The Dutch EEG Speech Register Corpus

(DESRC)

M. Bentum[1)], L. ten Bosch[1)], A van den Bosch[2)], M. Ernestus[1)]*
[1]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[2] Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands
*Corresponding author
{martijn.bentum, louis.tenbosch, mirjam.ernestus}@ru.nl, a.p.j.vandenbosch@uu.nl

Abstract

The Dutch EEG Speech Register Corpus contains 207 hours of EEG recordings from 48 participants listening to natural connected speech. The speech recordings were sampled from spontaneous dialogues, news broadcasts and read-aloud stories, and contain 50,277 word tokens per participant, time-locked to the EEG recordings. We cleaned the data with a novel approach by training a convolutional neural network artefact classifier on EEG recordings with manually labeled artefacts. We applied the artefact classifier on all EEG recordings and manually checked all automatically identified artefacts to ensure high data quality. Eye-related activity was removed with independent component analysis. The EEG recordings (raw and cleaned), contain 1.5 million word epochs, are freely available (license: CC BY NC 4.0) and offer research opportunities to investigate neural correlates of natural connected speech processing.

Keywords: speech perception, speech registers, electroencephalography, corpus, statistical language models

1. Introduction

This article presents a new corpus of EEG recordings, the Dutch EEG Speech Register Corpus (henceforth DESRC). For the DESRC, we recorded EEG data from 48 participants who listened to long (4 – 15 minutes) continuous stretches of natural speech and we time-locked the EEG data to *all* words in the speech materials. In this manner, we collected 207 hours of EEG recordings, containing 1.5 million word epochs. The large amount of freely available, and carefully cleaned EEG data in the DESRC corpus hopefully provides many new opportunities to investigate the neurophysiological correlates of speech perception of natural connected speech.

The setup for the EEG recordings for the DESRC differs from classic EEG experiments. We used long continuous stretches of natural speech, which 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 approach as the *naturalistic sample approach*. The naturalistic sample approach is based on three ideas. The first idea is to use naturalistic linguistic stimuli to improve the ecological validity of experimental results (see also Willems, 2015).

The second idea is to dramatically increase the number of stimuli by considering *all* words in the language materials, rather than a subset of target words. This affords a relaxation of stimuli control, tackling an important drawback of EEG, namely, the sensitivity to the precise surface form (e.g. a specific recording of a spoken word) of the experimental materials (for more challenges, see Luck, 2014). The effects of stimuli surface forms will average out over the large number of stimuli, i.e. hundreds of thousands of stimuli, versus tens to hundreds in classical experiments. The use of large numbers of stimuli is aided by statistical analysis techniques such as linear mixed effects modelling (Bates et al., 2015).

The third idea concerns the type of predictors. With a big dataset, categorical predictors can be replaced with continuous predictors, thereby foregoing 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).

With the use of longer stretches of natural speech it becomes feasible to consider the register of the language materials as a factor influencing processing. Register refers to the influence of the communicative situation on language use (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; formal occasions encourage more careful pronunciation than informal occasions (Ernestus et al., 2015) and the use of more formal words. Bentum et al. (2019a) found that word surprisal, as estimated by a statistical language model, depends both on the preceding words and speech register. The differences between speech registers could influence a listener's speech processing. To capture the effect of speech register variation, we sampled speech materials from different registers and used the surprisal findings reported by Bentum et al. (2019a) to select three distinct speech registers (for further details see Bentum et al., 2022).

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. 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, because any z-score thresholding rejects outliers, but is not informative about the quality of the rejected data. For example, when a participant's dataset is noisy, z-score thresholding will only remove extremely noisy subsets and keep potentially corrupted data, while when the participant's 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 was trained to discover features that distinguish between clean and artefact data without relying on statistics of simple measures (e.g. amplitude, channel correlation), which only imperfectly capture that distinction.

In the following sections, we detail the speech materials, the EEG recording and processing procedure, the training and validation of the automatic artefact classifier, and discuss validity of the corpus.

2. Corpus

The Dutch EEG Speech Register Corpus (DESRC) consists of 207 hours of EEG materials recorded from 48 participants listening to Dutch speech, sampled from three different registers: spontaneous dialogues, news broadcasts, and read-aloud stories. The EEG recordings for a participant was split into three sessions; during a single session a participant

listened to 90 minutes of speech materials form a specific register (e.g. spontaneous dialogues). The orthographically transcribed speech material is time-locked to the EEG recordings and we further enriched the data set with various types of information about the words in the natural speech stretches: part-of-speech tags, word frequency and several information theoretic measures such as word surprisal, entropy and cross entropy (for details see Bentum et al., 2022 and Bentum et al., 2019b, respectively).

2.1 Speech materials

The speech materials were sampled from different corpora; the news broadcasts and read-aloud stories were taken from the Spoken Dutch Corpus (Oostdijk, 2002). 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 and EEG recordings.

Table 1 lists descriptive 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), recorded in 2006. They freely talked about any topic that came to mind. One of the 11 speakers is present in two dialogues. The read-aloud stories materials consist of seven 12-minute-long excerpts from read-aloud Dutch audio books published between 1991 and 1999. 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.

Table 1. Overview of the materials per speech register: 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 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.

2.3 EEG Procedure

The participants came to the lab on three separate occasions, separated by at least one week. They were fitted with the correct size electrode cap and seated in a sound-attenuating booth. The audio materials were presented via in-earphones (Etymōtic ER1) at a comfortable listening volume; a short audio sample (not used during the experiment) was used to set the volume. Participants listened to 90 minutes of speech from one register (see Table 1). The order of the registers was counter-balanced across participants. Participants were requested 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. At the end of each block a participant could take a break.

To encourage attentive listening, we visually presented yes-no comprehension questions. For both the dialogues and books sessions, comprehension questions were presented at the end of a 15- and 12-minute block, respectively, while for the news session,

questions were presented at the end of 4-minute sections within each block. In total we presented 36 questions for dialogues, 42 questions for books and 84 questions for the news materials. The news materials were split up in shorter segments to ease cognitive load, because the news materials contain many different topics. Participants responded with a button box.

2.4 EEG recording

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

2.5 EEG 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 training and applying a deep neural network artefact classifier (see Section 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 DESRC dataset. See Table 2 for an overview of the EEG materials.

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)**

3. Automatic EEG artefact detection with a convolutional neural network

In the following subsections, we describe how we annotated part of the EEG materials, and based on these annotated materials, trained and tested a CNN for artefact detection.

3.1 Manual artefact annotation

We manually annotated approximately 60 hours of EEG data, 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 two 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 that part of the data.

3.2 Training, test and validation materials

The EEG data was first downsampled from 1000 to 100 Hz for training and classification purposes. Based on the manually annotated data, 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 per window) 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 thus consisted of a matrix of 25 channels by 100 samples and had a label: *clean* or *artefact.*

For the *channel* dataset we again excluded the Fp2 channel. For this dataset we created a separate window for each channel, which had the label *clean* or *artefact* based on the given target channel. Every window consisted of a matrix of 32 channels by 100 samples. We created this matrix by copying the target channel to the rows 1, 7, 13, 19, 25, 31. All other rows were filled by the 25 channels in fixed order. In this manner, each channel had a fixed position in the matrix while the target channel also had a fixed position in the matrix (i.e. row 1, 7 , …). We duplicated the target channel on these rows to mark it as the target and to ensure that the second layer kernel (see Section 3.3) would always be exposed to the target channel.

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

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., 2017). The typical CNN architecture for image classification specifies multiple convolutional layers of *n* by *n* (e.g. *n* = 5) kernels. For EEG data this kernel specification appears to be suboptimal, arguably because the time and channel dimensions have a different impact and statistical behavior. We adapted the model following Schirrmeister et al. (2017), who reported good results with EEG data classification where the first two convolutional layers of their model specify the time and channel dimensions, respectively. We found that this time-channel separation approach also strongly improved the performance of our classifier.

We defined a separate stretch and channel classification 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 is exposed to six EEG channels at each time point. Subsequently, the output is pooled and followed by a kernel of 5 by 5 for the stretch model and 6 by 6 for the channel model, followed by 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 parentheses). Conv. stands for convolutional layer, Relu for rectified linear unit.

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

3.4 Training, classification and manual correction

We trained both models with stochastic gradient descent. Each training epoch, a model was exposed to 200 windows drawn randomly from a given training set. We sampled down in favor of the artefact windows to a 50/50 ratio (approximately 7% of windows contain artefacts in the original data), to ensure the classifiers have a high recall of artefacts. We repeated training cycles until the classifier performance plateaued on the validation set.

The resulting 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 two sections of artefact annotations were separated by less than two seconds, we combined the two artefact annotations. All automatic artefact annotation boundaries were corrected based on manual inspection. During manual inspection, we did not consider sections labeled as clean by the automatic classifiers of 40 seconds or longer because long clean stretches are unlikely to contain artefacts (artefacts tend to cluster). Therefore, 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, otherwise channel artefacts were relabeled as stretch artefact

(i.e. labeling the stretch of EEG data as artefact). These manually corrected annotations were used to exclude EEG data contaminated with artefacts from the EEG dataset.

3.5 Classifier validation

We analyzed the quality of the CNN classifier by comparing the automatic artefact annotations with the manually corrected annotations and with a simple threshold approach. This latter approach functioned as baseline, which we detail below.

We chose the *word epoch* as the validation unit. 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 the different annotation sets. As ground truth, we used the manually corrected automatic annotations (see Section 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 (a standard value for thresholding EEG data).

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. As intended, we boosted the recall (0.87) of artefacts at the cost of a slight drop in precision (0.83).

The validation results show that there is a clear trade-off between time spent cleaning the EEG materials versus the quality and amount of usable EEG materials. The threshold approach is very fast, because no prior labelling of EEG data is required. However, this 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 for labelling training data. Manually correcting the classifier output further improved the quality of the EEG materials; however, this took 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**

4 Corpus validation

The DESRC is the first corpus with a large amount of EEG recordings time-locked to each word in the natural speech materials presented to the participants. The corpus is therefore ideally suited to investigate neural correlates of speech perception. The speech materials are manually transcribed at the orthographic level and automatically transcribed at the phonemic level, enabling more fine-grained studies into phoneme perception in natural speech. For each participant, there are approximately fifty thousand word epochs.

In the following subsections we briefly discuss previous research that utilized the DESRC so far and that shows that the corpus is indeed suited for investigating a range of research questions. In addition, we discuss the potential of the corpus for future research.

4.1 Initial findings

The data in the DESRC was already used in two published studies. In Bentum et al., 2019b, we investigated the phonological mismatch negativity (PMN), an event-related potential (ERP) that indexes unexpected compared to expected speech sounds in word onsets (see for example, Connolly et al., 1990 & 1992; Brunellière & Soto-Faraco, 2013). An important

criticism of previous PMN studies concerns the artificialness of the experimental design used to elicit the PMN, comparing very likely to very unlikely words (Huettig, 2015).

The EEG data in the DESRC were ideal to address the issue of artificial experimental stimuli, since it is based on participants listening to natural speech and the corpus contains a large amount of EEG data to study the PMN, which is only a small effect. We analyzed all words in the natural speech stimuli with a novel continuous measure (Bentum et al., 2019c) that quantifies the unexpectedness of the speech sounds in a word's onset (the computed values are also part of the DESRC), based on the preceding words and an analysis of the speech sounds with an automatic speech recognition system; see Bentum et al. (2019c) for the details and validation.

. We were able to successfully predict the PMN amplitude for the word epochs in the EEG data of the corpus with this new measure (see Bentum et al., 2019b). This is the first study that indicates that the PMN is also present in EEG data recorded from participants listening to natural speech, and provides evidence that listeners continuously anticipate upcoming speech sounds.

In a second study (Bentum et al., 2022), we investigated the influence of speech register on word expectations during listening. Previous research (Frank et al, 2015) had found that word surprisal, a measure that captures the unexpectedness of a word, can predict the N400 amplitude in a reading task. The N400 is a well-known ERP and can be characterized as a negative deflection of the ERP signal 400 ms after word onset, indexing the unexpectedness of a word given the preceding context (for an overview see Kutas & Federmeijer, 2011).

In the Bentum et al. (2022) study, we used the EEG data collected in the DESRC to investigate whether listeners use the wider context of speech register to adjust their word anticipations. The EEG data in the DESRC were recorded from participants listening to long

stretches of natural speech materials sampled from different speech registers. We modelled the N400 amplitude with different word surprisal values that reflect processing strategies that were either based on general language use or on specific speech registers. The word surprisal values for the different conditions are also part of DESRC. The results indicate that listeners use the wider context of speech register to adjust their word anticipations. This study indicates again that the DESCR contains valid data and can be used to answer research questions that cannot easily be addressed on the basis of other data sets.

4.2 Future research

The audio materials for the dialogue speech materials from the IFADV corpus are freely available and the audio materials for the news and read aloud books are freely available for researchers as part of the Spoken Dutch Corpus. The DESRC contains information on the location of the appropriate audio files. This implies that the corpus can be used by researchers from different fields to answer a range of research questions.

The large amount of data for each of the participants enables a study of individual differences in speech perception.

Further, new developments within automatic speech recognition, such Wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2023) allow for new analysis of the speech materials; the latent representations of these advanced models could be fruitfully applied to further analyze the EEG data in the DESRC.

The DESRC contains manually validated EEG artefact annotations. In Section 3, we described our approach to train and apply an automatic EEG artefact classifier. The set of artefact annotation could be used to further develop artefact or other EEG classifiers.

5 Conclusion

The Dutch EEG speech register corpus (DESRC) contains EEG recordings from participants listening to long (4 – 15 minutes) stretches of natural speech. The DESRC is available under license CC BY NC 4.0 and contains a rich set of meta-data with orthographic and phonemic transcriptions time-locked to the EEG data. We enriched the transcriptions with part-of-speech tags, word frequency and information theoretic measures such as word surprisal, entropy and cross entropy.

Furthermore, we annotated the EEG data with a novel automatic CNN artefact classifier. All automatically identified artefacts were manually checked. The artefact annotations allow easy exclusion of data contaminated with artefacts. The DESRC was already used to show that the mismatch between anticipated and actual word forms predicts the N200 amplitude (Bentum et al., 2019b) and to show that listeners anticipate words based on preceding words and speech register when listening to natural speech (Bentum et al., 2022). We hope that the dataset will be used to investigate many other (linguistic) phenomena.

Acknowledgements

We would like to thank Lou Boves and Tineke Snijders for their advice and Tim Zee for his help with artefact annotation.

Funding

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

Open Practices statement

The data and materials are available at:

https://data.ru.nl/collections/ru/cls/dutch_eeg_speech_register_corpus_dsc_807

The DOI of the dataset is: https://doi.org/10.34973/97pv-jw72

None of the experiments were preregistered.

References and links

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). Tensorflow: a system for large-scale machine learning. In: *OSDI 16*, 265-283.

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

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Bentum, M., Ten Bosch, L., Van den Bosch A., & Ernestus, M. (2019a). Do speech registers differ in the predictability of words? *IJCL 24*(1), 98-130. https://doi.org/10.1075/ijcl.17062.ben

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: *Proc. Interspeech 2019*, 2265-2269. https://doi.org/10.21437/Interspeech.2019-2741

Bentum, M., Ten Bosch, L., Van den Bosch A., & Ernestus, M. (2019c). Quantifying expectation modulation in human speech processing. In: *Proc. Interspeech 2019*, 2270–2274. https:/doi.org/10.21437/Interspeech.2019-2685

Bentum, M., Ten Bosch, L., van den Bosch, A., & Ernestus, M. (2022). Speech register influences listeners' word expectations. *Brain and Language, 235*, 105197. https://doi.org/10.1016/j.bandl.2022.105197

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and language*, 125(1), 82-93. https://doi.org/10.1016/j.bandl.2013.01.007

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. New York: Cambridge University Press.

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. https://doi.org/10.1016/0093-934x(90)90016-a

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. https://doi.org/10.1016/0093-934X(92)90018-A

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*, *134*(1), 9-21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *J. Phon. 48*, 60-75. https://doi.org/10.1016/j.wocn.2014.08.001

Federmeier, K.D. & Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang. 41*, 469–495. https://doi.org/10.1006/jmla.1999.2660

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 Lang., 140*, 1-11. https://doi.org/10.1016/j.bandl.2014.10.006

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkonen, L. & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446-460. https://doi.org/10.1016/j.neuroimage.2013.10.027

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research*, 1626, 118-135. https://doi.org/10.1016/j.brainres.2015.02.014

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. https://doi.org/10.1146/annurev.psych.093008.131123

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. https://doi.org/10.1145/3065386

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

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: fully automated statistical thresholding for EEG artifact rejection. *J. Neurosci. Methods*, *192*(1), 152-162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Oostdijk, N. (2002). The design of the spoken Dutch corpus. In *New frontiers of corpus research* (pp. 105-112). Brill. https://doi.org/10.1163/9789004334113_008

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intelli Neurosci*, *2011*, 1. https://doi.org/10.1155/2011/156869

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

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. *Hum. Brain Mapp.*, *38*(11), 5391-5420. https://doi.org/10.1002/hbm.23730

Van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: a Free Dialog Video Corpus. In *LREC* (pp. 501-508).

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

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cereb. Cortex*, 26(6), 2506-2516. https://doi.org/10.1093/cercor/bhv075

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 *EMBC* (pp. 4101-4105). IEEE. https://doi.org/10.1109/EMBC.2015.7319296