Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 195–203
Marseille, 11–16 May 2020

Using the RUPEX Multichannel Corpus in a Pilot fMRI Study on Speech Disfluencies

Katerina V. Smirnova[1], Nikolay A. Korotaev[2], Yana R. Panikratova[3], Irina S. Lebedeva[3], Ekaterina V. Pechenkova[4], Olga V. Fedorova[1]
[1]Lomonosov Moscow State University, [2]Russian State University for the Humanities, [3]Mental Health Research Center, [4]National Research University Higher School of Economics
[1]Leninskie Gory 1, [2]Miusskaya sq. 6, [3]Kashirskoye sh. 34, [4]Armyanskiy ln. 4, bld. 2, Moscow, Russia
kategold1@gmail.com, n_korotaev@hotmail.com, panikratova@mail.ru, irina.lebedeva@ncpz.ru, evp@virtualcoglab.org, olga.fedorova@msu.ru

Abstract

In modern linguistics and psycholinguistics speech disfluencies in real fluent speech are a well-known phenomenon. But it's not still clear which components of brain systems are involved into its comprehension in a listener's brain. In this paper we provide a pilot neuroimaging study of the possible neural correlates of speech disfluencies perception, using a combination of the corpus and functional magnetic-resonance imaging (fMRI) methods. Special technical procedure of selecting stimulus material from Russian multichannel corpus RUPEX allowed to create fragments in terms of requirements for the fMRI BOLD temporal resolution. They contain isolated speech disfluencies and their clusters. Also, we used the referential task for participants fMRI scanning. As a result, it was demonstrated that annotated multichannel corpora like RUPEX can be an important resource for experimental research in interdisciplinary fields. Thus, different aspects of communication can be explored through the prism of brain activation.

Keywords: multichannel corpus, speech disfluency, functional magnetic-resonance imaging.

1. Introduction

In natural communication, people often encounter problems in their discourse production. This may result in speech disfluencies, that is, various deviations from the "ideal", fluent delivery (Clark and Clark, 1977). Numerous classifications of disfluencies have been proposed and / or implemented in annotating speech corpora, see, Maclay and Osgood, 1959; Levelt, 1993; Schriberg, 1994; Eklund, 2004, inter alia. They are mostly based on the speaker's perspective. When planning their production, speakers may hesitate on a better way to express their intentions and search for a verbalization they will not have to reject afterwards. Alternatively, speakers may find an already-uttered discourse fragment inappropriate, stop their speech production "and then abort, recast or redo" their utterances, thus realizing self-repair (Fox et al., 2009: 59). Combined types of disfluencies are also possible, as will be shown in Section 3. Still, speech disfluencies also influence perception, as listeners may interpret them as signals of planning difficulties, possible places for turn-taking, and so on (see, e.g., Fox Tree, 2001; Barr and Seyfeddinipur, 2010). In this paper, we address various types of speech disfluencies from yet another perspective. We conducted a pilot event-related functional magnetic-resonance imaging (fMRI) study to reveal the neural correlates of disfluency perception in the listener's brain.

This research comes in line with the interdisciplinary initiative for analyzing multichannel discourse (see Kibrik and Fedorova, 2018a, b). We use data from the "Russian Pears Chats and Stories" (RUPEX) corpus, which was created as a resource for conducting multimodal studies. So, contrary to many experimental studies, the participants were presented excerpts of natural communication, but not scripted or enacted material. This methodology is similar to that described in Eklund and Ingvar's (2016) fMRI-based account on filled vs. unfilled hesitation pauses. According to their results, both filled and unfilled hesitation pauses cause extra activation of the primary auditory cortex (PAC) and the motor components of the speech system in the brain. Specifically, for filled pauses, a more expressed activation was detected in the additional motor cortex (SMA), which is known to be involved into the initiation of utterances. In our study, we tested these results against Russian data and analyzed more types of disfluencies.

The paper is organized as follows. In Section 2, the RUPEX corpus is described and its theoretical premises are briefly discussed. In Section 3, principles for disfluency annotation are presented. Section 4 describes the conducted fMRI study and presents its results. Section 5 concludes the body of the paper.

2. RUPEX Multichannel Corpus

The "Russian Pear Chats and Stories" corpus (RUPEX; see Kibrik and Fedorova, 2018c; https://multidiscourse.ru/main/?en=1) was created within the framework of a multichannel approach to natural communication. As is often pointed out in studies on multimodality (see McNeill, 2005; Kibrik, 2010; Loehr, 2012; Adolphs and Carter, 2013; Goldin-Meadow, 2014; Müller et al. eds., 2014; Church et al. eds., 2017, Kibrik, 2018, inter alia), we interact with one another using not only verbal material, but also additional means such as intonation, gestures, facial expressions, and eye gaze. Two major modalities can be distinguished. The vocal modality involves the segmental verbal channel, as well as a wide spectrum of non-segmental, prosodic, sound phenomena. The kinetic modality involves all kinds of movements (or significant absence of movement) performed by various body parts of an interlocutor: eyes, face, head, hands and arms, etc. In RUPEX, recordings of natural communication among several participants are integrated with their vocal and kinetic annotations.

The RUPEX received its name after the so-called "Pear Film", a stimulus material widely used in linguistics. The Pear Film was created by a research group directed by Wallace Chafe in Berkeley in the 1970s (Chafe ed., 1980).

The film presents an interaction of several characters, among them a gardener who is picking pears, a boy who steals a basket of pears, and others along the way. The film does not contain any talk but has some natural sounds, such as a rooster crowing. This film offers a well-structured chain of physical and social events and has long established itself as an excellent way to obtain compact and comparable retellings.

The RUPEX consists of separate communication episodes, each involving four participants: a Narrator, a Commentator, a Reteller, and a Listener. At the beginning of each session the film was watched by two participants: those who subsequently acted as the Narrator and the Commentator. After the film viewing was finished, the Narrator, the Commentator, and the Reteller were seated as shown in Figure 1.

Then communication as such started. It consists of three consecutive stages.

- Monologic stage: First Telling. The Narrator tells the Reteller the content of the film.
- Interactive stage: Conversation. The Commentator elaborates on the story, and the Reteller asks questions addressed to both interlocutors who have seen the film.
- (After that, the Listener joins in.)
- Monologic stage: Retelling. The Reteller retells the film to the Listener.

In the additional final stage, the Listener writes down the second retelling of the film. Note that the two final stages (Retelling and writing down) are important as they encourage all participants to engage in motivated, comprehensive and meaningful communication.

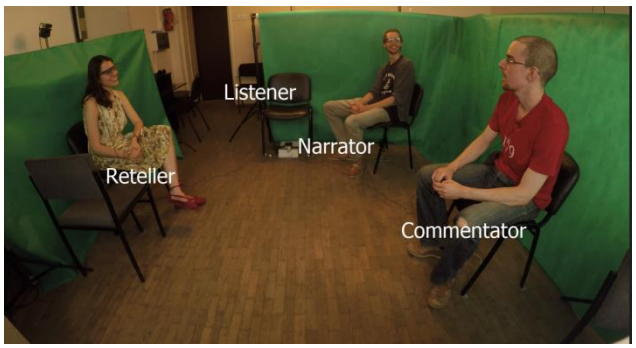

Figure 1. General design of the communication situation.

The RUPEX was recorded in two installments. The first part was collected in the summer of 2015. It contains 24 sessions with a total duration of approximately 9 hours, with an average duration of 23 minutes per session (ranging from 12 to 38 minutes), and a volume of about 100,000 words; a total of 96 people aged from 18 to 36 years old, 34 of whom were men and 62 of whom were women, took part in the sessions. The second part of the corpus was collected in the summer of 2017. It contains 16 sessions with a total duration of approximately 6 hours, with an average duration of 21 minutes (ranging from 8 to 41 minutes), and a volume of about 60,000 words; 64 people aged from 18 to 36 years old, including 16 men and 48 women, took part in the sessions.

For this study, we used a subcorpus of three sessions recorded in 2015 (marked as 04, 22, and 23 in the examples below). To facilitate the perception during the fMRI study, only monologic sequences were selected. These include the complete First Telling and Retelling stages of all the three sessions, as well as several monologic contributions of the Commentators during the Conversation stages of sessions 04 and 23. Overall, 8 speakers (6 F, 2 M) contributed to the analyzed data, the total duration is approximately 32 minutes.

3. Disfluency Annotation

As indicated in Section 1, most classifications of disfluencies recognize the difference between hesitation and repair; see also (Podlesskaya and Kibrik, 2009: 178-181) for an account based on Russian data. We also adopt this basic distinction, though we find that combined disfluencies are also quite frequent.

We used vocal transcripts of the analyzed subcorpus carried out according to the principles described in (Kibrik and Podlesskaya eds., 2009; Kibrik et al., 2020). In the transcripts, each text line corresponds to an elementary discourse unit (EDU), that is, a minimal step in natural speech production. In the examples below, the IDs of the involved EDUs are provided in parentheses after the translation. Other conventions can be found in (Kibrik et al., 2020) and on https://multidiscourse.ru/annotation/?en=1. For this study, we additionally annotated the following types of disfluencies.

1. Silent pauses (SPs). As it has often been discussed, SPs cannot be unequivocally treated as hesitation markers (see Eklund, 2004: 160-162 and references provided there). While they do occur in hesitation contexts, SPs also play an important role in discourse segmentation (see Chafe, 1994, inter alia) and there is no clear-cut distinction between these two functions. Formal and perceptive criteria are sometimes used to delimit the hesitation uses of SPs (see, e.g., Bouraoui and Vigouroux, 2005; Trouvain et al., 2016). We followed the same line and annotated SPs as disfluency markers only when they occurred inside a highly integrated syntactic phrase and / or seemed exceedingly long in the given context. Since there are too many factors that influenced the annotators' perception in this case (e.g., syntactic position, individual variation, etc.), we did not use a specific threshold value, but relied on the inter-annotator agreement[1]. For instance, in (1), the SP of more than 600 ms appears inside the VP, the most plausible interpretation being that the speaker seeks for an appropriate verb to express the idea of placing the pears into the indicated receptacle.

(1)	on	èti	/gruši	tuda	(0.61)	\skl̲adyvaet.
	he	these	pears	there	SP	puts

[1] There were two annotators for each fragment; in case of disagreement, a discussion took place that could involve other annotators. The same procedure applied to the cases of lengthening.

‘He puts these pears there’(23_R-vE156)

Pauses filled with inhalations were treated the same way as ‘bare’ SPs.

2. Filled pauses (FPs). There is little agreement on whether items like *uh* and *um* should be treated as filled pauses or as lexical units (see somewhat opposite accounts in Clark and Fox Tree, 2002; Corley and Stewart, 2008). However, their intrinsic hesitation function is beyond doubt. Similar to other languages, Russian FPs may have different phonological shapes with, potentially, different distribution patterns. In this study, we neglected these variations and treated all FPs the same. In (2), an *uh*-like FP transcribed as *(ə)* appears between the subject and the VP of a simple clause. In terms of information structure, the hesitation marks here the transition from the topical NP *mal’čik* ‘boy’ to the focal predicate. The epistemic adverb *vidimo* ‘apparently’ also indicates the speaker’s uncertainty.

(2) /Mal’čik (ə 0.38) vidimo /ušibsja,
boy FP apparently hurt.himself

‘The boy apparently got hurt’ (22_R-vE166)

3. Lengthening, or phoneme prolongation, is functionally and formally close to FPs, as it also marks hesitation by means of extended vocalization (Eklund, 2001). In example (3), the speaker produces the singular verb form *pokazyvaetsja* ‘is shown’ and then realizes that the subsequent noun *derev’ja* ‘trees’ is meant to be plural. She hesitates on whether to avoid this grammatical inconsistency — by correcting the verbal form or by choosing an appropriate singular noun — and decides not to. The hesitation is signaled by the final lengthening of the reflexive verbal suffix *-sja* (see the hyphenated notation in the transcript).

(3) nam pokazyvaetsja-a /derev’ja,
to.us is.shown trees

‘Trees are shown to us’ (22_N-vE011)

However, contrary to FPs (but similar to SPs), lengthening doesn’t necessarily signal hesitation. In Russian, it may also express emphasis, or intensification, or be part of a standard prosodic pattern associated with the meaning of inexhaustiveness (see Kibrik et al., in print). Here, we relied on perception and co-occurrence (see below); only those cases of lengthening that had an obvious hesitation flavor were annotated as disfluencies.

4. Self-repairs. In contrast to the previous types, self-repairs involve explicit correction (or repetition) of a previously aborted vocalization. We rely on the classical model of self-repair proposed by Schriberg (1994), which includes three compulsory elements: reparandum, interruption point (often signaled by word truncation), and repair (or, reparans). In example (4), the speaker corrects the case marking of the reciprocal pronoun *drug druga* ‘each other’. In the reparandum, the form *drug s drugom* ‘with each other’ is being constructed, but as soon as the speaker realizes that this form is inappropriate, she interrupts her production (see Levelt, 1993: 478) and provides the repair, i.e. the corrected form *drug drugu* ‘to each other’. The interruption point is indicated by a || symbol in transcripts and glossed as BR (from “break”).

(4) oni \edut drug s dru= ||
they go with.each.other BR

drug drugu /navstreču,
to.each.other towards

‘They are moving towards each other’ (22_R-vE150)

Classifications of self-repairs often account for different relations between reparandums and reparanses and structural levels of repair (see, i.a., Schegloff, 2013; Podlesskaya, 2015). For the sake of simplicity, we neglect these differences here (but see Podlesskaya et al., 2019 for a distribution of self-repairs in RUPEX). Also, only self-repairs with an easily perceptible interruption point were selected for this study.

5. Other types of disfluencies include special lexical items like *nu* ‘well’, placeholders, editing terms like *net* ‘no’, and so on. They are less frequent in our data and will not be discussed further.

Examples (1) – (4) above demonstrate ***isolated***, or simple, disfluencies, i.e. cases where a disfluency appears inside an otherwise fluent speech fragment. Quite often, however, disfluencies come in ***clusters***. In (5), a disfluency cluster starts with a short SP (150 ms), which is followed by a prolongated numeral *d-dve-e* ‘two’, an interruption point, a complex sequence of SPs and FPs, and, finally, the repair, *tri* ‘three’ instead of ‘two’.

(5) i u nego stoit (0.15) /–d-dve-e || (0.23)
and by him stand SP two BR SP

(ɯ 0.14) (ə 0.17) (0.16) /\tri nebol’šix /korziny,
FP SP three not.big baskets

‘And he has two three middle-sized baskets there’ (23_R-vE153)

To discriminate between isolated and clustered disfluencies, we used the adjacency principle. A disfluency was annotated isolated if it was separated from the closest disfluency by at least two fluent words. Otherwise, it was annotated as an element within a cluster. As shown in Table 1, clusters are more frequent in our data than strictly isolated disfluencies (cf. similar results for English and French in Crible et al., 2017). However, the most frequent are somewhat intermediate cases where an otherwise isolated disfluency is accompanied by a silent pause. To facilitate the design of the event-related fMRI study, these cases were treated the same way as strictly isolated disfluencies.

Disfluency category	**Occurrences**
Isolated disfluencies	116
Isolated disfluencies with SPs	158
Disfluency clusters	135
Total	***404***

Table 1. Number of isolated and clustered disfluencies in the annotated subcorpus.

For each isolated and clustered disfluency its starting and ending times were indicated. As RUPEX provides time codes for every word and pause, we relied on those for FPs

and SPs. For lengthening, we manually localized prolongated segments in PRAAT (Boersma and Weenink, 2012; http://www.praat.org/). For self-repairs, the interruption point was taken as both the starting and the ending time. Minor (non-hesitant) SPs that accompanied disfluencies affected their ending, but not their starting time. For instance, the disfluency cluster in example (5) above was annotated as starting at where the word *dve* 'two' begins and ending at where the last SP ends.

4. Pilot fMRI Study

4.1 Design

A pilot functional neuroimaging study was conducted to reveal possible neural correlates of perceived isolated speech disfluencies and their clusters in the listener's brain. An event-related fMRI design was used, and individual instances of speech disfluencies were treated as events with fluent speech used as a baseline condition (cf. similar design used by Eklund and Ingvar, 2016). Unlike Eklund and Ingvar study, we used a visual channel to introduce the speaker to the participants. We also used a referential task instead of an instruction to listen to the discourse as if it were addressed to the study participants.

4.2 Participants

Fourteen volunteers gave a written informed consent to participate in the study (6 males, 8 females; mean age 25±4 years; native language is Russian). They reported no contraindications to MRI scanning, no history of neurological or mental disease, and no hearing problems. All participants were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). No participant of the present experiment took part in RUPEX recording or was familiar with this corpus.

4.3 Materials

To select the stimulus materials, we searched for fragments containing annotated disfluencies separated by at least 2000 ms from one another (the lowest interstimulus interval that allows to distinguish the BOLD signal evoked by different events; see Buckner, 1998). The fragments were required to be semantically coherent, last for at least 10 seconds, and contain at least two instances of disfluencies. We used the ELAN software (see Wittenburg et al., 2006; https://tla.mpi.nl/tools/tla-tools/elan/) to facilitate the search procedure. The basic vocal annotation and the additional disfluency annotation were imported into the eaf files used in ELAN. Specifically, time boundaries and type code of every annotated disfluency were imported into the intervals of the "N/C/R-Dislf" tiers (where N, C or R stands for Narrator, Commentator, and Reteller, respectively). Next, the auxiliary tiers "N/C/R-NoDisfl" were created and filled using the in-built "Create annotation from gaps" function, with the values corresponding to the duration of the gaps in ms. For instance, the disfluency cluster in the example (5) above appears after an 8540 ms interval after the previous disfluency of this speaker, and it is also separated with a 5348 ms gap from the next one; see Figure 2. We retrieved all the gaps of at least 2000 ms using the in-built ELAN search engine. After that, we manually selected the coherent fragments that satisfied the initial requirements. (The screenshot on Figure 2 represents a part of one such fragment). Overall, we obtained 32 fragments, with a total duration of 1030 s. The fragments contain 154 disfluencies, their distribution across categories discussed in Section 3 is shown in Table 2.

The fragments were combined in two lists of comparable duration. Fragments contained from two to nine disfluency instances (Md = 4.5) and varied in duration from 12 to 84 seconds. The fragments were produced by eight out of nine speakers who provided the material for the subcorpus, and seven of them appeared on each list. The audio recording of each fragment was accompanied by a static screenshot introducing the speaker to the participants in order to facilitate compatibility with future research implementing both speech disfluencies and manual gesture stimuli. Fragments were separated by 2-second silent black screens.

Isolated disfluencies (with or without SPs)		**Disfluency clusters**	
Type (code)	**Occur.**	**Type (code)**	**Occur.**
Silent pauses (00_S)	7	Filled pause + lengthening (12_F+L)	18
Filled pauses (01_F)	46	Filled pause + self-repairs (13_F+B)	9
Lengthening (02_L)	17	Filled pause + lengthening + self-repair (123_F+L+B)	8
Self-repairs, or Breaks (03_B)	25	Lengthening + self-repair (23_L+B)	3
Other	4	Other	17
TOTAL	***99***	***TOTAL***	***55***

Table 2. Types of disfluencies in the fragments selected for the event-related fMRI study.

Besides 32 main fragments with speech disfluencies, 9 extra fragments were identified in the subcorpus using the following criteria: containing the pronoun 'he' in any form (*on*, *emu*, *ego*, etc. in Russian); no speech disfluencies occurring for 4 seconds before the pronoun; and no disfluencies for 6 seconds after the pronoun has been uttered. Among such fragments, eight overlapped with the main materials and were used for the referential task construction; the ninth was used for a practice video to instruct the participants before the scanning. When the speaker uttered the pronoun in the selected fragments, a red frame appeared on the screen, and the participant was asked to complete the referential task trial (to indicate to which movie character the pronoun referred). Four trials per list were administered with the intertrial intervals unpredictable to participants. To minimize the effect of the task on the BOLD signal evoked by the disfluencies, one fragment per list was presented twice – as the very first and the very last item on the list; the referential task was administered only during the second repetition, and the imaging data from this second presentation of the fragment was not analyzed. In total, the video for each session lasted for about ten minutes.

4.4 Procedure

The participants were briefed on MRI safety, given the task instructions and the practice video, and then proceeded to the scanner. The instructions were focused on the referential task and never mentioned speech disfluencies.

The protection from the noise of the scanner was accomplished with the headphones, and foam padding was used to restrict head motions. The stimuli sound level was tested with every participant individually to make sure that they were able to distinctly hear the speech against the

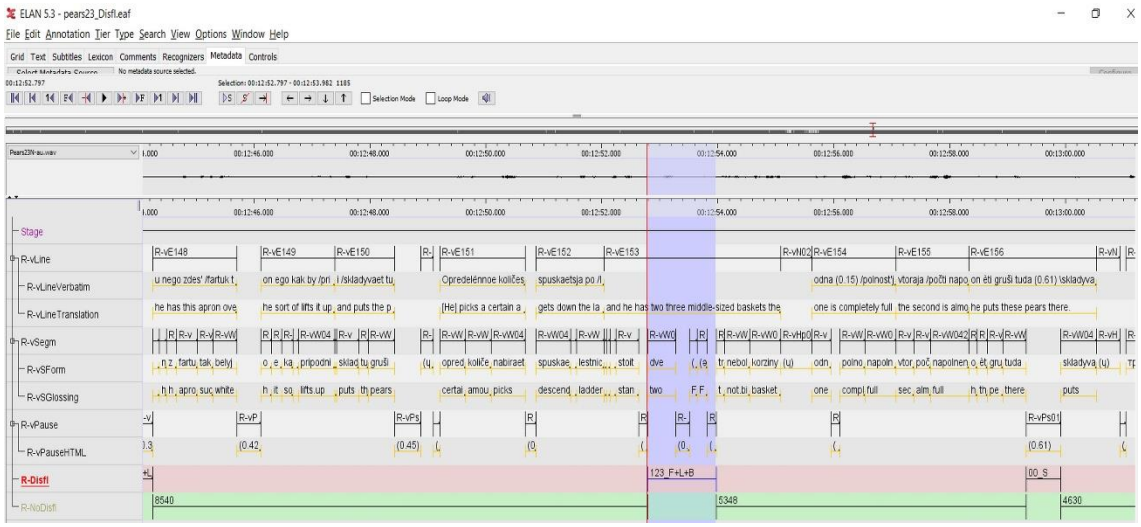

Figure 2. ELAN representation of the disfluency cluster in example (5); values in the green intervals indicate gaps between disfluencies, ms.

scanner noise and feel comfortable. During the structural imaging, the participants were shown the original "Pear Film" video (see Section 2). After that, they completed two sessions of the functional imaging while performing the referential task, one session per stimulus list. While listening to the audio recordings, participants waited for the appearance of the red frame on the screen, and then responded whether the pronoun 'he', which they heard at the red frame moment, referenced the boy who had stolen the pears or another person. The motor responses were made by pressing one of the two buttons. Between the two sessions, a short break was given for the participants to rest and for the fieldmap acquisition. After the task, the scanning procedure was continued by the collection of other data not falling within the scope of this pilot study.

4.5 Equipment and Imaging Parameters

The images were acquired with the Philips 3T Ingenia scanner at the Mental Health Research Center (Moscow, Russia). The standard 15-channel dS head coil was used. The stimuli were administered and the participants' responses were recorded with the InVivo presentation equipment for MRI environment, custom communication system and Cedrus Lumina synchronization box and response pads. The participants watched the video shown on a display located next to the magnet bore through a mirror attached to the head coil, and listened to the speech through the headphones. The stimulus presentation was controlled by the VLC player for MacOS, and the responses were recorded by the in-house python script based on the PyHID library.

4.6 Neuroimaging Data Analysis

Data analysis was performed with SPM12 (Wellcome Institute of Cognitive Neurology), FSL5.0.9 (FMRIB, Oxford, UK). The preprocessing included the following stages: slice timing correction, combined head motion and magnetic field inhomogeneity artifact correction on the basis of the fieldmap (SPM Realign & Unwarp), spatial coregistration of the functional and structural images, segmentation of the structural images into tissue volumes, spatial normalization of both functional and structural images to the Montreal Neurological Institute (MNI) space, spatial smoothing of the functional images with FWHM 8×8×8 mm. The fieldmaps were preprocessed with the FSL topup procedure, with the voxel displacement maps calculated in the SPM.

The general linear model for each participant's voxelwise data was constructed in SPM using the convolution of the predictors with the canonical hemodynamic response function (HRF). Six head motion parameters were entered into the model as regressors. The following 12 conditions (types of events) were modeled: isolated silent pauses (00_S), filled pauses (01_F), lengthening (02_L), breaks (03_B), and other disfluencies; clusters: filled pause + lengthening (12_F+L), filled pause + self-repairs (13_F+B), lengthening + self-repair (23_L+B), filled pause + lengthening + self-repair (123_F+L+B)), other clusters; fragment borders; red frames; motor responses. Although SPM treats the duration of short events (below 2 seconds) as zero, we entered the actual onsets and durations of all events for clarity. Motor responses were ascribed the duration of 0.5 second. Due to the small number of events in 23_L+B and isolated other disfluency types (see Table 1), these two predictors were used only to regress out the corresponding activation and were not further analyzed. Video fragment borders, red frames and motor responses were introduced for the same purpose. For the remaining eight types of disfluencies, one-sided T-test contrasts were computed and further entered into the second-level random-effects model. Two additional contrasts (00_S > 01_F and 01_F>00_S) were assessed to facilitate comparison with literature. Due to the pilot nature of the study, the results were visualized at a considerably liberal statistical threshold: uncorrected $p < 0.005$, cluster extent

of more than 10 voxels. All coordinates are reported in MNI space, and the anatomical areas were attributed according to the Neuromorphometrics atlas in SPM. T2*-weighted functional images were collected using the FFE EPI pulse sequence with the following parameters: TR/TE/FA = 2500ms/35ms/80°; 3.2 mm isotropic voxels; 80×80×34 acquisition matrix; SENSE with acceleration factor of 1.8. The slices were oriented parallel to AC/PC plane and acquired in the ascending order. Per participant, 267 and 271 volumes were acquired and 243 and 247 volumes used for the analysis from the first and the second session respectively; 3 extra volumes were scanned at the beginning of each session and then discarded by the scanner console software in order to reach magnetic equilibrium before acquiring the actual data. The functional data were complemented by the T1-weighted structural images (TFE sequence, TR/TE/FA = 8ms/4ms/8°, 170 near-transverse slices, 1-mm isotropic voxels) and two SE EPI sequences with opposite directions of the phase encoding (AP vs. PA) and with the same slice prescription and voxel resolution as the functional images; these volumes were further used for the calculation of a fieldmap.

4.7 Results and Discussion

No participants were excluded from the analysis due to excessive head motion, imaging artifacts or inability to perform the task. Although participants reported the referential task to be a difficult one, and three people performed at chance level (0.5 accuracy), everyone provided timely responses which ensures that all participants were attending to the stimuli all the time.

Two types of isolated speech disfluency (Fig. 3) and three types of disfluency clusters (Fig. 4) have demonstrated some characteristic neural correlates in the listener's brain. Although these results should be treated with caution due to the pilot nature of the study and the lenient statistical thresholds used, we believe that they may be important for future research.

Activation elicited by listening to the isolated silent pauses compared to the fluent speech baseline was revealed in the right occipital fusiform gyrus (peak coordinates: {33 −76 −13}), next to the left insula ({−36 −1 −13}) and in the right postcentral gyrus ({48 −19 44}). Breaks were characterized by extra activation in the inferior occipital gyri bilaterally ({36 −88 −1} and {−45 −82 −4}), next to the right central operculum ({−9 −43 26}), next to the left planum polare ({−48 −16 −1}), and next to the posterior cingulate gyrus bilaterally ({−9 −43 26}, {24 −46 17}).

Clusters that combined filled pauses and lengthenings evoked activation in the visual areas next to occipital pole bilaterally ({−21 −97 1}, {27 −94 −1}), in the superior temporal gyrus bilaterally ({−54 −16 −1}, {57 −10 −4}), and in the left temporal pole ({−39 8 −25}).

Surprisingly, when the break was added to a disfluency cluster (F+L) to produce a (F+L+B) cluster, less activation was observed. Nevertheless the activation was found within much the same bilateral temporal regions covering areas around superior temporal gyri and planum temporale ({−42 −34 8}, {54 10 −4}), but it was not accompanied by visual cortex activation.

Perception of the other disfluency clusters involved predominantly right planum temporale ({63 −25 14}) and bilateral insula ({39, 5, 5}, {−39, −22, 2}).

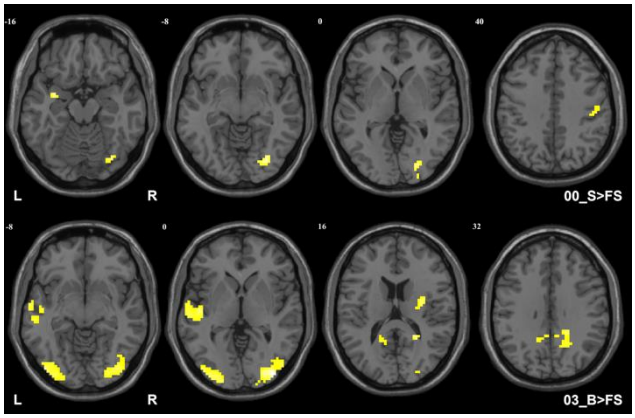

Figure 3. Activation maps obtained from one-sided t-contrasts for perception of the isolated disfluency instances (compared to the fluent speech baseline). 00_S — silent pauses, 03_B — breaks, FS — fluent speech. Z coordinates are given in the MNI space. Statistical threshold: voxelwise $p < 0.005$ uncorrected, cluster extent > 10 voxels. Maps are overlaid on a standard individual brain structural image (MNI space).

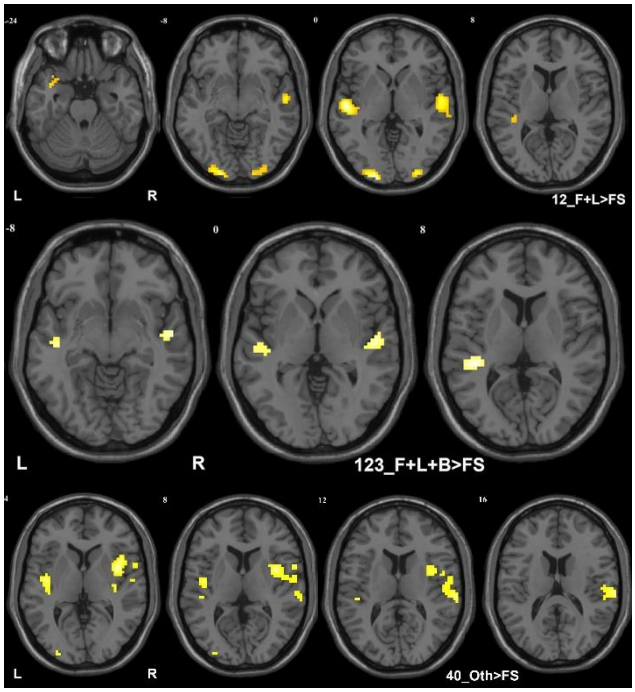

Figure 4. Activation maps obtained from one-sided t-contrasts for perception of the disfluency clusters (compared to the fluent speech baseline). 12_F+L — filled pauses and lengthening,123_F+L+B — filled pauses, lengthening, and breaks, 40_Oth — disfluency clusters other than combinations of filled pauses, lengthening and breaks, FS — fluent speech. Z coordinates are reported in the MNI space. Statistical threshold: voxelwise $p < 0.005$ uncorrected, cluster extent > 10 voxels. Maps are overlaid on a standard individual brain structural image (MNI space).

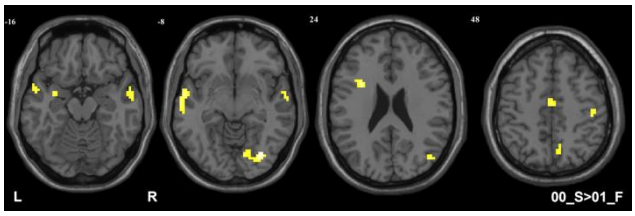

Figure 5. Activation maps obtained from one-sided t-contrast for perception of silent vs. filled pauses. 00_S — silent pauses, 01_F — filled pauses. Z coordinates are reported in the MNI space. Statistical threshold: voxelwise $p < 0.005$ uncorrected, cluster extent > 10 voxels. Maps are overlaid on a standard individual brain structural image (MNI space).

To summarize, when compared to the fluent speech, perception of speech disfluencies mainly involved extra activation of extrastriate visual cortex as well as auditory and speech-related temporal cortices, and insula. Activation of visual cortices was not observed in Eklund and Ingvar study (2016), and this difference seems to be likely to the difference in paradigms used, since our paradigm included some visual information on the speaker, although very limited, while Eklund and Ingvar used purely auditory paradigm. Eklund and Ingvar attributed the increased activation in the auditory cortex they observed for the silent and filled pauses to increasing auditory attention to the speech stimuli when the listener faces uncertainty; in our paradigm, the same logic may apply with the only difference that participants might started paying greater visual attention to the photograph of the speaker.

Although we observed greater activation in the temporal cortices for various types of disfluencies, we have obtained neither extra activation in the primary auditory cortex for the pauses, nor activation in the supplementary motor area for the filled pauses specifically. Therefore, we haven't replicated these two findings by Eklund and Ingvar (2016). Moreover, the comparison of the BOLD responses to the silent and filled pauses in our study have shown greater activation in the auditory and visual cortices as well as in the right precentral areas and in the middle cingulate cortex for the silent pauses compared to filled pauses (see Fig. 3) which is the opposite of the Eklund and Ingvar results. Possible explanation for such differences may lie in a different definition of the silent pauses we used. While we presented only few isolated silent pauses to our participants, the disfluency instances were very pronounced, which might have reversed the effect. The task at hand (listening to the discourse as if being one of the dialog participants vs. referential task) might also constitute an important factor to be tested in the future research. It is also interesting to compare the neural correlates of disfluency production and comprehension since their potential similarity may be used as an indirect argument for the involvement of the simulation mechanisms into speech disfluency perception. Since neuroimaging research of speech production is even more limited than that of speech comprehension, we were only able to compare our results with those by Kircher et al. (2004) who have shown the left temporoparietal junction to be involved in production of within-clause speech pauses (Kircher et al., 2004) versus continuous speech. In our data, no activation was found in this region in any of the contrasts described above.

5. Conclusion

Bridging corpus linguistics and neuroimaging methodology may open new perspectives in language research. The present paper demonstrates how such an approach may be implemented in the field of natural communication studies. A neuroimaging study of such a subtle phenomenon as the perception of speech disfluencies by the listener was made possible due to the RUPEX multichannel corpus amplified by search techniques that helped to select the natural speech stimulus materials satisfying the requirements of the fMRI study design.

A pilot functional neuroimaging study was conducted to reveal possible neural correlates of perceived isolated speech disfluencies and their clusters in the listener's brain. An event-related fMRI design was used, and individual instances of speech disfluencies were treated as events with fluent speech used as a baseline condition. Compared to the only previous study by Eklund and Ingvar (2016), we used the visual channel to introduce the speaker to the participants, and included several extra types of disfluencies not previously used in neuroimaging experiments.

We plan to conduct further research with a study on a bigger sample to analyze speech disfluencies phenomena at the neurophysiological level more accurately. Also, we are going to use vocal and kinetic channels from RUPEX in order to investigate speech-accompanying gestures, specifically in their interaction with speech disfluencies, and to analyze neural correlations of such interaction. Inasmuch as this domain is quite new, we have a lot to discover on the basis of developed methodology.

Acknowledgements

This study is supported by Russian Foundation for Basic Research (complex grant #18-00-01598 with subgrants #18-00-01485 and #18-00-01592). We thank Alena Rumshiskaya, M.D., Darya Bazhenova, M.D., Liudmila Makovskaya, M.D., and Liudmila Litvinova for their assistance with the organization of the study and data collection.

Bibliographical References

Adolphs, S. and Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. N.-Y.: Routledge.

Barr, D. and Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, 25(4), pp. 441–455.

Boersma, P. and Weenink, D. (2012). Praat: Doing phonetics by computer [Computer Program]. Version 5.3.04. Online: http://www.praat.org/ (Accessed on November 4, 2019).

Bouraoui, J.-L. and Vigouroux, N. (2005). Disfluency phenomena in an apprenticeship corpus. *The 4th Workshop on Disfluency in Spontaneous Speech*, pp. 33–37, Aix-en-Provence, France.

Buckner, R.L. (1998). Event-related fMRI and the hemodynamic response. *Human Brain Mapping*, 6(5–6), pp. 373–377.

Chafe, W. (Ed.) (1980). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood: Ablex.

Chafe, W. (1994). *Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing*. Chicago.

Church, R. B., Alibali, M. W., and Kelly, S. D. (Eds.) (2017). *Why gesture? How the hands function in speaking, thinking and communicating*. Benjamins.

Clark, H. H. and Clark, E. V. (1977). *Psychology and Language: An Introduction to Psycholinguistics*. New York: Harcourt Brace Jovanovich.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84, pp. 73–111.

Clark, H. H. and Wilkes-Gibbs, D. (1986) Referring as a collaborative process. *Cognition*, 22, pp. 1–39.

Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass* 2(4), pp. 589–602.

Crible, L., Degand, L., and Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast*, 17(1), pp. 69–95.

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *Proceedings of DiSS '01 Disfluency in Spontaneous Speech, ISCA Tutorial and Research Workshop*, pp. 33–37, August 29–31 2001, University of Edinburgh, England.

Eklund, R. (2004). Disfluency in Swedish human–human and human–machine travel booking dialogues. *PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science*, Linköping University, Sweden.

Eklund, R. and Ingvar, M. (2016). Supplementary Motor Area Activation in Disfluency Perception. An fMRI Study of Listener Neural Responses to Spontaneously Produced Unfilled and Filled Pauses. In *Proceedings of Interspeech 2016*, pp. 1378–1381, September 8-12 2016, San Francisco, USA.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory and Cognition*, 29(2), pp. 320–236.

Fox, B., Wouk, F., Hayashi, M., Fincke S., Tao, L., Sorjonen, M.-L., Laakso, M., and Flores Hernandez, W. (2009). A cross-linguistic investigation of the site of initiation in same-turn self-repair. In Jack Sidnell (ed.), *Conversation Analysis: Comparative Perspectives* (pp. 60–103). Cambridge: Cambridge University Press.

Goldin-Meadow, S. (2014). Widening the lens: What the manual modality reveals about language, learning, and cognition. *Philosophical Transactions of the Royal society*, 369 (1651).

Kibrik, A. A. (2010). Multimodal linguistics. In Yu.I. Aleksandrov, V.D. Solov'jev (Eds.) *Cognitive studies*, IV. Moscow: Institute of psychology.

Kibrik, A. A. (2018). Russian multichannel discourse. Part I. Setting up the problem. *Psixologičeskij žurnal*, 39(1), pp. 70–80.

Kibrik, A. A. and Podlesskaya, V. I. (Eds.) (2009). *Night Dream Stories: A corpus study of spoken Russian discourse*. Moscow: Jazyki slavjanskix kul'tur.

Kibrik, A. A. and Fedorova, O. V. (2018a). Language production and comprehension in face-to-face multichannel communication. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"* (pp. 305–316). Moscow: RGGU.

Kibrik, A. A. and Fedorova, O. V. (2018b). A "Portrait" approach to multichannel discourse. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of LREC 2018, Fifth International Conference on Language Resources and Evaluation*, pp. 1908–1912, Miyazaki, Japan 7-12 May 2018. European Language Resources Association (ELRA).

Kibrik, A. A. and Fedorova, O. V. (2018c). An empirical study of multichannel communication: Russian Pear Chats and Stories. *Psixologija. Žurnal Vysšej Školy ėkonomiki*, 15(2), pp. 191–200.

Kibrik, A. A., Korotaev, N. A., and Podlesskaya, V. I. (2020, in print). Russian spoken discourse: Local structure and prosody. In: Shlomo Izre'el, Heliana Mello, Alessandro Panunzi, and Tommaso Raso (Eds.), *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam: John Benjamins.

Kircher, T., Brammer, M., Levelt, W., Bartels, M., Mcguire, P. (2004). Pausing for thought: Engagement of left temporal cortex during pauses in speech. *NeuroImage*, 21, pp. 84–90.

Levelt, W. J. M. (1993). *Speaking: From Intention to Articulation*. The MIT Press.

Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1), pp. 71–89.

Maclay, H. and Osgood, C.E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word*, 5, pp. 19–44.

McNeill, D. (2005). *Gesture and thought.* Chicago.

Müller, C., Fricke, E., Cienki, A., and McNeill, D. (Eds.) (2014). *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction*. Berlin: Mouton de Gruyter.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9, pp. 97–113.

Podlesskaya, V. I. (2010). Parameters for typological variation of placeholders. In N. Amiridze, B.H. Davis, M. Maclagan (Eds.), *Fillers, Pauses and Placeholders. (Typological Studies in language (TSL)*, vol. 93, pp. 11–32). Amsterdam/Philadelphia: John Benjamins.

Podlesskaya, V. I. (2015). A corpus-based study of self-repairs in Russian spoken monologues. *Russian Linguistics*, 39(1), pp. 63–79.

Podlesskaya, V. I., Korotaev, N. A., and Mazurina, S. I. (2019). A corpus study of self-repairs in Russian monologues and dialogues. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"* (pp. 547–561). Moscow: RGGU.

Schegloff, E. A. (2013). Ten operations in self-initiated, same-turn repair. In Makoto Hayashi, Geoffrey Raymond, Jack Sidnell (Eds.), *Conversational repair and human understanding* (pp. 41–70). Cambridge: Cambridge University Press.

Shriberg, E. E. (1994). Preliminaries to a Theory of Speech Disfluencies. *PhD thesis*, University of California, Berkeley, USA.

Trouvain, J., Fauth, C., and Möbius, B. (2016). Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. In *Proceedings of Speech Prosody (SP8)*, pp. 31–35, May – 3 June 2016, Boston.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, pp. 1556–1559, Genoa, Italy, May 22-28, 2006. European Language Resources Association (ELRA).

Wouk, F. (2005). The syntax of repair in Indonesian. *Discourse Studies*, 7(2), pp. 237–258.