Neural processing of sentences

Amelia Burroughs[*] AMELIA.BURROUGHS@BRISTOL.AC.UK
Nina Kazanina[**,†] NINA.KAZANINA@BRISTOL.AC.UK
Conor Houghton[*] CONOR.HOUGHTON@BRISTOL.AC.UK

[*] *Department of Computer Science, University of Bristol, UK*

[**] *School of Psychological Science, University of Bristol, UK*

[†] *Institute of Cognitive Neuroscience, National Research University Higher School of Economics, Moscow, Russian Federation*

Abstract

How the human cognitive system is able to comprehend language has been a matter of recent debate. On the one hand the brain may make use of learned grammatical rules to decompose sentences into a hierarchy of syntactic structures to generate meaning. On the other hand the brain may rely on simpler, statistical methods where the generation of meaning relies on sequential processing of lexical information. To investigate these we recorded EEG from participants as they listen to streams of isochronously-presented words. In our different conditions there are grammatical and semantic manipulations: the words in the stream can or cannot be parsed into phrases and sentences which are, or are not, semantically sensible. These manipulations affect the cortical activity synchronised with the rate at which syllables, phrases and sentences were presented.

1. Introduction

The ability of the human brain to rapidly generate meaning from an incoming stream of sentences during natural language is impressive. There are two competing, but not necessarily exclusive, theories describing how we are able to process words. Both of these theories suggest that the smaller discrete units of language, such as words and phrases, are concatenated by the brain into larger units, such as sentences. What is under debate is the principles that govern this organisation: one theory argues for the primacy of a learned rule-based grammar, the other, for the primacy of statistical relationships between the discrete units.

It has been demonstrated, using MEG in (Ding et al. 2016) and using EEG in (Ding et al. 2017) that cortical activity can entrain to the rate of syllable, phrase and sentence presentation. In the EEG experiments participants were played continuous streams of four-word sentences, where each word was 320 ms long in duration and consisted of only a single syllable. As in Fig. 1 each sentence was composed of a noun phrase and a verb phrase, which both contained two words. Thus these stimuli have a specific frequency at three levels of linguistic structure: syllables at 4/1.28 Hz, phrases at 2/1.28 Hz and sentences at 1/1.28 Hz. The neural responses were analysed using time-frequency decomposition and measures of inter-trial phase coherence (ITPC). Cortical activity was found to be phase-locked to the rate of presentation of syllables, phrases and sentences even though only the syllable frequency is present in the auditory signal itself, the other two frequencies rely on the meaning of the words and the structure of the sentences. This can be interpreted as evidence for neural entrainment to discrete higher-level syntactic structures, as opposed to neural tracking of, for example, transition probabilities during sentence processing.

However, a considerable amount of behavioural data highlights the significance of statistics during language comprehension. At the level of word-statistics, word recognition times can largely be determined by their frequency of occurrence in the language, word reading times have been found to

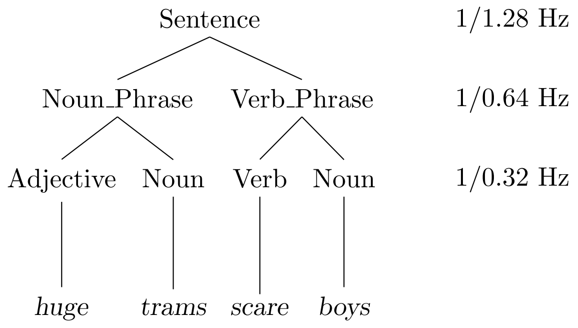

Figure 1: **Tree with frequencies** Each word has a 0.32 second duration. The tree also shows a hierarchical parsing of the sentence; one interpretation of the observed frequencies is that they correspond to these grammatical structures.

closely correlate with word probability and when an ambiguous sentence has to be interpreted, the most probable interpretation is most likely to be chosen (for review see (Jurafsky 2002)). Additionally, both reading times and the amplitude of neural responses are graded by the strength of constraints imposed by prior context on possible sentence continuations (Gibson and Pearlmutter 1998). Language statistics must therefore play a role in human language processing.

In a statistical description, the brain, in a Bayesian manner, exploits the rich statistical structure of language to predict possible identities for syllables, words and phrases and uses these predictions to aid the identification of the actual linguistic input. In this picture, the interpretation of language involves the production, re-evaluation and resolution of predictions, and grammar has evolved out of a kind of 'language game' where speakers aid each other by propagating conventions about how words are arranged, enriching the statistical structure of utterances and syntactic categories are not psychologically real, but epiphenomenal to a statistical-based response to linguistic stimuli.

In this context (Frank and Christiansen 2018) propose a computational model that assumes no higher level of abstraction than a semantic clustering of word representations. Their model represents each word as a high-dimensional vector chosen so that the proximity structure of the vectors matches the statistical relationship between the corresponding words in a large language corpus. The frequency tagged experiment from (Ding et al. 2016), described above, were simulated and the model generated power peaks at the same frequencies as those observed experimentally. This suggests that the peaks could be generated using only the lexical information within the word stimuli, without the need for any knowledge of syntax.

Here human electroencephalography (EEG) is used to measure responses to simple meaningful sentences in comparison to sentences with two different manipulations; nonsense sentences which have been deliberately chosen to have little sensible meaning, and ungrammatical sentences, in which syntax has been destroyed by re-ordering the words. The aim of this experiment is to help to distinguish between the two competing explanations for the phrase- and sentence-level phase locking observed in (Ding et al. 2016, Ding et al. 2017).

Methods

Participants

Eighteen right-handed, native English speakers (11 female, mean age 26 years (range 22 - 32 years)) participated in this study. All participants gave written, informed consent prior to undertaking the

	sensible	nonsense
grammatical	**GS**: huge trams scare boys	**GN**: bored mugs write beds
ungrammatical	**US**: scare trams boys huge	**UN**: write mugs beds bored

Table 1: A summary of the four conditions; the condition label used in the text is in bold, the sentence beside this gives an example.

study and were paid 20 GBP. Participants were screened for dyslexia and hearing impairments. Ethical approval for our experimental procedures were obtained from the University of Bristol Faculty of Science ethics board. All methods were performed in accordance with the relevant guidelines and regulations.

Stimuli

The experimental procedures were similar to those used in a recent EEG study (Ding et al. 2017). Listeners were played English sentences composed of four single-syllable words. Each word was synthesised independently using the MacinTalk Synthesizer (male voice Alex, in Mac OS X 10.7.5). All of the synthesised words (226 - 365 ms) were adjusted to 320 ms duration and volume normalised using the freely available Praat software (Boersma and Weenink 1995–2018).

20 single-syllable words were chosen for each of the four word categories: adjective, subject, verb, object. Words were selected if they were synthesised clearly by the speech synthesizer and if they could be easily categorised into a distinct word category. This was to avoid verbs, such as “ride”, which can also be used as nouns. Nouns were pluralised and all sentences were played in the present tense.

Four conditions were created (see Table 1): sensical and nonsensical grammatical sentences (GS and GN) as well as two ungrammatical conditions in which the words for sensical and nonsensical sentences were re-ordered: the US and UN conditions respectively. For the GS and GN grammatical sentences words from the noun, adjective and verb categories were randomly selected and ordered as

$$\text{adjective, noun}_1\text{, verb, noun}_2.$$

The resulting sentences were then independently ranked in terms of how much sense they made by 290 on-line participants recruited through Prolific Academic. Participants were presented with 110 pairs of sentences, ten of these were an attention trap; participant were asked to press ‘F’ in response to sentences containing the word ‘fish’ and were punished with a time out if they made a mistake. For the remaining 100 pairs they selected the sentence which ‘sounds more normal in everyday speech. Elo chess ranking (Elo 1978) was used to derive individual scores for each sentence from the pairwise comparisons. This established a ranking from sense to nonsense. The top 20 sentences were chosen to form the sensical GS condition and the bottom 20 were chosen to form the nonsensical GN conditions. To obtain the ungrammatical US and UN conditions the words from the GS and GN conditions respectively were re-ordered as

$$\text{verb, noun}_1\text{, noun}_2\text{, adjective.}$$

Experimental Procedures

Each trial contained a sequence of thirteen four word sentences played back to back in a continuous stream. Each word was 320 ms long and trials lasted 16.64 seconds. In total, participants listened to 120 trials, with 30 trials for each of the four conditions. Blocks were made up of five trials from the same condition. Within each block, trials were presented to the participants one after the other with an 800 ms gap between trials. At the end of each block, participants were asked to rate the sentences on a scale of one to five in terms of how much sense the sentences within the trials they had just heard made to them on average using a button press. Following the button press, the next block was played after a delay of 1200 ms. Blocks were presented in a random order and the order of the blocks was counterbalanced across participants. At the end of each block participants were given a ten second break, with a longer two minute break at the halfway point.

EEG Recording

EEG signals were sampled at 1000 Hz from 32 Ag/AgCl electrodes fitted on a standard electrode layout elasticised cap using a BrainAmp DC amplifier (Brain Products GmbH). The EEG was recorded in DC mode, using a low-pass filter of 1000 Hz (fifth-order Butterworth filter with 30 dB/octave). FCz was used as a reference channel. The impedance of the electrodes was kept below 5 kOhms. Recordings were analysed off-line using `Matlab` (Mathworks Inc.) and the `Fieldtrip` toolbox (Oostenveld et al. 2011). Eye blink artefacts were removed using ICA. An independent component was removed if in its topography the mean power over the most frontal four channels (Fp1, Fp2, F7 and F8) was two times greater than the mean power over all other channels, as in (Ding et al. 2017). As our signals of interest are in the low-frequency region, at 1/1.28 (sentences), 2/1.28 (phrases), and 4/1.28 Hz (syllables), the EEG signals were filtered off-line using a 25 Hz low-pass filter (sixth-order Butterworth IIR). Data were re-referenced off-line to a common average reference. For each condition, individual trials (16.64s long) were epoched. Upon sound onset there is a transient EEG response and so the first four syllables (1.28 seconds) in each epoch were removed from the analysis. This meant that the overall length of the analysed part of each trial was 15.36 seconds, corresponding to 48 syllables x 0.32 s.

Data analysis

The EEG signal for each trial was converted into the frequency domain using the discrete Fourier transform with a frequency resolution of 0.065 Hz, that is, 1/15.36 Hz. The complex-valued Fourier coefficient of the trail k, $X_k(f)$, is then used to calculate the inter-trial phase coherence .

The inter-trial phase coherence is defined as

$$R(f) = \frac{1}{K}\left[\left(\sum_k \cos\theta_k(f)\right)^2 + \left(\sum_k \sin\theta_k(f)\right)^2\right] \tag{1}$$

where $\theta_k(f)$ is the phase angle of each complex-valued Fourier coefficient $X_k(f)$ and K is the total number of trials. As with the evoked power, ITPC measures the time-locked response; because it uses only the phase-angle rather than the whole response it does not show the same $1/f$ noise that is present in the evoked power. As such, it is a convenient measure of EEG response to a stimulus with fixed frequency.

As in (Ding et al. 2016) a one-tailed paired t-test was used to test whether the inter-trial phase coherence value in a given frequency bin was significantly stronger than the average of the neighboring four frequency bins, two bins on either side. This was applied to all frequency bins below 5 Hz and a FDR correction for multiple comparisons was applied.

The hypothesis that the grammatical conditions have a larger ITPC peak at the phrase rate than ungrammatical, and that sensible conditions have a higher peak than nonsensical was tested by performing pairwise comparisons using a one-side Wilcoxon signed-rank test. Since there is no hypothesis for comparing GN and US this gives five comparisons.

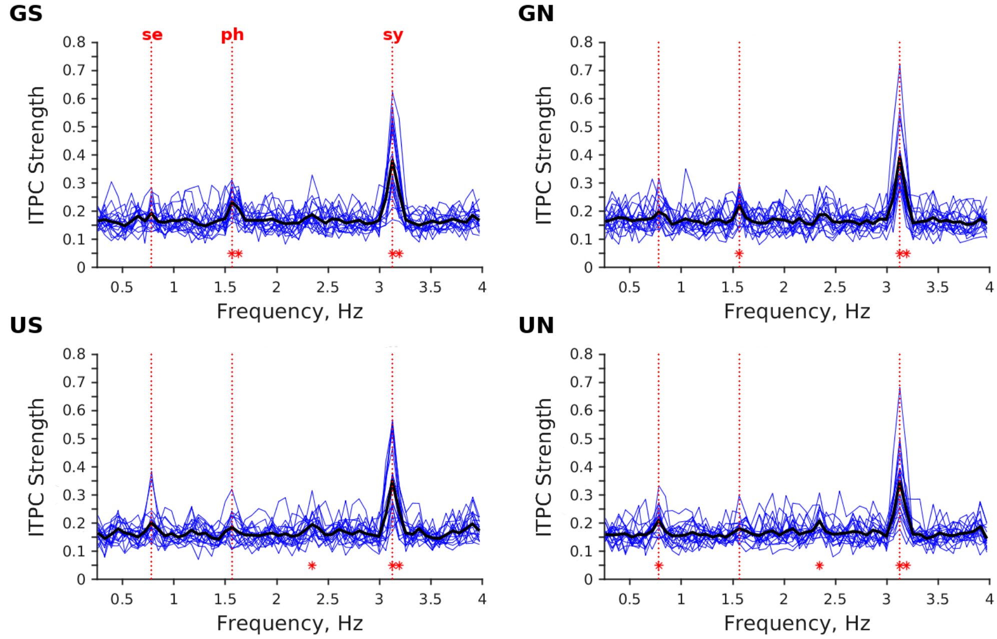

Figure 2: **The spectrum of inter trial phase coherence in the EEG response to sentences from each of the four conditions.** These figures show the grand average of ITPC over all participants and electrodes to each of the four condition; the grand average is in black, the blue lines are averages over all electrodes for the 18 individual participants. The vertical dotted lines are used to locate the frequencies of interest; for GS these are also labelled as 'se' for sentence, 'ph' for phrase and 'sy' for syllable; these labels were not added to the other graphs to prevent clutter. Stars represent statistical significance, after correction, which is equivalent to $p < 0.05$, , **:$p < 0.005$

Results

Significant peaks in the strength of the ITPC were observed in the EEG response at the syllabic (4/1.28 Hz) rate while subjects listened to four word sentences in each of the four conditions (Fig. 2) (GS; $\mu = 0.3806$, $\sigma = 0.1319$, $p < 0.005$, GN; $\mu = 0.3922$, $\sigma = 0.1115$, $p < 0.005$, US; $\mu = 0.3511$, $\sigma = 0.1316$, $p < 0.005$, UN; $\mu = 0.3484$, $\sigma = 0.1265$, $p < 0.005$). Significant peaks in the ITPC were also observed at the phrase rate in both of the grammatical conditions (GS; $\mu = 0.2334$, $\sigma = 0.046$, $p < 0.005$, GN; $\mu = 0.2205$, $\sigma = 0.0547$, $p = 0.0173$) but not in the US and UG conditions. A significant sentential rate peak in the strength of the ITPC was only observed during the UN condition (UN; $\mu = 0.2097$, $\sigma = 0.0509$, $p = 0.0024$).In addition there is a statistically significant peak at 3/1.28 Hz in the US condition (US; $\mu = 0.1958$, $\sigma = 0.0442$, $p = 0.0234$).

The ITPC response at both the sentential and syllabic rate was similar across all of the four conditions (Figure 3). We next compared the strength of the ITPC at the rate of phrase presentation between each of the four experimental conditions. The strength of the ITPC at the phrasal rate is significantly greater in response to the grammatically well-formed sensical sentences (GS) when compared to both the ungrammatical conditions (US and UN); the grammatical nonsensical condition (GN) is significantly larger than UN, ungrammatical nonsense (GS: $\mu = 0.2334$, $\sigma = 0.0447$, GN: $\mu = 0.2205$, $\sigma = 0.0532$, US: $\mu = 0.1845$, $\sigma = 0.0538$, UN: $\mu = 0.1798$, $\sigma = 0.0415$; GS>US: $p = 0.003$, GS>UN: $p = 0.002$, GN>UN: $p = 0.005$). No other comparisons were significant at the

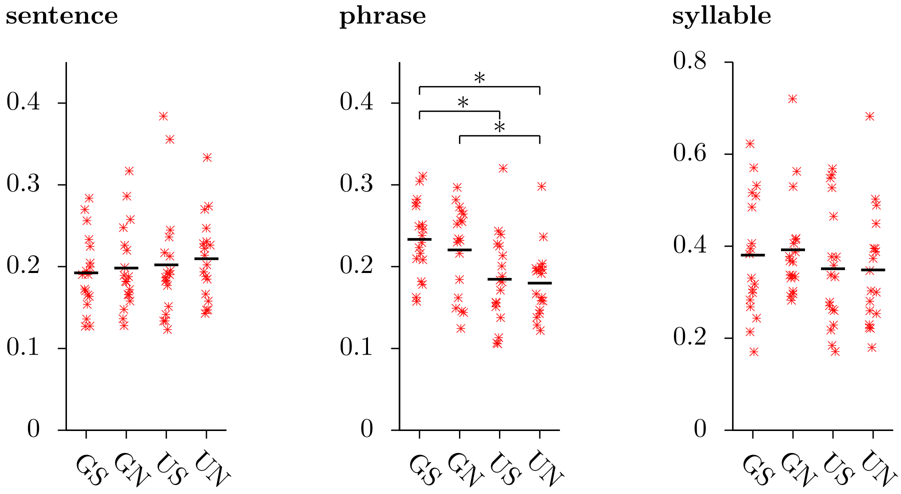

Figure 3: **Comparing ITPC values at frequencies of interest across conditions** Graphs showing average ITPC values at the sentence, phrase and syllable rates (1/1.28 Hz, 1/0.64 Hz, 1/0.32 Hz, respectively) for each of the four conditions tested. Each red cross represents one participant: a small amount of random jitter in the horizontal direction has been added to aid visualization. The thick black lines represent the averages. At the phrasal rate both GS and GN are greater than both US and UN. Stars represent statistical significance, after correction, which is equivalent to *: $p < 0.05$.

phrase rate. There is no hypothesis for the comparing conditions at the sentence and syllable rate; an *ad hoc* comparison using a two-sided Wilcoxon signed-rank test showed no significant differences.

No effect of time was observed in the present study. The ITPC graphs were similar when comparing trials that occurred at the beginning of each experiment compared with trials that occurred later, and when comparing the responses to sentences that were presented during the first half of each trial with the last half of each trial. In cases where conditions US and UN were presented to participants before either conditions GS or GN, the peak in ITPC at the 1/1.28 Hz rate was still evident. Therefore, the peak observed in ITPC at the sentence rate to presentation in the ungrammatical conditions were not the result of the expectancy of four word sentences given by the prior presentation of sentences from the grammatical conditions.

Discussion

In our experiment we examined the effect of grammatical and semantic manipulations on the entrainment of cortical responses to phrases in four-word sentences. In the ungrammatical conditions there were no significant peak in the cortical response at the phrase rate, there are for both grammatical conditions. Furthermore the response for the grammatical condition is significantly higher than for the equivalent ungrammatical condition. There is no significant difference between the responses to equivalent sensical and nonsensical conditions. This indicates that there is a stronger cortical response to grammatically well-formed phrases.

It is possible that the severity of the grammatical manipulation is greater than the semantic one: it is, of course, difficult to compare the two. It is, nonetheless striking that the semantic manipulation appears to have only a very mild effect on the cortical response: the brain appears adept at deducing the phrase structure in “bored mugs write beds” but does not recognize any phrase structure in “scare trams boys huge”. There was, however, a peak in the ITCP at the sentence rate for the ungrammatical nonsense condition: unexpectedly, this was the only condition which showed a significant peak at the sentence rate. The absence of a sentence peak for three of the four conditions can be attributed to noise. It is interesting though that the sentence peak is observed for UN, it would appear that the repetition every fourth word of words in the same lexical category is sufficient to produce a measurable response.

The occurrence of a particular word does not usually allow for good prediction of the words that follow it. As (Pulvermüller 2002), states, it is likely that the regularities governing word sequences likely operate over lexical category. This means that the presentation of a pronoun for example can predict, with high probability, the later occurrence of a compliment verb. For this to be possible abstraction over word category is necessary. Our results are consistient with this but appear, beyond that, to support the view that this abstraction also applies to simple phrases.

Data availability

Data and details of the stimuli can be found at `github.com/conorhoughton/CLiN2020`.

Acknowledgements

CJH and AB acknowledge support through a James S. McDonnell Foundation Scholar Award in Cognition (JSMF #220020239). NK acknowledges the support by the International Laboratory of Social Neurobiology ICN HSE RF Government grant 075-15-2019-1930. We thank Tom Vaughan who ran the Prolific Academic study used to rank the sentences.

References

Boersma, Paul and David Weenink (1995–2018), Praat: doing phonetics by computer [computer program].

Ding, Nai, Lucia Melloni, Aotian Yang, Yu Wang, Wen Zhang, and David Poeppel (2017), Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG), *Frontiers in Human Neuroscience* **11**, pp. 481.

Ding, Nai, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel (2016), Cortical tracking of hierarchical linguistic structures in connected speech, *Nature Neuroscience* **19**, pp. 158–164.

Elo, Arpad E (1978), *The rating of chessplayers, past and present*, Arco Pub.

Frank, Stefan L. and Morten H. Christiansen (2018), Hierarchical and sequential processing of language, *Language, Cognition and Neuroscience* **0**, pp. 1–6.

Gibson, Edward and Neal J Pearlmutter (1998), Constraints on sentence comprehension, *Trends in Cognitive Sciences* **2**, pp. 262–268.

Jurafsky, D (2002), Probabilistic modeling in psycholinguistics: Linguistic comprehension and production, *Probabilistic Linguistics* **30**, pp. 1–50.

Oostenveld, R., P. Fries, E. Maris, and JM Schoffelen (2011), Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data, *Computational Intelligence and Neuroscience* **2011**, pp. 156869.

Pulvermüller, F (2002), A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order, *Progress in Neurobiology* **67**, pp. 85–111.