Milena Rabovsky

Change in a probabilistic representation of meaning can account for N400 effects on articles: a neural network model

Suggested citation referring to the original publication:
Neuropsychologia 143 (2019) , Art. 107466
DOI https://doi.org/10.1016/j.neuropsychologia.2020.107466

Neuropsychologia 143 (2020) 107466

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model

Milena Rabovsky

Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476, Potsdam, Germany

ARTICLE INFO

Keywords:
N400
ERPs
Prediction
Neural networks
Cue validity
Meaning

ABSTRACT

Increased N400 amplitudes on indefinite articles (a/an) incompatible with expected nouns have been initially taken as strong evidence for probabilistic pre-activation of phonological word forms, and recently been intensely debated because they have been difficult to replicate. Here, these effects are simulated using a neural network model of sentence comprehension that we previously used to simulate a broad range of empirical N400 effects. The model produces the effects when the cue validity of the articles concerning upcoming noun meaning in the learning environment is high, but fails to produce the effects when the cue validity of the articles is low due to adjectives presented between articles and nouns during training. These simulations provide insight into one of the factors potentially contributing to the small size of the effects in empirical studies and generate predictions for cross-linguistic differences in article induced N400 effects based on articles' cue validity. The model accounts for article induced N400 effects without assuming pre-activation of word forms, and instead simulates these effects as the stimulus-induced change in a probabilistic representation of meaning corresponding to an implicit semantic prediction error.

1. Introduction

The N400 component of the event-related brain potential (ERP) has received a great deal of attention because it provides an electrophysiological indicator of meaning processing in the brain (Kutas and Federmeier, 2011). One issue that has been addressed using the N400 component and has triggered intense debates is the question to what extent and in how much detail (i.e., at which levels of representation) upcoming input is predicted and thus pre-activated during language comprehension. Specifically, in a landmark study DeLong and colleagues obtained larger N400 amplitudes to articles incompatible with an expected noun such as e.g., 'The day was breezy so the boy went outside to fly an … ' where 'an' is incompatible with the expected continuation 'kite' (DeLong et al., 2005). This incompatibility between "an" and "kite" is based on a phonological regularity in English, whereby the singular indefinite article is phonologically realized as "an" before words beginning with a vowel sound and as "a" before words beginning with a consonant sound (e.g., "an airplane" and "a kite"). To decide whether an indefinite article is compatible with an expected noun it is necessary to know the phonological form of the next word (i.e., whether it starts with a consonant or a vowel). Therefore, the result of larger N400 amplitudes on indefinite articles incompatible with expected nouns has often been taken to indicate prediction of phonological word forms.

Earlier studies had already shown that N400 amplitudes are reliably reduced for predictable language input (see Kutas and Federmeier, 2011, for a review) such as for instance demonstrated by influences of cloze probability, which refers to the percentage of participants continuing a sentence fragment with a specific word in offline sentence completion tasks. For example, N400 amplitudes are smaller for high cloze probability continuations such as "Don't touch the wet *paint*" as compared to plausible low cloze probability continuations such as "Don't touch the wet *dog*" (Kutas and Hillyard, 1984). However, in most studies, it is difficult to unequivocally decide whether reduced N400 amplitudes reflect facilitated processing due to prediction/pre-activation of upcoming input (Kutas and Federmeier, 2000; Lau et al., 2008), or whether reduced N400 amplitudes reflect facilitated bottom-up processing because the incoming input better fits the preceding context (Brown and Hagoort, 1993). Importantly, N400 effects induced by articles that do not differ in meaning but just in their form (i.e., 'a' and 'an') are hard to explain by differences in bottom-up processing, as they seem to fit the semantic context equally well. Therefore, the observed N400 effects on articles have been taken as strong evidence for pre-activation of upcoming input during language

E-mail address: milena.rabovsky@uni-potsdam.de.

https://doi.org/10.1016/j.neuropsychologia.2020.107466
Received 17 March 2019; Received in revised form 10 April 2020; Accepted 12 April 2020
Available online 18 April 2020

comprehension (DeLong et al., 2005).

Pre-activation of which aspect of the upcoming input is reflected in these article-induced N400 effects? As noted above, article induced N400 effects are often cited as evidence for prediction of upcoming language input at the level of phonological word forms (e.g., Altmann and Mirković, 2009; Hagoort, 2017; Pickering and Garrod, 2013). The current study aims to offer an alternative possible perspective, suggesting that article-induced N400 effects do not require prediction of word forms, but rather can be accounted for by the change in the conditional probabilities of semantic features, which is cued by encountering the articles (Fig. 1).[1] In a recent study we simulated a broad range of 16 empirically observed N400 effects by treating N400 amplitudes as the change in a neural network model's hidden layer activation state, which probabilistically represents expected sentence meaning (Rabovsky et al., 2018; see also Rabovsky and McClelland, 2020, for discussion). From the perspective implemented in the model, at any given point in sentence presentation, the listener or reader probabilistically predicts all aspects of meaning of the described event based on the experience of statistical regularities in the environment, and N400 amplitudes reflect the change in this prediction induced by the new incoming stimulus, corresponding to an implicit prediction error or Bayesian surprise (Itti and Baldi, 2006) at the level of meaning (see also Kuperberg and Jaeger, 2016; Rabovsky and McRae, 2014). From this view, encountering an indefinite article increases the represented probabilities of semantic features of things consistent with the article and decreases the represented probabilities of semantic features of things inconsistent with the article. This shift in the represented probabilities of semantic features is suggested to underlie N400 effects on articles. The model does not assume separate stages of lexical access to word meaning versus semantic integration of word meaning into the sentence context, but instead assumes that each incoming word provides cues constraining a probabilistic representation of sentence meaning, and N400 amplitudes reflect the change in this probabilistic representation induced by the word. The main goal of the current study is to demonstrate via explicit simulations how this perspective can mechanistically account for article induced N400 effects, thus offering an alternative to the original and still very common interpretation of the empirical effects. A number of alternative models of N400 amplitudes have been proposed (Brouwer et al., 2017; Fitz and Chang, 2019; Frank et al., 2015; Cheyette and Plaut, 2017; Laszlo and Armstrong, 2014; Laszlo and Plaut, 2012), but as yet none of them has simulated article induced N400 effects.

An issue concerning article induced N400 effects, which has recently been intensely debated, is that these effects have not always been replicated (for discussion see DeLong et al., 2017; Ito, Martin and Nieuwland, 2017b, 2017a), and in any case seem to be considerably smaller than observed in the original study (Nieuwland et al., 2018). It has been suggested that this might be due to the fact that articles do not deterministically predict specific nouns in natural language (Ito et al., 2017b; Nieuwland et al., 2018). A second goal of the current study was to directly manipulate this factor, i.e. the cue validity of the articles concerning upcoming meaning in the long-term linguistic environment, to demonstrate its impact on article induced N400 effects in the model. This was done by including one simulation where articles deterministically predict the upcoming nouns in the training environment (Simulation 1) and one simulation where this predictive relationship vanishes due to adjectives presented between articles and nouns during training (in analogy to, e.g., 'an old kite'; Simulation 2).

[1] Please note that this perspective is compatible with the idea of word form prediction, it just suggests that the observed N400 effects on articles may not speak to the issue (see Discussion section).

2. Methods

2.1. Model architecture and processing

N400 amplitudes are simulated using the Sentence Gestalt (SG) model (Rabovsky et al., 2018; St. John and McClelland, 1990), as displayed in Fig. 1.

2.2. Environment and training

The model environment consists of sentences (presented word by word at the input layer) such as 'In the afternoon, the man waters a pine in the garden' each paired with a corresponding event description, specifying who does what to whom in the described event, i.e., who is the agent of the event, what is the action in the event, what is the patient or object in the event, etc. Thus, the event specification consists of a set of pairs of thematic roles (e.g., agent, action, etc.) and their fillers (e.g., man, watering, etc.). The pairs of sentences and corresponding event specifications are probabilistically generated online during training according to pre-specified constraints. After each presented word (represented by a word-specific unit at the input layer), the model is probed concerning all aspects of the event described by the sentence in the query network (see Fig. 1). Responding to a probe consists in completing a role-filler pair when probed with either a thematic role (i.e., agent, action, patient, location, or situation; each represented by an individual unit at the probe and output layer) or a filler of a thematic role (e.g., the man, to water, the pine, etc.). Each filler concept is represented by a number of semantic feature units at the output layer; the semantic features are handcrafted to create graded similarities between concepts roughly corresponding to real world similarities (Rabovsky et al., 2018). For each response, the model's activation at the output layer is compared with the correct output, the gradient of the cross-entropy error measure for each connection weight and bias term in the query network is back-propagated through this part of the network, and the corresponding weights and biases are adjusted accordingly. At the SG layer, the gradient of the cross-entropy error measure for each connection weight and bias term in the update network is collected for the responses on all the probes for each word before being back-propagated through this part of the network and adjusting the corresponding weights and biases.

For the current simulations, changes to the previous model environment and training (described in detail in Rabovsky et al., 2018) were kept to a minimum, while including the characteristics necessary to address N400 effects on articles. Specifically, for Simulation 1, the model environment was adjusted to include articles, and for Simulation 2, it was further extended to include adjectives in addition to the articles. As for our previous simulations, 10 independently initialized models (with initial weights randomly varying between ± .05) were exposed to 800,000 probabilistically generated example sentences, and a learning rate of 0.00001 and momentum of 0.9 was used throughout.

Adjustments for Simulation 1. For Simulation 1, two input units were added, representing the two indefinite articles ('a' and 'an'), which were not associated with specific semantic features at the output layer. These indefinite articles were presented during training at the sentence position prior to the objects in the sentences, such as e.g. 'The man waters a pine'. Crucially, during training each of ten specific actions (e.g. 'water') is followed by either a high probability object (e.g., 'pine' with a probability of .7) or a low probability object (e.g., 'oak' with a probability of .3), and the training environment was constructed such that for each action, the high and low probability objects differ in terms of the appropriate article (e.g., 'The man waters a pine' vs. '… an oak'). Across all actions, both articles are linked equally often to the high versus low probability objects so that the articles do not differ in overall frequency. Importantly, because the indefinite articles always preceded the objects during training, after being presented with sentence beginnings such as 'The man waters an … ' the trained model can predict that the sentence

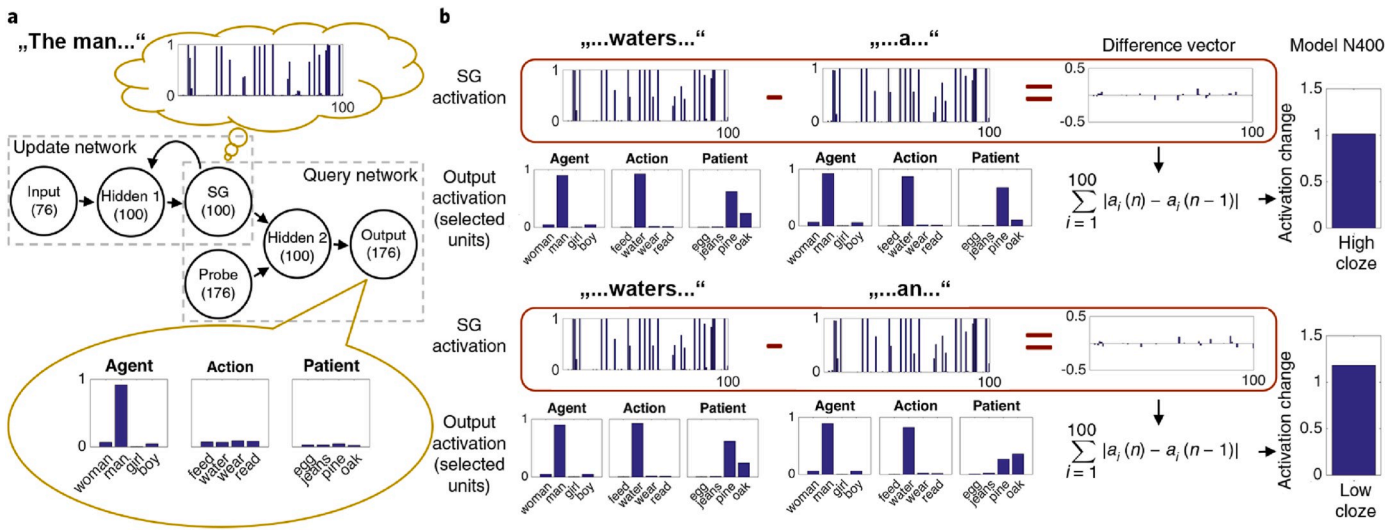

Fig. 1. The Sentence Gestalt (SG) model architecture, processing a sentence with a high or low cloze probability article, and the model's N400 correlate. The model (left) consists of an update network and a query network (highlighted by grey boxes). Ovals represent layers of units (numbers give the number of units in each layer). Arrows represent all-to-all modifiable connections; each unit applies a sigmoid transformation to its summed inputs, where each input is the product of the activation of the sending unit times the weight of that connection. In the update part of the model, each incoming word is processed through layer Hidden 1 where it combines with the previous SG activation to produce the updated SG activation (shown as a vector above the model), corresponding to the model's current probabilistic representation of the meaning of the sentence (i.e., the described event). During training, after each presented word, the model is probed and given feedback concerning all aspects of the described event (e.g. agent, "man", action, "water", patient, „pine", etc.) in the query network (please note that the model is trained on a synthetic environment where the man man waters pines more often than oaks). Here, the activation from the probe layer combines via layer Hidden 2 with the current SG pattern to produce output activations. Selected output units activated in response to the agent, action, and patient probes are shown; each query response includes a distinguishing feature (e.g. 'man', 'woman', as shown) as well as other features (e.g., 'person', 'adult', not shown) that capture semantic similarities among event participants). After presentation of "The man" (leftmost), the SG representation (thought bubble at top left) supports activation of the correct features when probed for the agent and estimates the probabilities of action and patient features. After the word „waters" (shown twice in the middle) the SG representation is updated and the model now activates the correct features given the agent and action probes, and estimates the probability of alternative possible patients, based on its experience. If the next word is „a" (top), which is compatible with the high cloze probability continuation „pine", the change in SG activation (summed magnitudes of changes in 'Difference vector') is smaller than if the next word is „an" (bottom), which is incompatible with „pine" and instead to be followed by the low cloze probability continuation „oak". The change, called Semantic Update (SU) is the proposed N400 correlate (right). It is larger for the article compatible with the less as compared to the more probable described event.

will not continue with the high probability continuation ('pine') but instead with the lower probability continuation ('oak'; see Figs. 1b and 3).

Adjustments for Simulation 2. For Simulation 2, adjectives were presented between the articles and the nouns. Specifically, two further input units were added, which represent two adjectives, each of which followed one of the indefinite articles; they can be thought of as 'old' and 'new' as in 'an old x'/'a new x' (please note that the units' labels do not influence model behavior, but instead just serve to help the reader to map the roles of the units to the roles of words in natural sentences). The adjectives were not associated with specific semantic features at the output layer; this implementation does not correspond to a theoretical claim, but rather serves the goal of keeping a clear focus on the main issues at stake here and avoiding to introduce additional variation. As both indefinite articles occurred equally frequently across the training environment, the corresponding adjectives occurred equally frequently as well. To impede associations between the articles and the subsequently presented objects, article cloze probability (.7 vs. 0.3) was counterbalanced across agents for each action, and the probability of the articles was independent of the probability of the objects. Both adjectives were compatible with both the high cloze and the low cloze probability objects, i.e. each object was preceded equally frequently by either adjective and the corresponding article. Thus, in this situation the articles' cue validity concerning noun meaning was zero.

Please note that these simulations aim to isolate and illustrate the influences of specific manipulations and specific aspects of the environment while minimizing changes to our previous synthetic training environment (Rabovsky et al., 2018), rather than capture the complexity and richness of natural language environments. While this controlled and synthetic approach to training has the advantage of being transparent concerning the factors influencing model behavior, it does not allow to simulate N400 amplitudes observed in response to specific experimental sentences used in empirical experiments (an approach afforded by large scale training; Frank et al., 2015). Future complementary work should scale up the model to large scale language environments to achieve a more direct link to neural data, which will however come at the cost of decreased transparency concerning the factors influencing model behavior.

2.3. Simulations

The goal of including two simulations is to demonstrate the influence of predictive relationships in the learning environment on the model's N400 correlate. Therefore, Simulations 1 and 2 differed in terms of the training environment (corresponding to the linguistic environment during human language development as well as later in life, as adaptation is assumed to occur continuously over the life span) as described above, but did not differ in terms of the simulation experiments themselves (corresponding to the situation in the empirical experiments with human participants). For each of the simulations, we presented an agent ('man'), followed by each of the ten specific actions (e.g., 'waters') and, for the high cloze condition the high cloze probability article (e.g., 'a' presented with a probability of .7 in this situation during training), and for the low cloze condition the low cloze probability article (e.g., 'an' presented with a probability of .3 in this situation during training),

before the presentation of the object (e.g., 'pine' or 'oak').[2] It is important to note that for models exposed to the training environment for the first simulation, the articles were predictive of the upcoming nouns, while this predictive relationship was removed in the training environment for the second simulation. For both simulations, there were 10 items per condition, and the model's N400 correlate was computed as the summed magnitude of the difference in SG layer activation between presentation of the action (word *n*-1) and the article (word *n*), as illustrated in Fig. 1b.

3. Results

The model's N400 correlate to the articles in Simulation 1, where the articles provide reliable cues to meaning, is displayed in Fig. 2 (Supplementary Figure S1 shows the results by item). We used linear mixed model analyses (LMMs) implemented in the packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) in R (www.r-project.org) to analyze influences of article cloze probability on the model's N400 correlate. We entered a fixed effect for the article cloze probability factor (sum coding: 1 /+1) and added crossed random effects for models and items, with uncorrelated random intercepts and random slopes (for the cloze probability factor) varying across models and varying across items. (Adding random effects correlations yielded invalid estimates of −1, indicating overfitting.)

As can be seen in Fig. 2, the model's N400 correlate showed reliable article-induced modulations in Simulation 1, with larger semantic update for low cloze probability articles as compared to high cloze probability articles ($b = 0.085$, $SE = 0.017$, $t_{(180)} = 4.998$, $p < 0.0001$). Please note that despite the deterministic relationship between articles and nouns, these effects are smaller (Cohen's d = 1.87) than the simulated N400 effects for noun cloze probability, which we report in Rabovsky et al. (2018; Cohen's d = 2.71). The effects on articles crucially depend on the predictive relationship between the articles and subsequent meaning, as they vanish in Simulation 2 when the cue validity of the articles is removed by the adjectives presented between articles and nouns during training ($b = 0.0009$, $SE = 0.013$, $t_{(180)} = 0.072$, $p = 0.943$).

Fig. 2. Displayed is the model's N400 correlate as a function of article cloze probability after training on an environment where articles provide valid cues to meaning, because the articles are always directly followed by the nouns (in analogy to e.g., 'a kite'/'an airplane'; Simulation 1). Blue dots represent results for 10 independent runs of the model (averaged across 10 items per condition). Red dots represent condition means, +/- standard error of the mean (SEM) is represented by red error bars (barely visible because they barely protrude the area of the dot). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

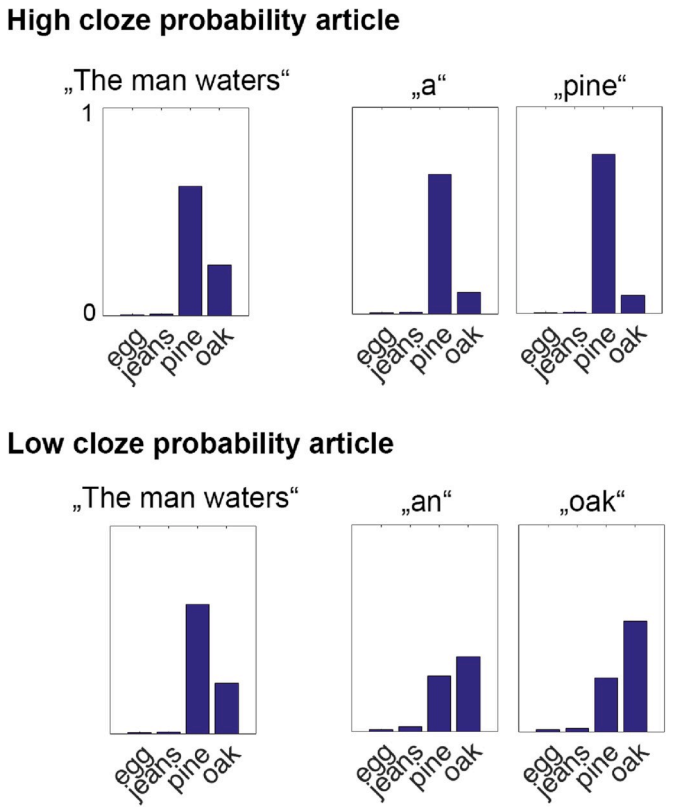

Fig. 3. Activation of selected output units when the model is probed for the filler of the patient role (i.e. "what is the patient/object in the event?") during sentences with high cloze (top) and low cloze (bottom) probability continuations in Simulation 1 (please note that the model is trained on a synthetic environment where the man man waters pines more often than oaks). After each presented word, the model gradually adjusts the represented probabilities of event features (for simplicity, only distinguishing features (names) are shown, but each concept is additionally represented by a number of other semantic features, e.g. 'can grow', 'food', etc.). The change in the represented probabilities of semantic features is larger for the low cloze probability article (bottom) that for the high cloze probability article (top), and from the perspective implemented in the model, this difference (measured at the hidden Sentence Gestalt layer, where the feature probabilities are implicitly represented) underlies article cloze probability effects on N400 amplitudes. Please also note that despite the deterministic relationship between articles and nouns in the training environment for this simulation, represented probabilities get further adjusted at the noun (see main text for discussion).

4. Discussion

The current study simulates article induced N400 effects (DeLong et al., 2005) as the change in a probabilistic representation of meaning, which corresponds to an implicit semantic prediction error, cued by encountering the articles (Figs. 2 and 3). In addition, the study explores the impact of the predictive relationship between articles and nouns (the articles' cue validity in the long term linguistic learning environment) on the model's N400 correlate. Results show that in the model, N400 effects on articles are reliable when the articles provide valid cues to meaning, but vanish when the articles' cue validity is low due to adjectives presented between articles and nouns during training. Thus, the

[2] As noted above, the model is trained on a synthetic environment where , pine' is a high probability continuation and ,oak' is a low probability continuation of the sentence beginning ,the man waters'.

model predicts that (given sufficient power) article induced N400 effects should be modulated by cross-linguistic differences in articles' cue validity.

4.1. Simulating article induced N400 effects without assuming word form prediction

The simulations provide a mechanistically explicit account of article induced N400 effects that does not depend on prediction at the level of phonological word forms, but rather treats N400 amplitudes as the change in the conditional probabilities of semantic features cued by encountering the articles (Fig. 3; Rabovsky et al., 2018; Rabovsky and McRae, 2014; Yan et al., 2017). This explanation differs from the most common interpretation of these effects (Altmann and Mirković, 2009; DeLong et al., 2005; Hagoort, 2017; Pickering and Garrod, 2013; but see Fleur et al., 2019). Importantly, the goal of the current study is not to suggest that word form prediction cannot account for the observed effects; instead, the goal is to offer an alternative explanation based on the change in a probabilistic representation of meaning, which seems interesting because the effects have been commonly taken as strong evidence for prediction at the level of word form. The perspective offered here suggests that if there is probabilistic prediction at the level of word forms, which may be expected if prediction is a fundamental aspect of brain function (Clark, 2013; Friston, 2005; McClelland, 1994; Rao and Ballard, 1999; Schultz et al., 1997), the consequences of probabilistic form prediction may be reflected in other (presumably earlier) ERP components (Gagl et al., 2020; see Nieuwland, 2019, for review).

Interestingly, as can be seen in Fig. 3, despite the deterministic relationship between articles and nouns in the training environment of Simulation 1, there was a further change in the probabilistic representation upon presentation of the nouns. This can be explained by the fact that the articles are also followed by other words during training and that the model also activates the semantic features of those other words to some degree upon presentation of the articles. This shows that the model does not perfectly track conditional probabilities based on the sentence context and that both more local (word) and more global (sentence) context contribute to its activations and probability estimates. This is interesting also because it suggests that even though ‚a' and ‚an' both mean ‚some one thing', they differ in terms of the associated distributional information in semantic space (see e.g. Mikolov et al., 2013, for large-scale naturalistic word embeddings), based on their respective contributions to the probabilistic estimation of meaning. In principle, this would mean that if some context would induce strong expectations for either ‚a' or ‚an', presenting the unexpected article might induce a larger shift in semantic activations even in the absence of specific contextual expectations concerning noun meanings. However, as indefinite articles will simultaneously strengthen or weaken probabilistic expectations of compatible or incompatible noun meanings, respectively, in any particular context (even if there is no specific high cloze noun), this seems difficult to test empirically.

4.2. The impact of cue validity

A factor suggested to contribute to the difficulty to reliably observe article induced N400 effects in empirical studies is that the predictive relationship between indefinite articles and subsequent nouns is weakened by the fact that indefinite articles are not necessarily directly followed by nouns in natural language (Nieuwland et al., 2018). In the current study the articles' cue validity is directly manipulated by including simulations with very high (perfect) versus very low (non-existent) cue validity, the assumption being that the cue validity in natural language presumably lies somewhere in between these extremes and can vary across languages (according to the Corpus of Contemporary American English and British National Corpus, in English indefinite articles are directly followed by nouns in about a third of the cases; based on Nieuwland et al., 2018), presumably contributing to the small size of the empirically observed effects (Nicenboim et al., 2020; Nieuwland et al., 2018).

Cue validity is manipulated here between training environments (and hence models), thus conceptualizing and implementing cue validity as a characteristic of a specific long term linguistic environment (i.e., a language) rather than a specific experimental condition (e.g., experimental blocks with high versus low cue validity, which would correspond to more short-term adjustments), or specific contexts (e.g., specific verbs after which cue validity of articles may be high or low). Thus, when translating the results of the current simulations into an empirical prediction, this prediction most naturally concerns differences between languages. For instance, if there is a language with phonological marking of the article where articles are always (or almost always) directly followed by nouns (e.g., because adjectives are presented after the nouns), article induced N400 effects should be stronger as compared to English. A slightly different but relevant case are gender marked articles (as in e.g., Spanish, German, Dutch), which have been shown to induce similar N400 effects, i.e., typically increased N400 amplitudes at articles incompatible with expected nouns (see Kochari and Flecken, 2019 for a recent review, Nicenboim et al., 2020 for a meta-analysis, and Fleur et al., 2019 for evidence consistent with the view that word form prediction may not be essential). In general, gender marked articles would be expected to provide stronger cues to meaning (i.e., with higher cue validity) as compared to phonologically marked articles, because the predictive relationship between gender marked articles and nouns does not depend on intervening adjectives.[3] Therefore, the model predicts that article induced N400 effects should be stronger for gender marked as compared to phonologically marked articles (see also Yan et al., 2017). However, please note that even with perfect cue validity the model predicts smaller effects for articles as compared to nouns (Cohen's $d = 2.71$ vs. 1.87), and additional semantic update at the occurrence of the noun (see Fig. 3). Please also note that in their meta-analysis using publicly available data, Nicenboim et al. (2020) did not find evidence either for or against an interaction between article cloze probability and the type of manipulation (gender vs. phonological marking) on N400 amplitudes. However, given that the main effect of article cloze probability was very small, the lack of evidence for a significant interaction does not provide conclusive evidence against such an interaction (Nicenboim & Vasishth, personal communication), and furthermore, focusing just on the publicly available datasets, the meta-analysis did not include any of the earlier Spanish studies employing gender marked articles, most of which observed reliable N400 effects (Wicha, Bates, Moreno and Kutas, 2003a; Wicha, Bates and Kutas, 2003b).

In this context, it is relevant to note that the model's predictions are rather qualitative: it predicts whether an effect (such as an article induced N400 effect) should be present in principle, but not whether the effect will be observed empirically in a specific study (which depends on additional factors such as EEG noise, which are not considered in the model). Similarly, the model predicts whether an effect should be larger or smaller than some related effect (e.g., article cloze probability effects smaller than noun cloze probability effects, and stronger article cloze probability effects when the articles' cue validity concerning upcoming noun meaning is high in a language), but it does not predict the absolute size of an effect and whether (or more specifically, based on which sample size) a difference in the effect sizes should be detected

[3] However, please note that in some languages (e.g., German and Dutch) there are a few additional complications that need to be taken into account. For instance, in German the feminine article is the same as the plural article (e.g., *die* Frau [the woman, feminine article] and *die* Männer [the men, plural article]), and the neutral article can be used to indicate miniaturisation of things that usually have feminine or masculine articles (e.g. *das* Tischlein [the small table, usually masculine article] and *das* Täschlein [the small bag, usually feminine article]), which can impact the need to update semantic predictions.

empirically. Based on the very small main effect of article cloze probability, an interaction with cue validity (e.g., in the sense of phonological versus gender marking) presumably requires a very large sample size to be substantiated empirically.

4.3. Probabilistic prediction of sentence meaning

It seems interesting to note that the same predictive comprehension system implemented in our model does produce article induced N400 effects for situations with high but not low cue validity (corresponding to Simulations 1 and 2), and that this difference does not speak to the predictive nature of the system, but simply reflects the statistical regularities in the environment. The lack of an effect of article cloze probability in Simulation 2 is not due to the fact that the model does not predict, but rather due to the fact that the predictions of the model are not differentially adjusted based on the two articles (i.e., if the model predicts upon encountering the word 'waters' that the man will likely water a pine and less likely an oak, this prediction will not change upon encountering the indefinite article). If the system probabilistically predicts sentence meaning, as implemented in our model, it does not necessarily matter whether other words (e.g., adjectives) are presented in between as long as they do not change conditional probabilities of aspects of meaning, because the goal is not to predict the next word, the goal is to predict all aspects of meaning of the described event.

Please note that pre-activation in the model is probabilistic and graded, in accordance with conditional probabilities in the environment. So, as apparent in Fig. 3, the model does not pre-activate the high probability object completely. Instead, it activates different objects according to their conditional probabilities. In the current version of the model trained on a simple synthetic environment (as described in Methods), this means that after being presented with 'The man waters … ' it represents a relatively high probability for a pine and a relatively low probability for an oak.[4] In a natural environment, where men can water many more different things (and boys can fly many different things including kites, airplanes, drones, toy helicopters, etc.), the semantic features of all these things would be pre-activated according to their conditional probabilities in the context. Presentation of the article is assumed to increase the represented probabilities of semantic features of all objects supported by the context, which are compatible with this article, and decrease the represented probabilities of semantic features of all objects incompatible with this article. This shift in the implicitly represented conditional probabilities of semantic features, which corresponds to an implicit semantic prediction error, is proposed to be reflected in N400 amplitudes.

Are there qualitatively different aspects of predicted sentence meaning reflected in different subcomponents of the N400 in the sense that qualitatively different neural generators underlie N400 amplitudes during different time windows? This issue was raised by a recent large-scale ERP study observing influences of different types of context (cloze probability and plausibility) in different time windows within the broad N400 segment (Nieuwland et al., 2019). From the perspective implemented in the model, this finding might be explained by assuming different aspects of the variance of the change in a probabilistic representation of meaning being captured by the different measured variables, i.e. cloze probability versus plausibility. The model assumes a single (but distributed) generator, mainly located in the temporal lobe (Lau et al., 2008). Specifically, in the model the N400 can be seen as corresponding to a change in the activation state of the semantic system. There seems to be a semantic hub in the anterior temporal lobe (Rogers and McClelland, 2008), but based on embodied accounts of meaning representations (e.g., Kiefer and Pulvermüller, 2012), activation changes in different parts of the brain might be assumed to contribute to the overall magnitude of activation change. Importantly, while the change in neural activation may happen across distributed brain areas (with a main hub in the temporal lobe), the model does not assume different functional subprocesses contributing to the N400.

[4] The minimum of cross-entropy error across the training environment is reached when the represented probabilities (i.e., the activations of the output units) correspond to the conditional probabilities in the environment (Rumelhart et al., 1995), i.e., in the model's synthetic environment this corresponds to .7 for the pine and .3 for the oak.

5. Conclusion

In conclusion, the current simulations provide a mechanistically explicit account of article induced N400 effects as reflecting the change in a probabilistic representation of meaning corresponding to an implicit semantic prediction error. Furthermore, the simulations predict that – with sufficient power - article induced N400 effects should be modulated by cross-linguistic differences in the predictive relationship between articles and nouns (i.e., the articles' cue validity), a factor potentially contributing to the small size of the effects observed in empirical studies. This account is in line with the view that the brain probabilistically predicts upcoming input based on the experience of statistical regularities in the environment, and that the prediction error or Bayesian surprise at the level of meaning is reflected in N400 amplitudes (Kuperberg and Jaeger, 2016; Rabovsky et al., 2018; Rabovsky and McRae, 2014).

Data and code availability

Data and code for this project are available on the Open Science Framework (OFS): https://osf.io/5tsjw/

Declaration of competing interest

None.

Acknowledgments

Funding was provided by an Emmy Noether grant from the German Research Foundation (grant RA 2715/2-1) to Milena Rabovsky. Many thanks to Jay McClelland for helpful discussion and to Marta Kutas, Mante Nieuwland, and Shaorong Yan for helpful and constructive comments on an earlier version of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107466.

References

Altmann, G.T.M., Mirković, J., 2009. Incrementality and prediction in human sentence processing. Cognit. Sci. 33 (4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Brouwer, H., Crocker, M.W., Venhuizen, N.J., Hoeks, J.C.J., 2017. A neurocomputational model of the N400 and the P600 in language processing. Cognit. Sci. 41, 1318–1352. https://doi.org/10.1111/cogs.12461.

Brown, C., Hagoort, P., 1993. The processing nature of the N400: evidence from masked priming. J. Cognit. Neurosci. 5 (1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34.

Cheyette, S.J., Plaut, D.C., 2017. Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. Cognition 162, 153–166. https://doi.org/10.1016/j.cognition.2016.10.016.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36 (3), 181–204. https://doi.org/10.1017/S0140525X12000477.

DeLong, K.A., Urbach, T.P., Kutas, M., 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nat. Neurosci. 8 (8), 1117–1121. https://doi.org/10.1038/nn1504.

DeLong, K.A., Urbach, T.P., Kutas, M., 2017. Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). Language,

Cognition and Neuroscience 32 (8), 966–973. https://doi.org/10.1080/23273798.2017.1279339.
Fitz, H., Chang, F., 2019. Language ERPs reflect learning through prediction error propagation. Cognit. Psychol. 111, 15–52. https://doi.org/10.1016/j.cogpsych.2019.03.002.
Fleur, D.S., Flecken, M., Rommers, J., Nieuwland, M.S., 2019. Definitely saw it coming? An ERP Study on the Role of Article Gender and Definiteness in Predictive Processing. BioRxiv https://doi.org/10.1101/563783 Retrieved from.
Frank, S.L., Galli, G., Vigliocco, G., 2015. The ERP response to the amount of information conveyed by words in sentences. Brain Lang. 140, 1–25.
Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 360 (1456), 815–836. https://doi.org/10.1098/rstb.2005.1622.
Gagl, B., Sassenhagen, J., Haan, S., Gregorova, K., Richlan, F., Fiebach, C.J., 2020. An orthographic prediction error as the basis for efficient visual word recognition. Neuroimage. https://doi.org/10.1016/j.neuroimage.2020.116727 (in press).
Hagoort, P., 2017. The core and beyond in the language-ready brain. Neurosci. Biobehav. Rev. 81, 194–204. https://doi.org/10.1016/j.neubiorev.2017.01.048.
Ito, A., Martin, A.E., Nieuwland, M.S., 2017a. How robust are prediction effects in languagecomprehension? Failure to replicate article-elicitedN400 effects. Lang. Cognit.Neurosci. 32 (8), 954–965. https://doi.org/10.1016/j.jml.2013.08.001.
Ito, A., Martin, A.E., Nieuwland, M.S., 2017b. Why the A/AN prediction effect may be hard to replicate: a rebuttal to Delong, Urbach, and Kutas (2017). Language, Cognition and Neuroscience 32 (8), 974–983. https://doi.org/10.1080/23273798.2017.1323112.
Itti, L., Baldi, P., 2006. Bayesian Surprise Attracts Human Attention, vols. 1–8. Nips. https://doi.org/10.1016/j.visres.2008.09.007.
Kiefer, M., Pulvermüller, F., 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex 48, 805–825.
Kochari, A.R., Flecken, M., 2019. Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch. Language, Cognition and Neuroscience 34 (2), 239–253. https://doi.org/10.1080/23273798.2018.1524500.
Kuperberg, G.R., Jaeger, T.F., 2016. What do we mean by prediction in language comprehension? *Language*. Cognit. Neurosci. 31 (1), 32–59. https://doi.org/10.1080/23273798.2015.1102299.
Kutas, M., Federmeier, K.D., 2000. Electrophysiology reveals semantic memory use in language comprehension. Trends in Cognitive Sciences. https://doi.org/10.1016/S1364-6613(00)01560-6.
Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62 (August), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.
Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 101–103.
Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Software 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.
Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics: (de) constructing the N400. Nat. Rev. Neurosci. 9 (12), 920–933. https://doi.org/10.1038/Nrn2532.
Laszlo, S., Armstrong, B.C., 2014. PSPs and ERPs: applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended Event-Related Potential reading data. Brain Lang. 132, 22–27. https://doi.org/10.1016/j.bandl.2014.03.002.
Laszlo, S., Plaut, D.C., 2012. A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. Brain Lang. 120, 271–281.
McClelland, J.L., 1994. The interaction of nature and nurture in development: a parallel distributed processing perspective. Int. Perspect. Psychol. Sci. ume 1 (Leading Themes).
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Neural Information Processing Systems 1–9. https://doi.org/10.18653/v1/d16-1146.
Nicenboim, B., Vasishth, S., Rösler, F., 2020. Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. Neuropsychologia.
Nieuwland, M.S., 2019. Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. Neurosci. Biobehav. Rev. 96 (May 2018), 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019.
Nieuwland, M.S., Barr, D.J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D.I., et al., 2019. Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. In: Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences. Advance online publication. https://doi.org/10.1098/rstb.2018.0522.
Nieuwland, M.S., Politze-Ahles, S., Heyselaar, E., Segaer, K., Bartolozzi, F., Kogan, V., et al., 2018. Large-scale replication study reveals a limit on probabilistic prediction in language. ELIFE 1–24, 5030732.
Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behav. Brain Sci. 36 (4), 329–347. https://doi.org/10.1017/s0140525x12001495.
Rabovsky, M., Hansen, S.S., McClelland, J.L., 2018. Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nat. Human Behav. 2, 693–705. https://doi.org/10.1038/s41562-018-0406-4.
Rabovsky, M., McClelland, J.L., 2020. Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 375, 20190313. https://doi.org/10.1098/rstb.2019.0313.
Rabovsky, M., McRae, K., 2014. Simulating the N400 ERP component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. Cognition 132 (1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010.
Rogers, T.T., McClelland, J.L., 2008. A précis of semantic cognition: a parallel distributed processing approach. Behav. Brain Sci. 31, 689–714.
Rao, R.P.N., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature 2 (1), 79–87.
Rumelhart, D.E., Durbin, E., Golden, R., Chauvin, Y., 1995. Backpropagation: the basic theory. In: Chauvin, Y., Rumelhart, D.E. (Eds.), Backpropagation: Theory, Architectures, and Applications. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pp. 1–34.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275 (5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
St John, M.F., McClelland, J.L., 1990. Learning and applying contextual constraints in sentence comprehension. Artif. Intell. 46 (1–2), 217–257. https://doi.org/10.1016/0004-3702(90)90008-N.
Wicha, N.Y.Y., Bates, E.A., Moreno, E.M., Kutas, M., 2003a. Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. Neurosci. Lett. 346 (3), 165–168. https://doi.org/10.1016/S0304-3940(03)00599-8.
Wicha, N.Y.Y., Moreno, E.M., Kutas, M., 2003b. Expecting gender: an event related brain potential study on the role of grammatical gender in comprehending a line drawing within a written sentence in Spanish. Cortex 39 (3), 483–508. https://doi.org/10.1016/S0010-9452(08)70260-0.
Yan, S., Kuperberg, G.R., Jaeger, T.F., 2017. Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). https://doi.org/10.1101/14375. BioRxiv. Retrieved from.