In: A. Goel, C. Seifert, & C. Freska (2019, eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Bayesian Pragmatics Provides the Best Quantitative Model of Context Effects on Word Meaning in EEG and Cloze Data

Markus Werning[1*§], Matthias Unterhuber[1*§], and Gregor Wiedemann[2]
[*]Corresponding authors: {matthias.unterhuber, markus.werning}@rub.de
[1] Department of Philosophy, Ruhr University Bochum
[2] Department of Informatics, University of Hamburg

Abstract

We contrast three views of how words contribute to a listener's understanding of a sentence and compare corresponding quantitative models of how the listener's probabilistic prediction on sentence completion is affected in online comprehension. The Semantic Similarity Model presupposes that the predictor of a word given a preceding discourse is their semantic similarity. The Relevance Model maintains that utterances are chosen to maximize relevance. The Bayesian Pragmatic Model assumes a relevance-guided modulation of a word's lexical meaning that can be regarded as a Bayesian update of statistical regularities stored in memory. In addition to a Cloze test, we perform an EEG study, recording the event-related potential on the predicted word and take the N400 component to be inversely correlated with the word's predictive probability. In a multiple regression analysis, we compare the three models with regard to Cloze values and N400 amplitudes. The Bayesian Pragmatic Model best explains the data.

Keywords: Bayesian Pragmatics, EEG, N400, Cloze Test, Semantic Similarity, Relevance, Generative Lexicon, Probabilistic Prediction, Online Comprehension, Modulation, Predictive Completion Task

Introduction

A preceding discourse can influence the way words contribute to a listener's understanding of a sentence (Recanati, 2012). This contextual influence also affects a listener's implicit probabilistic predictions on the completion of a discourse in the process of online comprehension. The listener's implicit task of predicting the next word w_{n+1} following a discourse that consists of the word sequence $w_1 \dots w_n$ can be described by a predictive probability of the following form:

(1) $P(w_{n+1}|w_1 \dots w_n)$.

In an EEG study, Nieuwland and Van Berkum (2006) showed that discourse contexts can interact with the lexical animacy feature of concrete nouns. A preceding context with a peanut being fictitiously described as dancing and singing can, e.g., invert comparative predictive probabilities such that the predicate *(was) in love* now has a higher probability for the listener than the otherwise more likely predicate *(was) salted*. The inversion of comparative predictive probabilities was revealed by a crossing-over of the N400 components between the two conditions measured on the critical predicates. The N400 component is defined as a negatively-going deflection of the event related potential over centro-parietal electrodes occurring around 400ms after stimulus onset (Kutas & Federmeier, 2011). As reviewed by Kuperberg and Jaeger (2016), the N400 component measured on a word is typically inversely correlated with its conditional probability given the preceding context.[1] There are two dominant interpretations of the underlying neuro-cognitive functions reflected in the N400 component: the semantic integration (e.g., Hagoort, Baggio, Willems, 2009) and the lexical retrieval view (e.g., Brouwer, Crocker, Venhuizen, & Hoeks, 2017). On these views, a negative increase of the N400 amplitude reflects higher processing demands associated with either (a) the integration of the target word's meaning into the compositional meaning of a sentence, or, respectively, (b) the retrieval of the target word's lexical meaning from memory.

Neural network models of the N400's underlying neural mechanisms have been proposed (e.g., Brouwer et al., 2017), some of which take the correlation of the N400 component with its predictive probability as a key explanandum (Rabovsky, Hansen, & McCelland, 2018; Fitz and Chang, to appear).

What has remained unclear in the peanut study is which factors of the context are responsible for the crossing over of the N400 components. The Semantic Similarity View maintains that the contextual influence is due to the degree of the semantic similarity between parts of the discourse context and the words in the target sentence (Otten & Van Berkum, 2008). The semantic similarity between two expressions can be determined by statistical regularities on co-occurrences in large corpora, as described in Distributional Semantics. In the above example the expression *in love*, e.g., has a greater semantic similarity to words in the context story (e.g., *dancing*, *singing*) than *salted*.[2]

The Relevance View, in contrast, holds that the crossing-over and the associated changes in the listeners' predictions

§ MW and MU contributed equally.

[1] A problem for this interpretation of the N400 are so-called semantic illusions, where zero-Cloze cases do not yield an increase in the N400 amplitude (cf. Kuperberg, 2007; Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer, Fitz, & Hoeks, 2012).

[2] Existing accounts in Distributional Semantics base semantic similarity values merely on co-occurrences of syntactically unstructured lexical primitives and have problems coping, especially, with certain logical contexts such as negation or negative quantifiers. It would thus not be surprising if semantic similarity values attained by existing Distributional Semantics accounts had problems predicting the N400 in those contexts (e.g., Urbach & Kutas, 2010; Nieuwland, 2016). However, ongoing research in Distributional Semantics attempts to also capture complex logical contexts.

	+ALex	−ALex
AStdCtx	Maria richtet ein Kuchenbuffet her, das ihre Freunde beeindrucken soll, und bereitet alles Notwendige dafür vor. *Maria prepares a cake buffet to impress her friends and makes ready everything necessary for it.* Sie ist schon dabei **Sahne** zu **schlagen**. *She is already about **cream** to **whip**.*	Maria richtet ein Kuchenbuffet her, das ihre Freunde beeindrucken soll, und bereitet alles Notwendige dafür vor. *Maria prepares a cake buffet to impress her friends and makes ready everything necessary for it.* Sie ist schon dabei **Sahne** zu **zeichnen**. *She is already about **cream** to **draw**.*
ANewCtx	Maria übt für ein Bild von einem Kuchenbuffet und benutzt ihr Notizbuch für ihre Vorstudie. *Maria practices for a picture of a cake buffet and uses her notebook for her preliminary study.* Sie ist schon dabei **Sahne** zu **schlagen**.] *She is already about **cream** to **whip**.*	Maria übt für ein Bild von einem Kuchenbuffet und benutzt ihr Notizbuch für ihre Vorstudie. *Maria practices for a picture of a cake buffet and uses her notebook for her preliminary study.* Sie ist schon dabei **Sahne** zu **zeichnen**. *She is already about **cream** to **draw**.*

Table 1. Example of stimuli with English translation (Experiments 1 & 2). In the 2 × 2 design, Agentive (+Alex) and Non-Agentive (– Alex) verbs are combined with a standard (AStdCtx) or a new context (ANewCtx). The word order of the target sentence in the English translation is adjusted to the German original.

are only due to relevance considerations, regarding the relation between the context and the target expression. In the process of comprehension, the listener may assume that the speaker has chosen a particular combination of words in the discourse to maximize relevance (Sperber & Wilson, 1996). In the above example, the preceding fictitious story appears to be more relevant if the noun *peanut* is interpreted as animate. Consequently, a completion with the predicate *in love* should be more probable than one with *salted*. The Relevance View might also be associated with the view that there are no identifiable word meanings in the mental lexicon, at all (Elman, 2004). Consequently, the notion of semantic similarities between lexical meanings would be void.

Both, the Semantic Similarity and the Relevance View contrast with the Bayesian Pragmatic View. The latter accounts for the rational cooperation between speaker and listener by Bayes's Theorem. The predictive probability is identified with the posterior probability of a word, which results from updating its prior probability with its likelihood. The prior is simply a function of the semantic similarity (reflecting overall statistical co-occurrences) between the target word and the words in the preceding context. By being able to update the prior, listeners can incorporate pragmatic considerations on speakers' intentions, such as the thrive for relevance, into to their interpretation and, consequently, adjust their predictions about speakers' continuation of a sentence. Bayesian pragmatics has been successfully used to explain results in behavioral experiments on simple referential games (Frank & Goodman, 2012), scalar implicatures (Degen, Tessler, & Goodman, 2015; Goodman & Stuhlmüller, 2013), gradable adjectives (Lassiter & Goodman, 2013; Qing & Franke, 2014), modal expressions (Lassiter & Goodman, 2015), and figurative meaning (Kao, Wu, Bergen, & Goodman, 2014). It has so far only been validated in EEG by Werning & Cosentino (2017).

In this paper, we test and compare the empirical adequacy of the three different views. For each of the views, a quantitative model of the listener's predictive probability of a word given a preceding discourse is developed. To feed the model with data, values for semantic similarity and relevance are attained. As a measure of semantic similarity, we use GloVe values (see below), a computer linguistic measure in the framework of Distributional Semantics. Values for relevance are collected via relevance ratings in an online questionnaire. To determine listeners' predictive probabilities, we perform two experiments with the same stimulus material: a forced-choice Cloze study employing an online questionnaire and an N400 study in EEG. In a multiple linear regression analysis, finally, the proportions of variance explained by each of the three models are compared with regard to both experiments.

Models

To design the experiment, we build on Pustejovsky's (1995) Generative Lexicon Theory, according to which the lexical entry of a concrete noun (e.g. *cake*) contains a "Qualia Structure", which, among others, specifies an Agentive component (e.g. *bake*). The Agentive component represents the typical way of bringing about the denoted object – it contrasts with the Telic component that relates to a typical purpose or function, e.g., *cake-eat* (Cosentino, Baggio, Kontinen, & Werning, 2017). Triggered by verbs like *begin* and *finish*, the Agentive component of a noun co-composes with the noun in sentence meaning composition (Werning, 2004, 2005). This explains why sentences such as (a) and (c) are typically understood as having the meaning of (b) and, respectively, (d):

(a) Granny finished the cake.
(b) Granny finished baking the cake.
(c) The artist began the statue.
(d) The artist began sculpturing the statue.

Since nouns often co-occur with verbs expressing their Agentive component – so-called Agentive verbs – the semantic similarity of a noun and the respective Agentive verb is usually high. For our stimuli, the high semantic similarity was explicitly confirmed by GloVe values (see below).

Each quadruple of our 2×2 experimental design $\{+\text{ALex}, -\text{ALex}\} \times \{\text{AStdCtx}, \text{ANewCtx}\}$ (see Table 1) was built around a fixed concrete noun n (*cream*). In condition +ALex as [opposed to −ALex] the critical word was chosen as a verb (*whip* [*draw*]) that expressed [did not express] the

Agentive component in the lexical entry of the preceding noun n and had a high [low] semantic similarity to it. In condition AStdCtx, a standard context preceded the target sentence, whereas in condition ANewCtx the preceding discourse sentence suggested a new way of bringing about the object. The semantic similarity between the verbs and each of the context sentences was held invariant over all four combinations of conditions.

The Semantic Similarity Model presupposes that the only predictor for the verb given its preceding discourse is the semantic similarity of the former to the latter (word frequency held constant). The predictive probability $P_n(v|c)$ of the verb v following the noun n given the preceding context sentence c is a monotonously increasing function f_n^+ of the semantic similarity $S(v,n)$ between the verb and the noun and the (however invariant) semantic similarity $S(v,c)$ between the verb and the context sentence c. Accordingly, the listener's predictive probability $P_n(v|c)$ hence comes to:

$$(2)\quad P_n(v|c) = f_n^+\,(S(v,n)).$$

The Relevance Model, in contrast, maintains that listeners assume that speakers aim at maximizing relevance by choosing their utterances. The sole predictor is the relevance of the situation expressed by the context sentence c for the action (expressed by the verb v) to be performed on the object (denoted by the noun n):

$$(3)\quad P_n(v|c) = g_n^+\,(Rel_n(c,v)).$$

The Bayesian Pragmatic Model allows listeners to update their priors regarding the verb following the noun with pragmatic considerations on speakers' intentions, thus, arriving at probabilistic predictions of the verb. The prior, i.e., $P_n(v)$, strictly increases with the semantic similarity between the verb v and the noun n. The update as described by the likelihood, i.e., $P_n(c|v)$, is modelled as the conditional probability of speakers' choice of a context c given their communicative intentions, namely, their intentions to attribute, to a protagonist, an action (denoted by the verb v) to be performed on a given object (denoted by the noun n). The speaker, in other words, has to choose a preceding context sentence (c) to let this action appear relevant such that the choice of c given v strictly increases with the relevance of c for v. This leads to the following identifications: $P(c|v) = g_n^+\,(Rel_n(c,v))$, $P_n(v) = f_n^+\,\big(S(v,n)\big)$, and by Bayes Theorem we get:

$$(4)\quad P_n(v|c) = K \cdot P_n(c|v)\,P_n(v) = K \cdot g_n^+\,(Rel_n(c,v)) \cdot f_n^+\,\big(S(v,n)\big).$$

The Bayesian update of the semantic similarity can be regarded as reflecting a relevance-guided modulation of the Agentive component in the lexical entry of the noun (Recanati, 2012).

Model Predictions. Both, Cloze values and the amplitude of the N400 component, can be assumed to correlate with the predictive probability $P_n(v|c)$. To model the two variables, we assume monotonous functional relations between $P_n(v|c)$ and the values of the Cloze test and, respectively, the N400 amplitude measured on v.

Account	Model
Bayesian Pragmatic View	$a\,Rel_n(c,v) + b\,S(v,n) + k$
Relevance View	$a\,Rel_n(c,v) + k$
Semantic Similarity View	$b\,S(v,n) + k$

Table 2. Linear parametric model predictions for Cloze values and the amplitude of the N400 component measured on the critical verb.

Logarithmization and subsequent linear approximation of the model predictions (2), (3) and (4) lead to the linear parametric model predictions described in Table 2. The negative logarithm of the predictive probability of a word has been interpreted as word surprisal and is directly correlated with the amplitude of the N400, measured on the word (Frank, Otten, Galli, & Vigliocco, 2015).

Experiment 1: Cloze Test

Method

Participants: Cloze Test. Forty German native speakers were recruited via Prolific. Three participants, who failed to have a high-school degree, and two further ones, who answered the test in below four minutes (estimated time: ten minutes), were excluded. Of the remaining thirty-five participants, 57.1% were male, 42.9% female. The average age was 30.20 years (SD=10.86).

Participants: Relevance Rating. Forty participants were recruited via Prolific. Of those, thirty-eight people were included which satisfied the requirement of being German native speakers and having attained a high school diploma.

General Material. For the forced-choice Cloze test and the relevance ratings, forty quadruples of the 2×2 design were generated in German (Table 1). Regarding the Cloze test, we used a forced-choice format to ensure that only Agentive verbs were available for comparison. The critical verbs did neither repeat, nor occur in context sentences.

$\langle +\text{ALex}, -\text{ALex}\rangle$-*verb pairs*. For each noun (e.g., *cream*), an +ALex verb (*whip*) and an −ALex verb (*draw*) were chosen so that the semantic similarity of the +ALex verb to the noun was .15 higher than that of the −ALex verb. $\langle +\text{ALex}, -\text{ALex}\rangle$-pairs did not differ significantly in frequency class (https://wortschatz.uni-leipzig.de/), nor in character length.

Context sentences. For each quadruple, a standard context was chosen (AStdCtx) that was highly relevant for the $\langle \text{n}, +\text{ALex}\rangle$ combination and less relevant for $\langle \text{n}, -\text{ALex}\rangle$. The new contexts (ANewCtx) were designed to reverse this order of relevance. The semantic similarity of $\langle \text{Context}, \text{Verb}\rangle$-pairs was not allowed to differ significantly over all four conditions. The character length of contexts AStdCtx and ANewCtx did not differ significantly either.

Fillers. Twenty filler discourses with congruent and twenty with incongruent $\langle \text{Noun}, \text{Verb}\rangle$-pairs were generated and had the same structure as the test material. The frequency

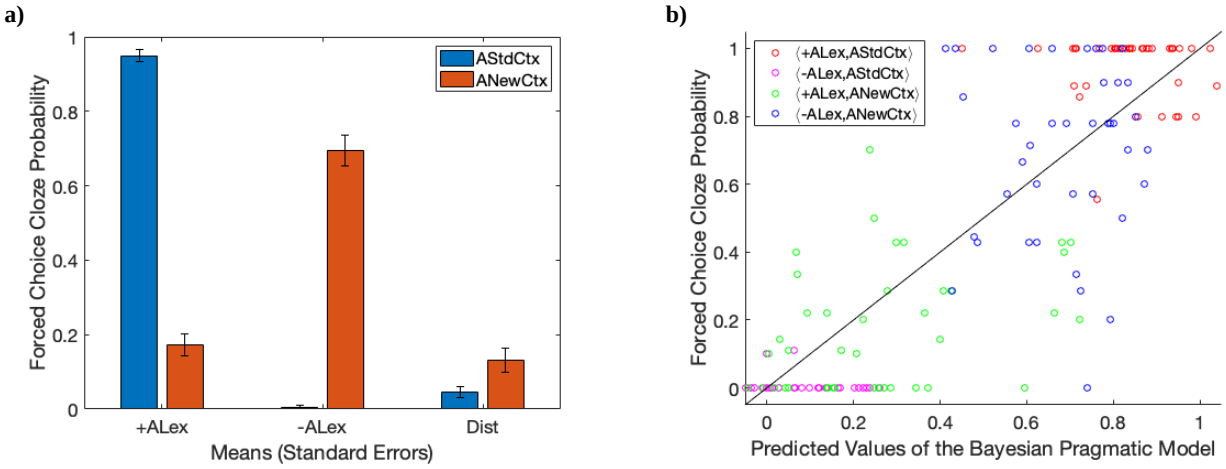

Figure 1. (a) Forced-choice Cloze probabilities. (b) Scatterplot for predictions of the Bayesian Pragmatic Model.

class and the length of the verbs did not differ significantly from the critical verbs of the test stimuli. Nor did the character length of the respective contexts differ significantly. *Semantic Similarity* was measured by an implementation of GloVe (Global Vectors, Pennington, Socher, & Manning, 2014) based on all articles from German Wikipedia and ca. three million news articles from Leipzig Corpora Collection (Goldhahn, Eckart, & Quasthoff, 2012).

Material and Design (Cloze Test). For the forced-choice Cloze test, the stimulus material was randomly distributed over four questionnaires in a counterbalanced way to avoid repetition. Context sentence and target sentence were presented as described in Table 1, with a blank instead of the critical verb. Subjects were instructed to choose the best fit among six alternative verbs. The alternatives always included the +ALex and −ALex verbs, as well as four alternative, agentive verbs drawn from the stimulus material not used in the questionnaire.

Material and Design (Relevance Ratings). The stimulus material including fillers was randomly distributed over four questionnaires in a counterbalanced way to avoid repetition. For each of the vignettes the context sentence surrounded by a box and marked as "A" was shown above the target sentence marked as "B", with the critical verb underlined. Participants were instructed to rate relevance on a 7-point Likert scale by answering the question: How plausible is the situation described in box A, given the action in sentence B.

Procedure. The Cloze test and relevance ratings were done using online questionnaires, via Qualtrics. The relevance ratings were preceded by two practice items.

Results and Discussion

Mean Cloze probabilities are summarized in Figure 1. The following results ensued: (1) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .0001$, $d = 9.01$), (2) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .0001$, $d = 2.65$), (3) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .0001$, $d = 3.88$), and (4) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .0001$, $d = 1.31$). To compare the Semantic Similarity, the Relevance and the Bayesian Pragmatic Models, we performed multiple regression analyses of Cloze data (see Table 3). The correlation between relevance ratings and corresponding similarity values was very small ($r = .10$).

The Bayesian Pragmatic Model was the clear winner according to BIC and AIC values, which take the unequal number of predictors into account. Within this model, the relevance values ($\beta = 1.01, p < .001$) and semantic similarity values ($\beta = .31, p < .01$) were significant predictors (for the scatterplot see Figure 1).

Experiment 2: EEG Study

Method

Participants. Twenty-eight participants were recruited. Two participants were excluded due to noisy EEG data and pain medication prior to the experiment. Of the resulting twenty-six participants, 30.8% were male, 69.2% female. The average age was 24.23 years (SD=3.40).

Design and Material. The stimulus material and fillers described in Experiment 1 were used. One of eight three-word phrases succeeded each target sentence (e.g., *und ist fröhlich* [and is happy]) to avoid the critical word being sentence final. Half of those appended phrases expressed positive and, respectively, negative emotional states, randomly chosen for each quadruple in a counterbalanced way.

The stimulus material was randomly split in two parts so that every context sentence and every critical verb appeared only once in each part. The two parts were administered to participants in two separate sessions, at least two weeks apart, to avoid carry-over and contrast effects. All participants had been presented with the entire stimulus material plus fillers. The order was counterbalanced. The complete set of fillers described in Experiment 1 was used in each session, resulting in eighty vignette (twenty per condition) and forty fillers (twenty congruent and twenty incongruent). The order was randomized.

Procedure. Each trial started with a fixation cross (1300ms) followed by the presentation of the context sentences. The context sentences were split up in roughly equal-sized chunks (character lengths: M = 17.91, SD =

Account	Model	N	df	RMSE	r	r_{adj}	BIC	ΔBIC	AIC	AICc
Predicted Variable: Cloze Probabilities (Experiment 1)										
Bayesian Pragmatic Model	Y~A+B+1	160	157	.22	.850*	.848*	-16.40	–	-25.64	-25.47
Relevance Model	Y~A+1	160	158	.23	.839*	.838*	-11.12	+5.30	-17.27	-17.19
Semantic Similarity Model	Y~B+1	160	158	.40	.223*	.209*	175.54	+191.93	169.38	169.46
Predicted Variable: EEG Amplitude (370–500ms, Experiment 2)										
Bayesian Pragmatic Model	Y~A+B+1	160	157	1.90	.489*	.479*	671.06	–	661.84	661.99
Relevance Model	Y~A+1	160	158	1.96	.426*	.419*	677.71	+6.64	671.56	671.63
Semantic Similarity Model	Y~B+1	160	158	2.08	.283*	.273*	696.26	+25.20	690.11	690.19

Table 3. Model comparisons for the regression analysis of Cloze probabilities (Experiment 1) and EEG amplitudes (370–500ms, Experiment 2). The EEG Amplitude was averaged across electrodes CP1, CPz, CP2, P1, Pz and P2. AICc = Aikaike Information Criterion corrected for sample size. *p<.001.

3.41). Each chunk was presented for 1300ms with a random interval of 200–400ms.

The target sentence, including the three-word phrase, was then presented word by word. In order to allow for a sufficiently long time interval for measuring ERP components, the target word was always presented for 450ms followed by an inter-stimulus interval of 450ms. All other words in the target sentence were presented for either 400 or 450ms, with a random inter-stimulus interval of 250–450ms. 700ms after the vignette's last word, participants had to judge, on a keypad, whether the complete vignette – consisting of context and target sentence (including the three-word phrase) – was plausible. The left-right orientation was randomized. The plausibility judgment should ensure that participants read the stimulus material carefully. Three examples with clearly congruent and incongruent ⟨Noun, Verb⟩-pairs served as practice material, prior to the testing phase.

Electroencephalogram recording and data processing. A BrainAmp Acticap system was used to record the electroencephalogram (EEG) from 66 active electrodes including four electro-oculogram electrodes for monitoring horizontal and vertical eye movements. The Brain Vision Analyzer 2.0 was employed to filter the data, correct for eye movements via independent component analysis (ICA). We used automatic artifact rejection to remove an episode per electrode if the change in currency exceeded 150μV in a 150ms interval. In case more than 10 electrodes were affected, or the trial had been interrupted before the critical word occurred, the whole episode was removed.

ERP data on the critical word (for each participant and trial) was exported to Matlab and individual trials were sorted and averaged across participants for each of the 160 vignettes, based on the Fieldtrip format for EEG data.

Results and Discussion

Mean amplitudes in the interval 370–500ms for posterior-central electrodes (CP1, CPz, CP2, P1, Pz, P2) across the four conditions are described in Figure 2. Bonferroni-corrected t-tests revealed significant differences between (1) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .01$, $d = 1.80$), (2) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .01$, $d = 1.13$), and (3) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .05$, $d = .61$). However, no significant difference was found for (4) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($d = .17$). For the respective ERP results as measured on the CPz see Figure 3.

To compare the Semantic Similarity, the Relevance and the Bayesian Pragmatic Models, we performed multiple regression analyses of the mean, baseline-corrected ERP in

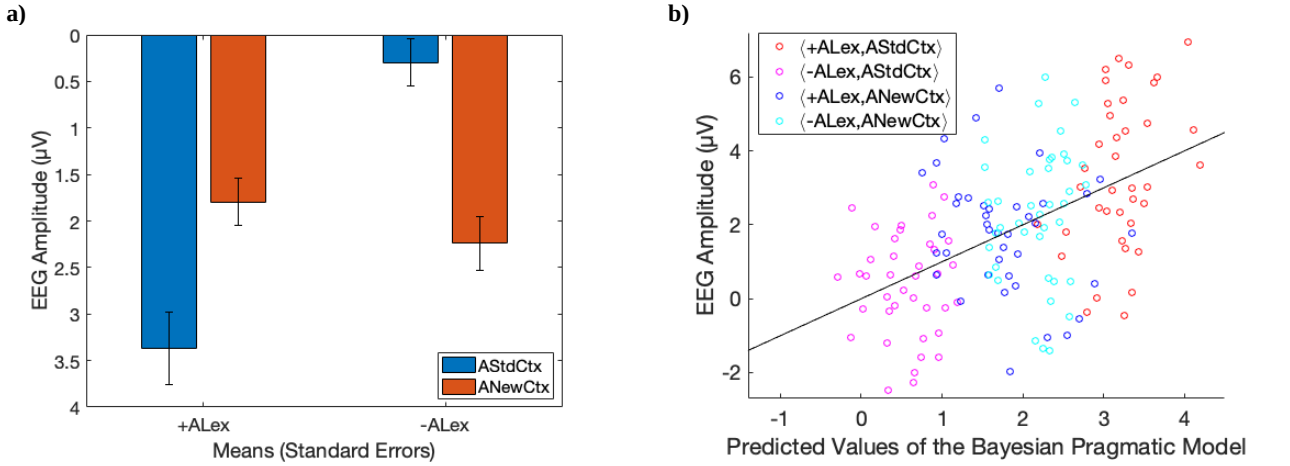

Figure 2. (a) EEG amplitudes (370–500ms, μV) for the pooled central-posterior electrodes (CP1, CPz, CP2, P1, Pz, and P2) after baseline correction. (b) Scatterplot for predictions of the Bayesian Pragmatic Model.

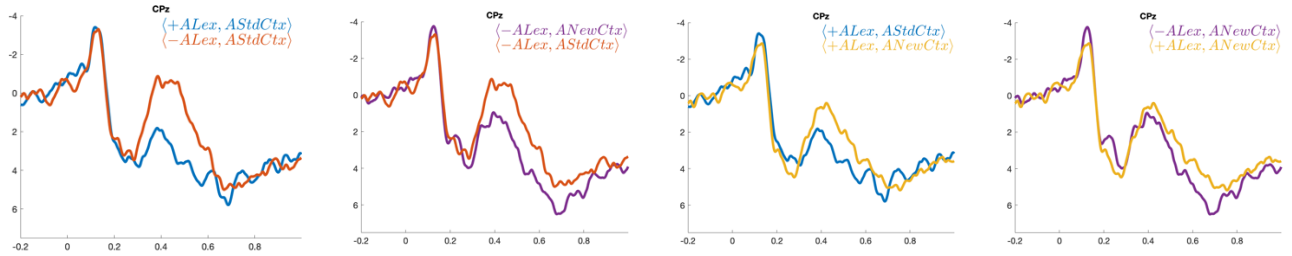

Figure 3. ERPs as measured on the CPz (µV).

the interval 370–500ms of the pooled posterior-central electrodes (see Table 3). Again, the Bayesian Pragmatic Model turned out as the clear winner according to BIC and AIC. Within this model, the relevance values ($\beta = 2.54, p < .0001$) as well as the semantic similarity values ($\beta = 2.81, p < .001$) were significant predictors. See Figure 2 for the scatterplot of the averaged N400 amplitudes and the values predicted by the Bayesian Pragmatic Model.

General Discussion

We contrasted three views of how words contribute to a listener's understanding of a sentence and compared three corresponding quantitative models of how the listener's implicit probabilistic predictions on the completion of a discourse is affected in online comprehension. The Semantic Similarity Model presupposes that the only predictor for a word given a preceding discourse is the semantic similarity between the two. The Relevance Model maintains that listeners assume that speakers aim at maximizing relevance by choosing their utterances. The Bayesian Pragmatic Model assumes a relevance-guided modulation of a word's lexical meaning that can be regarded as a Bayesian update of learnt statistical regularities stored in semantic memory.

To compare the explanatory power of the three models with regard to our Cloze and EEG data, we used the BIC and AIC criteria. Unlike r^2, BIC and AIC penalize for the unequal number of predictors. The clear winner in the comparisons, regarding both, the Cloze and EEG data, was the Bayesian Pragmatic Model. The Bayesian Pragmatic Model is not merely a combination of relevance and semantic similarity, but relates the two as the relevance-guided Bayesian update of a similarity-based prior probability. Interestingly, the two factors, relevance and similarity, did not contribute equally to the success of the Bayesian Pragmatic Model. Relevance outperformed semantic similarity, as indicated by the relative success of the Relevance over the Semantic Similarity Model. With regard to the EEG data, relevance explains 2.27 times as much variance as semantic similarity does. With regard to the Cloze data, however, this ratio is dramatically higher and equals 14.16. This pattern is also evidenced by comparing the EEG and the Cloze data with respect to the relative differences in BIC and AIC values of the three models. At first sight, this suggests that relevance is the dominant factor for the predictive probability of a word. A closer look reveals that semantic similarity still plays a larger role in truly incremental online comprehension, as observed in EEG, than in the Cloze test, which allows for backward-looking and untimed deliberation.

The lack of a significant difference of the ERPs in the time window of the N400 (370–500ms) for the comparison of the conditions ⟨+ALex, ANewCtx⟩ and ⟨−ALex, ANewCtx⟩ – with an effect size of only $d = .17$ – indicates that a greater semantic similarity value can compensate for a lower relevance value. This qualitatively illustrates the still prevalent importance of semantic similarity and thus the superiority of the Bayesian Pragmatic Model over the Relevance Model.

Our results also have clear implications for both the retrieval and the integration account of the N400. In light of our results, both approaches have to be modified to explicitly address the relevance-guided modulation of lexical meaning, described as updating by the Bayesian Pragmatic Model. Following the retrieval account, it will be the *modulated* lexical meaning of a preceding word that facilitates or impedes the retrieval of a subsequent word's meaning. Within the integration account, it is the *modulated* lexical meaning that determines the ease of integrating a subsequent word's meaning into the compositional meaning of the sentence.

Acknowledgments

We are grateful to the German Research Foundation for financial support through the grant WE 4984/4-1 as part of the Priority Program XPrag.de (SPP1727).

References

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews, 59*, 55–73.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research 1446*, 127–143.

Brouwer, H., Crocker, M. W, Venhuizen, N. J., Hoeks, J. (2017). A Neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41*(Suppl. 6), 1318–1352.

Cosentino, E., Baggio, G., Kontinen, J., & Werning, M. (2017). The time-course of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in Psychology, 8*, 813.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky

worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 548–553.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306.

Fitz, H. & Chang, F. (to appear). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 759–765.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Hagoort, P., Baggio, G., and Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (ed.), *The Cognitive Neurosciences* (pp. 819–836). Boston, MA: MIT Press.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, 12002–12007.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positiveform adjectives. *Proceedings of SALT*, *23*, 587–610.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, *136*, 123–134.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potential. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

Otten, M., & Van Berkum, J. J. (2008). Discourse-based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, *45*(6), 464–496.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use. *Proceedings of SALT*, *24*, 23–41.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in probabilistic representation of meaning. *Nature Human Behavior*, *2*, 693–705.

Recanati, F. (2012). Compositionality, Semantic Flexibility, and Context-Dependence. In M. Werning, W. Hinzen, & E. Machery (eds.), *Oxford Handbook of Compositionality* (pp. 175–191). Oxford: Oxford University Press.

Sperber, D., & Wilson, D. (1996). Precis of "relevance: communication and cognition." In H. Geirsson & M. Losonsky (eds.), *Readings in Language and Mind* (pp. 460–486). Oxford: Blackwell.

Urbach, T. P., & Kutas (2010). Quantifiers more or less quantify online: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179.

Werning, M. (2004). Compositionaltity, context, categories and the indeterminacy of translation. *Erkenntnis*, *60*(2), 145–178.

Werning, M. (2005). Right and wrong reasons for compositionality. In M. Werning, E. Machery, & G. Schurz (eds.), *The Compositionality of Meaning and Content* (Vol. I, pp. 285–309). Frankfurt: Ontos Verlag.

Werning, M., & Cosentino, E. (2017). The Interaction of Bayesian Pragmatics and Lexical Semantics in Linguistic Interpretation: Using Event-related Potentials to Investigate Hearers' Probabilistic Predictions. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 3504–3509.