrase peak corresponding to the phrase rate (e.g. *new plans, give hope*) and a 1 Hz peak corresponding to the sentence rate (*new plans give hope*; Fig. 1c). Unlike the 4 Hz syllable peak, the 2 Hz phrase and 1 Hz sentence peaks could not have emerged due to any regularity in the acoustic signal and instead were considered to reflect the listener's syntactic knowledge which results in automatic parsing of a word stream into syntactically meaningful constituents whenever such are available.

Frank and Yang (2018) challenged Ding et al.’s (2016) interpretation of the findings and proposed an alternative explanation formulated in purely lexical terms without involving syntax. They simulated activity elicited by the stimuli used in Ding et al. (2016) using distributional semantic vectors (Mikolov, Chen, Corrado, & Dean, 2013). For each word, they constructed a ~300-dimensional distributional semantic vector based on a *word2vec* model (Mikolov et al., 2013). For each word, a distributional semantic vector is calculated on the basis of the words that surround the word in question in a large text corpus (and without recourse to higher-level syntactic representations such as phrases and sentences) so that words that occur in similar lexical contexts receive similar vectors. When each word (or syllable in case of the multisyllabic Chinese stimuli) in Ding et al.’s sequences was simulated by its corresponding distributional semantic vector, the power spectrum for each condition proved to be qualitatively identical to the neural spectral data in Ding et al. (2016). Thus, Frank and Yang (2018) provided an alternative account for Ding et al.’s neural findings whereby the observed spectral peaks are accounted by lexical properties of the stimuli rather than their hierarchical syntactic features.

As noted by Frank and Yang, a likely reason behind successful replication of the neural data via distributional semantic vector approach lies in the nature of Ding et al.’s experimental materials. For example, in Ding et al.’s English sentence condition, every other word is a noun that typically denotes an entity and every fourth word is a transitive verb that denotes an action. Because words that share syntactic and/or semantic properties tend to have a similar surrounding lexical context, their corresponding distributional semantic vectors are also similar. As a result of similar vectors occurring regularly, the power spectrum of the simulated sentence condition shows spectral peaks reflecting these regularities.

In order to distinguish lexical vs. syntactic-level contribution to neural tracking one needs to test conditions for which predictions based on distributional semantics and those based on syntactic constituency are distinct. Languages with rich morphology, in which a superficially subtle manipulation of a single suffix added to the same lexical root can have a considerable effect on syntactic structure, come to help. Consider, for example, minimally distinct sentences of Russian in Fig. 2. The two conditions are identical in terms of the lexical roots and most of affixes; the difference lies in a single phoneme, i.e. the final phoneme of the second noun, which is a suffix marking the case of the noun: the noun's case is either genitive (*Diny* “of-Dina”) or dative (*Dine* “for-Dina”).

The case of the second noun has considerable effects on the syntactic structure of the whole sentence. The Genitive condition is “symmetrical” in that it contains two two-word phrases, i.e., the noun phrase (NP) *povar Diny* “Dina's cook” and the verb phrase (VP) *pechot bliny* “is making pancakes”; we will refer to it as “2-2” condition hereinafter. The Dative condition, referred to as “1–3” condition, consists of a one-word NP *povar* “cook” and a 3-word VP *Dine pechot bliny* “is making pancakes for Dina”. Importantly, neither condition involves a garden-path effect. In the Genitive condition, upon encountering *Diny*^{GEN} the listeners can immediately form a noun phrase ([$_{NP}$ *povar Diny*]). In the Dative condition, *Dine*^{DAT} cannot be attached to povar to form an NP; rather a VP has to be projected [$_{NP}$ *povar*] [$_{VP}$ *Dine* ...].

Yet, because the roots and all affixes except the case marker on the second noun (*Dina*) are identical, it is likely that the model based on distributional semantics vectors yields similar results for the two conditions.

In Experiment 1, we modeled the activity for a sequential stream consisting of multiple sentences from either Genitive or Dative condition in Fig. 2 using distributional semantic vectors. This computational simulation has already been completed, as its outcome is a pre-requisite for the MEG experiment. Anticipating the findings from Experiment 1, the modeling yielded peaks at the syllable, word, 2-word and sentence rates. Most importantly, the 2-word peak in the Genitive condition was not larger than in the Dative condition. That is, higher-level groupings emerging on the basis of

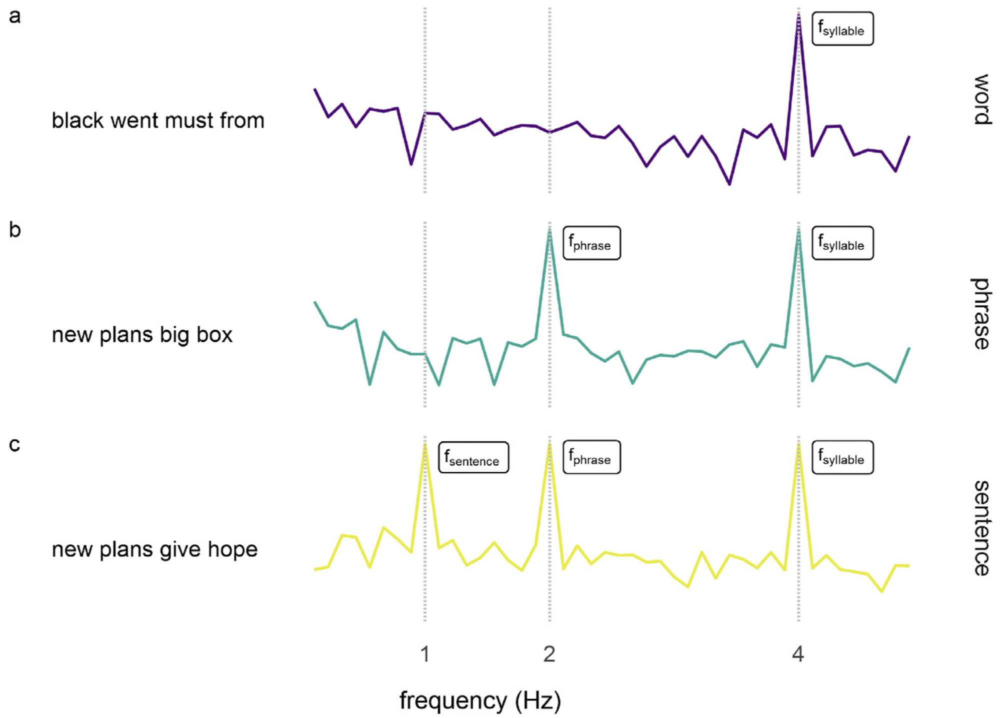

Fig. 1 – Visualization of the results from Ding et al. (2016): peaks in power spectra correspond exactly to the acoustic (syllable) and syntactic (phrase, sentence) units present in the stimuli. NB: the data are not original, the graph is for illustrative purposes only.

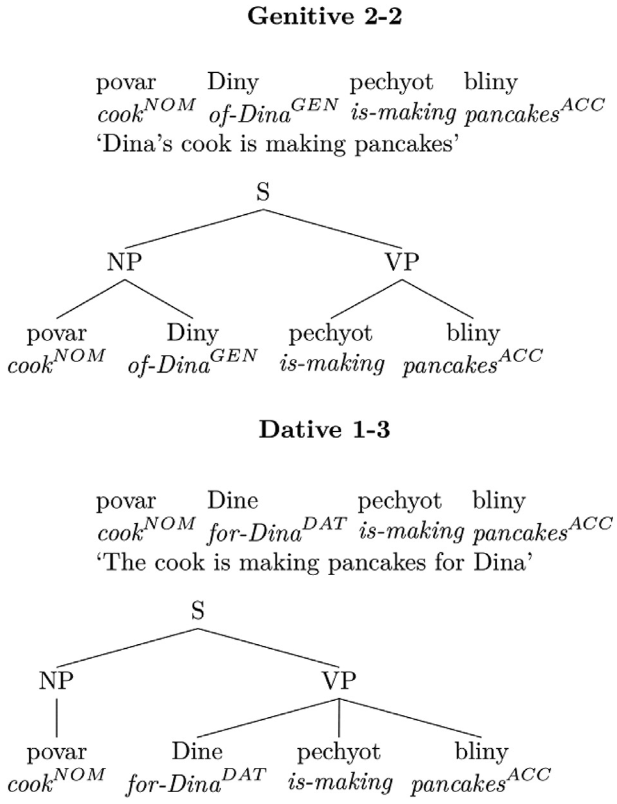

Fig. 2 – A sample set of sentences containing a Genitive and Dative condition.

lexical properties of the stimuli in the Genitive and Dative conditions did not mimic the syntactic constituency, making it possible to dissociate syntactic and distributional semantic accounts.

In Experiment 2, we used MEG to record brain activity of native Russian listeners while they listened to the same streams of sentences as in Experiment 1, presented isochronously at the syllable rate of 3.125 Hz (=320 ms/syllable). In both conditions, we expected a peak at the syllable rate of 3.125 Hz reflecting a regular nature of the auditory stimulus, a peak at the word rate of 1.56 Hz reflecting regular occurrence of lexical items (as in Makov et al., 2017 who found a word peak for bisyllabic words), and a peak at the sentence rate of .39 Hz reflecting regularity at the 4-word level (either because a sentence-sized syntactic constituent is built or due to regular occurrence of grammatical categories, e.g., every fourth word being a transitive verb). Critically, only if listeners entrained to the syntactic structure, we would have expected a 2-word peak at .78 Hz in the Genitive 2-2 condition that would be significantly stronger than in the Dative 1–3 condition, reflecting regular occurrence of phrases in the former but not the latter condition. Several subjects were recorded before Stage 1 of this report, which enabled us to verify the quality of our setup without checking the critical difference between the conditions described above. These preliminary results are described in the *Pilot data* section. Results based on the full dataset are described in the *Results* section under *Experiment 2*.

2. Experiment 1: computational simulation using the model by Frank and Yang (2018)

2.1. Materials

Sixty-four sentence sets of two conditions such as in Fig. 2 were constructed. All of the sentences consisted of four bisyllabic words and followed the pattern *Noun1* + *Noun2* + *Verb*+*Noun3*. *Noun1* was always nominative, *Noun3* was always accusative and served as a direct object of the transitive verb that preceded it. *Noun2* was a proper or common noun in either genitive (*Genitive 2-2* condition) or dative case (*Dative 1–3* condition). *Noun2* words were selected so that the case marking was unambiguously distinguishable phonologically and orthographically, i.e., we did not include nouns for which genitive and dative case-marked suffixes sound similar due to phonological reduction (such as the name *Petja* "Pete" for which the genitive ($Peti^{GEN}$) and dative ($Pete^{DAT}$) case forms are both pronounced as [peti]). Each word appeared in exactly one sentence pair.

Four sentence pairs that received the lowest scores as a result of auditory pre-screening (see section *Generation and pre-screening of auditory stimuli* under *Experiment 2*) were excluded. Experiment 1 employed the remaining 60 sets of sentences.

2.2. Simulation

For the simulation, we closely followed the procedure in Frank and Yang (2018). Twelve participants were simulated. For each participant, all sentences from each condition were reshuffled to produce a 60 sentence long sequence. The sequences were thus 480 syllables long (60 sentences × 4 words × 2 syllables = 480 syllables).

Next, each 480-syllable long sequence had to be represented as a chain of distributional semantic vectors. This was done by imitating the process of word segmentation as performed by the human brain exposed to the same syllable sequence auditorily, at an isochronous syllable rate without any cues to word boundaries. Following Frank and Yang (2018), the procedure was as follows: the first syllable s_1 in the 480-long syllable sequence (s_1, s_2, ...,s_{480}) activated a cohort of words that start with that initial syllable with each word activated in proportion to its frequency. Then the next syllable from the sequence was added to yield s_1s_2, and the cohort was reduced to only those words that started with that string; the procedure repeated for as long as the sequence s_1s_2 ... s_k yielded a non-empty cohort. Once the cohort was empty, the segmentation procedure restarted from the last syllable, i.e. s_k. To exemplify using an English example *can-dy-mel-ted* (*candy melted*), when the syllable *can* becomes available it activates a cohort that includes words *can*, *candy*, *candle*, *canton*, *candid*, *cantaloupe*, *candidate*, *candyfloss* among others. Once *can-dy* becomes available, the cohort is reduced to *candy* and *candyfloss* and few other words; when the following syllable is added, the resulting string *can-dy-mel* results in an empty word cohort. The segmentation process then restarts from the last syllable (*mel*). When applied to our Russian materials, 97% of the word boundaries had been identified correctly; incorrect identifications were left as they were (i.e. not fixed manually).

Each cohort in the sequence produced by the step above was then represented by the frequency-weighted sum of the distributional semantic vectors of the words it contained. We used vectors from the Russian Distributional Thesaurus project (Panchenko, Ustalov, et al., 2017) which we downloaded from a publicly available dataset (Panchenko, Arefyev, et al., 2017). The authors trained a skipgram word2vec model (Mikolov et al., 2013) on a corpus of Russian-language books (13 billion words). In a skipgram word2vec model, a neural network with a single hidden layer is trained to predict neighbors of each word within a surrounding context window of a set size (e.g. within 5 words to the left or right from the word). After training, the weights connecting a word to the hidden layer are taken as the word's distributional semantic vector. The training was done using the skip-gram model with 500 units in the hidden layer and a context window size of 10 words and iterated 3 times.

The segmentation procedure was applied to each participant's sequences (one for each condition). Each syllable was mapped to a 500-dimensional vector **v** representing the distributional vector of a cohort activated upon hearing that syllable. In order to represent the temporal dynamics of the syllable stream in a way comparable to the auditory presentation (as in the subsequent Experiment 2), vector **w**(*t*) representing activation elicited at time *t* from the syllable onset was modeled as background noise from which **v** emerges after time τ ms (time τ was randomly drawn from a uniform distribution with range 40 ± 25). For each dimension *i* out of 500 and for each *t* between 1 ms and 320 ms (320 ms is the syllable duration in the MEG experiment) the activation was calculated as follows:

$$w_i(t) = \begin{cases} \varepsilon_i(t) & \text{if } t < \tau \\ \varepsilon_i(t) + v_i(t) & \text{if } t \geq \tau \end{cases}$$

where $\varepsilon_i(t)$ was normally distributed with $\mu = 0$, $\sigma = .1$.

For each participant and each condition, the procedure described above yielded a 500 × 153,600 matrix, where 153,600 = 60 sentences × 8 syllables × 320 ms is the duration of each condition stream in milliseconds.

2.3. Data analysis

Matrices output by the simulation were Fourier-transformed along the time dimension, only the coefficients above DC and below 5 Hz were retained. The squared norms of the resulting coefficients were then averaged along the 500-long dimension to calculate average power at each frequency bin.

To quantify the statistical significance of any peaks in the resulting power spectra, we calculated signal-to-noise ratios (SNR) by dividing power at each frequency bin by the mean power at the four immediately neighboring frequency bins (two on each side). A flat spectrum corresponds to SNR close to 1. Peak presence at each frequency bin was tested by comparing whether the normalized SNR is significantly larger than 1 using a one-sample one-tail *t*-test. False discovery rate (FDR, Benjamini & Hochberg, 1995) correction was applied with .001 as the significance level.

2.4. Results

Simulated power spectra for the Genitive 2-2 and Dative 1–3 conditions are shown in Fig. 3 and clearly feature peaks at frequencies corresponding to the syllable, word, two-word and sentence rates, as well as their harmonics.

We compared SNRs at the syllable (3.12 Hz) and word (1.56 Hz) rates across conditions using paired-samples two-tail *t*-tests. The syllable peaks predictably did not differ ($M_d = -.47$, 95% CI [-3.34, 2.39], $t(11) = -.36$, $P = .723$); the word peaks, however, did differ ($M_d = 1.99$, 95% CI [1.51, 2.48], $t(11) = 9.04$, $P < .001$) due to differences in the distributional semantic vectors corresponding to genitive and dative forms of *Noun2*. Because the amplitude of the word peak in each condition might have influenced the amplitude of higher (2-word and sentence) peaks in the SNR spectra, leading to differences that are irrelevant to the critical manipulation, we normalized all SNRs by the SNR at the word frequency. This was done in the logarithmic space to keep the baseline noise SNR at 1 (see Equation (1)).

$$SNR_{norm} = \exp\left(\frac{\ln SNR}{\ln SNR_{word}}\right) \quad (1)$$

Normalized SNRs, shown in Fig. 4, significantly differed from 1 at frequencies corresponding to the syllable, word, 2-word and sentence rates, as well as their harmonics (all $P < .001$). Importantly, the word-normalized 2-word peak in the Genitive 2-2 condition was not larger than in the Dative 1–3 condition ($M_d = .05$, 95% CI [$-\infty$, .09], $t(11) = 2.53$, $P = .986$).

2.5. Discussion

Modeling using distributional semantic vectors showed a similar pattern of activity at frequencies corresponding to 2-word combinations and sentences in the Genitive 2-2 and Dative 1–3 conditions. Thus this pair of conditions presents a case that can help to dissociate whether neural tracking can be explained by a model that solely relies on word-level statistics, or whether recourse to hierarchical syntactic structure is needed. In the former case, in accordance with the outcomes of the current simulation, a statistically indistinguishable neural response is expected for the two conditions in Experiment 2 with human participants, most critically at the .78 Hz frequency rate corresponding to 2-word combinations. On the other hand, difference in the .78 Hz response, that corresponds to the rate of phrases in the Genitive but not in the Dative condition, is expected on the syntactic view.

3. Experiment 2–the MEG experiment

3.1. Methods

3.1.1. Participants

According to our sequential sampling plan (see *below*), we planned to collect enough participants so that the first statistical test was run on data from 20 participants. Due to technical reasons and the recruitment procedure whereby participants were recorded in blocks of up to six participants, the test was first run on data from 27 participants (we additionally report

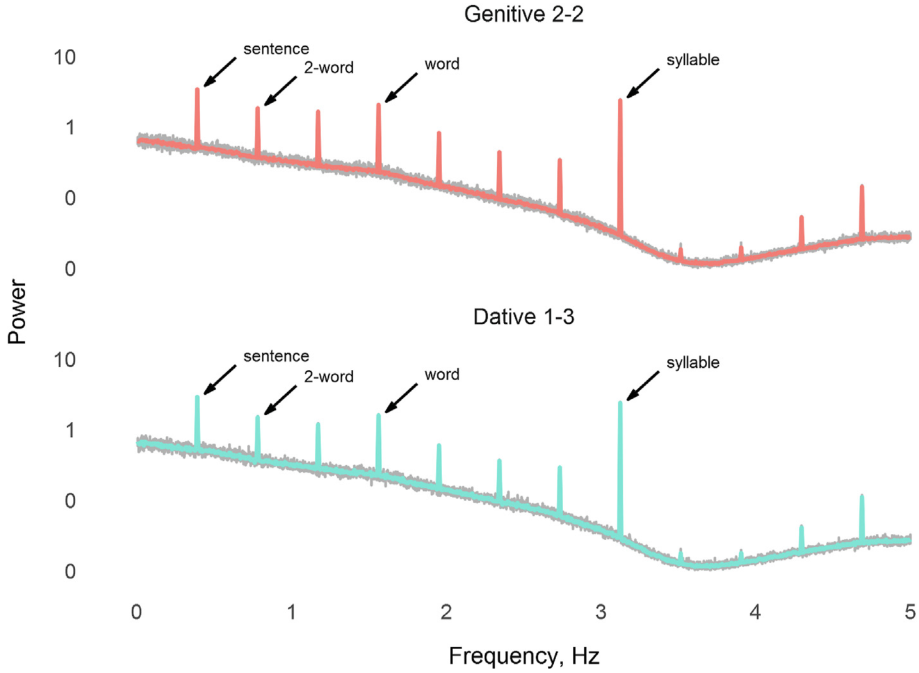

Fig. 3 – Simulated average power spectra for the Genitive and Dative conditions from the distributional semantics vector model. Peaks at the frequencies corresponding to syllables, words, 2-word combinations and sentences are marked by the arrows. Thin grey lines represent simulation of individual participants. Pink or teal lines represent the mean of all simulated participants.

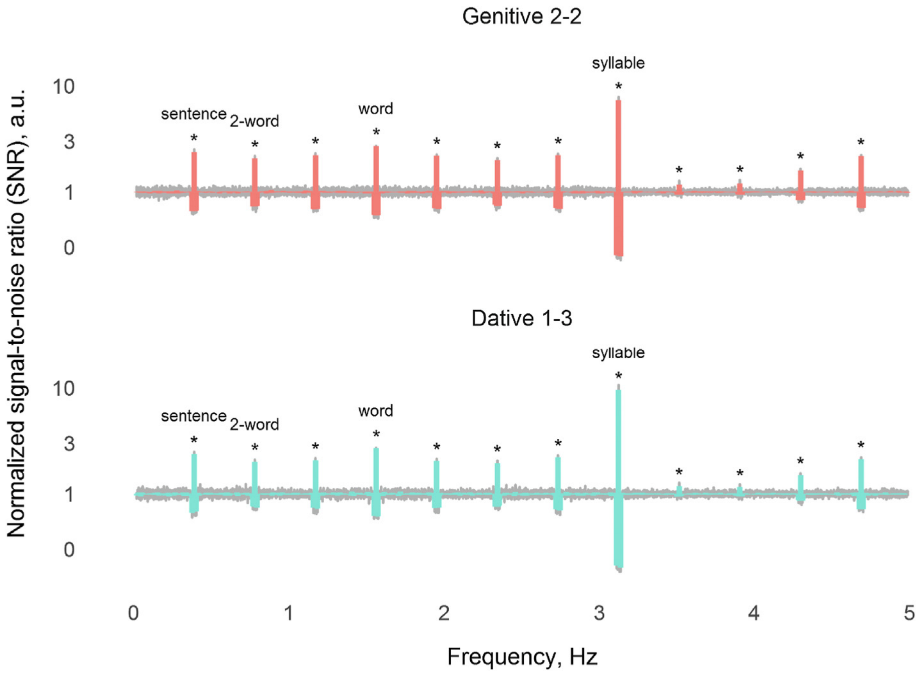

Fig. 4 – Normalised signal-to-noise ratios (SNR) calculated on the basis of the simulated power spectra from the distributional semantics vector model for the Genitive and Dative conditions. '*' mark frequencies at which the power is larger than mean power at the four neighboring bins with FDR-adjusted p-value smaller than .001. Peaks at the frequencies corresponding to syllables, words, 2-word combinations and sentences are labelled. Thin grey lines represent simulation of individual participants. Pink or teal lines represent the mean of all simulated participants.

Table 1 – Sample critical and control questions used in auditory pre-screening for the sample Genitive 1–3 ("Dina's cook is making pancakes") and Dative 2-2 ("The cook is making pancakes for Dina") conditions from Fig. 2. Two response options for each question type and the correct ( ) and incorrect (×) responses for each condition are also shown.

Question		Response options	
Critical	Who are the pancakes for?	certainly for Dina (Gen × , Dat)	cannot tell definitively (Gen ✓, Dat ×)
Control 1	Who is making the pancakes?	Cook (Gen ✓, Dat ✓)	Dina (Gen × , Dat ×)
Control 2	What is the cook making?	Pancakes (Gen ✓, Dat ✓)	Pies (Gen × , Dat ×)

what would the test have shown had we stopped at 20 as planned, see *MEG data* under *Results*. In total, we collected the data from 40 participants. Of these 40 participants, six were removed due to various technical reasons (participants falling asleep, sound not recorded, data not recorded at all, etc.), and another 3 participants were removed due to missing triggers in the MEG data. Additional data-based participant exclusion is reported in the *Behavioral data* section under *Results*.

3.1.2. *Materials*

Materials in Experiment 2 were the auditory versions of the 60 sentence sets of 2 conditions used in Experiment 1. As mentioned earlier, the 60 sets were chosen from a larger pool via auditory prescreening, described below together with other relevant details.

3.1.2.1. GENERATION AND PRE-SCREENING OF AUDITORY STIMULI. Sixty-four sets of Genitive 2-2 and Dative 1–3 conditions as in Table 1 were created, phonetically transcribed and split into syllables. Each unique syllable was then synthesized using the MacinTalk Synthesizer (Russian female voice Milena, macOS High Sierra Version 10.13.6). All silent intervals at the beginning and end of the synthesized syllables were removed. This resulted in syllable durations ranging from 200 to 550 ms. All syllables were then slowed down or speeded up to become as close to the target duration of 320 ms as possible (pitch preserved). The exact duration of 320 ms was obtained by truncating the resulting syllables by several milliseconds or padding them with several milliseconds of silence. Auditory words were constructed by concatenating 2 syllables together without any gaps; sentences were constructed by concatenating 4 words. The stimulus intensity spectra averaged over 24 trials for each condition are depicted in Fig. 5. Note that unlike in Ding et al. (2016), our spectra do feature small peaks at the word, 2-word, and sentence rates. However, an independent-samples *t*-test showed no significant difference in the peak amplitude at the 2-word rate between conditions ($t(43.38) = -.42$, $P = .674$, 90% CI on Cohen's d $[-.61, .36]$[1]). Therefore,

[1] The 90% CIs on Cohen's d are reported here because they correspond to the equivalence and noninferiority/nonsuperiority tests at the significance level of 5% (Walker & Nowacki, 2011). Positive d values correspond to the peaks in Genitive condition being larger.

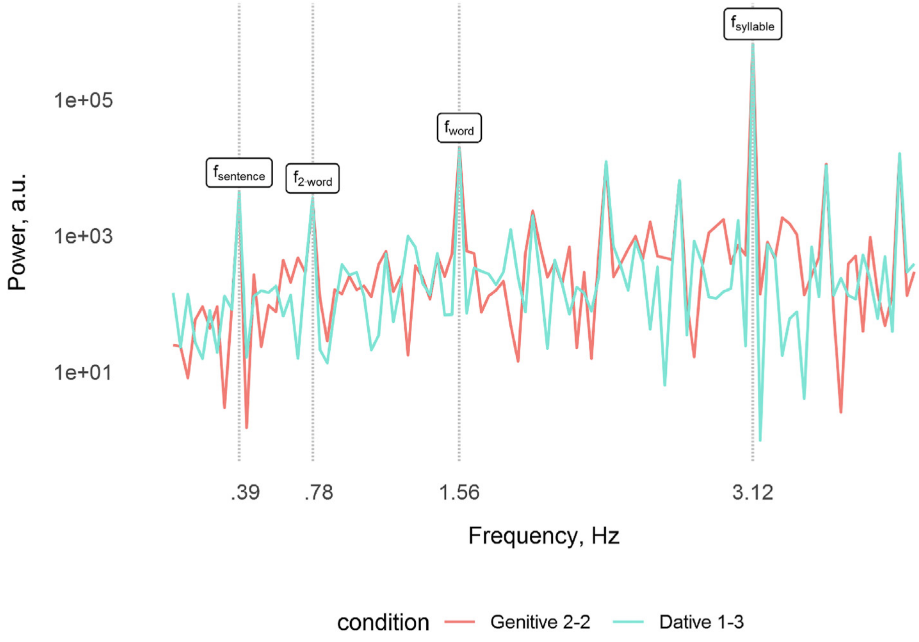

Fig. 5 – Spectra of stimulus intensity for the Genitive 2-2 (pink line) and Dative 1–3 (teal line) conditions averaged over 24 trials

this acoustic regularity does not confound the comparison of the 2-word rate peaks in MEG data. Incidentally, the peaks at the three other rates of interest were similar as well (syllable rate: *t* (41.42) = .22, *P* = .825, 90% CI on Cohen's *d* [– .42, .55]; word rate: *t* (43.01) = – .20, *P* = .846, 90% CI on Cohen's *d* [– .54, .43]; sentence rate: *t* (40.38) = – .51, *P* = .612, 90% CI on Cohen's *d* [– .63, .34]).

The quality of the auditory sentences and whether they can be correctly understood by listeners was verified in a prescreening by an independent set of 11 native Russian-speaking volunteers aged 18–29 (7 females, 4 males) who did not take part in the MEG experiment. For each of 64 sets, we constructed three questions: a critical question and two controls. Each question was paired with two response options, of which exactly one was correct on a given trial. All questions served to test general intelligibility of the sentences, while the critical ones specifically assessed whether participants could correctly hear and interpret the case of the critical noun (*Noun2*). Control questions were added to add variability and prevent strategies on behalf of the participants and ensured that participants had to listen to the whole sentence and not just to the critical noun. Table 1 shows questions for the set of sentences from Fig. 2.

On each trial, a sentence was presented auditorily via headphones and followed by a question presented visually on the screen together with two response options. The participants were asked to choose the correct response. With 64 sentence pairs and three sentence types, there was a total of 64 sets × 2 conditions × 3 questions = 384 possible trials.

Participants were tested individually, in a sound-resistant cubicle. In total, the data from 11 participants were collected. The first 6 participants were tested on half of all trials (192 trials, counterbalanced across participants). As the data collection was quicker than expected, the remaining 5 participants were assigned a full set of 384 trials presented in random order. The data were analyzed by fitting a mixed-effects logistic regression with a random intercept of *participant* and estimating marginal means, confidence intervals and comparing across conditions based on the fitted model. The estimated overall percentage of correct answers to all the critical questions combined was generally high (87%, 95% CI 81%–91% across participants) although it was higher in the Dative condition (estimated marginal means for Dative: 90%, Genitive: 84%, odds ratio 1.71, *P* = .005). Three sets of sentences that yielded the lowest accuracy (all 62%) were removed. A fourth set was removed that contained an unintelligible word noted by multiple participants during debrief at the end of the test. The remaining 60 sets of conditions had an overall accuracy of 88% across participants (95% CI 83%–92%) and were used as stimuli in Experiments 1 and 2.

3.1.3. *Procedure*

Participants' MEG activity was recorded while they listened to isochronous speech presented in 10-sentence-long trials. Sixty sentences from each condition were randomly split into 6 trials, each containing 10 sentences from a single condition. This resulted in a total of 12 trials (6 Genitive and 6 Dative) that constitute a single block. Each trial was presented auditorily at a rate of 320 ms/syllable, without any pauses between syllables or any other prosodic segmentation cues to word or sentence boundaries. The block was repeated 4 times for each participant (with pauses in between blocks); the composition of trials within a block and the order of trials was randomized for each block and each participant. All in all, there were 24 Genitive and 24 Dative trials.

In order to minimize the evoked response to the auditory onset in the MEG experiment, a fade-in/fade-out was added at the beginning/end of each trial. To create them, a random sentence representing the same condition that was not used in the trial was chosen and split into halves, i.e. *Noun1* + *Noun2* and *Verb*+*Noun3*. For the trial-initial fade-in, the 4-syllable sequence corresponding to *Verb*+*Noun3* was manipulated as follows: (i) the initial 0–1.5 syllables (the exact duration was chosen randomly) was rendered silent, (ii) the intensity of the following 2.5–4 syllables built up linearly from silence to the original level, (iii) the intensity of the remaining 0–1.5 syllables did not change. The procedure was applied in the mirrored order to the *Noun1* + *Noun2* sequence to create a fade-out at the end of the trial. The overall duration of the 10-sentence long trial including the fade-in and fade-out was 25.6 sec.

A memory task was be presented at the end of each trial. The participants were shown a sentence on the screen, and had to judge whether it had been played during the trial by choosing between "This sentence was among the ones just played" and "There was no such sentence" response options. The participants judged two sentences at the end of each trial. Each time, the sentence was either a full sentence presented during the trial (e.g., Q1 in Fig. 6a and Q1 and Q2 in Fig. 6b) or a grammatically correct novel sentence that combined words from three different sentences presented during the trial (e.g., Q2 in Fig. 6a).

At the start of each trial, a beep (a 240-ms-long diamond-shaped 250-Hz sinusoid) was played as a cue to trial beginning, followed by 760 ms of silence and then by the rest of the trial sequence (fade-in, 10 sentences of the same condition, fade-out, memory task). The audio level of the sound was adjusted to a comfortable level for each participant.

The experiment started with a practice session of 3 trials (additional sentences similar to the experimental ones were used). Each participant then was exposed to 4 blocks of 12 trials, with a pause in between blocks. The experiment took approximately 40 min.

3.1.4. *MEG recording*

The magnetoencephalographic (MEG) recording was done using a 306-channel Neuromag Vector View (Elekta Oy, Finland) in a magnetically-shielded room at the Moscow MEG Centre (Moscow State University of Psychology and Education campus).

An online band-pass filter of .1–330 Hz was applied during the recording. Head position was monitored continuously by means of 4 head position indicator (HPI) coils attached above the forehead and behind the participant's ears. HPI coils and additional head point positions in reference to the nasion, left and right pre-auricular points coordinate frame were digitized using Polhemus FASTRAK device. The sampling frequency was 1 kHz.

3.1.5. *MEG data analysis*

A .1–40 Hz band-pass filter was applied to the data. The data were divided into 23.04 sec long epochs starting at the onset of the second sentence and ending at the offset of the last (i.e., 10th) sentence. The first sentence was excluded in order to avoid the response to the acoustic onset of each trial (as in Ding et al., 2016).

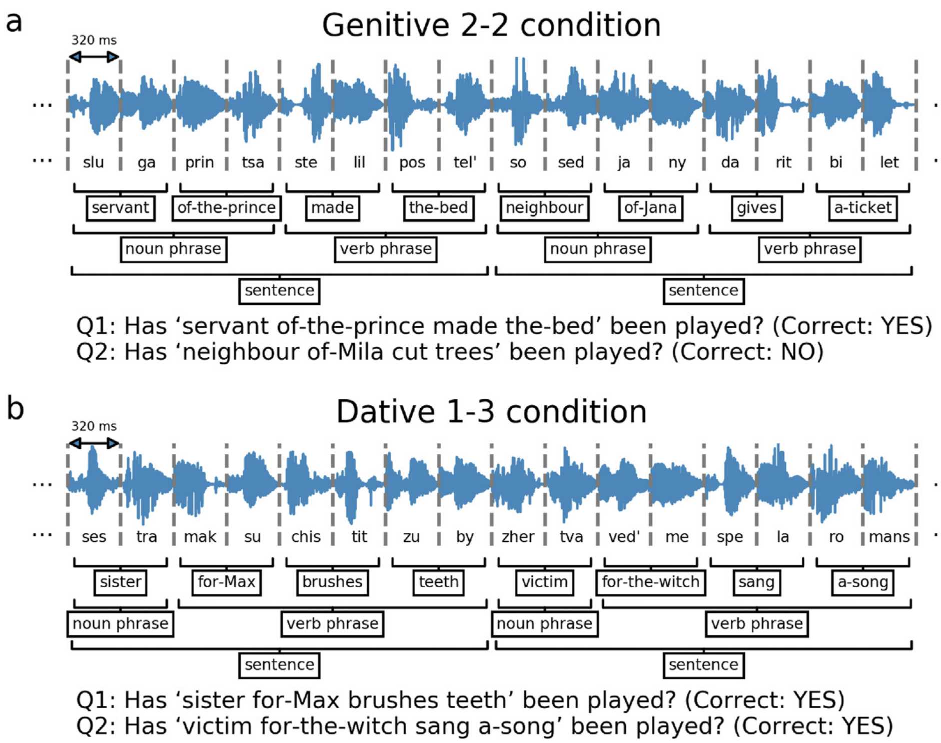
a
Genitive 2-2 condition
320 ms
...
slu ga prin tsa ste lil pos tel' so sed ja ny da rit bi let
servant
of-the-prince
made
the-bed
neighbour
of-Jana
gives
a-ticket
noun phrase
verb phrase
noun phrase
verb phrase
sentence
sentence
Q1: Has 'servant of-the-prince made the-bed' been played? (Correct: YES)
Q2: Has 'neighbour of-Mila cut trees' been played? (Correct: NO)
b
Dative 1-3 condition
320 ms
...
ses tra mak su chis tit zu by zher tva ved' me spe la ro mans
sister
for-Max
brushes
teeth
victim
for-the-witch
sang
a-song
noun phrase
verb phrase
noun phrase
verb phrase
sentence
sentence
Q1: Has 'sister for-Max brushes teeth' been played? (Correct: YES)
Q2: Has 'victim for-the-witch sang a-song' been played? (Correct: YES)

Fig. 6 – A sample MEG trial. Two-sentence long excerpts from a 10-sentence long Genitive 2-2 and Dative 1–3 condition trials are shown, together with their underlying syntactic constituent structure. Questions Q1 and Q2 are part of the memory task presented at the end of each trial.

In the following steps, we employed *denoising based on spatial filtering* (Cheveigné & Simon, 2008). This technique partitions data into stimulus-related and stimulus-unrelated activity based on their trial-to-trial phase-locked reproducibility. As in Ding et al. (2016), we utilized this technique twice: once to denoise the data in the original time space and once to accentuate the peaks in the frequency space. The initial denoising was done with the following parameters: 60 components kept during two applications of *principal component analysis* (PCA), the proportion of evoked power explained–90%. Epoched data was then Fourier-transformed into the frequency domain where peaks corresponding to different syntactic units could be compared. The frequency resolution was 1/23.04 $sec^{-1} = .04$ Hz. We then applied *denoising based on spatial filtering* again, once for each frequency bin using the formula in Ding et al. (2016) (subsection *Data Analysis* in *Online Methods*). As a consequence of applying the outlined procedure, the spatial dimensionality of the data was reduced from 306 (the number of sensors) to 1, i.e., each trial was represented by a single frequency-indexed vector.

3.1.6. *Exclusion criteria*

We excluded participants whose performance on the memory task was not above chance at the significance level of .05. With 96 questions (2 question after each of the 48 trials) this criterion corresponds to having less than 56 correct answers. This was the only data-based exclusion criterion.

3.1.7. *Statistical analysis*

For each participant, after processing the data as described in subsection *MEG data analysis*, we applied the following steps:

1. The power at each frequency bin was averaged over trials separately for each condition.
2. Averaged power was then converted into SNR.
3. SNRs were normalized by the SNR at the word-rate frequency as in Equation (1).
4. The natural logarithm of the SNRs was taken to obtain *log-SNRs*.

For each participant, the log-SNRs at the 2-word frequency in Gen 2-2 and in Dat 1–3 conditions were compared with a one-tailed paired Bayes factor (*BF*) *t*-test with the boundaries 1/6 and 20. The asymmetry in boundaries was introduced in order to balance the probabilities of false positive and false negative errors (as proposed in Schönbrodt & Wagenmakers, 2018; and Weiss, 1997). The null hypothesis for the test was that there was no effect of condition, and the alternative hypothesis was that there was an effect of condition with an informed prior used (a shifted and scaled t-distribution). See subsection *Sequential sample analysis* in the Appendix for details on the parameter selection.

3.1.8. *Sequential sampling plan*

We planned to initially collect 20 participants and then sequentially collect additional participants in the increments of 5 until we got a Bayes factor less than 1/6 or larger than 20, or reach the sample size of 50. We applied Bayes Factor Design Analysis (*BFDA*, Schönbrodt & Wagenmakers, 2018) in order to assess this plan and took the results to suggest that our plan had a high probability of yielding compelling evidence towards the correct hypothesis. See subsection *Sequential sample analysis* in Appendix for details.

3.1.9. *Outcome-neutral quality assurance*

To test the quality of our setup and data collection, we planned to check that the sentence, word and syllable peaks were present in the collected data, i.e. the corresponding unnormalized SNRs were all significantly larger than 1 at the significance level of .01.

3.1.10. *Code and data availability*

The MEG and behavioral data are available at https://openneuro.org/datasets/ds003703/versions/1.0.0. The code and the files necessary to recreate this manuscript are available at https://osf.io/kdpcs/.

3.1.11. *Preregistration*

Registered Report Protocol Preregistration is available at https://osf.io/qhg9z. The accepted Stage 1 manuscript is available at https://osf.io/project/qhg9z/files/osfstorage/60ddd82431881a025463d91c.

3.2. **Results**

3.2.1. *Behavioral data*

Of the 31 participants remaining after exclusion for non-data-related reasons (see *Participants* under *Methods*), further 4 had to be removed because the number of the correct answers they gave was less than the prespecified threshold of 56 (see *E. criteria* under *Methods*). The accuracy of the remaining 27 participants ranged from 56 to 91 correct answers with the median of 69.

3.2.2. *MEG data*

The sentence, word and syllable peaks were all present in the data thus confirming the data quality (all p's $< .01$, see *Outcome-neutral quality assurance* under *Methods*). The averaging and the tests were done using the logarithms of unnormalized SNR due to the right-skewed nature of the SNR distribution. Fig. 7 shows SNR spectra from several representative participants (pooled across the Gen 2-2 and Dat 1–3 conditions) and demonstrates strong peaks at the sentence, word and syllable rates, as well as at the 2-word phrase rate.

Looking at the data from the Genitive 2-2 and Dative 1–3 condition separately, the four peaks are again clearly visible in the individual and averaged power spectra (Fig. 8). A critical comparison is that between the 2-word peaks in the Genitive 2-2 vs. Dative 1–3 conditions (Fig. 9). A quick glance is enough to see that the conditions look similar both at the group level, and at an individual level for most participants.

The interpretation was confirmed via a planned one-tailed paired BF *t*-test with the boundaries 1/6 and 20 (see *Sample size estimation*) applied to the word-normalized log-SNRs from two conditions of the 27 eligible participants which resulted in a BF of .009 or, approximately, 1/112. Because this number is smaller than 1/6 we considered the current sample as final and concluded that there was evidence of no difference between the conditions (Note that the conclusion is the same

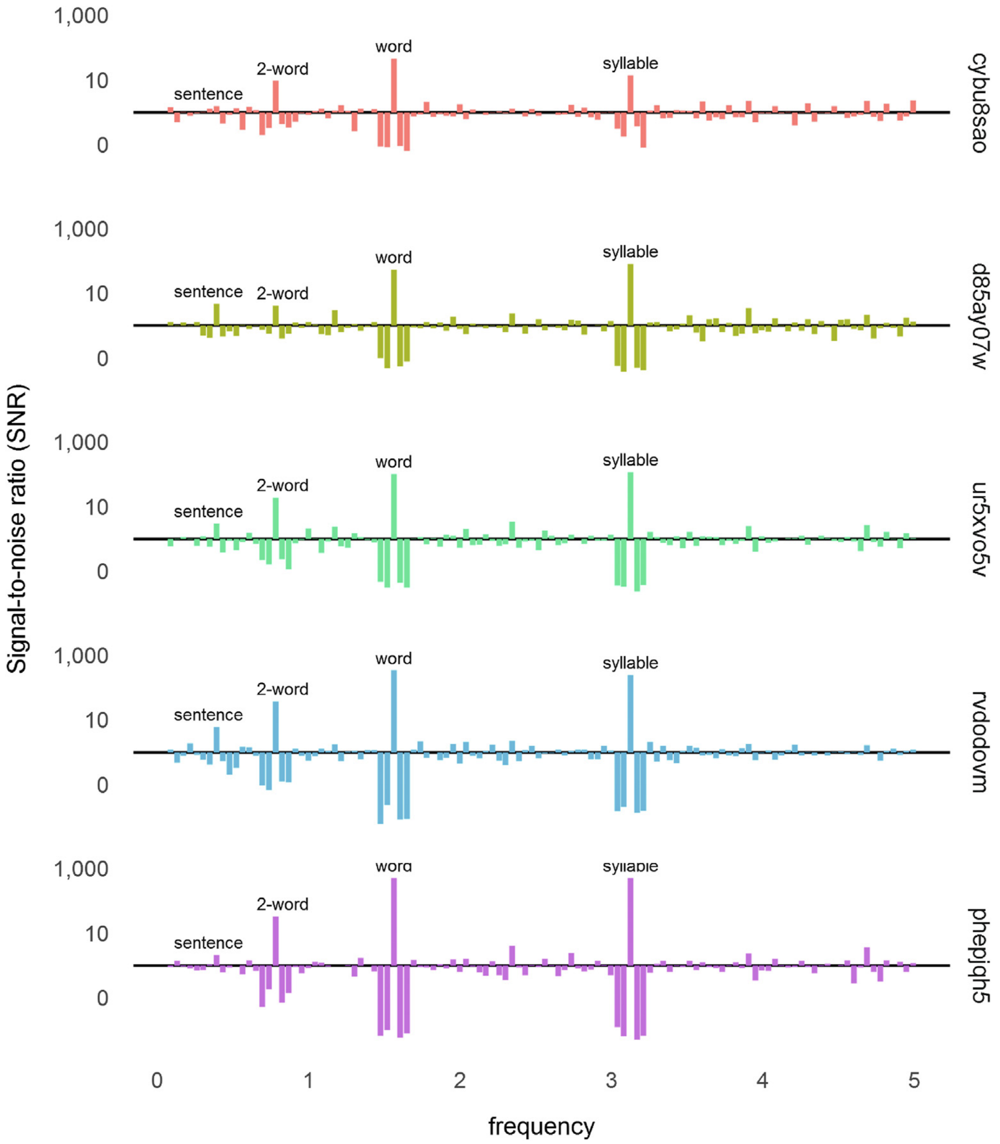

Fig. 7 – SNR spectra of several representative participants spanning the range from the lowest to the highest SNRs averaged over the four frequencies of interest. Power was first averaged over the trials pooled from both conditions. Then the averaged power at each frequency bin was divided by the average of the power in the two neighboring bins on both sides.

when a subsample of the first 20 participants is analysed: BF of .011 or, approximately, 1/93).

Unnormalized SNRs at the sentence, 2-word, word, and syllable rates are additionally depicted in Fig. 10 for visual comparison.

3.3. *Exploratory analysis*

3.3.1. *Response topographies*

At the request of one of the reviewers, here we provide response topographies for the labeled peaks in Fig. 8 The particular analysis we performed following Ding et al. (2016), does not allow for comparing the topographies of the peaks across conditions. Applying the *denoising based on spatial filtering* (DSS) procedure for the second time—in the frequency domain—reduced the spatial dimensionality of our data to 1 (see subsection *MEG data analysis* for details). Thus, response topographies of the peaks could only possibly differ in magnitude but not in distribution over the sensors. To overcome this problem, here we ran a slightly different calculation:

- In the original analysis, the second DSS was used to find an optimal filter to use for averaging across DSS components. Here, we first calculated power spectra after applying DSS for the first time and then averaged the results both over

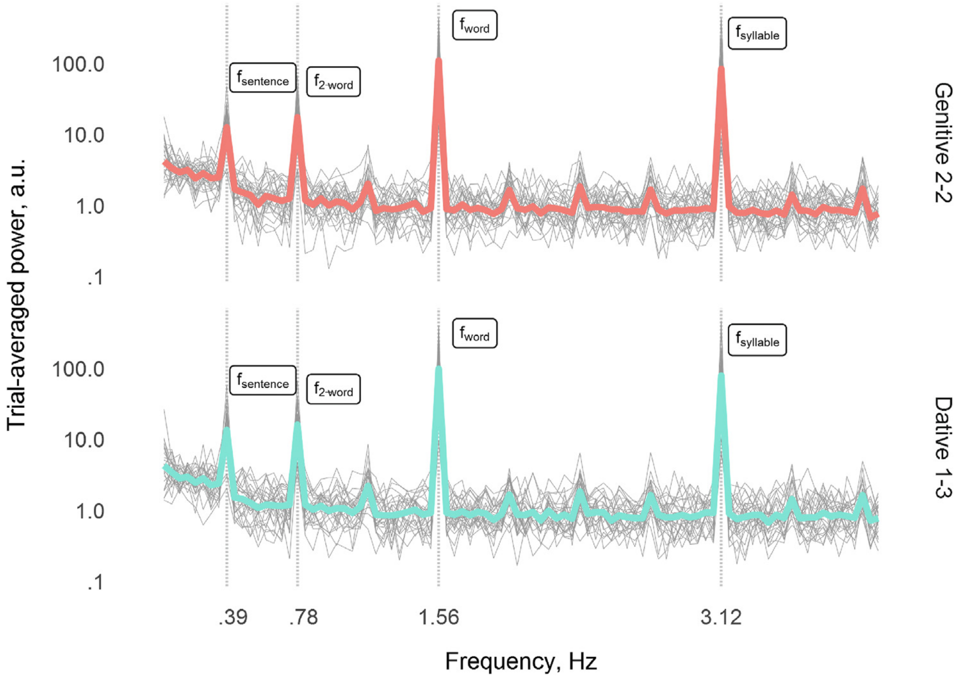

Fig. 8 – Power spectra of individual participants (light grey) and their grand averages (thicker lines). All the four expected peaks (sentence, 2-word, word, syllable) are clearly present in both conditions.

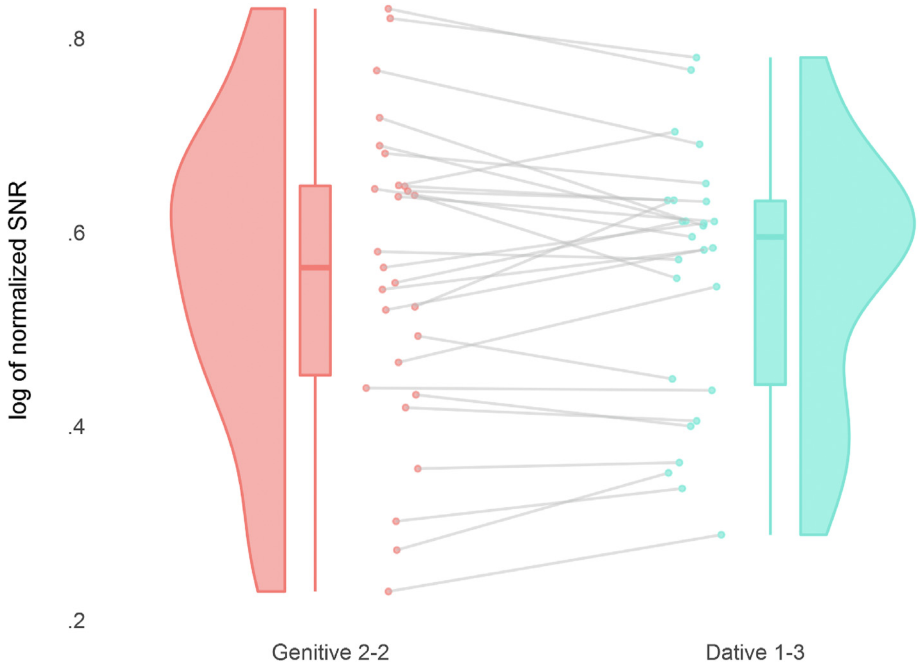

Fig. 9 – The peaks at the 2-word frequency operationalized as logs of normalized SNR of the power spectrum. The two vertical box and whiskers plots show distributions of peak sizes in the Genitive 2-2 (pink) and Dative 1–3 (teal) condition. The black lines connect within-subject points (*n* = 27).

trials and DSS components to estimate the power spectra for each participant–condition combination. These individual power spectra are depicted in Fig. 11 (thin lines) together with the average per-condition spectra (the thick lines). Note that the peaks are much smaller than in Fig. 8 and the sentence peaks cannot be discerned at all.

- We then inverted the first DSS transformation to return to the sensor space and then averaged power over trials

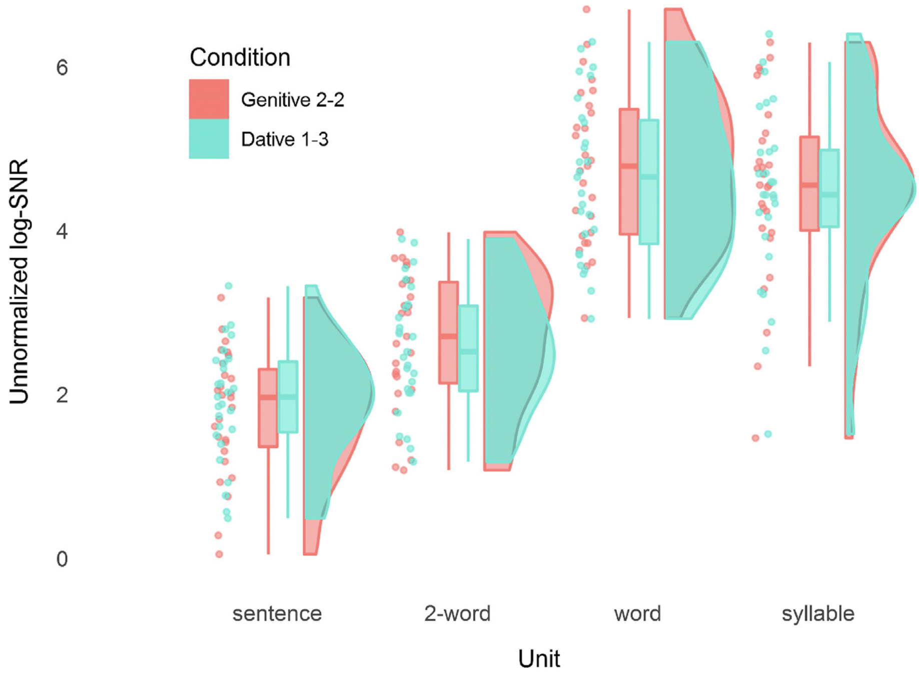

Fig. 10 – Comparison of peaks at the frequencies of the four units of interest (sentence, 2-word, word, syllable) operationalized as SNR of the power spectrum. The vertical box and whiskers plots show distributions of peak sizes in the Genitive 2-2 (pink) and Dative 1–3 (teal) condition.

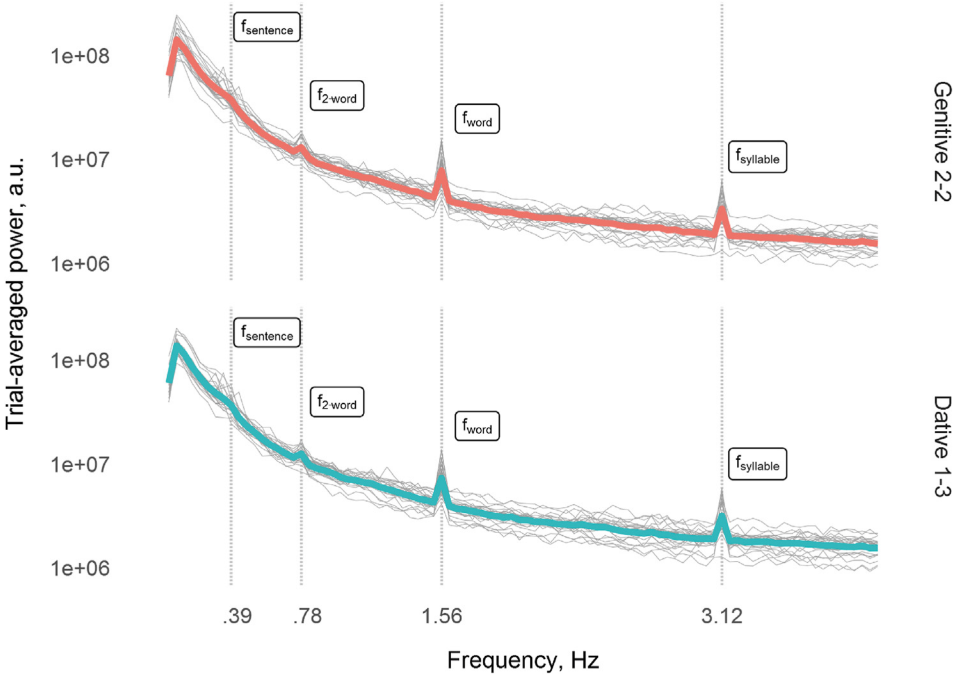

Fig. 11 – Power spectra of individual participants after the first application of DSS (light grey) and their grand averages (thicker lines). Only three out of the four expected peaks (2-word, word, syllable) are clearly present in both conditions, the sentence peak is not.

only, to estimate the power spectra for each participant–condition–sensor combination. We then normalized each power topography independently and averaged the results across the participants to obtain an estimate of a general topography for each condition–frequency combination. The topographies for the four frequencies of interest are depicted in Fig. 12. The topographies are very similar across conditions and frequencies, which is likely a consequence of the DSS filtering. Such a similarity is also present in Ding et al. (2016).

4. Discussion

In the present study, we used Russian case marking to create pairs of 4-word sentences that differed in a single phoneme (corresponding to the Genitive vs. Dative case marker) which led to differences in the syntactic structure of the sentences but not in their lexico-semantic characteristics (as modeled by Frank and Yang (2018)). The Genitive 2-2 condition sentences contained two 2-word long constituents: a 2-word long subject (e.g., *cook of-Dina*) followed by a 2-word long verb phrase (*is-making pancakes*). In the Dative 1–3 condition, on the other hand, the subject was a single word (e.g., *cook*) and was followed by a 3-word long verb phrase (indirect object, verb and direct object, e.g., *for-Dina is-making pancakes*). As participants listened to sequences of sentences from the same condition we recorded their MEG and found spectral power peaks corresponding to the rates of syllables, words, sentences, as well as 2-word pairs. According to previous frequency tagging studies (Ding et al., 2016, 2017) that interpreted such peaks as reflecting the sentence syntactic structure, we should have observed a larger peak at the 2-word rate in the Genitive 2-2 condition that contained well-formed syntactic constituents at that rate. Yet there was no difference in the power of the 2-word peak in the Genitive 2-2 vs. Dative 1–3 condition. A plausible and, arguably, simpler alternative interpretation is that the peaks resulted from the lexico-semantic regularities in the stimuli as proposed by Frank and Yang (2018). As can be seen in Fig. 3, according to Frank and Yang's model our Genitive vs. Dative conditions do not exhibit reliable differences at the 2-word rate peak (as well as at the syllable or sentence rate peaks). The lack of difference between conditions observed in the human data is in line with this model.

At the same time, we would like to emphasize that our results should not be taken as providing strong support for the model by Frank and Yang (2018) as a model of the EEG/MEG response produced during auditory sentence comprehension. The model predicts large peaks at the harmonic frequencies of the sentence rate which are much less salient in the MEG data (compare Figs. 3 and 8). This suggests that lexico-semantic properties alone may not be sufficient to explain a full pattern of results (see Jin, Lu, and Ding (2020) and Lo (2021) for a similar point). In particular, as we discuss next, it was proposed that factors other than those discussed so far (i.e. syntactic and lexico-semantic factors) may also contribute to the EEG/MEG response during sentence comprehension.

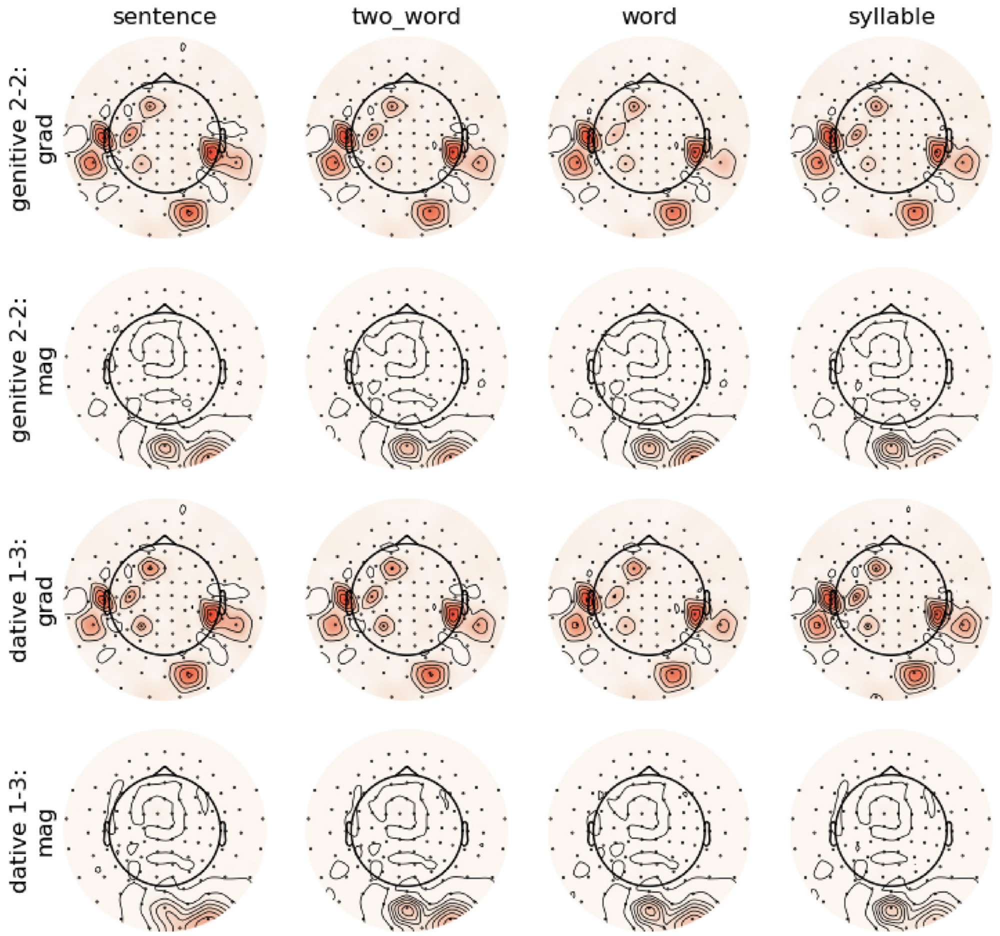

Fig. 12 – Response topographies of the power spectra after the single application of DSS at the four frequencies of interest (in columns) for both conditions (Genitive 2-2 in the first two rows, Dative 1–3 in the last two) and the two types of sensors (gradiometers in the odd rows, magnetometer in the even ones). Note that the topographies are barely distinguishable across frequencies and conditions.

Glushko, Poeppel, and Steinhauer (2020) draw attention to the role of prosody in sentence comprehension and argue that Ding et al.'s (2016) findings may strongly reflect prosodic factors, i.e. stem from prosodic properties of the stimuli. For sentences in the 2-2 syntactic condition, the 2-2 grouping was a prosodic default, e.g., (*new plans |give hope*). At the same time, the sentences in the 1–3 syntactic condition could only have the 1–3 prosodic grouping (*drink |lemon juice*) but not (*drink le- | mon juice*). Thus, in the critical 2-2 and 1–3 conditions, the prosodic groupings paralleled the syntactic structure. Even though Ding et al. (2016) explicitly neutralized prosodic cues, listeners are known to activate covert, implicit prosody (see Glushko et al. (2020) for references of previous research demonstrating this). Therefore, Ding et al.'s (2016) findings could also be explained by the prosodic account. Glushko et al. (2020) tested this alternative by comparing the 2-2 condition with a new 1–3 condition in which the prosodic grouping into two 2-word chunks was plausible, e.g., (*John likes |big trees*). As in the current study, the 2-word peak in this new 1–3 condition did not differ from that in the 2-2 condition, contradicting the syntactic interpretation in Ding et al. (2016).

Both our study and the Glushko et al. (2020) study adapted the 2-2 and 1–3 conditions from Ding et al. (2016). This is not a coincidence: the 1–3 was a crucial control condition. Without it, it would have been impossible to tell whether the 2-word peak in the 2-2 condition had anything to do specifically with the phrases. Indeed, it could have been argued that the largest meaningful chunks (sentences in the case of the 2-2 condition) not only produced the peak at their corresponding frequency but also at its harmonics. By changing the 2-2 syntactic structure to the 1–3 structure and then not observing a peak at the 2-word frequency, Ding et al. (2016) refuted this argument. Tavano et al. (2021) took a different approach to differentiating phrase-level peaks from those arising as harmonics of a slower (sentence) rhythm: they employed sentences that had 2-3 and 3-2 structures in addition to 2-2 and 1–3 structures. Critically for the present discussion, Tavano et al. (2021) did not find any difference in the 2-word peaks between the 2-2 and 1–3 conditions. Whereas their findings do not make it possible to distinguish between the lexico-semantic vs. prosodic accounts (as this was not the study goal), their findings are clearly at odds with the syntactic account.

An anonymous reviewer notes that the memory task employed in our study did not require syntactic processing and that a different task could produce a stronger syntactic response. We agree that a properly syntactic task would enhance the degree of syntactic processing undertaken by the listeners, which may have effects on MEG spectral responses and, consequently, on general conclusions. Arguably, our memory task could be solved on the basis of word sequences only, with no involvement from syntax. Yet we point out that our memory task was formulated in terms of sentences (i.e. "This sentence was among the ones just played–yes/no") and that building a sentence requires establishing syntactic relations between (groups of) words. Thus we believe that the task difference was an unlikely reason for the difference in the conclusions of Ding et al.'s (2016) study and the current one. Ding et al.'s (2016) task for the Chinese participants was to detect outlier trials containing sentences/phrases that were syntactically correct but semantically implausible. Ding et al.'s (2016) task for the English participants required spotting outliers that were syntactically ill-formed and semantically implausible. Like our memory task, both variants require establishing syntactic relations between (groups of) words. The part of Ding et al.'s task that is different from our memory task–evaluating how plausible the sentence/phrase is–is largely semantic in nature.

Summarising, our findings do not support a strong syntactic interpretation of frequency peaks proposed in Ding et al. (2016). Together with other corroborating evidence (Glushko et al., 2020; Tavano et al., 2021), we have to conclude that the frequency-tagging paradigm as used in this and other studies does not successfully isolate the syntactic structure and is subject to other–for example, prosodic and lexico-semantic–influences. This conclusion has important practical repercussions, i.e. frequency-tagging data cannot serve as a specific marker of intact or impaired syntactic processing in developmental studies or in clinical studies of patients. Theoretical repercussions of our findings are limited: they show that the frequency-tagging findings cannot be taken as evidence for hierarchical syntactic structure and return the field to where it was prior to Ding et al. (2016), with some who take a hierarchical syntactic structure as an integral part of sentence comprehension and others who do not.

Funding

The study has been funded by the International Laboratory for Social Neuroscience of the Institute for Cognitive Neuroscience HSE, RF Government grant # 075-15-2019-1930.

Credit author statement

Evgenii Kalenkovich: Software, Validation, Formal analysis, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. Anna Shestakova: Resources, Project administration, Funding acquisition, Writing – review & editing. Nina Kazanina: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision

Open practices

The study in this article earned Open Data, Open Materials and Preregistetred badges for transparent practices. Data and Materials for this study can be found at: https://openneuro.org/datasets/ds003703/versions/1.0.0.

Acknowledgements

We would like to thank Liya Merzon for her help with the MEG equipment and data collection, and her consultation concerning the MEG data analysis. We would also like to thank Oleg Shevtsov at the Center for Cognition and Decision-Making, NRU HSE and Andrey Prokofiev at the Moscow MEG Center, MSUPE for their generous technical support.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2021.09.012.

REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berwick, & Weinberg. (1984). *The grammatical basis of linguistic performance*. Book. Retrieved from https://dl.acm.org/citation.cfm?id=5620 papers://710a1fe1-62fe-4ca9-ac92-c334728110c1/Paper/p38.

Cheveigné, A. de, & Simon, J. Z. (2008). Denoising based on spatial filtering. *Journal of Neuroscience Methods, 171*(2), 331–339. https://doi.org/10.1016/j.jneumeth.2008.03.015

Chomsky, N. (2002). *Syntactic structures*. Mouton de Gruyter. Retrieved from https://books.google.ru/books?hl=en{\&}lr={\&}id=SNeHkMXHcd8C{\&}oi=fnd{\&}pg=PR5{\&}ots=AVauRDZotP{\&}sig=AKKNLytip{_}9{_}qqgbK6OnVyTFwGc{\&}redir{_}esc=y{\#}v=onepage{\&}q{\&}f=false.

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. https://doi.org/10.3389/fnhum.2017.00481

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

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences, 19*(12), 729–743. https://doi.org/10.1016/j.tics.2015.09.008

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. *Plos One, 13*(5), e0197304. https://doi.org/10.1371/journal.pone.0197304

Glushko, A., Poeppel, D., & Steinhauer, K. (2020). Overt and covert prosody are reflected in neurophysiological responses previously attributed to grammatical processing. *bioRxiv*, 301994. https://doi.org/10.1101/2020.09.17.301994, 2020.09.17.

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

Lo, C.-W. (2021). Testing low-frequency neural activity in sentence understanding *(PhD Thesis)*. University of Michigan.

Makov, S., Sharon, O., Ding, N., Ben-Shachar, M., Nir, Y., & Golumbic, E. Z. (2017). Sleep disrupts high-level speech parsing despite significant basic auditory processing. *Journal of Neuroscience, 37*(32), 7772–7781. https://doi.org/10.1523/JNEUROSCI.0168-17.2017

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CrossRef Listing of Deleted DOIs, 1*, 1–12. https://doi.org/10.1.1.764.2227.

Panchenko, A., Arefyev, N., Ustalov, D., Loukachevitch, N., Paperno, D., Biemann, C., et al. (2017). Russian distributional Thesaurus (RDT): Word embeddings. https://doi.org/10.5281/ZENODO.400631.

Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., et al. (2017). Human and machine judgements for Russian semantic relatedness. In *Analysis of images, social networks and texts: 5th international conference* (pp. 221–235). Yekaterinburg, Russia: Springer International Publishing. https://doi.org/10.1007/978-3-319-52920-2_21. aist 2016, *yekaterinburg, russia, april 7-9, 2016, revised selected papers*.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Tavano, A., Blohm, S., Knoop, C. A., Muralikrishnan, R., Scharinger, M., Wagner, V., et al. (2021). Neural harmonics of syntactic structure. *bioRxiv*, 31575. https://doi.org/10.1101/2020.04.08.031575, 2020.04.08.

Walker, E., & Nowacki, A. S. (2011). *Understanding equivalence and noninferiority testing*. Springer. https://doi.org/10.1007/s11606-010-1513-8

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (the Statistician), 46*(2), 185–191. https://doi.org/10.1111/1467-9884.00075