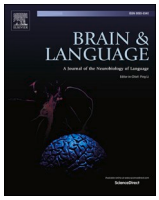
Brain & Language 235 (2022) 105197

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Speech register influences listeners' word expectations

M. Bentum [a,*], L. ten Bosch [a,b], A van den Bosch [c], M. Ernestus [a,b]

[a] Center for Language Studies, Radboud University, Nijmegen, The Netherlands
[b] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[c] Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Predictive language processing
Speech register
N400
Statistical language models
Electroencephalography

ABSTRACT

We utilized the N400 effect to investigate the influence of speech register on predictive language processing. Participants listened to long stretches (4 – 15 min) of naturalistic speech from different registers (dialogues, news broadcasts, and read-aloud books), totalling approximately 50,000 words, while the EEG signal was recorded. We estimated the surprisal of words in the speech materials with the aid of a statistical language model in such a manner that it reflected different predictive processing strategies; generic, register-specific, or recency-based. The N400 amplitude was best predicted with register-specific word surprisal, indicating that the statistics of the wider context (i.e., register) influences predictive language processing. Furthermore, adaptation to speech register cannot merely be explained by recency effects; instead, listeners adapt their word anticipations to the presented speech register.

1. Introduction

Human perception of sensory input involves more than passive registration. A rich body of research (e.g., Bar, 2007; Friston, 2005, 2012, 2018; De Lange, 2018) shows that prediction is a core aspect of perception. Similarly, humans engaged in reading or listening show sensitivity to the statistical structure of the language input (e.g., Ellis, 2002). Importantly, as studies investigating register variation show (e.g., Staples et al., 2015), patterns of language use differ extensively between registers, influencing the statistical distributions of words of the different varieties (e.g., Bentum et al., 2019). Consequently, expectations on the occurrences of words that are valid for one register might be invalid for a different register. In the current study, we utilize the N400 effect to investigate whether listeners adapt their word expectations as a function of the speech register they are listening to.

1.1. Register variation

The three examples below illustrate a range of registers: chatting with friends (1), coverage from a news reporter (2) and a novelist telling a story (3).

(1) It just irritated me and then Joanne, Joanne's like "did you hear someone page Dan's brother-in-law?" I said "he wouldn't give his name." And she just started laughing. (Barbieri, 2005).
(2) The leader's gunshot wounds are taking their toll, complicating efforts to persuade him to surrender. (Biber, 1999).
(3) Last summer, a short time before my son was due to leave home for college, my wife woke me in the middle of the night. (Nicholls, 2014).

The examples illustrate that language use varies in relation to the communicative context (Borrillo, 2000) and purpose (Biber & Conrad, 2001): Conversational speech is produced in real time, without much time to prepare disfluencies are prevalent, and it is characterized by a lower type-token ratio and a frequent use of pronouns. News reportage is typically prepared and intended to convey information about a certain event, which results in frequent use of time and place adverbials as well as proper nouns. A novel is typically written with ample time to revise and refine the language use, affording a rich vocabulary and complex sentence structure.

Differences in language use between registers are well documented (see Biber & Conrad, 2009, for an overview). For example, word choice differences (Biber, 1999), such as the use of *like* in (1) is typical of informal conversation (Barbieri, 2005). Variation in grammatical

* Corresponding author.
E-mail addresses: martijn.bentum@ru.nl (M. Bentum), louis.tenbosch@ru.nl (L. ten Bosch), a.p.j.vandenbosch@uu.nl (A. van den Bosch), mirjam.ernestus@ru.nl (M. Ernestus).

https://doi.org/10.1016/j.bandl.2022.105197
Received 25 April 2022; Received in revised form 5 October 2022; Accepted 19 October 2022
Available online 4 November 2022
0093-934X/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

constructions (Staples et al., 2015), such as the retention of the complementizer in *that*-clauses, as in (4), which differs between conversational speech and academic prose. In conversation *that*-omission is typical, while academic prose typically retains it (Biber, 1999).

(4) I hope [that] Paul tells him off. (Staples et al., 2015).

The lexical and grammatical variation between registers gives rise to register-specific word co-occurrence statistics. Bentum et al. (2019a) indeed found that word predictability differs between speech registers. The probability of a word thus not only depends on the directly preceding words but also on the wider context of register. This raises the research question of the current paper, namely, whether listeners adapt their word expectations based on the register of the speech input.

1.2. Predictive language processing and the N400 effect

Evidence for predictive language processing is well established in the literature (see e.g. Elman, 2009; Huettig, 2015; Kuperberg & Jaeger, 2016; Pickering & Gambi, 2018 for overviews). Importantly, there is converging evidence from many different experimental paradigms. For example, self-paced reading studies show that unlikely words are read more slowly compared to more likely words (Rayner, 1998; Kliegl et al., 2006). The visual word paradigm used in eye-tracking studies shows that listeners more often gaze in anticipation to a picture that matches the verb of the sentence, among multiple objects (e.g., they more often look at a picture of a cake when they hear *The boy eats* compared to *The boy moves*; Altmann & Kamide, 1999).

The so-called N400 effect is also associated with anticipatory language processing (see Kutas & Federmeier, 2011 for an overview). The N400 is a negative deflection of the event related potential (ERP), which peaks about 400 ms after word onset at central posterior electrode sites. When participants read short sentences, such as (5) and (6), with occasionally an anomalous final word, as in (6), the semantically incongruous word (here *socks*) results in a more negative deflection of the ERP compared to the congruent word (here *work*; Kutas & Hillyard, 1980).

(5) It was his first day at *work*.
(6) He smeared the warm bread with *socks*.

Later experiments revealed that semantic incongruency is not required for an N400 effect (e.g. Hagoort & Brown, 1994). For example, constraining sentence pairs such as (7), which raise a strong expectation for a specific word (i.e., *palms*), elicit a graded N400 effect, with the unexpected but semantically related word *pines* resulting in an attenuated N400 amplitude compared to the unexpected and unrelated *tulips* (Federmeier & Kutas, 1999). Importantly, the different sentence final words (i.e., *palms*, *pines* and *tulips*) are all possible non-anomalous endings, indicating the N400 effect is not dependent on semantic anomaly.

(7) They wanted to make the hotel look more like a tropical resort. So, along the driveway, they planted rows of [palms / pines / tulips].

Several experiments document that the N400 also provides evidence for anticipatory activation of words (e.g., Wicha et al., 2004; Van Berkum et al., 2005). These experiments use a paradigm where the anticipatory effects are measured before the expected word is presented. For example, when participants read sentence (8), the determiner *an* resulted in a more negative deflection of the N400 waveform compared to *a*, indicating that readers expected the following word to start with a consonant (DeLong et al., 2005; see also Yan et al., 2017, and Nieuwland et al., 2018). Furthermore, they found that word predictability (as estimated with a cloze test) correlates with the N400 amplitude, indicating that people generate probabilistic expectations of upcoming language input.

(8) The day was breezy so the boy went outside to fly [a kite / an airplane] ….

The findings for the determiner reported in DeLong et al. (2005) failed to be replicated in a large-scale study conducted by nine different labs, reported in Nieuwland et al. (2018). Nieuwland et al. (2018) argue that this failure to replicate challenges an empirical cornerstone of the 'strong prediction' view (i.e. people pre-activate words at all levels of representation in a routine and implicit fashion, and pre-activation is thus not limited to a word's meaning but includes its grammatical features and orthographic or phonological form). Yan et al. (2017) discuss both Delong et al. (2005) and a preprint version of Nieuwland's study and argue that the findings from this replication study can also be interpreted as in line with a prediction account; for example, the correlation between cloze values and the N400 was replicated. We interpret the combined literature (see also Wicha et al., 2004; Van Berkum et al., 2005) as supportive of anticipatory processing, especially because new evidence supporting phonological pre-activation can be found in Bentum et al. (2019b); see also Poulton & Nieuwland (2022) for a critical view).

Despite the wealth of evidence, predictive language processing remains (to some extent) controversial. For example, Huettig (2015) notes that much evidence for prediction is based on studies that only use the extremes of predictability and questions whether prediction plays an important role during natural language perception across the entire range of word probabilities. For example, the N400 effect is typically elicited by comparing highly likely versus highly unlikely words (e.g., Hagoort & Brown, 2000), which does not reflect normal language use.

We follow Kuperberg & Jaeger (2016) and use *prediction* to mean *graded probabilistic prediction*, whereby multiple candidates (e.g., words) have probabilities assigned based on the preceding context. In our interpretation, predictive language processing at the lexical level can be conceived of as generating a probability distribution over all words in the mental lexicon. Consequently, there will (almost) always be prediction error since not all probability is assigned to a single word. This prediction error is indexed by the N400-effect. For example, it is well attested that even with highly constraining sentences (e.g., Federmeier & Kutas, 1999), words other than the expected word show an attenuated N400 compared to unrelated words.

1.3. Word predictability estimation, cloze tests, and statistical language modeling

Word predictability is typically established with cloze tests, whereby participants fill in blanks in sentences, such as *So, along the driveway, they planted rows of …* The percentage of participants that fill in a specific word, such as *palms*, is referred to as the word's cloze probability. This percentage measures the 'expectedness' of a word given the context. This approach has two drawbacks: It is labor intensive to gather cloze probabilities and the method cannot distinguish among the predictability of low cloze probability words (Yan et al., 2017).

A different approach to estimate word predictability is the use of statistical language modelling. Work on statistical language modelling shows that, given a set of n preceding words, it is possible to assign a probability to each word to be the next word (e.g., Chen & Goodman, 1999; Och & Ney, 2003; Kilgarriff, 2001). In their most basic implementation, a statistical language model (SLM) is based on counting 'word n-grams' (henceforth n-gram) in corpora. An n-gram is a sequence of n consecutive units. For example, *the fast horse* is a word trigram with the bigrams *the fast* and *fast horse*; and the unigrams *the, fast* and *horse*. Based on counts of these n-grams in a large body of text, context-dependent word probabilities can be estimated:

$$P(W_i) = P(W_i|W_{i-n}, \cdots, W_{i-1})$$

P denotes the conditional probability of word W_i given a sequence of n preceding words. The automation of word predictability estimation

allows for the investigation of predictability effects for many words across the whole predictability spectrum. For example, Smith & Levy (2013) used this approach to determine that reading time is log-linearly related to the probability of a word on the basis of a dataset of approximately 50,000 words.

The log-linear relation between word probability and reading time fits well with Surprisal Theory of language processing (Hale, 2001; Levy, 2008), according to which language processing costs relate to surprisal. Surprisal is an information theoretic measure that captures the amount of Shannon information an item (i.e., word) in a message conveys. It is defined as the negative logarithm of the probability of a word given its pre-context and can informally be thought of as the 'unexpectedness' of a word given its pre-context. Frank et al. (2015) used statistical language modelling to estimate word surprisal for all content words in sentences from several novels. In this manner they could analyze a large set of approximately 30,000 word tokens. They used these sentences in an EEG experiment. Participants read sentences word-by-word while their EEG was recorded. Less expected words (i.e., words yielding high surprisal) elicited a larger negative amplitude in the N400 time window compared to more expected words.

Pickering & Gambi (2018) argue that surprisal effects (e.g. Smith & Levy, 2013; Frank et al., 2015) do not constitute evidence of predictive language processing, since surprisal and experimentally correlated effects are found on the perceived word. We disagree with this interpretation. Word surprisal is based on the preceding context (i.e. words), by computing the probability for each word in a lexicon to follow that context. The surprisal value is therefore not a static attribute of a specific word but a value derived from preceding context with respect to all words in the lexicon. The prediction consists of distributing probability over all words in the lexicon. In the 'activation' vernacular, each word in the mental lexicon is 'pre-activated' to the extent the context makes the word a probable continuation.

1.4. Discourse based ERP research

Most ERP studies investigating language processing use sentences presented in isolation. However, there have been discourse-level studies, whereby discourse is typically interpreted as anything more than one sentence, for example, short narratives such as (9 & 10).

(9) The brave knight saw that the dragon threatened the benevolent sorcerer.
He quickly reached for a [sword / lance] ….

(10) The benevolent sorcerer saw that the dragon threatened the brave knight.
He quickly reached for a [sword / lance] … (Van Berkum, 2012).

The short narratives were carefully matched on prime words. In the examples (9) and (10), only *brave knight* and *benevolent sorcerer* switched position. The sentence *He quickly reached for a …* by itself does not constrain in favor of either *sword* or *lance*. The preceding sentence in (9) favors *sword* while in (10) it does not. The unexpected words (e.g., *lance* in (9)) resulted in a larger N400 effect than the expected words (e.g., *sword* in (9)), whereas the words elicited similar N400s in less constraining sentences (e.g., 10) (Otten & Van Berkum, 2007). This and other results (see Van Berkum, 2012 for an overview) show that readers and listeners use the wider context of discourse to build up predictions of upcoming input.

One understudied aspect of predictive language is the effect of discourse beyond *multi*-sentence short narratives. In more natural communication situations, readers or listeners are engaged with reading or listening within a much wider context, which is itself modulated by the register. In the following section, we explain how we studied the influence of register variation on listeners' language processing.

1.5. Current study

In the current study we investigate whether listeners' word anticipations are influenced by speech register information. We test long stretches (4 – 15 min) of natural speech from different registers. Following Frank et al. (2015), we use statistical language modeling to estimate the word surprisal of all content words in our language materials and use word surprisal to predict the N400 amplitude for the content words. We estimate and compare four different 'types' of word surprisal: *register-specific, register-mismatch, generic,* and *recency-based* word surprisal.

The different 'types' of word surprisal reflect different processing strategies, which we compare to investigate the role of register in predictive language processing. *Register-specific* surprisal reflects the word predictability in a specific register. We hypothesize that if listeners adapt their word expectations based on register information, this register-specific surprisal will best predict the N400 amplitude. *Register-mismatch* word surprisal is used as a sanity check and reflects the word predictability based on an incorrect (mismatching) register. It should therefore predict the N400 amplitude less accurately than a register-specific model if listeners adapt their predictions to the register at hand. *Generic* word surprisal reflects the word predictability of register-unspecific, average language use. If listeners do not adapt to a register, this word surprisal should perform at least on par with register-specific word surprisal. Finally, *recency-based* word surprisal reflects generic word surprisal updated with information on recent words, of which the likelihood of recurring is temporarily boosted. If listeners do not use register characteristics, but instead recent language input, recency-based word surprisal may better predict the N400 amplitude.

The different word surprisal 'types' can be estimated by training SLMs on a specific set of language materials, as the estimated word surprisal depends crucially on the selected language materials the SLM is trained on. For example, an SLM trained on a book corpus will perform worse when tested on news materials as compared to when tested on an unseen book corpus. We therefore train SLMs on register-specific language materials to estimate *register-specific* word surprisal. *Register-mismatch* word surprisal is estimated by using an SLM trained on language materials from a mismatching register (see Section 2.2).

Generic word surprisal is more difficult to operationalize, because sampling language materials always introduces bias in some manner (Kilgarriff, 2007; Biber & Conrad, 2001); i.e., there is no 'general' corpus to train a bias-free SLM. To address this issue, we train an SLM on a large corpus (see Section 2.1.1) that does not overlap with the register-specific language materials. The resulting SLM can be considered *generic* (register-unspecific) to the extent that the register-specific SLMs are expected to show improved performance on the register-specific materials, i.e., the register-specific SLMs can be expected to assign overall higher probabilities to the next words in register-specific texts as compared to the generic SLM. Lastly, we estimate recency-based word surprisal with a cached SLM, a standard extension of the generic SLM, whereby the SLM is updated with the most recent n unigrams (i.e., words).

The current study also tests whether the effect that word surprisal predicts the N400 amplitude (for reading, Frank et al., 2015) generalizes to a *listening* study. There are two methodological reasons why this effect may be difficult to detect in a listening study. First, word onsets are harder to accurately determine in connected speech compared to the onsets of visually presented words. The uncertainty in word onset determination could potentially lead to temporal 'smearing' of the ERP (Van Berkum et al., 2005) and thereby to less clear temporal patterning of ERP components. Second, while it is possible to use fixed-paced presentation for a reading experiment (with a predetermined pause between words), this is neither feasible nor desirable with auditory presentation of natural speech. For example, due to co-articulation in speech, it would sound wholly unnatural to insert pauses between the words of a recorded sentence. The continuous nature of speech therefore likely results in overlapping, temporally smeared word effects in the

Table 1
Overview of the materials per speech style. The table shows the number of word tokens and types per register (word type is defined as the orthographic surface form), the average word duration in milliseconds, the number of speakers and the speakers' age range.

speech register	word tokens (word types)	average word duration	speakers (male)	Speaker age range
Dialogues	21,718 (2,435)	206 ms	11 (2)	20 – 62 years
news	15,350 (3,526)	289 ms	8 (7)	23 – 46 years
Books	13,209 (2,349)	256 ms	7 (3)	38 – 75 years
Total	50,277 (5,866)	245 ms	26 (13)	20 – 75 years

Table 2
Overview of the text materials for SLMs training. Word count lists the number of words used for training and testing and Source describes the source of the text materials.

Copus	Words tokens	Source
COW	1 billion	web crawled text
SoNaR books	17 million	excerpts from books
SoNaR news	2.5 million	news broadcast transcripts
CGN dialogues	1.7 million	transcribed dialogues
CGN books	500 thousand	transcribed read aloud books
CGN news	200 thousand	transcribed radio and television news broadcasts
IFADV	50 thousand	transcribed dialogues

EEG signal. As a result, the N400 could be attenuated when this ERP is elicited with all content words in long stretches of natural connected speech.

To counterbalance the issue of smaller expected effect sizes, we collected a large amount of data. We used audio recordings of speech from three different speech registers: dialogues, (read-aloud) books, and (broadcast) news. The registers were selected to be distinct in word predictability, based on the findings by Bentum et al. (2019a), and were assigned to three separate experiments. The reasons for conducting separate experiments are twofold. First, an experiment dedicated to one register allows the participant to adapt their anticipations to that speech register. Second, it is possible to present more materials of each register by spreading them over three experiments, fulfilling our requirement of a large dataset.[1]

In summary, in this study we test whether listeners anticipate words in long stretches (4 – 15 min) of natural speech, sampled from three speech registers. We estimate word surprisal and test whether this predicts the N400 amplitude and compare how well register-specific, register-mismatched, generic, and recency-based word surprisal estimates predict the N400 amplitude. With this comparison, we test whether listeners adapt their anticipations of upcoming words based on speech register; i.e., whether register-specific word surprisal is a better predictor of the N400 amplitude compared to the other word surprisal estimates.

2. Methods

2.1. Participants

Forty-eight neurologically unimpaired right-handed native speakers of Dutch (18–29 years, mean age = 21.7 years), 14 men and 34 women, participated in the three EEG experiments of the study. All participants gave informed consent for the experiments and the subsequent publication of the EEG recordings. They were paid 80 Euros for their participation.

[1] This dataset will be made freely available as the EEG Speech Register Corpus, ESRC for short.

2.1.1. Materials

The stimuli for the EEG experiments consisted of audio recordings of Dutch speech from different registers, with approximately 90 min of speech materials for each register. The recordings were extracted from two corpora: the *Spoken Dutch Corpus* (Oostdijk, 2001) and the *Institute of Phonetic Sciences Amsterdam Dialogue Video Corpus*, henceforth IFADV (Van Son et al., 2008), see also Section 2.2.1). The books and the news speech materials were extracted from the Spoken Dutch Corpus, the dialogues were extracted from IFADV.

Six distinct dialogues of approximately 15 min each were included for the dialogues experiment. Each dialogue was between two well-acquainted interlocutors (e.g., friends, colleagues), who freely talked about any topic that came to mind (see Van Son et al., 2008, for details). Seven 12-minute excerpts from read-aloud Dutch books were included in the books experiment. Finally, the news experiment consisted of 21 sections of approximately-four minutes long. Each section contains multiple news items presented by the same broadcaster. We inserted 0.9 s of silence between news items and combined the four-minute sections into seven 12-minute blocks.

All recordings used in the experiments were orthographically and phonemically annotated, which allowed for the time-locking of each individual word to the EEG-recording. All recordings were equalized at 60 dB with Praat (Boersma & Weenink, 2018). See Table 1 for an overview of the speech materials presented in the EEG experiments.

2.2. Estimating generic, register-mismatch, recency and register-specific word surprisal

2.2.1. Training and test materials

To estimate word surprisal of each content word in the experimental materials we trained and applied multiple statistical language models (SLMs). To train these SLMs, we used language materials from four corpora, NLCOW14, SoNaR, the Spoken Dutch Corpus, and IFADV. The NLCOW14 corpus, henceforth COW (Schäfer, 2015; Schäfer & Bildhauer, 2012), is a collection of web-crawled Dutch texts consisting of approximately 4,7 billion words. The SoNaR corpus (Oostdijk et al., 2013) is a collection of written Dutch texts of approximately 500 million words. From this corpus, we used a subset of the Dutch teleprompt texts (SoNaR news) and Dutch books (SoNar books). The Spoken Dutch Corpus (Oostdijk, 2001) is a corpus of recorded and transcribed Dutch speech from different registers containing approximately 10 million word tokens. We used three components from the Spoken Dutch Corpus: the spontaneous dialogue component (CGN dialogues), the news broadcasts (CGN news) and the read-aloud books (CGN books). Finally, we used the IFADV corpus (Van Son et al., 2008), a collection of recorded and transcribed dialogues, containing approximately 70,000 word tokens.

We preprocessed the COW corpus by excluding sentences with three or more word or character repetitions, or with characters not used in standard Dutch orthography. The following preprocessing steps were performed for all language materials from all corpora. We replaced characters with diacritics to the equivalent characters without diacritics, and mapped all numbers, websites and tagged words (e.g., #tag#) to special word codes. We removed punctuations, except for commas. We normalized shortened words with apostrophes to a standard spelling (e.g., *'t'* becomes *het* 'the'). For an overview of the processed text materials see Table 2.

IFADV, CGN news and CGN books contain language materials used in the EEG experiment. For the purpose of SLM training, we removed these particular materials. Subsequently, we created register-specific sets by combining CGN dialogues with IFADV, CGN books with SoNaR books and, finally, CGN news with SoNaR news. We will refer to these sets as *dialogues*, *books,* and *news* respectively. Each set was split randomly into a training set with 80 % of the materials and a test set with the remaining 20 % of the text materials. We used all preprocessed materials from COW for training purposes.

Table 3
Performance of SLMs expressed as the rounded perplexity score on the experimental and (test) materials. Lower scores indicate better performance in terms of perplexity. Best performance per register (column) in bold face, second best underlined.

SLM	dialogues	news	books
Generic	3943 (4340)	807 (1312)	1736 (1834)
Recency-based	**328** (453)	**287** (325)	**343** (371)
Dialogues	454 (460)	723 (955)	722 (923)
News	1188 (1384)	314 (327)	828 (639)
Books	1463 (1775)	601 (623)	714 (417)

2.2.2. Statistical language modeling

We trained the SLMs with the aid of the SRILM toolkit (Stolcke, 2002) and used the same settings for each language model; a tetragram SLM with Kneser-Ney discounting (Chen & Goodman, 1999) for smoothing.

We trained separate SLMs on the following training materials: COW, dialogues, news, and books. The SLM trained on the COW materials will be referred to as the *generic* SLM. This SLM was also used for the computation of the *recency-based* SLM and as the background language model which we interpolated with the SLMs trained on the dialogues, news and book training materials to create *register-specific* SLMs.

To find the best interpolation weights for the register-specific SLMs, we interpolated each with the background SLM (trained on the COW corpus) and tested a series of weights. We chose the weight resulting in the lowest perplexity on the register-specific held-out test materials (perplexity is a performance metric for SLMs whereby a lower score indicates better performance). The optimal weights for the background model were 0.3 for both news and books and 0.13 for the dialogues model.

Finally, we created a recency-based SLM, based on the generic model trained on the COW materials. We determined the optimal cache size (number of preceding words used to update the SLM) by testing different sizes (i.e., 2, 4, 8, …, 512, 1024 words) on the test materials of the different registers. The SLM performance asymptotes quickly with increasing cache sizes and we therefore selected a cache size of 64.

Table 3 shows an overview of the perplexity scores for each SLM on the materials used in the EEG experiment and (between brackets) the score on the test materials. We observe that each register-specific model performs better on the corresponding register material compared to the other materials (the mismatching register materials), and the recency-based model performs better still. The book SLM performed worse on the materials used in the EEG experiment than the testing materials, indicating a discrepancy between the training and test materials and the language materials used in the EEG experiment. However, this model is still better than the generic SLM (i.e., 714 versus 1736).

2.2.3. Word surprisal estimation

To estimate word surprisal, we used the generic, recency-based, and register-specific SLMs described in Section 2.2.2. The different SLMs were used to estimate the surprisal of each word in the experimental speech materials. We used the generic and recency-based SLM to estimate *generic* and *recency-based* word surprisal, respectively. The register-specific SLMs were used to compute *register-specific* word surprisal for

Table 4
Comparisons between the generic, register-mismatch, recency probability estimates on the one hand and the register-specific word probability estimates on the other hand, based on the AIC of LME models (AIC difference between parenthesis). The p-value indicates the probability that a model with generic, mismatch, or recency estimates is a better fit than the register-specific word surprisal.

LME model	**relative likelihood (Δ AIC)** register-specific
generic	p <.001 (74)
register-mismatch	p <.001 (99)
recency-based	p <.001 (31)

Table 5
Overview of the fixed effects of the linear mixed effect model with the N400 as dependent variable. For each fixed effect, its name, the beta (β), the standard error (SE β) and the *t*-value (t) are reported. The predictor of interest (register-specific word surprisal) is bolded.

fixed effect	**β**	**SE β**	**t**
intercept	−0.378	0.044	−8.65
baseline	5.364	0.009	582.58
log word frequency	0.140	0.030	4.70
experiment news	0.371	0.029	12.72
experiment books	0.256	0.026	9.76
surprisal	**−0.171**	**0.029**	**−5.83**
word duration	0.083	0.017	4.98
word position in sentence	0.183	0.010	17.91
exp. news: **surprisal**	**0.183**	**0.033**	**5.46**
exp. books: **surprisal**	0.038	0.029	1.30

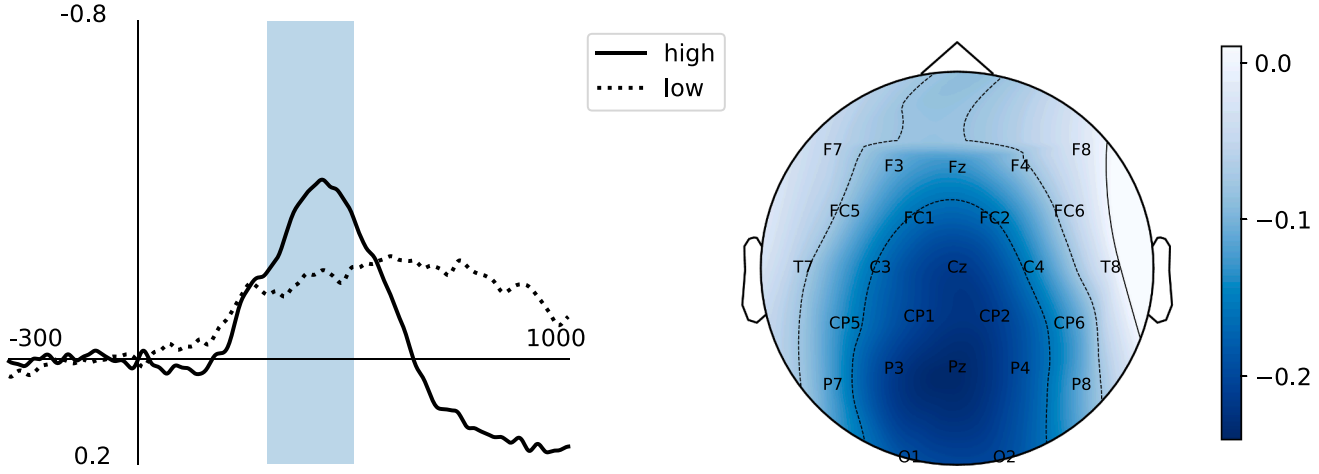


Fig. 1. (left) Grand average plot of the ERP response averaged over all content words, participants and channel set. The blue shaded area indicates the analysis window (300 – 500 ms from word onset). X-axis shows time in milliseconds and y-axis amplitude in μvolt. The solid line shows the average of words with highest tertile generic word surprisal, the dotted line the lowest tertile. (right) Topographic difference plot between content words with the highest tertile generic word surprisal values versus words in the lowest tertile. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

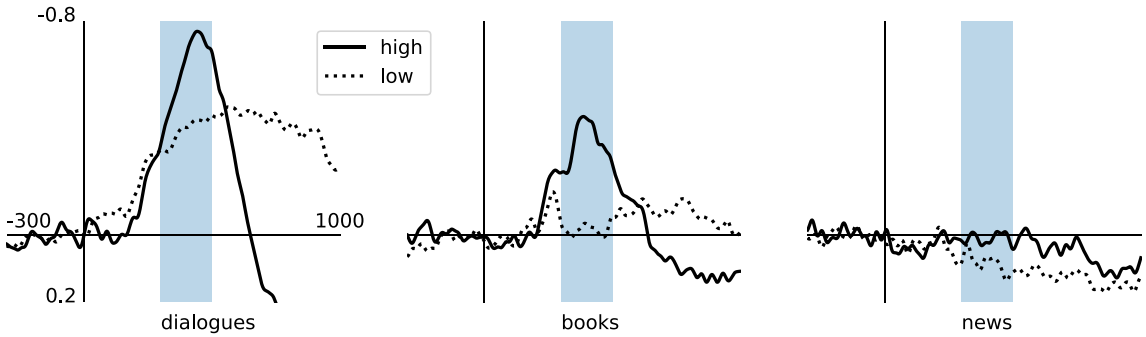


Fig. 2. Grand average plots of the ERP response averaged over all content words, participants and channel set, but split between speech registers: dialogues, books and news. The solid line shows the average of words with highest tertile register-specific surprisal, dotted line the lowest tertile. The blue shaded areas indicate the analysis window (300 – 500 ms from word onset). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table B1
Number of words, word types and median duration of words in milliseconds in the final dataset.

Experiment	# words	# word types	median duration in milliseconds	word in sentence median (max)
Dialogues	258,971	1,600	210	5 (20)
Books	175,880	1,814	302	4 (16)
News	134,080	1,428	289	5 (18)
All	568,931	3,423	254	5 (20)

Table B2
Number of word types overlapping between experiments in the final dataset. Percentage overlap is computed based on the total number of word types of an experiment of a given row.

	Dialogues	Books	News
Dialogues (1,600)		653 (41 %)	569 (36 %)
Books (1,814)	653 (36 %)		593 (33 %)
News (1,428)	569 (40 %)	593 (42 %)	

Table B3
Number of word tokens overlapping between experiments in the final dataset. Percentage overlap is computed based on the total number of words of an experiment of a given row.

	Dialogues	Books	News
Dialogues (258,971)		212,936 (82 %)	191,684 (74 %)
Books (175,880)	123,818 (70 %)		113,359 (65 %)
News (134,080)	90,427 (67 %)	91,620 (68 %)	

Table B4
Twenty most common words per experiment, with approximate English translations and the number of occurrences in the final dataset.

Dialogues (word count)	Books (word count)	News (word count)
is ‘is’ (8,445)	niet ‘not’ (5,103)	is ‘is’ (5,772)
wel ‘good’ (7,966)	is ‘is’ (3,856)	zijn ‘be’ (3,133)
ook ‘also’ (7,945)	was ‘was’ (3,345)	niet ‘not’ (2,993)
niet ‘not’ (6,947)	had ‘had’ (2,581)	jaar ‘year’ (2,493)
zo ‘later’ (6,357)	nog ‘still’ (2,271)	nog ‘still’ (2,037)
dan ‘than’ (6,315)	wel ‘good’ (1,736)	heeft ‘has’ (1,990)
was ‘was’ (5,724)	ook ‘also’ (1,,724)	worden ‘become’ (1,835)
nog ‘still’ (4,241)	zo ‘later’ (1,699)	ook ‘also’ (1,317)
echt ‘really’ (4,185)	vader ‘father’ (1,673)	mensen ‘people’ (1,302)
gewoon ‘just’ (3,889)	heeft ‘has’ (1,611)	al ‘already’ (1,274)
heb ‘have’ (3,601)	zijn ‘be’ (1,468)	wordt ‘becomes’ (1,195)
heel ‘whole’ (3,400)	nu ‘now’ (1,369)	was ‘was’ (1,175)
dus ‘so’ (3,034)	al ‘already’ (1,317)	hebben ‘have’ (1,003)
nou ‘well’ (2,855)	moeder ‘mother’ (1,281)	nieuwe ‘new’ (898)
beetje ‘just’ (2,747)	toch ‘however’ (1,196)	kunnen ‘can’ (879)
zijn ‘be’ (2,710)	dan ‘than’ (1,050)	weer ‘again’ (862)
maar ‘but’ (2,609)	heb ‘have’ (1,043)	gaan ‘go’ (832)
goed ‘good’ (2,608)	huis ‘house’ (996)	moeten ‘must’ (816)
toen ‘then’ (2,353)	eens ‘once’ (992)	procent ‘percentage’ (788)
had ‘had’ (2,338)	weer ‘again’ (955)	wil ‘want’ (768)

the different register-specific materials, i.e., the dialogues SLM was used to estimate word surprisal in the dialogue materials, etcetera. Finally, we used mismatching pairs of registers, e.g., books SLM to estimate probability for words in the news materials. We used the following mismatch pairs: books-news, news-dialogues, news-books. We refer to this as *register-mismatch* word surprisal.

2.3. Procedure

Participants came to the lab on three separate occasions. Consecutive visits were a week or more apart. Participants were fitted with the correct size electrode cap and the electrodes were placed. The participants were seated in a sound-attenuating booth and listened to approximately 90 min of speech from a specific register (i.e., dialogues, books, or news), 270 min in total for the three experiments. The order of the speech registers was counterbalanced across participants. The audio materials were presented via in-earphones (Etymōtic ER1) at a comfortable listening volume. To this end, a short speech fragment (corresponding to the register but not part of the further experiment) was played to check the volume. When necessary, the ear-plugs or volume were adjusted. The participants were asked to sit still and keep eye movements and blinks to a minimum.

The audio materials were presented in blocks of approximately 15 min. The order of block presentation was counterbalanced across participants. After each block, the participant could take a break before the experiment continued.

To ensure participants listened attentively, yes–no comprehension questions were visually presented during breaks in the experiment and participants responded with a button box. For example, a participant could be asked: *Heeft ze het British Museum bezocht?* ‘Did she visit the British Museum?’ For both the dialogues and books the questions were presented at the end of each block. For the news experiment the questions were present at approximately 4-minute intervals, to compensate the higher information density of the materials compared to the other registers. During the dialogues, books and news experiments 36, 42 and 84 yes–no comprehension questions were presented, respectively.

2.4. EEG recording

The electroencephalogram (EEG) was recorded from 26 silver-chloride cap-mounted electrodes. The electrodes were placed according to the Standard International 10–20 System (Fp2, Fz, F3, F4, F7, F8, FC1, FC2, FC5, FC6, Cz, C3, C4, T7, T8, P3, Pz, P4, P7, P8, CP1, CP2, CP5, CP6, O1, O2). Four additional electrodes were used to monitor eye-related artifacts (eye movements and blinks), placed at the outer left and right canthi, and below and above the left eye (converted off line to horizontal and vertical electro-oculogram (EOG) signals). Two additional electrodes were placed on the left and right mastoids. All electrodes were referenced to the left mastoid electrode and electrode impedances were below 15 kΩ before recording started. The EEG-data was amplified with an Easycap system and band-pass filtered with 0.01 and 100 Hz cut off frequencies and digitized at a 1000 Hz sample frequency.

2.5. Preprocessing

The data were re-referenced off-line to the mean of the left and right mastoids and filtered with a 5th order Butterworth bandpass filter with cut-off frequencies at 0.05 and 30 Hz. We removed sections containing artefacts from the data in a semi-automatic fashion. To automatically detect artefacts in the EEG materials, we first manually annotated 60 hours (out of a total of 207 hours) of EEG materials for artefacts. Based on the manual annotations we trained a convolutional neural network classifier with Tensorflow (Abadi et al., 2016) to detect these artefacts. The classifier was trained such that it was very sensitive to artefacts, erring on the side of classifying more EEG materials as artefact to find as many as possible. The classifier achieved an F1 score of 0.89 on unseen EEG materials. We used this classifier to classify all EEG materials for artefacts. Subsequently, we manually checked all found artefacts and made corrections when needed.

Individual channels were removed when a channel was contaminated with artefacts for minimally 40 % of an experimental block. Otherwise, we removed the section (all channels) where one or more channels showed artefact corruption. The Fp2 channel was removed for all recordings, due to poor overall signal quality.

After artefact removal, independent component analysis (ICA) was used to filter out activity related to eye blinks and eye movement. Following Winkler et al. (2015), the ICA was computed on the EEG data band-passed filtered at cut off frequencies of 1–30 Hz. Subsequently, components were selected that reflected eye blinks and eye movements based on visual inspection of topographic and time-course plots. The ICA solution was then used to recompose the EEG data (band-pass filtered at cut off frequencies of 0.05–30 Hz) without the eye-activity-related components. This approach attenuates the sensitivity of ICA to slow drift (Winkler et al., 2015) without adversely affecting ERP analysis (see Tanner et al., 2015).

We extracted EEG-data in the time window −300 to 1000 ms relative to word onset, for each content word (i.e., nouns, verbs, adverbs and adjectives) in our dataset. We used the following exclusion criteria to construct the dataset. We excluded items which overlapped with artefactual EEG data or if the signal exceeded ± 75 μV in the previously defined time window of the word. We excluded all data from nine participants because less than 40 % of the data remained after artefact removal. We excluded all words from a stop list of words (see Appendix A), and excluded words that occurred in overlapping speech (only relevant in the dialogues experiment). We excluded the first word of each sentence, to lower the correlation between word surprisal and word frequency (a covariate in our statistical model, see below). We excluded words shorter than 50 ms or longer than 700 ms and words that occurred less than 24 times in the dataset, to lower the number of word types in the experiment (from 5,866 to 3,423). A smaller set of word types was needed to achieve convergence of the statistical model. Across all experiments, these steps resulted in a dataset of 568,931 word epochs.

No participants were excluded based on the yes–no comprehension results that were not already excluded based on the EEG data quality. Overall, the participants performed well on the yes–no comprehension question: with 83 % correct for the news experiment, 96 % correct for the books experiment and 94 % correct for the dialogues experiment.

2.6. Analysis

Based on previous literature (see Frank et al., 2015), we defined the N400 amplitude as the average of the channel set C3, C4, Cz, CP5, CP1, CP2, CP6, P7, *P*3, Pz, *P*4, P8, O1, O2 within the time window 300 – 500 ms after word onset. Following Frank et al. (2015), we did not subtract the baseline from the ERP. Instead, the baseline was used as a covariate in the statistical model. We computed the baseline by averaging the amplitude over the time window −150 – 0 ms (relative to word onset) and the same channel set.

We estimated several linear mixed effect (LME) models (Bates et al., 2015) with the statistical package R (R Core team, 2015) to predict the N400 amplitude. We first estimated a *null* LME model with the following standardized covariates: the aforementioned *baseline*, the log *word frequency* (based on the COW corpus), the *word duration*, the *word position in the sentence* and finally a factor for *experiment* (with three levels, one for each register). In addition, we added *participant* and *word* as random effects with random slopes for surprisal for both participant and word.

The predictor of interest (word surprisal) was added to the null model to create a *generic, recency-based, register-specific,* and *register-mismatch* LME model, based on the corresponding word surprisal type (i.e., generic word surprisal corresponds with a generic LME model). We also added an interaction term between word surprisal and experiment to allow for differences between speech registers. We considered to include a random slope for word surprisal by participant but this resulted in convergence issues.

3. Results

Model comparison with the anova likelihood-ratio test revealed that the LME model with generic word surprisal improved compared to the null model ($\chi^2 = 553.46$, $p < .001$). The N400 amplitude is more negative with increasing values of word surprisal (see Fig. 1).

Subsequently we compared the generic, register-mismatch, recency-based, and register-specific LME models. For these comparisons, we were precluded from using the anova likelihood-ratio test since these models were not nested versions of each other. We therefore compared the AIC of each LME model and computed the corresponding relative likelihood. This comparison revealed that the register-specific word surprisal values best predict the N400 amplitude (see Table 4, left). The recency-based model performed better than the generic model, and the register-mismatch model.

In the register-specific LME model (see Table 5), the interaction term for the news materials and word surprisal has a *t*-value of 5.46. To further investigate this interaction effect, we split the data according to register (dialogues, books and news) and fitted LME models to each subset. The LME models for the news materials failed to converge. We therefore computed the news LME models without random slopes for surprisal for participant and words. Model comparison with the anova likelihood-ratio test revealed that the LME model with register-specific word surprisal improved compared to the null model for both dialogues ($\chi^2 = 260.73$, $p < .001$) and books ($\chi^2 = 279.97$, $p < .001$), while this was not the case for the news materials ($\chi^2 = 0$, see also Fig. 2).

4. Discussion

In the current study, we recorded the EEG signal from participants who listened to long (4 – 15 min) stretches of natural speech sampled form different speech registers: dialogues, news broadcasts, and read-aloud books. The speech materials were analyzed with statistical

language models (SLM) estimating word surprisal. We found that the N400 amplitude was more negative for words with higher surprisal (i.e., unexpected words).

We investigated the influence of speech register on prediction in speech comprehension by estimating and comparing different word surprisal 'types'. We compared *generic* with *register-specific* word surprisal and found that register-specific word surprisal best predicted the N400 amplitude. This finding indicates that listeners are sensitive to the specific statistical structure of the speech register they listen to, and that they adjust their anticipations accordingly. To test whether the adaptation of word anticipations was the result of register, we also compared register-specific word surprisal with register-mismatch word surprisal. This comparison provided a sanity check to test whether any 'specific' word surprisal would better predict the N400 amplitude compared to generic word surprisal. Register mismatch was defined as the surprisal estimated on mismatching register materials, e.g., the SLM was trained on books but used to estimate surprisal for the news materials. We found that register-mismatch word surprisal did not improve upon generic word surprisal, providing further evidence that register-specific information influenced participants' word expectations.

Furthermore, we tested whether the register-specific effects could be explained merely by tracking recent input. In theory, listeners could adapt their expectations not based on register characteristics, but just on recent language input. We therefore also compared the register-specific word surprisal to recency-based word surprisal. The recency-based word surprisal is computed by updating the generic SLM with caching of a number ($n = 64$) of recent words. As Table 4 shows, the recency-based word surprisal better predicts the N400 amplitude compared to the generic word surprisal. Importantly, the register-specific word surprisal does better still. This finding indicates that listeners do not only use recent language input to adjust their predictions of upcoming words but also register information. Listeners may have stored representations of the statistical structure of registers, whereby different expectations are generated when listening to a story than when listening to a dialogue.

Our results are also relevant for the question whether prediction occurs during normal language processing (Huettig, 2015; Nieuwland et al., 2018). In our experiments, we used long stretches (4 – 15 min) of naturalistic speech. Therefore, there are no artificial pauses between the presentation of words, which could potentially influence predictive processing (Luka & Van Petten, 2014). Our finding shows that listeners anticipate words in normally-paced language input. Furthermore, we investigated most words in the speech materials, which allows for the investigation of predictive language processing across the whole probability spectrum, from very unexpected to highly expected words. This is relevant in light of Huettig's (2015) criticism that most evidence for prediction comes from comparing extremes of predictability.

Do the presented results provide evidence for prediction? Pickering & Gambi (2018) note that effects found on the target word, such as surprisal effects, can alternatively be explained by an integration account. This interpretation might be influenced by the traditional assumption that prediction entails one or at most a few words. However, we argue that surprisal effects should be interpreted in a different way (leading, in our opinion, to a clearer interpretation of the underlying mechanism). The surprisal value is not a static attribute of a single word. Instead, it derives from a probability distribution that is computed over an entire lexicon based on the preceding context. The prediction is not a specific word or small set of words. The prediction *is the specific probability distribution over the lexicon*. To state it differently, each word in the mental lexicon is 'pre-activated' to the extent it is a probable continuation given the preceding context.

In our experiment, each target word (with corresponding surprisal value) samples the predicted probability distribution of the lexicon. The word does not have this surprisal value in isolation, a probability can only be defined over a set of options, in this case a lexicon. Sampling these predicted probability distributions explains variance in the amplitude of the N400: it is more negative for when sampling less probable continuations and more positive when sampling more probable continuations. Therefore, our results can be interpreted as evidence for these predicted word probability distributions, and we propose that our data are therefore best explained by a predictive account.

Furthermore, probability effects found during language perception, especially those found over a wide probability spectrum such as Smith & Levy (2013), Frank et al. (2015), and the effects reported in this article, fit comfortability within more general and well attested predictive perception framework, such as predictive coding (Friston, 2005; 2012), in which top-down predictions result in prediction error from bottom-up input. And while it might be a logical possibility that an integration account could explain these specific effects, it would need further specification beyond the bare claim that more probable words are easier to integrate (see Pickering & Gambi, 2018). Furthermore, if we apply Occam's razor, a predictive account is preferable, since it is more parsimonious when we assume language perception is similar to more general perception mechanisms.

We want to emphasize that the preceding argument does not exclude the N400 effect as an index for integration mechanisms. Indeed, several studies (e.g. Frank & Willems, 2017; Nieuwland et al., 2020) suggest that these processes are dissociable, where the variance of the N400 effect is partly captured by word predictability and partly by semantic plausibility operationalized by either participant rating tests or automatic techniques such as latent semantic analysis.

The current results do show that listeners engage in predictive language processing while listening to natural everyday speech (without artificially constraining sentences). This result is in line with the results reported for reading by Smith & Levy (2013) and Frank et al. (2015). Interestingly, a recent article by Heilbron et al. (2022), reported similar findings with speech materials using a state of the art GTP-2 language model (i.e. a language model based neural networks).

We found an unexpected difference between the speech registers: we did not observe an N400 effect for the news broadcast speech materials (see Fig. 2). It is unlikely that this difference was caused by news broadcasts being less predictable than the other speech materials: The perplexity scores for SLMs tested on news materials were comparable to scores for dialogues and books (see Table 3), indicating that the SLMs could predict upcoming words in the news materials with performance similar to for the other register materials. If news broadcasts were less predictable, the SLMs performance should drop accordingly.

An explanation for the interaction effect between word surprisal and register could be participants' attention to the speech materials. Participants possibly found it harder to concentrate on the news materials compared to dialogues and book materials. Attention difficulties for the news materials could be caused by the high topic density in this register. The news materials consisted of sequences of short news items on many different topics. In fact, because of this high density of topics, we decided to segment the news materials into 4-minute sections, while books and dialogues materials were segmented into 12- and 15-minute sections, respectively. Still the participants performed worse on average for the comprehension questions on news (83 % correct) than on books (96 % correct) and dialogues (94 % correct), indicating that they indeed found it harder to pay attention to the news materials. There is evidence that attention can modulate the N400 (for a discussion, see Kutas & Federmeier, 2011), but it is unclear to what extent lack of attention would completely suppress the N400 effect.

We found an unexpectedly high correlation between word surprisal and log word frequency. A high correlation between the predictor of interest (word surprisal) and a covariate make statistical results less reliable (e.g., effects can flip, because the variance can be ascribed to either of the variables). An explanation for the unexpectedly high correlation is related to the first word in a sentence. The dialogues materials contain a high number of very short sentences resulting in a relatively high proportion of first words. Statistical language models (SLM) generally do not use cross-sentence-boundary pre-context. Therefore, the word surprisal of the first word in a sentence will tend to the

frequency of that word. We therefore removed the first word of each sentence for our analyses. In future studies, it would be interesting to test whether SLMs could be used that take cross-sentence-boundary pre-context into account.

Our study raises questions for future research. First, how do listeners adjust their expectations to a specific register? Our results show that simply using the most recent words to adjust anticipations does worse in modelling N400 amplitude in listeners compared to using register-specific information. This indicates that listeners do not merely use recent context to adjust expectations, and could imply that registers are represented in the listener's mind in some form and can be utilized to adapt expectations to upcoming input. This could mean that multiple generative models (e.g., registers, schema's) are represented and language perceivers switch between these models (see also Kuperberg, 2016).

A second question for future research is whether speech register provides the correct level of granularity for a predictive model of language? The current study found evidence that listeners can use register-specific information to adjust their anticipations. However, register is a high-level construct that correlates with, for example, topic. It could be that topic differences are also an important factor in structuring language perceivers' expectations.

Third, how to interpret the success of SLMs in modelling language perceivers' processing costs? SLMs are an implausible cognitive model for language prediction. For example, an SLM could not model prediction effects produced by humans with sentences 9 & 10 (Section 1.4), because these effects are based on long range dependencies. What aspects of predictive human language processing do SLM capture that make them successful in modelling processing costs and when would they fail?

5. Conclusion

We analyzed ERPs elicited with spoken words from long stretches (4–15 min) of naturalistic speech and found that word surprisal predicts the N400 amplitude. Listeners anticipate words while listening to natural speech that is not highly constrained nor limited to very likely or very unlikely words. Moreover, by comparing generic, recency-based, and register-specific word surprisal, we showed that listeners broadly adapt their expectations to the register of the speech they are perceiving, which indicates that listeners also use cues from the wider context to predict upcoming words.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We would like to thank Lou Boves and Tineke Snijders for helpful discussions and Tim Zee for his invaluable help with the EEG annotations.

Funding

This work was supported by Radboud's Centre for Language Studies, a Consolidator grant from the European Research Council [grant number 284108] and a Vici grant from the Netherlands Organization for Scientific Research. Both grants were awarded to Prof. dr. M.T.C. Ernestus.

Appendix A

Stop word list, a list of words to be excluded from the dataset.

w	single letter word in transcription
i	single letter word in transcription
l	single letter word in transcription
e	single letter word in transcription
n	single letter word in transcription
c	single letter word in transcription
k	single letter word in transcription
a	single letter word in transcription
r	single letter word in transcription
b	single letter word in transcription
j	single letter word in transcription
uh	interjection, mislabeled by tagger
wo	abbreviation
ah	interjection, mislabeled by tagger
za	abbreviation
ha	interjection, mislabeled by tagger
uhm	interjection, mislabeled by tagger
tvs	typo in transcription
tja	interjection, mislabeled by tagger
goh	interjection, mislabeled by tagger
kof	typo in transcription
des	typo in transcription
hot	English word
kch	typo in transcription
cnv	abbreviation
joh	interjection, mislabeled by tagger
geg	typo in transcription
war	typo in transcription

(*continued on next page*)

(*continued*)

bah	interjection, mislabeled by tagger
bad	English word
fol	unknown, typo in transcription
hum	interjection, mislabeled by tagger
pub	English word
wow	interjection, mislabeled by tagger
oll	typo in transcription)
ggg	transcription code for uninterpretable speech
jee	typo in transcription
oke	'ok'
hup	interjection, mislabeled by tagger
mina	typo in transcription
juno	name
datis	unknown, typo in transcription
cd-rom	abbreviation
ns-top	abbreviation
marjak	typo in transcription
molsla	typo in transcription
eu-top	abbreviation
do-door	typo in transcription
wao'ers	abbreviation
hondsdr	typo in transcription
thijsen	name
ge-goed	typo in transcription
ing-bank	abbreviation
esf-geld	abbreviation
ex-beatle	'ex beatle'
barteling	name
mkz-virus	abbreviation
fnv-leden	abbreviation
mkz-crisis	abbreviation
mkz-boeren	abbreviation
ij-kantine	name
knsm-eiland	abbreviation
cbs-cijfers	abbreviation
eu-collegas	abbreviation
ns-stations	abbreviation
ns-directie	abbreviation
lufthansa-piloten	name
lockheed-affaire	name
mkz-gebieden	abbreviation
radio-1-journaal	name
vn-klimaattop	abbreviation
encarta-encyclopedie	name
vn-vluchtelingenverdrag	abbreviation
landbouw-uh-universiteit	stuttering
endemol-aandelen	name
mkz-problemen	abbreviation
vn-klimaatconferentie	abbreviation
klauwzeercontroles	old and very specific word
nipo-enquete	abbreviation
kyotoafspraken	name
cao-onderhandelingen	abbreviation
asterix-stripalbum	comic book name
nipo-onderzoek	abbreviation
mkz-maatregelen	(abbreviation
xtc-laboratorium	abbreviation
pvda-partijleider	abbreviation
gsm-abonnement	abbreviation
mkz-uitbraken	abbreviation
cda-fractieleider	abbreviation
pvda-politicus	abbreviation

Appendix B

Descriptive statistics for the language materials in the dataset. See Tables B1-B4.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. *OSDI, 16*, 265–283.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in cognitive sciences, 11*(7), 280–289.

Barbieri, F. (2005). Quotative use in American English: A corpus-based, cross-register comparison. *Journal of English Linguistics, 33*(3), 222–256.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bentum, M., ten Bosch, L., van den Bosch, A., & Ernestus, M. (2019a). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics, 24*(1), 98–130.
Bentum, M., Ten Bosch, L., Van den Bosch, A., & Ernestus, M. (2019b). Listening with great expectations: An investigation of word form anticipations in naturalistic speech. In *Interspeech 2019: 20th Annual Conference of the International Speech Communication Association* (pp. 2265–2269).
Biber, D. (1999). A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse studies, 1*(2), 131–150.
Biber, D., & Conrad, S. (2001). Register variation: A corpus approach. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 175–196). Malden, Mass: Blackwell Publishers.
Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. New York, NY: Cambridge University Press.
Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer [Computer program]*.
Borrillo, J. M. (2000). Register Analysis in Literary Translation: A Functional Approach. *Babel Revue internationale de la traduction / International Journal of Translation, 46*(1), 1–19.
Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language, 13*(4), 359–393.
De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences, 22*(9), 764–779.
DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience, 8*(8), 1117–1121.
Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition, 24*(2), 143–188.
Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science, 33*(4), 547–582.
Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language, 41*(4), 469–495.
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.
Frank, S. L., & Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience, 32*(9), 1192–1203.
Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences, 360*(1456), 815–836.
Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology, 83*(2), 248–252.
Friston, K. (2018). Does predictive coding have a future? *Nature neuroscience, 21*(8), 1019.
Hagoort, P., & Brown, C. (1994). Brain responses to lexical ambiguity resolution and parsing. *Perspectives on sentence processing, 14*, 45–80.
Hagoort, P., & Brown, C. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia, 38*(11), 1518–1530.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001* (pp. 159–166).
Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences, 119*(32). e2201968119.
Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research, 1626*, 118–135.
Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics, 6*(1), 97–133.
Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics, 33*(1), 147–151.
Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135*(1), 12–35.
Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience, 31*(5), 602–616.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and 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.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.
Luka, B. J., & Van Petten, C. (2014). Prospective and retrospective semantic processing: Prediction, time, and relationship strength in event-related potentials. *Brain and Language, 135*, 115–129.
Nicholls, D. (2014). *Us*. New York, NY: HarperCollins Publishers.
Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., et al. (2020). Dissociable effects of prediction and integration during language comprehension: Evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B, 375*(1791), 20180522.
Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife, 7*.
Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics, 29*(1), 19–51.
Oostdijk, N. (2001). The design of the Spoken Dutch Corpus. *Language and Computers, 36*(1), 105–112.
Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns, & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch* (pp. 219–247). Berlin: Springer.
Otten, M., & Van Berkum, J. J. A. (2007). What makes a discourse constraining? Comparing the effects of discourse message and scenario fit on the discourse-dependent N400 effect. *Brain research, 1153*, 166–177.
Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological bulletin, 144*(10), 1002–1044.
Poulton, V. R., & Nieuwland, M. S. (2022). Can you hear what's coming? Failure to replicate ERP evidence for phonological prediction. *Neurobiology of Language, 3*(4), 556–574. https://doi.org/10.1162/nol_a_00078
R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for statistical Computing. http://www.R-project.org/.
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.
Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)* (pp. 28–34). Mannheim: Institut für Sprache.
Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *LREC* (pp. 486–493).
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319.
Staples, S., Egbert, J., Biber, D., & Conrad, S. (2015). Register Variation A Corpus Approach. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (pp. 505–525). Malden, Mass: Wiley Blackwell.
Stolcke, A. (2002). SRILM-an extensible language modelling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.
Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology, 52*(8), 997–1009.
Van Berkum, J. J. A. (2012). The electrophysiology of discourse and conversation. In M. Spivey, M. Joanisse, & K. McRae (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 589–612). Cambridge: Cambridge University Press.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443.
Van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: A Free Dialog Video Corpus. In *LREC* (pp. 501–508). Marrakech: ELRA.
Wicha, N. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of cognitive neuroscience, 16*(7), 1272–1288.
Winkler, I., Debener, S., Müller, K. R., & Tangermann, M. (2015). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (pp. 4101–4105). IEEE.
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). *BioRxiv*, 143750.