Check for updates

DATA NOTE

A studyforrest extension, an annotation of spoken language in the German dubbed movie “Forrest Gump” and its audio-description [version 1; peer review: 1 approved, 2 approved with reservations]

Christian Olaf Häusler[1,2], Michael Hanke[1,2]

[1]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Nordrhein-Westfalen, 52425, Germany
[2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University, Düsseldorf, Nordrhein-Westfalen, 40225, Germany

v1 **First published:** 28 Jan 2021, **10**:54
https://doi.org/10.12688/f1000research.27621.1
Latest published: 28 Jan 2021, **10**:54
https://doi.org/10.12688/f1000research.27621.1

Abstract

Here we present an annotation of speech in the audio-visual movie “Forrest Gump” and its audio-description for a visually impaired audience, as an addition to a large public functional brain imaging dataset (studyforrest.org). The annotation provides information about the exact timing of each of the more than 2500 spoken sentences, 16,000 words (including 202 non-speech vocalizations), 66,000 phonemes, and their corresponding speaker. Additionally, for every word, we provide lemmatization, a simple part-of-speech-tagging (15 grammatical categories), a detailed part-of-speech tagging (43 grammatical categories), syntactic dependencies, and a semantic analysis based on word embedding which represents each word in a 300-dimensional semantic space. To validate the dataset’s quality, we build a model of hemodynamic brain activity based on information drawn from the annotation. Results suggest that the annotation’s content and quality enable independent researchers to create models of brain activity correlating with a variety of linguistic aspects under conditions of near-real-life complexity.

Keywords

annotation, language, speech, narrative, naturalistic stimulus, fMRI, studyforrest

Open Peer Review

Reviewer Status ✓ ? ?

	Invited Reviewers		
	1	**2**	**3**
version 1 28 Jan 2021	✓ report	? report	? report

1. **Giada Lettieri**, IMT School for Advanced Studies Lucca, Lucca, Italy
2. **Martin Wegrzyn**, Bielefeld University, Bielefeld, Germany
3. **Roberta Rocca**, The University of Texas at Austin, Austin, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Christian Olaf Häusler (der.haeusler@gmx.net)

Author roles: Häusler CO: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Hanke M**: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: COH was supported by a graduate stipend from the German federal state of Saxony-Anhalt and MH was, in-part, supported by funds from the German federal state of Saxony-Anhalt and the European Regional Development Fund (ERDF), Project: Center for Behavioral Brain Sciences (CBBS). Development of the data management technology used to publish these data on the Open Science Framework was co-funded by the US National Science Foundation (NSF 1912266), and the German Federal Ministry of Education and Research (BMBF 01GQ1905).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Häusler CO and Hanke M. **A studyforrest extension, an annotation of spoken language in the German dubbed movie "Forrest Gump" and its audio-description [version 1; peer review: 1 approved, 2 approved with reservations]** F1000Research 2021, **10**:54 https://doi.org/10.12688/f1000research.27621.1

First published: 28 Jan 2021, **10**:54 https://doi.org/10.12688/f1000research.27621.1

Introduction

Cognitive and psychiatric neuroimaging are moving towards studying brain functions under conditions of lifelike complexity[1,2]. Motion pictures[3] and continuous narratives[4,5] are increasingly utilized as so called "naturalistic stimuli". Naturalistic stimuli are usually designed for commercial purposes and to entertain their audiences. Thus, the temporal structure of their feature space is usually not explicitly known, leading to an "annotation bottleneck"[6] when used for neuroscientific research.

Data-driven methods like inter-subject correlation (ISC)[7] or independent component analysis (ICA)[8] are often used to analyze such fMRI data in order to circumvent this bottleneck. However, use of data-driven methods alone falls short of associating results with particular stimulus events[9]. Model-driven methods, like the general linear model (GLM), which are based on stimulus annotations can be useful to test hypotheses on specific brain functions under more ecologically valid conditions, to statistically control confounding stimulus features, and to explain not just "how" the brain is responding to a stimulus but also "why"[10]. Studies using GLMs based on annotations of a stimulus' temporal structure have elucidated, for example, how the brain responds to visual features of a movie[11] or speech-related features of a narrative[12]. Furthermore, stimulus annotations can inform data-driven methods about a stimulus' temporal dynamics, or model-driven and data-driven methods can be combined to improve the interpretability of results[13].

Here we provide an annotation with exact onset and offset of each sentence, word and phoneme (see Table 1 for an overview) spoken in the audio-visual movie "Forrest Gump"[14] and its audio-description (i.e. the movie's soundtrack with an additional narrator)[15]. fMRI data of participants watching the audio-visual movie[16] and listening to the audio-description[17] are the core data of the publicly available *studyforrest* dataset (studyforrest.org). The current publication enables researchers to model hemodynamic brain responses that correlate with a variety of aspects of spoken language ranging from a speaker's identity, to phonetics, grammar, syntax, and semantics. This publication extends already available annotations of portrayed emotions[18], perceived emotions[19], as well as cuts and locations depicted in the movie[20]. All annotations can be used in any study focusing on aspects of real-life cognition by serving as additional confound measures describing the temporal structure and feature space of the stimuli.

Materials and methods

Stimulus

We annotated speech in the slightly shortened "research cut"[17] of the movie "Forrest Gump" and its temporally aligned audio-description[16] that was broadcast as an additional audio track for visually impaired listeners on Swiss public television[15]. The plot of the original movie is already carried by an off-screen voice of the main character Forrest Gump. In the audio-description, an additional male narrator describes essential aspects of the visual scenery when there is no off-screen voice, dialog, or other relevant auditory content.

Annotation procedure

Preliminary, manual orthographic transcripts of dialogues, non-speech vocalizations (e.g. laughter or groaning) and the script for the audio-description's narrator were merged and converted to Praat's[21] TextGrid format. This merged transcript contained rough onset and offset timings for small groups of sentences, and was further edited in Praat for manual validation against the actual content of the audio material. The following steps were performed by a single person, already familiar with the stimulus, in several passes to iteratively improve the quality of the data: approximate temporal onsets and offsets were corrected; intervals containing several sentences were split into intervals containing only one sentence; when two or more persons were speaking simultaneously the less dominant voice was dropped; low volume

Table 1. Overview of the annotation's content for the audio-description of "Forrest Gump" (i.e. the audio-only variant of the movie) that comprises the additional narrator. Counts are given for the whole stimulus (`all`) and its individual segments used during fMRI scanning. The category `sentences` comprises complete grammatical sentences which are additionally marked in the annotation with a full stop at the end ("my feet hurt."). It also comprises questions ("do you want a chocolate?"), exclamations ("run away!"), or non-speech vocalizations in quick succession ("ha, ha, ha"), or in isolation (e.g. "Forrest?", "Forrest!", "ha") at time points when speakers switch rapidly. The category `words` comprises each word or non-speech vocalization (N=202) in isolation.

Category	All	1	2	3	4	5	6	7	8
Sentences	2528	292	366	320	352	344	289	365	200
Words	16187	2089	2162	2115	2035	2217	2033	2322	1214
Phonemes	66611	8802	8727	8770	8557	9197	8353	9351	4854

non-speech vocalizations or low volume background speech (especially during music or continuous environmental noise) which were subjectively assessed to be incomprehensible for the audience were also dropped.

We then used the Montreal Forced Aligner v1.0.1[22] to algorithmically identify the exact onset and offset of each word and phoneme. To enable the aligner to look up the phonemes embedded within each word, we chose the accompanying German pronunciation dictionary provided by Prosodylab[23] that uses the Prosodylab PhoneSet to describe the pronunciation of phonemes. To improve the detection rate of the automatic alignment, the dictionary was manually updated with German words that occur in the stimuli but were originally missing in the dictionary. The pronunciation of English words and phonemes occurring in the otherwise German audio track was taken from the accompanying English pronunciation dictionary (following the ARPAbet PhoneSet). The audio track of the audio-description was converted from FLAC to WAV via FFmpeg v4.1.4[24] to meet the aligner's input requirements. This WAV file, the merged transcription, and the updated dictionary were submitted to the aligner that first trained an acoustic model on the data and then performed the alignment.

The resulting timings of words and phonemes were corrected manually and iteratively in several passes using Praat v6.0.22[21]: in a first step, onsets and offsets on which the automatic alignment performed moderately were corrected. Some low volume sentences that are spoken in continuously noisy settings (e.g. during battle or hurricane) were removed due to poor overall alignment performance. In a second step, the complete sentences of the orthographic transcription were copied into the annotation created by the aligner. In a third step, a speaker's identity was added for each sentence (see Table 2 for the most often occurring speakers). During every step previous results were repeatedly checked for errors and further improvements.

We employed the Python package spaCy v2.2.1[25] and its accompanying German language model (`de_core_news_md`) that was trained on the TIGER Treebank corpus[26] to automatically analyze linguistic features of each word in their corresponding sentence. Non-speech vocalizations were dropped from the sentences before analysis to improve results. We then performed analyses regarding part-of-speech (i.e. grammatical tagging or word-category disambiguation), syntactic dependencies, lemmatization, word embedding (i.e. a multi-dimensional meaning representation of a word), and if the word is one of the most common words of the German language (i.e. if the word is part of a stop list).

Data legend

The annotation is available in two different versions, both providing the same information: a) as a text-based Praat TextGrid file, and b) as a text-based, tab-separated value (TSV) formatted table. The following descriptions refer to the ten columns of the TSV file, namely `onset`, `duration`, `person`, `text`, `pos`, `tag`, `dep`, `lemma`, `stop`, `vector`.

Start (`start`)

The onset of the sentence, word or phoneme. Time stamps are provided in the format seconds.milliseconds from stimulus onset.

Table 2. Sentences spoken by the ten most often occurring speakers sorted alphabetically. The narrator only occurs in the audio-description. Overall 97 persons were identified. Names are mostly identical to the names used in[18].

Name	All	1	2	3	4	5	6	7	8
Bubba	74	0	16	40	18	0	0	0	0
Forrest	354	22	37	22	48	50	61	49	65
Forrest (child)	19	17	2	0	0	0	0	0	0
Forrest (v.o.)	369	61	48	53	51	37	40	63	16
Hancock	16	16	0	0	0	0	0	0	0
Jenny	177	0	46	30	3	25	0	57	16
Jenny (child)	23	7	16	0	0	0	0	0	0
Lt. Dan	183	0	0	49	33	65	28	0	8
Mrs. Gump	53	38	2	0	0	0	13	0	0
Narrator	903	111	134	78	139	93	115	147	86

Duration (duration)

The duration of the sentence, word or phoneme provided in the format seconds.milliseconds.

Speaker identity (person)

Name of the person that speaks the sentence, word or phoneme. See Table 2 for the ten most often occurring speakers.

Text (text)

The text of a spoken sentence or word, or the pronunciation of a phoneme. Phonemes of German words follow the Prosodylab PhoneSet, English words follow the ARPAbet PhoneSet.

Simple part-of-speech tag (pos)

A simple part-of-speech tagging (grammatical tagging; word-category disambiguation) of words. The tag labels of this simple part-of-speech tagging follow the Universal Dependencies v2 POS tag set (universaldependencies.org). See Table 3 for a description of the labels and the respective counts of all 15 labels. Nouns that spaCy mistook for proper nouns or vice versa were corrected via script. Additionally in cells of this column, sentences are tagged as SENTENCE, and phonemes are tagged as PHONEME to facilitate filtering in potential further processing steps.

Detailed part-of-speech tag (tag)

A detailed part-of-speech tagging of words following the TIGER Treebank annotation scheme[26] which is based on the Stuttgart-Tübingen-Tagset[27]. See Table 4 for a description of the labels and the respective counts of the 15 most often occurring labels (overall 43 labels). Nouns that spaCy mistook for proper nouns or vice versa were corrected via script.

Syntactic dependency (dep)

Information about a word's syntactic dependencies with other words within the same sentence. Information follows the TIGER Treebank annotation scheme[26] and is given in the format: "arc label;word's head;word's child1, word's

Table 3. Simple part-of-speech tagging (pos) performed by the Python package spaCy[25]. All 15 labels sorted alphabetically. Descriptions were taken from spaCy.explain(). Non-speech vocalizations (NONSPEECH) were manually identified. Counts for the whole stimulus (all) and for each of the eight stimulus segments refer to the audio-description.

Label	Description	All	1	2	3	4	5	6	7	8
ADJ	adjective	916	138	126	106	96	130	118	128	74
ADP	adposition	1429	181	176	176	194	188	183	213	118
ADV	adverb	1332	166	169	220	162	178	169	193	75
AUX	auxiliary	807	102	120	92	96	125	110	112	50
CONJ	conjunction	525	74	63	71	49	61	80	86	41
DET	determiner	1754	257	243	198	219	220	222	254	141
NONSPEECH	non-speech vocalization	202	23	21	9	23	55	44	13	14
NOUN	noun	2620	361	341	332	343	331	356	351	205
NUM	numeral	66	8	11	11	7	4	9	14	2
PART	particle	572	60	100	90	62	83	53	86	38
PRON	pronoun	2348	275	321	328	260	348	262	362	192
PROPN	proper noun	1012	131	135	119	168	162	116	117	64
SCONJ	subordinating conjunction	172	19	18	20	15	31	27	26	16
VERB	verb	2317	285	308	320	319	289	274	349	173
X	other	108	8	10	21	21	11	9	17	11

Table 4. Detailed part-of-speech tagging (`tag`) performed by the Python package spaCy[25]. The 15 most often occurring labels (overall 43 labels) sorted alphabetically. Descriptions were taken from spaCy.explain(). Counts for the whole stimulus (`all`) and for each of the eight stimulus segments refer to the audio-description.

Label	Description	All	1	2	3	4	5	6	7	8
ADJA	adjective, attributive	478	73	58	58	51	77	58	70	33
ADJD	adjective, adverbial or predicative	438	65	68	48	45	53	60	58	41
ADV	adverb	1181	146	145	201	143	157	149	174	66
APPR	preposition; circumposition left	1192	156	146	156	152	157	150	178	97
ART	definite or indefinite article	1340	199	183	140	178	159	176	191	114
KON	coordinate conjunction	475	58	58	66	45	58	76	78	36
NE	proper noun	1012	131	135	119	168	162	116	117	64
NN	noun, singular or mass	2620	361	341	332	343	331	356	351	205
PPER	non-reflexive personal pronoun	1638	183	210	221	168	246	176	287	147
PPOSAT	attributive possessive pronoun	274	34	47	36	23	39	32	40	23
PTKVZ	separable verbal particle	353	34	63	49	46	41	33	60	27
VAFIN	finite verb, auxiliary	767	96	108	89	92	116	106	110	50
VVFIN	finite verb, full	1512	181	213	201	202	172	181	228	134
VVINF	infinitive, full	271	37	25	51	32	42	27	40	17
VVPP	perfect participle, full	329	37	40	35	58	44	51	50	14

child2, ...", where the "arc label" (see Table 5) describes the type of syntactic relation that connects a "child" (the current word) to its "head".

Lemmatization (`lemma`)

The base form (root) of a word.

Common Word (`stop`)

This column's cell provides information if the word is part of a stop list, hence one of the most common words in the German language or not (`True` vs. `False`).

Word embedding (`vector`)

A 300-dimensional word vector providing a multi-dimensional meaning representation of a word. Out-of-vocabulary words with a vector consisting of 300 dimensions of zeroes were set to # to save space.

Dataset content

The annotation comes in two different versions. First, as a text-based TextGrid file (`annotation/fg_rscut_ad_ger_speech_tagged.TextGrid`) to be conveniently edited using the software Praat[21]. Second, as a text-based, tab-separated-value (TSV) formatted table (`annotation/fg_rscut_ad_ger_speech_tagged.tsv`) in accordance with the brain imaging data structure (BIDS)[28]. The dataset and validation data are available from Open Science Framework, DataLad and Zenodo (see Underlying data)[29,30,31]. The source code for all descriptive statistics included in this paper is available in `code/descriptive-statistics.py` (Python script).

Dataset validation

In order to assess the annotation's quality, we investigated if contrasting speech-related events to events without speech lead to increased activation in areas known to be involved in language processing[32]. Moreover, we tested if two similar

Table 5. Syntactic dependencies (`dep`) performed by the Python package spaCy[25]. The 15 most often occurring labels (overall 37 labels) sorted alphabetically. Descriptions were taken from spaCy.explain(). Counts for the whole stimulus (`all`) and for each of the eight stimulus segments refer to the audio-description.

Label	Description	All	1	2	3	4	5	6	7	8
cd	coordinating conjunction	335	48	44	48	34	41	53	42	25
cj	conjunct	524	65	74	88	53	65	80	65	34
cp	complementizer	160	17	17	20	16	29	25	21	15
da	dative	170	15	30	27	19	23	18	27	11
ju	junctor	130	10	13	16	12	16	22	31	10
mnr	postnominal modifier	245	30	29	33	44	31	28	27	23
mo	modifier	2634	349	345	355	327	356	334	384	184
nk	noun kernel element	3763	516	482	448	475	507	485	551	299
oa	accusative object	1036	117	139	149	148	146	126	134	77
oc	clausal object	732	98	86	97	94	105	97	115	40
pd	predicate	301	39	50	40	25	45	41	38	23
pnc	proper noun component	154	36	19	15	14	28	22	15	5
ROOT	root of sentence	2417	285	349	322	336	317	267	358	183
sb	subject	2231	280	306	271	276	301	281	340	176
svp	separable verb prefix	355	36	65	45	49	43	33	56	28

linguistic concepts (proper nouns and nouns) providing high semantic information contrasted with a concept providing low semantic information (coordinate conjunctions) lead to increased activation in congruent brain areas.

We used a dataset providing blood oxygenation level-dependent (BOLD) functional magnetic resonance imaging (fMRI) data of 20 subjects (age 21–38 years, mean age 26.6 years, 12 male) listening to the 2 h audio-description (7 Tesla, 2 s repetition time, 3599 volumes, 36 axial slices, thickness 1.4 mm, 1.4 × 1.4 mm in-plane resolution, 224 mm field-of-view)[17]. Data were already corrected for motion at the scanner computer. Further, individual BOLD time-series were already aligned by non-linear warping to a study-specific T2*-weighted echo planar imaging (EPI) group template (cf.[17] for exact details).

All further steps for the current analysis were carried out using FEAT v6.00 (FMRI Expert Analysis Tool)[33] as part of FSL v5.0.9 (FMRIB's Software Library)[34]. Data of one participant were dropped to due to invalid distortion correction during scanning. Data were temporally high-pass filtered (cut-off 150 s), spatially smoothed (Gaussian kernel; 4.0 mm FWHM), and the brain was extracted from surrounding tissue. A grand-mean intensity normalization of the entire 4D dataset was performed by a single multiplicative factor.

We implemented a standard three-level, voxel-wise general linear model (GLM) to average parameter estimates across the eight stimulus segments, and later across 19 subjects. At the first level analyzing each segment for each subject individually, we created 26 regressors (see Table 6) based on events drawn from the annotation. The 20 most often occurring detailed part-of-speech labels (`nn` with N=2620 to `prf` with N=157) were modeled as boxcar function from onset to offset of each word. The remaining other part-of-speech labels were pooled to a single new label (`tag_other`; N=1123) and modeled as a boxcar function from a word's onset to offset. The 80 most often occurring phonemes (`n` with N=6053 to `IY1` with N=32) were pooled to `phonemes` (N=65251) and modeled as boxcar function from a phoneme's onset to offset. The end of each complete grammatical sentence was modeled as an impulse event (N=1651) to capture variance correlating with sentence comprehension. "No-speech" events (`no-sp`; N=264) serving as a control condition were created such that a sufficient number of events and a minimum separation of speech and non-speech events were achieved. Events were randomly positioned in intervals without audible speech that lasted at least 3.6 s. Each event of the no-speech condition had to have a minimum distance of 1.8 s to any onset or offset of a word, and to any onset of another no-speech event. A length of 70 ms was chosen for no-speech events matching the average length of phonemes. Lastly, we used continuous bins of information about low-level auditory features (left-right difference in volume and root mean square energy) that was averaged across the length of every movie frame (40 ms) to capture variance correlating with assumed low-level perceptual processes. Time series of events were convolved with FSL's "Double-Gamma HRF" as a

Table 6. Overview of events that were used to create the 26 regressors of the GLM analysis. The respective counts are given for the whole stimulus and the eight segments that were used during fMRI scanning. The 20 most often occurring labels from the detailed part-of speech tagging (`tag`) were used as such. Words belonging to all other labels were pooled to `tag_other`. The label `sentence` contains the end of complete grammatical sentences. The label `phones` contains events of the 80 most often occurring phonemes (phoneme `n` with N=6053 to phoneme `IY1` with N=32). The label `no-sp` represents moments when no speech was audible. `fg_ad_lrdiff` (left-right volume difference) and `fg_ad_rms` (root mean square energy) were compute for and averaged across every movie frame (40 ms) via Python script. Events were convolved with FSL's "Double-Gamma HRF" to create the regressors. The correlation of these regressors over the time course of the whole stimulus can be seen in Figure 1.

Label	Description	All	1	2	3	4	5	6	7	8
adja	adjective, attributive	478	73	58	58	51	77	58	70	33
adjd	adjective, adverbial or predicative	438	65	68	48	45	53	60	58	41
adv	adverb	1181	146	145	201	143	157	149	174	66
appr	preposition; circumposition left	1192	156	146	156	152	157	150	178	97
apprart	preposition with article	233	24	28	20	42	31	32	35	21
art	definite or indefinite article	1340	199	183	140	178	159	176	191	114
kon	coordinate conjunction	475	58	58	66	45	58	76	78	36
ne	proper noun	1012	131	135	119	168	162	116	117	64
nn	noun, singular or mass	2620	361	341	332	343	331	356	351	205
pds	substituting demonstrative pronoun	192	16	32	31	25	33	27	17	11
pis	substituting indefinite pronoun	217	36	30	35	28	30	21	23	14
pper	non-reflexive personal pronoun	1638	183	210	221	168	246	176	287	147
pposat	attributive possessive pronoun	274	34	47	36	23	39	32	40	23
prf	reflexive personal pronoun	157	18	25	15	23	17	26	19	14
ptkvz	separable verbal particle	353	34	63	49	46	41	33	60	27
vafin	finite verb, auxiliary	767	96	108	89	92	116	106	110	50
vmfin	finite verb, modal	183	28	21	29	24	28	15	30	8
vvfin	finite verb, full	1512	181	213	201	202	172	181	228	134
vvinf	infinitive, full	271	37	25	51	32	42	27	40	17
vvpp	perfect participle, full	329	37	40	35	58	44	51	50	14
tag_other	all other TAG categories	1123	153	165	174	124	169	121	153	64
sentence	complete grammatical sentences	1651	205	231	200	215	212	198	249	141
phones	80 most often occurring phonemes	65251	8589	8534	8597	8387	8976	8184	9232	4752
no-sp	no-speech	264	16	20	23	50	25	27	56	47
fg_ad_lrdiff	left-right volume difference	180133	22574	22075	21925	24425	23125	21975	27175	16859
fg_ad_rms	root mean square	180133	22574	22075	21925	24425	23125	21975	27175	16859

model of the hemodynamic response function to create the actual regressors. The Pearson correlation coefficients of the 26 regressors across the time course of all stimulus segments can be seen in Figure 1. Temporal derivatives were also included in the design matrix to compensate for regional differences between modeled and actual HRF. Finally, six motion parameters were used as additional nuisance regressors and the design was subjected to the same temporal filtering as the BOLD time series. The following three *t*-contrasts were defined: 1) words (all 21 `tag`-related regressors) > no-speech (`no-sp`), 2) proper nouns (`ne`) > coordinate conjunctions (`kon`), and 3) nouns (`nn`) > coordinate conjunctions (`kon`).

The second-level analysis that averaged contrast estimates across the eight stimulus segments per subject was carried out using a fixed effects model by forcing the random effects variance to zero in FLAME (FMRIB's Local Analysis of Mixed Effects)[35,36]. The third level analysis which averaged contrast estimates across subjects was carried out using a mixed-effects model (FLAME stage 1) with automatic outlier deweighting[36,37]. Z (Gaussianised T/F) statistic images were thresholded using clusters determined by Z>3.4 and a corrected cluster significance threshold of p<.05[37]. Brain regions

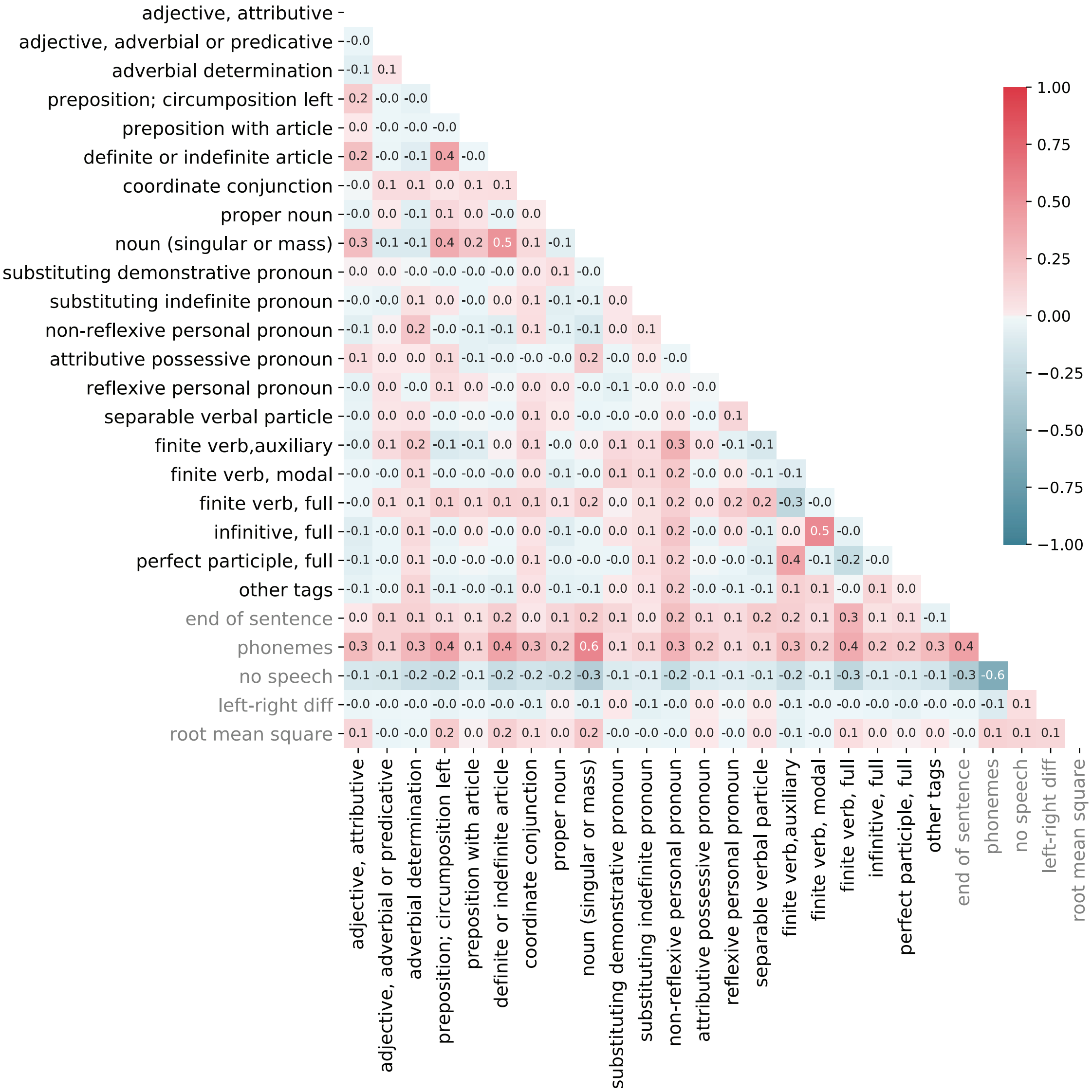

Figure 1. Pearson correlation coefficients of the 26 regressors used in the analysis to validate the annotation. Regressors were created by convolving the events with FSL's "Double-Gamma HRF" as a model of the hemodynamic response function, temporally filtered with the same high-pass filter (cut-off 150 s) as the BOLD time series, and concatenated across runs before computing the correlation.

associated with observed clusters were labeled using the Jülich Histological Atlas[38,39] and the Harvard-Oxford Cortical Atlas[40] provided by FSL.

Figure 2 depicts the results of the three contrasts (z-threshold Z>3.4; p<.05 cluster-corrected). The contrast words > no-speech yielded four significant clusters (see Table 7): one left-lateralized cluster spanning from the angular gyrus and inferior posterior supramarginal gyrus across the superior and middle temporal gyrus, including parts of Heschl's gyrus and planum temporale. A second left cluster in (inferior) frontal regions, including precentral gyrus, pars opercularis (Brodmann Areal 44; BA44) and pars triangularis (BA45). Similarly in the right hemisphere, one cluster spanning from the angular gyrus across the superior and middle temporal gyrus but including frontal inferior regions (pars opercularis and pars triangularis). A fourth significant cluster is located in the left thalamus.

The contrast proper nouns > coordinate conjunctions yielded nine significant clusters (see Table 8): one left-lateralized cluster spanning from the angular gyrus across planum temporale and superior temporal gyrus, partially covering the Heschl's gyrus, into the anterior middle temporal gyrus. A largely congruent but smaller cluster in the right hemisphere. Two clusters in posterior cingulate cortex and precuneus of both hemispheres. Three small clusters in the right occipital pole, right Heschl's gyrus and left superior lateral occipital pole.

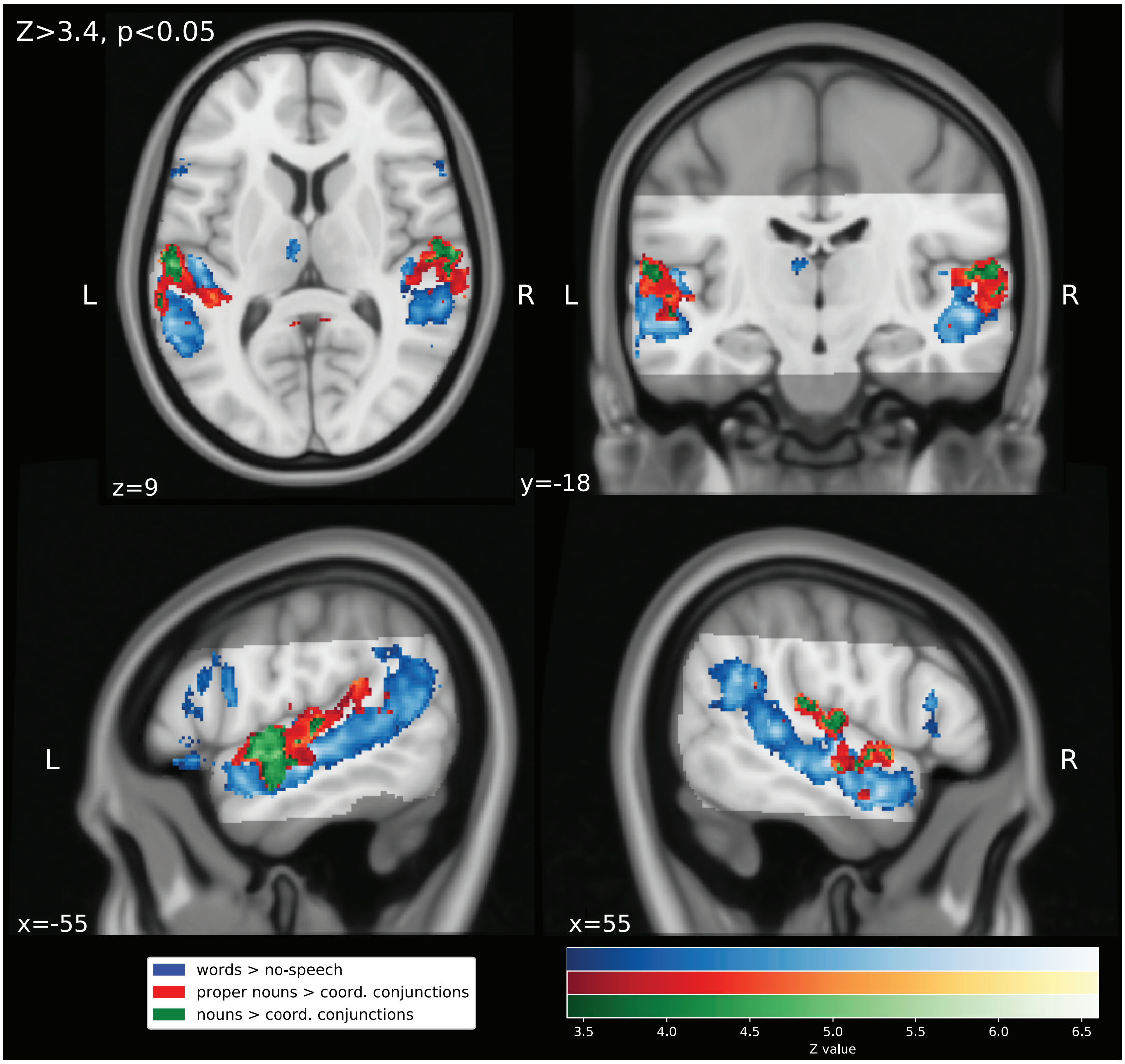

Figure 2. Results of the mixed-effects group-level (N=14) GLM *t*-contrasts for the audio-description of the movie "Forrest Gump". Significant clusters (Z>3.4, p<0.05 cluster-corrected) are overlaid on the MNI152 T1-weighted head template (grey). Light grey: the audio-description dataset's field-of-view (cf.[17]).

Table 7. Significant clusters (z-threshold Z>3.4; p<.05 cluster-corrected) for the contrast words (all 21 `tag`-related regressors) > no-speech. Clusters sorted by voxel size. The first brain structure given contains the voxel with the maximum Z-Value, followed by brain structures from posterior to anterior, and partially covered areas (l. = left; r. = right; c. = cortex; g. = gyrus).

			Max location (MNI)			Center of gravity (MNI)			
Voxels	$p_{corr.}$	**Z-max**	**x**	**y**	**z**	**x**	**y**	**z**	**Structure**
14990	<.001	6.31	-49	-24.7	6.35	-54.8	-32.5	3.73	l. Heschl's g.; lateral superior occipital c., angular g., superior & middle temporal g. (posterior to anterior); parts of supramarginal g. & planum temporale
14469	<.001	6.48	55	-14.9	-6.9	54.1	-23.1	0.374	r. superior temporal g.; angular g., superior (and middle) temporal g. (posterior to anterior), Heschl's g.; parts of supramarginal g., planum temporale, pars opercularis (BA44) & pars triangularis (BA45)
1971	<.001	5.26	-51.1	25.6	-10.5	-53.6	17.8	10.2	l. frontal orbital c.; pars opercularis (BA44), pars triangularis (BA45); parts of precentral g.
217	.002	4.55	-4.48	-13.7	10.3	-6.46	-14.9	9.96	l. thalamus

Table 8. Significant clusters (z-threshold Z>3.4; p<.05 cluster-corrected) for the contrast proper nouns (`ne`) > coordinate conjunctions (`kon`). Clusters sorted by voxel size. The first brain structure given contains the voxel with the maximum Z-Value, followed by brain structures from posterior to anterior, and partially covered areas (l. = left; r. = right; c. = cortex; g. = gyrus).

			Max location (MNI)			Center of gravity (MNI)			
Voxels	$p_{corr.}$	**Z-max**	**x**	**y**	**z**	**x**	**y**	**z**	**Structure**
7691	<.001	6.23	-61.2	-22.3	11.6	-55.9	-20.7	4.03	l. planum temporale; posterior inferior supramarginal g., superior temporal g., planum polare, parts of posterior angular g., Heschl's g., middle temporal gyrus
5928	<.001	5.5	57.5	-26.2	15.9	58.2	-15.8	3.55	r. planum temporale; Heschl's g., superior temporal g., planum polare, temporal pole; parts of angular g. & posterior inferior supramarginal gyrus
479	<.001	4.62	-5.42	-32.3	25.3	-4.28	-39.4	22.8	l. posterior cingulate g.
420	<.001	4.85	-4.76	-71.4	40.1	-3.74	-68.5	36.2	l. precuneus
407	<.001	5.07	6.83	-40.1	24.5	6.67	-38.7	23.1	r. posterior cingulate c.
294	<.001	4.57	17	-69.1	34.6	17.7	-67.1	34.9	r. precuneus
121	.024	3.95	8.12	-98.2	0.359	8.75	-97.7	-3.15	r. occipital pole
117	.027	4.38	36.9	-24.8	4.55	37.4	-23	3.09	r. Heschl's g.
115	.029	4.08	-44.6	-71.7	21.7	-43.6	-70.8	23.4	l. superior lateral occipital c.

Table 9. Significant clusters (z-threshold Z>3.4; p<.05 cluster-corrected) for the contrast nouns (`nn`) > coordinate conjunctions (`kon`). Clusters sorted by voxel size. The first brain structure given contains the voxel with the maximum Z-Value, followed by brain structures from posterior to anterior, and partially covered areas (l. = left; r. = right; c. = cortex; g. = gyrus).

			Max location (MNI)			Center of gravity (MNI)			
Voxels	$p_{corr.}$	Z-max	x	y	z	x	y	z	Structure
3166	<.001	5.75	-61.3	-10.6	-2.93	-57.7	-14.3	1.47	l. anterior superior (and middle) temporal g.; planum temporale, planum polare, anterior superior temporal g.; part of posterior supramarginal g., Heschl's g.
1753	<.001	4.99	63.3	-15.1	8.41	58	-13	4.02	r. planum temporale, anterior superior temporal g., planum polare; part of& part of Heschl's G.
166	.004	4.5	6.83	-40.1	24.5	7.01	-39.7	24.2	r. posterior cingulate g.
149	.008	4.13	18.2	-67.8	36	19.8	-66.4	34.6	r. precuneus

The contrast nouns > coordinate conjunctions yielded four significant clusters (see Table 9): two clusters that are slightly smaller than the lateral temporal clusters of contrast nouns > coordinate conjunction. In this case, spanning from angular gyrus in the left hemisphere and from planum temporale in the right hemisphere into the anterior part of superior temporal cortex. Finally, two small right-lateralized clusters in the right posterior cingulate gyrus and right precuneus.

For the contrast words > no-speech, results show increased hemodynamic activity in a bilateral cortical network including temporal, parietal and frontal regions related to processing spoken language[32,41,42]. These clusters resemble results of previous studies that implemented an ISC approach to analyze fMRI data of naturalistic auditory stimuli[5,43,44]. We do not find significantly increased activations in midline areas (like the posterior cingulate cortex and precuneus or anterior cingulate cortex and medial frontal cortex) which showed synchronized activity across subjects in previous studies. In this regard, our results are similar to[4] who implemented both an ISC and a GLM analysis. In this study, the ISC analysis showed synchronized activity in midline areas but the GLM analysis contrasting blocks of listening to narratives to blocks of a resting condition showed significantly decreased activity in these areas.

The two contrasts that contrasted nouns and proper nouns respectively to coordinate junctions yielded increased activation partially located in early sensory regions (Heschl's Gyrus;[45]) and most prominently adjacent regions bilaterally (planum temporale; superior temporal gyrus;[46,47]). We chose nouns and proper nouns for these two contrasts because they represent linguistically similar concepts but are uncorrelated in the German language and stimulus (cf. Figure 1). We contrasted nouns and proper nouns respectively to coordinate conjunctions because nouns and proper nouns are linguistically different to coordinate conjunctions as well as uncorrelated. Despite the fact that nouns and proper nouns are uncorrelated, both contrasts lead to largely spatially congruent clusters. Results suggest that models based on our annotation of similar linguistic concepts correlate with hemodynamic activity in spatially similar areas. We confirmed the validity of these interpretation by testing if the spatial congruency could be attributed to a negative correlation of coordinate conjunctions with the modeled time series which turned out not to be the case. In summary, results of our exploratory analyses suggest that the annotation of speech meets basic quality requirements to be a basis for model-based analyses that investigate language perception under more ecologically valid conditions.

Data availability

Underlying data

Zenodo: A studyforrest extension, an annotation of spoken language in the German dubbed movie "Forrest Gump" and its audio-description (annotation). https://doi.org/10.5281/zenodo.4382143[29].

Dataset 1. The annotation (v1.0; registered) as a tab-separated-value (TSV) formatted table and a text-based TextGrid file (the native format of the software Praat).

Zenodo: A studyforrest extension, an annotation of spoken language in the German dubbed movie "Forrest Gump" and its audio-description (validation analysis). https://doi.org/10.5281/zenodo.4382188[30].

Dataset 2. The data of the analysis (v1.0; registered) that we ran as a validation of the annotation's content and quality.

Open Science Framework: studyforrest-paper-speechannotation. https://doi.org/10.17605/OSF.IO/GFRME[31].

The paper as LATE X document, and accompanying datasets 1 and 2 (up-to-date; unregistered) accessible as DataLad (RRID:SCR_003931) datasets.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Author contributions

COH designed, performed, and validated the annotation, and wrote the manuscript. MH provided critical feedback on the procedure and wrote the manuscript.

Acknowledgements

COH is grateful to Valeri Kippes who took care of the author's mental sanity by providing excellent training at his gym in Jülich during the mentally draining period of manual corrections of the annotation.

References

1. Sonkusare S, Breakspear M, Guo C: **Naturalistic Stimuli in Neuroscience: Critically Acclaimed.** *Trends Cogn Sci* 2019; **230**(8): 699–714.
   PubMed Abstract | **Publisher Full Text**
2. Eickhoff SB, Milham M, Vanderwal T: **Towards clinical applications of movie fMRI.** *Neuroimage* 2020; page 116860.
   PubMed Abstract | **Publisher Full Text**
3. Hasson U, Landesman O, Knappmeyer B, *et al.*: **Neurocinematics: The Neuroscience of Film.** *Projections* 2008; **20**(1): 1–26.
   Publisher Full Text
4. Wilson SM, Molnar-Szakacs I, Iacoboni M: **Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension.** *Cereb Cortex* 2008; **180**(1): 230–242.
   PubMed Abstract | **Publisher Full Text**
5. Lerner Y, Honey CJ, Silbert LJ, *et al.*: **Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story.** *J Neurosci* 2011; **310**(8): 2906–2915.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
6. Aliko S, Huang J, Gheorghiu F, *et al.*: **A "Naturalistic Neuroimaging Database" for understanding the brain using ecological stimuli.** *BioRxiv* 2020.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
7. Hasson U, Nir Y, Levy I, *et al.*: **Intersubject Synchronization of Cortical Activity During Natural Vision.** *Science* 2004; **3030**(5664): 1634–1640.
   PubMed Abstract | **Publisher Full Text**
8. Bartels A, Zeki S: **The chronoarchitecture of the human brain–natural viewing conditions reveal a time-based anatomy of the brain.** *Neuroimage* May 2004; **220**(1): 419–433.
   PubMed Abstract | **Publisher Full Text**
9. Kauttonen J, Hlushchuk Y, Tikka P: **Optimizing methods for linking cinematic features to fMRI data.** *Neuroimage* 2015; **110**: 136–148.
   PubMed Abstract | **Publisher Full Text**
10. Hamilton LS, Huth AG: **The revolution will not be controlled: natural stimuli in speech neuroscience.** *Lang Cogn Neurosci* 2020; **350**(5): 573–582.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Bartels A, Zeki S: **Functional brain mapping during free viewing of natural scenes.** *Hum Brain Mapp* 2004; **210**(2): 75–85.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
12. Rocca R, Coventry KR, Tylén K, *et al.*: **Language beyond the language system: dorsal visuospatial pathways support processing of demonstratives and spatial language during naturalistic fast fMRI.** *Neuroimage* 2020; **216**: 116128.
    PubMed Abstract | **Publisher Full Text**
13. Lahnakoski JM, Salmi J, Jääskeläinen IP, *et al.*: **Stimulus-Related Independent Component and Voxel-Wise Analysis of Human Brain Activity during Free Viewing of a Feature Film.** *PLoS ONE* 2012; **70**(4): e35215.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
14. Zemeckis R *Forrest Gump [motion picture]* United States: Paramount Pictures; 1994.
15. Koop O, Michalski H, Beckmann R, *et al*. **[German audio description of the motion picture].** *Hörfilm e.V. Berlin and Schweizer Radio und Fernsehen*produced by Bayrischer Rundfunk; 2009.
16. Hanke M, Adelhöfer N, Kottke D, *et al.*: **A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation.** *Sci Data* 2016; **30**: 160092.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
17. Hanke M, Baumgartner FJ, Ibe P, *et al.*: **A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie.** *Sci Data* 2014; **1**: 140003.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
18. Labs A, Reich T, Schulenburg H, *et al.*: **Portrayed emotions in the movie "Forrest Gump" [version 1; referees: 2 approved].** *F1000Res* 2015; **40**(92).
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
19. Lettieri G, Handjaras G, Ricciardi E, *et al.*: **Emotionotopy in the human right temporo-parietal cortex.** *Nat Commun* 2019; **100**(1): 1–13.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
20. Häusler CO, Hanke M: **An annotation of cuts, depicted locations, and temporal progression in the motion picture "Forrest Gump".** *F1000Res* 2016; **5**.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**
21. Boersma P, Weenink D: **Praat: doing phonetics by computer [Computer program] Version 6.0.22.** 2019 retrieved 1 September 2019 from
    Reference Source
22. McAuliffe M, Socolof M, Mihuc S, *et al.*: **Montreal Forced Aligner [Computer program] Version 1.0.1.** 2019 retrieved 5 September 2019 from
    Reference Source
23. Gorman K, Howell J, Wagner M: **Prosodylab-aligner: A tool for forced alignment of laboratory speech.** *Canadian Acoustics* 2011; **390**(3): 192–193.
24. "FFmpeg Developers": **FFmpeg tool [Computer program] Version 4.1.4.** 2019
    Reference Source
25. Honnibal M, Montani I: **spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing [Computer program] Version 2.2.1.** 2017
    Reference Source
26. Brants S, Dipper S, Eisenberg P, *et al.*: **TIGER: Linguistic Interpretation of a German corpus.** *Research on Language and Computation* 2004; (4): 597–620.
    Publisher Full Text

27. Schiller A, Teufel S, Stöckert C, *et al.*: **Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).** *Technical report* 1999.

28. Gorgolewski KJ, Auer T, Vince D, *et al.* **The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments.** *Sci Data* 2016; **3**: 160044.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Häusler CO, Hanke M: **A studyforrest extension, an annotation of spoken language in the German dubbed movie "Forrest Gump" and its audio-description (annotation).** December 2020.
Publisher Full Text

30. Häusler CO, Hanke M: **A studyforrest extension, an annotation of spoken language in the German dubbed movie "Forrest Gump" and its audio-description (validation analysis).** December 2020.
Publisher Full Text

31. Häusler CO, Hanke M: **studyforrest-paper-speechannotation.** December 2020.
Publisher Full Text

32. Hickok G, Poeppel D: **The cortical organization of speech processing.** *Nat Rev Neurosci* 2007; **80**(5): 393–402.
PubMed Abstract | **Publisher Full Text**

33. Woolrich MW, Ripley BD, Brady M, *et al.*: **Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data.** *Neuroimage* 2001; **140**(6): 1370–1386.
PubMed Abstract | **Publisher Full Text**

34. Smith SM, Jenkinson M, Woolrich MW, *et al.*: **Advances in functional and structural MR image analysis and implementation as FSL.** *Neuroimage* 2004; **23**: 208–219.
PubMed Abstract | **Publisher Full Text**

35. Beckmann CF, Jenkinson M, Smith SM, *et al.*: **General multi-level linear modelling for group analysis in FMRI.** *Neuroimage* 2003; **200**(2): 1052–1063.
PubMed Abstract | **Publisher Full Text**

36. Woolrich MW, Behrens TEJ, Beckmann CF, *et al.*: **Multilevel linear modelling for FMRI group analysis using Bayesian inference.** *Neuroimage* 2004; **210**(4): 1732–1747.
PubMed Abstract | **Publisher Full Text**

37. Woolrich M: **Robust group analysis using outlier inference.** *Neuroimage* 2008; **410**(2): 286–301.
PubMed Abstract | **Publisher Full Text**

38. Eickhoff SB, Stephan KE, Mohlberg H, *et al.*: **A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data.** *Neuroimage* 2005; **250**(4): 1325–1335.
PubMed Abstract | **Publisher Full Text**

39. Eickhoff SB, Paus T, Caspers S, *et al.*: **Assignment of functional activations to probabilistic cytoarchitectonic areas revisited.** *Neuroimage* 2007; **360**(3): 511–521.
PubMed Abstract | **Publisher Full Text**

40. Desikan RS, Ségonne F, Fischl B, *et al.*: **An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.** *Neuroimage* 2006; **310**(3): 968–980.
PubMed Abstract | **Publisher Full Text**

41. Friederici AD: **The Brain Basis of Language Processing: From Structure to Function.** *Physiol Rev* 2011; **910**(4): 1357–1392.
PubMed Abstract | **Publisher Full Text**

42. Price CJ: **A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading.** *Neuroimage* 2012; **620**(2): 816–847.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

43. Honey CJ, Thompson CR, Lerner Y, Hasson U: **Not Lost in Translation: Neural Responses Shared Across Languages.** *J Neurosci* 2012; **320**(44): 15277–15283.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Silbert LJ, Honey CJ, Simony E, *et al.*: **Coupled neural systems underlie the production and comprehension of naturalistic narrative speech.** *Proc Natl Acad Sci U S A* 2014; **1110**(43): E4687–E4696.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Saenz M, Langers DRM: **Tonotopic mapping of human auditory cortex.** *Hearing Research* 2014; **307**: 42–52.
PubMed Abstract | **Publisher Full Text**

46. Arsenault JS, Buchsbaum BR: **Distributed Neural Representations of Phonological Features during Speech Perception.** *J Neurosci.* 2015; **350**(2): 634–642.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

47. Mesgarani N, Cheung C, Johnson K, *et al.*: **Phonetic Feature Encoding in Human Superior Temporal Gyrus.** *Science* 2014; (6174): 1006–1010.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✓ ? ?

Version 1

Reviewer Report 01 March 2021

https://doi.org/10.5256/f1000research.30529.r78808

?

Roberta Rocca

Psychoinformatics Lab, Department of Psychology, The University of Texas at Austin, Austin, TX, USA

The authors introduce a new set of linguistic annotations for the *studyforrest* dataset, an open naturalistic fMRI dataset where subjects are presented audio-visual or audio-only versions of the movie Forrest Gump. This new corpus of annotations is an extremely valuable addition to the dataset. The availability of high-quality annotations of linguistic content enables researchers to easily set up and conduct analyses tapping into the neural correlates of phonetic, semantic and syntactic processing. Furthermore, making multi-level (word-, phoneme-, and sentence-level) time-stamped transcripts of the stimulus makes it possible to extract any linguistic features beyond those directly provided by the authors. Access to the new annotations is made simple through DataLad. Additionally, the authors report the results of a few validation analyses where annotated features are used in a GLM setting to retrieve established effects (e.g., neural correlates of speech and semantic processing), which speaks in favor of their reliability.
Overall, the manuscript is clear, the annotation pipelines used are solid and the data sharing format is easy to navigate. It is a great contribution to the field, and to open science culture more generally. I have a few suggestions that mostly concern: a) increasing the overall clarity and readability of the manuscript; b) making it easier for readers who are not familiar with the dataset to understand how annotations map onto the various parts of the studyforrest dataset and use them; c) making it easier for readers with limited expertise in linguistic annotations to track the meaning of individual features.

Introduction

1. In the introduction, the authors make a general point about model-driven GLM analyses providing a series of advantages over data-driven methods, especially in terms of interpretability. They write that “use of data-driven methods alone falls short of associating results with particular stimulus events”, while GLMs “allow to test hypotheses on specific brain functions under more ecologically valid conditions, to statistically control for confounding stimulus features, and to explain not just “how” the brain is responding to a

stimulus, but also why". A few objections can be moved to these claims (or to the way they are presented here). While GLMs are, on the surface, more easily interpretable, their use (especially in naturalistic contexts) poses a few challenges. For example, a) global characteristics of the stimulus make interpretation of baselines tricky (and their commensurability across datasets challenging), with consequences on the interpretation of the effects of specific features; b) the choice of covariates, while allowing to control for collinearity, can also alter the interpretation of the effects of individual features – in other words, interpretation is conditional on the model; c) causal claims (aka the *why*) are notoriously hard to make on the basis of linear models alone. While I am sympathetic to the idea that a lot can be gained by the use of these simple tools, authors could consider nuancing some or further specifying some of their claims.

2. My next (related) point concerns the link between the annotations presented by the authors and GLMs. In the introduction, the authors seem to present the usability of their annotations as almost constrained to analytic contexts where GLMs are used or as heuristics tools for data-drive methods. But nothing prevents these annotations from being used as either predictor or target features in other analysis frameworks (e.g., predictive models). The authors could consider being a bit more ambitious, and slightly reframing the first paragraphs along the lines of these considerations.

3. In the introduction, the authors claim that "stimulus annotations can inform data-driven methods about a stimulus' temporal dynamics". I think the gist of it is clear, but maybe a more concrete example or wording would help.

4. A minor point about wording, should "the current publication" be "the present publication" or "this publication"? Not a native speaker, so it may just be an issue with my English (in which case, please ignore this comment).

Materials and methods

1. It may be beneficial to mention early in the paper that the movie is presented in German. It is also a great feature of the dataset, being it one of the very few dataset – to my knowledge – with stimuli *not* in English, which is quite crucial for matters of cross-linguistic generalization.

2. One aspect that is not entirely clear from the paper is which component of the *studyforrest* dataset the annotations refer to / can be used for. Can they only be used for data from subjects where the movie was only presented auditorily, or can the same annotations (filtering out "narrator" rows) also be used for those parts of the dataset where participants are presented with the movie both auditorily and visually? I think it will be highly beneficial to the manuscript to a) provide a little recap of what studyforrest is, its sub-components, and where to find them; b) make it more explicit for which batches of subjects/tasks these new annotations can be used.

3. More of a clarification question than a suggestion: the onsets in the annotations tsv file refer to the full movie file. It should be possible to cross-reference these onsets with run- and subject-specific events files from the BIDS dataset. If so, could more details on how to get from onsets in the annotation files to run-specific GLM-ready onsets be provided (at a high level, e.g., "onsets in the annotation files can be cross-referenced with time-stamps in

the event files x and x to retrieve subject-specific and run-specific onsets ...")? Disclaimer: I am relatively new to BIDS, and this may be a trivial point (in which case ignore this comment).

4. Once again a minor point, but the fact that annotations are BIDS-compliant and shared in tsv could already be mentioned in the introduction (or even in the abstract).

5. In "Annotation procedure", the authors introduce the fact that the stimulus is also annotated at the sentence-level. However, a definition of a sentence is only given in the description for Table 1. Could a quick reference of what counts as a sentence be added to the main text?

6. The authors mention having dropped the "less dominant" voice when speakers overlap. Could a couple of words be added on how that was defined (volume? character prominence?)? And since some could argue these events are also potentially interesting bits of information, do instances of overlaps represent a significant portion of the stimulus, or does this occur only in few instances?

7. On page 4, the authors talk about feature extraction in terms of analysis (e.g., "automatically analyze linguistic features of each word in their corresponding sentence" and performed "analyses" regarding part of speech). Probably a matter of taste, but would rather refer to these steps as "extraction" of linguistic features; to me, the use of "analysis" suggests that more has been performed on the features (e.g., visualization or some form of validation) than mere feature extraction.
8. Which type of word embeddings were extracted should be specified.

9. SpaCy has very good documentation explaining the interpretation of each feature and how pipelines for extraction work (e.g. https://spacy.io/usage/linguistic-features). References to that (e.g., simply links to the documentation) could be added, for example in the sentence "We then performed analyses regarding part-of-speech... " whenever each feature/pipeline is introduced. It would make it easier for non-linguists to get a feel for what features mean, and for experts to dig into the details of the specific pipelines used.

10. The paragraph "Annotation procedure" may benefit from adding an example, e.g., a sample sentence from the transcript with the corresponding annotation. It would help readers (especially those who do not have deep expertise in linguistics) to visualize what features are about and visualize the different levels of annotation. I understand this may be tricky to do without disrupting the flow of the text though and will leave it to the authors to decide whether to implement it or not.

11. In the caption for table 2, can the expression "sentences spoken by the ten most often occurring speakers" be streamlined?

12. Can a sample of the resulting annotation dataset (whose components are described in "data legend") be displayed as exemplification, e.g., just the head of the table or something similar?

Dataset validation

1. Can the (run-level) design matrix be shown, possibly with labels for the regressors? Would make for a nice visual, and help readers better understand the structure of the analysis.

2. It is not entirely clear to me why authors annotated for the onset/duration of phonemes, but only using the 80 most frequent phonemes. What is the role of that regressor in the model, what is it meant to code and/or control for? Can this be motivated more in detail in the text?

3. Another aspect of the analysis which is not entirely clear to me is why “no speech” events were created ad hoc. Wouldn’t the group-level z-map for a binary regressor coding for the onset of all speech events (or even just all words) do the same job without the need for mock events? Or couldn’t a comparison between word events and events tagged as NONSPEECH in the ‘pos’ column of the dataset be an easier implementation of the same question? The analyses are convincing and sound to me considering the purpose of this paper, so I’m not necessarily suggesting to change them. But the authors could consider further clarifying the rationale for their choices.

4. When mentioning annotations of low-level auditory features, the authors refer to “continuous bins of information about low-level feature”. The wording here could be made more transparent.

5. Can the authors provide some motivation for the choice of thresholding at z>3.4?

6. I find the representation of contrasts in Figure 2 a bit confusing, because of the combinations of high overlaps between effects and of the colormap (not a lot of contrast between the three colors). If that is not uniquely my concern, the authors could consider either splitting the map into separate figures or experimenting with alternative color combinations.

It was a pleasure to review this work. Great set of resources, and great contribution to the field!

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: naturalistic fMRI, natural language processing, psychoinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 February 2021

https://doi.org/10.5256/f1000research.30529.r78810

Martin Wegrzyn
Department of Psychology, Bielefeld University, Bielefeld, Germany

The authors present a word-for-word annotation of an audio version of the movie “Forrest Gump”, as used in the “studyforrest” project, for which 7T fMRT data of 20 participants is available (as described in other publications). The presented work provides a number of syntactic and semantic labels, which can be useful for studying language using this naturalistic stimulus. An additional validation of the dataset, by using a subset of its parameters as a design for fMRI analysis, illustrates its applicability. The rationale for creating the dataset is clearly described. The protocols used are appropriate and the work is technically sound. Overall, sufficient details of methods and materials are provided for the main syntactic markers of speech. However, some of the additional semantic annotations are not described in enough detail, in my opinion. The created datasets are clearly presented and are in a usable and accessible format, especially if one is a Python user.

This is a high-quality work that I think will substantially boost the usability of the 2014 “studyforrest” publication by Hanke and colleagues. I think naturalistic stimuli are very valuable; Especially for data sharing, their openness to different scientific questions is a strong asset. Right now (2021), many PhD-students will have a hard time collecting fMRI data or are unable to collect data altogether. I would not be surprised if this publication (together with Hanke 2014 Scientific Reports) will give rise to a dozen or more valuable publications on language-fMRI.
Having worked with some “studyforrest” data myself, and having tried to annotate the stimulus, I think I can appreciate the level of dedication required to complete an annotation at the level of the present one (I know I couldn't do it). Because a naturalistic stimulus is not designed for experimental usage, many parts of it will be more ambiguous than one would like it to be. Stemming in part from my own experience, I therefore also have a number of specific questions regarding how the authors solved those challenges.

Questions and comments:

- At the end of the first paragraph/beginning of the second, the authors write: *“Thus, the temporal structure of their feature space is usually not explicitly known, leading to an 'annotation bottleneck'[6] when used for neuroscientific research. Data-driven methods like inter-subject correlation (ISC)[7] or independent component analysis (ICA)[8] are often used to analyze such fMRI data in order to circumvent this bottleneck. However, use of data-driven methods alone falls short of associating results with particular stimulus events [9]”*

I understand the point the authors are making. Namely, that it usually takes much more time to annotate the data than to collect them. However, I would like to suggest that the wording ("temporal structure", "feature space", "bottleneck"), especially in the first paragraph of the article, might be a bit too challenging to immediately understand. Maybe a gentler introduction to the topic would be helpful for some readers.
Also, I think that the point the authors make is not 100% valid: The study by Hasson et al. (cited in [7]) actually circumvented the problem of annotating the whole stimulus, by using "reverse correlation": That is, only interpreting the time points where activity is highest. This allowed Hasson and colleagues to interpret effects on a substantive level, just by showing the frames of the movie for which brain activity was highest. I think this is also a useful and economical approach – but certainly more limited than a full annotation.

○ In the description of Table 1 the authors write: *"The category 'sentences' comprises complete grammatical sentences which are additionally marked in the annotation with a full stop at the end ('my feet hurt.'). It also comprises questions ('do you want a chocolate?'), exclamations ('run away!'), or non-speech vocalizations in quick succession ('ha, ha, ha'), or in isolation (e.g. 'Forrest? ', 'Forrest!', 'ha') at time points when speakers switch rapidly. The category words comprises each word or non-speech vocalization (N=202)."*
I completely understand that the additional coding of whether a speech episode is a legal sentence in German would go beyond the scope of the present work. Also, users will be able to code this themselves, according to their needs, using the detailed tagging provided by the authors (and the dot at the end is also useful, though not a perfect marker (e.g. "lief und lief." or "Hörfilm e. V.")). However, I think that calling a variable or feature "sentences", when it really is not coding the presence of a sentence, could be a source of future errors. The same goes for "words". So maybe the authors would want to consider using different labels for those two features.

○ Related to the point above, I think that sometimes a single sentence has longer pauses and is then divided into multiple "sentences". Take the narration at the very end of the movie: *"S ie wird wieder angehoben, ..."* , *"und schwebt auf die Kamera zu."* is coded as two sentences. Researchers interested in syntax might benefit from being aware that they have some additional work to do (which is only fair!), in order to merge those speech episodes which are really a single sentence. Additionally, if they define the sentences differently, they will have to do a different tagging of the words, too. I do not suggest that the authors need to do any additional work, but I just want to point out that the label "sentences" could mislead some language researchers.
Apart from sentences being divided by pauses, there is the question of punctuation. Given that the stimulus is audio only, my personal experience is that it is sometimes difficult to decide on a definitive punctuation (where commas should occur or if a comma or a full stop should be used). I think that in German commas play a bigger role in defining the syntactic structure of a sentence than they might do e.g. in English. Therefore, I would suggest that some brief discussion of the general limitations of deriving syntactic information from an audio stimulus might be a useful addition to the article (of course only if the authors agree with my premise).

○ "Data Legend" > "Common Word (stop)": The source or the definition of the stop list would be helpful for users who would want to cite it in their publications

- “Data Legend” > “Word embedding”: This is a very rich and useful resource and I was immediately able to do some interesting analyses using the 300-dimensional word vector. For example, correlating the values of all words with each other gave some very interesting and plausible results (e.g. color words clustering together etc.). I think this itself could merit a separate publication with a focus on semantics. As it currently stands, there is however no information about how these data were generated and how future users could cite them and find out more about them. Therefore, I think that a bit more information could significantly increase the impact of this resource.

- Also, when playing around with it, I noticed that sometimes two different words have the exact same 300-dimensional vector. For example:
  strahlend-bewölkt
  überqueren -durchqueren
  braun-olivgrün
  gepackten-schlamm
  schütteln-wischen
  Is this correct behavior and why would two different words have the exact same values? Especially for a pair like “gepackten”-”Schlamm” that might be surprising.

- “Dataset validation”: maybe some brief examples for the “coordinating conjunction” and different “nouns” could be given for illustrative purposes

- “Dataset validation”: why were non-speech events not modeled with a boxcar, spanning their whole duration, when this was otherwise possible for single words. It would seem to me that the duration of individual words would make a boxcar function much more problematic than using a boxcar for speech-free periods (which are often short, but there are some really long ones in there as well, which might not be fully utilized when modeled with a default length of 70ms – which would amount to a stick function).

- “Dataset validation”: how do the non-speech events correlate with the other features (i.e. could they be included in the heatmap in Fig 1?)

- Tables 7-9: could it be better to round the values to full mm? I would argue that even for 7T data (smoothed with 4mm FWHM), sub-mm precision for the MNI-coordinates might imply an accuracy that is not obtainable

- Figure 2 legend (and pages 7, 12): I recently learned that “cf.” is used to point to contrasting information (for example, if there is literature making an opposing point; https://blog.apastyle.org/apastyle/2010/05/its-all-latin-to-me.html).

Thank you!

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: Invited the authors to a symposium on naturalistic fMRI, scheduled for March 2021; using their data from studyforrest.org for my own work

Reviewer Expertise: fMRI, language

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 16 February 2021

https://doi.org/10.5256/f1000research.30529.r78806

✓ **Giada Lettieri**
Social and Affective Neuroscience Group, MoMiLab, IMT School for Advanced Studies Lucca, Lucca, Italy

The current Data Note reports and describe an extensive dataset including speech annotation in the audio-visual movie "Forrest Gump". To assess the consistency of annotation, the brain hemodynamic activity elicited by the contrast of speech events versus no speech was also investigated.
The dataset here provided is richly detailed and is a great addition to the work already done in the studyforrest project. The significant effort carried out by the authors is particularly relevant for data sharing and for future investigations on language and lifelike experiences.

In light of all this, I would only have two suggestions:

1. I think it would help readers that are not familiar with the original dataset to have in the header of tables the indication "run 1, run 2..." or "segment 1, segment 2...", instead of just the numbers from 1 to 8.

2. It was not fully clear to me which kind of algorithm the authors used to obtain the word-embedding. As different algorithms provide different results for word embedding, I think it would be interesting to specify which one of these was employed in their analyses (e.g., GloVe, word2vec).

In conclusion, this extension enriches the already available annotations of the original neuroimaging dataset and concur in building a comprehensive and valuable description of a naturalistic stimulation.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroscience

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.