Listening with great expectations

A study of predictive natural speech processing

The research reported in this dissertation 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.

2021, Martijn Bentum
ISBN: 978-94-6421-286-0
Cover image: Inspired by the photograph *Hidden dalmation* by RC James
Cover design: Fenna Oosterhoff
Printed and bound by Ipskamp Drukkers b.v.

Listening with great expectations

A study of predictive natural speech processing

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 22 april 2021
om 10.30 uur precies

door
Martijn Bentum
geboren op 28 mei 1982
te Alkmaar

Promotoren:	Prof. dr. M.T.C. Ernestus
	Prof. dr. A.P.J. van den Bosch
Copromotor:	Dr. L.F.M. ten Bosch

Manuscriptcommissie:

Prof. dr. F. Huettig

Prof. dr. A.S. Meyer (Max Planck Institute Nijmegen)

Prof. dr. R.W.N.M. van Hout

Dr. T.M. Snijders (Max Planck Institute Nijmegen)

Dr. S.L. Frank

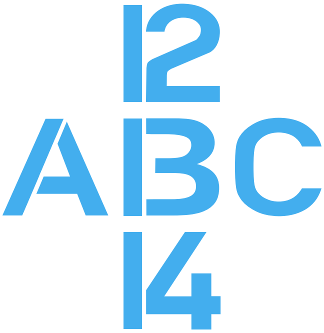
You shall know a word by the company it keeps

John R. Firth, 1957

Contents

Introduction

Chapter 1



Perception is not limited to the simple sensory analysis of transduced physical input, such as photons hitting the retina or air pressure fluctuations on the ear drum. The above picture (seen in Clark, 2015) nicely illustrates that context shapes perception: In a column of numbers the middle character jumps out as *13*. In a row of capital letters, the middle character morphs into *B*. In general, a number is more likely to follow after another number and the same holds for letters[1]. The brain uses this contextual information to anticipate likely upcoming input and resolve ambiguity, such as, *13* in the context of numbers and *B* in the context of letters.

Context sensitivity also influences language processing. The human language processing system is highly sensitive to statistical regularities found in language. An example of this statistical sensitivity is the word frequency effect, a ubiquitous finding across different experimental paradigms, whereby readers and listeners more easily process a frequent word, such as *thief*, compared to an infrequent word, such as *rogue*. I interpret this and other findings, which I discuss further below, as evidence for predictive language processing.

[1] Based on counts from the web-crawled corpus NLCOW2014 (Schäfer, 2015), it is ~ 65 times more likely that a number follows a number (compared to a letter following a number) and ~ 300,000 times more likely that a letter follows a letter compared to a number following a letter.

In the current thesis, I extend the research into predictive language processing by studying the influence of context on listeners' anticipation of words and their auditory form while listening to natural speech.

1.1 Predictive language processing

There is ample experimental evidence that the human language processing system uses anticipatory mechanisms to process language input (for overviews see Ellis, 2002; Elman, 2009; Luka & Van Petten, 2012; Huettig, 2015; Kuperberg & Jeager, 2016; Norris, McQueen, Cutler, 2016). Importantly, the evidence for predictive language processing is based on many different experimental paradigms. I discuss a few examples.

Eye-tracking studies use a camera to continuously register participants' gaze location. With the visual world paradigm, the concurrent processing of auditory speech input and visual input of pictures is investigated. Altman & Kamide (1999) used this setup to investigate predictive processing. They presented, for example, a picture with a boy, a cake and a toy car while the participant listened to a sentence such as *The boy eats…* or *The boy moves…* The verb *eats* elicits more looks towards the cake compared to *moves*, indicating that listeners anticipate an edible object when the verb affords this. Many subsequent studies show that multiple sources of information are used to incrementally update the mental model of the linguistic and visual input (see Huettig, Rommers & Meyer, 2011, for an overview of visual world paradigm studies).

Smith & Levy (2013) used eye-tracking during reading to investigate the relation between word predictability and reading time. They estimated the probability of words given the preceding context with the aid of statistical language models (on which I will go into more detail in Section 1.6). Their analysis showed that word probability is log-linearly related to reading times, such that likely words are read faster compared to unlikely words. They found that word surprisal, the negative logarithm of the probability of a word, predicts the reading time of a word. Surprisal is an information-theoretic measure which encodes the amount of Shannon information an item (i.e., word) in a message conveys. Informally, surprisal can be thought of as the 'unexpectedness' of a word, whereby a higher surprisal results in longer reading times.

Electroencephalography (EEG) measures electrical potential differences across the scalp, and it is used to investigate cognitive processes in relation to stimuli. DeLong et al. (2005) used the N400 effect to test whether listeners pre-activate upcoming words. The N400 is responsive to the semantic expectancy of a word, whereby an unexpected word elicits a more negative deflection of the event-related potential (ERP) around 400 milliseconds from word onset. Delong et al. (2005) visually presented words in a fixed-paced reading task, whereby words were presented one at a time on a computer screen. Stimuli consisted of sentences such as *The day was breezy so the boy went outside to fly ... in the park.* The blank was either filled with *a kite* or *an airplane.* They found an N400 effect on the article, whereby *an* evoked a more negative amplitude, indicating that listeners expected the word *kite* and therefore did not expect *an.* Similar findings were also reported in related studies (e.g., Wicha et al., 2004; Van Berkum et al., 2005). However, a recent study (Nieuwland et al., 2018) failed to replicate the effect; see DeLong et al. (2017) for a response[2]. For an overview of predictive language processing as investigated with the N400, see Federmeier (2007) and, for an overview of the N400, see Kutas and Federmeier (2011).

In N400 studies, word predictability is typically estimated with the aid of a cloze test. In a cloze test, participants fill in the empty slot '_' in a sentence such as *The day was breezy so the boy went outside to fly _ in the park.* The responses are counted and if, for example, 80% of the participants filled in a certain word, that word has an 80% cloze tested probability. The highest scoring word also determines the 'constraint' of the sentence, whereby a score such as 80% is considered a highly constraining sentence. More recently, the aforementioned 'word surprisal' has been linked to the N400 by Frank et al., 2015. In a fixed-paced reading task, they found that words with high surprisal (i.e., unexpected words) result in a more negative N400 amplitude (see Kuperberg et al. 2017 for a comparison between word probability and surprisal).

[2] The response from DeLong et al. (2017) was not peer reviewed and commented on a pre-print version of Nieuwland et al. (2018). DeLong et al. (2017) argued that the failure to replicate could be caused by differences between Nieuwland et al. (2018) and DeLong et al. (2005) experimental setup. A reanalysis of the data from the Nieuwland et al. replication by Kuperberg (2017) did reveal an effect similar to Delong et al. (2005). In combination with the findings from Wicha et al. (2004) and Van Berkum et al. (2005) we argue that the evidence still supports the proposition that word (forms) are anticipated based on preceding context.

Notwithstanding the evidence for predictive language processing, there are critical views of the role or relative importance of anticipatory language comprehension. Huettig (2015) reviews the evidence for the role of prediction in language comprehension and notes that evidence is typically based on contrasting highly probable with highly improbable sentence continuations (i.e., low versus high cloze probability words). Also, the presence of many highly constraining sentences in experimental materials may artificially boost the benefit and therefore the use of prediction. Huettig (2015) concludes that it is unclear to what extent predictive language processing occurs when listeners or readers process natural language.

Another confounding factor is the manner of stimulus presentation. For example, many language-related EEG experiments use the fixed-paced reading paradigm, whereby participants read sentences one word at a time, presented sequentially at the center of a computer monitor. This setup has several benefits. Participants do not have to move their eyes, which helps to avoid artefacts related to eye movements. Furthermore, the presentation rate is controlled without input from the participants (e.g., button presses) that could introduce artefacts in the EEG data. This setup does have a drawback, however: The experimenter has to choose a specific inter-stimulus interval (ISI), which potentially influences the found effects (Luka & Van Petten, 2014). For example, prediction effects might be found with a longer inter stimulus interval (ISI) but not with a shorter one (Ito et al., 2016). This raises the question whether evidence for predictive processing generalizes to more natural comprehension situations or is a by-product of the artificial testing conditions.

1.2 Ecological validity or the challenge of studying natural language

A robust approach to experimental research is the precise manipulation of only one experimental variable while keeping everything else constant. Any difference found in the measurements is then only ascribable to the experimental manipulation. In many psycholinguistics and neurolinguistics studies, this is devilishly hard to attain or even approximate: Matching words on duration, frequency of occurrence, stress pattern, number of syllables, syllable structure, etcetera, may leave an experimenter with a very restricted and idiosyncratic stimulus list. Still, the challenge is even harder, because, as Section 1.1 argued, listeners and readers are very sensitive to the statistical structure of the context (see also Lau et al., 2013). In many experiments, non-words, pseudowords, or semantic or grammatically erroneous sentences are presented to participants. Filler items are used to obfuscate the intention and idiosyncrasies of the experiment but these cannot be too different from target items or they may emphasize the experimental manipulation. All in all, typical experimental materials will be far removed from natural language use. Presenting these types of materials to participants could therefore well elicit atypical language processing strategies.

New experimental and statistical methods allow experimenters more freedom in the experimental design, which affords the use of natural language materials (see Willems, 2015 for an overview). Linear mixed effects (LME) models (Bates et al. 2015) can model both fixed and random effects, i.e., the model can incorporate nuisance variation from, for example, participants and items. For predictability effects, statistical language modelling can automatically estimate the word predictability of many more words compared to what is feasible with cloze tests. This allows for sampling a huge number of experimental stimuli from existing language materials (e.g., Frank et al., 2015; Willems et al., 2016), instead of constructing dichotomous sets of probable and improbable sentence continuations (e.g., DeLong, 2005).

1.3 Register variation

One understudied aspect of predictive language studies is the influence of the wider context, that is, beyond short narratives[3]. Registers are characterized by different patterns of language use resulting from differences in communicative context and purpose. The examples below illustrate a range of register variation: (1) conversational speech, (2) formal address, (3) news commentary, and (4) a machine learning text book.

[3] Van Berkum (2012) provides an overview of studies investigating predictive language processing with short narratives. These studies typically limit the context to one or a few sentences.

(1) A: An' uh an' each ti- eh boy did I hesitate,
but I thought now she knows uh the
Goren rule, an' when you say "two"
It's a cut-off, an' sh- an' uh so uh -
B: Yeah.

Schegloff et al., 1977

(2) The world is very different now. For man holds in his mortal hands, the power to abolish all forms of human poverty and all forms of human life.

John F. Kennedy's inaugural speech, 1961

(3) A U.N. monitoring mission, whose presence the United States hoped might help quell the strife, on Saturday suspended its operations.

Tiedemann, 2012

(4) The notation [*a*, *b*] is used to denote the *closed* interval from *a* to *b*, that is the interval including the values *a* and *b* themselves, while (*a*, *b*) denotes the corresponding open interval, that is the interval excluding *a* and *b*.

Bishop, 2006

Conversational speech (1) typically involves communicators in close proximity, which affords immediate feedback signaling successful communication. However, the immediacy of alternatively listening and speaking precludes extensive revision or finetuning of utterances, leading to disfluencies and more frequent use of formulaic speech. In contrast, prepared speech (2 & 3) or written texts (4) do not allow for communicative feedback, since a monologue only allows for minimal interaction and the writer and reader or listener typically share neither time nor space. The asynchrony between preparing a message and receiving it allows the (speech) writer to edit and refine prepared speech or written text, resulting in structurally more complex and lexically richer language (Biber & Conrad, 2001).

To study register variation, one approach consists of comparing the distribution of lexico-grammatical features across registers by counting them in language materials form different registers. These comparisons reveal lexical variation between registers (e.g., Kennedy, 1991; Biber et al., 1994). For example, there is systematic difference in the choice of adverbs in a conversation. Compare, for instance, *It did look pretty bad* with *The mother came away somewhat bewildered* from a news reportage (Biber et al. 1999). Similarly, the distribution of grammatical features differs across registers. For example, the complementizer *that* in a *that*-clause *I think (that) he likes you* can be retained or omitted, depending on the register. Biber et al. (1999) found that conversations preferentially omit *that* while news reportages are more likely to retain it (see Staples et al., 2015; Biber & Conrad, 2009 for overviews).

Register variation could potentially be relevant for predictive language processing. If the choice of words and grammatical construction differ across registers, it is plausible that registers differ in their word predictability. For example, in a conversation, the adverb *pretty* will be more likely than in a news broadcast (similarly with certain grammatical constructions). If these patterns differ systematically among registers, listeners and readers might be sensitive to these differences.

1.4 Research questions

Below I present an overview of the research questions which I address in this thesis.

1. *How can we investigate predictive language processing with event-related potentials (ERP) evoked by words in long stretches of natural speech?*

   In this thesis I investigate predictive language processing with natural language sampled from speech corpora. Participants listen to natural connected speech from different registers while their EEG signal is recorded. The use of natural speech materials and the inclusion of words from across the range of the predictability spectrum potentially improves the ecological validity of the evidence for predictive language processing.

2. *To what extent do listeners anticipate the auditory word form while listening to natural speech?*

   As discussed above, evidence for auditory word form anticipation is based on experiments using highly constraining sentences and a binary grouping of highly likely versus unlikely sentence continuations, which could potentially boost predictability effects. To improve upon this, I develop a *continuous* measure which captures the unexpectedness of the auditory word form given the preceding context. This mismatch measure will be used to model the ERP of participants' EEG signal evoked by words in natural speech. If listeners anticipate auditory word forms, I expect to find a correlation between the mismatch measure and the ERP amplitude.

3. *To what extent do listeners adapt their expectations of upcoming words based on the speech register they are listening to?*

   Language use varies among registers and listeners may adapt to these differences in their predictive processing. To investigate whether speech register variation affects the predictability of words, I conduct a corpus study. If there are systematic register-specific differences in word predictability, listeners might use this information to adapt their word expectations. To test this, I investigate whether listeners adapt their expectations of upcoming words based on the register they are listening to. I estimate word predictability with the aid of statistical language models, which reflect either *register specific* or *non-specific* word anticipatory processing strategies and I compare how well they predict the ERP evoked by content words in the natural speech materials.

1.5 Theoretical framework

1.5.1 Predictive coding

I use predictive coding (Friston, 2005, 2012, 2018) for the theoretic framework of my research. This framework proposes that the mismatch between expected and actual input is important for perception, in a sense made explicit below.

Our senses receive input, such as air pressure modulations or photons. These inputs are transduced to neuronal excitations. Importantly, human perception is not limited to the simple detection of air pressure differences or photons. Instead, we perceive a rich world of meaningful sounds and objects. Predictive coding proposes that to achieve the transformation from sensory input to rich percepts, the brain infers putative causes from the sensory input: e.g., a pattern of air pressure differences as a spoken word *hello* or a pattern of photons striking the retina as a *tree*.

Perception is challenging because it is rather difficult to infer causes from sensory input due to the variability in the sensory input. Words can be uttered in silence or a noisy café; a tree may be partially occluded by a house. Different causes interact and the sensory input we receive is a mixed jumble (i.e., words spoken by different speakers in the café, or the house in front of the tree) and because of this, it is not always possible to compute the underlying causes solely from the mixed sensory input (Friston, 2005).

A solution to this perceptual conundrum is to generate expectations about the input. Predictive coding (Friston, 2005) proposes a hierarchical cognitive architecture, where a generative model at cognitive level *i* modulates activity at cognitive level *i*-1. The mismatch between expectation and bottom-up input is referred to as prediction error. The generative model is adjusted in such a way that it minimizes prediction error. In this manner, the most likely cause for sensory input can be inferred by a combination of expectation (prior experience) and the sensory input.

1.5.2 Prediction

What entails prediction in the context of language processing? In the debate about predictive language perception prediction is sometimes conceptualized as predicting one word or a small set of words (e.g., Van Petten & Luka, 2012). This interpretation leads to an important criticism, namely the potential cost of mispredictions (Jackendoff, 2002; Lau et al., 2013; Kutas et al., 2011). If a specific word is predicted, this prediction is either right or wrong and given the number of options, odds are stacked against correct prediction.

However, prediction does not have to be an *all or nothing* process. Prediction can be probabilistic. It is possible to assign many words a probability, thereby estimating a probability distribution over a set of words (Kuperberg, 2016; Frank & Willems, 2017; Aurnhammer & Frank, 2019). Importantly, probabilistic prediction does not have a dichotomous outcome (i.e., right or wrong). Instead, by distributing probability over many words, it is possible to hedge your bets. For example, imagine you hear the utterance *This is an excellent …*, words such as *idea, house, time* would be assigned a relative high probability, while words such as *rhinoceros, mistake, disease* a relative low probability, and ungrammatical continuations such as *by, over, the* an even lower probability, though not zero since people do not always produce perfect prose. Imagine that the next word was *mistake*, which was assigned a low probability. In this case the prediction was mostly wrong and a little bit right (it was still assigned some probability). In this manner, the cost of being wrong is less dire compared to the all or nothing interpretation of predication. Furthermore, being mostly wrong provides valuable information. If you heard *rhinoceros* while expecting the word *house* this tells you that you are in a different conversation than you thought. The prediction error affords an update of the model generating the prediction (Friston, 2005).

Another issue in the prediction debate concerns timing. Prediction is sometimes conceptualized as something that has to happen *just* before the actual input arrives (e.g., Lau et al., 2013; Brother et al. 2015; Aurnhammer & Frank, 2019), for example, by pre-activating a word before sensory evidence for that word can be perceived. I propose a less strict view of what constitutes prediction.

To clarify my position let us make an analogy with Morse code. Morse code consists of sequences built from three symbols: a dot, a dash, and a break. Concatenating the symbols results in a code that refers to letters. Morse developed this code in such a way as to minimize work for the telegraphists by assigning the shortest code to the most frequent letters (Gleick, 2011). This organizational scheme can be viewed as a static predictive model of a language's orthography. Importantly, nothing happens just before a telegraphist sends a Morse code message, but still the code ensures that letters are encoded with (close to) minimal effort by predicting which letters will be most frequent. No agency is required, and no action or activation is involved for a system to be predictive. Similarly, the ubiquitous word frequency effect could be viewed as a static predictive model of a language at the word level.

A static predictive system can be refined by making it dynamic, incorporating context into the prediction. Sentential context can be modelled in such a way that most probability is assigned to likely words. Such a system can generate a probability distribution over a set of words and it is this probability distribution that constitutes the prediction. In the current thesis, I use prediction to mean probabilistic prediction.

1.6 Methodology

In this thesis I use electroencephalography (EEG) and statistical language modelling (SLM) for multiple studies. Below, I will give a short introduction to these methods.

1.6.1 Statistical language modelling

When listening to speech or reading a text, a listener or reader encounters a sequence of words. The next word W_i in such a sequence does not occur randomly, but has a relation to the preceding words. The probability distribution over a large vocabulary for word W_i is therefore not uniform (i.e., identical probabilities for each word). Instead, there will often be a small set of more likely words and a long tail of less likely words.

Statistical language models (SLM) based on n-grams leverage this structure in natural language by counting sequences in corpora of language materials. These sequences, classically known as n-grams, may stand for any kind of sequence (e.g., characters, words, part of speech tags, etcetera). For example, a word trigram *the blue sky* consists of two word bigrams *the blue* and *blue sky* and three word unigrams *the, blue* and *sky*. By counting n-grams in large, sufficiently representative text corpora, it is possible to estimate the probability of *sky* given the bigram *the blue* or, more generally, any word following any context. The probability of a word given the preceding words is estimated by:

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

which denotes the conditional probability of word W_i given n preceding words. In this manner, context-sensitive word probabilities can be automatically derived from a large text corpus.

1.6.2 Electroencephalography

Electroencephalography (EEG) measures electrical potential differences on the scalp to investigate cognitive processes (see Luck, 2014 for an introduction). To measure these potential differences, a reference point is chosen, for example, the mastoids (a bony structure behind the ears), and electrodes are placed on the mastoids and predefined locations distributed across the scalp. The simultaneous firing of many neurons results in a measurable potential difference between the reference and scalp electrodes. The signal is very weak; it is measured in µ-volts (1 millionths of a volt) and is easily distorted by noise. Furthermore, since it travels through both the skull and the scalp, the electrical signal is also topographically distorted and therefore it is non-trivial to reconstruct the origin of potential differences measured at the scalp.

To counteract the influence of noise, one approach is to compute event-related potentials (ERP). Participants in an EEG-experiment are exposed to many stimuli of a specific type, for example, words. The precise presentation time of the stimuli is time-locked to the EEG-signal, and an epoch related to the stimulus is extracted from the EEG-signal. The ERP is created by averaging over all the extracted epochs. In the ERP, components can be identified representing the exogenous brain activity, i.e., the activity evoked by the stimulus. In this manner many different ERP components have been found and described in the literature (e.g., Luck, 2014).

In this thesis, the ERP paradigm will be used to investigate predictive language processing. I will focus on two ERP components: the phonological mismatch negativity (PMN, also referred to as N200) and the N400. The PMN is an early anterior (near the forehead) negativity peaking around 200 milliseconds after word onset, and is sensitive to expectations about auditory word forms (e.g., Connolly et al., 1994). The N400 is a negative component peaking around 400 milliseconds after word onset mostly at the central parietal electrodes (the middle back of the head). The N400 can be elicited by presenting words in context to participants (see Kutas & Federmeier, 2011, for an overview).

1.7 Outline of the thesis

The objective of this thesis is to study predictive language processing during the perception of natural speech sampled from different registers.

In Chapter 2, I describe the dataset of EEG-recordings I collected from participants listening to long stretches (4 – 15 minutes) of natural speech from different registers, consisting of approximately 200 hours of EEG data. Participants came to the lab on three separate occasions to listen to approximately 90 minutes of speech from three distinct speech registers (~ 270 minutes in total). The registers – dialogues, news broadcasts and read-aloud stories – were chosen based on the results from Chapter 5 (the corresponding study was conducted before the EEG recordings). To preprocess the 200 hours of EEG data, I developed a novel approach by training and applying a convolutional neural network to detect artefacts in the EEG signal. The EEG dataset will be used in the subsequent studies and published as an open access corpus, named the Dutch EEG speech register corpus (DESRC).

In Chapter 3, I develop a novel continuous mismatch measure that captures the mismatch between high-level expectations (word probabilities estimated with an SLM) and low-level perceptual input (the actual speech segment). This mismatch measure can help to improve the study of listeners' word form anticipations. I introduce the concept of a word probability distribution (WPD), which is a lexicon of words with assigned probabilities. One WPD is purely based on preceding context, whereby the word probabilities are estimated with an SLM. The second WPD is an updated version of the first WPD. The update is based on the initial auditory segment of a word. The mismatch measure is formulated in terms of the cross-entropy between these two WPDs, i.e., the WPD before and after the auditory update. The measure is validated with several tests, for example, whether the update with auditory input increases the probability of the correct word.

In Chapter 4, I use the mismatch measure developed in Chapter 3 on the EEG-data (described in Chapter 2) to investigate whether listeners are sensitive to the mismatch between expected and actual word form inputs. The cleaned EEG dataset contains over a million word epochs for the analysis. I expect to find a PMN effect, whereby a higher value for the mismatch measure results in a more negative PMN amplitude.

In Chapter 5, I investigate word predictability differences between speech registers. The output of register-specific SLMs is used to train a register classifier. With this classifier, I test whether it is possible to distinguish between speech registers based on word predictability (as estimated with SLMs). Furthermore, I study the influence of potential confounds such as sentence length and topic.

In Chapter 6, I test whether listeners adapt their expectations based on the speech register they are listening to. SLMs are trained on different register-specific language materials to estimate word surprisal that reflects different contexts: a generic context and a register-specific context. With these SLMs, I estimate different variants of word surprisal (based on generic or register-specific contexts) for each content word in the corpus and compare how well they predict the N400 amplitude.

In Chapter 7, I discuss the findings presented in the preceding chapters and relate them to the existing literature. Furthermore, I discuss possibilities for future research related to predictive language processing.

The Dutch EEG Speech Register Corpus

Chapter 2

Abstract

The Dutch EEG Speech Style Corpus contains 207 hours of EEG recordings from 48 participants listening to natural connected speech. The speech materials were sampled from: spontaneous dialogues, news broadcasts and read-aloud stories, and contain 50,277 word tokens, time-locked to the EEG data. We cleaned the EEG data by labelling and removing artefacts with the aid of an automatic artefact classifier. Eye-related activity was removed with independent component analysis. The EEG data (raw and cleaned), containing 1.5 million word epochs, is freely available (license: CC BY 4.0) and offers new research opportunities to investigate neural correlates of speech processing.

2.1 Introduction

This article presents the Dutch EEG Speech Register Corpus (henceforth DESRC), which consists of 207 hours of recorded electroencephalography (EEG) from 48 participants listening to Dutch speech sampled from three different registers: spontaneous dialogues, news broadcasts, and read-aloud stories. In each of three sessions, participants listened to 90 minutes of each register, approximately 270 minutes of EEG were recorded.

We recorded EEG data for the corpus with a paradigm different from classic ERP experiments (see Luck 2014 for an overview). Participants listened to long (4 – 15 minutes) continuous stretches of natural speech, while the EEG signal was recorded. We time-locked the EEG materials to *all* words in the speech materials. The use of long continuous stretches of natural speech is similar to, for example, the approach taken in Willems et al. (2016), who conducted an fMRI study with participants listening to excerpts from audio books. We will refer to this new approach as the *naturalistic sample approach*.

The naturalistic sample approach is based on three ideas. First, it is important to also conduct experiments on naturalistic stimuli to improve the ecological validity of experimental results (see also Willems, 2015). Second, by increasing the number of stimuli (i.e., by considering all words in the language materials, rather than a subset of target words), it is possible to relax strict control over the stimuli. The nuisance effects of, for example, stimuli surface forms (see Luck, 2014) will average out over the large number of stimuli (i.e., hundreds versus hundreds of thousands of stimuli). Third, the large amount of data allows us to replace categorical predictors with continuous predictors and forego the need of artificially binning linguistic materials. For example, Frank et al. (2015) recorded EEG during a forced-paced reading task with sentences sampled from novels. They used a continuous predictor of word surprisal to predict the amplitude of the N400, while in classical N400 experiments words are typically grouped categorically into congruent and incongruent sets. The use of continuous predictors fits well with the graded effects observed with, for instance, the N400 (e.g., Federmeier & Kutas, 1999).

The naturalistic sample approach requires a large amount of data to be collected. For EEG-recordings, this results in a non-trivial amount of work concerning the preprocessing of the data. To remove artefacts from 207 hours of EEG-data, we used a novel approach by training and employing a convolutional neural network to detect artefacts (see Section 2.3).

We enriched the orthographically transcribed speech materials used in the EEG experiment with part-of-speech (POS) tags, word frequency and several information theoretic measures such as word surprisal, entropy and cross-entropy (see Bentum et al. 2019). The resulting dataset can be used to answer many different research questions. This paper provides the accompanying description of the DESRC. In the following sections, we explain the speech material selection, the EEG recording procedure, and the steps taken to clean the EEG data.

2.2 Corpus

The DESRC contains EEG materials recorded from participants listening to speech sampled from different registers. In general, the communicative situation influences the register adopted by a speaker and research shows that there are important differences between registers (see Biber and Conrad, 2009 for an overview). For example, people chatting socially use a different vocabulary compared to a person giving a formal address. Bentum et al. (2019a) found that word surprisal as estimated by a statistical language model depends both on the preceding words and speech register. Furthermore, the communicative situation influences pronunciation; formal occasions trigger more careful pronunciation than informal occasions (Ernestus et al. 2015). The differences between speech registers could influence a listener's speech processing. To capture register variation, we sampled from multiple registers and used the findings reported by Bentum et al. (2019a) to select three distinct speech registers.

2.2.1 Materials

The speech materials were sampled different corpora; the news broadcasts and read-aloud stories were taken from the Netherlandic Dutch part of the Spoken Dutch Corpus (Oostdijk, 2001). The spontaneous dialogues were taken from the IFADV corpus (Van Son et al., 2008). Both corpora provide manual orthographic and automatically obtained phonemic annotations and segmentations, which allowed us to align the speech input with the EEG recordings. Short excerpts from each speech style are part of the supplementary materials of this article. Table 1 lists general statistics for the speech materials used for the EEG recording sessions.

The spontaneous dialogue materials consist of six 15-minute dialogues between well acquainted dyads (e.g., friends, colleagues). They freely talked about any topic that came to mind. One of the 11 speakers is present in two dialogues. The news broadcast materials consist of radio news broadcasts from the late nineties and early 2000s, which were grouped into seven blocks of 12-minutes, with each block further subdivided into sections of four minutes. The read-aloud stories materials consist of seven 12-minute-long excerpts from read-aloud Dutch audio books.

Table 1. Overview of the materials per speech style: 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
spontaneous dialogues	21,718 (2,435)	206 ms	11 (2)	20 - 62
news broadcasts	15,350 (3,526)	289 ms	8 (7)	23 - 46
read-aloud stories	13,209 (2,349)	256 ms	7 (3)	38 - 75
total	50,277 (5,866)	245 ms	26 (13)	20 - 75

2.2.2 EEG Participants

Forty-eight neurologically unimpaired right-handed native speakers of Dutch (18-29 years), 34 women and 14 men, participated in all three sessions of EEG recordings. All participants gave informed consent to participation and the public release of the recorded EEG-signal. Participants were paid 80 euros for their participation.

2.2.3 EEG Procedure

The participants came to the lab on three separate occasions; which were separated by at least one week. They were fitted with the correct size electrode cap and seated in a sound-attenuating booth. Participants listened to 90 minutes of speech from one register (see Table 1). The order of the registers was counter-balanced across participants. The audio materials were presented via in-earphones (Etymōtic ER1) at a comfortable listening volume; a short audio sample was used to set the volume. Participants were asked to sit still and keep eye-movement and blinks to a minimum.

The audio materials were presented in blocks of approximately 15 minutes and the order of blocks was counter-balanced across participants. During pauses in the audio materials (every four minutes for the news broadcasts and approximately every 15 minutes for the other registers), yes-no comprehension questions were visually presented to encourage attentive listening. Participants responded with a button box. At the end of each experimental block, the participant could take a break before resuming the experiment.

2.2.4 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 artefacts (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). Two additional electrodes were placed on the left and right mastoid. All electrodes were referenced to the left mastoid electrode and all electrode impedances were below 15 kΩ before recording started. The EEG signal 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.

2.2.5 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 artefacts from the data semi-automatically by using a deep neural network (see Section 2.3). Subsequently, we used independent component analysis (ICA) to filter out activity related to eye-movement and blinks. Following Winkler et al. (2015), we computed the ICA on 1-30 Hz bandpass filtered data (after removing the artefacts). We visually determined EOG related ICA components based on the topography and the correlation with the EOG channels. We recomposed the 0.05-30 Hz bandpass filtered data without these components. The original recordings, the artefact annotations and the ICA-decompositions are all available in the online dataset (see Table 2 for an overview of the EEG materials).

In the following section we describe how we trained and applied an EEG artefact classifier to automatically detect and remove artefact sections and channels to clean the EEG data.

Table 2. Overview of EEG materials, (before) and after artefact removal. Word epochs are defined as EEG materials from 300 milliseconds before to 1000 milliseconds after word onset

speech registers	hours	word epochs	content word epochs
spontaneous dialogues	51 (70)	701,335 (1,020,305)	368,407 (537,470)
read-aloud stories	47 (66)	438,086 (631,970)	229,807 (332,322)
news broadcasts	44 (71)	371,603 (731,649)	204,269 (401,130)
total	142 (207)	1,511,024 (2,383,924)	802,483 (1,270,922)

2.3 Automatic EEG artefact detection with a convolution neural network

Artefact removal from EEG data is a time-consuming process. Several neuroimaging packages, such as EEGLAB (Delorme et al., 2004), MNE (Gramfort et al., 2014) and FIELDTRIP (Oostenveld et al., 2011), provide statistical means to aid artefact detection. Statistical artefact rejection is also described in Nolan et al. (2010). These methods use various measures to describe the data (e.g., amplitude, amplitude range, variance, correlation between channels), which are typically transformed to z-scores. The measures are thresholded at a conservative value (e.g., $|z| > 3$) to find data that contain artefacts.

Unfortunately, the use of statistics on simple measures (e.g., amplitude range) for artefact removal has serious limitations. The z-score is typically calculated separately for each participant, which results in a different rejection criterion per participant. Furthermore, any z-score thresholding only rejects outliers, but is not informative about the quality of the rejected data. For example, if a dataset is noisy it will only remove extremely noisy subsets and keep potentially corrupted data, while if the dataset is clean, it will remove potentially usable data.

Instead of using threshold statistics to detect artefacts, we trained a convolutional neural network (CNN) to distinguish between clean and artefact EEG data. The classifier is trained to discover features that distinguish between clean and artefact data without relying on statistics of simple measures (e.g., amplitude, channel correlation) that only imperfectly capture that distinction. In the following subsections, we describe how we trained and tested this CNN for artefact detection.

2.3.1 Manual annotation

We manually annotated approximately 60 hours of EEG data, by marking artefacts by their start and end boundaries. We divided the artefacts in two types: *stretch* and *channel* artefacts. Stretch artefacts are visible on all or most EEG channels during a stretch of time. The artefacts can be due to muscle activity, a sweaty scalp, etcetera. Channel artefacts occur on individual channels, due to poor connection with the scalp, technical problems (e.g., faulty electrode), etcetera. The solutions for these artefact types differ. If all or most channels show artefacts (i.e., stretch artefacts), it is best to remove a complete section of EEG data (i.e., all channels). If a specific channel shows artefacts over an extended period of time (i.e., channel artefacts), that single channel should be removed from the dataset.

2.3.2 Training, test and validation materials

The EEG data was first downsampled from 1000 to 100 Hz for training and classification purposes. We created separate datasets for the stretch and channel artefacts and performed the following steps for each. We windowed the EEG data into 1-second windows (i.e., 100 samples) with 99% overlap (i.e., at every sample a window was started). We labelled each window as *artefact* when half or more of the samples overlapped with the manually annotated artefacts. All other windows were labelled *clean*. After this labeling procedure, we assigned each window randomly to one of a 100 sets. Ninety sets were used for training and ten sets were held out for validation and testing.

For the *stretch* dataset we selected 25 channels, excluding the Fp2 channel due to overall poor signal quality. Each window consisted of a matrix of 25 channels by 100 samples and had a label: *clean* or *artefact.* For the *channel* dataset we created a separate window for each channel. Every window consisted of a matrix of 32 channels by 100 samples. We created this matrix by copying the target channel to the following rows 1, 7, 13, 19, 25, 31 of the window matrix. All other rows were filled by the 25 channels in fixed order. This approach was used to fix the order of the channels while ‘marking’ the target channel as target.

We normalized the EEG signal within each window to a value between 0 and 1. Before normalizing the stretch artefact windows, we set the threshold value to ± 100 µV, and for the channel artefacts we used ± 300 µV (i.e., all larger values were set to these threshold values). Lastly, we multiplied the resulting windows with a Hamming window.

2.3.3 Model specification

We specified the CNN in Tensorflow (Abadi et al., 2016) and started with a standard CNN model architecture inspired by its use in image classification (e.g., Krizhevsky et al., 2012). The typical CNN architecture for image classification specifies multiple convolutional layers of *n* by *n* (e.g., *n* = 5) kernels. For EEG data this appears to be suboptimal, arguably because the time and channel dimensions have a different impact and statistical behavior compared to the height and width dimensions of images. We therefore adapted the model according to Schirrmeister et al. (2017), who reported good results with EEG-data classification where the first two convolutional layers of their model separately specify the time and channel dimensions, respectively. We found that this separation approach also strongly improved the performance of our classifier.

We defined a separate stretch and a channel model. The structure of these models is presented in Table 3. The first layer (1 by 25 kernel) is exposed to 25 samples (i.e., from 25 consecutive time points) from one EEG channel. The second layer, a 6 by 1 kernel, steps through the window exposed to six EEG-channels at each time point. Subsequently, the output is pooled and followed by a kernel of 5 by 5 (6 by 6 for the channel model) and a second round of pooling, followed by a fully connected layer, which is mapped to an output class vector of length 2 (i.e., clean or artefact).

Table 3. Overview of convolutional neural network architecture for the section and channel models. (Values that are different for the channel model are between brackets). Conv. stands for convolutional layer, Relu for rectified linear unit, ch for channel.

Layer	Type	In ch.	Out ch.	kernel size	feature map (channel model)	Stride	Activation
1	conv.	1	32	1 × 25	25 (32) × 100 × 32	1	ReLu
2	conv.	32	64	6 × 1	25 (32) × 100 × 64	1	ReLu
3	pool	64	64	2 × 2	13 (16) × 50 × 64	2	
4	conv.	64	128	5 (6) × 5 (6)	13 (16) × 50 × 128	1	ReLu
5	pool	128	128	2 × 2	7 (8) × 25 × 128	2	
6	linear	128	1		2400		ReLu
7	softmax				2		

2.3.4 Training, classification and manual correction

We trained both models with stochastic gradient descent. Each training cycle, a model was exposed to 200 windows drawn randomly from a specific training set. To ensure an equal number of clean and artefact windows, we downsampled in favor of the artefact windows to a 50/50 ratio (approximately 7% of windows contain artefacts in the original data). We repeated training cycles until the classifier performance plateaued on the validation set.

The stretch and channel models were used to classify the complete set of EEG-materials. Subsequently, we transformed the windows classified as artefacts to start and end boundaries in the EEG signal. If a start boundary followed an end boundary within two seconds of a previous annotation, we combined the two artefacts. The resulting artefact boundaries were corrected based on visual inspection. We skipped sections of materials (40 seconds or longer) considered clean by the automatic classifiers because of time constraints. Artefacts tend to cluster, and therefore, long clean stretches are less likely to contain artefacts. We trained the classifiers to have a high recall of artefacts (by class downsampling in favor of the artefacts), to reduce the chance of missing artefacts. Nevertheless, since we did not check all materials it is possible that some artefacts remained unidentified.

After manual correction, we marked channels as bad (i.e., to be removed from the data for subsequent processing) if the data from a channel contained artefacts for more than 40% of an experimental block (see Section 2.2.3), otherwise channel artefacts were relabeled as stretch artefact. These manually corrected annotations were used to remove EEG data contaminated with artefacts (see Section 2.2.5).

2.3.5 Classifier validation

We analyzed the quality of the CNN classifier by comparing the resulting artefact annotations with manually corrected annotations and a simple threshold approach which we detail below. As a unit of validation of the classification procedure, we chose the *word epoch.* Word epochs were defined as EEG materials from 300 milliseconds before word onset to 1000 milliseconds after word onset. We extracted all word epochs from the EEG materials and labelled each as clean or artefact based on different annotation sets. As ground truth, we used the manually corrected automatic annotations (see Section 2.3.4). We compared the labelling based on the automatic CNN annotations with a labelling based on thresholding, a procedure whereby word epochs were considered clean if the maximum value of the word epoch EEG materials was between ± 75 μV.

We computed the precision, recall and F1-scores for the threshold and automatic CNN labeling of word epochs. The automatic classification based on the CNN classifier outperformed the threshold approach (see Table 4) with an F1-score of 0.89 compared to 0.73.

The validation results show that there is a clear trade-off between time spent cleaning the EEG materials and quality or amount of the EEG materials. The threshold approach is very fast, because no prior labelling of EEG data is required. However, this time gain comes at the cost of missing 28% of the usable data and 27% of the artefacts. The uncorrected output of the CNN classifier performed better (missing only 10% and 13% respectively), however, this came at the cost of approximately 300 hours of annotation work. Manually correcting the classifier output further improves the quality of the EEG materials; however, this entailed another 240 hours of work.

Table 4. Overview of word epoch labelling performance for different classification strategies.

	threshold			CNN		
	precision	recall	f1-score	Precision	recall	f1-score
artefact	0.60	0.73	0.66	0.83	0.87	0.85
clean	0.82	0.72	0.77	0.92	0.90	0.91
average	0.74	0.72	0.73	0.89	0.89	0.89

2.4 Conclusion

The EEG speech style corpus (DESRC) provides a large database of EEG recordings from participants listening to long (4 – 15 minutes) stretches of natural speech. The DESRC will be made available under license CC BY 4.0 and provides a rich set of meta-data; complete orthographic and phonemic transcriptions time-locked to the EEG data. We enriched the transcriptions with part-of-speech (POS) tags, word frequency and information theoretic measures such as word surprisal, entropy and cross-entropy. Furthermore, we annotated the EEG data for artefacts to allow easy exclusion of data contaminated with artefacts.

Quantifying expectation modulation in human speech processing

Chapter 3

This chapter is a reformatted version of:

Martijn Bentum, Louis ten Bosch, Antal van den Bosch & Mirjam Ernestus (2019). Quantifying expectation modulation in human speech processing. *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019

Abstract

The mismatch between top-down predicted and bottom-up perceptual input is an important mechanism of perception according to the predictive coding framework (Friston, 2005). In this paper we develop and validate a new information-theoretic measure that quantifies the mismatch between expected and observed auditory input during speech processing. We argue that such a mismatch measure is useful for the study of speech processing. To compute the mismatch measure, we use naturalistic speech materials containing approximately 50,000 word tokens. For each word token we first estimate the prior word probability distribution with the aid of statistical language modelling, and next use automatic speech recognition to update this word probability distribution based on the unfolding speech signal. We validate the mismatch measure with multiple analyses, and show that the auditory-based update improves the probability of the correct word and lowers the uncertainty of the word probability distribution. Based on these results, we argue that it is possible to explicitly estimate the mismatch between predicted and perceived speech input with the cross-entropy between word expectations computed before and after an auditory update.

3.1 Introduction

Listeners are able to extract words from speech input in a wide range of (adverse) listening conditions. The difficulty of this task is attested by the many decades of research aimed at creating artificial systems with similar performance. The details of the cognitive processes underlying human speech processing are still contentious. A long-standing debate revolves around the importance and timing of top-down versus bottom-up influence for word recognition during speech comprehension (Magnuson et al., 2018; Norris et al., 2018). Certain autonomous models, for example Shortlist A & B (Norris, 1994; Norris & McQueen, 2008) claim that early speech processing is exclusively bottom-up, and top-down influence is only exerted at the lexical phase of word recognition. Other interactive models, for example Trace (McClelland & Elman, 1986) allow for a certain degree of top-down influence, congruent with the predictive coding framework (Friston, 2005).

The predictive coding framework (Friston, 2005) assumes that perception entails anticipation based on a generative model, whereby cognitively higher levels generate predictions about upcoming (low-level) perceptual input. The mismatch between the prediction and the actual input provides an error signal, which informs to what extent the hypotheses generated by the generative model need to be adapted. If we think about human speech processing in this framework, we need a model to assign a probability to upcoming words, given the preceding words, and a mechanism to quantify the mismatch between bottom-up observations and top-down expectations. The first part, the probability of upcoming words, can be estimated according to Equation 1, which lies at the basis of a statistical language model (SLM), whereby P denotes the conditional probability of word Wi given a sequence of n preceding words:

$$\hat{P}(W_i|context) = P(W_i|W_{i-n}, \dots, W_{i-1}) \qquad (1)$$

Several studies (e.g., Smith & Levy, 2013; Frank et al., 2015; Willems et al., 2016) have successfully used statistical language modelling to study human language processing. They employed an SLM to compute word probabilities from a text corpus and show that listeners and readers are indeed sensitive to the probability of a word given the preceding words. These results suggest that listeners anticipate likely upcoming words. The predictive coding framework makes an additional prediction, namely that human listeners generate low-level auditory expectations based on the anticipated words.

This paper addresses the estimation of the mismatch (i.e., the error signal) between the expected word form and the observed word form as it comes in as speech input. To estimate this error signal, we make use of the concept of a word probability distribution (WPD), consisting of a list of words, whereby each word is assigned a probability. We compute two types of WPD. The *prior* WPD is based on the top-down expectation at word onset without any auditory input. In this WPD, each word is assigned a probability given the preceding words as estimated by an SLM. A *post* WPD is based on the prior WPD in combination with the bottom-up acoustic evidence received so far: i.e., the word probabilities are updated according to the unfolding auditory information.

We analyze the auditory input with statistical paradigms developed in the automatic speech recognition (ASR) domain, to generate a probability distribution on a large set of phone sequences that could all potentially match a possible word start. These phone sequence probabilities are used to update word probabilities matching these phone sequences, resulting in a post WPD. The error signal can then be defined as the cross-entropy between the prior and post WPD, which captures the mismatch between the high-level expectation (word probabilities) and the sensory input (a spoken word). The cross-entropy between prior and post WPD can be computed with Equation 2, whereby H denotes cross-entropy, p the prior WPD, q the post WPD and X the WPD word list.

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x) \qquad (2)$$

To summarize, the prior WPD captures high-level expectations (based on preceding words). The post WPD differs only from the prior WPD in the added auditory information. We therefore propose that the cross-entropy between prior and post WPD quantifies the mismatch between high-level expectations and auditory input.

To validate the computation of the mismatch measure, we test if the auditory update improves the post WPD in relation to the prior WPD. We expect the auditory update to decrease the entropy of the post WPD and increase the probability of the correct word. In addition, we test whether these measures improve with more auditory input. Since our goal is to compute a mismatch measure which is relevant for human speech processing, we also test the optimal amount of auditory materials for cross-entropy computation. In the following sections, we will describe the language materials and methods used to compute both the prior and post WPDs and the subsequent analyses. After these sections, results are presented, followed by a discussion and a future outlook.

3.2 Method

3.2.1 Materials

We used materials from three corpora, namely, the Spoken Dutch Corpus (Oostdijk, 2001), IFADV (Van Son et al., 2008) and NLCOW14, henceforth COW (Schäfer, 2015; Schäfer & Bildhauer, 2012). The first two corpora consist of audio recordings and transcriptions of spoken Dutch materials. The COW corpus consists of 4,7 billion words of web-crawled Dutch text.

We pre-processed the COW corpus by excluding all non-Dutch sentences, removing sentences with three or more repeating words or characters, or characters that are not used in standard Dutch orthography. We replaced characters with diacritics to the equivalent characters without diacritics. Furthermore, we mapped all numbers, websites and tagged words (e.g., @tag@) to special word codes. We removed all punctuation, except for commas. We normalized all apostrophe words to a standard spelling (e.g., *'t* becomes *het*, 'the'). The Spoken Dutch Corpus and IFADV were already appropriately tokenized (see Goedertier et al., 2000); we only applied the apostrophe normalization and diacritic removal to these texts.

For our experiments we extracted a subset of the Spoken Dutch Corpus and the IFADV containing 50,277 word tokens (see Table 1). This subset, henceforth called Speech Corpus, consists of annotated speech from different speech registers (i.e., spontaneous dialogues, news broadcasts, and read aloud stories). The selection criteria for our materials were based on a different experiment. The differences in speech styles reflected in our materials will not be important in the current study.

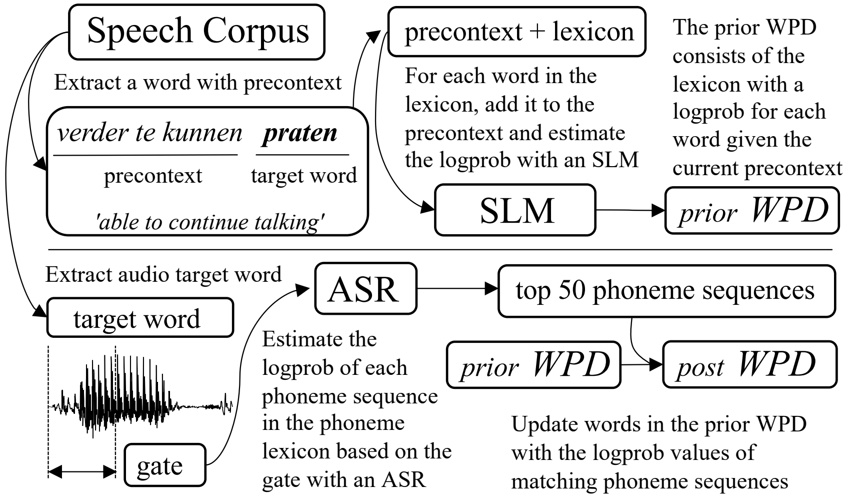
Table 1: Overview of the materials in the Speech Corpus.

speech style	word tokens (word types)	average word duration (milliseconds)
spontaneous dialogues	21,718 (2,435)	206
read-aloud stories	13,209 (2,349)	256
news broadcast	15,350 (3,526)	289
total	50,277 (5,866)	245

3.2.2 Procedure

For each word in the Speech Corpus we created two types of word probability distributions (WPD), one prior and one post auditory information integration (see Figure 1). We will explain how we created these WPDs for a given word (henceforth ‘target word’) in the Speech Corpus. To create the prior WPD, we used an SLM and a lexicon (i.e., the set of words in the WPD). We trained a 4th order Markov SLM on the Dutch COW corpus by using SRILM (Stolcke, 2002) with Kneser-Ney discounting for smoothing (Chen & Goodman, 1999). For the lexicon we selected approximately 200,000 Dutch phonemically transcribed words that are in the top .9 cumulative probability of the word unigrams of the SLM. We estimated the probability of each word in this lexicon based on the words preceding the target word in the Speech Corpus.

Figure 1: Diagram of prior and post WPD construction.



Speech Corpus
Extract a word with precontext
verder te kunnen praten
precontext
target word
'able to continue talking'
precontext + lexicon
For each word in the lexicon, add it to the precontext and estimate the logprob with an SLM
SLM
The prior WPD consists of the lexicon with a logprob for each word given the current precontext
prior WPD
Extract audio target word
target word
gate
ASR
Estimate the logprob of each phoneme sequence in the phoneme lexicon based on the gate with an ASR
top 50 phoneme sequences
prior WPD
post WPD
Update words in the prior WPD with the logprob values of matching phoneme sequences

Post WPD construction was done in multiple steps. In the first step, we used the forced aligned phonemic transcriptions (present for all materials in the Speech Corpus) to determine the word onset of the target word in the audio materials and defined 28 gates of different durations (110, 130, …, 650 ms), starting from word onset. In step 2, each gate was used to create a post WPD, resulting in 28 post WPDs per target word. Figure 1, bottom part, shows post WPD construction for one gate.

To create the post WPDs, we used KALDI (Povey et al., 2011) to estimate a phonemic probability distribution for the gated speech input. We did this by first creating a ‘Phoneme Lexicon' consisting of all lexically licensed phoneme sequences up to length 8, approximately 400,000 entries. For example, the word *universiteit* ‘university’ with phonemic representation /y n i v ɛ r s i t ɛi t/ yields the eight cohort forms /y/, /y n/, ..., /y n i v ɛ r s i/ to be included into the Phoneme Lexicon. This Phoneme Lexicon, in combination with a flat language model (i.e., each phoneme sequence has an equal prior probability), was used in the KALDI decoding of the gated speech chunks. For each gate, this decoding leads to a weighted phone lattice. The 500 best paths through this lattice were chosen as a decoding result. This step resulted in scaled logprob scores for each of the 500 phoneme sequences.

The scaled logprob scores were 'descaled' to a genuine probability distribution. The descaling factor determines the decay of the phoneme string probabilities (i.e., the probability difference between the winning hypothesis, runner-up, etc.). This descaling factor was estimated by investigating the entropy of the phonemic probability distribution for different gate durations. We assume that the entropy of the phoneme probability distribution should decrease for increasing gate lengths, because more acoustic material should yield a better identification and thereby a sharper distribution of the phoneme sequence probabilities. We therefore chose the factor which resulted in the highest entropy decrease across gates to descale the logprobs.

After descaling the logprobs, we inspected the phoneme *n*-best lists for multiple words from the Speech Corpus to determine a useful value of *n*. The top-50 appeared to be a sufficient threshold to exclude implausible phoneme sequence strings.

The logprobs of the top-50 phoneme strings were used to update the prior WPD to the post WPD. However, directly adding logprob values has (for our purposes) an unfortunate effect of generating the biggest difference in unlikely candidates. Since we truncated our *n*-best phoneme sequence set, this would result in a bad update. We therefore *shifted* the logprobs by adding the absolute value of the logprob of phoneme sequence 51 (from the *n*-best list) to the top-50 phoneme sequences. The most likely phoneme sequence now causes the biggest shift in the post WPD and normalization of this distribution ensures that unlikely words are shifted downwards appropriately.

To perform the auditory update, we matched each of the top-50 candidate phoneme sequences to all words in the lexicon (i.e., the set of words in the WPD). For example, the word *kat* ‘cat’*,* represented in the Dutch lexicon as ‘kat, k ɑ̲ t’ would match with the phoneme sequences /k/, /k ɑ̲/, /k ɑ̲ t/ and mismatch with /ɑ̲/, /ɑ̲ t/ or /k ɑ̲ t s/. We computed the word probabilities of the post WPD by adding the shifted logprob values of phoneme sequences to the logprob values of matching words in prior WPD.

3.2.3 Analysis

We performed two analyses to validate our approach and one to investigate the amount of auditory materials needed for the best cross-entropy computation. For Analysis 1, we tested whether the auditory update from prior to post WPD lowered the surprisal of the correct word, which tests whether the auditory update assigns more probability to the correct word. Furthermore, we test whether the entropy of the post WPD was lower compared to the prior WPD, indicating that there is less uncertainty in the post WPD, which is expected if the auditory update functions correctly.

To make the comparison between prior and post WPD, we conducted two tests to check that both surprisal and entropy decrease after the auditory update. The first test was a conservative test that compares surprisal of the correct word and entropy of the prior WPD to the highest (i.e., worst) surprisal and entropy values of the set of 28 post WPDs for a given word. The less conservative test compared the surprisal of the correct word and entropy of the prior WPD to the mean of the surprisal and entropy over the same set of post WPDs. In both cases (conservative and less conservative), the post WPD surprisal and entropy values are compared with the corresponding prior WPD values.

For Analysis 2, we tested whether the surprisal value of the correct word and the entropy of the post WPDs decreased with increasing gate duration. We tested this by first computing the difference in surprisal of the correct word between prior and post WPD for each gate. Longer gates should improve the post WPD more, because a longer gate provides more information about the upcoming word. Of course, this only holds if the gate is shorter than the word, because otherwise information of following words is also incorporated in the auditory update. We therefore excluded all cases where the word was shorter than the gate.

Finally, Analysis 3 investigated which gate should be used for the cross-entropy computation. We want to use the cross-entropy to predict human speech processing cost and therefore we tested which gate duration performs the best update for all words (including words shorter than a given gate). This analysis reflects the situation for a human listener, who does not know the duration of upcoming words. For this analysis, we computed the difference in surprisal for the correct word between prior and post WPD for all words and gate durations.

3.3 Results

We used R (R Core Team, 2015) for all analyses. For Analysis 1, we compared the surprisal of the correct word between the prior and post WPD with a simple linear regression model. The regression model was fitted on 80% of the data and tested on 20% unseen data. Based on the results of the unseen data, we computed the R^2_{CV}

(cross validated). Similar R^2 and R^2_{CV} values indicate that the model generalizes well to unseen data and was not overfitted to the current sample. We created separate models for the conservative (i.e., worst) and less conservative (average) test, as detailed in Section 3.2.2. We used the same approach to compare the entropy between prior and post WPD. As expected, both surprisal and entropy decrease (i.e., have negative betas) after the auditory update, as can be seen in Table 2. This appears both from the conservative and less conservative test.

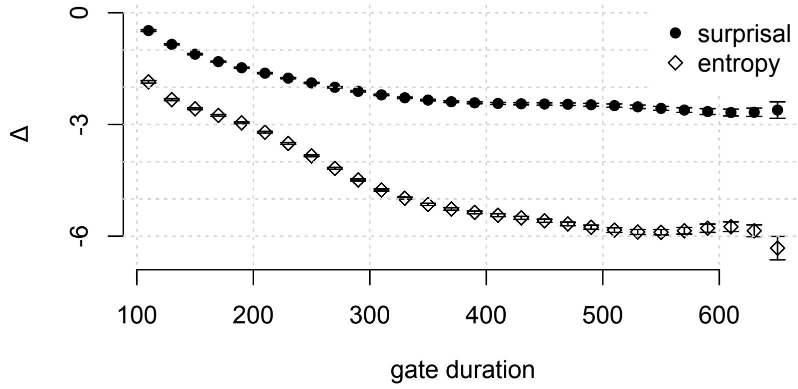
Table 2: Simple linear regression models for surprisal and entropy comparison between prior and post WPD.

	R2 (R2CV)	B	SE B	P
worst surprisal	0.02 (0.02)			< .001
update*		-0.49	0.01	< .001
avg.† surprisal	0.64 (0.64)			< .001
update*		-2.80	0.01	< .001
worst entropy	0.14 (0.14)			< .001
update*		-1.85	0.02	< .001
avg.† entropy	0.80 (0.80)			< .001
update*		-6.30	0.01	< .001

*Difference between prior and post WPD, †average

For Analysis 2 we tested whether the surprisal of the correct word of the post WPD improves with increasing gate length. We fitted a linear regression model on the difference in surprisal between prior and post WPD for each gate length. We modelled the relationship between surprisal difference and gate length with a 7th order polynomial, to capture possible non-linear relationships and established the order of the polynomial with model comparison by selecting the highest uneven order that still improved the model. We used the same approach to test entropy difference in relation to gate length; for this model we used an 11th order polynomial on gate duration. Both the surprisal and the entropy model were fitted on 80% of the data. Again, we used the remaining 20% unseen data to compute the R^2_{cv}, to test whether the model generalizes well.

Figure 2: Predicted difference in surprisal and entropy as a function of gate duration, with 99% confidence intervals.



We do not report the betas for all polynomials in the surprisal and entropy model, because they are hard to interpret. Instead, we visualized the results of both models in Figure 2. The surprisal of the correct word shows a clear negative trend with increasing gate length (R^2 = 0.13, R^2_{CV} = 0.14, p < .001). Similarly, the entropy of the post WPD also shows a clear negative trend with increasing gate length (R^2 = 0.22, R^2_{CV} = 0.22, p < .001). The negative trend for surprisal means that with increasing gate length the probability of the correct word increases (if only words longer than the gate duration are considered). The negative trend for entropy means that the amount of uncertainty in the post WPD keeps decreasing when more relevant acoustic information becomes available.

Finally, we investigated which gate duration most improved the surprisal of the correct word for all words. We fitted a linear regression model on 80% of the data to predict the difference in surprisal by gate duration with a 7th order polynomial, R^2 = 0.13, R^2_{CV} = 0.13, p < .001. In Table 3 we report the top 3 gate durations that most improved the surprisal of the correct word after the auditory update. In addition, we fitted the same regression model on a randomly selected subset (10% of the data) a 1000 times. For each model we ranked the predicted surprisal difference, with rank 1 for best performance. Table 3 shows that the auditory update of 190 milliseconds resulted in the largest reduction in the surprisal of the correct word.

Table 3: Predicted surprisal difference and number of times a gate duration (milliseconds) showed best improvement in surprisal between prior and post WPD.

gate	predicted	99% CI	# rank 1
170	-0.978	-1.004, -0.952	285
190	-0.982	-1.006, -0.958	715
210	-0.944	-0.968, 0.920	0

3.4 Discussion

The goal of this study was to quantify a mismatch measure between high-level expectations and low-level input in speech perception. We created two types of word probability distributions (WPD), one prior and one post auditory update. The prior WPD is completely based on preceding words and represents the high-level expectations. The post WPD is an update of the prior WPD integrating auditory information. We hypothesized that the difference between prior and post WPD captures the mismatch between expectations and speech input and could be quantified by cross-entropy.

To validate the mismatch measure, we investigated whether the auditory update performed as expected. In Analysis 1, we showed a decrease in both the surprisal of the correct word and the entropy of the post WPD, in line with our expectations. Furthermore, we showed in Analysis 2 that the surprisal and entropy further decrease with increasing gate length (only considering words that are longer than the gate duration). This was also expected; longer gate durations provide more information for the auditory update and should therefore improve update results.

The results show that the difference between the prior and post WPD reflects auditory information, which improved the probability of the correct word and lowered the uncertainty (entropy) of the post WPD. Prior and post WPD differ in word probabilities based on the extra information that the auditory input provides. We therefore argue that the cross-entropy between both distributions captures the mismatch between expected and observed auditory input.

After validating our results, we investigated the amount of auditory materials needed to compute the mismatch measure in Analysis 3. For this analysis we included all words, because this more closely resembles the situation of a human listener (who does not know how long the next word will be). We compared surprisal improvement of the correct word between different gate durations and found that a gate of 190 milliseconds performed best. We also confirmed this with smaller subsets of the data, suggesting that this result generalizes to unseen data.

The mismatch measure we developed can be usefully applied in language research and could inform the discussion about autonomous versus interactive language processing. Norris et al. (2016), arguing for the autonomous word recognition models, discusses the evidence pertaining predictive coding and suggests that more evidence is needed to provide insight for the role of predictive coding in language processing. The mismatch measure can elucidate whether cognitively high-level anticipations are relevant during the processing of low-level incoming speech sounds in human listeners, which, if found, would provide evidence against a strong autonomous bottom-up-only mechanism for speech perception (see below for a possible experiment).

A further question concerning the role of prediction in language processing is to what extent listeners predict speech input in regular non-experimental situations. Huettig (2015) notes that most evidence for prediction in language processing comes from experiments that only investigated the extremes of predictability, comparing, for example, highly predictable words with unpredictable words. Recent studies (e.g., Smith & Levy, 2013; Frank et al., 2015; Willems et al., 2016) using information-theoretic measures, such as word surprisal and entropy to predict processing costs during language processing, investigate the whole spectrum of predictability. These studies show that human listeners and readers are sensitive to these information-theoretic measures across the whole predictability spectrum. Similarly, the mismatch measure we developed quantifies the whole range of mismatch between high-level expectations and low-level input. This will allow us to investigate the importance of predictive coding in regular speech processing.

A key test of the mismatch measure is to analyze its relation to data from human listeners. For example, in an experiment using electroencephalography (EEG) it has been shown that listeners are sensitive to violations of expected auditory forms (Connolly & Phillips, 1994; Brunellière & Soto-Faraco, 2013); this effect is referred to as the phonological mismatch negativity (PMN). We hypothesize that our measure should predict the amplitude of the PMN, whereby higher cross-entropy between prior and post WPD would result in a more negative deflection of the EEG-signal.

3.5 Conclusions

The predictive coding framework proposes that the mismatch between cognitively high-level expectations and low-level perceptual input is an important mechanism in perception. We showed that we can quantify this mismatch for speech perception with the aid of statistical language modelling and an automatic speech recognition system. We used naturalistic speech recordings, containing approximately 50,000 words, to compute the mismatch measure. This opens up the possibility of investigating the importance of predictive coding during normal speech processing. We propose that the mismatch measure could be used to predict processing measures of listeners during speech perception. The results can inform the discussion about autonomous versus interactive models of speech perception.

Listening with great expectations: An investigation of word form anticipations in naturalistic speech

Chapter 4

This chapter is reformatted version of:

Martijn Bentum, Louis ten Bosch, Antal van den Bosch & Mirjam Ernestus (2019). Listening with great expectations: An investigation of word form anticipations in naturalistic speech. *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019

Abstract

The event-related potential (ERP) component named *phonological mismatch negativity* (PMN) arises when listeners hear an unexpected word form in a spoken sentence (Connolly & Phillips, 1994). The PMN is thought to reflect the mismatch between expected and perceived auditory speech input. In this paper, we use the PMN to test a central premise in the predictive coding framework (Friston, 2005), namely that the mismatch between prior expectations and sensory input is an important mechanism of perception. We test this with natural speech materials containing approximately 50,000 word tokens. The corresponding EEG-signal was recorded while participants (n = 48) listened to these materials. Following Chapter 3, we quantify the mismatch with two word probability distributions (WPD): a WPD based on preceding context, and a WPD that is additionally updated based on the incoming audio of the current word. We use the between-WPD cross-entropy for each word in the utterances and show that a higher cross-entropy correlates with a more negative PMN. Our results show that listeners anticipate auditory input while processing each word in naturalistic speech. Moreover, complementing previous research, we show that predictive language processing occurs across the whole probability spectrum.

4.1 Introduction

Human listeners easily perceive words when listening to continuous speech in normal circumstances. This apparent ease hides the profound difficulty of extracting words from a speech stream, which is for example attested by the challenge of developing artificial speech recognition systems with human-like performance. The details of the human speech processing system are still contentious. A long-standing debate concerns the timing and importance of top-down and bottom-up processing. Autonomous models of word recognition, for example Shortlist (Norris, 1994; Norris & McQueen, 2008), claim that early phases of speech processing are exclusively bottom-up and that top-down information can only exert influence at the lexical level. In contrast, interactionist models allow for top-down influence to affect lower-level acoustic processing, for example TRACE (McClelland & Elman, 1986). We investigate the influence of top-down expectations on the processing of low-level auditory speech input with an event related potential (ERP) called the phonological mismatch negativity (PMN).

The PMN (also referred to as N200, N250 or phonological mapping negativity) was first reported by (Connolly & Phillips, 1994; Connolly et al., 1990; Connolly et al., 1992). They presented spoken sentences to participants while recording the electroencephalography (EEG) signal. The sentences were highly constraining, for example, "*the gambler had a streak of bad …*" and the final word either initially matched the expected word (e.g., *luggage*, when *luck* is expected), or, alternatively, mismatched the expected word. A mismatching word showed an early negativity, around 200 milliseconds from word onset compared to the initially matching word.

The PMN is interpreted by Connolly & Phillips (1994) as reflecting a mismatch between expected word forms (based on the context) and observed auditory input (for a different interpretation of the PMN see Van den Brink et al., 2001; Hagoort, 2007). As noted by Brunellière & Soto-Faraco, 2013), this interpretation is closely related to an important claim in the predictive coding framework (Friston, 2005), namely, that higher-level cognitive processes generate predictions about low-level perceptual input. The mismatch between these predictions and the perceived input results in an error signal, useful for generating new expectations. For speech perception, this could mean that listeners generate word form expectations based on the preceding context, and violations of these expectations incur a processing cost.

In the current study, we aim to investigate predictive language processing. As Huettig (2015) notes, many experiments investigating predictive language processing only test extremes of predictability (see also Van Berkum et al., 2005; Norris et al., 2016). For example, the N400, an ERP component thought to reflect word predictability (Kutas et al., 2011; but see Hagoort, 2007 for a different interpretation) is typically based on a comparison of very likely versus very unlikely words. The aforementioned PMN is for example elicited with highly constraining sentences as the example above. This leaves open the question of whether language processing normally involves prediction, or only in these extreme cases.

Recently, several studies (e.g., Frank et al., 2015; Smith & Levy, 2013) have shown that the whole spectrum of word predictability can be investigated by utilizing information-theoretic measures. For example, Frank et al. (2015) used word surprisal, estimated with a statistical language model, to successfully predict the amplitude of the N400 measured while participants were reading sentences. Their approach is not based on the dichotomy of likely versus unlikely words, but instead uses the whole range of word probabilities. Furthermore, they used a large set of naturalistic language materials (see also Willems, 2016), improving the ecological validity of their findings. These studies provide stronger evidence for prediction during normal language processing.

In the current study we expand on this type of research with a mismatch measure inspired by the predictive coding framework. The measure developed in Chapter 3 quantifies the mismatch between expected and actual sensory speech input. To implement this mismatch measure, we need to quantify the mismatch between top-down expectations and bottom-up observations. The top-down expectations are estimated with a statistical language model (SLM). The SLM estimates the probability P of a word W_i given the preceding words W_{i-n} ... W_{i-1} and thus captures the top-down expectations (see Equation 1).

$$\hat{P}(W_i|context) = P(W_i|W_{i-n}, \dots, W_{i-1}) \qquad (1)$$

The bottom-up speech input will be represented with an auditory fragment of the initial part of the current word. This audio fragment needs to be processed in such a way that it can update word expectations. To achieve this, we use automatic speech decoding techniques used in speech recognition software and adapt these to estimate the probability of a phoneme sequence given the partial auditory fragment of the current word. We use the resulting phoneme sequence probabilities to update the top-down expectations (i.e., word probabilities estimated with the SLM). In this manner we compute two word probability distributions (WPD): one prior WPD based only on the SLM output (which is based on the previous words), and one post WPD, which is the prior WPD updated with the phoneme sequence probabilities based on the audio fragment of the current word.

The post WPD differs only from the prior WPD in the added auditory information. We therefore propose that the cross-entropy between prior and post WPD quantifies the mismatch between high-level expectations (based on previous word context) and auditory input. The cross-entropy can be computed according to Equation 2, whereby H denotes cross-entropy, p the prior WPD, q the post WPD and X the WPD word list.

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x) \qquad (2)$$

In the current study we test whether we can predict the amplitude of the PMN with the cross-entropy between prior and post WPDs. Based on the predictive coding framework and the EEG literature, we hypothesize that with increasing cross-entropy a listener incurs a higher processing cost (i.e., the sensory speech input is more surprising), which is reflected in a more negative amplitude in the 200 millisecond latency range. In the following sections, we will describe the EEG experiment and the methods used to test our hypothesis, followed by the results, a discussion and a conclusion.

4.1 Method

4.1.1 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 sessions of EEG recordings. All participants gave informed consent to participation.

4.1.2 Materials

For the experimental stimuli, we used materials from two corpora: the Spoken Dutch Corpus (Oostdijk, 2001) and IFADV (Van Son et al., 2008). These corpora contain audio recordings of Dutch speech. We extracted stretches of speech from these corpora, varying in duration from 4 to 15 minutes). The extracted speech stretches contain 50,277 word tokens (see Table 1). This subset, henceforth called Speech Corpus, consists of annotated speech from three speech registers (see Table 1). The different registers were selected for a different experiment and will not be relevant for the current study. For the estimation of the cross-entropy (see Section 4.2.3) we used NLCOW14, henceforth COW (Schäfer, 2015; Schäfer & Bildhauer, 2012), which is a large collection of web-crawled Dutch texts (4,7 billion words).

Table 1: Overview of the materials in the Speech Corpus.

speech style	word tokens (word types)	average word duration (ms)
spontaneous dialogues	21,718 (2,435)	206
read-aloud stories	13,209 (2,349)	256
news broadcast	15,350 (3,526)	289
total	50,277 (5,866)	245

4.1.3 Computing cross-entropy

We computed the cross-entropy (as detailed in Chapter 3) for all words longer than 60 and shorter than 700 milliseconds (46,734 word tokens, 5,254 word types). All subsequent analyses are performed on this subset of the Speech Corpus. To compute the cross-entropy, we need a word probability distribution (WPD) at the start of a word and a WPD after the auditory update. Therefore, we estimated for each word one WPD *prior* and one WPD *post* auditory update. These WPDs consists of a list of approximately 200,000 word types with associated probabilities. We created the word type list by selecting the most frequent word types in the COW corpus. We use the term word type to refer to the surface form of a word, i.e., *boy* and *boys* are two different word types.

The prior WPD is an estimation of word probabilities at the start of a word, given the preceding words (see Eq. 1). For example, consider the phrase *he played the guitar,* the prior WPD for the word *guitar* consists of the 200,000 word types with corresponding probabilities, given the preceding words *he played the.* To estimate the probabilities, we trained a 4th order Markov SLM on the COW corpus with the SRILM (Stolcke, 2002) toolkit (for smoothing we used Kneser-Ney discounting (Chen & Goodman, 1999)). We used this SLM to create a prior WPD for each word in the Speech Corpus.

We created the post WPD by updating the prior WPD with an auditory fragment of part of the current word (i.e., *guitar*). To perform this update, we transformed the audio fragment into probabilities of phoneme sequences. To estimate these probabilities, we extracted an audio fragment from word onset. By testing different durations for the auditory update, Chapter 3 showed that a fragment of 190 milliseconds, is the optimal duration for the cross-entropy computation. We analyzed audio material by using KALDI (Povey et al., 2011) as a speech decoding framework. We provided KALDI with a dedicated decoding lexicon whereby each entry was a sequence of 1 – 8 phonemes. We limited the set of phoneme sequences (n ≈ 400,000) to those that are found in Dutch words. The KALDI analysis resulted in a 50-best list of phoneme sequences with corresponding probabilities. The phoneme sequences were matched with the word types in the prior WPD and the probabilities were adjusted accordingly by the conventional Bayes rule, with the post WPD as result.

4.1.4 Procedure

Participants visited the lab on three occasions. Consecutive visits were separated by at least a week. Participants were fitted with the correct size electrode cap and seated in a sound-attenuated booth. They were asked to sit still and keep eye-movement and blinks to the minimum. During each visit, participants listened to approximately 90 minutes of speech, 270 minutes in total. The speech materials were presented over in-ear headphones (Etymōtic ER1) on a comfortable listening level (tested with a short audio fragment). The speech materials were presented in blocks of approximately 15 minutes, followed by a short break. During breaks in the experiment, yes-no comprehension questions were visually presented and participants responded via a button box.

4.1.5 EEG recording

We placed 26 cap-mounted silver-chloride electrodes according to the 10 - 20 international 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). We used four additional electrodes to monitor eye-related artefacts (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 EOG signals). Two additional electrodes were placed on the left and right mastoid. All electrodes were referenced to the left mastoid electrode and all electrode impedances were below 15 kΩ before recording started. The EEG-data was amplified with an Easycap system, band-pass filtered with 0.01 and 100 Hz cut-off frequencies, and digitized at a 1000 Hz sample frequency.

4.1.6 Preprocessing

The EEG-signal was re-referenced off-line to the left and right mastoid channels and filtered with a 5th order Butterworth bandpass filter with cut-off frequencies at 0.05 and 30 Hz. We removed artefacts from the EEG data semi-automatically, whereby all suggested artefacts were manually checked. We determined per block whether EEG channels with poor signal quality should be removed from the dataset. The Fp2 channel was completely removed from all recordings, due to poor overall signal quality.

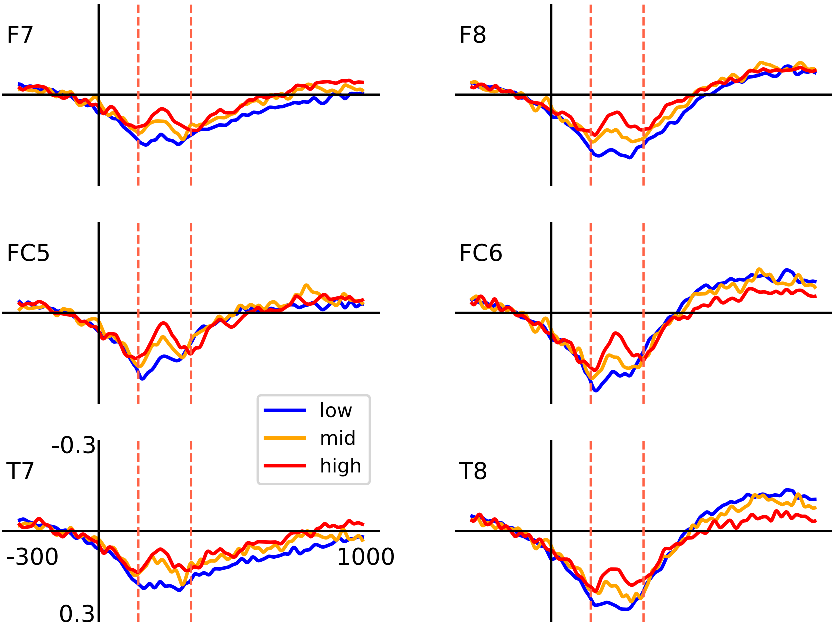
Subsequently, we removed activity related to blinks and eye movements from the EEG signal with the aid of independent component analysis (ICA). Following Winkler et al. (2015), the ICA was computed on data bandpass filtered with cut-off frequencies at 1 – 30 Hz, with aid of the MNE toolkit (Gramfort et al., 2014; Gramfort et al., 2013). We visually inspected the resulting components and selected those that were related to blinks and eye movement. We recomposed the EEG-data, bandpass filtered at 0.05 – 30 Hz, without the selected components.

The cleaned EEG-signal was time-locked to the words in the Speech Corpus. We extracted an epoch from 300 milliseconds before to a 1000 milliseconds after word onset for each word. All word epochs exceeding ± 75 μV on any channel were excluded from the dataset. We excluded the data of 9 (of the 48) participants because of poor signal quality (i.e., less than 40% of the data remaining after artefact removal). This resulted in a dataset of 1,172,894 word epochs (52.3% of all data). This dataset is part of the Dutch EEG speech register corpus (see Chapter 2)

4.2 Results

Based on previous literature (e.g., Connolly & Phillips, 1994; Van den Brink et al., 2001; Brunellière & Soto-Faraco, 2013), we expected a divergence of the grand average ERPs for low, middle and high cross-entropy (i.e., we split the data into terciles based on cross-entropy) at around 200 milliseconds from word onset at frontal sites. To select a latency range and a set of channels for analysis, we also inspected the grand average ERPs for all channels (see Figure 1 for a subset and Figure 2 for topographic plot of the relevant time window). We computed the average over the 150 – 350 millisecond time window and the following channels: F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7 and T8. The result of this averaging was one value for each word epoch. Following Winkler et al. (2015), we did not subtract the baseline from the ERP. Instead, we used the baseline as predictor in our statistical model. We computed the baseline by averaging over the same channel set for the -150 – 0 millisecond time window.

Figure 1: Grand average ERPs for words with low (blue), middle (yellow) and high (red) cross-entropy. The x-axis shows time in milliseconds and the y-axis shows amplitude in μV (negative is plotted upwards). The vertical dashed lines indicate the window between 150 and 350 milliseconds.



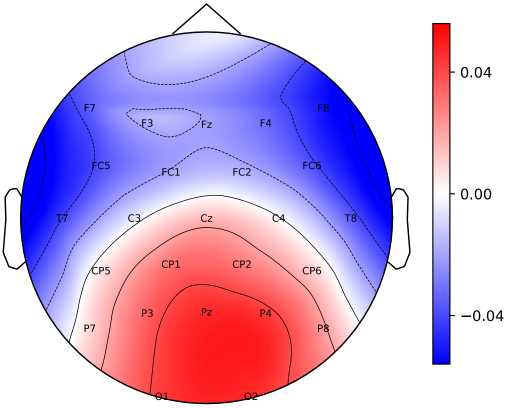
We analysed the data with linear mixed effects models Bates et al., 2015 in R (R Core Team, 2015), with the per-word ERP amplitude as dependent variable and cross-entropy as predictor of interest. The standardized covariates are the aforementioned baseline, the surprisal of the word, the entropy of the prior WPD, log frequency of the word in the COW corpus, the duration of the word, the word number in the sentence, and the word number in the block. Furthermore, we added participant and word as random effects. We considered a random slope for cross-entropy by participant but did not include it in the final model, because it resulted in a convergence error.

Table 2: Overview of the fixed effects in the linear mixed effect model with the PMN as dependent variable. The variable names, the beta (B) value, the standard error (SE B) and the t-value (t) are reported.

name	B	SE B	t
intercept	-0.23	0.04	6.5
entropy	0.11	0.01	16.3
surprisal	-0.02	0.01	-1.9
baseline	6.02	0.01	1145.9
log frequency	0.07	0.02	3.8
word duration	-0.06	0.01	-6.7
word in sentence	-0.04	0.01	-6.4
word in block	-0.05	0.01	-8.5
cross-entropy	-0.02	0.01	-3.6

We computed two models: a simple model without the predictor of interest, cross-entropy, and a second model with this predictor. Table 2 lists the fixed effects of the second linear mixed effect model. Model comparison reveals that the model with cross-entropy significantly improves compared to the simple model χ^2 (1) = 4.86, $p < .05$. The PMN is more negative with increasing values of cross-entropy.

Figure 2: Topographic difference plot between words with a high cross-entropy versus words with a low cross-entropy, averaged over the time window 150 - 350 milliseconds after word onset.



4.3 Discussion

According to the predictive coding framework Friston (2005), the mismatch between prior expectations and sensory input is an important mechanism for perception. We tested this claim for speech perception with an event related potential (ERP) component named the phonological mismatch negativity (PMN). This component is thought to reflect the mismatch between expected and actual auditory word forms. Following Chapter 3, we quantified the mismatch between expectations and sensory input as the cross-entropy between two-word probabilities distributions (WPD), one based on preceding words, and an updated version based additionally on the auditory input. We found that in line with our hypothesis, the cross-entropy has a negative correlation with the PMN amplitude, i.e., higher cross-entropy corresponds with a more negative amplitude of the PMN (see Figure 1 & 2 and Table 2). We propose that speech processing involves a comparison between high-level expectations and the auditory speech input. When the input mismatches with the expectations, this incurs a processing cost.

We recorded the electroencephalography (EEG) signal while participants listened to naturalistic speech. Following Willems (2015), the speech materials were extracted from corpora and represented normal language. Participants listened to long stretches of speech (approximately 15 minutes), instead of individually presented sentences or words. This has the distinct advantage that more relevant EEG data can be acquired. However, the downside is an increase of artefacts, because it is not possible for participants to sit perfectly still or refrain from blinking for a block of 15 minutes. As a result, more materials needed to be removed (~ 50% of the word epochs) than in more classical EEG experiments. However, because we analysed most words (93%) in the speech materials, we were able to create a dataset with approximately one million word epochs, orders of magnitude larger than analysed in classical EEG experiments.

The amount of data is important for this experiment. We investigated an ERP based on stimuli (i.e., words) while we could not control for sensory form or the preceding context, i.e., the target stimuli were the spoken words occurring in naturalistic speech. ERPs are sensitive to these kinds of differences (Luck, 2014) and it is only by averaging over very many tokens, which we had available, that this diversity averages out in the EEG signal.

Since we investigate many words in many contexts, we could investigate predictive language processing not just in highly constraining or artificially (un)likely contexts. As Huettig, 2015) noted, most experimental evidence for predictive language processing is based on experiments using these kinds of artificial language input. Recent studies (e.g., Frank et al., 2015; Smith & Levy, 2013; Willems et al., 2016) have been using a new approach, whereby language processing costs are predicted based on information theoretic measures, investigating a wide range of the probability spectrum. We extended these findings by using the mismatch measure (developed in Chapter 3) to predict processing costs of word forms during natural language perception. Our study shows that listeners do indeed anticipate word forms over the whole range of predictability, in line with the idea of *graded predictions* (see Van Berkum, 2005).

We propose that our finding is best explained by top-down feedback. This explanation is at odds with autonomous models of word recognition (e.g., Shortlist (Norris, 1994; Norris & McQueen, 2008)), which claim that early speech perception consists of bottom-up-only processing and do not allow for top-down feedback. Norris et al. (2016) defend the idea of no feedback by stating that processing of the acoustic signal is already optimal (by means of Bayesian inferencing), and therefore cannot be improved by feedback. However, feedback can be highly informative as an error-detection device, informing the listener whether the current word priors are on point (i.e., how well do they explain the current input). If the overall error increases, then the generative model needs to be adapted. The difference between the expected and perceived speech input (the error signal) provides a mechanism to dynamically adapt perceptual processing (and a built-in learning system to boot). This ‘error signal’-based reasoning is in line with predictive coding, which proposes that early stages of sensory input processing involves propagating the mismatch between the expected and actual sensory input.

4.4 Conclusions

We used a novel experimental approach in which participants listened to naturalistic speech while their EEG signal was recorded. Based on one million EEG word epochs, we showed that the ERP named the PMN has a negative correlation with cross-entropy, which quantifies the mismatch between expected and perceived auditory input. We showed that naturalistic speech stimuli can be used in an EEG experiment, and that it is possible to analyse most words (93%) in these speech materials. Furthermore, we extended research using an information-theoretic measure to predict processing costs of word forms, and provided additional evidence for extensive predictive language processing.

Do speech registers differ in the predictability of words?

Chapter 5

This chapter is reformatted version of:

Martijn Bentum, Louis ten Bosch, Antal van den Bosch & Mirjam Ernestus (2019). Do speech registers differ in the predictability of words? *International Journal of Corpus Linguistics*, *24*(1), 98-130.

Abstract

Previous research has demonstrated that language use can vary depending on the situational context. The present paper extends this finding by comparing word predictability differences between 14 speech registers ranging from highly informal conversations to read-aloud books. We trained 14 statistical language models to compute register-specific word predictability and trained a register classifier on the perplexity score vector of the language models. The classifier distinguishes perfectly between samples from all speech registers and this result generalizes to unseen materials. We show that differences in vocabulary and sentence length cannot explain the speech register classifier's performance. The combined results show that speech registers differ in word predictability.

5.1 Introduction

People communicate in different situations and modalities, ranging from casual conversations between friends to formal lectures or public addresses. Many previous studies have shown that these different situations elicit different language use; see Biber & Conrad (2009) for an overview. The term 'register' is used to provide a link between a communicative act and the context of the situation it occurs in (Marco, 2000). Likewise, we will use the term register to refer to language variation in relation to the situation of use (see Lee, 2001 for a discussion). In this paper, we investigate register-specific differences in word predictability, defined as the conditional probability of a word given the preceding words. We conducted five experiments to test whether speech registers differ in word predictability.

To investigate register differences in word predictability, we use statistical language modelling, a technique widely used within the discipline of Natural Language Processing (NLP). We compute register-specific word predictability scores with the aid of statistical language models (SLM) and use these scores to train a speech register classifier. The performance of the classifier shows, to what extent speech registers differ in word predictability. In Section 5.2, we will explain how this NLP approach complements register analysis. In Section 5.3, we introduce the corpora we use for this study and outline our analysis approach. In the following sections, we describe the experiments we conducted. In Study 1, we investigate how to create SLMs that allow cross register comparison. In Study 2, we train and test register-specific SLMs to estimate register-specific word predictability. These word predictability scores are then used to train a speech register classifier. In Study 3, we validate the results from Study 2, by testing the speech register classifier on the validation corpus. In Study 4, we investigate the amount of data necessary for classification. Finally, in Study 5, we investigate the influence of average sentence length on word predictability. We end with a general discussion of our findings.

5.2 Characterizing text in register analysis and natural language processing

The most fundamental approach in register analysis is to count lexico-grammatical features (e.g., demonstrative pronouns), and compare their prevalence across registers. The studied materials may be written, or consist of orthographic transcription of speech samples. For example, Tottie (1991) investigates differences between spoken and written British English and found that negatives are twice as prevalent in spoken language as in written text. Van Gijsel et al. (2006) compare excerpts from different speech registers in Dutch and show that word type-token ratio (TTR) is lower for informal dialogues than for formal monologues.

Biber (1988, 1995) develops an approach for register analysis known as multidimensional analysis, which aims at identifying co-occurring linguistic features and discovering underlying dimensions of language use by means of factor analysis. For example, Biber (1988) finds that discourse particles, first and second person pronouns, and present tense verbs are typical of *involved* language. Conversely, a high frequency of nouns and prepositions and a high word type-token ratio are typical of *informational* language. The dimensions can be used to group or distinguish between different registers and give a functional interpretation to the patterns of lexico-grammatical features (Biber & Conrad, 2009).

In contrast to the interpretative approach of register analysis, text classification methods as developed within the discipline of NLP, characterize texts by a large set of (mostly) automatically generated features (see Killgariff, 2001 for an overview). The feature set typically consists of *n*-grams, which can for example consist of POS tags, or words. Based on *n*-grams, an SLM can be created that estimates the probability of a word given the preceding words. SLMs are a staple technology for applications such as machine translation, automatic speech recognition, and document retrieval (Jurafsky & Martin, 2009).

Both register analysis and NLP have advantages and disadvantages. For example, because register analysis uses a relatively small set of lexico-grammatical features to describe and interpret differences between registers (Biber & Conrad, 2009), it precludes data-driven research. Registers can only be characterized with features that are defined beforehand, based on previous research or on the researcher's intuitions. Statistical language modelling avoids this and opens up the possibility of a data-driven search of patterns in a corpus.

From the perspective of register analysis, there is a disadvantage to SLMs; because typically many features are used, the interpretation of patterns of textual differences is difficult. The feature set is essentially a long list of item co-occurrence statistics, and therefore ill-suited for human interpretation. Still, *n*-grams are a valuable tool for various types of analysis. For example, Gries (2001) successfully uses the statistics of word co-occurrences to disambiguate the meanings of near synonyms. Denoual (2006) uses character n-grams (i.e., based on graphemes instead of words) to classify texts on a dimension ranging from literary to oral.

We propose that investigating the distribution of word *n*-grams across speech registers may reveal register differences not accessible with current register analysis tools. We use word *n*-grams because they are theory neutral; only minimal assumptions have to be made to count and compare *n*-grams of words (see also Gries & Ellis, 2015: 231). Moreover, previous research shows that listeners are sensitive to the statistics of word *n*-grams.

For our current study, word predictability is a crucial concept. We define word predictability as the probability of a word given the previous context (i.e., the preceding words). For example, the predictability of the word *gun* given the context *The policeman pulled out his ...* is high compared to a word like *socks.* Word predictability is thus the conditional probability P(*word*|*context*) of a word *word* given the preceding context *context*, which can be estimated with an SLM (e.g., Smith & Levy, 2013).

Word predictability plays an important role in language comprehension (e.g., Kutas et al., 2010). Converging evidence from studies using different methodologies such as self-paced reading (e.g., Monsalve et al., 2012; Smith & Levy, 2013), eye-tracking (e.g., Frisson et al., 2005), EEG (e.g., Van Berkum et al., 2005), and fMRI (e.g., Willems et al., 2016) show that that the processing of speech and text is influenced by the predictability of a word given the previous context. For an overview of frequency effects in language processing, see Ellis (2002).

Word predictability also plays a role in language production. For example, Bell et al. (1999) found that the pronunciation of English function words depends on word predictability, whereby less predictable words are pronounced in fuller form. Similarly, Pluymaekers et al. (2006) found that the duration and number of segments of Dutch suffixes are influenced by the predictability of the carrier word.

The widespread and converging evidence for the importance of word predictability in language comprehension and production led us to investigate to what extent word predictability differs across registers. One reason to suspect differences is the aforementioned finding that lexical richness differs across speech registers (e.g., Van Gijsel et al., 2006); more formal registers have higher word type-token ratios than more informal registers. If one register contains more word types compared to other registers, it is likely that this influences word predictability. We will use SLMs to compute register-specific word predictability and test whether predictability patterns distinguish between registers.

5.3 Methodology

We describe the corpus we use in Section 5.3.1 and our methods of analysis in 5.3.2.

5.3.1 Corpus

We used a subset of the Spoken Dutch Corpus (Oostdijk, 2001). This corpus is ideally suited to investigate speech register differences, because it consists of *components* reflecting speech in different situations of use, ranging from spontaneous conversations to television news broadcasts and read-aloud stories. We used the orthographically transcribed recordings of adult native speakers in the Netherlands. We excluded the Flanders part (approximately one-third) of the corpus, because possible differences between Northern Dutch and Flemish Dutch speech styles are outside the scope of our study. In addition, we excluded one component ("Masses and solemn speeches") because it is comparatively small (fewer than 6,000 word tokens). This left 14 components for analysis (see Table 1). This subset consists of approximately five million word tokens of Netherlandic Dutch speech, a variety of Dutch spoken in the Netherlands.

Table 1. Overview of the 14 components in the Spoken Dutch Corpus used for Studies 1 - 4

ID	Component description
a	Spontaneous conversations (face-to-face)
b	Interviews with teachers of Dutch
c	Spontaneous telephone dialogues via a platform
d	Spontaneous telephone dialogues via a minidisc recorder
e	Business negotiations
f	Radio and television interviews and discussions
g	Debates, discussion and meetings (especially political)
h	Classes
i	Spontaneous radio and television commentaries (e.g., sports)
j	Radio and television newsroom and documentaries
k	News broadcast on radio and television
l	Reflections and commentaries broadcast on radio and television
n	Lectures and speeches
o	Read-aloud stories

We also created a validation corpus to validate our findings and ensure they generalize beyond the materials in the Spoken Dutch Corpus. It consists of materials from three different corpora: two corpora of Dutch spontaneous speech, the Institute of Phonetic Sciences Amsterdam Dialogue Video Corpus, henceforth IFADV (Van Son et al., 2008), and the Ernestus Corpus of Spontaneous Dutch, henceforth ECSD

(Ernestus, 2000), and two components of the *STEVIN Dutch Reference Corpus*, henceforth SoNaR, (Oostdijk et al., 2013), namely a subset of Dutch teleprompt texts (news broadcasts) and Dutch books. We will refer to the combination of these new materials as the validation corpus, which consists of approximately 2.2 million word tokens.

The materials in the validation corpus were chosen because they correspond to three specific components in the Spoken Dutch Corpus. The two corpora of spontaneous speech (IFADV and ECSD) correspond to component “a” (“Spontaneous conversations”), the set of Dutch teleprompt texts correspond to component “k” (“News broadcasts on radio and television”) and, finally, the Dutch books correspond to component o (“Read-aloud stories”).

The SoNaR texts are not an orthographic transcription of speech, while this is the case for all other corpora that were used in this study. They are nevertheless similar to the respective components “k” and “o” in the Spoken Dutch Corpus, because news broadcasts (component “k”) are typically read from teleprompts and should conform to the teleprompt texts closely, and read-aloud stories (component “o”) are a collection of read-aloud audiobooks. Still, differences could occur between the SoNaR materials and the orthographically transcribed texts, for instance, in the placement of sentence boundaries.

5.3.2 Analysis

We used SLMs to investigate whether speech registers influences word predictability. The reasoning is as follows. SLMs are sensitive to the difference between the language materials they are trained on and the materials they are tested on. The performance of a language model in terms of predicting the next word correctly on the basis of a sequence of previous words is known to suffer in general if the difference between the training and test set increases. We assert that this is also likely to apply to differences in speech register. For example, if an SLM is trained on spontaneous conversations and subsequently tested on read-aloud stories, the model’s predictive performance (i.e., its ability to assign the correct probability to the next word given the preceding context) is likely to be worse than in a test on an unseen set of spontaneous conversations. SLM performance can thus be utilized to assess the similarity of different registers to the register the model was trained on. We use this language model characteristic to determine word predictability differences between speech registers.

To test whether speech registers systematically differ in word predictability, we train a classifier on the SLM performance measures. If word predictability differs between speech registers, the classifier should be able to differentiate these registers and achieve good register classification results. In addition, we investigate the amount of data necessary to achieve accurate classification of speech registers. Furthermore, we aim to rule out that our classifier results are driven by sentence length differences between speech registers. This is important because sentence length could influence the SLM results, as SLMs tend to assign higher likelihood scores to shorter sentences. Furthermore, registers can differ in sentence length (Wiggers & Rothkrantz, 2007).

Because we aim to compare SLM word predictability scores between registers, the SLM vocabulary (i.e., a list of all words used to train the SLM) deserves special consideration. An SLM's vocabulary is typically based on the texts it is trained on, referred to as a 'training set'. The 'out-of-vocabulary words' (i.e., words not part of the language model, also referred to as 'OOV words') are typically ignored in performance evaluation. However, we train SLMs on different registers and want to compare between them. If the number of OOV words differs between SLMs trained on different speech registers, this can influence test results of the SLM; for instance, if a register contains many OOV words, the SLM could attain an artificially boosted performance. Therefore, for a fair comparison between all register-specific SLMs, they should have the same register-insensitive vocabulary.

For the creation of the fixed SLM vocabulary, we need a corpus containing multiple registers and an approach for vocabulary word selection. Two extreme approaches are possible: 'greedy selection', that is, selection of all or nearly all words occurring in the corpus; or 'robust selection', that is, selection of only those words that are most likely present if the corpus would be created again, regardless of register. For example, consider the word *gamble*, which can be used in many different registers, while the word *inning* typically occurs in sports commentaries. In this example, the word *gamble* is a good candidate for a robust vocabulary, while *inning* may not be.

The advantages of greedy selection are the maximum use of available data and a straightforward inclusion criterion, which typically consists of the selection of all words occurring above a certain frequency threshold (e.g., word frequency of 5) in the corpus. The disadvantage of greedy selection relates to the unreliability of the decision to include a word. For example, the 'burstiness' of words, the phenomenon that a word's likelihood increases if it has been used recently (Church & Gale, 1995), lead to an uneven distribution of tokens throughout a corpus. These findings make word frequency an unreliable measure to base word selection criteria on (Kilgariff, 2001; Gries & Ellis, 2015).

Robust selection addresses the word burstiness problem. Savický & Hlaváčová (2002) developed a metric called average reduced frequency (ARF), which adjusts word frequency based on the word's dispersion in a corpus, whereby a word with low dispersion (i.e., with a bursty distribution) results in a lower ARF as compared to a word that is more evenly distributed (cf. Section 5.4.1). If a word is used regularly throughout the corpus, it is more likely it will be found again in a newly sampled corpus, whereas a word that only occurs in local bursts may be an idiosyncratic (e.g., topical) characteristic of a specific corpus. Therefore, a vocabulary based on the highest scoring ARF words could improve word selection quality.

A potential disadvantage of robust selection is the reduction of the available data, because the resulting vocabulary will be significantly smaller than the vocabulary resulting from greedy selection. In addition, the word exclusion criterium is more complex and the quality of the vocabulary depends on the viability of these criteria. In sum, both approaches have their advantages and disadvantages, and it is unclear whether greedy or robust selection is the best way to create an SLM vocabulary for our purposes. Therefore, we test the greedy and robust SLM vocabulary selection strategies in Study 1 and select the best approach. The four subsequent studies use this approach to create SLMs. In these studies we test word predictability differences between registers, rule out confounds, and test the robustness of the found differences between registers. The methodological details of each study will be discussed in the respective sections.

5.4 Study 1: SLM vocabulary selection

In study 1 we tested whether robust or greedy selection is better suited for the creation of a SLM vocabulary.

5.4.1 Procedure

We extracted the orthographic transcriptions from the Spoken Dutch Corpus and removed the special corpus-specific word codes (explicitly marking foreign words, dialectal words, regionally accented words, new words, interjections, onomatopoeia, hesitations and mispronunciations, see Goedertier et al., 2000). Further text normalization was not necessary because the orthographic transcriptions were already tokenized and normalized according to the protocol described in Goedertier et al. (2000).

We defined word type as the word surface form (i.e., *run* and *runs* are two different word types) and created the greedy vocabulary by selecting the 50,000 most frequent word types from the corpus. We created the robust vocabulary by ranking word types based on their average reduced frequencies (see below) and selected all word types with an average reduced frequency (ARF) of at least 50. This lower bound of the ARF was based on the trade-off between coverage and the constraint that word types should be present in most components of the corpus. This resulted in a list of 585 words types, covering 77.5% of all word tokens in the corpus.

To compute the ARF of each word in the corpus, we extracted the first 61,834 word tokens (i.e., the number of tokens in the smallest component) from each component, which ensures that the ARF scores are not influenced by the amount of materials of each component in the corpus. We then calculated the reduced frequency (RF) of each word (Savický & Hlaváčová, 2002). The RF (Equation 1) equals the word's frequency if the word is evenly distributed throughout the corpus, while it has a lower bound of one if the word is clustered in one location in the corpus (Hlaváčová & Rychly, 1999). That is, words with the highest ARF are those words that occur evenly throughout the corpus and are therefore neither topic-specific nor register-specific.

To compute the RF for each word *w*, the corpus is divided into a number of intervals ($N_{intervals}$) equal to the frequency of word *w*. The RF is then computed as the number of intervals word *w* occurs in. Therefore, it is important to keep the original word order of texts and to group register-specific texts together.

$$RF = \sum_{i}^{N_{intervals}} f_w(i)$$
$$\begin{cases} f_w(\mathrm{i}) = 1, & if\ the\ word\ w\ occurs\ in\ the\ i^{th}\ interval \\ f_w(\mathrm{i}) = 0, & if\ the\ word\ w\ does\ not\ occur\ in\ the\ i^{th}\ interval \end{cases} \quad (1)$$

The RF depends on the start and end points of the intervals and the start point of the first interval determines the start and end points of all other intervals. There are many possible starting points for the first interval. To avoid this arbitrariness, the RF is calculated for all non-redundant starting points in the corpus, that is, for the first word of the corpus, up to and including the word with number $v = \lfloor N_{words}/N_{intervals} \rfloor$, where v denotes the number of starting points, N_{words} the number of word tokens in the corpus and $N_{intervals}$ denotes the number intervals the corpus is divided into. We computed the average reduced frequency for each word by averaging over all RFs.

To compare the greedy and robust vocabularies, we created two versions of our corpus. All OOV words were mapped to the dummy string *unk.* In one version, we used the greedy vocabulary to determine the OOV words and in the other version we used the robust vocabulary.

We used frequency profiling, described in Rayson & Garside (2000), to discover those *n*-grams (restricted to unigrams, bigrams or trigrams) in each component that distinguish a given component from the other components, for both the greedy and robust corpus versions. Frequency profiling compares the frequency of a *n*-gram in different corpora by computing the log-likelihood (Equation 2) of the *n*-grams frequency in one corpus compared to the frequency in one or more other corpora. For the computation of the log-likelihood we used the regular frequency (not the ARF) of the *n*-gram.

$$LL_{ngram} = 2\left(\sum\nolimits_i O_i \; ln\left(\frac{O_i}{E_i}\right)\right) \qquad (2)$$

In Equation 2 O_i denotes the *n*-gram frequency in the *i*-th corpus. E_i denotes the expected value of the *n*-grams frequency in the *i*-th corpus and is computed according to Equation 3,

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \qquad (3)$$

where N_i refers to the total number of *n*-gram tokens in the *i*-th corpus.

To compute the log-likelihood statistic we used the Colibri-Core toolkit (Van Gompel & Van den Bosch, 2016), which includes an implementation of frequency profiling. To investigate *n*-grams that are specific for a component compared to the rest of the corpus, we used the leave-one-out approach; we compared all *n*-grams in each component against the combination of the 13 other components. The log-likelihood statistic was calculated for all word unigrams, bigrams and trigrams, for both the robust and greedy corpus.

5.4.2 Results and discussion

The Colibri-Core toolkit returns *n*-gram lists ranked on log-likelihood, whereby the *n*-grams that distinguish a component most compared to the others are ranked at the top. We checked whether the greedy vocabulary resulted in a more uneven distribution of word forms across components compared to a robust vocabulary. We found that this was indeed the case. To illustrate this, we list the highest ranking unigrams for a sample of four components, in Table 2 for the greedy vocabulary version, and in Table 3 for the robust vocabulary version.

Table 2. Overview of the most distinguishing word forms of four speech registers based on frequency profiling, greedy vocabulary (50,000 most frequent word forms)

Casual conversation	Interviews with teachers of Dutch	Political debates	Sports commentary
ja (“yes”)	*uh* (“ehm”)	*het* (“it”)	*bal* (“ball”)
nee (“no”)	leerlingen (“pupils”)	*de* (“the”)	Kluivert*
oh (“oh”)	Nederlands (“Dutch”)	voorzitter (“chairman”)	Bergkamp*
‘k (“I”)	*lezen* (“read”)	*motie* (“motion”)	Zenden*
zo (“later”)	*onderwijs* (“education”)	*u* (“you”)	*de* (“the”)
echt (“really”)	*literatuur* (“literature”)	*heer* (“gentleman”)	balbezit (“ball possession”)
mmm (“ehm”)	*klas* (“class”)	van (“of”)	
wel (“well”)	*school* (“school”)	mevrouw (“lady”)	Overmars*
gewoon (“just”)	*vak* (“course”)	*minister* (“minister”)	Boer*
maar (“but”)	*ik* (“I”)	*vraag* (“question”)	Cocu*

NOTE: * name of Dutch soccer player

Table 3. Overview of most distinguishing words of four speech registers based on frequency profiling and a robust vocabulary (585 top ranking ARF words)

Casual conversation	Interviews with teachers of Dutch	Political debates	Sports commentary
ja (“yes”)	uh (“ehm”)	het (“it”)	unk*
nee (“no”)	lezen (“read”)	voorzitter (“chairman”)	de (“the”)
oh (“oh”)	school (“school”)	de (“the”)	speelt (“plays”)
'k (“I”)	ik (“I”)	u (“you”)	nul (“zero”)
zo (“later”)	vind (“think”)	heer (“gentleman”)	nu (“now”)
wel (“well”)	mmm (“ehm”)	van (“of”)	helft (“half”)
mmm (“ehm”)	heel (“very”)	minister (“minister”)	voor (“before”)
echt (“really”)	dus (“so”)	unk*	meter (“meter”)
maar (“but”)	ben (“am”)	vraag (“question”)	tweede (“second”)
gewoon (“just”)	kinderen (“children”)	om (“to”)	gaat (“go”)

NOTE: * *unk* is the dummy string that out-of-vocabulary words are mapped to.

We observe that the greedy selection approach produces a topicality confound (i.e., differences in *n*-gram frequency between components due to the topics discussed in the components). For example, the component containing interviews with teachers of Dutch contains many words specifically related to education (e.g., the Dutch equivalents of *pupils*, *school*, *class*), while the sports commentary component contains many proper names of Dutch soccer players (e.g., *Kluivert*, *Zenden*). A similar pattern is present in the higher order *n*-grams (i.e., bigrams and trigrams). Consequently, if we create SLMs based on a greedy vocabulary, it will not be possible to ascertain whether components are distinguished based on register or topic. The robust strategy, as illustrated in Table 3, attenuates the topicality confound. For example, the most distinguishing words for the sports commentary do not include proper names, and we see only few terms specifically related to education for the component containing interviews with teachers of Dutch. Note that, by using ARF to select words, we do not restrict the vocabulary to function words. As can be observed in Table 3, content words are also present in the robust vocabulary.

In sum, Study 1 showed that a greedy vocabulary introduces a topicality confound. Such a vocabulary contains many words that are specific for topics that happened to be discussed in one or several components of the Spoken Dutch Corpus. As a consequence, when we train the speech register classifier based on the SLM results obtained with the greedy vocabulary, we do not know whether speech registers are distinguished based on genuine register-specific word predictability or the coincidental distribution of topic-specific words. The robust vocabulary remedies this confound by excluding words that are not evenly distributed across the corpus.

5.5 Study 2: Training and testing of the speech register classifier

In Study 2 we test whether we can distinguish between register-specific components of the Spoken Dutch Corpus with a classifier based on word predictability.

5.5.1 Procedure

We used the same subset of the Spoken Dutch Corpus as described in Study 1 to train SLMs and create the speech register classifier. The Spoken Dutch Corpus was pre-processed as described in Study 1. We trained register-specific tri-gram models with the SRILM-toolkit[1] (Stolcke, 2002), using the robust vocabulary created in

Study 1. For smoothing, we used Witten-Bell discounting with interpolation (Witten & Bell, 1991). We could not use the standard smoothing technique, that is, modified Kneser-Ney discounting (Chen & Goodman, 1998), because of our small vocabulary of relatively frequent words. Kneser-Ney discounting needs counts of infrequent *n*-grams to assess the probability mass needed for unseen *n*-grams. Witten-Bell is able to deal with truncated count-of-count lists[2] because it uses the first occurrence of *n*-grams to assess the probability mass needed for unseen *n*-grams.

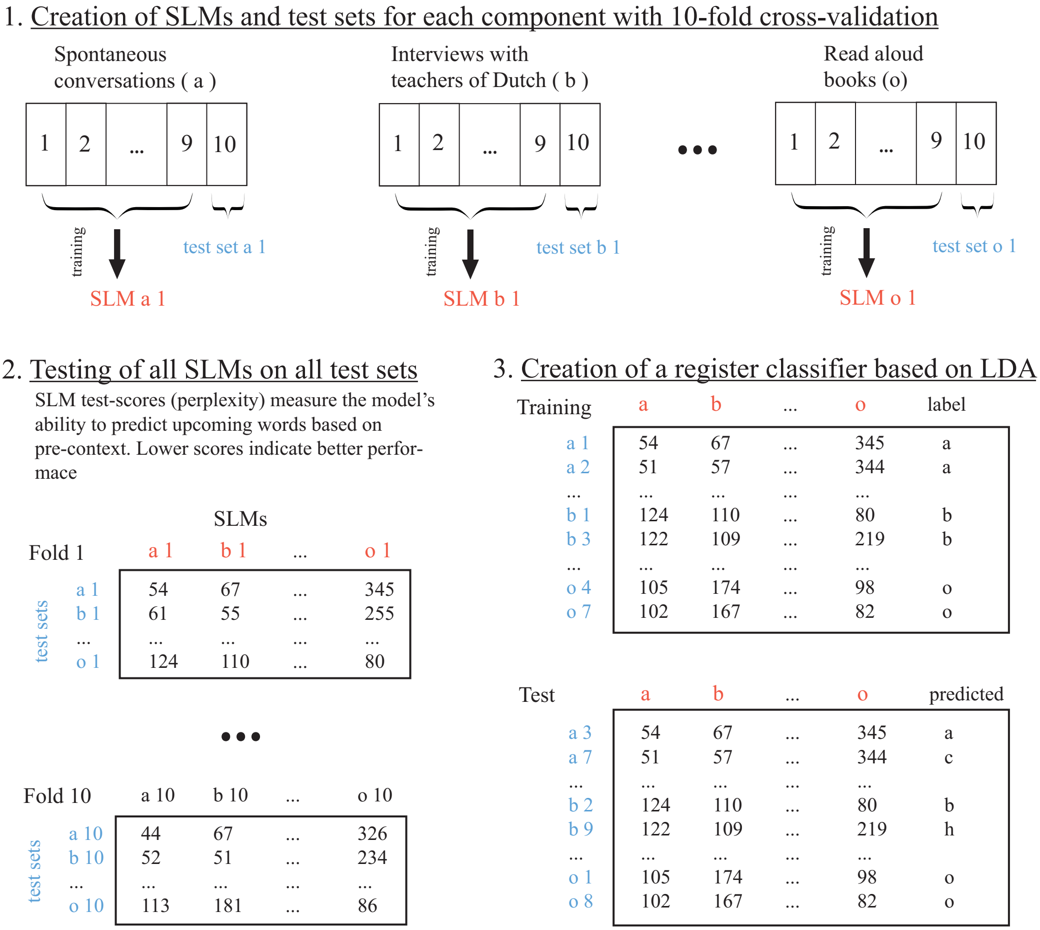
To create register-specific SLMs, we first mapped all OOV word tokens to the dummy string *unk.* The mapping was used to maintain the serial structure of the sentences. Next, we created training and test sets for each component in the Spoken Dutch Corpus by grouping all sentences of a given component into a single text file. Subsequently, the sentences of a given component were randomly assigned to one of ten equally-sized partitions to ensure a fair sampling of the register in all of the partitions.

For each component we ran a ten-fold cross-validation experiment on the partitions, using nine parts for training and one part for testing in a rotating fashion (see also Figure 1). The ten-fold cross-validation experiments yield perplexity scores for each of the ten folds. Perplexity is a measure of how well a register-specific SLM predicts words (based on the preceding words) in new, unseen texts. Importantly for our study, registers similar to the SLM will generate lower perplexity scores than less similar registers.

The perplexity scores were computed with Equation 4, where *word* stands for a specific word token in the test file and *context* stands for the preceding words (maximally a bigram). N_{words} and $N_{sentences}$ represent the number of word tokens and sentences in the test set, respectively, and N_{OOV} represents the number of out-of-vocabulary words, which always equal 0 in our test sets, because all OOV words were mapped to the *unk* token.

$$perplexity = \frac{\sum_{all\ words} 10_{log} P(word|context)}{(N_{words} - N_{OOV} + N_{sentences})} \qquad (4)$$

Figure 1. Workflow overview for the creation of a speech register classifier based on word 1 predictability



1. Creation of SLMs and test sets for each component with 10-fold cross-validation

2. Testing of all SLMs on all test sets

SLM test-scores (perplexity) measure the model's ability to predict upcoming words based on pre-context. Lower scores indicate better performace

SLMs

Fold 1	a 1	b 1	...	o 1
a 1	54	67	...	345
b 1	61	55	...	255
...	...	...	...	...
o 1	124	110	...	80

Fold 10	a 10	b 10	...	o 10
a 10	44	67	...	326
b 10	52	51	...	234
...	...	...	...	...
o 10	113	181	...	86

3. Creation of a register classifier based on LDA

Training	a	b	...	o	label
a 1	54	67	...	345	a
a 2	51	57	...	344	a
...	...	...	...	...	
b 1	124	110	...	80	b
b 3	122	109	...	219	b
...	...	...	...	...	
o 4	105	174	...	98	o
o 7	102	167	...	82	o

Test	a	b	...	o	predicted
a 3	54	67	...	345	a
a 7	51	57	...	344	c
...	...	...	...	...	
b 2	124	110	...	80	b
b 9	122	109	...	219	h
...	...	...	...	...	
o 1	105	174	...	98	o
o 8	102	167	...	82	o

For each test file (ten from each of the 14 components), we created a 14-dimensional vector of perplexity scores (i.e., a list of 14 perplexity scores, one for each SLM) by applying all 14 trained language models to that test file. The resulting perplexity vector describes how well the test file is predicted by the 14 register-specific language models. The perplexity vectors for the 140 test files form a 140-by-14 similarity matrix, whereby each row describes the location of a test file in a 14-dimensional space, while the columns correspond to the register-specific SLMs in

)

the Spoken Dutch Corpus. The perplexity similarity matrix shown in Figure 1 (step 3) is a subset of the complete similarity matrix we created based on the 14 components in the Spoken Dutch Corpus.

We used Linear Discriminant Analysis (LDA) to create a speech register classifier based on the similarity matrix. LDA finds a linear combination of features that maximizes class separation (see Equation 5).

$$\hat{x} = \underset{x}{argmax} \; \frac{x^t \sum_b x}{x^t \sum_w x} \qquad (5)$$

The between-class and within-class scatter matrices are represented by $\sum_b$ and $\sum_w$ respectively. A vector of weights $\hat{x}$ is found that maximizes the coefficients of the between-class and within-class scatter matrices, which results in an optimal class separation when two assumptions hold about the data: homoscedasticity (identical within-class scatter matrices) and within-class multivariate Gaussian distributions. Because our data do not conform to these assumptions, we validated our classifier, as will be discussed in Study 3.

5.5.2 Results and discussion

The speech register classifier was able to distinguish perfectly (accuracy 100%, on the held-out test sets) between all registers within the Spoken Dutch Corpus material. Compared to chance performance (accuracy 7.14%), the speech classifier performed considerably better. Performance metrics in terms of precision, recall and f1 can be found in Appendices 1–8.

5.6 Study 3: Validation of the speech register classifier

We showed that a classifier based on register-specific word predictability can distinguish between speech register-specific components (Cf. Study 2). In study 3 we tested whether the classifier is indeed sensitive to register differences between components. Furthermore, the LDA assumptions do not hold for our dataset and it was therefore important to test the robustness of our results. First, we compared the results of the speech register classifier with a classifier trained on a random version of the corpus, and second, we tested the speech register classifier on materials from different corpora, to test whether the performance of the classifier generalizes to new data.

5.6.1 Procedure

We constructed 1,000 pseudorandom corpora with materials from the Spoken Dutch Corpus to validate the speech register classifier. For each pseudorandom corpus the sentences from the Spoken Dutch Corpus were randomly assigned to one of 14 components. The random components were made to contain as many word tokens as the original components in the corpus. We trained component-specific SLMs and tested these on held out test sets with ten-fold cross validation as in Study 2). Subsequently, we trained an LDA classifier for each pseudorandom corpus, also following the procedure of Study 2. If the speech register classifier based on the real corpus outperforms the classifiers based on the pseudorandom corpora, then the classification accuracy of the register classifier must be due to the grouping of sentence according to speech register.

The four components in the validation corpus were pre-processed individually. Since IFADV was annotated with the same protocol as used for the Spoken Dutch Corpus (Van Son et al., 2008; Goedertier et al., 2000), we used the same pre-processing steps as in Study 1. The ECSD used a slightly different annotation style with more elaborate punctuation. To approximate the annotation and tokenization of the Spoken Dutch Corpus, we created sentences by splitting the text materials on question marks, exclamation marks, commas and points. We replaced the capital letter at the start of each sentence with the lowercase equivalent, even if it was part of a proper name, since proper names were not included in the SLM vocabulary.

All sentences in the teleprompt texts and Dutch books from the SoNaR corpus already start with lower-case characters. We split on questions marks, exclamation marks, colons, commas and points and removed all remaining punctuation. For the set of teleprompt texts, we also removed special recording instructions (e.g., *start audio*).

The four components in the validation corpus were each split into ten equally sized partitions, equal to the ten-fold cross-validation structure we created for the Spoken Dutch Corpus. On each partition we applied the corresponding SLMs trained on the Spoken Dutch corpus. The resulting perplexity vectors were used as classifier test sets for the register classifier trained on the materials from the Spoken Dutch

Corpus. Importantly, the validation materials did not influence the SLMs (which were exclusively trained on the Spoken Dutch Corpus) and did not influence the register classifier (which were trained only on the perplexity feature vectors from the Spoken Dutch Corpus). This validation therefore provides a strong test of whether our approach generalizes to new unseen data.

5.6.2 Results and discussion

The classifiers based on the pseudorandom corpora performed poorly, with a mean accuracy of 12% and a standard deviation of 5%. The performance is close to chance level performance (accuracy 7%). The result shows that a classifier based on perplexity scores cannot distinguish between random collections of sentences. The high performance of the classifier developed in Study 2 therefore indicates that the components of the Spoken Dutch Corpus are more homogeneous than those in the pseudorandom Corpora and that they differ in word predictability.

The speech register classifier developed in Study 2 yields an accuracy score of 93% on the validation corpus, compared to 100% accuracy on the held out classifier test sets of the Spoken Dutch Corpus. The classifier thus attained a high accuracy on materials from new corpora, which shows that the speech register classifier is not overfitted to idiosyncratic aspects of the Spoken Dutch Corpus. The accuracy score on the validation corpus was not perfect, however. The confusion matrix in Table 4 shows that all classification errors are made on the ECSD corpus of spontaneous speech. Interestingly, the ECSD is confusable with component "b", ("Interviews with Teachers of Dutch"). There is considerable overlap between ECSD and component "b", as both are unscripted dialogues, which suggests that the classification mistakes are not random.

Table 4. Confusion matrix of the speech register classifier test on the validation corpus

corpora	a	b	k	o
SoNaR-books	0	0	0	10
ECSD	7	3	0	0
IFADV	10	0	0	0
SoNaR-teleprompt	0	0	10	0

In conclusion, a speech register classifier based on word predictability can distinguish between genuine speech registers, but not between randomly sampled sets of sentences. In addition, we showed that the register classifier cannot only classify materials from the training corpus (the Spoken Dutch Corpus), but also materials from the validation corpus. The combined results suggest that word predictability differs across speech registers.

5.7. Study 4: How much text material is needed for speech register classification?

The aim of Study 4 was to investigate the amount of text materials needed for a reliable register classifier. We divided the speech registers into differently sized subsets. Classifiers trained on smaller subsets are expected to especially confuse more similar registers, which would provide further evidence that classification is based on register characteristics.

5.7.1 Procedure

We used materials from the Spoken Dutch Corpus and the validation corpus as described in Section 5.3. We used a similar procedure to that described in Study 2 except that we created perplexity vectors based on sets containing the following number of sentences from a specific speech register: 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. We did this by dividing the text materials of each register from the Spoken Dutch Corpus into sets of a specific number of sentences. We computed the perplexity vectors for all sentence sets according to Equation 5 with the SLMs we created in Study 2. We trained and tested separate classifiers on the perplexity vectors for sentence sets with a given cardinality (i.e., 2, 4, … or 1024 sentences). For each register we randomly grouped half of the perplexity vectors for training and the other half for testing each classifier.

In addition, we used the text materials from the validation corpus obtained in Study 3. We divided each register into sentence sets containing the same number of sentences as before (2, 4, … 1024) and computed the perplexity vectors for all sentence sets. The register classifiers we trained on the Spoken Dutch Corpus materials were used to classify the sentence sets from the validation corpus. Again a classifier trained on sentence sets with a given cardinality (i.e., 2, 4, … or 1024) was used to test sentences sets with the same cardinality.

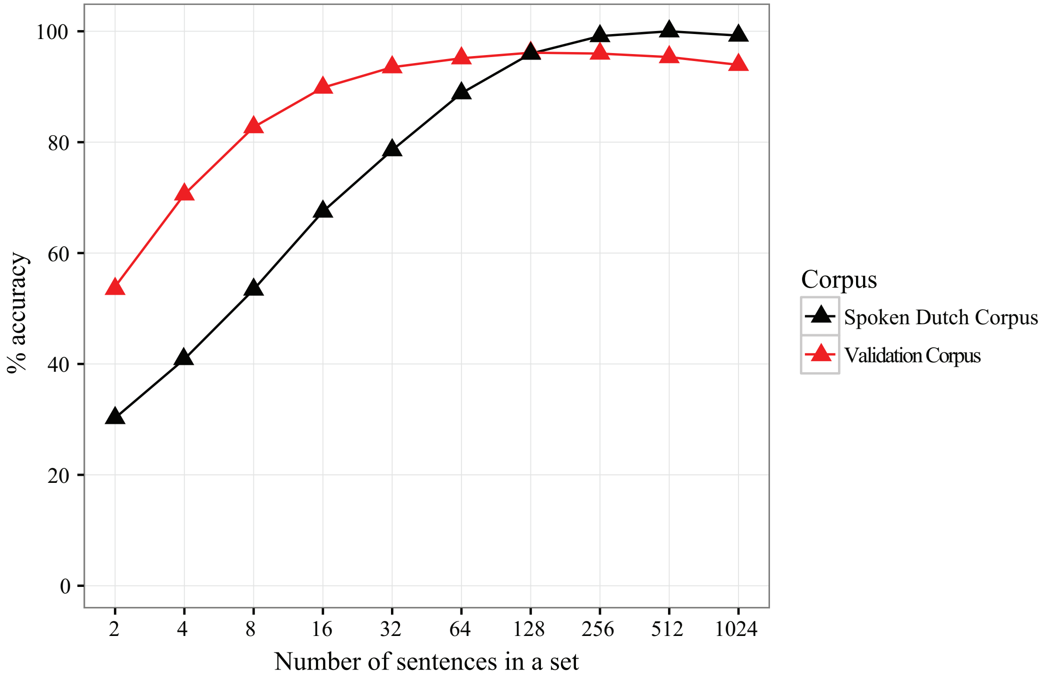
5.7.2 Results and discussion

The results, shown in Figure 2, show that the speech register classifier reaches ceiling performance (100%) when using sets of 512 sentences, while the classification of the validation corpus reaches its maximum performance (95%) with sets of 256 sentences. The accuracy results based on sets of 128 sentences are similar (92%) for the validation and Spoken Dutch Corpus. Larger sentence sets show slightly better performance for the Spoken Dutch Corpus, possibly a result of overfitting.

For the smaller sets of 2 – 64 sentences, the accuracy results for the validation corpus are higher than for the Spoken Dutch Corpus, which might come as a surprise. However, the components of the validation corpus belong to three very distinct speech registers, while the Spoken Dutch Corpus consists of 14 speech registers, including closely related registers (e.g., spontaneous conversations and telephone dialogues). This makes classification of the registers in the Spoken Dutch Corpus harder.

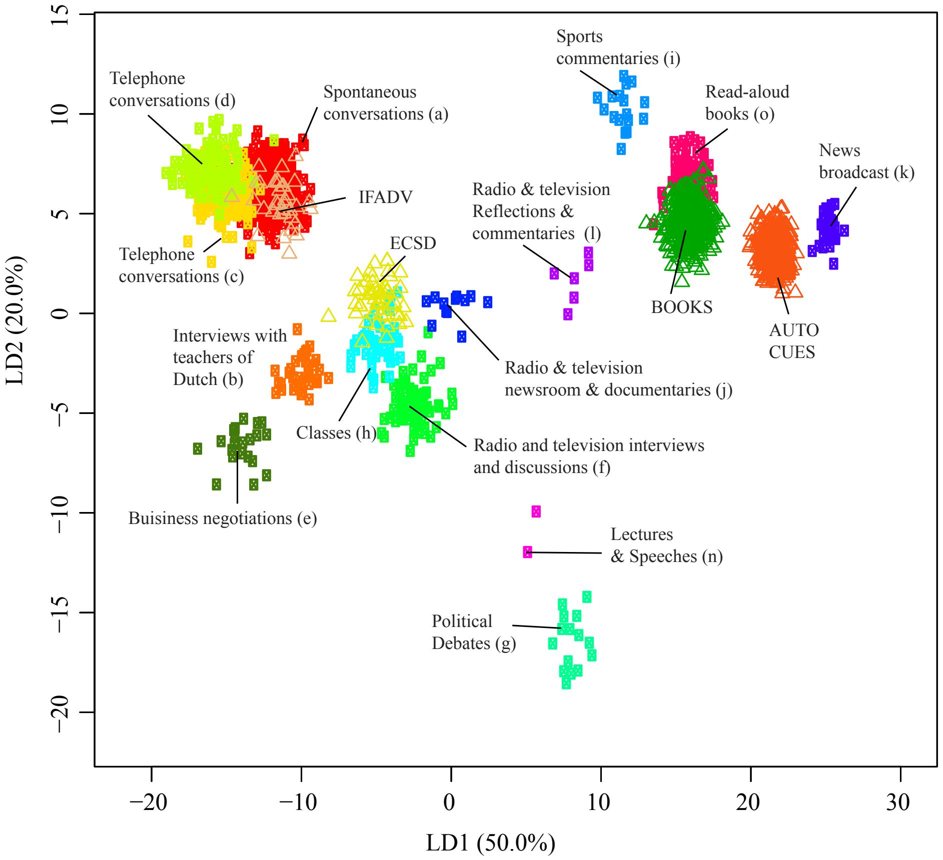
Importantly, with small sets of sentences reasonably high accuracy is achieved. For the Spoken Dutch Corpus only 64 sentences are needed for 90% accuracy and for the validation corpus only 16 sentences are needed for a similar accuracy.

Figure 2. Speech register classifier accuracy as a function of the number of sentences in a set



To investigate whether some speech registers are more similar in word predictability compared to others, we created a scatterplot based on the first two Linear Discriminants from the register classifier based on sets of 128 sentences (see Figure 3). Each point in the scatterplot is based on a set of 128 sentences. The squares represent sentence sets from the Spoken Dutch Corpus and triangles represent sentence sets from the validation corpus (with the validation corpus components shown in capitals). The scatterplot shows that the four components of the validation corpus are located closely to the counterparts in the Spoken Dutch Corpus. Most registers are separated from all other registers except for the spontaneous dialogues (components “a”, “c”, “d”), which show considerable overlap.

Figure 3. Scatterplot of all registers in the Spoken Dutch corpus and the validation corpus plotted on the first two Linear Discriminants



Taken together the results show that it is possible to classify registers with a small amount of speech (i.e., 128 sentences) with high accuracy (92%). The scatterplot and the classification errors show that the spontaneous registers are similar, while all other components in the Spoken Dutch Corpus are more distinct.

5.8. Study 5: The sentence length confound

Study 5 addressed the potential sentence length confound, because the different components in the Spoken Dutch Corpus show a wide range in the average length of sentences, which could influence perplexity scores. The classifier may therefore be based on average sentence length rather than on word predictability. We investigated this possibility by selecting a subset of our materials in such a way to reduce the difference in average sentence length between components. Furthermore, we created a classifier based on sentence length to test to what extent such a classifier can successfully distinguish between registers.

5.8.1 Procedure

We used the same materials as in Study 4, with the exception that, across all components, we only selected sentences containing 2 - 25 words. We excluded one-word sentences because they are mostly backchannels, which occur predominantly in more spontaneous speech registers and may therefore have a strong influence on overall perplexity score differences between speech registers. We excluded sentences longer than 25 words to restrict the range in average sentence length over all components.

To show the extent of average sentence length variability across registers, we tabulated, in Table 5, the average sentence length for the different speech registers in the Spoken Dutch Corpus and the validation corpus (its components are capitalized in the table). The average sentence length differs quite extensively (range 6 – 28 words on average per sentence). The range was reduced to 7 – 15 words on average per sentence in the subset restricted by sentence length. Table 5 also shows that similar registers can differ in sentence length in different corpora. For example, components “k” and “o” from the Spoken Dutch Corpus have a high average sentence length, while the validation corpus equivalents (i.e., books and teleprompt texts) do not.

Table 5. Number of words, sentences and the average sentence length across datasets

All Sentences				Sentences with 2 - 25 words		
com-ponent	word tokens	sentences	average sentence length	% of total word tokens	% of total sentences	average sentence length
a	1,745,854	303,186	6	88	70	7
b	249,844	23,835	11	67	68	10
c	738,794	129,351	6	88	68	7
d	509,960	83,514	6	87	70	8
e	136,438	179,14	8	77	69	9
f	538,795	52,274	10	68	73	10
g	217,626	110,63	20	42	68	12
h	278,749	34,496	8	83	78	9
i	130,336	124,12	10	76	94	9
j	90,614	7,620	12	73	82	11
k	285,278	21,176	14	96	98	13
l	80,081	6,210	13	72	85	11
n	61,799	2,190	28	28	54	15
o	551,441	47,944	12	79	90	10
BOOKS	1,000,042	121,256	8	93	91	8
TP*	1,000,044	107,080	9	95	95	9
IFADV	70,170	12,203	6	92	74	7
ECSD	157,106	19,197	8	71	72	8

NOTE: *teleprompt texts

We trained the speech register classifiers using the same procedure and sentence sets as described in Study 4. In addition, we created a speech register classifier based solely on sentence length. To create the latter classifier, we computed sentence length counts (counts of sentences with specific numbers of words) for each speech register in the Spoken Dutch Corpus. The histogram of sentence lengths per register represents a register-specific sentence length model analogous to the SLM used before. We created test sets for the sentence length model by computing sentence length counts for all sentence sets (of 2, 4, 8 … 1024 sentences) for both the Spoken Dutch Corpus and the validation corpus. We compared these test sets with the (speech register-specific) sentence length models using the Kullback-Leibler divergence (D_{KL}), presented in Equation 6. We used the D_{KL} as a similarity metric analogous to how we used perplexity scores.

$$D_{KL}(p||q) = \sum_{i} p(i)\, log \frac{p(i)}{q(i)} \qquad (6)$$

In Equation 6 q denotes the observed distribution (test set) and p the modelled distribution. The D_{KL} is a measure of the asymmetric difference between q and p. In our case the observed distribution q is the sentence length counts for a given set of sentences and the modelled distribution p is the sentence length counts of a given register (i.e., a component in de the spoken Dutch corpus).

We calculated the D_{KL} for each combination of a sentence set and speech register, similar to the approach used with the SLMs. We used the resulting D_{KL} similarity vectors for each sentence set to train and test register classifiers based on the Spoken Dutch Corpus. We validated these classifiers with sentence sets from the validation corpus. Classifiers for the smaller sentence sets (sets of 2,4,…,16 sentences) were not created, because of the prohibitively long computing time necessary for the calculation of all the D_{KL} values.

To quantify performance difference between word predictability and sentence length based classifiers, we calculated the average cross-entropy (ACE) for both the sentence length and word predictability classifiers (both LDA based). The cross-entropy reflects the difference between the probability the classifier assigns to each possible class (the fourteen different registers in this case) and the correct class. If a classifier assigns a high probability to the correct class, this results in a low cross-entropy. The cross-entropy is calculated according to Equation 7, where p denotes the probability of the class for the current test set (i.e., the correct class equals one and all other classes equal zero) and q denotes the probability for each class according to the classifier.

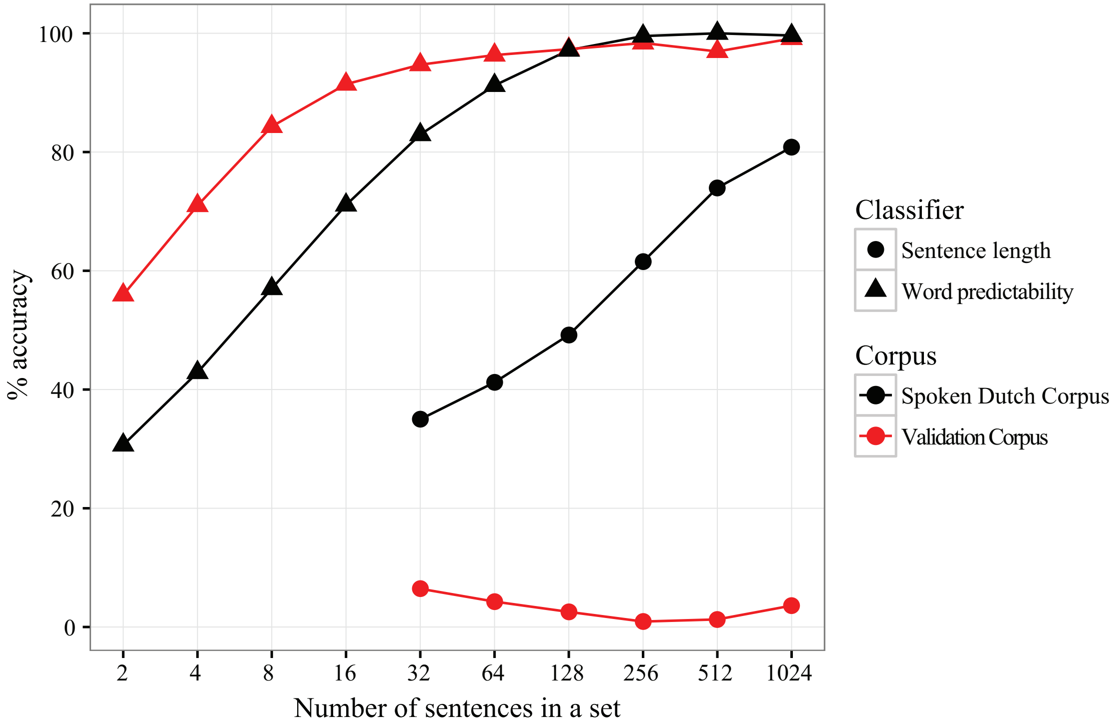
$$H(p, q) = -\sum_{x} p(x)\, log\, q(x) \qquad (7)$$

We computed the cross-entropy for all sentence sets for both the classifier based on word predictability and the one based on sentence length. Subsequently, we computed the ACE by averaging the cross-entropy across all sentence sets of specific cardinality for each classifier (based either on word predictability or sentence length) and compared the results.

5.8.2 Results and discussion

Figure 4 shows the results of the two different types of speech register classifiers, the one based on word predictability and the one based on sentence length. The results are provided for both the Spoken Dutch Corpus and for the validation corpus.

Figure 4. Classification accuracy of speech registers based on word predictability and sentence length, as a function of sentence set cardinality



The speech register classifiers based on word predictability reach ceiling performance (accuracy 100%) with sets of 512 sentences. The validation corpus reaches maximum performance (accuracy 98%) with sets of 1024 sentences. The results are comparable to the results obtained in Study 3, which were based on all sentences (see Figure 2). This is a first indication that average sentence length differences across registers do not underlie the accuracy of our classifiers assumed to be based on word predictability, because differences in average sentence length were reduced in the current experiment.

The classifiers based just on sentence length were able to classify speech registers in the Spoken Dutch Corpus with reasonable accuracy. The classification performance does not generalize to the validation corpus. Furthermore, the ACE results also show that the word predictability based classifiers outperform the sentence length classifiers (Table 6): The comparison between classifier types shows a clear advantage for the word predictability classifier. We conclude that sentence length differences between registers cannot explain the results found with the classifiers based on word predictability.

Table 6. Performance comparison of the speech register classifiers

	ACE scores for each register classifier	
Sentence set	Sentence length	Word predictability
32	1.74	0.43
64	1.49	0.23
128	1.23	0.08
256	0.94	0.01
512	0.68	0.0001
1024	0.38	< 0.0001

NOTE: ACE scores are based on the data from the Spoken Dutch Corpus. Lower scores indicate better performance.

In conclusion, the results from Study 4 show that the performance of the speech register classifier based on word predictability cannot be attributed to sentence length differences between the components in the Spoken Dutch Corpus. When we restrict the corpus to sentences of 2-25 words (to attenuate differences in sentence length between speech registers), the accuracy results are very similar to the results based on all sentences. Additionally, when we trained a classifier based on sentence length, the classifier performance did not generalize to the validation corpus and this classifier was also clearly outperformed by a classifier based on word predictability, as shown by the ACE comparison.

5.9 General Discussion and Conclusion

We conducted five studies to investigate differences in word predictability between speech registers in Dutch. We used statistical language modelling (SLM) to quantify register-specific word predictability and trained a LDA classifier on the SLM output. In fives studies we determined the best approach to create the SLMs, whether the classifier can distinguish speech registers and the robustness of the results.

The aim of Study 1 was to test the best approach to create a balanced SLM vocabulary for training register-specific SLMs. We found that there were substantial differences in word token frequency for some word types between speech registers. We used averaged reduced frequency (ARF) to filter out bursty words (i.e., words that only occur in concentrated bursts in the corpus). This approach was able to attenuate speech register vocabulary differences related to topic specificity. Future studies that investigate differences in register or genre thus best use a word selection criterion that penalizes topic-specific words. For the current study, we treated word burstiness and the topic-specificity of words as equivalent. Future research may investigate their relationship and the possibility to create a measure that more specifically targets topic-specificity, which may result in an improved inclusion criterion for a robust vocabulary.

The aim of Study 2 was to create a speech register classifier based on word predictability. We used register-specific SLMs in combination with LDA to create the classifier and found that it was able to distinguish perfectly between 14 (register-specific) components of the Spoken Dutch Corpus. This result shows that the classifier can distinguish between texts grouped into components. We conducted Study 3 to test whether the classifier is indeed sensitive to register differences between the components. We performed the procedures from Study 2 on 1000 pseudo-random variants of the components. The resulting classifiers performed poorly and could not distinguish the randomized components. This result shows that a classifier trained on perplexity scores cannot distinguish between random (heterogeneous) sets of sentences and that the perfect accuracy results obtained in Study 2 are based on systematic word predictability differences between speech registers.

Furthermore, we created a validation corpus, with materials from other corpora. The validation of our register classifier is important in light of the finding by Miller & Biber (2015), who showed that the number of word types keeps growing with the addition of new texts to a corpus, even if they are from a restricted domain (i.e., psychology textbook). It is therefore important to test whether results hold across corpora. We tested the speech register classifier (trained on material from the Spoken Dutch Corpus) on the validation corpus and found that it can also accurately classify registers in this corpus. This shows that our speech register classifier is not overfitted to idiosyncratic aspects of the Spoken Dutch Corpus. The combined results support our hypothesis that word predictability differs across speech registers.

The aim of Study 4 was to investigate the amount of text materials needed to classify the register of a text based on word predictability. We found that sets of 128 sentences are sufficient to train a classifier with a classification accuracy of 92% on the Spoken Dutch Corpus (with similar performance on the validation corpus). We conclude that register differences can be identified with a small amount (approximately 1000 words) of materials.

Figure 3 shows speech register differences captured by our classifier by means of a scatterplot based on the first two linear discriminants of the LDA. The plot illustrates that, compared to other registers, the spontaneous registers cluster together closely. This is corroborated by the confusion matrices of the classifiers; most classification errors are made between the spontaneous conversations *a* and the two telephone dialogue components "c" and "d". All other registers are well separated.

The clustering of spontaneous speech registers corresponds well with previous literature. Multiple factors contribute to the similarity of spontaneous speech registers (e.g., Leech, 2000: 697-701; Ellis, 2002: 156). For example, shared context between interlocutors reduces the need for specificity. Another contributing factor is the available processing time. Speakers only have limited time for processing and no possibility of editing, which typically results in a limited and reused repertoire. (i.e., the use of formulaic language to achieve a certain speech act; e.g., Schmitt, 2010: 8-12). These factors work together to produce spontaneous speech registers that are similar, as is attested by the result from our study.

Previous research reported a distinction between informational and involved dimension in language use (Biber, 1988, 1995) with factor analysis. Our cluster of spontaneous registers could be interpreted as registers that use involved language; however, the other registers do not cluster together in an informational counterpart. This could be because instead of using comparatively small sets of lexico-grammatical features, we used large sets of *n*-grams with statistical language modelling. It is possible that a large feature set such as *n*-grams is sensitive to differences between registers that use informational language, which would explain why we did not find a cluster of informational registers. Our results suggest that register differences are not exclusively related to lexico-grammatical features, because word *n*-grams reveal subtle but robust differences across registers. We propose that register analysis based on lexico-grammatical features, could be fruitfully complemented by this new approach.

Speech registers differ in the average length of sentences (see Table 5). In Study 5 we tested whether sentence length influences the performance of the speech register classifier. We used a subset of the corpus with reduced differences in average sentence length between registers. We found results similar to those in Study 4, which suggests that sentence length differences cannot account for the performance of the speech register classifier. Furthermore, we trained a register classifier based solely on sentence length, which could distinguish between speech registers to some extent, similar to Wiggers & Rothkrantz (2007) findings. However, the classifier based on sentence length was clearly outperformed by the classifier based on word predictability. Additionally, the performance of the classifier based on sentence length did not generalize to the validation corpus, indicating that sentence length is not a robust basis for a register classifier. The results showed that the classifier performance is best explained by word predictability differences and cannot be explained by sentence length differences between registers.

Our results have implications for studies investigating word predictability in relation to language comprehension. Given the sensitivity of readers and listeners to the predictability of words (e.g., Smith & Levy, 2013), it is plausible that they are also sensitive to register-specific differences in word predictability. In addition, Study 4 showed that the differences in word predictability between registers are already substantial in only 128 sentences (i.e., approximately 5 minutes of speech materials). It is therefore plausible that human listeners can notice these substantial differences as well. Future research has to show whether readers and listeners adapt their expectations based on the wider context of situation of use when comprehending written or spoken language.

Our results also raise important questions about the nature of lexical representations. For example, what type of lexical representation allows speakers to systematically adapt their word use to the appropriate register? Are different word predictabilities stored for every speech register and if so, how many registers are lexically represented? If listeners use register-specific word predictability to tune their anticipations of upcoming words, the question is again how these register-specific word predictabilities are mentally represented.

The study shows that the combination of register analysis and text classification with the aid of statistical language modelling provides important new insights about registers and the requirements needed for speech processing and the mental lexicon. Importantly, the study extends the finding that situation of use determines language variation, by reporting differences across speech registers in word predictability.

Notes

1. SRILM release 1.5.12, http://www.speech.sri.com/project

2. A count-of-count list lists the number of *n*-grams occurring a specific number of times (i.e., there are 15 unigrams that occur 3 times) in the training data.

Appendix 1. Precision, recall and f1 scores for the speech register classifier in Study 2

Component	precision	recall	f1
Spontaneous dialogues	1.00	1.00	1.00
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues (platform)	1.00	1.00	1.00
Spontaneous telephone dialogues (minidisc)	1.00	1.00	1.00
Business negotiations	1.00	1.00	1.00
Radio and television interviews & discussions	1.00	1.00	1.00
Debates, discussion and meetings	1.00	1.00	1.00
Classes	1.00	1.00	1.00
Spontaneous radio & television commentaries	1.00	1.00	1.00
Radio & television newsroom & documentaries	1.00	1.00	1.00
News broadcast on radio & television	1.00	1.00	1.00
Reflections & commentaries broadcast	1.00	1.00	1.00
Lectures and speeches	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

Appendix 2. Precision, recall and f1 scores for the speech register classifier tested on the validation materials in Study 3

Component	precision	recall	f1
Spontaneous dialogues	1.00	0.85	0.92
News broadcast on radio & television	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

Appendix 3. Precision, recall and f1 scores for the speech register classifier tested on the validation corpus materials in Study 4

Component	precision	recall	f1
Spontaneous dialogues	1.00	0.61	0.76
News broadcast on radio and television	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets.

Appendix 4. Precision, recall and f1 scores for the speech register classifier in Study 4 tested on the Spoken Dutch corpus materials

Component	precision	recall	f1
Spontaneous dialogues	0.98	0.96	0.97
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues (platform)	0.90	0.89	0.90
Spontaneous telephone dialogues (minidisc)	0.81	0.90	0.85
Business negotiations	1.00	1.00	1.00
Radio and television interviews & discussions	1.00	1.00	1.00
Debates, discussion & meetings	1.00	1.00	1.00
Classes	1.00	0.99	1.00
Spontaneous radio & television commentaries	1.00	1.00	1.00
Radio & television newsroom & documentaries	0.97	1.00	0.98
News broadcast on radio & television	1.00	1.00	1.00
Reflections and commentaries broadcast	1.00	1.00	1.00
Lectures & speeches	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets.

Appendix 5. Precision, recall and f1 scores for the speech register classifier (based on word predictability scores) tested on the validation corpus materials in Study 5

Component	precision	recall	f1
Spontaneous dialogues	1.00	0.73	0.84
News broadcast on radio & television	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets.

Appendix 6. Precision, recall and f1 scores for the speech register classifier (based on sentence length) tested on the validation corpus materials in Study 5

Component	precision	recall	f1
Spontaneous dialogues	0.75	0.19	0.30
News broadcast on radio & television	0.00	0.00	0.00
Read-aloud stories	0.01	0.01	0.01

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets.

Appendix 7. Precision, recall and f1 scores for the speech register classifier (based on word predictability) tested on the validation corpus materials in Study 5

Component	precision	recall	f1
Spontaneous dialogues	0.99	0.98	0.99
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues (platform)	0.94	0.91	0.93
Spontaneous telephone dialogues (minidisc)	0.86	0.94	0.90
Business negotiations	1.00	1.00	1.00
Radio and television interviews & discussions	1.00	1.00	1.00
Debates, discussion and meetings	1.00	1.00	1.00
Classes	1.00	1.00	1.00
Spontaneous radio & television commentaries	1.00	1.00	1.00
Radio and television newsroom & documentaries	1.00	1.00	1.00
News broadcast on radio & television	1.00	1.00	1.00
Reflections & commentaries broadcast	1.00	1.00	1.00
Lectures & speeches	1.00	1.00	1.00
Read-aloud stories	1.00	1.00	1.00

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets.

Appendix 8. Precision, recall and f1 scores for the speech register classifier (based on sentence length) tested on the validation corpus materials in Study 5

Component	precision	recall	f1
Spontaneous dialogues	0.77	0.56	0.65
Interviews with teachers of Dutch	0.36	0.38	0.37
Spontaneous telephone dialogues (platform)	0.40	0.47	0.44
Spontaneous telephone dialogues (minidisc)	0.23	0.29	0.26
Business negotiations	0.24	0.48	0.32
Radio and television interviews & discussions	0.55	0.47	0.51
Debates, discussion & meetings	0.84	0.87	0.85
Classes	0.36	0.36	0.36
Spontaneous radio & television commentaries	0.18	0.51	0.27
Radio & television newsroom & documentaries	0.12	0.53	0.20
News broadcast on radio and television	1.00	0.99	0.99
Reflections & commentaries broadcast	0.15	0.30	0.20
Lectures and speeches	0.67	0.67	0.67
Read-aloud stories	0.66	0.37	0.48

NOTE: The scores are provided for the classifier trained and tested on 128-sentence sets

Speech register influences listeners’ word expectations

Chapter 6

Abstract

We investigate the influence of speech register on predictive language processing using the N400 effect. Participants listened to long stretches (4 – 15 minutes) of naturalistic speech from different registers (dialogues, news broadcasts, and read-aloud books), totaling approximately 50,000 words, while the EEG signal was recorded. We estimated the surprisal of words in the speech materials with the aid of statistical language models. Word surprisal was estimated in such a manner that it reflected different processing strategies; generic, register-specific, or recency. 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, the comparison between processing strategies shows that adaption to speech register cannot merely be explained by recency effects; instead, listeners adapt their word anticipations to the presented speech register.

6.1 Introduction

Human perception of sensory input involves more than passive registration. A rich body of research (e.g., Bar, 2007; Friston, 2005, 2012) 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 the different varieties. 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 use the N400 effect to investigate whether listeners adapt their word expectations as a function of the speech register they are listening to.

6.1.1 Register variation

Below we present three examples to show a range of registers: chatting with friends (1), a reporter providing coverage of a news event (2), 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 (1) allows for more interaction and opportunity for clarification and is produced in real time, with no preparation or opportunity to revise the utterance and is typically more personal, leading to, for example, disfluencies, a lower type-token ratio and a frequent use of pronouns. News reportage (2) is typically prepared and intended to convey information about a certain event, which results in a more frequent use of time and place adverbials as well as proper nouns. A novel (3) is written without interaction between writer and readers, which allows the writer to revise and refine their language use, affording a rich vocabulary and complex sentence structure.

Many studies found evidence for systematic differences in patterns of language use between registers (see Biber & Conrad, 2009, for an overview). For example, word choice differs between registers (Biber, 1999); the use of *like* in (1) is typical for informal conversation (Barbieri, 2005). The usage of grammatical constructions also varies between registers (Staples et al., 2015), for example, the retention of the complementizer in *that*-clauses, as in (4), 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

This lexical and grammatical variation results in register-specific word co-occurrence statistics. In Chapter 5 we 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 question whether language perceivers adapt their word expectations based on the register of the language input.

6.1.2 Predictive language processing and the N400 effect

Evidence for predictive language processing is well established in the literature (see Elman, 2009; Huettig, 2015; Kuperberg & Jaeger, 2016 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 gaze in anticipation to a picture of a cake (among multiple objects) when they hear *The boy eats* compared to *The boy moves* (Altmann & Kamide, 1999).

The N400 effect also provides important evidence for anticipatory language processing (see Kutas & Federmeier, 2011 for an overview). The N400 is a negative deflection of the event related potential (ERP), which peaks 400 milliseconds after word onset at central posterior electrode sites. When participants read short sentences, such as (5), with occasionally an anomalous final word, as in (6), the semantically incongruous word *socks* results in a more negative deflection of the ERP compared to the congruent word *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. The unexpected but semantically related word *pines* results 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]

The N400 also provides strong 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 were expecting the following word to start with a consonant (DeLong et al., 2005). 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]

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 this interpretation, there will (almost) always be prediction error since not all probability is assigned to a single word. For example, it is well attested that even with highly constraining sentences (e.g., Federmeier & Kutas, 1999), words semantically related to the expected word show an attenuated N400 compared to unrelated words. This interpretation of prediction is congruent with the predictive coding framework (e.g., Friston, 2005, 2018).

6.1.3 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 provides a measure of how expected that word is. 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 (Kuperberg 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 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 with Equation 1.

$$\hat{P}(W_i|context) = P(W_i|W_{i-n}, \dots, W_{i-1}) \qquad (1)$$

Whereby 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. 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. It can informally be thought of as the 'unexpectedness' of a word. Frank et al. (2015) used statistical language modelling to estimate word surprisal for all content words in sentences from unpublished 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.

6.1.4 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; only *brave knight* and *benevolent sorcerer* switch position in (9 & 10). The sentence *He quickly reached for a ...* does not constrain in favor of either *sword* or *lance.* The preceding sentence in (9) favors *sword* while in (10) it does not. In (9) the unexpected word *lance* resulted in an N400 effect compared to the expected *sword*, whereas in (10) this did not occur (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.

6.1.5 Current study

In the current study we investigate whether listeners' word anticipations depend on speech register information. We use long stretches (4 – 15 minutes) 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 of 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 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 should best predict the N400 amplitude.

The 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 6.2.2).

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 6.2.1.1) 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 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[4].

In summary, in this study we test whether listeners anticipate words in long stretches (4 – 15 minutes) 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.

6.2 Methods

6.2.1.1 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.

6.2.1.2 Materials

The stimuli for the EEG experiments consisted of audio recordings of Dutch speech from different registers, with approximately 90 minutes 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 6.2.2.1. The books and the news speech materials were extracted from the Spoken Dutch Corpus, the dialogues were extracted from IFADV.

[4] This dataset will be made freely available as the Dutch EEG Speech Register Corpus, DESRC for short.

The IFADV dialogues materials used for the EEG experiment consisted of six dialogues of 15 minutes each. All dialogues were 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). The books experiment consisted of 12-minute excerpts from seven read-aloud Dutch novels. 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 seconds 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 allows 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.

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)	avg. 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

6.2.2 Estimating generic, register-mismatch, recency and register-specific word surprisal

6.2.2.1 Training and test materials

To train statistical language models 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. 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').

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 (approximately 1 billion words).

6.2.2.2 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 .3 for both news and books and .13 for the dialogues model.

Finally, we created a recency-based SLM, henceforth cache 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 2 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 mismatching register material, and the cache model performs better still. The book SLM performed worse on the materials used in the EEG experiment compared to the testing materials, indicating a discrepancy between the training and test materials and the language materials used in the EEG experiment. However, this model still improved compared to the generic SLM (i.e., 714 versus 1736).

Table 2. 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 in bold face, second best underlined.

SLM	dialogues	news	books
generic	3943 (4340)	807 (1312)	1736 (1834)
cache	**328** (453)	**287** (325)	**343** (371)
dialogues	<u>454</u> (460)	723 (955)	722 (923)
news	1188 (1384)	<u>314</u> (327)	828 (639)
books	1463 (1775)	601 (623)	714 (<u>417</u>)

6.2.2.3 Word surprisal estimation

To estimate word surprisal, we used the generic, cache and register-specific SLMs described in Section 6.2.2.2. The different SLMs were used to estimate the surprisal of each word in the experimental speech materials. We used the generic and cache SLM to estimate *generic* and *recency-based* word surprisal respectively. The register-specific SLMs were used to compute *register-specific* word surprisal for 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[5]. We refer to this as *register-mismatch* word surprisal.

6.2.3 Procedure

Participants came to the lab on three separate occasions. Consecutive visits were separated by minimally a week. 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 minutes of speech from a specific register (i.e., dialogues, books, or news), 270 minutes in total. 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 audio fragment was played to check the volume (speech in the

[5] We used the following mismatch pairs (SLM-materials): books-news, news-dialogues, news-books

audio fragment corresponded to the register of the experiment). 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 minutes. 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.

6.2.4 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 artefacts (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.

6.2.5 Preprocessing

The data was 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 whereby all proposed artefacts were manually checked. 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 a channel 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 milliseconds relative to word onset, for each content word (i.e., nouns, verbs, adverbs and adjectives) in our dataset that did not overlap in time with other words (only relevant in the dialogues experiment) and was not the first word of a sentence. The second exclusion criterion - the removal of the first word of a sentence - was applied to lower the correlation between word surprisal and word frequency (a covariate in our statistical model, see below). Furthermore, we excluded those words which overlapped with artefactual EEG data or if the signal exceeded ± 75 µV in the previously defined time window of the word. Finally, we excluded all data from nine participants because less then 40% of the data remained after artefact removal. Across all experiments, these steps resulted in a dataset of 600,276 word epochs. This dataset is part of the Dutch EEG speech register corpus, see Chapter 2 for further details.

6.2.6 Analysis

Based on previous literature, we defined the N400 amplitude as the average of the channel set C3, C4, Cz, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, O2 within the time window 300 – 500 milliseconds 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 over the time window -150 – 0 milliseconds (relative to word onset) and the same channel set.

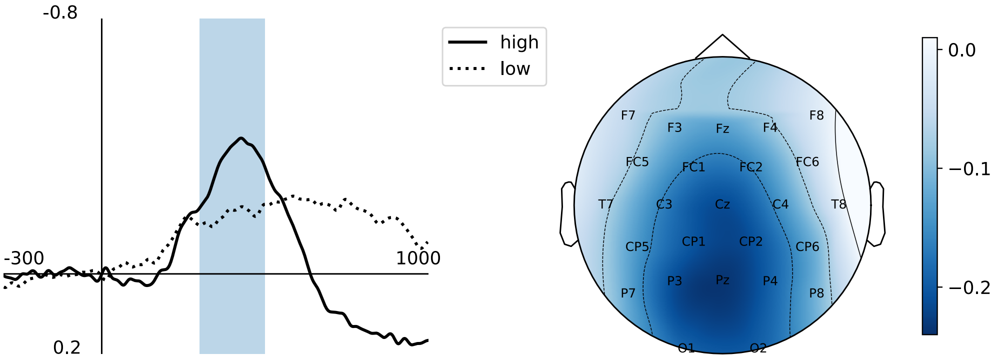
We estimated several linear mixed effect (LME) models (Bates et al., 2015) with the statistical package R (R Core team, 2015) to predict the N400 amplitude. We first estimated a *null* LME model with the following standardized covariates: the aforementioned *baseline*, the *log word frequency* (based 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.

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 his resulted in convergence issues.

6.3 Results

Model comparison with the anova likelihood-ratio test revealed that the LME model with generic word surprisal improved compared to the null model χ^2 (3) = 38.73, p < .001. The N400 amplitude is more negative with increasing values of word surprisal (see Figure 1).

Figure 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 milliseconds 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.



Subsequently we compared the generic, register-mismatch, recency 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 3, left). The recency-based model performed better than the generic model, while the register-mismatch model performed similar to the generic model.

Table 3. Comparison between {generic, register-mismatch, recency} and register-specific word probability estimates 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 is a better fit compared to the register specific word surprisal. (left) Compares register-specific LME model to the generic, mismatch and recency models on all experimental data. (right) Compares the register-specific LME model with both the null and generic model on subsets of the experimental materials (dialogues, news and books).

LME model	relative likelihood (Δ AIC) register-specific
generic	p < .001 (44)
register-mismatch	p < .001 (44)
recency-based	p < .001 (25)

<table>
<tr><th>experimental materials</th><th>relative likelihood (Δ AIC) (null | generic) vs register-specific</th></tr>
<tr><td>dialogues</td><td>p < .05 (8) | p < .001 (15)</td></tr>
<tr><td>news</td><td>(-7) | (-1)</td></tr>
<tr><td>books</td><td>p < .01 (13) | p < .01 (10)</td></tr>
</table>

In the register-specific LME model (see Table 4) the interaction term for the news materials and word surprisal has a t-value of 8.03. To further investigate this interaction effect, we split the data according to register (dialogues, books and news) and fitted LME models to each subset. Table 3 (right) shows the result of the comparisons between the register-specific LME model and both the generic and null model for the dialogues, books and news materials. For both the book and dialogue materials, the register-specific model outperformed both the null and generic model, while for the news materials the register-specific model did not improve compared to either the null or the generic model (see also Figure 2).

Figure 2. Grand average plots of the ERP response averaged over all content words, participants and channel set 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.

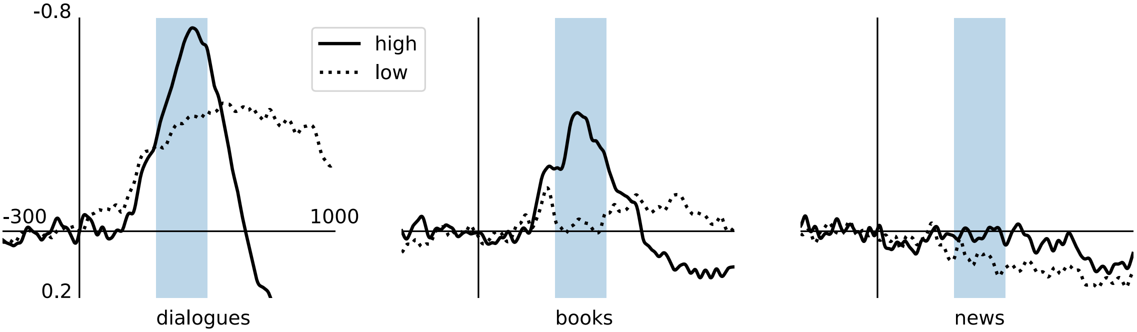


Table 4. Overview of the fixed effects of the linear mixed effect model with the N400 as dependent variable. The fixed effect names, the beta (**β**), the standard error (SE **β**) and the t-value (t) are reported. The predictor of interest (register-specific word surprisal) is bolded.

fixed effect	β	SE β	t
intercept	-0.409	0.046	-8.97
baseline	5.407	0.009	603.44
log word frequency	0.171	0.029	5.95
experiment news	0.400	0.028	14.36
experiment books	0.268	0.025	10.61
surprisal	-0.149	0.020	-7.38
word duration	0.075	0.017	4.41
word position in sentence	0.184	0.010	18.72
exp. news: **surprisal**	0.212	0.026	8.03
exp. books: **surprisal**	0.022	0.026	0.87

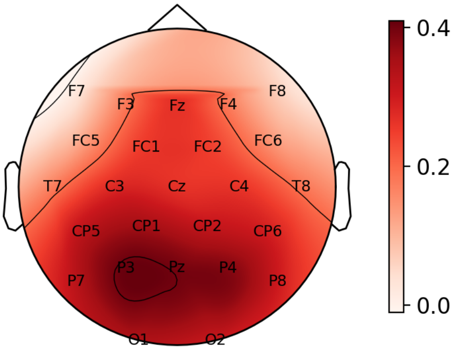
6.3.1 Exploratory analysis

The results informed us about an effect we did not foresee. The grand average plots revealed an unexpected post N400 positivity (PNP), see Figures 1 & 2, which was most pronounced for the dialogue materials, and absent for news. Based on the time course information, topographic plot Figure 3, and observations in the literature (e.g., Van Petten & Luka, 2012), we decided to analyze the average amplitude over the channel set (CP5, CP1, CP2, CP6, P3, Pz, P4) and time window 600 – 900 milliseconds. We excluded the news materials, since these showed no effect.

We used a similar *null* LME model as before, however, this time the PNP average was the dependent measure. We created a second LME model (see Appendix A, Table 1 for an overview of the fixed effects) by adding generic word surprisal as predictor of interest to the null model. We considered a random slope between participant and word surprisal but did not include it due to convergence issues.

Model comparison with the anova likelihood-ratio test revealed that the model with word surprisal outperformed the null model χ^2 (1) = 7.39 (which would amount to $p < .01$ in a confirmatory testing analysis). In addition, we tried an interaction between experiment and surprisal, however, this did not show an improvement. Also, replacing generic word surprisal by register-specific word surprisal did not further improve the model. The exploratory analysis indicates that posterior PNP amplitude is more positive with increasing values of word surprisal, i.e., when words were less expected.

Figure 3. Topographic difference plot of the post-N400-positivity between content words with the highest tertile generic surprisal values versus words in the lowest tertile in the time window 600 – 900 milliseconds.



6.4. Discussion

In the current study, we recorded the EEG signal from participants who listened to long (4 – 15 minutes) stretches of natural speech sampled form different speech registers: dialogues, news broadcasts, and read-aloud books. The speech materials were analyzed with the aid of statistical language models (SLM) to estimate word surprisal. We found that the N400 amplitude was more negative for words with high surprisal (i.e., unexpected words) and was best predicted by register-specific word surprisal estimates.

To investigate the influence of speech register on prediction in speech comprehension, we compared 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 provides 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 solely by utilizing recent 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 3 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 relevant for the question whether prediction occurs during normal language processing (Huettig, 2015). In our experiments, we used long stretches (4 – 15 minutes) 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. The current result shows that listeners do indeed engage in predictive language processing while listening to natural everyday speech (without artificially constraining sentences). This result is in line with the results reported with reading studies by Smith & Levy (2013) and Frank et al. (2015).

We found an unexpected difference between the speech registers: we did not observe an N400 effect for the news broadcast speech materials (see Figure 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 results on dialogues and books (see Table 2), indicating that the SLMs could predict upcoming words in the news materials with performance similar to 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.

Unexpectedly, we found a post N400 positivity (PNP), also referred to as the late positivity complex or the semantic P600 (Van Petten & Luka, 2012). Exploratory analysis revealed a *posterior* PNP effect (see Figure 3). Van Petten & Luka (2012) hypothesized that two late positivities can be distinguished based on the topography: an anterior positivity, mostly related to the difficulty of integrating unexpected but plausible words, and a posterior positivity, related to the processing of implausible words, such as *summer* in *He pounded the nails with a book/summer.* DeLong et al. (2014) found evidence supporting this hypothesis, whereby implausible words show a more posterior effect compared to a more anterior effect for unlikely but plausible words (e.g., *book* in the previous example).

We found a posterior positivity, which is puzzling, since this would indicate that our materials contained many implausible sentence continuations. This is unexpected because these speech materials were sampled from natural language use in books and dialogues (we excluded the news materials for the PNP analysis). It is therefore unlikely that there were many anomalous or implausible sentences. A possible explanation could be the presence of counterfactual stories in our experimental materials. Kolk et al. (2003) found that a (semantic) P600 can be elicited with a sentence such as *The mouse chased the cat.* In general, we can say that this was not the case for our experimental materials. Unfortunately, we cannot analyze this in detail, since we do not have plausibility ratings for our materials. Another explanation for the posterior PNP could be that the distinction between anterior and posterior late positivities is not ironclad and that posterior late positivity also occurs in response to an unexpected but plausible continuation. Further investigation is necessary to distinguish between the alternative explanations.

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 (effects can flip, because the variance can be ascribed to either of the variables). The reason for the unexpected 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 would indicate that listeners do not merely use recent context to adjust expectations, and would imply that registers are represented 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). Second, does speech register provide 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 found with sentences 9 & 10 (Section 6.1.4) 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?

6.5 Conclusion

We analyzed ERPs elicited with spoken words from long stretches (4 - 15 minutes) 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.

Appendix A. Overview of the fixed effects of the linear mixed effect model with the PNP as dependent variable. The fixed effect names, the beta (B), the standard error (SE B) and the t-value (t) are reported. The predictor of interest is bolded.

fixed effect	B	SE B	t
intercept	0.099	0.063	1.57
baseline	4.491	0.012	379.63
log word frequency	0.140	0.042	3.33
experiment books	0.085	0.030	2.82
generic surprisal	0.074	0.020	3.66
word duration	0.169	0.023	7.50
word in sentence	0.285	0.013	21.40

It is difficult to make predictions, especially about the future

Niels Bohr

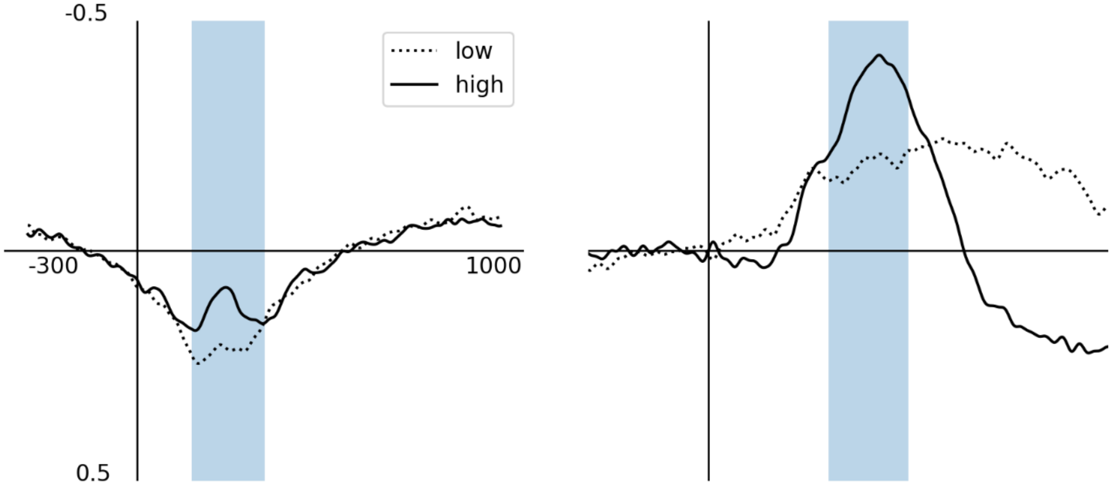
Discussion

Chapter 7

In this thesis, I investigated the role of predictive language processing by modelling event-related potentials (ERP) based on information-theoretic measures that capture different aspects of predictability. In this manner, I investigated listeners' anticipations of words and word forms based on the preceding context and speech registers.

The electroencephalography results of this thesis are summarized in the plots presented in Figure 1. On the left-hand side, results are shown from Chapters 3 and 4, where I investigated auditory word form anticipation based on preceding words. The plot shows a grand average of the event-related potential (ERP) averaged over words, participants and EEG channels. The ERP data was split in tertiles based on the cross-entropy value for each word, whereby averages were taken across words with low and high cross-entropy values and plotted with the dotted and solid line, respectively. The plots and statistical analysis (Chapter 4) reveal that the phonological mismatch negativity (PMN) is more negative with increasing values of cross-entropy. As I detail in the following sections, this finding is in line with the idea that listeners anticipate auditory words forms while listening to natural speech.

Figure 1. Grand average plot of the ERP response averaged over words, participants and channels. The data was split in tertiles, based on cross-entropy (left) and surprisal (right).



On the right-hand side of Figure 1, results are shown from Chapters 5 and 6, where I investigated the influence of speech register on predictive language processing. Again, the grand average of the ERP response (averaged over content words, participants and EEG channels) is plotted but now for different values of word surprisal (the unexpectedness of a word). The data was split in tertiles based on word surprisal values. Again, the low and high valued words were averaged and plotted with the dotted and solid line, respectively. The plot and statistical analysis (see Chapter 6) reveal that higher surprisal values result in a more negative N400 amplitude. In addition, the comparison of several processing strategies (see Chapter 6) suggests that human listeners use the wider context of speech register to adapt their word anticipations.

In the following sections, I discuss the results presented in this thesis per research question as formulated in the Introduction (Section 1.4). First, I discuss the EEG data corpus and the convolutional neural network that was trained and applied to clean 200 hours of EEG data. Second, I discuss the mismatch measure and the study investigating auditory word form anticipation with the PMN. Third, I discuss the corpus study investigating word predictability differences between speech registers and the EEG study investigating the influence of register on listeners' word anticipations. Fourth, I give a rough sketch of possible theoretical implications of the results presented in this thesis.

7.1 How can we investigate predictive language processing with event-related potentials (ERP) evoked by words in long stretches of natural speech?

It is possible to investigate human speech processing of long stretches of natural connected speech with ERPs. The secret is to collect a lot of data. I recorded approximately 200 hours of EEG data and extracted approximately 1.5 million word epochs from this data. Based on two predictability measures, I successfully modelled the amplitude of the N400 for content words and the phonological mismatch negativity (PMN) for all words. I will elaborate on these results in Section 7.2 and 7.3. In the current section, I focus on big data collection for ERP research and the use of naturalistic language materials in relation to predictive language processing.

7.1.1 EEG data cleanup and the convolutional neural network artefact classifier

For ERP research it is important to clean the EEG data prior to statistical analysis. Typically, recorded EEG materials must be cleaned by removing trials (i.e., word epochs) contaminated by artefacts. This can be accomplished with manual or automatic approaches or a combination of both. Manual artefact rejection is more precise; however, due to the prohibitively large number of epochs, a completely manual approach was not feasible. A drawback of automatic methods is that they are less precise. For example, it is possible to automatically threshold the EEG data on a specific value, for example, ± 75 µV, and to exclude all EEG data which exceeds that threshold. The drawback of this method is that it rejects usable data, for example, data contaminated by eye-blinks, which can easily be attenuated with independent component analysis (ICA).

Experimental paradigms with well-defined short trials achieve relatively clean EEG data by asking participants to refrain from blinking during those short trials. In my research this option was unavailable because participants listened continuously to the experimental materials and could not be expected to refrain from blinking for up to fifteen minutes at a time. As a consequence, my EEG dataset was more heavily contaminated by eye-blink related activity compared to data from classic style EEG experiments. It was therefore important to be able to attenuate eye-blink related activity without rejecting all this data. To deal with this problem for a large amount of data, I developed a more sensitive algorithmic approach to clean the EEG data by training and applying a convolutional neural network (CNN) to automatically classify EEG artefacts.

To create training material for training the classifier, I manually annotated 60 hours of EEG materials (i.e., about 30 percent) by marking the start and end boundaries of EEG data that I judged to be contaminated by artefacts. For my annotations, I distinguished between two artefact annotation types: *channel* and *stretch* artefacts. The channel artefacts were indicated per channel when only one or a few channels showed artefacts, while the stretch artefacts were marked for all channels when most channels showed artefacts. I distinguished between channel and stretch artefacts to achieve a balance between specificity and speed in the annotation process. The channel artefact annotations allowed for the removal of channels with poor signal quality without removing the data from the other channels, thereby saving data. The stretch artefact annotations speeded up annotation time, because all channels could be annotated simultaneously (instead of individually). Still, manual artefact annotation is time-consuming work: annotating 60 hours of EEG data for both the channel and stretch artefacts took approximately 300 hours of work.

The CNN classifier was trained and tested on 1-second chunks of EEG data labeled *clean* or *artefact*. These labels were based on the manual annotations. I found that the CNN classifier clearly outperformed the thresholding procedure mentioned before (see Chapter 2, section 3.5). To further ensure the quality of the EEG dataset, all automatically annotated artefacts were manually checked. Correcting automatically generated artefacts speeded up the annotation work compared to complete manual annotation. This is because artefacts tend to cluster, and I could therefore skip long stretches of clean EEG data.[6]

The performance evaluation (see Chapter 2, Table 4) revealed that the current CNN artefact classifier is best used in combination with manual post-correction to ensure the quality of the resulting annotation. The results (see Chapter 2) show that the use of a CNN artefact classifier provides a good approach for cleaning a large amount of EEG data. However, there is room for improvement. In the following paragraphs,

[6] This correction approach (skipping sections classified as clean) is a tradeoff between speed and accuracy. I trained the classifier to err on the side of classifying something as artefact to reduce the chance of missing them, however, there is still a chance that artefacts in clean sections were missed. Since artefacts tend to cluster, I think overtraining on artefacts and skipping clean sections is the best approach, because there are sharply diminishing returns for the extra effort to remove the last remaining artefacts embedded in clean sections.

I discuss challenges and propose improvements for CNN based artefact classifiers related to generalizability, artefact definition and CNN architecture.

7.1.1.1 Generalizability

The real benefit of a CNN artefact classifier would lie in its generalizability, using a pre-trained classifier to classify newly recorded data (preferably without the need for manual correction). Currently, the classifier is trained and utilized on the same dataset and can probably only classify EEG data from the same or (maybe) a very similar EEG channel layout. This is because I used windowed EEG data, whereby the input data consisted of one second of data from all EEG channels. The consequence of this setup is that the CNN artefact classifier will not be able to process EEG data with a different number or a different set of EEG channels.

Generalizability of the classifier can be improved by training solely on channel level annotations. As mentioned in the previous section, in the current approach I annotated stretch and channel artefacts. I distinguished between stretch and channel artefacts to speed up annotation. Annotating all channels separately (i.e., without stretch artefact annotations) would dramatically increase annotation times, because for every stretch artefact every channel would need to be annotated.

By using annotations only on channel level, a more generalizable classifier could be trained because, in this case, the input to the CNN model could consist of a target channel with a set of *k* nearest neighboring channels. This data structure (a channel with a set of neighboring channels) can be created irrespective of the channel layout or the number of channels. However, there are two potential problems. First, the channel data is likely to systematically differ across channels, depending on their location (e.g., anterior channels will show more eye-related activity compared to more posterior channels). Second, the correlation between the target channel and its neighbors will differ for low compared to high density recording setups (e.g., 32 versus 256 channels). It is an empirical question whether these problems might frustrate the goal of a generalizable artefact classifier.

The complete channel level approach would also provide a more fine-grained insight into the quality of the EEG data and more flexibility with respect to removing channels or stretches. The annotations provided with the EEG speech register corpus (DESRC) could be extended to create such a classifier. However, to test whether such a classifier genuinely generalizes, other EEG datasets need to be annotated as well.

7.1.1.2 Artefact definition

The windowed EEG data were labelled as either *clean* or *artefact* based on the amount of overlap with manually annotated artefacts; windows with 50% or more overlap were labelled as artefact and all other windows were labelled clean.

In retrospect, this labelling scheme was not ideal. The windows were created in a sliding window fashion whereby each window started at the datapoint following the start point of the previous window and neighboring windows overlapped by 99 %. Each window created in this manner was randomly assigned to a training, test, or validation set. However, since neighboring windows overlapped by 99%, the separation between training, test and validation was not as strict as it could have been. Very similar (but not identical) windows could be incorporated in the training and test sets.

An alternative approach would be to label a window based on the status of the central column of the window (i.e., at half a second). The window would be labelled artefact if the central column falls within an artefact annotation and labelled clean otherwise. This scheme clearly separates the windows and prevents contamination between the classifier training and test data. Furthermore, labelling clean and artefact windows in this manner would also better reflect the translation of window classifications to start and end boundaries in the EEG signal. In Chapter 2, the start and end boundaries were automatically generated by taking the middle timepoints of the first and last window, in a sequence of artefact windows. This approach was most suitable, due to the overlap between neighboring windows. The new labelling scheme would more closely match this translation between window classification and start and end boundary annotation, thereby potentially further improving the automatic annotations.

7.1.1.3 CNN architecture

The present classifier used a simple CNN architecture inspired by Krizhevsky et al. (2012) and Schirrmeister et al. (2017). The main architecture is similar to a standard image classification CNN. An important difference is the stepwise processing of the time and channel dimensions (Schirrmeister et al., 2017) compared to the one-step processing of the combination of width and height of images. Without this adaptation, the model never improved much beyond chance level.

The automatic artefact annotations of the CNN artefact classifier needed to be improved with manual correction. Ideally, the classifier would work wholly or mostly autonomously. One weakness of the current CNN classifier is the very limited context awareness in the time dimension (i.e., 1 second). This is suboptimal, because artefacts can occur over much longer timescales and tend to cluster together.

I hypothesize that a better artefact classifier should be able to deal with the time dimension at multiple levels of granularity and propose that an architecture inspired by Wavenet (Oord et al., 2016) could improve the artefact classifier performance. Wavenet utilizes *dilated causal convolutional layers* to achieve sensitivity over different temporal scales (i.e., increasing the receptive field in the time dimension of the model).

Wavenet is a deep neural network developed to perform speech synthesis.[7] It is a *generative* model which attempts to predict the value of an unknown sample based on context. For artefact classification, however, a *discriminative* model is needed. Fortunately, a generative model can easily be converted into to a discriminative model by training separate models on the classes that need to distinguished. Separate models could be trained on clean and artefact contaminated EEG data. The data could then be classified by testing which model best predicts the data (the clean or artefact model).

7.1.2 Event-related potentials and big data

An important limitation of ERP research is the poor signal-to-noise ratio (Luck, 2014). Many trials are needed to reveal an ERP component in the EEG signal. Experimental paradigms that allow for the collection of a large amount data open up new research lines. In the current experimental setup (inspired by Frank et al., 2015; see also Willems, 2015), I analyzed ERPs elicited by words in natural speech. Analyzing most of the words in the speech materials presented to participants, increased the size of the EEG dataset. For the word form anticipation study (Chapter 4), I analyzed over a million word epochs and for the speech register study (Chapter

[7] For examples of the speech synthesis produced by the model, visit the following website: https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

6) I analyzed well over half a million word epochs. The size of these datasets is orders of magnitudes larger than classic EEG experiments.

The setup has important ecological consequences. It affords the study of more naturalistic language materials. If an experiment contains dozens of target words distributed over a small number of conditions, it becomes very important to carefully match these target stimuli. The ideal for classical ERP research is to use the same stimulus, but change the context to ensure that the surface form does not influence the results. This careful matching makes it difficult to use ecologically valid materials. Fortunately, this matching becomes less important as the size of the dataset increases because nuisance variation can be averaged out over the large number of stimuli.

The results presented in this thesis suggest that *big data collection* is a feasible and fruitful approach for ERP research. Possible new applications might be the collection of data from a more diverse participants pool, by incorporating participants from a wider age range and a more diverse educational background (our current dataset only contains data from university students within specific age bandwidth without reported health issues). These more general data could elucidate whether our predictability results generalize across human listeners in general.

7.2 To what extent do listeners anticipate the auditory word form while listening to natural speech?

Do listeners anticipate speech sounds while they are listening to someone talk? To investigate this, I developed a novel mismatch measure to capture the probability of a speech sound given the preceding speech input (see Chapter 3). This mismatch measure was used to predict the amplitude of the phonological mismatch negativity (PMN). I analyzed a large EEG dataset with over a million EEG word epochs. The results (see Chapter 4) indicate that listeners anticipate auditory word forms based on preceding language input.

7.2.1 Measuring mismatch with the cross-entropy between word probability distributions

To investigate predictive speech processing, I developed a mismatch measure (Chapter 3) inspired by the predictive coding framework (Friston, 2005). This framework proposes that cognitively high-level expectations feed backward to low-level perceptual input. Low-level sensory input is explained to a greater or lesser extent by the high-level expectations. The remaining unexplained *prediction error* travels upward to adjust the generative model (generating expectations). I operationalized this idea for speech perception by applying the concept of a word probability distribution (WPD).

The WPD is implemented as a lexicon whereby each word has an associated probability and a phonological representation. I used two WPDs, a *prior* WPD based solely on top-down word probabilities as estimated with the aid of an SLM, and a *post* WPD, which is the prior WPD but updated with information from the auditory onset of a word (the first 190 milliseconds). The prior WPD captures the high-level expectation based on the preceding context; the post WPD differs from the prior WPD to the extent that the auditory word onset was unexpected. If the speech signal does not support the prior WPD, the difference between prior and post WPD is large, while if the audio matches the expected word, the difference is small. The mismatch measure was computed by taking the cross-entropy between the prior and post WPDs. In this manner, the mismatch measure captures the unexpectedness of the auditory input given the preceding context (i.e., the prediction error).

The cross-entropy between prior and post WPD captures a plausible mechanism of predictability. Consider, for example, the following cases: When the sentential context is highly constraining, i.e., a few words have a high probability[8], the cross-entropy will be high if the actual next word is *unexpected*. This is true because the word onset will not match the words with most probability mass in the prior WPD. As a consequence, the post WPD will have a markedly different probability distribution[9] resulting in a high cross-entropy. In contrast, when the context is highly constraining and the actual word is *expected*, the cross-entropy will be low because the post WPD will have a very similar probability distribution. Finally, when the context is less constraining, probability will be more evenly distributed among words, and therefore the cross-entropy will not be as high for mismatching word

[8] The WPD tends to follow a power law distribution; there is a quick decline in probability from the most likely to the second most likely word and from the second to third word, etcetera, ending in a fat tail (i.e., unlikely words are not astronomically unlikely). Furthermore, it is the case that the amount of probability assigned to the most likely words depends on the constraining character of the sentential context, whereby less constraining contexts result in less probability for the most likely words and vice versa for more constraining contexts.

[9] This is because the post WPD is updated with the mismatching auditory word onset, resulting in a large shift of word probabilities. For implementation details, see Chapter 3.

onsets because the shift in the probability distribution (i.e., mismatch) is smaller compared to highly constraining contexts.

The advantage of the mismatch measure is that it can be applied to all words in a sentence and is not dependent on having an (artificially) high constraining sentential context. Furthermore, it is no longer necessary to bin target words into artificial sets of expected or unexpected words, because the mismatch measure is continuous, which allows for investigating the whole spectrum of predictability. This approach provides evidence that anticipatory processes are not limited to artificially constraining sentences but are also employed while listening to natural speech.

7.2.2 The underlying cognitive process indexed by the phonological mismatch negativity

There are different accounts for the underlying cognitive process indexed by the PMN, which is why the component is referred to by different names (i.e., N200, N250 and the phonological mapping negativity). The accounts mainly differ with respect to the processing level at which the purported mismatch occurs at. One interpretation explains the PMN as a mismatch between expected and perceived phoneme input (e.g., Connolly et al., 1994; Desroches et al., 2009; Brunellière & Soto-Faraco, 2013). According to this account, the speech processing system anticipates specific speech sounds based on the preceding context. When the word onset mismatches with this anticipatory process, a larger PMN effect is elicited compared to a word onset that confirms the anticipated sounds. The current results are best explained by an account of anticipated and mismatching auditory word onsets.

A competing explanation is related to an assumed modular speech processing system, such as Shortlist (Norris, 1994), in which early speech processing stages are exclusively feedforward. From this perspective, there cannot be a mismatch between expected and perceived phonemes because this would imply feedback to early (i.e., prelexical) stages of speech processing. To account for a N200 effect, it is assumed that this ERP component is sensitive to a mismatch between the sentential context and the set of semantic features of the word cohort activated by the auditory onset of the word (e.g., Van den Brink et al., 2001; Hagoort, 2007).

Van den Brink et al. (2001:979) argue against a phonological mismatch explanation (they refer to an N200 effect). Their data showed an early negativity for both expected and unexpected word onsets, whereby unexpected word onsets showed an increased negativity compared to expected onsets. According to their argument, if the N200 effect indexes a phonemic mismatch, you would not expect to see an early negativity with an expected word onset. However, they argue specifically against the case whereby only a single word form would be predicted from the preceding context. We showed that it is also possible that the anticipatory process results in a probability distribution over a large set of words and, by extension a probability distribution over a large set of possible word onsets (for implementation details, see Chapter 3). If probability is distributed over different word onsets, it makes sense that even high probability word onsets results in an early negativity, because the anticipatory system also assigns some probability to the other word onsets.

The current results do not rule out the semantic account of the PMN (or N200) effect (e.g., Van den Brink et al., 2001; Hagoort, 2007). It is possible that the PMN indexes both anticipatory auditory and lexical interface processes (similar to the N400 effect, which shows sensitivity to both semantic congruency and contextual expectedness). The current results suggest that one part of the semantic account is less tenable, however, which is the proposed feedforward only processing, whereby the auditory onset results in a cohort of matching words that only subsequently can be matched with the sentential context. In the following section, I discuss this issue further.

7.2.3 Feedforward models of speech processing

A long-running debate in language perception concerns the status of feedback (Norris et al., 2016; Magnuson et al., 2018). Autonomous models such as the previously mentioned model Shortlist (Norris, 1994; Norris & McQueen, 2008) assume that speech input is processed in a strictly feedforward manner. In Shortlist, no feedback is allowed from the lexical to the prelexical stages. In contrast, interactive models such as TRACE (McClelland & Elman, 1986) allow feedback from lexical to prelexical levels.

Norris et al. (2016) argue for the feedforward position from a Bayesian perspective. They argue that Bayesian inference from the available evidence (i.e., speech input) cannot be improved by feedback. In their model, prior probability is based on word frequency and the likelihood is based on the speech input. Bayes' theorem provides an optimal method to combine the prior and likelihood to compute the posterior probability of a word (i.e., an update of word probabilities with newly available evidence). Since the prior and likelihood are both part of the equation, nothing can be gained by feedback from lexical (i.e., prior) to the prelexical level (i.e., likelihood).

We argue that Norris et al. (2016) underrate the practical relevance of what can be ignored or minimized with feedback. As they themselves note, predictive coding is a compressive algorithm. Instead of passing through the whole signal, only differences between expected and perceived input are passed along (Friston, 2005). Norris et al. (2016) note that the predictive processing code could be learned (offline), while during online processing only feedforward mechanisms are necessary. The ERP results presented in Chapter 4 are in conflict with this hypothesis. I found evidence that online processing of natural language results in an ERP-component sensitive to the mismatch between expected and actual word form input. This result can be explained if perception entails high-level probabilistic feedback to low-level processing.

7.2.4 The sensory hypothesis and word recognition hypothesis

In a recent overview article, Nieuwland (2019) provides a critical overview of early ERP effects in relation to predictive language processing. In this overview, he makes a distinction between a sensory hypothesis and a word recognition hypothesis. The sensory hypothesis is closely related to the predictive coding framework, based on matching predicted and perceived sensory input (Nieuwland, 2019). The word recognition hypothesis proposes that prediction effects occur during the recognition of a specific word form. The sensory hypothesis is further characterized as non-linguistic/perceptual and the word recognition hypothesis as linguistic processing. The PMN/N200 effect is categorized under the word recognition hypothesis.

The current results do not neatly correspond to the sensory or word recognition hypothesis. I designed the mismatch measure (i.e., cross-entropy) in such a manner that it captures the mismatch between expected and actual auditory input (see Chapter 3). In this sense, it is closely related to the sensory hypothesis. However, I would not claim that the found effect is purely non-linguistic, nor would I assume a sharp divide between non-linguistic and linguistic processing. I argue that the results are best explained by the predictive coding framework and that this should not be viewed as a purely non-linguistic explanation because different levels of processing are coupled; high-level anticipations (based on preceding words) feed backward to auditory processing.

Nieuwland (2019) further argues that there is insufficient evidence for dissociating the PMN/N200 from the N400 and proposes that the results reported in the literature could be explained solely by the N400 component (by assuming the N400 has an early onset when elicited with auditory stimuli). I propose that the current results provide further evidence that the PMN component is dissociable from the N400. In this thesis, I used the same EEG dataset to analyze both the PMN and the N400 and found that the components can be dissociated by different information theoretic measures: The mismatch measure developed in Chapter 3 predicted the PMN amplitude, while word surprisal predicted the N400 amplitude.

The PMN is smaller in amplitude and closely precedes the N400. The overlapping component problem could explain why the PMN was not always reported when one would expect it (e.g., Van Petten, 1999). In addition, classical EEG experiments only collect a limited number of observations. Since EEG suffers from a poor signal-to-noise ratio, and the PMN effect has limited amplitude difference, it is unsurprising that the PMN effect is not always observed.

The dissociability of the PMN and N400 components could be further explored with the topographic distribution of found effects. For the PMN, I found a frontal lateral effect, and for the N400 a central parietal effect. I did not explore this further since a topographical analysis was not within the scope of our experiments, but I assume that topographic distinctions underpin the difference between PMN and N400. Further evidence that the PMN and N400 effects are dissociable could be obtained by performing a similar experiment with MEG recordings. Since MEG data has a far better spatial resolution compared to EEG data, this would provide more interpretable results of the underlying brain regions responsible for processing different aspects of natural speech.

7.3 To what extent do listeners adapt their expectations of upcoming words based on the speech register they are listening to?

If listeners are sensitive to the local context of preceding words, is it also the case that they are sensitive to the broader context? To investigate this, I used statistical language modelling to compare word predictability differences between speech registers and found that they differ systematically. Based on these findings, an EEG study was conducted in which participants listened to speech from different registers. The analysis of the EEG data showed that listeners adapt their expectations of upcoming words to the register they are listening to.

7.3.1 Word predictability differences between speech registers

In Chapter 5, word predictability differences between speech registers were investigated. To model register-specific word predictability, I used statistical language models (SLMs) by training individual SLMs on different register-specific language materials (e.g., dialogues, news broadcasts, read-aloud stories, etcetera). These register-specific SLMs were then applied to unseen language materials from multiple registers. The performance of the register-specific SLMs was computed and used to train a register classifier. The logic behind this approach is the following: SLMs performance depends on the similarity of the training and testing materials. An SLM trained on language materials from dialogues will likely perform worse when applied on language materials from news broadcasts compared to when it is applied to dialogue materials. I used the SLM performance measure (i.e., perplexity) as input to train a linear discriminant analysis (LDA) classifier to distinguish between 14 speech registers taken from the Spoken Dutch Corpus. The resulting classifier was able to perfectly distinguish between all speech registers (see Chapter 5).

To rule out confounds and test whether the classifier results were due to speech register differences in word predictability, I conducted several more experiments. First, I repeated the experiment with randomly grouped language materials instead of language materials grouped on register. SLMs were trained on the randomly grouped materials, and a classifier was again trained on the SLM performance measure. The resulting classifier could not distinguish between the randomly grouped sets of language materials. This result shows that the register classifier distinguishes between language materials grouped on the basis of some criterium (e.g., register), since it cannot distinguish randomly grouped sets of sentences. This result provides additional support that the speech register classifier results were based on systematic differences in word predictability among speech registers.

Second, I investigated whether the results would generalize beyond the Spoken Dutch Corpus. I extracted materials from other corpora and showed that the classifier could also correctly distinguish between registers with language materials from other corpora (see for example, Figure 3, Chapter 5). This shows that the classifier is not sensitive to differences between corpora and the results are not an artefact of an idiosyncratic aspect of the Spoken Dutch Corpus.

Third, I tested whether our classifier results could alternatively be explained by *sentence length* difference between registers. Sentence length might be an obvious predictor for register since it is related to speech style and therefore to speech register. Prepared speech, such as read-aloud stories or news broadcasts, contain longer sentences than more spontaneous speech, such as dialogues. Since sentence length differences potentially influence SLM performance, I investigated whether sentence length influenced the classifier results. I reran the experiment with a subset of the language materials, excluding very short and very long sentences, thereby reducing the average difference in sentence length among registers. The resulting classifier was still able to distinguish between the registers. In addition, I trained a sentence length classifier to test how well registers can be distinguished based on sentence length compared to a classifier based on word predictability. The classifier based on sentence length was clearly outperformed by the word predictability classifier (see Chapter 5). Furthermore, the classifier based on sentence length did not generalize to materials from other corpora. This indicates that the results of the word predictability-based classifier cannot be explained by sentence length difference between registers.

Fourth, I observed that topic correlates with register. For example, many soccer terms and names of famous soccer players occur very frequently in sports commentary, while they hardly occur at all in read aloud stories. To attenuate the influence of topic on the classifier results, I used a restricted SLM vocabulary (see Chapter 5.4) by only including words that were most evenly distributed across the different registers. In this manner, words strongly related to one specific register could not drive the classifier results.

The combined results from the different experiments provide strong evidence that speech registers indeed systematically differ in word predictability.

7.3.2 Register adaption during natural speech perception

In Chapter 6, I investigated the influence of register on predictive speech processing. The study was based on two observations: First, as shown by the corpus study in Chapter 5, speech registers systematically differ in word predictability. Second, as argued in the Chapter 1, language perception entails the anticipation of upcoming input. Most evidence of anticipatory language processing is limited to very local context (i.e., preceding words and sentences). I extended previous research by investigating the influence of the wider context represented by different speech registers.

The EEG dataset, described in Chapter 2, was used to test whether listeners adapt their word anticipations to the speech register they are listening to. Recall that the

EEG dataset consists of recordings from participants who listened to three distinct speech registers, namely dialogues, news broadcasts and read-aloud stories.

The language materials of the EEG dataset were analyzed with statistical language models (SLMs) to estimate *word surprisal*. Word surprisal increases in relation to how unexpected a specific word is given the preceding words. Different SLMs were trained in such a manner to reflect different anticipatory processing strategies, namely *generic, register-specific,* or *recency*. A generic strategy captures the case of no adaptation, register-specific captures register adaptation and recency captures adaptation to local context. The strategy-specific word surprisal values from these SLMs were used to model the EEG data.

I analyzed the event related potential (ERP) elicited by the content words ($n \approx 600{,}000$) and found that, as expected, the N400 amplitude was more negative for words with higher surprisal (i.e., more unexpected words) and best predicted by *register-specific* word surprisal. This finding indicates that listeners adapt their word anticipations to the wider context of the current language input.

7.3.3 How to explain the register effects on anticipatory word processing?

I propose two possible explanations for the results from Chapter 6: A *recognition* hypothesis and a *recency* hypothesis. The recognition hypothesis assumes that there are cognitive representations of the wider context, possibly in the form of registers. The recognition of a given register (or another kind of representation of the wider context) results in an update of the anticipatory processes. This update results in a different probability estimate of a word given the local context (i.e., a number of preceding words).

The recognition hypothesis received support from the current findings (the register-specific strategy outperformed all other processing strategies). Furthermore, recognition-based adaptation would be beneficial for anticipatory word processing: Recognizing the current context allows for faster updates compared to slowly continuously adapting to the context. However, it does entail positing multiple mental constructs representing different registers or other representational variants that encode the wider context.

In contrast, the recency hypothesis entails that the wider context is not explicitly represented and adaptation does not depend on the recognition of the wider context. Instead, anticipatory processes are continuously updated based on the received recent input. In Chapter 6, I found evidence that *simple* recency effects cannot explain the found results: The register-specific word surprisal values better predicted the N400 amplitude compared to the recency-based values. However, these surprisal values may be based on a too simplistic view of recency. I used a specific implementation of recency updating, whereby the generic SLM was updated with the last 64 encountered words. There are perhaps more optimal ways to implement recency effects. For example, *long short term memory* (LSTM) models (e.g., Sundermeyer et al., 2012) provide a state of the art method for language modelling. In these types of models, longer contextual effects can be modelled. Perhaps recency effects could be successfully modelled with this type of model, without the need of explicit reference to a specific register.

A viable approach to provide further evidence for either the recognition or recency hypothesis would be to study the speed of adaptation to the wider context. If the recognition hypothesis is correct, adaptation should show a sudden shift, because once the register is recognized, word anticipations can be adapted to it. In contrast, with continuous updating based on recency no sudden shift in word anticipations would be expected. Whether this can be investigated with the current approach is unclear; because ERP analysis requires many trials (i.e., many word epochs) to average out noise, it is an open question whether it would have enough temporal resolution to capture this sudden shift in adaptation.

Another approach to decide between both hypotheses would be to vary the familiarity the participants have with a given register. For example, a participant group of soccer enthusiasts and a neutral group could listen to soccer match commentaries. If the recognition hypothesis holds, then the soccer fans should show a greater register specificity effect compared to non-soccer fans. A potential pitfall would be attentional differences between the participant groups because soccer fans could be more engaged by the materials. Alternatively, participants could be exposed to unfamiliar registers and subsequently be tested on materials related or unrelated to these registers. Again, if the recognition hypothesis holds, the familiarized materials should result in a greater register specificity effect.

7.4 Future outlook

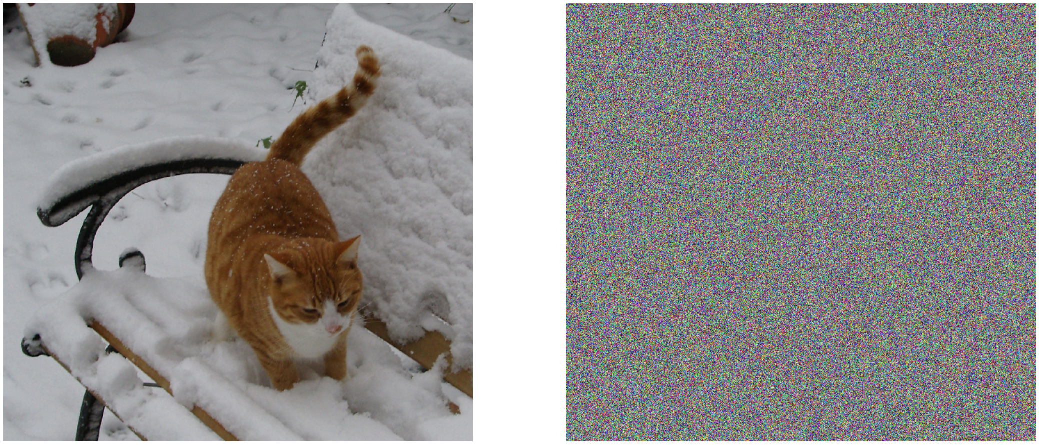
Given the results presented in this thesis, I briefly consider some broader theoretical implications which might be useful for future research.

7.4.1 Information theoretic measures and the perception of language

Information theoretic measures are central in this thesis, and they prove to be valuable for modelling linguistic predictive processing. What can this tell us about human speech perception?

First, it is important to keep in mind that *information* has multiple senses; information in the colloquial sense means something very different from the information as formal measure as put forward by Shannon (1948). Shannon information is purely defined in terms of channel capacity *without any notion of semantic content*, while the colloquial use of information does entail semantics. The difference is illustrated with the images of a cat and random numbers in Figure 2. The image of a cat contains *less* Shannon information compared to the random numbers because the cat image is structured. It depicts a cat, snow, and a bench. Consequently, some colors, like orange and white, occur more often than other colors and are grouped together in specific constellations. This *structure* affords a more efficient manner to record the information by, for example, assigning short codes to frequent colors and long codes to infrequent colors. Structure allows for code efficiencies, which reduces the amount of Shannon information. In contrast, in a set of random numbers, structure is absent by definition and no code efficiencies are possible, which is why a random sequence has the highest amount of Shannon information.

Figure 2. Left-hand side, a picture of a cat. This picture has a lower amount of Shannon information but a higher amount of colloquial information compared to the picture on the right-hand side which depicts a set of random numbers.



Colloquial information is a less well-defined concept. It overlaps with Shannon information e.g., a book contains more information than a paragraph, but colloquial information includes semantics, which Claude Shannon kept outside his definition of information (Shannon, 1948). Therefore, *colloquial information can conflicts with Shannon information*; for example, a book with a random sequence of letters contains more Shannon information compared to an equally long novel, while only the novel contains any colloquial information worth mentioning. There is no mathematical formula that captures colloquial information, and it is difficult to study with rigor (e.g., Mitchell, 2009).

With the distinction between colloquial and Shannon information in place, we are better positioned to relate the findings in this thesis to predictive language processing. I used information-theoretic measures to model the PMN (Chapter 4) and N400 (Chapter 6) effects elicited by words in connected speech. The question is: Why is the human speech processing system sensitive to Shannon information, which entails that an image of random numbers contains more information than an image of a cat? Intuitively, this feels completely backwards. Language is about the colloquial kind of information rather than the Shannon kind.

A possible solution to this puzzle is that there are multiple routes for predictive language processing. One route is sensitive to the sequential nature of language and is best modelled with information theoretic measures as investigated in the dissertation. However, this route does not capture the whole story. Take the following example:

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

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

Van Berkum, 2012

The target words *sword* and *lance* elicit a different ERP response dependent on whether the *knight* or the *sorcerer* is reaching for a weapon. Notice that this effect is observed while there are many intervening words between knight/sorcerer and sword/lance, which makes it unlikely that probabilistic anticipation modelled with information theoretic measures can account for this effect (see Aurnhammer & Frank, 2019). Importantly, as argued above, the different semantic relations between knight/sorcerer and sword/lance cannot be accounted for with information theoretic measures.

Frank and Willems (2016) compared *semantic similarity* with *word surprisal* and found evidence that semantic similarity and word surprisal *independently* predict word processing effort, whereby semantically more similar words and more expected words both result in less processing effort as indicated by reduced BOLD signal and N400 amplitude. Furthermore, the fMRI results indicate that the semantic similarity and word surprisal are processed in distinct neural areas.

Anticipatory language processing might thus be handled by a processing route based on the statistical structure in the sequential language input and another based on semantic similarity. I propose that the first route (modelled with information theoretic measures) is closely related to associative learning because it is based on the statistical structure in the environment and links co-occurring stimuli together. For example, a bell that reliably precedes food makes dogs salivate in expectation of food (Pavlov, 2010). Similarly, linguistic elements can be a signal for upcoming linguistic input if they co-occur reliably.

There is some experimental support linking associative learning to language processing. Baayen et al. (2011, 2013) used the Rescorla-Wagner equation to model language processing effects including *n*-gram effects. The Rescorla-Wagner model (Rescorla & Wagner, 1972), was developed to explain classical conditioning effects and correctly predicts many experimental results (e.g., Siegel & Allen, 1996). Furthermore, the link between associative learning and predictive languages processing provides a plausible explanation as to why Frank and Willems (2016) found that semantic similarity and word surprisal are complementary. This could be due to blocking (see Kamin, 1967). Blocking is the effect that with classical conditioning, associative links are only formed if they are informative. To the extent a word is already expected based on semantic similarity, blocking could prevent any additional associative learning.

A theoretic conjecture is only as good as the predictions it makes. Since this is only a sketch, I will end with one prediction based on the experiment by Bransford & Johnson (1972). In this experiment, participants were presented with enigmatic instructions to do a simple task (see below). The experimental manipulation consisted of whether or not the instruction was preceded with a clarification that the instructions concerned doing the laundry. The clarification improved the recall of participants. In a similar vein an fMRI experiment could be conducted whereby participants read a paragraph with or without a clarifying title. If the data were analyzed in a similar fashion as the Frank & Willems (2016) paper (see before), I predict a dissociation between word surprisal and semantic relatedness. The experimental manipulation (title or no title) would not influence the word surprisal findings, since the manipulation does not change the sequential structure of the stimulus. However, it would influence the semantic relatedness findings by causing a reduction in the BOLD response for the title compared to no title, because the semantic content of the input can be more easily integrated.

> The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavour. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many can easily arise. A mistake can be expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell. After the procedure is completed one arranges the materials into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will have to be repeated. However, that is part of life.

(Bransford & Johnson, 1972)

To recap, language input is sequential and structured; words follow each other, and some words are more likely to follow than others. Associative learning is sensitive to the statistical structure of the environment (in this case, language input) but can be blocked to the extent an input is already explained by other means. Words are semantically related to each other, and this leads to semantically-based expectations (e.g., a tree is cut down with a *chain saw* or an *axe* but not with a *scalpel* or a *banana*). The effect of blocking results in a complementary effect of prediction based on the statistical structure of language (modelled with information theoretic measures) and semantic similarity. The processing routes can be dissociated by manipulating a factor that only one route is sensitive to, as proposed in the previous paragraph.

7.5 Conclusion

This thesis focused on predictive language processing of natural speech. I used a novel approach to study natural speech processing. Participants were presented with long stretches of natural connected speech stimuli while their EEG was recorded. Natural speech stimuli are by their nature highly variable, which can be problematic for EEG analysis, since the EEG signal is highly sensitive to stimulus variation. However, by collecting a large EEG dataset, the nuisance variation could be averaged out over the large number of target items. The resulting EEG dataset contains well over a million word epochs and will be made freely available for further research.

The results presented in this thesis show that listeners anticipate speech sounds and words during the processing of natural speech. Listeners are not only sensitive to the local context of the immediately preceding words but also take into account the wider context of register. Predictive language processing is partially based on the sequential nature of the language input, which can be modelled with information theoretic measures.

Nederlandse samenvatting

Hoofdstuk 8

Gesproken taal verstaan is bijzonder uitdagend. Dit kan je ervaren wanneer je luistert naar een taal die je niet volledig machtig bent. Het lijkt dan alsof de spraakklanken onduidelijker zijn en het allemaal veel te snel gaat. Dit in tegenstelling tot het gemak waarmee je je eigen moedertaal verstaat. Een belangrijk verschil is de hoeveelheid ervaring die je hebt opgedaan met het luisteren naar verschillende talen. Veel ervaring helpt bij het classificeren van de spraakklanken en het herkennen van en anticiperen op woordreeksen. Een verklaring voor het gevoel dat een taal die je onvoldoende machtig bent te snel gesproken wordt, is dat het je slechter lukt om te anticiperen op wat er komen gaat.

Anticiperen op woorden en woordreeksen lijkt misschien een vergezocht idee. Er zijn immers zoveel manieren om je uit te drukken. De overvloed aan uitdrukkingsmogelijkheden gaf theoretici lang het idee dat taal voorspellen een zinloze exercitie is (Jackendoff, 2002). De redenatie is als volgt: Als iemand een bijna oneindige hoeveelheid uitdrukkingsmogelijkheden heeft, dan is de kans op één daarvan verwaarloosbaar klein en zal voorspellen meer kwaad dan goed doen. Deze veronderstelling gaat er alleen aan voorbij dat, hoewel er vele mogelijke manieren zijn om met woorden welgevormde zinnen te vormen, het echter niet zo is dat alle welgevormde woordreeksen met dezelfde waarschijnlijkheid voorkomen.

Sommige woorden komen namelijk vaker voor dan anderen, denk bijvoorbeeld aan *huis* of *mes* in vergelijking met *residentie* of *keukengerei*. Hetzelfde geldt voor woordreeksen, zoals *we gaan er nu vandoor* of *moet ik opstaan* in vergelijking met *we zullen terstond vertrekken* of *is het facultatief dat ik me verhef*. De voorbeelden illustreren dat hoewel we op verschillende manieren uitdrukking kunnen geven aan een voorwerp of een situatie, bepaalde manieren geëigender zijn dan anderen. Wanneer we het grondiger aanpakken en de computer inschakelen om woorden te tellen in grote hoeveelheid teksten, blijkt dat de frequentie van woorden en woordreeksen scheef is verdeeld. Een kleine groep komt zeer vaak voor en een zeer grote groep komt maar zelden voor.

Deze zeer scheve frequentieverdeling is een belangrijke vingerwijzing dat het mogelijk is om te anticiperen op taaluitingen. Woorden of woordreeksen die vaak voorkomen, hebben namelijk een grotere kans gehoord of gelezen te worden. Kennis van woordkansen kan taalperceptie ondersteunen door simpelweg de verwachting op woorden bij te stellen aan de hand van hoe waarschijnlijk ze zijn. Dat woordkansen ook daadwerkelijk belangrijk zijn bij taalperceptie blijkt uit experimenteel onderzoek. Bijvoorbeeld een leesexperiment (Smith & Levy, 2013) liet zien dat als een woord minder waarschijnlijk is, een lezer langer zal pauzeren en de lengte van de pauze langer duurt naarmate het woord onwaarschijnlijker is. Zo zijn er nog legio voorbeelden te noemen van experimenten die laten zien dat woord- en woordreekskansen luisteraars of lezers beïnvloeden (zie bijvoorbeeld de inleiding van dit proefschrift).

De experimenten die ik heb ondernomen en beschreven in dit proefschrift belichten nieuwe aspecten van anticipatieprocessen gedurende waarneming van **natuurlijke gesproken taal**. Met natuurlijke taal bedoel ik taal die niet specifiek bedacht en opgenomen is voor een experiment, maar bijvoorbeeld een gesprek tussen twee mensen of een voorgelezen boek. Het voordeel van dit soort taaluitingen te gebruiken voor experimenten is dat het minder kunstmatig is dan het taalmateriaal dat speciaal voor onderzoek wordt ontwikkeld. Het nadeel is dat het moeilijker is om vast te stellen waardoor een gevonden effect wordt veroorzaakt. Nieuwe statistische methodes maken het nu echter mogelijk ook effecten te achterhalen in dit soort moeilijkere gevallen. Het gebruik van natuurlijke taal in experimenten is een recente ontwikkeling (Willems, 2015) en is een belangrijke aanvulling op het gebruik van experiment-specifiek materiaal. Het laat namelijk zien in hoeverre eerder gevonden effecten ook te generaliseren zijn naar taalgebruik buiten het onderzoekslab.

Er zijn twee overkoepelende thema's in mijn onderzoek. Het eerste thema gaat over de anticipatie van spraakklanken. Woorden zijn opgebouwd uit spraakklanken. Als je luistert naar een gesprekspartner ben je typisch alleen bewust van de woorden of de bedoeling die wordt gecommuniceerd, maar onbewust verwerk je eerst de klanken waar woorden uit bestaan. Achtereenvolgens komen deze klanken binnen en onthullen successievelijk welke woorden gesproken worden. Als anticipatie een belangrijk mechanisme is bij het waarnemen van gesproken taal, dan zou het zo kunnen zijn dat dit niet enkel op woordniveau gebeurt, maar ook op het spraakklankniveau. Met een nieuw ontwikkelde metriek en experimenteel onderzoek heb ik bestudeerd of luisteraars inderdaad anticiperen op spraakklanken.

Het tweede thema gaat over de ruimere context waarin woorden gebruikt woorden, zoals bijvoorbeeld het verschil tussen een gesprek in de kroeg en een voordracht van een nieuwslezer. Het verschil in communicatiesituatie (kroeggesprek of nieuwsbericht) beïnvloed taalgebruik en daarmee mogelijk de waarschijnlijkheid van woorden. Ik heb onderzocht of woordwaarschijnlijk-heden systematisch verschillen tussen communicatiesituaties en of luisteraars daar gevoelig voor zijn.

Om te testen of luisteraars gevoelig zijn voor woord- en spraakklankkansen heb ik gebruikgemaakt van elektro-encefalografie (EEG). Dit is een techniek die hersenactiviteit meet met behulp van op de hoofdhuid geplakte elektrodes. Hersenen bestaan uit cellen genaamd neuronen en die communiceren met elkaar doormiddel van elektrochemische signalen. De elektrische activiteit van neuronen kan gemeten worden op de hoofdhuid. De EEG-meeting kan met behulp van statistiek gekoppeld worden aan experimentele manipulaties. Bijvoorbeeld, wanneer iemand het volgende hoort: *Hij smeert zijn brood met sokken*, dan zal bij het woord sokken een reactie optreden die is terug te zien in het EEG-signaal (Kutas en Federmeier, 2011). De uitdaging hierbij is dat de spanningsverschillen die je meet op de hoofdhuid zeer zwak zijn. Het is daarom nodig om metingen van vele observaties bij elkaar te nemen en te middelen om de effecten zichtbaar te maken.

Voor mijn onderzoek was er nog een extra uitdaging omdat ik natuurlijke gesproken taal onderzoek. Dit is uitdagend omdat EEG-metingen zeer gevoelig zijn, de gebruikte woorden of de manier van spreken kan het EEG-signaal beïnvloeden. Om ongewilde invloeden op het signaal te voorkomen wordt normaal gesproken het experimentele materiaal zo gemaakt dat er geen verschillen zijn, behalve hetgeen wat onderzocht wordt. In het voorgaande voorbeeld zal het woord *sokken* bijvoorbeeld ook in een zin gebruikt worden waarin het wel past *Hij vult de kleerkastlade met sokken*. Het nadeel van deze aanpak is dat het experimentele materiaal nogal kunstmatig wordt. Er zijn namelijk zoveel aspecten waarin woorden verschillen dat er maar een beperkt aantal mogelijkheden overblijft en met gesproken taal komt daar ook nog de invloed bij van de spreker bij. Natuurlijke spraak zit vol met verschillen.

De oplossing hiervoor was om zeer veel data te verzamelen. Zoals eerder vermeld, is het EEG-signaal zeer zwak en gevoelig. Het is daarom uitdagend om effecten te observeren in EEG-opnames. Door meerdere observaties bij elkaar te nemen kan je onderliggende effecten zichtbaar maken. In mijn onderzoek heb ik deze werkwijze een stap verder doorgevoerd. Een klassiek EEG-experiment bevat duizenden observaties, voor mijn onderzoek heb ik er meer dan een miljoen verzameld. De extra observaties helpen om met behulp van statistiek de onderliggende effecten, nu ook verstoord door de verschillen in het experimentele materiaal, zichtbaar te maken. De gevonden effecten zal ik hieronder verder beschrijven. De precieze beschrijving van de totstandkoming van de EEG-dataset is te vinden in Hoofdstuk 2.

Ik zal nu verder ingaan op het eerste thema, de anticipatie van spraakklanken. In de jaren negentig is hier al onderzoek naar gedaan met behulp van EEG-metingen. Connolly en collega's (1990, 1992, 1994) lieten een zin zoals de volgende horen, *de man leefde van een klein [pensioen, penseel, kussen].* Waarbij een zin werd afgemaakt met een van de drie woorden tussen haakjes. De experimentele manipulatie was dat de laatste woorden niet in de context van de zin passen en het tweede woord *penseel* met de eerste lettergreep *pen* overeenkomt met het passende woord *pensioen.* Ze vonden een verschil in het EEG-signaal tussen de niet passende woorden *penseel* en *kussen.* Dit effect is door onderzoekers verschillend geïnterpreteerd.

Sommige onderzoekers (Connolly en collega's 1992, 1994; Brunnelier en Soto-Faraco, 2013) veronderstellen dat het gerelateerd is aan het anticiperen op spraakklanken. Andere onderzoekers (Van den Brink en collega's, 2001; Hagoort, 2008) veronderstellen dat het gerelateerd is aan een semantisch effect, waarbij de eerste lettergreep van een woord een lijst van woorden activeert die overeenkomen met die lettergreep. Wanneer de semantische aspecten van de woorden in die woordenlijst niet overeenkomen met de voorgaande zin, zou het EEG-effect daardoor kunnen optreden. Een laatste interpretatie (Nieuwland, 2019) is dat het effect niet een verschil in verwerking laat zien, maar een verschil in timing. Het verschil tussen *penseel* en *kussen* is dat bij het woord *penseel* pas later ontdekt wordt dat het niet in de zinscontext past. Het zou dan niet liggen aan de spraakklanken, maar aan het niet passende woord, zoals bij het eerder vermelde voorbeeld van *sokken* waarmee een boterham werd belegd.

In mijn onderzoek ben ik uitgegaan van de eerste veronderstelling, dat het EEG-effect gerelateerd is aan het anticiperen op spraakklanken. Bij het waarnemen van taal creëer je op basis van hetgeen je gehoord hebt, verwachtingen over mogelijke volgende woorden. Op basis van deze verwachtingen is het mogelijk om verwachtingen te hebben over de spraakklanken die je zult horen. In het voorgaande voorbeeld zou je het woord *pensioen* kunnen verwachten en daarmee de lettergreep *pen.* In de hiervoor genoemde experimenten werd gebruikgemaakt van speciaal geformuleerde zinnen met woorden die ofwel zeer verwacht (pensioen) of zeer onverwacht waren (penseel of kussen) en precies overeenkwamen of verschilden in de eerste lettergreep van het woord. In mijn onderzoek maak ik gebruik van natuurlijke spraak en kan dus niet de taaluitdrukkingen aanpassen om de verschillen in woordkansen en spraakklanken te benadrukken.

Een alternatieve manier om tot een vergelijking te komen tussen een woord met een verwachte klank (penseel) en een onverwachte klank (kussen), is een metriek te ontwikkelen die aangeeft hoe verrassend een spraakklank is, gegeven welke woorden je verwacht. De metriek is gebaseerd op basis van een kansdistributie van woorden en spraakklanken. Een kansdistributie is een set van kansen die wordt verdeeld over verschillende mogelijkheden, bijvoorbeeld woorden. De woordkansen worden geschat op basis van tellingen van woorden in een grote hoeveelheid tekst. De spraakklankkansen worden geschat op basis van automatische spraakherkeningssoftware. Deze software analyseert audiomateriaal en geeft spraakklanken een waarde die omgezet kan worden tot een kans. Door de woord en spraakklank kansdistributies te combineren konden we waardes berekenen voor de kans op bepaalde spraakklanken gegeven de voorafgaande woorden. Deze spraakklank verrassingsmetriek is op verschillende manieren getest (voor meer details, zie Hoofdstuk 3), waarmee we konden valideren dat de metriek inderdaad een systematisch verloop van waardes laat zien voor verwachte tot onverwachte spraakklanken.

Zoals eerder genoemd hebben we EEG-data verzameld van mensen die luisterden naar natuurlijke spraak. Om te onderzoeken of luisteraars gevoelig zijn voor de kans op een bepaalde spraakklank hebben we het volgende gedaan. Voor elk woord in de natuurlijke spraak die de deelnemers van het experiment hebben gehoord, hebben we de hiervoor besproken metriek gebruikt om een verwachtingswaarde toe te kennen aan ruwweg de eerste lettergreep van het woord. Vervolgens hebben we voor elk woord de bijbehorende EEG-data genomen. Op basis van een statistisch model konden we laten zien dat er een verband is tussen de onverwachtheid van een spraakklank en de bijbehorende EEG-waardes (zie Hoofdstuk 4). We hebben dit effect gevisualiseerd in de grafiek op pagina 57, waarbij te zien is dat onverwachte spraakklanken een andere EEG-waarde opleveren ten opzichte van verwachte spraakklanken.

De resultaten komen overeen met de veronderstelling dat luisteraars onbewust verwachtingen hebben over spraakklanken. Dit is een interessant resultaat. Een theoretische stroming (Norris en collega's, 2018) veronderstelt namelijk dat de perceptie van spraakklanken niet beïnvloed worden door verwachtingen op basis van net gehoorde woorden. De redenering is dat het beter is om een zo accuraat mogelijke perceptie te hebben van de klanken die binnenkomen, in plaats van een door verwachtingen gekleurde perceptie. Anders zou je wellicht de ander niet meer verstaan en alleen maar horen wat je verwacht te horen. Hoewel dit soms weleens lijkt te gebeuren lukt het de meeste luisteraars toch echt om te verstaan wat de ander zegt. Het lijkt dus een redelijke veronderstelling dat spraakperceptie opgebouwd wordt door het nauwkeurig waarnemen van de simpelste eenheden - de spraakklanken - en dat je die waarneming beter niet kan verkleuren door hetgeen je al denkt te weten. Toch laten de resultaten van mijn onderzoek zien dat die verkleuring juist wel gebeurt. Hoe zit dat dan?

Een theorie over perceptie die dat zou kunnen verklaren is *predictive coding* (Friston, 2005, 2012, 2018). Deze theorie veronderstelt dat perceptie gebeurt aan de hand van voorspellingen omdat de prikkels die binnenkomen via de zintuigen zeer moeilijk te interpreteren zijn. Wat je hoort en ziet is namelijk een samenspel van vele invloeden. Als je bijvoorbeeld iemand hoort spreken zijn er vaak ook nog allerlei andere geluiden, zoals andere sprekers, muziek of het gezoem van een oude koelkast. Al deze invloeden maken het ingewikkeld om in de luchttrillingen die je trommelvlies bereiken de woorden te herkennen die gesproken zijn. Het heeft bijvoorbeeld een goede zestig jaar geduurd voordat het lukte om een computer gesproken woorden te laten herkennen en mensen zijn er vaak toch nog beter in.

Predictive coding veronderstelt dat om toch complexe ervaringen zoals het herkennen van woorden of het zien van een huis uit simpele prikkels (trilling van het trommelvlies of fotonen op de retina) te destilleren, er gebruik gemaakt kan worden van voorspellingen. De anatomie van de hersenen lijkt dit idee te ondersteunen. De verschillende waarnemingsgebieden in de hersenen voor bijvoorbeeld *zicht* of het *gehoor* zijn hiërarchisch opgebouwd, waarbij gebieden onderaan de hiërarchie simpele waarnemingskarakteristieken verwerken zoals contouren en gebieden aan de top complexe waarnemingen verwerken, bijvoorbeeld een *woord* of een *huis,* die zijn opgebouwd uit de simpele waarnemingen van lagere gebieden. Het is hierbij opvallend dat er in de hersenen vele verbindingen zijn die lopen vanaf hogere gebieden naar lagergelegen gebieden.

Het idee van predictive coding is dat gebieden aan de top van de hiërarchie voorspellingen doorgeven aan lagere gebieden. Op basis van een complexe waarneming, bijvoorbeeld een boom, kunnen er namelijk een voorspellingen gedaan worden over welke minder complexe waarnemingen daarbij waarschijnlijk zijn, welke contouren bijvoorbeeld verwacht worden bij een boom. De lagere gebieden geven dan vooral informatie door wanneer een voorspelling niet uitkomt. Op deze manier word je snel gewaar wanneer de dingen niet zo zijn zoals je dacht, zonder overweldigd te worden door de veelheid van indrukken die binnenkomen. Een voorbeeld van het effect dat een complexe waarneming een simpelere waarneming kan beïnvloeden, kan je ervaren als je naar de omslag van dit proefschrift kijkt. Als je niet weet wat er staat afgebeeld, zie je waarschijnlijk alleen een hoop zwarte vlekken, maar als je weet dat er in het midden een dalmatiër is afgebeeld, dan zie je opeens een hond verschijnen.

De resultaten die we gevonden hebben met het EEG-onderzoek en de spraakklank verrassingsmetriek komen overeen met hetgeen je zou verwachten op basis van de predictive coding theorie. Dit is verder nog interessant, omdat predictive coding een algemene theorie is over waarneming. In het taalonderzoek is lange tijd uitgegaan van het idee dat taal bijzonder is en niet zonder meer te vergelijken met andere vormen van perceptie. Het is natuurlijk zonder meer waar dat taal bijzonder is, maar het lijkt erop dat de perceptie van taal dus zeker ook overeenkomsten vertoond met andere vormen van perceptie.

Ik zal nu ingaan op het tweede thema; de invloed van de communicatiesituatie op woordkansen en de verwachtingen van luisteraars. Er is veel onderzoek gedaan (Biber en Conrad, 2001) waaruit blijkt dat de situatie waarin gecommuniceerd wordt het taalgebruik beïnvloed. Een kroeggesprek heeft bijvoorbeeld een heel andere dynamiek en vocabulaire dan een nieuwsbericht. Dit komt omdat de situatie communicatiemogelijkheden op allerlei manieren beïnvloed. In een gesprek wissel je spreekbeurten af en als er miscommunicatie is, dan is dit eenvoudig op te lossen door een vraag te stellen. Het is ook ongebruikelijk dat je veel tijd hebt om hetgeen wat je zegt voor te bereiden. Door deze mogelijkheden en beperkingen kenmerkt een spontaan gesprek zich door een beperkter vocabulaire en simpelere zinsconstructies. De sprekers hebben immers niet de tijd om het perfecte woord te vinden of een zin nog eens aan te passen. In tegenstelling, een nieuwslezer houdt een monoloog waarbij eventuele miscommunicatie niet gesignaleerd kan worden. De voordracht is meestal voorbereid waardoor er een ruimer vocabulaire gebruikt kan worden en het type woorden zal formeler zijn dan wat er in de kroeg gebruikt wordt.

Omdat het taalgebruik verschilt per situatie, zou het ook zo kunnen zijn dat woordkansen verschillen per situatie. We hebben daarom onderzocht of we systematische verschillen in woordkansen konden vinden tussen taalgebruik in verschillende communicatieve situaties. We hebben hiervoor onder meer het Corpus Gesproken Nederlands (Oostdijk, 2001) gebruikt. Dit een verzameling van audio-opnames van gesproken taal in verschillende situaties, zoals dialogen, voorgelezen verhalen en nieuwsberichten. De opnames zijn helemaal uitgeschreven. Hierdoor is het eenvoudig de voorkomende woorden te tellen. Op basis van deze woordtellingen konden we bij benadering woordkansen vaststellen voor het taalgebruik in de verschillende communicatiesituaties die zijn opgenomen in Corpus Gesproken Nederlands (zie Hoofdstuk 5).

Met behulp van een classificatie-algoritme genaamd lineaire discriminantanalyse, konden we vaststellen dat er inderdaad systematische verschillen zijn in woordkansen tussen het taalgebruik in verschillende communicatieve situaties. Een illustratie van het resultaat is te vinden op pagina 84. De afbeelding geeft de classificatie weer van verschillende soorten tekst. Elk puntje staat voor een stuk tekst (bijvoorbeeld een dialoog of een nieuwsbericht). Punten die dicht bij elkaar staan lijken volgens het algoritme meer op elkaar dan punten die verder weg staan. De afbeelding visualiseert dat het classificatie-algoritme het taalgebruik in de verschillende communicatieve situaties kan onderscheiden. Stukjes taalgebruik uit dezelfde communicatieve situatie clusteren namelijk samen, terwijl er duidelijk afstand te zien is tussen taalgebruik uit verschillende communicatieve situaties. In Hoofdstuk 5 staat de precieze methode beschreven en extra tests die we hebben uitgevoerd, waarmee we alternatieve verklaringen voor de verschillen (zoals zinslengte of situatie-gebonden onderwerpen) konden uitsluiten.

Nadat we hadden vastgesteld dat er systematisch woordkansverschillen zijn tussen taalgebruik in communicatieve situaties hebben we getest in hoeverre luisteraars hiervoor gevoelig zijn. We hebben hiervoor de eerder beschreven EEG-dataset gebruikt. De deelnemers luisterden naar verschillende soorten natuurlijk gesproken taal; dialogen, voorgelezen boeken en nieuwsberichten.

Voor alle inhoudswoorden zoals zelfstandige en bijvoeglijke naamwoorden, bijwoorden en werkwoorden hebben we de EEG-data verzameld. Voor elk van deze woorden hebben we ook de waarschijnlijkheid vastgesteld. Om de woordwaarschijnlijkheden te schatten hebben we gebruikgemaakt van taalmodellen. Een taalmodel gebruikt de voorgaande woorden om te bepalen wat de kans is op het volgende woord.

Een taalmodel kan woordkansen schatten door patronen te ontdekken in een grote hoeveelheid tekst. Een belangrijk gegeven hierbij is aan welke soort teksten het taalmodel wordt blootgesteld. Als het taalmodel bijvoorbeeld enkel leert te voorspellen op basis van teksten uit kookboeken, dan zal het woorden uit een recept goed kunnen voorspellen, maar veel moeite hebben met een krantenbericht. Dit gegeven hebben we gebruikt om verschillende soorten taalmodellen te maken voor dialogen, voorgelezen verhalen en nieuwsberichten. Daarbij hebben we ook een algemeen taalmodel gemaakt dat gebaseerd was op zeer veel verschillende teksten. Elk taalmodel schat woordkansen net anders in en met deze verschillende woordkansen konden we onderzoeken welk taalmodel het beste de EEG-data van de deelnemers van het experiment voorspelt.

Het zou bijvoorbeeld zo kunnen zijn dat luisteraars gelijke woordverwachtingen hebben ongeacht het taalgebruik dat ze horen, of ze nu naar een dialoog luisteren of naar een nieuwsbericht. Het zou ook kunnen zijn dat luisteraars hun verwachtingen bijstellen afhankelijk van het soort taalgebruik dat ze horen. Met behulp van statistiek hebben we vergeleken welke woordkansen de EEG-data het beste kon voorspellen. Het is namelijk zo dat woordkansen zich op een bepaalde manier verhouden tot de waarde van het EEG-signaal in reactie op een woord. Naarmate een woord minder verwacht is, kan je een negatievere waarde zien in het EEG-signaal. Het is wel zo dat dit alleen zichtbaar is in de gemiddelde EEG-waardes van vele woorden tezamen genomen.

Door te vergelijken welke woordkansen, gebaseerd op een specifiek of algemeen taalmodel, het beste de EEG-data voorspellen konden we achterhalen dat de woordkansen gebaseerd op een taalmodel getraind op eenzelfde taalgebruik (bijvoorbeeld dialogen) als hetgeen de luisteraars hoorden, het best overeenkwamen met de EEG-metingen. Dit resultaat komt overeen met onze hypothese dat luisteraars gebruikmaken van het soort taalgebruik bij het anticiperen op mogelijke woorden.

Aangezien er nog niet veel onderzoek is gedaan naar de invloed van het soort taalgebruik op woordverwachtingen is niet mogelijk om hier harde conclusies aan te verbinden en roept het ook veel vragen op. Hoe lukt het luisteraars om hun verwachtingen bij te stellen? Is het zo dat het soort taalgebruik herkend wordt en op basis van de herkenning verwachtingen worden bijgesteld? Of is het zo dat naarmate je meer luistert, je langzaamaan verwachtingen aanpast, zonder iets te herkennen? Dat laatste hebben we onderzocht door ook een taalmodel te maken dat gebruik maakt van recente input om verwachtingen bij te stellen. De vergelijking met het specifieke taalmodel liet zien dat de woordkansen gebaseerd op het specifieke taalmodel de EEG-metingen beter voorspelden. Dit geeft dus enige aanleiding om te veronderstellen dat soorten taalgebruik worden herkend en op basis daarvan woordverwachtingen worden bijgesteld.

In dit proefschrift is onderzoek beschreven naar de rol van anticipatie bij het verstaan van spraakklanken en woorden. De verschillende experimenten laten zien dat anticipatie plaatsvindt op verschillende tijdschalen, van heel kort bij *spraakklanken* tot zeer lang onder *de invloed van het soort taalgebruik*. De resultaten laten zien dat anticipatie een belangrijk onderdeel is van taalwaarneming.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: a system for large-scale machine learning. In: *OSDI* (Vol. 16, pp. 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.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., & Gildea, D. (1999). Forms of English function words-effects of disfluencies, turn position, age and sex, and predictability. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.), *Proceedings of ICPHS-99* (pp. 395-398). Berkley, CA: University of California. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0395.pdf (last accessed February 2019).

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 *Proceedings Interspeech 2019*, 2265-2269.

Biber, D. (1988). *Variation Across Speech and Writing*. New York: Cambridge University Press.

Biber, D. (1995). Dimensions of Register Variation: A cross-linguistic Comparison. New York, NY: Cambridge University Press.

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.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of verbal learning and verbal behavior*, *11*(6), 717-726.

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and language*, *125*(1), 82-93.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359-393.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*(2), 163-190.

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and language*, *39*(2), 302-318.

Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and language*, *43*(1), 1-18.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of cognitive neuroscience*, *6*(3), 256-266.

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.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150-162.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Denoual, E. (2006). A method to quantify corpus similarity and its application to quantifying the degree of literality in a document. *International Journal of Technology and Human Interaction*, *2*(1), 51-66.

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.

Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics* 48, 60-75.

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.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 862-877.

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.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goi, R., Jas, M., Brooks, T., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, *7*, 267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446-460.

Goedertier, W., Goddijn, S. M., & Martens, J. P. (2000). Orthographic transcription of the Spoken Dutch Corpus. In N. Calzolari, G. Carayannis, K. Choukri, H. Höge, B. Maegaard, J. Mariani, & A. Zampolli (Eds.), *Proceedings of LREC-2000*. Athens: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2000/pdf/87.pdf (last accessed February 2019).

Gleick, J. (2011). *The Information: A History, A Theory, A Flood.* New York: Random House, Inc.

Gries, S. T. (2001). A corpus linguistic analysis of English *-ic* vs *-ical* adjectives. *ICAME Journal*, *25*, 65-108.

Gries, S. T., Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*(1), 228-255.

Hagoort, P. (2008). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1055-1069.

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*, 159–166.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research*, *1626*, 118-135.

Hlaváčová J., Rychlý P. (1999). Dispersion of words in a language corpus. In V. Matousek, P. Mautner, J. Ocelíková, P. Sojka (Eds.), *Text, Speech and Dialogue: Second International Workshop, TSD'99 Plzen, Czech Republic, September 13-17, 1999 Proceedings* (pp. 321-324). Berlin: Springer.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954-965.

Jackendoff, R. (2003). Foundations of Language: Brain, Meaning, Grammar, Evolution. New York: Oxford University Press

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd ed.). Upper Saddle River, NJ: Pearson.

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.

Kolk, H. H., Chwilla, D. J., Van Herten, M., & Oor, P. J. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and language*, *85*(1), 1-36.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

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., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing.

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.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5*(3), 37-72.

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675-724.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.

Luck, S.J. (2014). An introduction to the event-related potential technique. Cambridge: MIT Press.

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.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in psychology*, *9*, 369.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.

Mitchell, M. (2009). *Complexity: A Guide d Tour.* New York: Oxford University Press.

Miller, D., Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, *20*(1), 30-53.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In W. Daelemans (Eds.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398-408). Avignon: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/E12-1041 (last accessed February 2019).

Nicholls, D. (2014). *Us*. New York, NY: HarperCollins Publishers.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of neuroscience methods*, *192*(1), 152-162.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, *115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience*, *31*(1), 4-18.

Norris, D., McQueen, J. M., & Cutler, A. (2018). Commentary on "Interaction in Spoken Word Recognition Models". *Frontiers in psychology*, *9*, 1568.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., ... & Mézière, D. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*.

Nieuwland M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and biobehavioral reviews*, *96*, 367–400.

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.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*, 1.

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.

Pavlov P. I. (2010). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, *17*(3), 136–141. https://doi.org/10.5214/ans.0972-7531.1017309

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Effects of word frequency on the acoustic durations of affixes. In *Proceedings of Interspeech 2006 - ICSLP* (pp. 953-956). Pittsburgh, PA: International Speech Communication Association. Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1241.pdf (last accessed February 2019).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hanneman, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, 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.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In A. Kilgarriff & T. Berber Sardinha (Eds.), *Proceedings of the Workshop on Comparing Corpora of ACL 2000* (pp. 1-6). Hong Kong: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W/W00/W00-0901.pdf (last accessed February 2019).

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning Ii: Current theory and research.* New York: Appleton-Century-Crofts.

Savický, P., & Hlavácová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215-231.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture.

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).

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391-5420.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314-321.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361-382.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. New York, NY: Palgrave Macmillan.

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 J. H. L. Hansen & B. L. Pellom (Eds.), *Proceedings of the International Conference on Spoken Language Processing*. Denver, CO: International Speech Communication Association. Retrieved from https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_0901.pdf (last accessed February 2019).

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

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.

Tottie, G. (1991). Negation in English Speech and Writing: A Study in Variation. San Diego, CA: Academic Press.

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 Den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of cognitive neuroscience*, *13*(7), 967-985.

Van Gompel, M., & Van den Bosch, A. (2016). Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software*, *4*(1), 1-10.

Van Gijsel, S., Speelman, D., & Geeraerts, D. (2006). Locating lexical richness: A corpus linguistic, sociovariational analysis. In J.M. Viprey (Eds.), *Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data* (pp. 961-971). Besançon: Presses universitaires de Franche-Comté. Retrieved from http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-085.pdf (last accessed February 2019).

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.

Van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: A Free Dialog Video Corpus. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *LREC* (pp. 501-508). Marrakech: ELRA. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/132_paper.pdf (last accessed February 2019).

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.

Willems, R. M. (Ed.). (2015). *Cognitive neuroscience of natural language use.* Cambridge: Cambridge University Press, 2015.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506-2516.

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.

Acknowledgements

The convention is to write a thesis in the first person; I did this and I thought of that, while the reality is of course, I had a lot of help with most of the work. I think in many places I still slipped up and wrote *we* anyway. So at the end, in the acknowledgements, I can try to set the record straight and express my gratitude to all those who helped me.

First and foremost, I would like to thank my (co)promotors: Mirjam Ernestus, I think it was maybe nine years ago that you hired me as a student assistant, which was a great opportunity to get acquainted with experimental research. In 2014 we decided to write a project proposal for a PhD position, which was granted by the Radboud's Centre for Language Studies. Thank you for giving me this wonderful opportunity. You always gave me a lot of freedom to pursue my own ideas and gave valuable feedback. When we strongly disagreed about something you were always very patient in discussing the issue and I even managed to convince you a couple of times.

Louis ten Bosch thanks for all your support. It was always nice to philosophize about whatever and sparring with you helped sharpen my ideas or discard the bad ones. I think we had a great collaboration on the metric we developed for the PMN research, with a great result.

Antal van den Bosch thank you for introducing me to computational linguistics, it grew to be an important part of my research work and I think the crossover between different disciplines really helped push my research further. You were always very supportive and open minded and provided great feedback on my drafts.

Lou Boves you were not one of my promotors but you were very generous with your time anyway. Thank you for all the feedback on my questions and ideas. Your door was always open and you always found time to explain or discuss anything.

Special thanks go out to all the student assistants: Melanie, Tom, Marein, Nadia, Lisa, Erik, Thera, Stef, Tim, that helped me run a seemingly endless amount of EEG experiments (well over a 150 two-hour sessions). Extra special thanks for Tim Zee for also helping me with the drudgery of cleaning all that EEG-data. It is extremely mind-numbing work and I am very grateful that I did not have to do it all by myself.

I would also like to thank all the colleagues on the eight and nineth floor of the Erasmusgebouw for a very friendly and welcoming working environment. The coffee and lunch breaks did not always boost productivity, but they were *erg gezellig*.

Of course, I would also like to thank my paranymphs; Emily Felker it was great fun and quite an adventure to go to the Interspeech conference together in Graz by train. Thanks a lot for your very helpful and detailed commentary of the introduction and discussion sections. You helped make it a lot better. Robert Chamalaun, thanks for being my officemate at the MPI and Erasmusgebouw. It was fun to attend the English evening course together. I do not know if our English improved much, but we did get a nice certificate.

I want to thank my parents for always supporting me during all my studies and a special thanks for my mother for proofreading a lot of my academic writing.

Last but definitely not least I want to thank Fenna for supporting me and being patient with me when I am not. I could not have done this without you.

For all those who I forgot to mention, my sincere apologies and of course my gratitude for not making a big deal out of this.

Curriculum Vitae

Martijn Bentum was born in Alkmaar, the Netherlands, on the 28th May 1982. He studied Fine Art art (BA) at the Artez art academy in Arnhem and Dutch language and culture (BA) and Cognitive neuroscience (MA) at Radboud University Nijmegen. In collaboration with Mirjam Ernustus, Martijn wrote a project proposal, which was granted by the Centre for Language Studies at Radboud University. Martijn started his PhD project in 2014. Since October 2019, Martijn works for Radboud's Humanities Lab.

List of publications

Bentum, M., ten Bosch, L., van den Bosch, A., & Ernestus, M. (2019). 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. (2019). Quantifying expectation modulation in human speech processing. In *Proceedings Interspeech 2019*, 2265-2269.

Bentum, M., ten Bosch, L., van den Bosch, A., & Ernestus, M. (2019). Listening with Great Expectations: An Investigation of Word Form Anticipations in Naturalistic Speech. In *Proceedings Interspeech 2019*, 2265-2269.