Decoding Part-of-Speech from Human EEG Signals

Alex Murphy[1], **Bernd Bohnet**[2], **Ryan McDonald**[3], **Uta Noppeney**[4]
[1]University of Birmingham, UK, [2]Google Research
[3]ASAPP, [4]Donders Institute for Brain, Cognition and Behaviour
murphyalex@gmail.com bohnetbd@google.com
rmcdonald@asapp.com uta.noppeney@donders.ru.nl

Abstract

This work explores techniques to predict Part-of-Speech (PoS) tags from neural signals measured at millisecond resolution with electroencephalography (EEG) during text reading. We show that information about word length, frequency and word class is encoded by the brain at different post-stimulus latencies. We then demonstrate that pre-training on averaged EEG data and data augmentation techniques boost PoS single-trial EEG decoding accuracy for Transformers (but not linear SVMs). Applying optimised temporally-resolved decoding techniques we show that Transformers outperform linear SVMs on PoS tagging of unigram and bigram data more strongly when information requires integration across longer time windows.

1 Introduction

Electro-/Magnetoencephalography (EEG/MEG), which measures neural activity at millisecond resolution, is a key neuroscientific method to assess how neural representations unfold dynamically in language processing. Early event related potential (ERP) studies that rely on averaging EEG activity across multiple trials have shown that EEG signal magnitude and topography depend on word length, frequency and open vs. closed class. Word length effects arose in EEG at about 150 ms, frequency effects at 200 ms and word class effects from 400-700 ms (Osterhout et al., 1997; Brown et al., 1999; Neville et al., 1992; Münte et al., 1998; Segalowitz and Lane, 2000; Münte et al., 2001; Dufau et al., 2015). Recent studies were able to predict these and other (e.g. semantic) aspects based on single trial multi-channel EEG/MEG activity (Ling et al., 2019; Chan et al., 2011; King et al., 2020). Importantly, the aim of cognitive neuroscience studies is to dissociate when (i.e. latency) and where (i.e. brain region) specific linguistic information is explicitly encoded in neural activity. Neuroscience studies therefore typically use linear decoders and try to disentangle neural activity for linguistic and non-linguistic dimensions that covary in natural language statistics (e.g. word class vs. length).

By contrast, engineering applications mainly aim to maximise performance accuracy, utilising all available information and more powerful non-linear classifiers. Intriguingly, recent studies have shown that adding human eye tracking data (Barrett et al., 2016) or morphosyntactic information extracted from human functional magnetic resonance imaging (fMRI) signals during sentence reading, can substantially improve PoS induction (Bingel et al., 2016). Yet, morphosynactic information obtained from fMRI is limited, because fMRI measures only the slow changes in blood oxygenation, peaking 5-6 s after stimulus onset, rather than the rapid neural activity during language processing.

Contributions. This interdisciplinary paper decodes PoS tags from EEG signals with linear SVMs and Transformers, pursuing several aims relevant for neuroscience and/or engineering.

Neuroscience-focused Section 3 uses linear SVMs to define the distinct neural representations of word length, frequency and class based on a new EEG data set, in which a single subject reads an extensive syntactically annotated corpus. To dissociate these linguistic and non-linguistic aspects, typically correlated in natural language statistics, Section 3 matched the stimulus distributions for each classification task with respect to the confounding dimensions of no interest. Consistent with previous reports, we show that word length, frequency and class can be decoded at different post-stimulus latencies based on single trial and trial-averaged data. This replication part serves to validate the new EEG data set.

Methods-focused Section 4 moves beyond open

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2201 - 2210
May 22-27, 2022 ©2022 Association for Computational Linguistics

vs. closed word class decoding that were the focus of previous MEG/EEG studies and decodes 6 PoS tags from EEG activity with linear SVMs and Transformers. We show that pretraining on trial-averaged data with subsequent fine-tuning on single-trial data, alongside data augmentation, boosts PoS decoding accuracy from single-trial EEG data selectively for Transformers (but not for linear SVMs).

Engineering-focused Section 5 finally uses linear SVMs and Transformers together with pretraining and augmentation techniques from Section 4 to assess how PoS information about unigrams and bigrams becomes progressively available in EEG activity across post-stimulus time. Comparing EEG decoding from sliding and incremental time windows suggests that Transformers outperform linear SVMs particularly when information needs being integrated across longer time windows. Our results raise the possibility of combining PoS-tagging based on EEG decoding with corpora and dependency tree annotation to obtain more reliable morphosyntactic information for low-resource languages.

2 General Methods

Our experiments used a new corpus annotated with EEG data, previously acquired at the University of Birmingham following ethical approval and participant's informed consent. The EEG annotated corpus is available[1] under a public license (CC BY-SA 4.0).

Data set. The stimulus set includes 4,479 sentences (74,953 tokens) selected from the English Web Treebank (Bies et al., 2012), covering the genres *weblogs, newsgroups, reviews and Yahoo Answers*. The mean sentence length is 16.7 words (standard deviation: 12.23). 75 sessions of EEG data are included over 20 days, each lasting 20-25 minutes, from a single subject who read approximately five and a half iterations of the stimulus set (i.e. 24,323 sentences and 404,205 tokens in total, thereby substantially exceeding current freely accessible data sets, e.g. (Bhattasali et al., 2020). Three sessions were excluded because of data corruption.

The EEG data for separate text passages were divided into training, dev and test sets to avoid any temporal overlap. Further, dev and test sets were matched for length of text passages, recording dates and sentence position (initial, mid & end). The training set contains 83% of the data (19,156 sentences; 317,753 tokens), dev set 8.5% (2,704 sentences; 45,822 tokens) and test set 8.5% (2,463 sentences; 40,630 tokens).

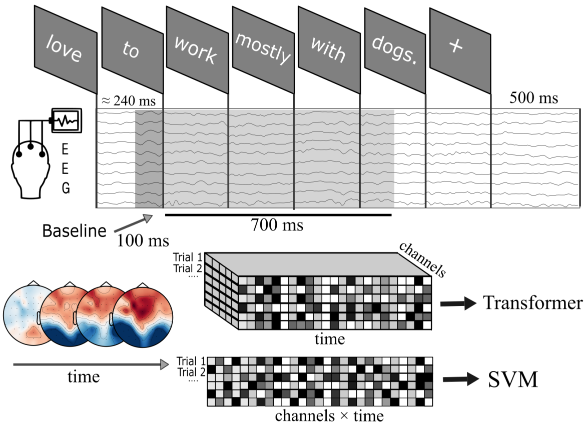

Figure 1: Example trial and EEG recording. Sentence words were presented on average approximately every 240 ms. EEG signals were extracted from -100 to 700 ms relative to word onset.

2.1 Experimental paradigm

In a Rapid Serial Visual Presentation (RSVP) paradigm, sentences were presented one word at a time, on average every ≈ 240 ms, in a white monospace font on a grey background approximately in the centre of the screen, at the optimal viewing position (Rayner et al., 2016). Each word subtended a horizontal angle 0.76° to the left and 11.81° to the right from the centre. Sentences were separated by 500 ms of a white central fixation cross (see Figure 1). On approximately 20% of the sentences in each session, the participant was prompted to verbalise the previous sentence back to the experimenter. An accuracy score of 93% across all sessions confirmed that the participant successfully attended the sentences. Stimuli were presented using PsychoPy (Peirce et al., 2019) on an LCD monitor with a resolution of 1920x1080 pixels and 60 Hz refresh rate. The subject's head was stabilised with a chin-rest.

2.2 EEG Data Acquisition and Preprocessing

Continuous EEG signals were recorded in 'reference-free' mode at a sampling rate of 1 kHz via BrainVision's PyCorder software using 64 Ag/AgCl active actiCAP slim electrodes arranged in a 10–20 layout (ActiCAP, Brain Products GmbH,

[1] https://edata.bham.ac.uk/617

tag	train	dev	test	total
ADJ	24,029	3,489	2,913	30,431
ADP	33,969	5,049	4,235	43,253
ADV	17,492	2,593	2,218	22,303
AUX	19,351	2,833	2,485	24,669
CCONJ	11,758	1,731	1,546	15,035
DET	31,429	4,589	3,962	39,980
INTJ	656	76	90	822
NOUN	59,991	8,691	7,501	76,183
NUM	5,062	712	677	6,451
PART	6,955	970	908	8.833
PRON	27,623	3,973	3,677	35,273
PROPN	28,867	3,737	3,641	35,245
PUNCT	3,716	485	501	4,702
SCONJ	7,116	1,046	943	9,105
VERB	39,710	5,723	5,186	50,619
X	1,029	125	147	1,301
total	317,753	45,822	40,630	404,205

Table 1: Number of samples for each PoS tag across train, dev and test set

Gilching, Germany). Channel impedances were kept below 15 kΩ.

Data were preprocessed using MNE-Python (Gramfort et al., 2013). Individual EEG sessions were band-pass filtered between 1-40 Hz, downsampled to 250 Hz and re-referenced to average reference. Noisy channels were determined based on visual inspection and interpolated. Non-neuronal components (e.g. ocular, muscular, electrical) were removed via Independent Component Analysis (ICA) individually for each recording session (an average of 4 components were removed per EEG session).

EEG signals were extracted from -100 to 700 ms relative to word onset. For baseline correction, we subtracted the channel-wise mean from -100 ms to 0 ms from the evoked post-stimulus EEG response ([0 700] ms separately for each word; see Figure 1). EEG data were spatially multivariate noise normalised using the noise covariance matrix estimated separately for each target class (Guggenmos et al., 2018). Each EEG trial was annotated with the gold part of speech tags of the current and subsequent words, their word lengths, and Zipf-logarithmic frequency scores from the Python package *WordFreq* (Speer et al., 2018).

2.3 Models and EEG decoding

parameter	value	parameter	value
encoder layers	4	mlp size	1024
learning rate	0.04	mlp dropout	0.1
batch size	16	qkv size	512
warm up steps	50k	attention heads	8
training steps	400k	attention dropout	0.1
Adam β_1	0.9	Adam β_2	0.98
Adam ϵ	10^{-9}	Adam weight decay	0.0

Table 2: Hyperparameters of the Transformer model

The decoding analyses used linear support vector machines (SVM) (Chang and Lin, 2011) and Transformers, which can capture complex interactions of EEG data across time points. All classifiers were trained on the EEG training data, assessed on the dev set and scored on the independent test set. Hyperparameters and early stopping were selected based on the dev set. We assessed linear SVMs and Transformers on the dev set using 10 different random seed points. We show mean classification accuracy with 68% confidence intervals (CI) over those 10 replications on the dev (Table 4, Figure 3) resp. test set (Figure 2, 4, 5). We compute statistics on test set classification responses from the model that scored the highest on the dev set (e.g. binomial or Wilcoxon signed rank tests).

For linear SVM, we used an online learning implementation of SCIKIT-LEARN (Pedregosa et al., 2011; Zhang, 2004), based on LIBSVM (Chang and Lin, 2011), with hinge loss and Stochastic Gradient Descent (SGD) optimiser. Hyperparameters were set to default except for the SGD regularisation parameter that was increased to $\alpha = 0.75$, which provided better classification accuracy on the dev set. The parameter α is inversely proportional to the C parameter in the standard SVM implementation. The online implementation also allowed us to select the best model using early stopping. The SVM was provided with EEG activity vectors as inputs, i.e. 1 x (EEG channels × time points).

For the Transformer (Vaswani et al., 2017), we conducted a model architecture and hyperparameter search (layers, learning rate, MLP dimensions, dropout rate, Encoder vs. Encoder-Decoder) on the dev set. The selected model was composed of four encoder-blocks and a final dense layer that projects the output of the last encoder-block onto the PoS tags via a softmax function. We used the Adam optimiser and early stopping. The implementation is based on the WMT example[2] of Google's novel ML frameworks *Flax/Jax*. Table 2 lists the selected hyperparameters. The Transformer received EEG channels x time points as inputs and provided a classification response for the entire time window.

We performed decoding based on (i) EEG for single-trials (i.e. no averaging), (ii) EEG averaged across three and (iii) ten trials. Averaging EEG

[2] https://github.com/google/flax/tree/master/examples/wmt

signals across trials increases the signal to noise ratio of the 'samples' (Grootswagers et al., 2017; Guggenmos et al., 2018; Roy et al., 2019; Tuckute et al., 2019), but ignores true variability across EEG data from different words of the same category (Münte et al., 2001). For training (resp. dev) set we generated the same number of samples for 3 and 10 trial averages as for the single trial test (resp. dev) sets via boostrapping. For the test set, we averaged data without replacement, so that examples can be entered as independent data points in statistical tests. Hence, the number of samples in the test set (but not in the training or dev sets) is smaller for 3 and 10 trial averages than single trials (Table 3).

3 Decoding word class, length, frequency

For comparison with previous research (Osterhout et al., 1997; Münte et al., 2001), we decoded word length, frequency and class with linear SVMs in a temporally-resolved fashion from 0 to 700 ms post-stimulus EEG, recorded during sentence reading.

Data set. We decoded word class from EEG via binary classification between open class words (i.e. NOUN, VERB, ADJ, and PROPN) vs. closed class words (i.e. DET, ADP, AUX, PRON, SCONJ and CCONJ). Likewise, for decoding word frequency and length, words were assigned to two classes based on the median values in the data set (i.e. Zipf-frequency > 5.91 = HIGH else LOW; word length > 4 characters = LONG, else SHORT). EEG decoding analyses were performed for single-trials, averages across 3 and 10 trials. To dissociate the distinct contributions word length, frequency and class that are highly correlated in natural language statistics, we decoded one variable by controlling for the other two variables. For instance, when decoding open vs. closed class words, we selected a subset of trials such that the joint distributions over the three confounding variables of word frequency (discretised to the nearest 0.25), length (number of characters) and sentence position (i.e. sentence initial, mid, end) were equated for the categories of open and closed class words.

To minimise confounds arising from the preceding word in the sentence, we balanced the test set with respect to the open/closed class status of the previous word. Similarly, we controlled the decoding of word frequency for word length, and the analysis of word length for word frequency, and both analyses for open/closed class and sentence position. Table 3 gives the number of examples for each analysis across training, dev and test sets.

	length	frequency	class
train	82,424	51,364	45,502
dev	12,402	7,590	5,670
test (single)	10,810	6,590	5,670
test (avg. 3)	3,603	2,196	1,890
test (avg. 10)	1,081	659	567

Table 3: Number of samples across train, dev and test set in the confound-controlled data set.

Methods. To temporally resolve how the brain encodes word length, frequency and class, we trained and tested linear SVMs on EEG signals separately for sliding windows of 64 ms (i.e. 16 time points) that shift in increments of 4 ms (i.e. one time sample). Figure 2 shows the mean accuracy values (averaged across 10 seed points) from the test set (centred on the last bin of each time window (Grootswagers et al., 2017)) with ± 68% CI. The classification responses for the test set from the model that performed best on the dev set were entered into a two-sided binomial test, separately for each time window. Solid lines in Figure 2 above the decoding accuracy time courses indicate time points that were significant at $(p < 0.05)$ False Discovery Rate (FDR) corrected for multiple comparisons (Rouam, 2013) across time (i.e. 160 tests).

Results. Figure 2 (top rows of A, B, C) shows butterfly plots for the effects of word length, frequency and class across 64 electrodes. Our linear SVM decoding analysis replicates the temporal cascade of word length, frequency and class effects previously reported for EEG responses averaged across a large number of trials. The word length effect arises early at about 100 ms, previously associated with visual word processing in occipitotemporal cortices (Hauk and Pulvermüller, 2004; Pulvermüller et al., 2009; Schuster et al., 2016). Word frequency influenced neural processing later from 200 ms onwards with a slight left-hemispheric predominance (Griffiths et al., 2012). The word class effect emerged in early and late time windows with the effect at about 550 ms in line with the well-known P600 as an ERP indicator for syntactic processing (Osterhout and Holcomb, 1992; Hagoort et al., 1993; ter Keurs et al., 1999). Word length and frequency effects were stronger than the word class effect; see King et al. (2020). As expected, decoding accuracy increased when EEG signals were averaged across trials. Thus, carefully controlling each comparison of interest (e.g. word class) for the effects of no interest (e.g. word length and frequency) enabled us to dissociate word length, frequency and class effects, despite their high correlation in natural language, thereby validating our new annotated EEG data set and analysis procedure.

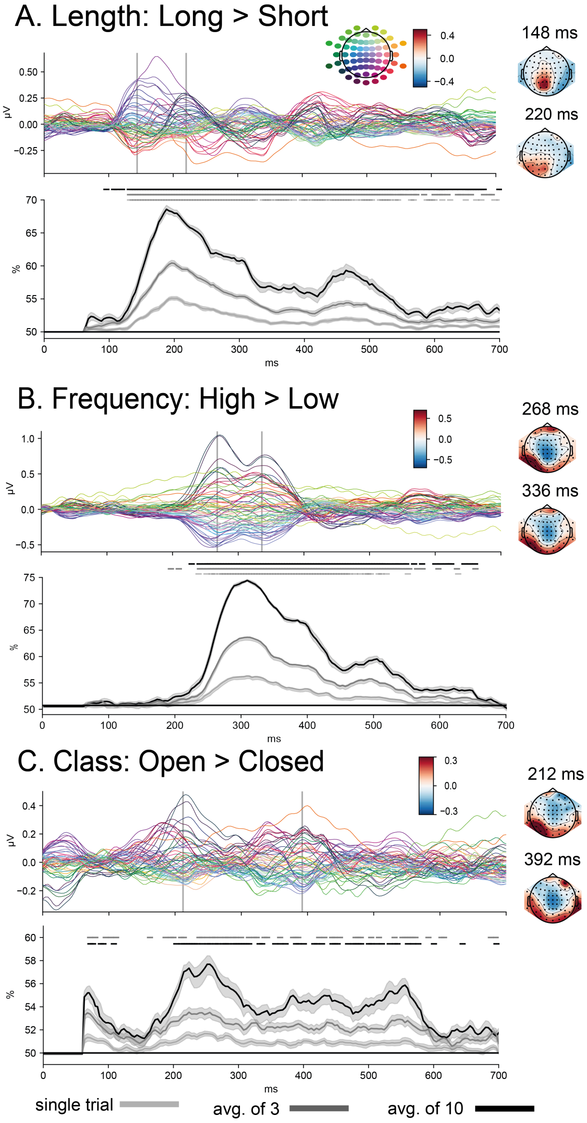

Figure 2: Butterfly plots for difference waves across all 64 channels (top rows, left), topographies (top rows, right) and time courses of decoding accuracy (bottom rows). (A) Length: LONG > SHORT, (B) Frequency: HIGH > LOW. (C) Class: OPEN > CLOSED. Each line in the butterfly plots represents an EEG channel, colour coded by its position. The EEG topographies are shown for the time points as indicated by the vertical lines. The mean decoding accuracy time courses ± 68% CI on the test set are shown for single-trials, averages of 3 and 10 trials. The horizontal lines (above the time courses) indicate the time points with decoding accuracy significantly different from chance at $p < 0.05$ FDR-corrected for multiple comparisons across time. Chance accuracy is denoted with a black line at 50%.

4 Improving training methods for PoS decoding

Moving beyond open/closed word class decoding, we assessed whether multi-class PoS decoding with SVMs and/or Transformers can be improved by (i) data augmentation, i.e. increasing the number of samples in the training set via bootstrapping and re-averaging (only applicable to 3 and 10 trial averages) and (ii) pretraining on trial-averages followed by fine-tuning of the model parameters on single-trial EEG data.

Data set. We focused on decoding of 3 open class (NOUN, VERB, PROPN) and 3 closed class (ADP, DET, PRON) PoS tags. From the word class dataset that was controlled for length and frequency effects (section 3), we selected an equal number of examples per PoS class (i.e. train: 3,470, dev: 335, test: 335, i.e. in total ≈ 20k data points; word frequency of each word > median frequency Zipf value (5.91). The samples for dev and test sets were matched for distribution of word lengths.

4.1 Data Augmentation

Methods. Using this 6-class unigram dataset, we assessed whether data augmentation via bootstrapping and re-averaging increases decoding performance for the 3 and 10 trial averages. We sampled 3 (resp. 10) individual trials with replacement from a particular PoS class and averaged them in 3 (resp. 10) trial averages. We thus trained SVMs and Transformers (over 20 random seeds) on 4 training set sizes: $N_{size} = \{20k, 100k, 250k, 500k\} \times 2$ levels of trial averaging (3 vs. 10) resulting in 8 training sets. The baseline training (resp. dev) set included as many 3 (resp. 10) trial averages as the initial single-trial training (resp. dev) set.

Results. Data augmentation systematically boosted the decoding accuracy of the Transformer but not of the SVM - most likely because of the former's greater model complexity. For both 3 and 10 trial averages the Transformer's decoding accuracy on the dev set increased from a training set size of 20k to 100k, peaking at 250k. It then declined for an even larger training size of 500k - potentially because continued bootstrapping pro-

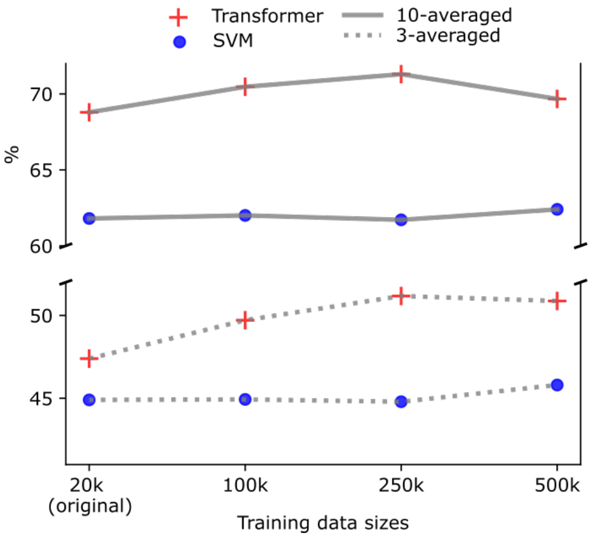

Figure 3: Dev set decoding accuracy (mean across seeds) for SVM (blue) and Transformer (red) separately for 3 and 10 trial averages and four levels of data augmentation: 20k (original), 100k, 250k and 500k. Chance decoding accuracy is 16.66%.

gressively generates dependencies amongst training samples thereby limiting their additional benefit beyond 250k. We formally assessed whether the Transformer that scored best on the dev set obtained better decoding accuracy for 250K than for the original 20k training set (n.b. we performed this statistical test on the test set, because the 3 and 10 trial averages within the dev set were not independent from one another as a result of boostrapping). Indeed, for both 3 and 10 trial averages the Transformer's (but not the SVM's) decoding accuracy was significantly better for 250k than the original 20k training set ($p < 0.01$; Wilcoxon signed-rank test).

4.2 Pretraining

Methods. The ultimate goal is to decode PoS from single-trial EEG data (rather than trial averages). We therefore assessed whether pretraining the SVM and/or Transformer on trial averages with subsequent fine-tuning on single-trial data increases decoding accuracy. Pretraining may be beneficial because trial averages have a greater signal to noise ratio. Specifically, we assessed the impact of pretraining in a 2×2 factorial design manipulating i) pretraining scheme: training in three steps (10-3-1) from 10 trial averages to 3 trial averages to single-trials vs. training in two steps (3-1) from 3 trial averages to single-trials and ii) data augmentation: training only on the original 20k vs. the 250k data set, which obtained highest dev set performance in section 4.1. We trained both SVMs and Transformers on the 2× 2 training conditions using 20 random seeds and report mean accuracy (± 68% CI, across those 20 seeds) in Table 4.

	SVM	Transformer
single-trials	31.93 (±0.62)	37.15 (±0.32)
10-3-1	31.74 (±0.51)	38.5 (±0.28)
10-3-1 (250k)	31.89 (±0.67)	39.17 (±0.33)
3-1	32.03 (±0.52)	37.83 (±0.24)
3-1 (250k)	31.79 (±0.58)	39.41 (±0.41)

Table 4: Single-trial decoding accuracies (%, mean across seeds ± 68% CI) on dev set for the SVM and Transformer: without pretraining, with 10-3-1 pretraining, with 10-3-1 and 250k data augmentation, with 3-1 pretraining, with 3-1 and 250k data augmentation

Results. For the SVM, the 3-1 pretraining without data augmentation resulted in the highest dev set accuracy (32.03%), though accuracy was only slightly better than for direct single-trial training (31.93%). For the Transformer, the 3-1 pretraining scheme with 250k data augmentation obtained the highest single-trial decoding accuracy (39.41%) on the dev set. Indeed, Wilcoxon signed-rank test (Pereira et al., 2009) confirmed that the best dev set Transformer performed significantly better on the test set after 3-1 pretraining than after direct single-trial training ($p < 0.01$).

5 Temporally-resolved PoS decoding

Sections 3 and 4 were driven by the neuroscience goal of dissociating neural representations associated with PoS from confounding factors such as word length or frequency, which are typically correlated with PoS in natural language statistics. To control for these confounds sections 3 and 4 generated data sets, in which e.g. PoS classes were equated with respect to word length. By contrast, Section 5 pursues the engineering goal of maximising PoS decoding accuracy. Here, correlations between word frequency, length and PoS class are no longer considered a confound but a useful source of information. Capitalising on the optimised 3-1 training scheme with 250k data augmentation from Section 4, section 5 will assess whether PoS information about unigrams and bigrams can be decoded from EEG signals (without any confound controls). For both unigrams and bigrams, we will first investigate how PoS information becomes available dynamically across post-stimulus time by training SVMs and Transformers on EEG signals from 64 ms sliding time windows. Second, we will assess how SVMs and Transformers integrate PoS information across post-stimulus time by training them on EEG signals from incremental time windows.

5.1 Unigrams

Data Set. We selected an equal number of examples for the 6 most frequent PoS tags (i.e. `NOUN`, `VERB`, `ADP`, `DET`, `PRON` & `PROPN`) from the data set matched for length of text passage, recording dates and sentence position, but not for word length or frequency. Each PoS class included the following number of samples - train: 28,265 , dev: 2,948, test: 3,183 examples.

Methods.We implemented the 3-1 pretraining with 250k data augmentation (section 4). For the sliding window analysis, we trained and tested SVMs and Transformers on EEG signals from 64 ms windows (i.e. 16 time points) that shifted in increments of 16 ms from 0 ms to 700 ms (i.e. resulting in a sequence of 41 decoding accuracies). For the incremental window analysis, we successively increased the initial [0 16] ms time window (i.e. 4 samples) by 4 additional sampling points resulting in a temporal sequence of 44 decoding accuracies. We computed decoding accuracies (mean across seeds, ± 68% CI) from the test set. Across time windows we compared the decoding accuracies on the test set of the best dev set SVM and Transformer using the Wilcoxon signed rank-test (reported at $p < 0.05$, FDR-corrected for multiple comparisons across time i.e. 41 resp. 44 tests).

Results. In the sliding window analysis, the decoding accuracies of SVMs and Transformers show two prominent peaks at 200 ms and 400 ms suggesting that PoS decoding relies on several aspects of information encoded in the EEG. Based on our confound-controlled analysis (section 3) the first peak reflects word length and frequency information, while the second peak is more closely related to semantic and syntactic aspects of the word. The incremental window analysis showed an accuracy benefit of 4.5% for the Transformer starting in the very first [0 16] ms time window. This difference in performance between the two models further widened, reaching a maximum advantage of 11.6% around 360 ms. Transformers thus benefit from integrating information about word frequency, length and class that arise at different post-stimulus latencies. Moreover, because PoS classes of subsequent words are correlated in natural language, the Transformer may also benefit via self-attention from information about the next word that is presented and progressively encoded in EEG activity about 240 ms after the current (i.e. to be decoded) word. Statistical testing confirmed that the Transformer significantly outperformed the SVM for all sliding and incremental windows (Wilcoxon-signed rank test, FDR-corrected at $p < 0.05$).

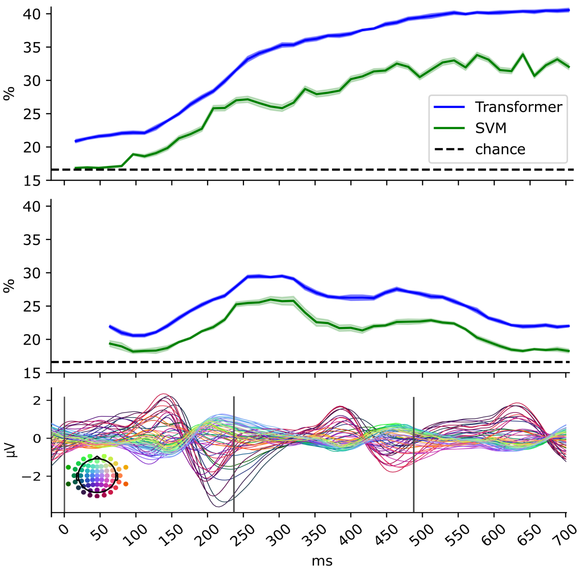

Figure 4: Unigram results: Test set decoding accuracies (mean across seeds ± 68% CI), aligned with last bin of each time window. Top: Incremental window analysis. Middle: Sliding window analysis. Bottom: ERP for `NOUN`, i.e. EEG averaged across all examples from the training set. Vertical lines indicate word onset times. All time windows are significant at $p < 0.05$ FDR-corrected.

5.2 Bigrams

To define the contributions of successive words to EEG PoS decoding in naturalistic text reading, we designed a bigram data set that artificially removes the correlations between PoS classes of subsequent words, though we note that this does not fully remove correlations between specific word tokens.

Data set. We selected 6 bigrams, in which each first word's PoS is combined equally often with two different PoS classes from the second word: `NOUN-PRON`, `NOUN-VERB`, `PRON-NOUN`, `PRON-VERB`, `VERB-NOUN` and `VERB-PRON`. As a result, the PoS class of word 1 is uninformative about the PoS class of word 2 and vice versa. Hence, prior to the presentation of word 2, the maximal possible decoding accuracy for a particular bigram is 50%. Each bigram class included the

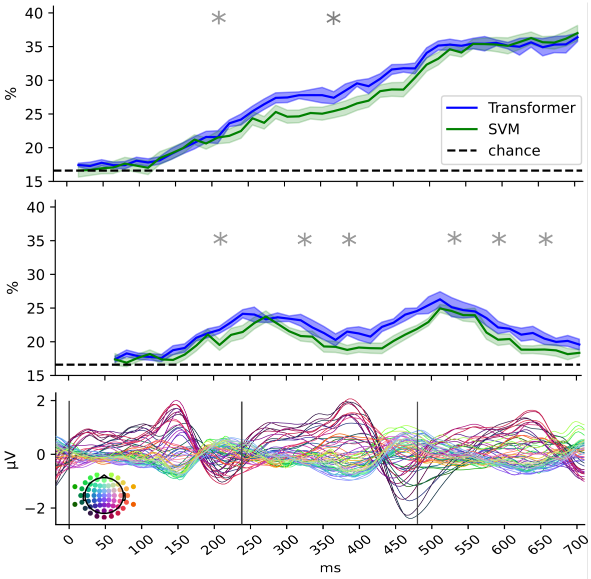

Figure 5: Bigram results: Test set decoding accuracies (mean across seeds ± 68% CI), aligned with last bin of each time window. Top: Incremental window analysis. Middle: Sliding window analysis. Bottom: ERP for VERB-PRON, i.e. EEG averaged across all examples from the training set. Vertical lines indicate word onset times. Asterisks (aligned with end of time window) indicate significance at $p < 0.05$ FDR-corrected.

following number of samples - train: 3,470, dev: 322, test: 349 examples.

Methods for details see unigram analysis.

Results Similar to the unigram results, the sliding window analysis revealed two prominent accuracy peaks at 200 ms and 500 ms. Yet, the 2nd peak was slightly later than in the unigram analysis and it was higher than the first peak only in the bigram, but not the unigram analysis. These differences between unigram and bigram decoding profiles arise, because EEG at 500 ms encodes semantic and syntactic aspects of word 1 and crucial information about word 2 of the bigram. As shown in Figure 5, the Transformer significantly outperformed the SVM in the sliding and incremental window analyses. Yet, in contrast to the unigram results, the Transformer outperformed the SVM in the incremental analysis only for windows 0-208 ms and 0-336 ms. The Transformer's smaller benefit arises mainly because our balanced design radically reduced the number of examples and thereby the Transformer's generalisation ability. It also removed the natural correlations between subsequent words on which the Transformer may have additionally capitalised in the unigram data.

6 Related Work

The confound-controlled analysis dissociated word length, frequency and class effects in EEG. This replication of earlier ERP (Osterhout et al., 1997; Münte et al., 1998) and MEG decoding work (King et al., 2020) validates a new EEG data set for an extensive morphosyntactically gold annotated corpus; c.f. Bhattasali et al. (2020). Transformers successfully decoded 6 PoS tags from single trial EEG with data augmentation and 3-1 pretraining ($\approx 40\%$ accuracy), raising the possibility to boost PoS induction with EEG-decoded PoS tags. While we acknowledge that our results are limited to EEG data from a single subject, given the spatial smoothness of EEG scalp topographies, we envision pretraining on EEG obtained from different participants. Further, because human brains generate similar neural signatures for word classes across different languages (c.f. Yudes (2016); Münte et al. (2001); Hagoort et al. (2003)), pretraining PoS-EEG decoders on large morphosyntactically annotated EEG datasets for English followed by fine-tuning on a smaller annotated EEG data set for a low-resource language may enable successful generalisation to EEG obtained from reading non-annotated texts in this low-resource language. PoS-induction jointly based on annotated texts and EEG signals could thus be transformative for corpus generation of low-resource languages.

7 Conclusion

Combining neural signals measured at millisecond resolution with EEG and a linguistically annotated corpus, this work shows - to the best of our knowledge - the first time that unigram and bigram PoS tags can be decoded successfully from single-trial EEG data. Temporally-resolved EEG decoding unraveled how information about linguistic and non-linguistic aspects evolved dynamically across time. Unsurprisingly, Transformers with self-attention mechanisms outperformed SVMs across all experiments. In particular, they benefited from integrating information across time, data augmentation and pretraining methods. Our work paves the way for future applications that incorporate human brain signals in traditional NLP methods.

References

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The alice datasets: fMRI & EEG observations of natural language comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 120–125, Marseille, France. European Language Resources Association.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium*.

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755, Berlin, Germany. Association for Computational Linguistics.

Colin M. Brown, Peter Hagoort, and Mariken ter Keurs. 1999. Electrophysiological signatures of visual lexical processing: Open-and closed-class words. *Journal of Cognitive Neuroscience*, 11(3):261–281.

Alexander M. Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S. Cash. 2011. Decoding word and category-specific spatiotemporal representations from meg and eeg. *NeuroImage*, 54(4):3028–3039. Date revised - 2012-03-01; Last updated - 2014-02-21; SubjectsTermNotLitGenreText - Semantics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Stéphane Dufau, Jonathan Grainger, Katherine J. Midgley, and Phillip J. Holcomb. 2015. A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological Science*, 26(12):1887–1897. PMID: 26525074.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. 2013. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267.

John D. Griffiths, William D. Marslen-Wilson, Emmanuel A. Stamatakis, and Lorraine K. Tyler. 2012. Functional Organization of the Neural Language System: Dorsal and Ventral Pathways Are Critical for Syntax. *Cerebral Cortex*, 23(1):139–147.

Tijl Grootswagers, Susan G. Wardle, and Thomas A. Carlson. 2017. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, 29(4):677–697.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. 2018. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *NeuroImage*, 173:434 – 447.

P. Hagoort, M. Wassenaar, and C. M. Brown. 2003. The time-course of processing of grammatical class and semantic attributes of words: Dissociation by means of erp. *Syntax-related ERP-effects in Dutch. Cognitive Brain Research*, 16(1):38–50.

Peter Hagoort, Colin Brown, and Jolanda Groothusen. 1993. The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, 8(4):439–483.

Olaf Hauk and Friedemann Pulvermüller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090 – 1103.

Jean-Rémi King, François Charton, David Lopez-Paz, and Maxime Oquab. 2020. Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, 220:117028.

Shouyu Ling, Andy C. H. Lee, Blair C. Armstrong, and Adrian Nestor. 2019. How are visual words represented? insights from eeg-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, 40(17):5056–5068.

Thomas F. Münte, Kolja Schiltz, and Marta Kutas. 1998. When temporal terms belie conceptual order. *Nature*, 395(6697):71–73.

Thomas F Münte, Bernardina M Wieringa, Helga Weyerts, Andras Szentkuti, Mike Matzke, and Sönke Johannes. 2001. Differences in brain potentials to open and closed class words: class and frequency effects. *Neuropsychologia*, 39(1):91 – 102.

Helen J. Neville, Debra L. Mills, and Donald S. Lawson. 1992. Fractionating Language: Different Neural Subsystems with Different Sensitive Periods. *Cerebral Cortex*, 2(3):244–258.

Lee Osterhout, Michael Bersick, and Richard McKinnon. 1997. Brain potentials elicited by words: word length and frequency predict the latency of an early negativity. *Biological Psychology*, 46(2):143 – 168.

Lee Osterhout and Phillip J Holcomb. 1992. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785 – 806.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Francisco Pereira, Tom Mitchell, and Matthew Botvinick. 2009. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209. Mathematics in Brain Imaging.

Friedemann Pulvermüller, Yury Shtyrov, and Olaf Hauk. 2009. Understanding in an instant: Neurophysiological evidence for mechanistic language circuits in the brain. *Brain and Language*, 110(2):81 – 94.

Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. 2016. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1):4–34. PMID: 26769745.

Sigrid Rouam. 2013. False discovery rate (fdr). In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 731–732. Springer New York, New York, NY.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001.

Sarah Schuster, Stefan Hawelka, Florian Hutzler, Martin Kronbichler, and Fabio Richlan. 2016. Words in Context: The Effects of Length, Frequency, and Predictability on Brain Responses During Natural Reading. *Cerebral Cortex*, 26(10):3889–3904.

Sidney J. Segalowitz and Korri C. Lane. 2000. Lexical access of function versus content words. *Brain and Language*, 75(3):376 – 389.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Mariken ter Keurs, Colin M. Brown, Peter Hagoort, and Dick F. Stegeman. 1999. Electrophysiological manifestations of open- and closed-class words in patients with Broca's aphasia with agrammatic comprehension: An event-related brain potential study. *Brain*, 122(5):839–854.

Greta Tuckute, Sofie Therese Hansen, Nicolai Pedersen, Dea Steenstrup, and Lars Kai Hansen. 2019. Single-trial decoding of scalp EEG under natural conditions. *Comput. Intell. Neurosci.*, 2019:9210785:1–9210785:11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Alberto Cuetos Fernando de Vega Manuel Yudes, Carolina Domínguez. 2016. The time-course of processing of grammatical class and semantic attributes of words: Dissociation by means of erp. *Psicológica*.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 116, New York, NY, USA. Association for Computing Machinery.