

OPEN

# Expectancy effects in the EEG during joint and spontaneous word-by-word sentence production in German

Tatiana Goregliad Fjaellingsdal<sup>1\*</sup>, Diana Schwenke<sup>3</sup>, Stefan Scherbaum<sup>3</sup> , Anna K. Kuhlen<sup>4</sup>, Sara Bögels<sup>5,6</sup>, Joost Meekes<sup>1</sup>  & Martin G. Bleichner<sup>1,2</sup> 

Our aim in the present study is to measure neural correlates during spontaneous interactive sentence production. We present a novel approach using the word-by-word technique from improvisational theatre, in which two speakers jointly produce one sentence. This paradigm allows the assessment of behavioural aspects, such as turn-times, and electrophysiological responses, such as event-related-potentials (ERPs). Twenty-five participants constructed a cued but spontaneous four-word German sentence together with a confederate, taking turns for each word of the sentence. In 30% of the trials, the confederate uttered an unexpected gender-marked article. To complete the sentence in a meaningful way, the participant had to detect the violation and retrieve and utter a new fitting response. We found significant increases in response times after unexpected words and – despite allowing unscripted language production and naturally varying speech material – successfully detected significant N400 and P600 ERP effects for the unexpected word. The N400 EEG activity further significantly predicted the response time of the subsequent turn. Our results show that combining behavioural and neuroscientific measures of verbal interactions while retaining sufficient experimental control is possible, and that this combination provides promising insights into the mechanisms of spontaneous spoken dialogue.

The exchange between two persons conversing is stunningly fast: interlocutors take turns speaking and listening at a rapid rate, requiring them to produce and process language simultaneously. The time needed for producing an utterance is commonly longer than the response times observed during natural conversations<sup>1</sup>. This speed can be accomplished by forming expectations, for example of the length of a turn and the meaning of the utterance<sup>2,3</sup>, and then preparing the own response based on these expectations<sup>4</sup>. When expectations are not met, however, language processing is often slowed down<sup>5,6</sup>. Aside from the time needed to process the unexpected event, interlocutors may need additional time to adapt a prepared response to the new unexpected context.

To understand the effect of expectations and expectation violations during interactions, not only behavioural but also neural underpinnings might provide a valuable frame. There is, indeed, a major interest in moving towards a neuroscience of social interaction<sup>7–10</sup>. Interactions, however, are characterized by their openness, while neuroscientific devices impose major constraints to measure sensible brain data. It is challenging to bring these two together<sup>11</sup>.

One setting in which interactive patterns can be observed in an experimentally controllable environment is the word-by-word exercise from improvisational theatre<sup>12,13</sup>. In this game, two persons construct a story together by taking turns for each word. A high degree of coordination is necessary, as the interacting partners have to adapt to each other turn-by-turn in order to produce a meaningful sentence. We believe this level of coordination is achieved by forming expectations about the partner's next utterance. Similar to natural interactions, one can observe that when expectations are not met the player hesitates to produce the next turn. The word-by-word

<sup>1</sup>Department of Psychology, European Medical School, University of Oldenburg, Oldenburg, Germany. <sup>2</sup>Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, Germany. <sup>3</sup>Department of Psychology, Technische Universität Dresden, Dresden, Germany. <sup>4</sup>Institute of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany. <sup>5</sup>Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. <sup>6</sup>Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands. \*email: [tatiana.goregliad@uol.de](mailto:tatiana.goregliad@uol.de)

setting further keeps the principal turn-taking structure of natural interactions intact while giving the possibility to manipulate systematically whether the preceding turn prompts an unexpected sentence completion.

Electroencephalography (EEG) has a high temporal resolution, which can capture the fine temporal structure of verbal interactions. The clear structure of the word-by-word paradigm lends itself for EEG recordings as it ensures valid segmenting of the time periods of interest. EEG studies including semantic expectancy violations have mainly reported modulations of the N400 event-related-potential (ERP), linked to semantic processing, and the P600 ERP linked to syntactic processing<sup>5,14–16</sup>. The N400 effect is characterized by an amplitude modulation in the averaged EEG approximately 400 ms after an unexpected compared to an expected word, indexing sensitivity to semantic expectancy<sup>17–25</sup>. The N400 is often followed by a so-called late positive complex or P600 that has been associated not only with syntactic analysis but also overall re-analysis<sup>16,26</sup> and even semantics<sup>27,28</sup>.

In the present word-by-word study, we want to allow unscripted – though controlled – language production of the participant while producing a sentence interactively with a confederate. We measure the underlying electrophysiological activity while systematically manipulating whether the participants' expectations are met or not. For this purpose, we make use of pictures showing objects that have more than one naming option, i.e., synonyms, with different gender-marked articles in German. To control utterances of participants without making them read aloud, we cued each sentence with a written verb and a picture of an object, both of which had to be included in the sentence. The confederate inserted unexpected sentence continuations (i.e., articles of unexpected gender), where the participant not only had to process the unexpected event but also needed to retrieve and produce a new response deviating from his or her preferred object name.

Behavioural studies on speech production suggest that final word selection among lexical competitors takes place rather late, around 300 ms after picture onset, meaning that multiple candidates are activated at first<sup>29–31</sup>. Neurophysiological studies of overt language production have found that this lexical access manifests as a positive deflection around 200 ms after event onset<sup>32–35</sup>. During speech comprehension, discourse (i.e., sentential context) can aid in pre-activating appropriate lexical representations, leading to behavioural costs and differing neurophysiological responses when expectations are violated<sup>5,13</sup>. Accordingly, we hypothesize that during interactive sentence production, such as in the word-by-word paradigm, a specific lexical candidate (e.g., the preferred naming of the object on the picture) will be pre-activated to produce the next turn as fast as possible<sup>1</sup>. Based on this pre-activation, the player will make predictions about the co-player's preceding turn. When the predictions are violated (i.e., when the confederate utters an unfitting gender-marked article with respect to the pre-activated noun), the player will need to recover by activating one of the less preferred lexical alternatives in order to produce a meaningful sentence.

The successful completion of the word-by-word task entails language comprehension, language production, and for instances of expectancy violations their detection along with the inhibition of a possibly pre-activated response. Our objectives here can be summarized as: (1) testing the feasibility of measuring sensible neural correlates along with behavioural markers in an interactive setup that allows unscripted language production, and (2) measuring some of the neural processes related to successful interactive language use and repair during spontaneous sentence production. For unexpected continuations (i.e., a different gender-marked article from the preferred object noun), we predicted an N400 and P600 ERP effect. Further, we predicted increased turn-times for the next response after encountering an unexpected article. To our knowledge, this is the first study to target interactive language use during EEG measurement with a paradigm that allows such a dynamic sentence production.

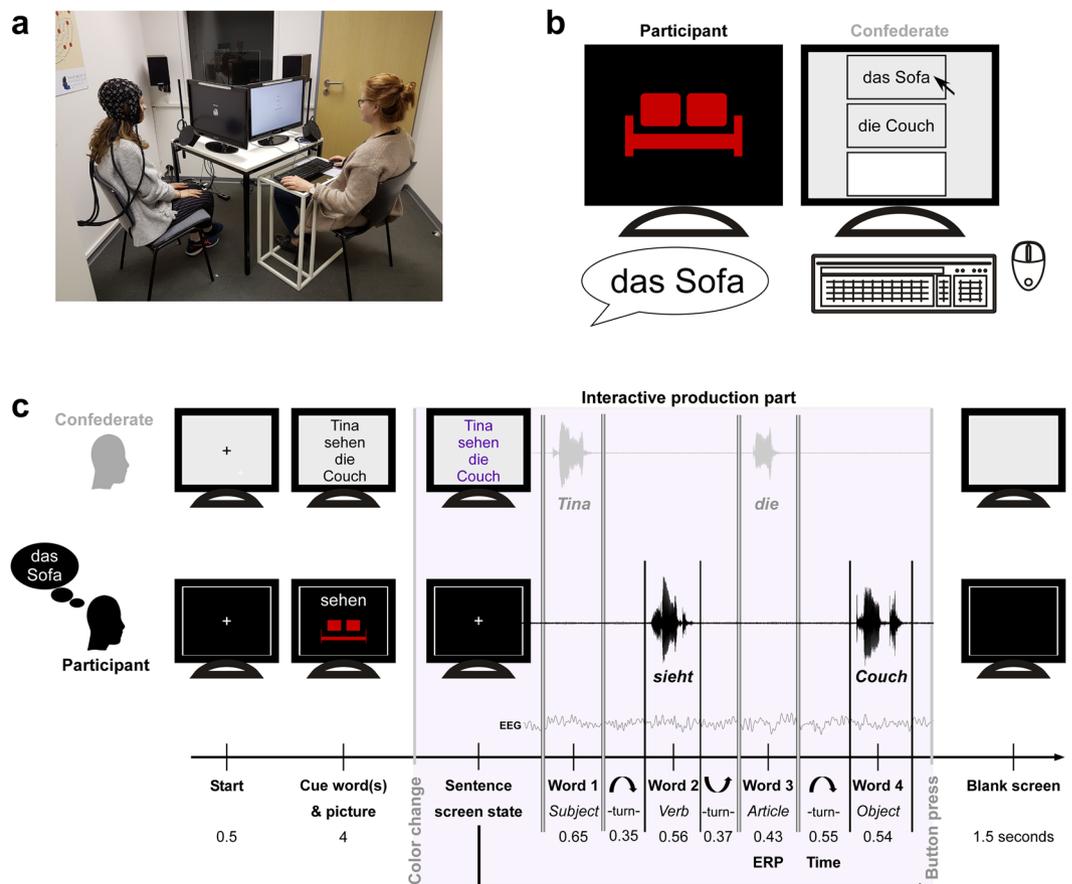
## Results

**Behavioural results.** In the word-by-word task (see Fig. 1c), participants successfully produced a sentence without grammatical errors and with the predefined verb and object in more than 98% of all trials. In critical trials, participants produced on average a correct sentence in 99.49% of expected trials and 97.62% of unexpected trials. Errors were predominantly present in the first of the twelve blocks, indicating a familiarization effect over the experiment and/or problems of the participant understanding the task of constructing a correct sentence. Three participants were reinstructed after the first block to ensure compliance of the task. Erroneous trials (e.g., sentences where the verb was forgotten, an ungrammatical fourth word was uttered or no fourth word was uttered) were excluded from further analysis. On average, the word duration, i.e., the time spent uttering a word, was  $559 \pm 12$  ms for word 2 and  $540 \pm 30$  ms for word 4 ( $543 \pm 38$  ms after expected articles and  $548 \pm 54$  ms after unexpected articles).

Participants needed on average  $351 \pm 59$  ms to produce word 2 (turn-time from offset of word 1 to onset of word 2) and  $554 \pm 193$  ms to produce word 4 (turn-time from offset of word 3 to onset of word 4) over all conditions (filler, congruent, incongruent). Split by critical conditions, participants needed on average after an expected article  $405 \pm 168$  ms and after an unexpected article  $958 \pm 273$  ms to produce word 4 (see Fig. 2a).

In the picture naming task (see Fig. 1b), participants had to name the picture three times. When the same name was used for a picture all three times it received an intra-individual frequency rating of 2, indicating the participant had a strong preference for a word. When the same name for a picture was used twice it received a rating of 1, and when it was named differently each time it received a rating of 0, indicating that the participant did not have a stable and preferred naming of this object. The results show that participants had overall stable naming preferences for all conditions ( $Mdn = 2$ ). For critical pictures (i.e., the ones used in expected and unexpected conditions in the main task), the median intraindividual frequency ranged from 1.5 ( $n = 1$ ) to 2 ( $n = 92$ , all remaining pictures) showing that each individual commonly had a clear preference for one naming. The interindividual frequency of naming was defined per naming used per picture (see also Supplementary Fig. 1). Figure 2c shows the relationship of the interindividual frequency of naming with the needed turn-time in the word-by-word task, indicating longer turn-times to produce words with an overall lower frequency.

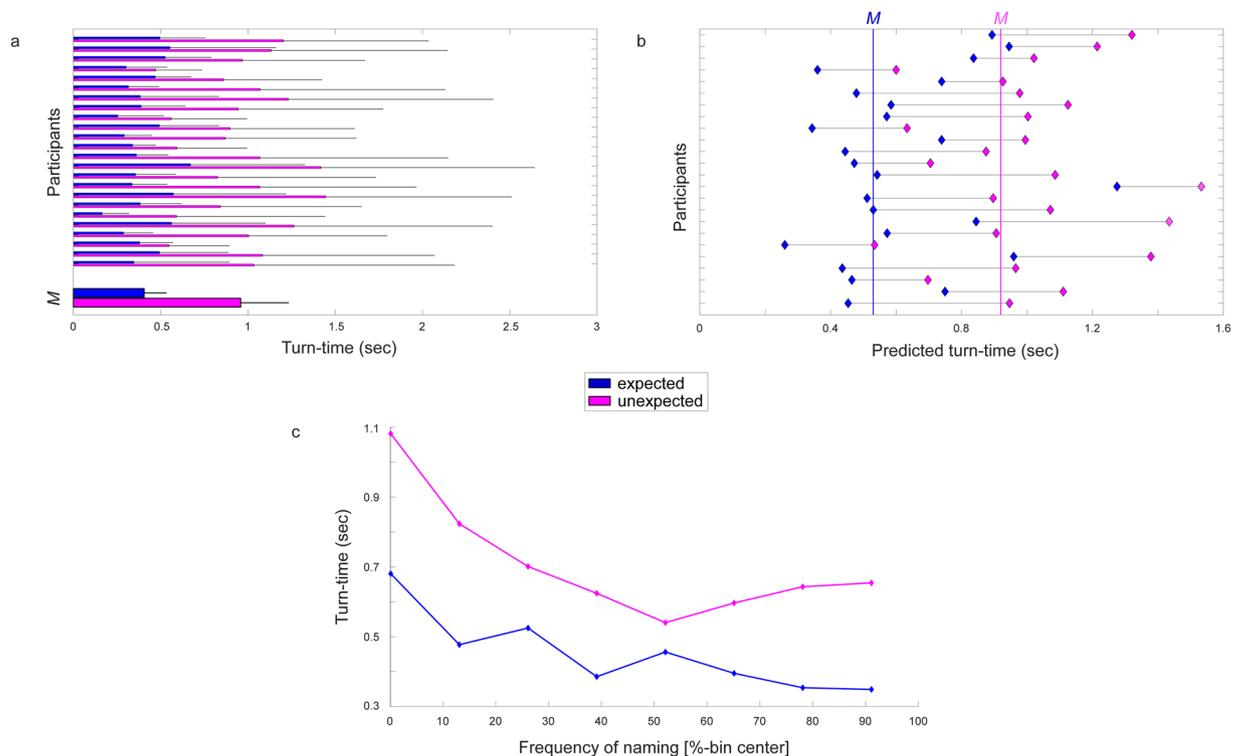
A GLMM with fixed factors expectancy, intra-individual frequency, inter-individual frequency, and random slope congruency nested in intercept participant, as well as random intercept word length showed that



**Figure 1.** Paradigm overview: setup, picture-naming task screen state, and word-by-word task and timeline. **(a)** Setup: Participant with EEG cap sat on the left hand side in front of a computer screen (no study participant is shown). A microphone was placed near her mouth. The confederate sat next to her in front of a second computer screen and a second microphone, as well as a table with keyboard and mouse. **(b)** Picture-naming task: As soon as the participant named the presented picture, the confederate clicked on the respective synonym or typed in a third option. Namings were automatically saved for the experiment. **(c)** Word-by-word timeline: Screen states and streams of participant and confederate. Example of an unexpected trial: participant named the picture ‘das Sofa’, expects the article ‘das’ and hears the article ‘die’. Steady presentation times of fixation cross and words and picture are given in seconds. During the interactive production part (lilac box), the participant saw a steady white fixation cross. The confederate started the sentence with word 1 (sentence subject), the participant then uttered word 2 (sentence verb), the confederate continued with word 3 (sentence article) and the participant ended the sentence with word 4 (sentence object). The interactive production part ended with the button press of the confederate. A blank screen was presented between trials. Average word duration for each word of a sentence, as well as turn-times from word to word are given in seconds. The ERP is computed for the expected/unexpected article (i.e., time-locked to the onset of word 3). The response time is contrasted between offset of word 3 until onset of word 4 after expected and unexpected articles. Sentence translation: “Tina sees the couch.” Cue verb is shown in the infinitive: ‘sehen’ – ‘to see’. Example sofa drawing not used in experiment (real-life object pictures were used).

intra-individual frequency of naming (i.e., intra-individual naming for the same picture) had no significant effect on turn-time ( $\chi^2(2) = 1.88, p = 0.391$ ). Therefore, it was dropped from the model for better model fit (see Supplementary Table 1). Results of the final GLMM (see Fig. 2b) showed that expectancy (expected vs. unexpected) had a significant effect on turn-times ( $\chi^2(1) = 40.69, p < 0.001$ ), with increased turn-times for unexpected events (see Supplementary Table 1). Further, a significant interaction between expectancy and frequency of the preferred naming (i.e., inter-individual naming distribution) was present ( $\chi^2(1) = 40.98, p < 0.001$ ), where higher frequencies of a naming lead to decreased turn-times for expected events and increased turn-times for unexpected events (see Fig. 2c). Excluding the nested random slope for congruency in participant significantly decreased model fit ( $\chi^2(2) = 62.90, p < 0.001$ ), indicating considerable variation between participants in the effect of the expectancy violation (see Fig. 2b).

**EEG results.** The grand average ERP of word 3 for expected and unexpected conditions shows an N400 effect between 250 to 450 ms with a centro-posterior scalp distribution (see Fig. 3). Further, a P600 is apparent between



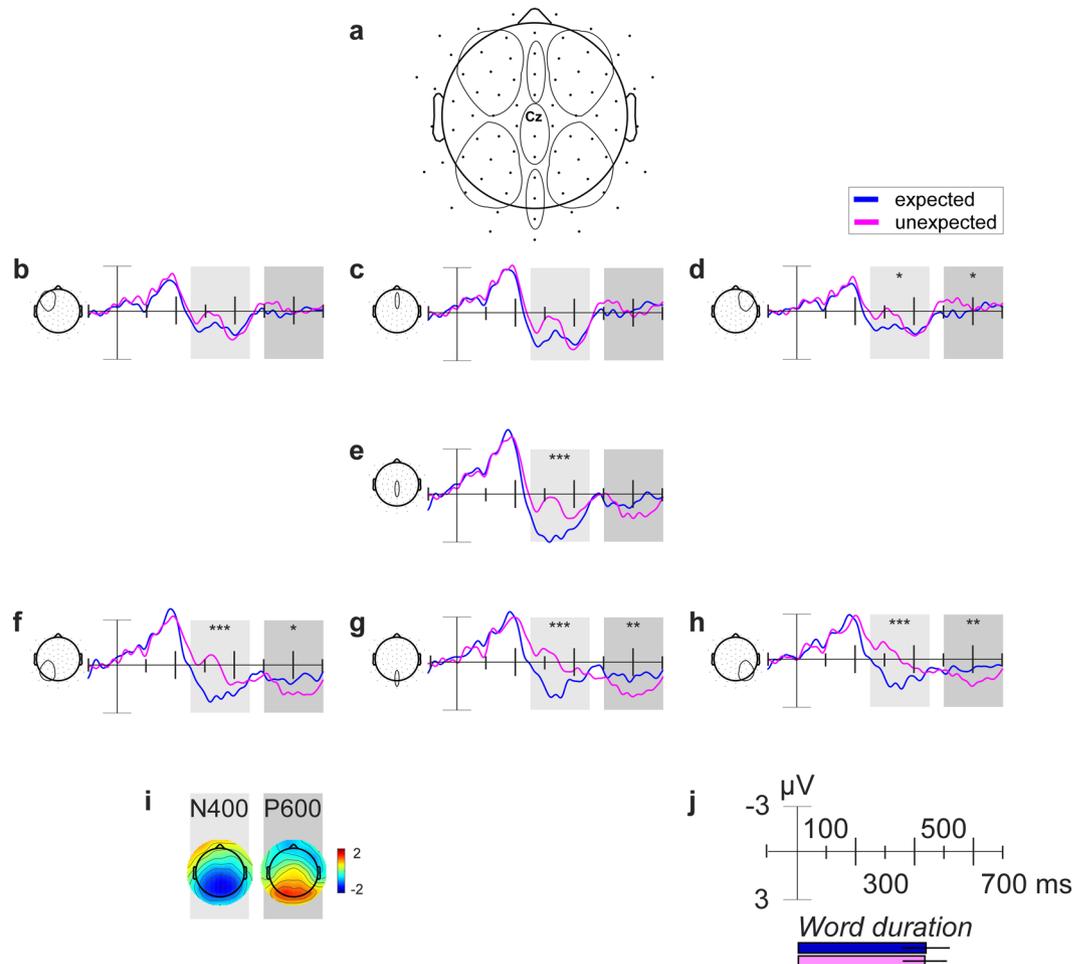
**Figure 2.** Turn-times from word 3 (expected/unexpected article) to word 4 (object noun). **(a)** Average turn-times from offset of word 3 to onset of word 4 are shown in seconds for each participant for expected (blue) and unexpected (magenta) conditions plus standard deviation (black lines). Grand average turn-time ( $M$ ) is shown below. **(b)** Predicted average turn-times from offset of word 3 to onset of word 4 from the calculated GLMM, shown for each participant (y-axis) for expected (blue) and unexpected (magenta) conditions. Grand average predicted turn-time for the expected condition is shown as blue vertical line and for unexpected conditions as magenta vertical line. **(c)** Average turn-times from offset of word 3 to onset of word 4 (y-axis; in seconds) are contrasted to interindividual frequencies of naming (x-axis; percentage bin center points connected by a line).

500 and 700 ms after word onset with a posterior topography (see Fig. 3). The overall distribution of the ERPs (N400 & P600) is shown in Supplementary Fig. 5.

Statistical analysis with Linear Mixed Models (LMM) confirmed that expectancy had a significant effect on the N400 amplitude for word 3 over five of the seven specified ROIs (see Fig. 3a for electrode locations and see Supplementary Table 2 for an overview of all results; e.g., posterior midline:  $\chi^2(1) = 15.48$ ,  $p < 0.001$ , left posterior quadrant:  $\chi^2(1) = 35.14$ ,  $p < 0.001$ , and right posterior quadrant:  $\chi^2(1) = 34.02$ ,  $p < 0.001$ ). The N400 amplitude was significantly more negative for the unexpected compared to the expected condition. The topographical distribution and statistical result indicate a widespread N400 effect with a posterior maximum (see Fig. 3 and Supplementary Table 2). To address concerns of multiple comparisons, we computed a single model with ROI as a fixed factor (all other factors were included like in the separate models). Results were not qualitatively different for expectancy ( $\chi^2(1) = 19.09$ ,  $p < 0.001$ ). ROI had a significant effect on N400 amplitude too ( $\chi^2(1) = 59.22$ ,  $p < 0.001$ ). Moreover, expectancy and ROI significantly interacted with each other ( $\chi^2(1) = 27.74$ ,  $p < 0.001$ ).

Expectancy also significantly modulated the P600 amplitude for word 3 (see Supplementary Table 3) over posterior midline ( $\chi^2(1) = 8.66$ ,  $p = 0.003$ ), left posterior quadrant ( $\chi^2(1) = 6.40$ ,  $p = 0.011$ ), right anterior quadrant ( $\chi^2(1) = 3.96$ ,  $p = 0.046$ ), and right posterior quadrant ( $\chi^2(1) = 8.34$ ,  $p = 0.004$ ). The P600 amplitude was significantly more positive for the unexpected condition compared to the expected condition over all mentioned ROIs, except over the right anterior quadrant, where it was more negative for the unexpected condition. These results indicate a posterior topographical distribution of the P600 effect.

**Brain-behaviour interaction results.** The GLMM calculated for turn-time from word 3 to word 4, including the mean N400 EEG activity from 250 to 450 ms and expectancy condition as explanatory factors showed that for ROIs at central midline, posterior midline, left posterior quadrant, and right posterior quadrant (see Fig. 3a) N400 EEG activity was significantly associated with turn-time (all  $p \leq 0.001$ , see Supplementary Table 4). Specifically, across conditions N400 EEG amplitude was larger on trials with lower turn-times. For ROIs at anterior midline, left anterior quadrant, and right anterior quadrant N400 EEG activity was not significantly associated with turn-time (all  $p > 0.05$ , see Supplementary Table 4). As expected, turn-times were significantly longer in the unexpected than in the expected condition (all  $p < 0.001$ , see Supplementary Table 4). There was no significant interaction between expectancy and N400 EEG activity (all  $p > 0.05$ , see Supplementary Table 4).



**Figure 3.** Grand average ERPs of word 3 shown for midline ROIs and quadrant ROIs for expected (blue) and unexpected (magenta) conditions. Same regions were used for statistical analysis. The N400 effect time window is highlighted in light grey (250–450 ms). The P600 time window is highlighted in dark grey (500–700 ms). Zero point is the onset of word 3, the expected or unexpected article. **(a)** Electrode map shows specific regions of all seven ROIs with highlighted Cz. **(b)** Grand average ERPs of word 3 averaged over left anterior quadrant. **(c)** Grand average ERPs of word 3 averaged over frontal midline. **(d)** Grand average ERPs of word 3 averaged over right anterior quadrant. **(e)** Grand average ERPs of word 3 averaged over central midline. **(f)** Grand average ERPs of word 3 averaged over left posterior quadrant. **(g)** Grand average ERPs of word 3 averaged over posterior midline. **(h)** Grand average ERPs of word 3 averaged over right posterior quadrant. **(i)** Difference grand average topographies for the specific time windows are shown for the N400 and P600 effect. **(j)** ERP plots axis description with mean word duration (of confederate) plus standard deviation below (same time scale). \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

## Discussion

In this study, two persons jointly produced a sentence, taking turns for each word. The word-by-word paradigm is inspired by a technique used in improvisational theatre, which models various aspects of natural interactions. The paradigm's structure allows for high experimental control, along with the ability to induce expectation violations during an interaction. These two pillars make the paradigm an effective tool to study neurophysiological (e.g., N400 and P600 ERP) and behavioural effects (e.g., turn-time) during verbal interaction. In the present study, we could successfully induce expectation violations by making a confederate utter a gender-marked article that did not fit the participants preferred object name's gender that had to be produced in the next turn.

The behavioural findings within this paradigm are what we predicted based on previous research<sup>4,6,36</sup> and our own observations from improvisational theatre, as well as everyday experiences of natural interactions. When expectations are not met, the time to produce the next response increased significantly as compared to when the expectations are met. This finding points to the fact that participants pre-activated their preferred object naming in order to produce the next turn as fast as possible. However, when they encountered an unexpected (unfitting gender-marked) article, they had to discard the pre-activated lexical entry of their preferred object naming and produce the fitting word. Behaviourally, we can capture the consequences that follow from this repair of a violated expectation. This behavioural effect might still reflect numerous underlying processes, which cannot be disentangled easily.

To further our understanding of the underlying mechanisms during word-by-word interactions, neurophysiological underpinnings can help in disentangling some of the crucial aspects for successful interactions. On the neural level, we predicted two main effects, the N400 and P600, to be modulated significantly by expectancy. This was indeed the case, unexpected articles led to a more negative amplitude of the N400 and a more positive amplitude of the P600. We will discuss in the following paragraphs what their presence in this particular setup can tell us about the interplay of language comprehension and language production during verbal interaction.

The first process observed in the EEG, the N400 effect, is known to index processing of expectation violations in various domains (for an overview see<sup>23</sup>). Seeing the N400 effect here is consistent with the idea that the participant pre-activates a specific lexical entry and accompanying grammatical gender during the interaction. The N400 effect is the response to encountering an unfitting gender-marked article to this pre-activated entry. Similar to grounding in conceptual pact studies<sup>37</sup>, i.e., where interacting players agree on a specific term for a specific object, the participant named the pictures pertaining to the objects in the co-constructed sentences prior to the experiment in the presence of the interacting confederate. We deduce that participants ascribed certain expectations to the confederate that she would name the objects the same way they had named them and would therefore utter a fitting gender-marked article. The confederate in fact uttered fitting gender-marked articles to the participants' expected object names in the majority of the trials (70%), rendering the remaining trials unexpected. Similar N400 effects on the article level (i.e., when the article renders a noun with high cloze probability grammatically incorrect) have been reported in a language comprehension task in 2005 by DeLong and colleagues (see also<sup>5</sup>). These N400 effects on article level have been interpreted by the scientific community as strong indicators for prediction during language comprehension<sup>23</sup>, since the article itself constrains the probability of following nouns without defining context in itself. However, DeLong *et al.*'s findings failed to be replicated in a large-scale replication analysis, suggesting that (phonological forms of) words are not necessarily pre-activated during language comprehension<sup>38</sup>. Our word-by-word setup combines language comprehension (of the article) with instant language production (of the following noun). We show that in this context a pre-activation of an object form is indeed present, which shows up as an N400 effect on the article when violated. The N400 was even predictive of the resulting turn-time needed to utter the next word. We conclude that pre-activation of a specific lexical entry aids in accomplishing the present word-by-word task in a rapid manner, common to the timely turn-taking structure of natural interactions<sup>39</sup>. It is an open question, if the pre-activated entry leads to a prepared word (i.e., in the speech production loop) or if it relates to pre-activation that aids in speech preparation after listening to the turn of the partner. In other words, it is unclear if speech production is planned during the turn of the partner or after the turn has finished. The later, positive going ERP we observed in the EEG for unexpected conditions could provide information to answer this question.

The classical account of this positivity we see would be that of a P600 ERP that has been linked to syntactic analysis<sup>26</sup> and discussed as an index for structural reanalysis, for example regarding semantics<sup>40</sup>. The P600 or late positive complex often follows an N400 effect (e.g.,<sup>16</sup>). In the present study, the P600 would then reflect the parsing of the unexpected article with a transfer to new retrieval. Sassenhagen and colleagues<sup>27</sup> for example found the P600 to be response-aligned to the reaction time of a button press. In this line of argument, the P600 reflects the point, when the event has been fully integrated in the sense-making system, opening the transfer to the most suitable response (be it a button press or speech preparation). An argument that links the P600 to the family of the P3 ERP component<sup>27</sup>.

Yet an alternative interpretation of this finding is possible. The late positivity might reflect lexical access: a starting point for the retrieval of the new response after an unexpected article is encountered. In line with speech production research, an early positive EEG component has been discussed as marker for the retrieval of words, which is usually largest at posterior electrode sites<sup>4,32–34</sup>. It is possible that this lexical retrieval component is present at different time points for the expected vs. unexpected condition (see topographies for the expected condition in Supplementary Fig. 4). For the expected condition, we would then interpret the positivity around 200 ms after the onset of the article as a lexical access response (see Supplementary Video), indicating speech preparation while still listening to the turn of the confederate (word duration of the article was on average > 400 ms long). In the unexpected condition, lexical access might be disrupted, marked by a larger N400 amplitude, which reflects the processing of the unexpected article. Lexical retrieval is then delayed (or re-activated) at a later stage, for example around 500 to 700 ms or even later, which can overlap with the interpretation of a P600. Given the considerable differences in task demands of earlier EEG studies on speech production and the present study (e.g., picture naming requiring immediate response vs. delayed response) this interpretation is rather speculative.

A way to test this interpretation is for example to add a control condition, where participants would listen to the unexpected article without having to produce a response thereafter. Having a condition, where the participant does not utter the final object noun leads to two possibilities: (a) one where no one utters the noun or (b) one where the confederate utters the last word instead. In the present study, such a control condition would have changed core aspects of the interactive word-by-word nature of this paradigm. Moreover, it would have prolonged the measurement time significantly. Therefore, we focused here on the word-by-word interaction in a parsimonious setting. We encourage future studies to target this question further.

Also in regard of language comprehension the question remains whether the noun is pre-activated due to the required speech production in the next turn or if the N400 effect on an article can also be found for pure language comprehension scenarios. Assessing the preferred object names of participants prior to a language comprehension task, where they listen to sentences with their preferred and dis-preferred object names could provide a scenario to study this question. However, the advantage of the present word-by-word paradigm is that it allows to study verbal interactions, combining language comprehension and language production. There is evidence for shared neural representations for listening and speaking<sup>41,42</sup>, specifically regarding the access to lexical memory<sup>43</sup>. During conversation, some accounts even suggest that amid language comprehension the prediction of upcoming words and turn in general is enabled by simulating (i.e., silently co-producing) the partner's turn<sup>24,44–46</sup>. In other

words, natural interactive language use would then be characterized by a constant overlap of language comprehension and language production circuits, allowing and relying on prediction of upcoming words.

The degree to which expectations are built and can hence be violated could differ between individuals. Future studies could target the role of such interindividual differences during interactions. We have seen that during the current word-by-word construction participants suffered to a different degree from the expectation violations (see Fig. 2). Such interindividual differences are also visible during joint story building in improvisational theatre. For example, one can observe differences in the response to expectation violations and their repair. Naïve players are often in situations where they cannot come up with a response, while proficient players manage smooth interactions also without knowing the partner beforehand. Grasping these differences with implicit and neurophysiological measures can pave the way to assess the role of learning in coping with unexpected events and its possible transfer to other social situations. The interaction of brain and behaviour in the present study further shows that neural correlates can be predictive of the behavioural outcome.

For manipulating expectancy experimentally, the present paradigm makes use of the fact that in the German language there are gender-marked nouns and accompanying gender-marked articles. At first sight, this approach might seem to restrict this paradigm to German or languages with gender-marked nouns only (e.g., Spanish, such as in<sup>22</sup>). However, one option in English (lacking gender-marked nouns) is to contrast the indefinite article ‘a’ and ‘an’ (see<sup>15</sup>), constraining upcoming nouns to start with either a consonant (i.e., with ‘a’) or a vowel (i.e., with ‘an’). Experimenters’ ingenuity should therefore make it possible to adapt this paradigm to a number of other languages, including those lacking gender-marked nouns.

We now have an experimental framework that combines experimental control with an unscripted verbal interaction. This setup accounts for social and dynamic aspects of language usage<sup>47</sup>, which are usually neglected in an individual setting. We show in this study that we can study neural activity in such interactive situations. The replication of well-established ERPs on expectancy violations underlines the validity of our approach and allows to directly relate our findings to an existing body of ERP studies. With this knowledge, we can now address further questions on how our brains work during interaction. For instance, Nieuwland and colleagues<sup>38</sup> highlighted that phonological forms of words are more likely activated by listeners than by readers, leading to the conclusion that novel approaches to study such predictions in natural settings are needed (see also<sup>48,49</sup>). To understand how expectation building is context dependent, we need to study these processes in realistic, i.e., interactive situations. In real life, e.g. when we only hear fragments of what the other person said due to background noise, we depend much more on our abilities to predict and fill in the missing information<sup>50,51</sup>. Also our understanding of ERP components related to semantic and syntactic processing (i.e., the N400 and P600 ERP) will profit from interactive setups, since a conversation requires a higher involvement than a sole listening situation<sup>52,53</sup>. Lastly, Pickering and Garrod<sup>24,54</sup> strongly argue for a unification of language comprehension and language production, not only within a theoretical framework, but also within empirical research. The word-by-word paradigm comprises both processes – language comprehension and production-, allowing to study the relationship between the two in more detail. The interactive nature of this paradigm, i.e., of two persons interacting on the fly, further adds to the current efforts of understanding the nature of language usage in a more natural interactive habitat.

## Conclusion

Social interactions are complex and marked by multiple levels of processing. Here, we successfully measured neural activity related to linguistic processing during verbal interaction. To our knowledge, this is the first study to measure EEG during expectation violation, where the participant is not only required to comprehend and detect the violation of freely produced speech, but also to inhibit a pre-activated response and retrieve a new response to complete an interactively produced sentence. Our EEG findings revealed two underlying processes of the handling of these expectation violations, with one significant outcome on the behavioural level, i.e., in turn-time. A link of these two measures could be established via a brain-behaviour model, i.e., the N400 effect on the article-level predicted the turn-time to produce the following object noun. We conclude that there is added value in combining both measures, behavioural and neural, to understand the mechanisms of social interactions. Our paradigm can be used to further our knowledge on the role of expectations during (verbal) interactions. This joint assessment was possible with the word-by-word paradigm, which combines verbal interaction with the necessary experimental control.

## Methods

**Participants.** Twenty-five healthy right-handed participants took part in this study. All were German-native speakers, students at the University of Oldenburg and were financially compensated for participation. All participants had normal audition, normal or corrected-to-normal vision, and no history of neurological disease, psychiatric disorder or language disorder (self-report). This was verified in a telephone-based screening interview prior to the invitation. Written informed consent was collected prior to participation. The study was approved by the local ethics committee of the University of Oldenburg. Research was conducted in accordance with the relevant guidelines. One participant was excluded from analysis due to technical difficulties during the recording. The remaining participants were on average  $23.6 \pm 2.5$  years old (14 female).

**Paradigm.** The word-by-word paradigm is inspired by improvisational theatre, where two persons act as one to produce a sentence together, taking turns for each word. In the current experiment, participants’ task was to construct a correct four-word sentence taking turns for each word, together with a confederate (see example 1).

(1)        *Tina*    *sieht*    *das*        *Sofa.*        (English: *Tina sees the sofa.*)  
              Subject   Verb    Article    Object

The participants were cued with a written word (the verb of the sentence in its infinitive form, in example 1: 'sehen'-'see') and a picture showing the object of the sentence (in example 1 the picture of a sofa). The confederate was cued with the whole sentence written out. The participant always uttered the verb and object of the sentence (i.e., second and fourth word). The confederate always uttered the subject and article of the sentence (i.e., first and third word). The paradigm relies on the fact that the German language has three grammatical genders for nouns with accompanying gender-marked definite articles (neuter 'das', feminine 'die', masculine 'der'). Hence, we could manipulate whether the confederate's use of the article matched the participant's individual preferred naming response (see next paragraph) for the sentence-final object. In 30% of the experimental trials, the confederate uttered an unexpected article (e.g., 'die' from 'die Couch') which did not match the participant's preferred object naming (e.g., 'das' from 'das Sofa'). In order to construct a correct sentence, the participant had to adjust to the unexpected article and produce the synonym that fitted this article (i.e., 'Couch' instead of the preferred naming 'Sofa'; see Fig. 1c).

Participants' preferred object naming was assessed prior to the main experiment by a picture-naming task (see Fig. 1b) with three non-sequential naming instances. Participants had to name the object on the picture with its accompanying gender-marked article. The last (third) naming instance was used as preferred naming within the experiment. Having three naming instances, gave an intra-individual measure of frequency (i.e., how stable the naming within participant was). Comparing participants' naming instances over the experiment gave an inter-individual frequency measure of picture naming (i.e., how frequent the used naming was over all participants).

**Material.** The 144 pictures used in this study are partly taken from Bögels, Barr, and colleagues<sup>55</sup>. The 94 critical pictures (49 expected, 45 unexpected) have at least two German naming options, i.e., synonym 1 and synonym 2, where each synonym has a different grammatical gender and therefore different German definite articles (e.g., 'das Sofa' vs. 'die Couch'; see Supplementary Fig. 1). The remaining 50 pictures were selected to have one common German naming option and were used as filler trials to achieve a 70% expected to 30% unexpected ratio.

The written word stimuli consisted of 144 first names (Number of letters:  $M = 5.48 \pm 1.48$ , range 3 to 9; 50% female) and 144 German verbs (Number of letters:  $M = 7.79 \pm 1.74$ , range 5 to 13; non-reflexive). Verbs were selected avoiding common co-occurrences of the verb with a particular object naming.

**Procedure.** The participant first signed informed written consent and filled in two questionnaires (Edinburgh Handedness Inventory<sup>56</sup> and a questionnaire on demographic and physiological information). Then the participant was briefed about the following tasks and that his interacting partner was a confederate. We controlled that participants did not know the confederate (the same person throughout the study).

For the measurement, participant and confederate sat next to each other, each in front of a computer screen (see Fig. 1a). The confederate additionally had a small desk with a keyboard and mouse. Two microphones (ETM-006, Lavalier Microphone) with an audio pop shield were located on an extension on the table and each one was placed approximately 15–20 cm in front of participant and confederate respectively. To keep the microphone recording as clean as possible, participants were instructed to utter only the words of the experiment and to avoid filling utterances (e.g., 'ehm', 'eh') and other vocal noises (e.g., laughs, throat clearing). The picture-naming task was programmed in Matlab R2017b. The word-by-word experiment was programmed with the psychophysics toolbox<sup>57,58</sup> in Matlab R2017b.

First, a picture-naming task with two naming instances per picture was conducted. Participants were asked to name each picture with its respective definite article. Three practice trials were used to familiarize the participant with the task. The 144 pictures (cropped image on black screen) were shown once in a randomized order and then again in a different randomized order (i.e., 288 trials in total) to the participant. The confederate sat alongside the participant and saved the respective responses by clicking on the predefined synonyms or typing and saving the new naming options (see Fig. 1b). The participant was not able to see the screen of the confederate (see seating in Fig. 1a) and was not informed about the task of the confederate.

Thereafter, the EEG cap was fitted and impedances controlled. A second picture-naming task with one naming instance per picture (i.e., 144 trials) was conducted (the same cropped images on black background). The confederate again saved the respective namings, which were used from this second run as target words for the expectancy manipulation in the word-by-word experiment.

Subsequently, the word-by-word experiment (see Paradigm and Fig. 1c) was conducted. Here, the task was to construct a correct four-word sentence, taking turns for each word. For the participant, each trial started with a fixation cross (0.5 sec), followed by the simultaneous presentation of the written verb in infinitive (white letters) in the upper middle centre and the cropped object picture in the lower middle centre of the screen (4 sec). During the interactive production part, the participant's screen displayed a steady fixation cross and the confederate's screen displayed the scripted words of the sentence. The confederate uttered the first word, 'Tina' (sentence subject), the participant then uttered the second word, which had to be conjugated, 'sieht' (sees, sentence verb), the confederate then uttered the third word, 'die' (the, sentence gender-marked article), and the participant uttered the fourth word, 'Couch' (sentence object). The trial was terminated with a button press by the confederate. A blank screen was presented for 1.5 seconds between trials. The article uttered by the confederate could either match the preferred naming of the participant (70% of trials – 'expected') or fit an alternative naming with different grammatical gender (30% of trials – 'unexpected'). We emphasized the importance of producing a correct sentence with the confederate using the verb and picture prior to the trial, without revealing that they would encounter unexpected sentence continuations. Three practice trials (without expectation violations) were conducted to clarify the task. Participants were instructed to keep movement minimal and to use the time of the blank screen between trials for necessary movements. Every 12 trials there was a pause and participants could decide when to continue.

After the experiment, the EEG cap was removed and participants were asked to fill in an evaluation questionnaire. Participants were for example asked to rate how natural the interaction seemed, how pleasant it was, and how pleasant the interaction partner was on a five-point scale from ‘not at all’ to ‘very’ (see Supplementary – Evaluation Results). A complete experimental session lasted around 3 to 3.5 hours.

**EEG recording.** Brain electrical activity was measured with a 96-channel EEG system (BrainProducts, Gilching, Germany). Ag/AgCl electrodes were placed equidistantly with a nose tip reference and centro-frontal ground (EasyCap, Herrsching, Germany). Impedances were kept below 20 k $\Omega$ . Data were digitized with a sampling rate of 500 Hz. EEG stream, marker stream, and audio streams were recorded synchronously using the Lab Recorder from Lab Streaming Layer<sup>59</sup>.

**Audio preprocessing.** Word onsets and offsets were first roughly estimated to create epochs around each single word. Next, the audio signal of these epochs was high-pass FIR filtered at 35 Hz and down sampled to 1470 Hz. The envelope was computed (filter length 300) and a low-pass Butterworth FIR filter at 730 Hz was applied. Thereafter, the root mean square (RMS) and cepstrum (using the Voicebox toolbox<sup>60</sup>) were calculated, and the first and second fundamental frequencies were extracted from the cepstrum (low-pass filtered at 600 Hz). To find the real speech onset, we applied the function ‘findchangepts’ (Matlab toolbox signal processing) on the RMS, which gives a series of markers showing changes in the RMS audio signal. The onset marker was set as valid, if changes were apparent in all calculated signals (RMS, envelope, first and second fundamental frequencies). For speech offset detection, the same procedure was applied with a time-reversed audio signal. Each word segment (from onset to offset) was inspected by ear and adjusted, if necessary.

**Behavioural analysis.** The results of the picture-naming task, where each picture had to be named three times, were assessed. When the same name was used for a picture all three times it received an intra-individual frequency rating of 2, when the same name was used twice it received a rating of 1, and when it was named differently each time it received a rating of 0, indicating that the participant did not have a stable and preferred naming of this object. In addition, we assessed the inter-individual frequency naming per picture across participants and naming instances. For the two predefined synonymic namings, we counted how often each synonym (synonym 1 & synonym 2) was chosen by the participants. All cases in which participants chose another option were combined into a third category (option 3). The respective naming percentage was then computed for each picture (see Supplementary Fig. 1 for an overview). Further, we calculated the frequency of naming for each word and trial to contrast interindividual frequency of naming and needed turn-time. Both measures, intra-individual frequency and inter-individual frequency, were used as additional explanatory variables for the turn-time prediction model (see statistical analysis below). The idea here is that individuals might cope better with expectation violations if they do not have a clear naming preference for the object on the picture.

Average speech durations and word length (number of letters) were calculated for all words. For word 4, the response accuracy was assessed. Trials were classified as erroneous and excluded from the behavioural analyses when the verb was forgotten or when the participant did not utter the fourth word or uttered an ungrammatical and/or wrong word (1.07% of all trials; 1.39% of critical trials, i.e., 0.51% from congruent and 2.38% from incongruent trials).

Turn-times were calculated from word offset of the previous word to word onset. Turn-times over 5 seconds and below 0 seconds were excluded from further analysis. To assess statistical differences of turn-times from word 3 to word 4 between conditions, a Generalized Linear Mixed Model (GLMM<sup>61</sup>); was calculated in R<sup>62</sup> with the lme4 package<sup>63</sup>. To account for non-normally distributed reaction time data, the model was fit with a probability distribution of the Gamma family and an inverse link function<sup>64</sup>. We included the fixed factors expectancy (expected vs. unexpected), intra-individual frequency (intra-individual frequency of naming: zero repetition vs. one repetition vs. two repetitions), and inter-individual frequency across participants (general frequency of naming of the picture with chosen naming; compare section Paradigm). Both latter factors, intra-individual frequency and inter-individual frequency, could have an effect on the time needed to recover from an expectation violation. To account for a possible interaction between inter-individual frequency and expectancy, an interaction term between these two factors was included in the model. A random slope congruency for the random intercept subject was added to model possible inter-individual differences in effect size of expected to unexpected condition. Furthermore, we included the random intercept factor length of word (i.e., the number of letters) in the model. P-values for each factor were calculated by comparing the simpler model (without the factor) to the model including the factor. P-values for an interaction between factors were calculated by comparing the full model with an interaction to the model without interaction.

**EEG analysis.** Preprocessing was performed with EEGLAB<sup>65</sup> in Matlab. For artefact attenuation, we applied extended infomax Independent Component Analysis (ICA) on finite impulse response (FIR) filtered data (1 to 40 Hz). To semi-automatically remove artifactual components, we applied the corrmmap toolbox<sup>66</sup>, as in the standardized procedure explained in Stropahl, Bauer, Debener, & Bleichner<sup>67</sup>. All outer ring electrodes (13 electrodes) were excluded from further analysis, due to increased muscle artefacts.

For ERP analysis, the data was FIR filtered between 0.1 to 30 Hz and re-referenced to average mastoids. The data was epoched from –500 to 1500 ms around word onset as determined by the microphone signal (see Audio Preprocessing) and baseline corrected from –100 to 0 ms. An automatic epoch rejection was applied, where all epochs exceeding three SDs from the mean signal were removed. Further, we applied an automatic artifactual channel detection (EEGLAB function) and interpolated channels exceeding a kurtosis value of 5. Per participant and grand average ERPs were calculated.

Two ERP components for the critical word (CW; word 3) were of main interest: the N400 ERP and the P600 ERP. N400 effects for nouns are usually defined between 300 and 500 ms<sup>23,68,69</sup>, however, N400 effects on articles have been reported to arise earlier, e.g., starting around 200 ms<sup>5,15</sup>. For this reason, we used an earlier time point and analysed the N400 effect using the mean amplitude between 250 and 450 ms. The P600 or late positive complex is commonly captured around 500 to 900 ms<sup>26,69</sup>. Due to the present task including speaking shortly after the article was perceived, we analysed the P600 effect using the mean amplitude between 500 and 700 ms. The cut-off after 700 ms circumvented speech production artefacts in the analysed time segment. The mean number of trials entering statistical analysis for word 3 were 43 and 38, for expected and unexpected conditions respectively (filler trials were excluded from analysis). A Linear Mixed Model (LMM) was calculated with the lme4 package<sup>63</sup> in R<sup>62</sup> separately for each component and each specified region of interest (ROI; see Fig. 3a for locations) along the midline (1 - anterior, 2 - central, and 3 - posterior; n = 3 electrodes per ROI) and along the quadrants (1 - left anterior, 2 - left posterior, 3 - right anterior, and 4 - right posterior; n = 8 electrodes per ROI). Fixed factor in each model was expectancy (expected vs. unexpected). We included a random intercept for participant and a random slope expectancy within participant. P-values for expectancy were calculated by comparison of the model with and without expectancy factor.

To address concerns of multiple comparisons, we computed a single model with ROI as a fixed factor (all other factors were included like in the separate models) for the N400 component. Further, we compared this single model with and without interaction of ROI and expectancy.

**Brain-behaviour interaction analysis.** To test the interaction between brain and behaviour (compare for example<sup>70</sup>), a GLMM was calculated in R<sup>62</sup> with the lme4 package<sup>63</sup>. Individual turn-times from offset of word 3 to onset of word 4 were the response variable and fixed factors were the respective mean N400 EEG activity between 250 and 450 ms over the specified ROIs (see Fig. 3) and expectancy (expected vs. unexpected). An interaction term between both factors was added to the model. Similar to the GLMM calculated for the behavioural analysis, random factors included word length (intercept) and expectancy (slope) in participant (intercept). P-values for each factor were calculated by comparing the simpler model (without the factor) to the model including the factor. P-values for an interaction between factors were calculated by comparing the full model with an interaction to the model without interaction.

## Data availability

Data cannot be made available due to data security requirements of ethic approval.

Received: 12 March 2019; Accepted: 10 March 2020;

Published online: 25 March 2020

## References

- Levinson, S. C. & Torreira, F. Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* **6**, 1–17 (2015).
- Bögels, S. & Torreira, F. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phon.* **52**, 46–57 (2015).
- Garrod, S. & Pickering, M. J. The use of content and timing to predict turn transitions. *Front. Psychol.* **6**, 1–12 (2015).
- Bögels, S., Magyari, L. & Levinson, S. C. Neural signatures of response planning occur midway through an incoming question in conversation. *Sci. Rep.* **5**, 1–11 (2015).
- van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V. & Hagoort, P. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 443–467 (2005).
- Ditman, T., Holcomb, P. J. & Kuperberg, G. R. An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology* **44**, 927–935 (2007).
- Pfeiffer, U. J., Timmermans, B., Vogeley, K., Frith, C. D. & Schilbach, L. Towards a neuroscience of social interaction. *Front. Hum. Neurosci.* **7**, 8–9 (2013).
- Bögels, S. & Levinson, S. C. The Brain Behind the Response: Insights Into Turn-taking in Conversation From Neuroimaging. *Res. Lang. Soc. Interact.* **50**, 71–89 (2017).
- Hari, R. & Kujala, M. V. Brain Basis of Human Social Interaction: From Concepts to Brain Imaging. *Physiol. Rev.* **89**, 453–479 (2009).
- Kuhlen, A. K., Allefeld, C., Anders, S. & Haynes, J. Towards a Multi-Brain Perspective on Communication in Dialogue. in *Cognitive neuroscience of natural language use* (ed. Willems, R. M.) 182–200 (Cambridge University Press, 2015). <https://doi.org/10.1017/CBO9781107323667.009>.
- van Berkum, J. J. A. The electrophysiology of discourse and conversation. in *The Cambridge handbook of psycholinguistics* (eds. Spivey, M., McRae, K. & Joanisse, M.) 589–612, <https://doi.org/10.1017/CBO9781139029377.031> (2012).
- Himberg, T., Hirvenkari, L., Mandel, A. & Hari, R. Word-by-word entrainment of speech rhythm during joint story building. *Front. Psychol.* **6**, 1–6 (2015).
- Goregliad Fjaellingsdal, T., Schwenke, D., Ruigendijk, E., Scherbaum, S. & Bleichner, M. G. Studying brain activity during word-by-word interactions using wireless EEG. *PLoS One*. (in press).
- Goregliad Fjaellingsdal, T., Ruigendijk, E., Scherbaum, S. & Bleichner, M. G. The N400 effect during speaker-switch-towards a conversational approach of measuring neural correlates of language. *Front. Psychol.* **7**, 1–10 (2016).
- DeLong, K. A., Urbach, T. P. & Kutas, M. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* **8**, 1117–1121 (2005).
- DeLong, K. A., Quante, L. & Kutas, M. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia* **61**, 150–162 (2014).
- Kutas, M. & Hillyard, S. A. Event-related brain potentials to grammatical errors and semantic anomalies. *Mem. Cognit.* **11**, 539–550 (1983).
- Bögels, S., Kendrick, K. H. & Levinson, S. C. Never say no... How the brain interprets the pregnant pause in conversation. *PLoS One* **10**, 1–15 (2015).
- Coch, D., Bares, J. & Landers, A. ERPs and morphological processing: the N400 and semantic composition. *Cogn. Affect. Behav. Neurosci.* **13**, 355–370 (2013).
- Hagoort, P. & Brown, C. M. ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia* **38**, 1518–1530 (2000).

21. Van Petten, C. & Luka, B. J. Prediction during language comprehension: Benefits, costs, and ERP components. *Int. J. Psychophysiol.* **83**, 176–190 (2012).
22. Wicha, N. Y. Y., Moreno, E. M. & Kutas, M. Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *J. Cogn. Neurosci.* **16**, 1272–1288 (2004).
23. Kutas, M. & Federmeier, K. D. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).
24. Pickering, M. J. & Garrod, S. An integrated theory of language production and comprehension. *Behav. Brain Sci.* **36**, 329–347 (2013).
25. Kutas, M. & Hillyard, S. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science (80-)*. **207**, 203–205 (1980).
26. Kaan, E., Harris, A., Gibson, E. & Holcomb, P. The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Process.* **15**, 159–201 (2000).
27. Sassenhagen, J., Schlesewsky, M. & Bornkessel-Schlesewsky, I. The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain Lang.* **137**, 29–39 (2014).
28. van de Meerendonk, N., Kolk, H. H. J., Vissers, C. T. W. M. & Chwilla, D. J. Monitoring in language perception: Mild and strong conflicts elicit different ERP patterns. *J. Cogn. Neurosci.* **22**, 67–82 (2010).
29. Jescheniak, J. D. & Schriefers, H. Discrete serial versus cascaded processing in lexical access in speech production: Further evidence from the coactivation of near-synonyms. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1256–1274 (1998).
30. Jescheniak, J. D., Hantsch, A. & Schriefers, H. Context effects on lexical choice and lexical activation. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 905–920 (2005).
31. Peterson, R. R. & Savoy, P. Lexical selection and phonological encoding during language production. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 539–557 (1998).
32. Aristei, S., Melinger, A. & Abdel Rahman, R. Electrophysiological chronometry of semantic context effects in language production. *J. Cogn. Neurosci.* **23**, 1567–1586 (2011).
33. Costa, A., Strijkers, K., Martin, C. & Thierry, G. The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proc. Natl. Acad. Sci.* **106**, 21442–21446 (2009).
34. Baus, C. *et al.* On predicting others' words: Electrophysiological evidence of prediction in speech production. *Cognition* **133**, 395–407 (2014).
35. Ganushchak, L. Y., Christoffels, I. K. & Schiller, N. O. The use of electroencephalography in language production research: A review. *Front. Psychol.* **2**, 1–6 (2011).
36. Levinson, S. C. Turn-taking in Human Communication - Origins and Implications for Language Processing. *Trends Cogn. Sci.* **20**, 6–14 (2016).
37. Brennan, S. E. & Clark, H. H. Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1482–1493 (1996).
38. Nieuwland, M. S. *et al.* Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *Elife* **7**, 1–24 (2018).
39. Stivers, T. *et al.* Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci.* **106**, 10587–10592 (2009).
40. Kuperberg, G. R. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Res.* **1146**, 23–49 (2007).
41. Menenti, L., Gierhan, S. M. E., Segaert, K. & Hagoort, P. Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* **22**, 1173–1182 (2011).
42. Awad, M., Warren, J. E., Scott, S. K., Turkheimer, F. E. & Wise, R. J. S. A common system for the comprehension and production of narrative speech. *J. Neurosci.* **27**, 11455–11464 (2007).
43. Segaert, K., Menenti, L., Weber, K., Petersson, K. M. & Hagoort, P. Shared syntax in language production and language comprehension—An fMRI study. *Cereb. Cortex* **22**, 1662–1670 (2012).
44. Pickering, M. J. & Gambi, C. Predicting while comprehending language: A theory and review. *Psychol. Bull.* **144**, 1002–1044 (2018).
45. Pickering, M. J. & Garrod, S. Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* **11**, 105–110 (2007).
46. Kuhlen, A. K. & Abdel Rahman, R. Having a task partner affects lexical retrieval: Spoken word production in shared task settings. *Cognition* **166**, 94–106 (2017).
47. Holtgraves, T. M., Hoeks, J. C. J. & Brouwer, H. Electrophysiological research on conversation and discourse processing. in *Oxford Handbook of Language and Social Psychology* 365–386 (The Oxford Handbook of Language and Social Psychology, 2014). <https://doi.org/10.1093/oxfordhb/9780199838639.013.024>.
48. Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
49. Hamilton, L. S. & Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Lang. Cogn. Neurosci.* **1–10**, <https://doi.org/10.1080/23273798.2018.1499946> (2018).
50. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–253 (2013).
51. Kuperberg, G. R. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Lang. Cogn. Neurosci.* **31**, 602–616 (2016).
52. Brouwer, H. & Hoeks, J. C. J. A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Front. Hum. Neurosci.* **7**, 1–12 (2013).
53. Brouwer, H., Fitz, H. & Hoeks, J. Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Res.* **1446**, 127–143 (2012).
54. Pickering, M. J. & Garrod, S. Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* **27**, 169–190 (2004).
55. Bögels, S., Barr, D. J., Garrod, S. & Kessler, K. Conversational interaction in the scanner: Mentalizing during language processing as revealed by MEG. *Cereb. Cortex* **25**, 3219–3234 (2015).
56. Oldfield, R. C. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
57. Pelli, D. G. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
58. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
59. Kothe, C. Lab streaming layer (lsl). (2014).
60. Brookes, M. Voicebox: Speech processing toolbox for Matlab. (2018).
61. Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390–412 (2008).
62. R Team. R: A language and environment for statistical computing. (2019).
63. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
64. Lo, S. & Andrews, S. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Front. Psychol.* **6**, 1–16 (2015).
65. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
66. Campos Viola, F. *et al.* Semi-automatic identification of independent components representing EEG artifact. *Clin. Neurophysiol.* **120**, 868–877 (2009).

67. Stropahl, M., Bauer, A. K. R., Debener, S. & Bleichner, M. G. Source-Modeling auditory processes of EEG data using EEGLAB and brainstorm. *Front. Neurosci.* **12**, 1–11 (2018).
68. Kos, M., Vosse, T., Van den Brink, D. & Hagoort, P. About edible restaurants: Conflicts between syntax and semantics as revealed by ERPs. *Front. Psychol.* **1**, 1–11 (2010).
69. Kaan, E. Event-related potentials and language processing: A brief overview. *Lang. Linguist. Compass* **1**, 571–591 (2007).
70. Frömer, R., Maier, M. & Abdel Rahman, R. Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Front. Neurosci.* **12**, 1–15 (2018).

## Acknowledgements

We thank Marike Maack for her help during data collection. We would also like to thank Reiner Emkes and Florian Wiese for technical support. We are grateful to Prof. Stefan Debener for his support, feedback, and provision of lab facilities. This research was funded by the Volkswagen Foundation (European Platform) and the Task Group 7 “BCI for Hearing Aids”, DFG Cluster of Excellence 1077 Hearing4all, Oldenburg, Germany.

## Author contributions

The study was conceptualized by T.G., M.B., S.S., D.S. and A.K. The materials used in this study were designed by T.G., D.S., M.B., S.S. and S.B. Experimental setup was enabled by T.G. and M.B. The data was analysed by T.G., M.B. and J.M. The main manuscript was written by T.G. and M.B. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-62155-z>.

**Correspondence** and requests for materials should be addressed to T.G.F.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020