Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), pages 43–49
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The Little Prince in 26 Languages: Towards a Multilingual Neuro-Cognitive Corpus

Sabrina Stehwien*, Lena Henke*, John T. Hale♠, Jonathan R. Brennan♣, Lars Meyer*
*Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, DE
♠University of Georgia, Athens, GA, USA
♣University of Michigan, Ann Arbor, MI, USA
{stehwien, henke, lmeyer}@cbs.mpg.de, jthale@uga.edu, jobrenn@umich.edu

Abstract

We present the *Le Petit Prince* Corpus (LPPC), a multi-lingual resource for research in (computational) psycho- and neurolinguistics. The corpus consists of the children's story *The Little Prince* in 26 languages. The dataset is in the process of being built using state-of-the-art methods for speech and language processing and electroencephalography (EEG). The planned release of LPPC dataset will include raw text annotated with dependency graphs in the Universal Dependencies standard, a near-natural-sounding synthetic spoken subset as well as EEG recordings. We will use this corpus for conducting neurolinguistic studies that generalize across a wide range of languages, overcoming typological constraints to traditional approaches. The planned release of the LPPC combines linguistic and EEG data for many languages using fully automatic methods, and thus constitutes a readily extendable resource that supports cross-linguistic and cross-disciplinary research.

1. Introduction

We present the *Le Petit Prince* Corpus (LPPC), a multi-lingual resource for experimental research in cross-linguistic (computational) psycho- and neurolinguistics. The corpus consists of translations of the children's story *Le Petit Prince* (*The Little Prince*), published by Antoine de Saint-Exupéry in 1943, in 26 languages. The corpus is built by combining current methods from speech and language technology, that is, state-of-the-art Text-to-Speech Synthesis (TTS) and dependency parsing, as well as electroencephalography (EEG).

This paper describes ongoing work. We describe the resource that we will release as well as the important aspects to consider while building this corpus. The final release of the dataset will include three main parts: The primary written data is given as raw text and annotated with dependency graphs in the Universal Dependencies (UD) standard (Nivre et al., 2016). A subset of the corpus will be provided as time-aligned synthetic speech. The speech data will be used as an auditory stimulus for recording EEG data, which comprises the final part of the release.

1.1. Motivation

Traditional psycho- and neurolinguistic research has employed factorial experimental designs that require a large number of trials with highly controlled stimuli. Such experimental designs thus limit the generalizability of findings, and it has been increasingly acknowledged in recent years that factorial experiments lack sufficient statistical power and ecological validity (Brennan, 2016; Willems et al., 2015). For this reason, more and more studies rely on naturalistic stimuli (Hamilton and Huth, 2018).

An additional shortcoming of factorial experiments is evident from recent findings in probabilistic language processing: Repetitive presentation of large numbers of matched stimuli can have the undesired effect of changing transitional probabilities during the experiment and thus, of obscuring neurobiological results (Kroczek and Gunter, 2017). The development of information-theoretic quantifications of speech and language processing (Hale, 2001) and their excellent fit to behavioral (Levy, 2008; Demberg and Keller, 2008) and neurobiological data (Hale et al., 2018; Rabovsky et al., 2018; Frank et al., 2015) supports this.

Traditional psycho- and neurolinguistic studies have typically been restricted to single or few individual languages. This results in limited generalizability beyond small typological domains, thereby hindering the understanding of cross-linguistic commonalities and differences in the cognitive apparatus and neural substrate of speech and language processing (Kandylaki and Bornkessel-Schlesewsky, 2019).

In contrast, the LPPC as a resource facilitates generalization across a range of languages (Kandylaki and Bornkessel-Schlesewsky, 2019), helping the psycho- and neurolinguistic fields to further overcome their current statistical and typological limitations. The motivation for building this dataset is in line with the recent development of openly accessible naturalistic stimulus sets in the neurolinguistic community, such as the *Mother of All Unification Studies* (Schoffelen et al., 2019), the *Narrative Brain Dataset* (Lopopolo et al., 2018), the *Alice Datasets* (Bhattasali et al., 2020) and the ongoing *Alice in Language Localizer Wonderland* project[1]. Unlike factorial and/or monolingual experimental datasets that are tailored to just one specific question, the LPPC's lexico-syntactic annotation in the UD standard fosters research that addresses a broad range of linguistic research questions. The LPPC is also sustainable in that its data is amenable to future re-analysis that addresses future research questions. Furthermore, the use of the dataset will facilitate the formulation of neurobiological frameworks that generalize across languages (Bornkessel-Schlesewsky and Schlesewsky, 2016), assuming that the structural and functional properties of the human brain that subserve language are shared among speakers of all languages (Futrell et al., 2015; Levy, 2008; Brennan et al., 2019). In turn, the dataset may also serve as re-

[1] https://evlab.mit.edu/alice

source for traditional linguistic research that aims to explain why languages are different, yet they all can be processed by brains that are unitary across humans.

1.2. The LPPC – an automatic corpus

Recent advancements in the field of speech and language processing, fueled by the striking success of deep learning models, have made it feasible to automatically create and annotate large amounts of data with a higher quality than previously possible. We exploit such methods for building our resource, that, given it comprises 26 languages, would require much effort using traditional manual methods. Apart from the primary text data, which is manually cleaned, the database is created using automatic dependency parsing, forced-alignment and speech synthesis. In addition to the state-of-the-art speech and language processing tools employed for building the corpus, the EEG data is preprocessed using a fully automatic pipeline setup. We are also planning to make the EEG data available to the community in an open format that facilitates further processing. To the best of our knowledge, the LPPC is the first resource for neurolinguistic research that is not only created by, but also combines such methods.

2. The LPPC multi-lingual resource

The corpus consists of translations of the children's story *Le Petit Prince* by Antoine de Saint-Exupéry. The text was originally written in 1943 and has since been translated into over 300 languages[2].

The languages chosen for the LPPC are Arabic, Chinese (Mandarin), Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Slovak, Spanish, Swedish, Turkish, Ukrainian, and Vietnamese.

The criteria for choosing these languages was their availability both as a significantly large treebank in the UD treebank (Nivre et al., 2016) (to allow for uniform syntactic parsing across languages) as well as in Google's Text-to-Speech API[3] voice selection. Both tools are part of automatic pipelines for creating the linguistic annotations and the speech data, respectively.

2.1. Primary written data

The primary data in the LPPC consists of one text version of the story in each of 26 chosen languages. The full story comprises 27 chapters in total (plus a short prologue) and the English version amounts to roughly 16k words. The LPPC includes the first six chapters in each language as spoken data, amounting to around 20 minutes of speech ($\approx$ 250 sentences).

We chose existing published translations of the story. Since the domain of the data is literary text, the versions for the various languages cannot be expected to be translated directly at the sentence level. Furthermore, we do not have any control on how close the different translations are to the French original, and we expect expect a certain degree of variation between the different translations. Nevertheless, given the fact that the book follows a clear story line and uses rather simple language, we consider the translations to be fairly parallel. The LPPC is therefore not a strictly parallel corpus, but a combination of comparable parts as well as parallel, but unaligned sentences[4].

2.1.1. Acquisition of text

For the written text part of the corpus, we acquired electronic translations of the text. In most languages, multiple translations have been published since the first issue, and newer translations continue to appear until today. Therefore, we carefully chose versions according to the following criteria: The first being the availability as an e-book[5], as these are readily obtained and easily converted to raw text. The second criterion was the availability of bibliographic data. Since the texts available on web differ in quality, we selected releases that contained information on the translator, year, and publishing house. We also discussed the choice of text with native speakers in cases where we were unsure about the quality of the translated versions.

2.1.2. Choice of translation

We placed an additional constraint on our choice of translations based on diachronic linguistic changes that may pose additional interfering factors during neuroscientific studies. Such problems may arise, for example, when performing EEG studies on canonical participant samples, e.g. in an age range of 20 to 30 years. Since participants in this age group are less familiar with the writing style used in the original version from 1943 and the early translations, we chose to collect more recent translations for the corpus. This decision was based on the outdated language in older translations, which may confound experimental measurements. For example, the Hungarian version uses the obsolete term *fölnőtt* for the word *grown-up*, whereas the new version uses *felnőtt*. An additional concern was the use of literary writing style, which has changed considerably over the years. Old words or syntactic constructions may be unfamiliar to participants and thus be experienced as unusual, thereby triggering meta-cognitive processing.

Conversely, due to the fact that *Le Petit Prince* is a well-known text, we expect that participants who are very familiar with the original translations may also display interference effects when confronted with a different translation. Therefore, this choice comprises a trade-off between a familiar story and a contemporary language style. To keep the corpus as consistent as possible, however, we chose the newest translation available that fits the aforementioned criteria. Apart from the French original text, the full collection therefore contains translations that were published after the

[2]It has thus been referred to as the most-translated non-religious text in the world (Le Figaro, 7. April 2017).

[3]https://cloud.google.com/text-to-speech/

[4]Cross-language sentence alignments may be carried out by hand to a limited extent. The research questions we seek to address with this corpus, discussed in section 3.3.3. do not require a strictly parallel corpus, and we therefore do not plan to include such alignments in this release.

[5]Obtaining digitized text from print versions was deemed too error-prone due to expected issues in using optical character recognition (OCR).

year 2000 except for Russian and Slovak, for which such recent translations are currently not available as e-books.

2.1.3. Preprocessing

In order to prepare the written text for the annotation pipeline, the documents needed to be cleaned of formatting errors, punctuation, typographical errors and other inconsistencies resulting from the conversion process. Additional text stemming from the title page, picture captions as well as biography sections or other supplemental sections was removed. Each sentence of the text is assigned a separate ID to facilitate further processing. We employed native speakers to preprocess and check the texts manually.

2.2. Synthesized speech data

The first six chapters of the story will be converted to spoken language via Text-to-Speech Synthesis (TTS). We chose to use synthetic voices over natural voices for two main reasons: First, due to the time and cost involved in recording professional speakers in a laboratory setting. Second, to have more control over the resulting speech output and to obtain voices that do not differ too much in voice quality, pitch and speaking rate. This allows for better experimental control over the effects of individual voice differences during neuroscientific studies.

2.2.1. Google Text-to-Speech API

In order to obtain synthesized speech that is as natural as possible, we chose to use the state-of-the-art WaveNet (Oord et al., 2016) voices provided by the Google Cloud Text-to-Speech API. We chose this synthesizer since it currently provides the largest selection of natural sounding voices. The client libraries are an efficient method of creating speech output in a wide range of languages in human-like quality. The API also allows the input to be further enhanced using the W3C Speech Synthesis Markup Language (SSML[6]), which enables the user to manually add additional instructions on how the input text is to be synthesized. The Google API supports a subset of SSML tags for generating different prosody or for reading out numerals.

2.2.2. Manual markup of input text

The first six chapters as cleaned written text files are used as raw input for TTS. The text is segmented into smaller parts, that is, single sentences or paragraphs, for easier handling during the processing pipeline.

We recruited native speakers with expertise in TTS to create SSML markup that increase the naturalness of the synthesis where necessary. This markup can be used to change the prosody, for example for making pitch modifications and inserting breaks. An example of the markup is illustrated in Figure 1.

Since the prosody across sentence boundaries can differ when sentences are entered individually or as part of a longer text, they were also asked to decide whether to synthesize the sentences individually or as grouped into paragraphs. The sentence IDs assigned to the raw text are kept track of during this step.

We let the native speakers choose the most natural sounding female WaveNet voices according to their opinion. The only current exception is Spanish, for which currently only one ”standard” female voice is provided.

2.2.3. Naturalness of synthetic speech

The naturalness of the speech recordings is constrained by feasibility: Based on prior experiences, we chose to employ TTS because the recruitment of professional speakers of comparable professionalism, speech training, and speech quality across languages is a hard-to-predict risk to a project of this size and scope. However, we ensured that the synthetic voices chosen for this corpus are of very high, and in part near-natural, quality. The mean opinion scores (MOS) obtained by using a WaveNet vocoder in the TTS system have been reported to greatly surpass those of traditional parametric or concatenative TTS systems (Shen et al., 2018; Oord et al., 2016).

In addition, we take two further measures to handle variability in synthesis quality: First, the native speakers in charge of SSML adjustment will report gross problems with the TTS output and SSML markup, such that corpus users can easily identify sentences of low synthesis quality. Second, we plan to include with each sentence the results of a rating study collected via crowdsourcing (e.g. Amazon Mechanical Turk), allowing users of the LPPC to include parametric covariates of naturalness in their statistical models or define individual naturalness thresholds.

2.2.4. Alignment of speech and text

The text and speech data will be time-aligned, that is, the timestamps that denote the start and end times of each word in the text will be automatically obtained and provided with the corpus. This step is especially necessary for aligning the spoken part of the data to the EEG recordings.

While standard available tools generally yield good performance in resource-rich languages such as English and German, we expect a poorer quality of the alignments in other languages, and that for certain languages there may not even exist suitable tools. Since Google's services do not provide timestamps for the synthesized output, we will use a workaround solution[7] using their multi-lingual Speech-To-Text API[8], which does provide word offset times.

2.3. Lexico-syntactic annotations

The LPPC will contain lexico-syntactic annotations for the written text part of the corpus that we will automatically obtain using natural language processing (NLP) tools. The full texts will be parsed according to the UD framework. This framework comprises a method of combining consistent annotations across languages. Furthermore, previous evidence has suggested a link between syntactic dependency and psycholinguistic processing (Brennan et al., 2019). The parsed output will be provided in a standard format (CoNLL), which includes part-of-speech (POS) tags and lemmatization. We will train the best state-of-the-art parser trained on the respective UD treebank for each lan-

[6] https://www.w3.org/TR/speech-synthesis11/

[7] This workaround had been suggested to us by Google.

[8] https://cloud.google.com/speech-to-text

```
<p>
  <s> He bent over the drawing. </s><break time="300ms"/>
  <s><prosody pitch="+2st" rate="110%"> "Not so small as all that. <break time="500ms"/>
   Look! <break time="300ms"/> He's gone to sleep!" </prosody></s><break time="700ms"/>
  <s> And that's how I made the acquaintance of the little prince. </s>
</p>
```

Figure 1: Example paragraph taken from the English translation of *Le Petit Prince* with SSML markup

guage for parsing. We refer to section 3.3.2. for a discussion on annotation quality estimation.

2.4. EEG data

We aim to collect EEG recordings from 20 participants for each of the languages in the LPPC. During EEG recording, the synthesized speech data (i.e., the first six chapters of the story) will be played via loudspeakers at a volume that is comfortable to the participants. To ensure that participants stay alert and focus on the content of the story, a set of multiple-choice comprehension questions will be asked after each chapter; questions and responses will be included in the corpus. This also enables corpus users to model inter-individual comprehension differences or define their own selection thresholds.

While we plan to include an active task, the paradigm behind the planned EEG recordings is mostly passive. We refer to a body of literature from the speech, language, and music fields (Cheung et al., 2019; Hale et al., 2018; Rabovsky et al., 2018; Frank et al., 2015; Armeni et al., 2019; Brennan and Martin, 2020; Weissbart et al., 2020; Meyer and Gumbert, 2018; Di Liberto et al., 2015) to expect variability of electrophysiological responses of interest to the user (e.g., evoked responses, changes in oscillatory phase and power) to exhibit enough variance for state-of-the-art statistical analysis (e.g., multiple regression, temporal response functions, speech–brain-coupling measures).

EEG data will be continuously recorded from 64 electrodes.The setup will be referenced against the left mastoid and grounded to the sternum. To facilitate subtraction of eye blink and movement artifacts, the horizontal and vertical electrooculograms will be acquired. Scalp electrodes will be placed according to the 10–20 system in an elastic cap. During recording, the word start and end markers of the audio will be stored as events in the EEG file.

Artifact cleaning will be automatic, combining functions from EEGLAB (Delorme and Makeig, 2004) and FieldTrip (Oostenveld et al., 2011) running in MATLAB®. We will use an absolute threshold to remove outlier recording channels. The 50-Hz artifact and resonance frequencies will be projected out via a combination of a perfect-reconstruction filter bank and a spatial filter (de Cheveigné, 2019). Remaining artifacts will be removed using independent-components analysis (ICA). To stabilize ICA, an 1-Hz highpass filter will be applied (Winkler et al., 2015), followed by wavelet ICA (Gabard-Durnam et al., 2018) and ICA (Makeig et al., 1996); artifact components will be automatically classified using MARA (Winkler et al., 2011), ADJUST (Mognon et al., 2011), and ICLABEL (Pion-Tonachini et al., 2019). Artifactual components will be removed from the data highpass-filtered at 0.01 Hz (Winkler et al., 2015). Then, channels removed from the initial thresholding will be interpolated.

3. Ongoing work

We are currently in the stage of acquiring cleaned versions of the text data as well as the SSML markup as input for our speech processing pipeline. The annotation of the text data and the recording of EEG data will occur in parallel once the acquisition of the primary data is completed.

3.1. Availability

We plan to release the corpus in three stages: (1) The release of the primary text data, synthesized speech and (word-level) time-alignments, (2) the lexico-syntactic annotations of the written text, and (3) the preprocessed EEG data recorded during listening and aligned with the speech data. The first version of the corpus release is expected to be available in parallel to this publication. The release of neuroimaging data is postponed for the third release due to pending legal issues regarding data privacy[9]. We plan to make as many EEG recordings available as possible under these constraints. For better re-usability, we also aim to convert the EEG data to openNeuro[10] format.

3.2. Metadata

The corpus release will include bibliographical information on the e-book publications (e.g., name of the translator, year of publication, and publishing house). We will provide the Google WaveNet voice ID as well as the SSML markup used to create the synthesized speech data. We will also provide detailed information on the NLP tools and methods used to create the lexico-syntactic annotations, as well as information on the estimated quality for each language.

The EEG subset of the corpus will include metadata such as the age, gender, native language and bilinguality of each subject. Complete EEG metadata (e.g., filter and ICA settings) will be provided with the respective release.

3.3. Discussion

Due to use of automatic annotation methods and the choice of using synthesized speech for our corpus, several open questions arise, which we discuss in the following. Furthermore, we welcome feedback on possible additional caveats and extensions while the corpus is under construction.

In addition, by means of an outlook, we will discuss some classes of research avenues that could be addressed by employing the LPPC in planned typological contrasts.

[9] Subjects must give written consent according to the European General Data Protection Regulation (GDPR).

[10] https://openneuro.org/

Corpus subset	Size	Annotations	Metadata
Text	27 chapters, ≈ 16k English words	Universal Dependencies	bibliographical data, NLP tools
Speech	chapters 1–6, ≈ 20 minutes of speech	time-alignments	Google voice, SSML
EEG	speech subset	time-alignments	subject metadata

Table 1: Overview of the planned LPPC resource in 26 languages.

3.3.1. Use of synthesized speech

The decision to use TTS to create the speech part of the LPPC was based on our aim to use the dataset for neurolinguistic studies that focus on higher-level syntactic processing. We would like to stress that we do not recommend the corpus for research on lower-level phonetic or auditory processing, since these would require human speech to rule out any confounds created by parts of the auditory stimulus that may be perceived as clearly non-human.

As discussed in section 2.2.3., the Google voices used to create the spoken part of the corpus have been judged to be of significantly higher quality than the best previous TTS systems and the SSML markup is used to further increase the naturalness of the synthesized speech. However, the synthesized speech still differs from human speech, especially when used to read out a literary text. We had chosen this method despite this drawback due to the fact that it enables us to efficiently obtain speech data for all chosen languages.

Depending on the outcome of the ratings obtained from crowdsourcing (see 2.2.3.), it may be necessary to include a recording of a human speaker for at least one language to perform a comparison in further neuroscientific studies. Expanding the selection of languages which include human speech can then be taken into account for possible future versions of the corpus.

3.3.2. Quality of automatic annotations

Since the linguistic annotations will be obtained using purely automatic NLP methods, they are expected to include errors. While the quality of the automatic time-alignments and the syntactic parses will likely be quite high for resource-rich languages such as English, we expect a higher degree of error in low-resource languages. By using tools that can be applied cross-linguistically, however, we aim to generate annotations with a high accuracy. Furthermore, domain differences between the data used to train the tools and the LPPC (children's literature) can be reduced by choosing treebanks from literary texts. The exact choice of tools is subject to current work and will consist of methods that meet this aim.

Possible methods to increase the quality of the linguistic annotations include hand-annotating small amounts of text as a gold-standard reference for automatic evaluation and for domain adaptation of annotation models, or employing native speakers to perform manual corrections in cases where the error rate is deemed too large to be acceptable. Previous efforts to increase the quality of automatic corpus annotation include, for example, a silver standard approach (Rebholz-Schuhmann et al., 2010; Schweitzer et al., 2018; Hale et al., 2019), in which several annotation layers can be combined to estimate confidence scores.

3.3.3. Outlook: an EEG typology

The main motivation for building the LPPC is to address the notion of overcoming the typological restrictedness of prior and current experimental designs in psycho- and neurolinguistics, which is a major obstacle for the generalizability of cognitive and neuroanatomical frameworks of language comprehension (Kandylaki and Bornkessel-Schlesewsky, 2019). While this work-in-progress paper cannot serve the purpose of providing an exhaustive list of cross-linguistic contrastive research questions, we here give a short set for inspiration.

First, cross-linguistic variance in evoked potentials and oscillatory power and phase changes associated with memory storage mechanisms of dependency formation could be tested (Meyer et al., 2013; Kluender and Kutas, 1993). Initial pilot work supports the feasibility of this (Brennan et al., 2019). Moreover, further open questions of models of dependency formation could be tested cross-linguistically, including retrieval cues and their weighting, as well as whether memory retrieval is activation-based or direct (Vasishth et al., 2019; McElree, 2000). Indeed, it has been shown that such fine-grained aspects can be dissociated; for instance, Search Effort as formalized in parsing algorithms was shown to model evoked components classically associated with syntactic processing difficulty (Hale et al., 2018). In addition to enhancing the validity of parsing algorithms proper from their statistical fit to the underlying electrophysiology, seminal work on the alignment between electrophysiological excitability and information content (Weissbart et al., 2020; Meyer and Gumbert, 2018) could be tested for its cross-linguistic generalizability, thus working towards an information-theoretic typology (Hahn et al., 2020; Gibson et al., 2019).

4. Conclusion

In this paper, we have presented the LPPC as a resource that combines linguistic data in the form of text and speech with EEG data for 26 languages. The corpus is currently being built semi-automatically; only the written story was acquired and the text cleaned by hand, and the synthetic speech data, linguistic annotations as well as EEG data is obtained using automatic state-of-the-art tools and methods. The LPPC bridges several gaps between traditional psycho- and neurolinguistic approaches and current data-driven research and enables researchers to investigate and generalize research questions across a wide range of languages. We hope to show that using corpora obtained using automatic methods is a realistic alternative to manual naturalistic stimuli, since this approach enables testing larger amounts of data and across a broader range of languages.

The corpus is work-in-progress. Apart from the planned release described here, we encourage future extensions of the corpus by the (computational) psycho- and neurolinguistics

communities to include additional languages as they become available in Google's TTS voice selection or in other synthesis systems of comparable quality. Furthermore, the speech part of the LPPC is limited to the first 6 chapters. Provided that the quality of the output is acceptable and that it proves to be a useful resource, the speech part can readily be extended.

As future work, we plan to further expand the scope of research questions that can be addressed with the LPPC by incorporating data from additional neuroimaging modalities, such as magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI; see Bhattasali et al. (2019) for an application using human speech). Our vision is for the LPPC to become an open infrastructure to which researchers from various communities can contribute by adding further modalities, such as functional near-infrared spectroscopy or electrocorticography. We also welcome further suggestions and contributions to help expand the utility of the LPPC across disciplines to facilitate innovative psycho- and neurolinguistic research.

5. Acknowledgements

The authors would like to thank Joakim Nivre for helpful discussion. Monique Horstmann carried out preparatory work during early stages. This work was funded by the Max Planck Society through the award of Max Planck Research Group *Language Cycles* to Lars Meyer.

6. Bibliographical References

Armeni, K., Willems, R. M., Van den Bosch, A., and Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, 198:283–295.

Bhattasali, S., Fabre, M., Luh, W.-M., Al Saied, H., Constant, M., Pallier, C., Brennan, J. R., Spreng, R. N., and Hale, J. (2019). Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510.

Bhattasali, S., Brennan, J. R., Luh, W.-M., Franzluebbers, B., and Hale, J. T. (2020). The Alice Datasets: fMRI EEG observations of natural language comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2016). The importance of linguistic typology for the neurobiology of language. *Linguistic Typology*, 20(3):615–621.

Brennan, J. R. and Martin, A. E. (2020). Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B*, 375(1791):20190305.

Brennan, J., Martin, A. E., Dunagan, D., Meyer, L., and Hale, J. (2019). Resolving dependencies during naturalistic listening. In *11th Annual Meeting of the Society for the Neurobiology of Language*.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Cheung, V. K., Harrison, P. M., Meyer, L., Pearce, M. T., Haynes, J.-D., and Koelsch, S. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, 29(23):4084–4092.

de Cheveigné, A. (2019). ZapLine: a simple and effective method to remove power line artifacts. *NeuroImage*.

Delorme, A. and 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.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

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

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336–10341.

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., and Levin, A. R. (2018). The harvard automated processing pipeline for electroencephalography (happe): standardized processing software for developmental and high-artifact data. *Frontiers in neuroscience*, 12:97.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*.

Hahn, M., Jurafsky, D., and Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

Hale, J., Dyer, C., Kuncoro, A., Brennan, J. R., and Acl, A. (2018). Finding Syntax in Human Encephalography with Beam Search. In *ACL*, pages 1–9.

Hale, J., Kuncoro, A., Hall, K., Dyer, C., and Brennan, J. (2019). Text genre and training data size in human-like parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5846–5852, Hong Kong, China, November. Association for Computational Linguistics.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Hamilton, L. S. and Huth, A. G. (2018). The revolution will not be controlled: natural stimuli in speech

neuroscience. *Language, Cognition and Neuroscience*, 0(0):1–10.

Kandylaki, K. D. and Bornkessel-Schlesewsky, I. (2019). From story comprehension to the neurobiology of language. *Language, Cognition and Neuroscience*, 4(4):405–410.

Kluender, R. and Kutas, M. (1993). Bridging the gap: Evidence from erps on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, 5(2):196–214.

Kroczek, L. O. and Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Scientific reports*, 7(1):1–9.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Lopopolo, A., Frank, S. L., Van den Bosch, A., Nijhof, A., and Willems, R. M. (2018). The Narrative Brain Dataset (NBD), an fMRI dataset for the study of natural language processing in the brain.

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D S Touretzky, et al., editors, *Advances in neural information processing systems 8*, pages 145–151. MIT Press, Cambridge.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, 29(2):111–123.

Meyer, L. and Gumbert, M. (2018). Synchronization of electrophysiological responses with speech benefits syntactic information processing. *Journal of cognitive neuroscience*, 30(8):1066–1074.

Meyer, L., Obleser, J., and Friederici, A. D. (2013). Left parietal alpha enhancement during working memory-intensive sentence processing. *Cortex*, 49(3):711–721.

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 48(2):229–240.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011:156869.

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197.

Rabovsky, M., Hansen, S. S., and McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, 2(9):693–705.

Rebholz-Schuhmann, D., Jimeno-Yepes, A. J., van Mulligen, E. M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., and Hahn, U. (2010). The calbc silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Schoffelen, J.-M., Oostenveld, R., Lam, N. H., Uddén, J., Hultén, A., and Hagoort, P. (2019). A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific data*, 6(1):1–13.

Schweitzer, K., Eckart, K., Gärtner, M., Falenska, A., Riester, A., Roesiger, I., Schweitzer, A., Stehwien, S., and Kuhn, J. (2018). German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Vasishth, S., Nicenboim, B., Engelmann, F., and Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in cognitive sciences*.

Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (2020). Cortical tracking of surprisal during continuous speech comprehension. *Journal of cognitive neuroscience*, 32(1):155–166.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Winkler, I., Haufe, S., and Tangermann, M. (2011). Automatic classification of artifactual ica-components for artifact removal in eeg signals. *Behavioral and Brain Functions*, 7(1):30.

Winkler, I., Debener, S., Muller, K. R., and Tangermann, M. (2015). On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2015-Novem, pages 4101–4105. IEEE.