The Narrative Brain Dataset (NBD), an fMRI Dataset for the Study of Natural Language Processing in the Brain

Alessandro Lopopolo[1], **Stefan L. Frank**[1],
Antal van den Bosch[1,2], **Annabel Nijhof**[3], **Roel M. Willems**[1,4,5]
[1]Center for Language Studies, Radboud University
Erasmusplein 1, 6525 HT Nijmegen, the Netherlands
[2]Meertens Institute, Royal Netherlands Academy of Arts and Sciences
Oudezijds Achterburgwal 185, 1012 DK Amsterdam, the Netherlands
[3]SGDP Centre, King's College London
16 De Crespigny Park, Denmark Hill, SE5 8AF London, UK
[4]Donders Institute, Radboud University
Kapittelweg 29, 6525 EN Nijmegen, The Netherlands
[5]Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, the Netherlands
{a.lopopolo, s.frank, a.vandenbosch}@let.ru.nl, annabel.nijhof@kcl.ac.uk, r.willems@donders.ru.nl

Abstract

We present the Narrative Brain Dataset, an fMRI dataset that was collected during spoken presentation of short excerpts of three stories in Dutch. Together with the brain imaging data, the dataset contains the written versions of the stimulation texts. The texts are accompanied with stochastic (perplexity and entropy) and semantic computational linguistic measures. The richness and unconstrained nature of the data allows the study of language processing in the brain in a more naturalistic setting than is common for fMRI studies. We hope that by making NBD available we serve the double purpose of providing useful neural data to researchers interested in natural language processing in the brain and to further stimulate data sharing in the field of neuroscience of language.

Keywords: fMRI, neuro-linguistics, naturalistic stimuli, narrative, perplexity, surprisal, PoS

1. Introduction

The Narrative Brain Dataset (NBD) is an fMRI dataset created by recording the brain activity of 24 native speakers of Dutch during passive listening to three narrative Dutch texts: excerpts from audiobooks. This task and these stimuli are intended to be as naturalistic as possible. The dataset is meant to be used by researchers interested in the study of natural language processing in the human brain using naturalistic, unconstrained linguistic material. This dataset has already been used in a number of neuroscientific studies combining computational linguistic models and brain imaging analysis techniques, as exemplified in Section 6 .
NBD comes with meta-data describing the temporal structure of the stimulus presentation (word onset, offset and duration) and with a series of supplementary annotation of the stimulus texts that might come useful as starting point for further analysis of the data.
We hope that by making NBD available we serve the double purpose of providing useful neural data to researchers interested in naturalistic language comprehension, and to further stimulate data sharing in the field of neuroscience of language.

2. Dataset Structure

The NBD dataset consists of three parts: fMRI data, text & meta-data, and supplementary annotation.
fMRI data (*/fMRI/*) contains 24 folders (*/S01/,..., /S24/*) – one for each subject. Each subject folder is divided in 6 run folders ((*/run1/,/run2/,/run3/,/run4/,/run5/, and /run6/*) containing .nii volume images constituting the magnetic resonance recording during the presentation of the stimuli. Table 1 explains the relation between runs and stimuli – in Section 3 2 we explain the procedure behind the 6 runs structure, whereas in Section 4 we give more details about the stimuli. The data is preprocessed according to the methods described in Section 3 and in a format that is compatible with SPM8 and later versions[1]. The current format can be easily converted into other formats according to the user's needs. We decided not to include the raw fMRI images for reasons of space and efficiency.

Run name	Stimulus	CGN name
run1	Narrative 1	fn1055
run2	Narrative 2	fn1100
run3	Narrative 3	fn1090
run4	Narrative 1 reverse	NA
run5	Narrative 2 reverse	NA
run6	Narrative 3 reverse	NA

Table 1: Correspondence between fMRI data runs, stimulus narratives (or reverse recordings of narratives) and original "Corpus Gesproken Nederlands" (CGN) file names.

Text & meta-data consists of three .txt tab separated files (*Narrative_1_wordtiming.csv, Narrative_2_wordtiming.csv, Narrative_3_wordtiming.csv*) containing the text of the three narratives presented to the subjects. Each row in the

[1] `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`

file corresponds to one word (or word + punctuation) of one of the three narrative text stimuli accompanied with the temporal parameters of each word of the textual stimuli with regard to the experimental paradigm described above. These consist of word onset, offset and duration in seconds. Table 2 provides an example of temporal meta-data for a sentence from the stimulus texts. These parameters are especially important given that a) fMRI volume acquisition and word onset are not synchronized; b) fMRI volume is acquired every 880 ms (see Section 3 3), whereas word duration is variable and can be shorter than that (as Table 2 exemplifies).

Word	Onset	Offset	Duration
plotseling	0.103	0.68	0.577
is	0.68	0.843	0.163
ze	0.843	0.976	0.133
er.	0.976	1.149	0.173

Table 2: Example of the timing information of the stimulus text.

NBD is also provided with a battery of **supplementary annotations** consisting of part of speech (PoS) tags and computational measures assigned to each word of the text stimuli (files: *Narrative_1_annotation.txt, Narrative_2_annotation.txt, Narrative_3_annotation.txt*). Each column of the annotation file contains: the word and its PoS, word frequency, PoS frequency, average phonetic frequency, word perplexity, PoS perplexity, average phonetic perplexity, word entropy, word semantic association. The procedures used to obtain these additional annotations are described in Section 5 .

3. Magnetic Resonance Data

3.1. Participants

Twenty-four healthy, native speakers of Dutch (8 males; mean age 22.9 years, range 18-31) without psychiatric or neurological problems, with normal or corrected-to-normal vision, and without hearing problems took part in the experiment. All participants except one were right-handed. Ethical approval was obtained from the CMO Committee on Research Involving Human Subjects, Arnhem-Nijmegen, The Netherlands (protocol number 2001/095), in line with the Declaration of Helsinki.

3.2. Procedure

The experimental paradigm consisted of passively listening to the three narratives (see Section 4) and their reversed versions (for a total of six sessions) inside the MRI scanner. That amounted to six experimental runs, all collected in one single fMRI session on the same day. Each story and its reversed speech counterpart were presented following each other. Reversed speech versions of the stories were created with Audacity 2.03[2]. Half the participants started with a non-reversed stimulus, and half with a reversed speech stimulus. Participants were instructed to listen to the materials attentively, which in practice is only possible for three narratives, and not for the reversed speech counterparts. There was a short break after each fragment.
Stimuli were presented with Presentation 16.2[3]. Auditory stimuli were presented through MR-compatible earphones. After the scanning session, participants were tested for their memory and comprehension of the stories.

3.3. Scanner Parameter

Images of blood-oxygenation level-dependent (BOLD) changes were acquired on a 3-T Siemens Magnetom Trio scanner (Erlangen, Germany) with a 32-channel head coil. Pillows and tape were used to minimize participants' head movement, and the earphones that were used for presenting the stories reduced scanner noise. Functional images were acquired using a fast T2-weighted 3D echo planar imaging sequence (Poser et al., 2010), with high temporal resolution (time to repetition: 880 ms, time to echo: 28 ms, flip angle: 14, voxel size: 3.5 $\times$ 3.5 $\times$ 3.5 mm, 36 slices). High resolution (1 $\times$ 1 $\times$ 1.25 mm) structural (anatomical) images were acquired using a T1 sequence.

3.4. Preprocessing

Preprocessing was performed using SPM8[4] and Matlab 2010b[5]. The first four volumes were removed to control for T1 equilibration effects. Rigid body registration was used to realign images. Images were realigned to the first image within each run. The mean of the motion-corrected images was then brought into the same space as the individual participant's anatomical scan. The anatomical and functional scans were spatially normalized to the standard MNI template, and functional images were re-sampled to 2 $\times$ 2 $\times$ 2 mm voxel sizes. Finally, an isotropic 8-mm full-width at half-maximum Gaussian kernel was used to spatially smooth the motion-corrected and normalized data.

4. Linguistic Data

Narrative text used as stimuli presented to the human subjects consisted of three excerpts from three distinct literary novels extracted from the Spoken Dutch Corpus, "Corpus Gesproken Nederlands" (CGN) (Oostdijk, 2000).[6]

The excerpts were spoken at a normal rate, in a quiet room, by female speakers (one speaker per story). Stimulus durations were: Narrative 1 (CGN file fn1005) 3:49 min, Narrative 2 (CGN file fn1100) 7:50 min, and Narrative 3 (CGN file fn1090) 7:48 min.

Table 3 contains summary information about the three narratives, including number of words, mean and range of word duration in milliseconds.

5. Annotation

Besides the temporal information, the linguistic data is accompanied by two additional types of annotation: linguis-

[2] http://www.audacityteam.org

[3] https://www.neurobs.com

[4] http://www.fil.ion.ucl.ac.uk/spm

[5] http://www.mathworks.nl

[6] Narrative 1: from Peper, R., *Dooi*, L.J. Veen, 1999; Narrative 2: from Van der Meer, V., *Eilandgasten*, Contact, 1999; Narrative 3: from Jakobsen, A., *De Stalker*, De Boekerij, 1999

		Word Duration (msec)	
	# Words	Mean (s.d.)	Range
Narrative 1	622	273 (181)	4-1174
Narrative 2	1291	252 (160)	31-949
Narrative 3	1131	274 (183)	40-1221

Table 3: Summary information of the three narrative texts used as stimuli.

tic – consisting of the PoS tags of the words in the text – and computational measures – consisting of stochastic and computational semantics measures computed on the word, PoS and phonological level of the texts.

5.1. Linguistic Annotation

The words in the stimuli are annotated with their syntactic categories, or parts of speech (PoS). The tagset employed here was the one employed by CGN (Oostdijk, 2000) and comprises 320 tag types[7]. Besides 13 base tags, this method explicitly assigns morpho-syntactic sub-category features to the base tags containing information such as gender, number, form and so on. This tagset closely follows the practices of the Dutch Grammar "Algemene Nederlandse Spraakkunst" (ANS) (Haeseryn et al., 1997).

5.2. Computational Annotation

All words in the linguistic data are assigned seven stochastic measures: word frequency and perplexity, PoS frequency and perplexity, average phonological frequency and perplexity, and word entropy. A measure of the semantic association between each word and its preceding textual context is also provided.

5.2.1. Stochastic Measures

Perplexity – the degree to which the actually perceived item x_t in a series deviates from expectation – is computed as an exponential transformation of the surprisal of encountering x_t given its previous context $x_1 \ldots, x_{t-1}$:

$$\text{ppl}(x_t) = 2^{\text{surprisal}(x_t)} = 2^{-\log P(x_t|x_1,\ldots,x_{t-1})}$$

The conditional probabilities required for obtaining perplexity are estimated by a second-order Markov model, also known as a trigram model. That is, $P(x_t|x_1, \ldots, x_{t-1})$ is simplified to $P(x_t|x_{t-2}, x_{t-1})$. Using SRILM (Stolcke, 2002), the model was trained on a random selection of 10 million sentences (comprising 197 million word tokens; 2.1 million types) from the Dutch Corpus of Web (NLCOW2012) (Schäfer and Bildhauer, 2012).

The PoS perplexity is computed analogously. Instead of using the surface forms of the training and stimulus set, the trigram model was trained on the PoS-tagged version of the same 10 million sentences subset of NLCOW2012. The tagging was performed using the Frog toolbox for natural language processing of Dutch text (Daelemans and van den Bosch, 2005; van den Bosch et al., 2007)[8].

Phonological perplexity was estimated from conditional probabilities $P(p_t|p_{t-1}, p_{t-2})$, where the ps refer to the phonological transcription of the words in the running texts into a sequence of phonemes using a memory-based grapheme phoneme converter (Busser et al., 1999) trained on CELEX 2 (Baayen et al., 1995). The probabilities are computed using WOPR[9] (van den Bosch and Berck, 2009) trained on CELEX 2 (Baayen et al., 1995). Once phoneme-wise perplexity is computed, the phonemic perplexity of each word of the stimulus is computed as the average value across the phonemes of that word.

Next-word **entropy** was also derived from the conditional probabilities of words given their preceding context. It is a function of the distribution of probabilities of all possible upcoming words. It is computed as:

$$E(x_{t+1}) = - \sum_{x_{t+1} \in V} P(x_{t+1}|x_t, x_{t-1}) \log P(x_{t+1}|x_t, x_{t-1}),$$

where V denotes the vocabulary (i.e., the set of word types in the training data). Entropy values were computed by WOPR (van den Bosch and Berck, 2009).

5.2.2. Semantic Similarity Measures

The semantic similarity between each content word w_t and its preceding context C is computed as the cosine between the distributional semantic vector representations of w_t and of C. Semantic vector representations of words were generated by the word2vec skipgram model (Mikolov et al., 2013). The representation of C is defined as the sum of the vector representations of the four content words preceding w_t (or fewer, if w_t is among the first four words of the text). If w_t is the first content word of the text then C is empty so semantic distance is undefined.

6. Published Analyses of the Current Dataset

The present fMRI data has already been analysed in several studies, demonstrating that naturalistic linguistic tasks and fMRI can yield interesting and meaningful results. Willems et al. (2016) have shown that entropy and surprisal predict brain activity in different brain areas. Frank and Willems (2017) demonstrated that predictive measures (surprisal) and semantic association measures can be distinguished with regard to brain area sensitivity. Similarly, PoS, lexical and phonological stochastic measures divide the cortical language network in non-overlapping sub-networks (Lopopolo et al., 2017). Part of the data was used by Nijhof and Willems (2015) to investigate how individuals differently employ neural networks important for understanding others' beliefs and intentions, and for sensori-motor simulation while processing narrative language.

7. Data Availability

The NBD is available at `https://osf.io/utpdy/`.

[7] more details at `http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/annot/pos_tagging/info.htm`

[8] `http://languagemachines.github.io/frog/`

[9] `https://ilk.uvt.nl/wopr/`

8. Acknowledgements

The work presented here was funded by NWO Gravitation Grant 024.001.006 to the Language in Interaction Consortium and by a grant from the Netherlands Organisation for Scientific Research (NWO-Vidi 276-89-007).

9. Bibliographical References

Busser, B., Daelemans, W., and van den Bosch, A. (1999). Machine learning of word pronunciation: the case against abstraction. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*.

Frank, S. L. and Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Lopopolo, A., Frank, S. L., van den Bosch, A., and Willems, R. M. (2017). Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE*, 12(5):1–18, 05.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May.

Nijhof, A. D. and Willems, R. M. (2015). Simulating fiction: Individual differences in literature comprehension revealed with fmri. *PLOS ONE*, 10:1–17, 02.

Poser, B., Koopmans, P., Witzel, T., Wald, L., and Barth, M. (2010). Three dimensional echo-planar imaging at 7 tesla. *NeuroImage*, 51(1):261 – 266.

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

10. Language Resource References

Baayen, H. R., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Daelemans, W. and van den Bosch, A. (2005). Memory-based learning in natural language processing. *Memory-Based Language Processing*, pages 3–14.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., and van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst.* ONijhoff and Deurne: Wolters Plantyn.

Oostdijk, N. (2000). The spoken dutch corpus. overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L00-1083.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, et al., editors, *LREC*, pages 486–493. European Language Resources Association (ELRA).

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. pages 901–904.

van den Bosch, A. and Berck, P. (2009). Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical LinguisticsE*, 91(17).

van den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In Frank V. Eynde, et al., editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. Leuven, Belgium.