A new fMRI localizer for preoperative language mapping using a sentence completion task: Validity, choice of baseline condition, and test-retest reliability

Kirill Elin[1†], Svetlana Malyutina[1†*], Oleg Bronov[2], Ekaterina Stupina[1], Aleksei Marinets[2], Olga Dragoy[1,3]

[1]Center for Language and Brain, HSE University, Moscow, Russia

[2]Department of Radiology, National Medical and Surgical Center Named after N. I. Pirogov, Moscow, Russia

[3]Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

[†] The authors contributed equally to this manuscript.

*** Correspondence:**
Svetlana Malyutina
smalyutina@hse.ru

Keywords: functional magnetic resonance imaging[1], presurgical mapping[2], language mapping[3], language localizer paradigm[4], test-retest reliability of fMRI[5]

Abstract

To avoid post-neurosurgical language deficits, intraoperative mapping of the language function in the brain can be complemented with preoperative mapping with fMRI. The validity of an fMRI 'language localizer' paradigm crucially depends on the choice of an optimal language task and baseline condition. This study presents a new fMRI 'language localizer' in Russian using overt sentence completion, a task that comprehensively engages the language function by involving both production and comprehension at the word and sentence level. The paradigm was validated in 18 neurologically healthy volunteers who participated in two scanning sessions, for estimating test-retest reliability. For the first time, two baseline conditions for the sentence completion task were compared.

At the group level, the paradigm significantly activated both anterior and posterior language-related regions. Individual-level analysis showed that activation was elicited most consistently in the inferior frontal regions, followed by posterior temporal regions and the angular gyrus. Test-retest reliability of activation location, as measured by Dice coefficients, was moderate and thus comparable to previous studies. Test-retest reliability was higher in the frontal than temporo-parietal region and with the most liberal statistical thresholding compared to two more conservative thresholding methods. Lateralization indices were expectedly left-hemispheric, with greater lateralization in the frontal than temporo-parietal region, and showed moderate test-retest reliability. Finally, the pseudoword baseline elicited more extensive and more reliable activation, although the syllable baseline appears more feasible for future clinical use.

Overall, the study demonstrated the validity and reliability of the sentence completion task for mapping the language function in the brain. The paradigm needs further validation in a clinical

sample of neurosurgical patients. Additionally, the study contributes to general evidence on test-retest reliability of fMRI.

1 Introduction

1.1 Language mapping in neurosurgical patients

When patients undergo neurosurgical interventions for brain tumors, refractory epilepsy, arteriovenous malformations et cetera, a crucial goal is to remove pathological tissue while sparing eloquent (functionally necessary) areas, so that the respective functions, including cognitive ones (Satoer et al., 2016), are not impaired following neurosurgery (Duffau, 2012). One critical function is language processing: it lies at the core of human communication, and its impairment negatively impacts return to work, social inclusion, and general quality of life (Gabel et al., 2019, Hilari et al., 2003).

Localization of the language network is highly variable across individuals (Ojemann, 1979), and the variability is further enhanced by functional re-organization that happens in case of brain pathology: for example, over the course of brain tumor growth (Almairac et al., 2018, Zhang et al., 2018). Thus, to avoid damage to brain areas critical for the language function and to prevent subsequent language impairment, the neurosurgical team performs mapping of 'language-eloquent' brain areas in individual patients. The gold standard for localizing language-eloquent brain areas is intraoperative mapping with direct electrical stimulation (DES) during awake craniotomy (Ojemann, 1979, Rofes et al., 2019). During this procedure, an electric current is applied to exposed brain tissue, causing a temporary disruption of neural activity. Meanwhile, the patient is awake from anaesthesia and is performing a language task. If application of DES to an area reliably leads to errors or speech arrest, this means that the area is eloquent and should be spared during neurosurgery, if possible.

While intraoperative DES is the standard procedure for language mapping, there are reasons to complement it with additional preoperative mapping. Firstly, preoperative language mapping allows to plan the surgical procedure in advance (Silva et al., 2018, Weng et al., 2018). Based on preoperative mapping data, the neurosurgeon can decide whether intraoperative DES mapping of the language function is necessary (for example, when operating over a presumably non-language-dominant hemisphere) or plan an optimal access route to bypass language-eloquent areas. Secondly, data from preoperative mapping can be used if DES cannot be completed: for example, if the patient does not cooperate or if there are epileptic seizures during DES (Hervey-Jumper et al., 2015). In such cases, the results of preoperative mapping, even though providing less direct information than DES, can still inform the neurosurgeon and reduce the risks of functional impairment.

Historically, preoperative language mapping was performed using the intracarotid sodium amobarbital procedure, known as the Wada test (Wada & Rasmussen, 1960). The test involves anesthetizing one hemisphere by injecting sodium amobarbital through a catheter while the patient is performing a language task. Failure to perform the task indicates that the anaesthesized hemisphere is crucial for language processing. However, the Wada test has major limitations. It is a highly invasive procedure that can cause complications: according to Loddenkemper, Morris and Möddel (2008), they emerge in up to 11% patients and include serious adverse events such as stroke. Another limitation is that effects of sodium amobarbital only last for a few minutes, providing very limited time for testing (Loring, Meador, & Lee, 2002). Critically, the Wada test can only assess hemispheric dominance of language processing but cannot address specific localization of language-eloquent areas within a hemisphere.

Therefore, other technologies have been replacing the Wada test for preoperative language mapping: navigated transcranial magnetic stimulation (nTMS; Picht et al., 2013), functional magnetic resonance imaging (fMRI; for review, see Agarwal et al., 2019, Silva et al., 2018), magnetoencephalography (MEG; Van Poppel et al., 2012), or combination thereof (Ille et al., 2015, Sollman et al., 2016). Unlike the Wada test, these methods are largely safe and non-invasive. On top of that, they have high spatial resolution and allow to identify language-eloquent brain areas with the precision of millimeters. To take full advantage of these methods, it is crucial to choose an optimal functional paradigm that would be sensitive, specific and reliable in identifying both lateralization (hemispheric dominance) and specific localization of brain networks comprehensively enabling the language function. In this paper, we present such paradigm for pre-operative language mapping using fMRI in Russian-speaking individuals and provide methodological evidence on its test-retest reliability and the optimal baseline condition.

1.2 Choice of a language task for fMRI mapping

The quality of a language mapping paradigm critically depends on the choice of a language task: that is, whether it is able to comprehensively engage all levels of linguistic processing while remaining feasible. Previous fMRI 'language localizer' paradigms have used a variety of tasks (for review, see Bradshaw et al., 2016, Manan et al., 2020). Below, we review most popular language tasks and summarize previous evidence on their success in identifying language networks in the brain.

Most 'language localizers' have used single-word tasks, particularly expressive single-word tasks (Manan et al., 2020). A traditional expressive single-word task adapted from early intraoperative batteries (Ojemann, 1993, Ruge et al., 1999) was number counting. However, counting is a highly automated process that engages linguistic processing only superficially and is no longer considered sufficiently sensitive to identify language networks (Morrison et al., 2016b; Petrovich Brennan et al., 2007). Today, other tasks are used to engage word retrieval, such as picture naming (Petrovich Brennan et al., 2007, Pouratian et al., 2002, Roux et al., 2003, Rutten et al., 2002) and verbal fluency, where the participant has to name as many words as possible from a given semantic category or starting with a given letter (Ruff et al., 2008 Sanjuan et al., 2010). Expressive single-word tasks are intuitive for the participant and are easily timed relative to fMRI scanning. Still, they only engage the sound and word levels of language production, leaving out any grammatical processing and any language comprehension. Thus, they cannot fully identify brain areas that are crucial for sentence-level communication. Indeed, compared to sentence-level tasks, expressive single-word tasks elicit less activation in both anterior and posterior left-hemispheric 'language regions' (Połczyńska et al., 2017, Manan et al., 2020). Additionally, picture naming shows a poor lateralizing ability (Bradshaw et al., 2017, Deblaere et al., 2002).

Similar problems are faced by receptive single-word tasks, such as phonemic judgement, requiring to make a decision about the sound structure of a word (for example, whether two words rhyme; Jones, Mahmoud & Philipps, 2011), or semantic judgement, requiring to make a decision about the meaning of a word (for example, whether it refers to an animate object, or whether two words are opposite in meaning; Binder et al., 1996, Szaflarski et al., 2008). Again, such tasks are easily integrated with the timing of fMRI scanning. Moreover, unlike expressive single-word tasks, they do not evoke head motion artifacts due to articulation. However, they are even more limited in engaging linguistic processes and cannot detect brain networks enabling language production or any grammatical processing beyond the word level. Empirically, these tasks have shown limited lateralizing ability (Deblaere et al., 2002, Jansen et al., 2006) and reliability (Jansen et al., 2006; reliability will be discussed in more detail below).

To more fully activate language networks, a number of fMRI protocols used sentence-level tasks requiring to process not only individual words but also their grammatical and semantic relations. Examples of expressive sentence-level tasks are describing a picture with a sentence (Mauler et al., 2017, Partovi et al., 2012) or generating a sentence with given words (Hakyemez et al., 2016). Such tasks appear to successfully activate both anterior and posterior language areas (Mauler et al., 2017, Partovi et al., 2012, Hakyemez et al., 2016). However, they are taxing for the patient and difficult to time relative to fMRI scanning, especially given interindividual variability in task completion speed across patients, so they are not widely used.

Much more popular are receptive sentence-level tasks. These are sentence or passage listening (Pillai & Zaca, 2011, Suarez et al., 2014, Wilson et al., 2017) or reading (Grummig et al., 2006, Fedorenko et al., 2010), which may be passive or accompanied by comprehension questions, such as to judge real-world plausibility of a sentence or match it to a picture (Kinno et al., 2014, Pillai & Zaca, 2011). These tasks are more easily timed than expressive sentence-level tasks but also successfully engage grammatical processing: that is, the participant has both to process individual words and analyse their relations. Passive receptive sentence-level tasks have an additional advantage of feasibility in patients with compromised language production or non-cooperative patients (for example, in the pediatric population, Suarez et al., 2014) but also an additional drawback: it is impossible to control or even monitor how much the participant is engaged in the task. Crucially, a major limitation is that none of receptive sentence-level tasks engage brain networks crucial for language production. Empirically, these tasks have shown low lateralizing abilities (Lehéricy, 2000; Pillai & Zaca, 2011).

Taken together, in order to comprehensively identify brain networks that enable real-life language use, an fMRI paradigm needs to engage both language production and comprehension in a task that goes beyond the word level. One solution is a conjunction analysis of multiple tasks targeting different language processes separately (Pouratian et al., 2002, De Guibert et al., 2010). However, interpretation of the conjunction analysis is not straightforward if tasks elicit largely different activations. Additionally, from the clinical viewpoint, multiple-task paradigms are time-consuming and less feasible in clinical settings. Thus, another solution is an fMRI localizer paradigm that uses a single task engaging both language production and comprehension beyond the word level.

One such comprehensive task is sentence completion, advocated in a recent white paper of the American Society of Functional Neuroradiology (Black et al., 2017) and a metaanalysis by Manan et al. (2020). In this task, the participant has to read aloud a sentence with a missing final word and complete it with a semantically and grammatically appropriate word. This task comprehensively involves many linguistic processes in both comprehension (orthographic processing, word access, grammatical parsing, semantic integration) and production (word search, grammatical inflection in morphologically complex languages such as Russian, phonological encoding and articulation). Empirically, previous works proved sentence completion superior to other tasks in assessing both lateralization and localization of language processing networks (Salek et al., 2017, Połczyńska et al., 2017, Zacà et al., 2012, Barnett et al., 2014, Wilson et al., 2017, Unadkat et al., 2019).

Inspired by these sentence completion paradigms in English, the present study presents a similar paradigm in Russian. Russian is the 8th most spoken language in the world, with about 120 million first-language speakers worldwide (Eberhard, Simons & Fennig, 2020), so a new clinical tool in Russian would serve the needs of a large Russian-speaking clinical population. So far, Russian-language paradigms for presurgical language mapping have been very few and have never used a sentence completion task (Litvinova et al., 2012, Rumshiskaya et al., 2014).

1.3 Choice of a baseline task for fMRI mapping

A crucial concept in classic fMRI analysis is 'subtraction logic': to isolate neural activation related to the process of interest, the analysis should 'subtract' the activation in a 'lower-level' baseline (control) task from the activation in a 'higher-level' experimental task (Huettel et al., 2008). Specifically in case of 'language localizer' paradigms, such subtraction allows to isolate language-related neural activity from activity due to sensorimotor processes, general alertness, et cetera. Due to the subtraction principle, not only the choice of an experimental language task but also the choice of a lower-level baseline task can vastly impact the findings of fMRI 'language localizers' (Bradshaw et al., 2017).

Some previous fMRI 'language localizers' used passive rest or viewing of a fixation cross as the baseline condition (Jones et al., 2011, Suarez et al., 2014). However, a passive baseline is problematic for several reasons. Firstly, a passive baseline does not require any sensorimotor or cognitive activity, so subtracting it from the experimental condition does not fully isolate language-related activity from lower-level processes. Secondly, it is not possible to control or monitor the patient's cognitive activity during passive rest, so mind-wandering or other patient-initiated cognitive activity may confound the results. Indeed, fMRI analyses using passive baselines have elicited less specific and less lateralized activation compared to analyses using active baselines (Dodoo-Schittko et al., 2012, Hund-Georgiadis et al., 2001, although see Miró et al., 2014, Newman et al., 2001).

Therefore, active baselines are typically recommended for more specific isolation of language-related activity from lower-level processes (Bradshaw et al., 2017). Choosing an optimal active baseline presents a challenge: for most language tasks, the respective lower-level processes can be addressed by several theoretically possible baselines. For example, for listening tasks, the baseline condition can involve listening to backwards speech (Lehéricy et al., 2000, Thivard et al., 2005), various types of noise (Rodd et al., 2005), or music (Bleich-Cohen et al., 2009). Although all these baselines involve auditory processing (lower-level sensory processing) and presumably no linguistic processing, an empirical comparison by Stoppelman et al. (2013) showed that different baselines yielded very different results. So far, such direct empirical comparisons in order to choose an optimal baseline have only been made for few language tasks and baseline types (Binder et al., 2008, Newman et al., 2001, Stoppelman et al., 2003).

To the best of our knowledge, our study is the first to make an empirical comparison between two different baselines for the sentence completion task. These are a syllable baseline, where the participant has to read aloud a sequence consisting of the same syllable and repeat the syllable once more, and a pseudoword baseline, where the participant has to read aloud a sequence of pseudowords and repeat any of them once. Both baselines are theoretically plausible. In contrast to the experimental condition, they do not consist of real words or resemble grammatical structures, so they do not elicit any linguistic processing. At the same time, they involve the same 'lower-level' processes as the experimental condition: visual and orthographic processing, motor planning and articulation, and initiation of a response (completion of a sequence). Thus, their subtraction allows to maximally isolate language-related from 'lower-level' neural activity. Previous sentence completion paradigms used other baselines that subtracted 'lower-level' activity less fully (rest: Połczyńska et al., 2017, passive viewing of nonsense symbols: Barnett et al., 2014, Zacà et al., 2012) or a conjunction analysis approach without an explicit baseline (Wilson et al., 2017). The present study is the first to employ and compare a syllable and pseudoword baseline for the sentence completion task.

1.4 Reliability of fMRI mapping

Another methodological contribution of the present study is estimating test-retest reliability of the fMRI paradigm. Reliability is critical for any clinical usage of fMRI 'language localizers': distribution of brain activity needs to be reproducible at multiple testing sessions in order to consider it clinically meaningful and draw any implications for neurosurgical treatment. Recent studies have raised concerns about test-retest reliability of task-based fMRI in general, due to inherent physiological noise, scanner noise, changes in concurrent non-task-related cognitive activity in participants, et cetera (Bennett & Miller, 2010, Elliott et al., 2020, Holiga et al., 2018). In light of these general concerns, it is important to quantify and report reliability of any paradigms suggested for clinical use.

Previous studies have started estimating test-retest reliability of fMRI 'language localizer' paradigms in healthy control participants (Fesl et al., 2010, Morrison et al., 2016a, Nettekoven et al., 2018, Wilson et al., 2016) and clinical populations (Fernández et al., 2003, Morrison et al., 2016a). For example, Morrison et al. (2016a) showed high individual variability in test-retest reliability, which was on average lower in a phonemic fluency task than in a rhyming task, and in patients with high-grade gliomas than patients with low-grade gliomas and healthy control participants. Using an overt object naming task in healthy participants, Nettekoven et al. (2018) showed high reliability of the activation peak location but low reliability of activation extent, particularly in the right hemisphere.

To the best of our knowledge, only two studies so far have estimated test-retest reliability of sentence completion paradigms. Whalley et al. (2009) used the Hayling sentence completion task in individuals with high genetic risk of schizophrenia and showed good test-retest reliability. Wilson et al. (2016) compared test-retest reliability of four language tasks in healthy participants and found the best reliability in picture naming, followed by naturalistic comprehension, sentence completion, and narrative comprehension. Despite this pattern, the authors concluded that sentence completion was one of two tasks offering the best balance of reliability and validity for an fMRI language localizer. Our study aims to add to these emerging data and provide more evidence on test-retest reliability of a sentence completion fMRI paradigm.

Previous studies used different metrics to quantify test-retest reliability: Dice coefficient (Fesl et al., 2010, Morrison et al., 2016a, Nettekoven et al., 2018, Wilson et al., 2016), Jaccard index (Morrison et al., 2016a), Euclidean distance (Morrison et al., 2016a, Nettekoven et al., 2018), voxelwise intraclass correlation coefficient (Fernández et al., 2003, Nettekoven et al., 2018, Whalley et al., 2009), correlation of lateralization indices (LIs; Morrison et al., 2016a, Fesl et al., 2010). The present study adopted two of them: between-session Dice coefficient and correlation of lateralization indices. Advantages of the Dice coefficient are that it is widely used in the literature, is straightforward to interpret and, unlike intraclass correlation coefficient, can provide a global whole-brain measure and is calculated individually with no reference to group data (Bennett & Miller, 2010, Wilson et al., 2016). Besides, using a Dice coefficient ensures comparability to Wilson et al. (2016), the only previous study that included a sentence completion task and compared its reliability to other tasks. Additionally, we measured the test-retest correlation of LIs because, just as individual activation maps, they also present a clinically relevant measure that can inform a neurosurgeon's decision on the necessity of awake surgery.

1.5 The present study

To summarize, this paper presents a new fMRI localizer paradigm for preoperative language mapping in Russian-speaking individuals with brain tumors, refractory epilepsy, and other conditions

when neurosurgery is indicated. Following the best practices in other languages, we used a sentence completion task that comprehensively engages language production and comprehension processes at the word and sentence level. We present the data from a control group of neurologically healthy individuals, test whether the paradigm can successfully identify the expected key language-related areas in this group, compare two different baseline conditions (syllables versus pseudowords), and quantify the test-retest reliability of the paradigm.

2 Method

2.1 Participants

The study included 21 right-handed native speakers of Russian with no history of neurological or psychiatric disorders. Data of three participants were excluded from analysis due to excessive head movement in the scanner (more than 5 mm), resulting in a sample of 18 participants (14 females; age: mean 41.3, SD 6.6, range 30–53 years; years of education: mean 16.2, SD 4.7, range 11–30 years; Edinburgh Handedness Inventory score: mean 52, SD 2.67, range 46–55). All participants had normal hearing and normal or corrected-to-normal vision. All participants gave written informed consent.

2.2 Task and Stimuli

During each scanning session, participants performed two identically structured language mapping paradigms. Both mapping paradigms comprised an experimental condition and a baseline condition. The experimental condition was identical in the two paradigms: participants were visually presented with a Russian sentence with a missing final word and instructed to read the sentence aloud and produce an appropriate final word aloud. The two paradigms differed with regard to the baseline condition, including either a syllable (henceforth, SYLL) or a pseudoword (henceforth, PW) baseline. In the SYLL baseline condition, participants read aloud a visually presented string consisting of one syllable repeated three times (e.g., «*Феее фееееее фееееее …*» - «*Feee feeeeee feeeeee* ...») and repeated the syllable aloud one more time. In the PW baseline condition, participants read aloud a visually presented string of pseudowords (pseudonouns) phonotactically legal in Russian (e.g., «*Уптилья пикаш измеха …*» - «*Uptilja pikaš izmeha …*») and repeated any single of the pseudowords.

The stimuli in the experimental condition were Russian sentences (60 per paradigm). The full stimuli list is publicly available online: https://www.hse.ru/en/neuroling/research/fmri-mapping. The sentences were three words long and had one of the following syntactic structures:

(1) Adjective + noun (subject) + verb...

Умная соседка прочла ….

A clever neighbour read…

(2) Noun (subject) + verb + adjective ...

Скрипачка сдала сложный …

A violinist passed a challenging …

(3) Noun (subject) + adverb + verb ...

Грабитель ловко украл …

A thief skillfully stole…

All verbs were used in the present or past tense and required a direct object. In sentences of structure (2), the inflectional form of the adjective unambiguously determined the gender and number of the direct object. Words were no longer than three syllables and at least one word in each sentence was no longer than two syllables, resulting in the mean length of 7.38 syllables per sentence (SD .75, range 5–8). In both paradigms, verb tense and subject gender could repeat in no more than two consecutive trials both within and across presentation blocks consisting of three sentences (see 2.4. Procedure), with one exception of three consecutive past tense forms per paradigm.

The sentences were selected from a set of 160 sentences tested in an online pilot study, where 100 participants (50 females, age: mean 38.3, SD 11.5, range 18–68 years) read the sentences and finished them aloud with a semantically plausible word within five seconds, matching the task timing in the fMRI study. Their responses were recorded and scored for accuracy by a single rater. A response was considered accurate if it was a semantically plausible single word in a grammatically correct form (direct object). Based on the results, 120 sentences were selected and split into two halves with similar accuracy, to be used in the SYLL and PW paradigm (SYLL: mean accuracy 90.8%, SD 5.8%, range 78% – 100%; PW: mean accuracy 89.9%, SD 5.4%, range 80% – 100%; $t(118) = .842$, $p = .401$).

The two final lists were matched for the gender of the subject (31 feminine, 29 masculine), sentence structure (type 1: n=20, type 2: n=19, type 3: n=21), number of present and past verb forms (SYLL: 30 present / 30 past forms, PW: 29 present / 31 past forms), length in syllables (SYLL: mean 7.43, SD .76, range 5–8; PW: mean 7.31, SD .74, range 5–8; $t(118) = .843$, $p = .400$) and overall word frequency (SYLL: mean 39.37, SD 55.18, range .4–424.1; PW: mean 33.75, SD 45.71, range .5 – 277; $t(358) = 1.051$, $p = .293$).

In the SYLL baseline condition, stimuli were 60 phonologically legal Russian syllables (consonant + vowel), where the vowel was spelled multiple times in order to match the experimental condition for length in letters (for example, for the syllable "*фе*" - "*fe*", the stimulus was «*Фееее фееееее фееееее* ...» - «*Feee feeeeee feeeeee* ...»). Participants were instructed to ignore the exact number of vowel letters and pronounce the syllable duration approximately. In the PW baseline condition, the stimuli were 60 pseudowords phonotactically legal in Russian, constructed to pairwise match the number of syllables in experimental sentences. All pseudowords were constructed as pseudonouns: that is, none of them had inflection typical for other parts of speech.

2.3 MRI data acquisition

MRI data were obtained on a Siemens Magnetom Skyra 3T scanner with a 20-channel head coil at the National Medical and Surgical Center named after N.I. Pirogov of the Ministry of Healthcare of the Russian Federation. Participants wore MRI-compatible headphones to reduce scanner noise and head movement. Visual stimuli were presented using head-coil mounted goggles (NordicNeuroLab, Bergen, Norway). Stimuli presentation was controlled with nordicAktiva, version 1.2.1. Oral responses were recorded with an MRI-compatible FOMRI III™ microphone (Optoacoustics LTD) using OptiMRI recording software, version 3.1.

For anatomical reference, T1-weighted MPRAGE structural images were acquired with the following parameters: voxel size 1.0×1.0×1.0 mm, 176 axial slices in ascending order, slice thickness 1.00 mm,

field of view (FoV) 320×320 mm, repetition time (TR) 2200 ms, echo time (TE) 2.43 ms, flip angle 8°. Functional blood-oxygenation-level-dependent (BOLD) data were obtained using the following parameters: voxel size 3.0×3.0×3.0 mm, 30 oblique slices in interleaved order, slice thickness 3 mm, FoV 205x205 mm, TR 7000 ms, TE 30 ms, delay in TR 5000 ms (sparse sampling), flip angle 90°, 128 volumes per paradigm.

2.4 Procedure

Each participant was scanned on two separate occasions with an interval of 14 days. The first session began with a short instruction and practice to familiarize participants with the task outside the scanner. Inside the scanner, acquisition of T1 anatomical images was followed by the two functional paradigms (SYLL and PW). Their order was balanced across participants and remained constant in the two sessions.

Each paradigm started with visually presented instructions followed by one training block per condition (not analyzed). Then, 120 stimuli (60 experimental and 60 baseline) were presented in blocks of three (Figure 1). A sparse sampling procedure was used (Hall et al., 1999). During a five-second delay in TR, participants gave oral response to the current stimulus (Figure 1). They were instructed to remain silent during the next two seconds, indicated by a «!» sign. During this time, MR images were acquired. The sparse sampling procedure allowed to minimize motion-induced artifacts due to articulation and to monitor participants' responses with no acoustic noise from scanning. The duration of each paradigm was 14 min 56 sec in total.

FIGURE 1 ABOUT HERE

2.5 Data analysis

2.5.1 Behavioral data

The participants' auditory responses from both sessions were transcribed, except for one participant's responses that were lost due to technical error. Response accuracy was assessed independently by two raters. In the experimental condition, a response was considered accurate if it was a grammatically correct and semantically appropriate sentence completion. In the baseline conditions, a response was considered accurate if the participant read and repeated a syllable/pseudoword without any phonological errors. Inter-rater reliability, as assessed using percent agreement and Cohen's kappa (Cohen, 1960), was high (1st session: 98.36% and .76, respectively; 2nd session: 98.62% and .69, respectively). All inconsistencies were resolved by discussion between the two raters.

2.5.2 Activation maps

MRI data were analyzed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) for MATLAB 2014b. Prior to data analysis, first eight volumes of each functional paradigm, corresponding to instructions and training blocks, were discarded. For data preprocessing, images of each participant were first manually reoriented to the AC-PC plane. Then, functional scans were re-aligned to correct for head motion. Participants with excessive head movement (more than 5 mm in any direction) were excluded from further analysis. Functional images were coregistered to the anatomical T1 image, followed by spatial normalization of images to the International Consortium of Brain Mapping (ICBM) space template–European brains (Mazziotta et al., 1995) based on segmentation into six tissue types (grey matter, white matter, cerebrospinal fluid, bone, soft tissue and air/background) defined by tissue probability maps in SPM12. This step was followed by spatial smoothing with an isotropic 8-mm Gaussian kernel.

Statistical analysis was performed separately for each paradigm and each session, resulting in four activation maps for each participant. In the first-level (individual-level) analysis, a high-pass filter with a cut-off period of 256 s was employed to remove slow signal drift. The model included two conditions: experimental (sentence completion) and baseline (SYLL or PW, depending on the paradigm). The duration of each event was set to 7 s. Six movement parameters obtained in re-alignment were entered as regressors. A canonical hemodynamic response function with no derivatives was used to model BOLD response. Model estimation was done using a restricted maximum likelihood fit. T-contrast maps were computed separately for each paradigm, subtracting activation in the baseline condition (SYLL or PW) from activation in the experimental condition.

Although the present paper focuses on individual localization of language-related areas, we still conducted second-level (group) statistical analysis for illustrative purposes, based on activation maps from the first session. The contrast maps from the first-level analyses were submitted to the second-level one-sample t-test. To visualize the results at different levels of statistical stringency, three types of statistical thresholding were applied to all group fMRI activation maps: the most conservative Bonferroni correction for multiple comparisons (family-wise error correction, FWE) at $a = .05$; a more liberal cluster-size correction with a minimum cluster size of $k \geq 200$ mm3 at $a = .001$; and adaptive thresholding (AT) method proposed by Gorgolewski et al. (2012), which combines Gamma-Gaussian mixture modeling with topological FDR thresholding at $a = .05$. The AT method takes into account the strength of the signal: it generates lower thresholds when the signal is weak, resulting in fewer false negative clusters, and higher thresholds when the signal is strong, resulting in fewer false positive clusters, aiming to ensure an optimal balance between Type I and Type II error rate (Gorgolewski et al., 2012).

Anatomical labels for activation clusters were determined based on the Brainnetome atlas (Fan et al., 2016) implemented in the ICN_atlas toolbox (Kozák et al., 2017) for SPM12. The Brainnetome atlas was selected due to detailed parcellation, as it includes 246 regions in total. The activation maps were visualized in MRIcroGL 1.2.2 (https://www.nitrc.org/projects/mricrogl).

2.5.3 Assessing individual-level activation of 'key language-related areas'

Since the ultimate goal of the localizer was to individually localize critical language areas, we estimated how well the paradigms were able to activate them in each participant. Following Benjamin et al. (2017), we focused on the following 'key language-related areas' in the left hemisphere: Broca's area, Exner's area, supplementary motor area (SMA), angular gyrus, Wernicke's area, and basal temporal language area. For a more detailed analysis, we divided the Broca's area into two smaller areas (pars triangularis and pars opercularis of the inferior frontal gyrus) and complemented the Wernicke's area (posterior superior temporal gyrus, pSTG) with the adjacent region in the posterior middle temporal gyrus (pMTG), another key region typically activated by language localizer paradigms (Połczyńska et al., 2017). At each of the three statistical thresholds applied in first-level analyses, we calculated which percentage of the resulting eight 'key language-related areas' was activated by the paradigm, using the Brainnetome atlas (Fan et al., 2016) implemented in the ICN_atlas toolbox (Kozák et al., 2017). The respective list of areas from the Brainnetome atlas is presented in Supplementary Table S1. The resulting values were visualized using the seaborn library (https://seaborn.pydata.org/) in Python 3.7.

To test how individual-level activation volume in eight 'key language-related areas' was affected by baseline and statistical threshold, a separate repeated-measures ANOVA was conducted for each area. The ANOVAs were conducted on the number of significantly activated voxels in the area and tested the main effects of baseline and statistical threshold and their interaction. Mauchly's test was

used to check the assumption of sphericity. In case it was violated, Greenhouse-Geisser correction was applied. The Bonferroni correction was applied to correct for the number of statistical models, resulting in a = .00625.

2.5.4 Test-retest reliability

Consistency of localizer paradigms with regard to individual-level activation is crucial for clinical applications (Binder et al., 2008), so we estimated test-retest reliability of our paradigms. Following Wilson et al. (2017), we used the Dice coefficient (Rombouts et al., 1997) to quantify the similarity of language-related activation in the first and second session of each participant. The Dice coefficient indicated a degree of overlap between the participant's activation maps in the first and second session and was calculated as follows: Dice = $2*V_{overlap} / (V_1 + V_2)$, where V_1 and V_2 denote the number of supra-threshold voxels in the first and second sessions of an individual and Voverlap is the total number of overlapping voxels. The overlap was calculated in Convert3D (http://www.itksnap.org/pmwiki/pmwiki.php?n=Downloads.C3D), which is an extension of the ITK-SNAP tool (Yushkevich et al., 2006). The Dice coefficients can be interpreted as low (.00 to .19), low-moderate (.20 to .39), moderate (.40 to .59), moderate-high (.60 to .79) or high (.80 to 1.00) (Wilson et al., 2017).

The Dice coefficients were calculated for each participant for the frontal, temporal-parietal and frontal-temporal-parietal regions, separately for the SYLL and PW paradigm. Regions were defined using brain lobe masks available in the LI toolbox (Wilke & Lidzba, 2007) based on the atlas by Hammers et al. (2003). The frontal, temporal-parietal and frontal-temporal-parietal regions were selected because they correspond respectively to anterior language regions, posterior language regions and combination thereof, excluding the occipital lobe that is not relevant for the language function. To test what factors affected test-retest reliability, a repeated-measures ANOVA was conducted on Dice coefficients, testing the main effects of brain region, baseline and statistical threshold, as well as all interactions thereof. Mauchly's test was used to check the assumption of sphericity. In case it was violated, Greenhouse-Geisser correction was applied.

2.5.5 Lateralization indices

Finally, we evaluated the hemispheric lateralization of individual language-related activation. This analysis aimed to confirm the validity of the paradigms and estimate their reliability in establishing individual lateralization of language processing.

Lateralization indices (LIs) were calculated with the LI toolbox for SPM (Wilke & Lidzba, 2007), based on the count and value of suprathreshold voxels and using adaptive thresholding. As implemented in the LI toolbox, adaptive thresholding uses averaged intensity of all voxels in the image as the internal threshold for a given participant, thus taking into account inter-subject variability of BOLD response. LI can take the values from +1 to -1, where +1 stands for full left lateralization of the activation, -1 indicates full right lateralization and 0 indicates bilateral activation. Similarly to Dice coefficients, LIs were calculated for the frontal, temporal-parietal and frontal-temporal-parietal regions for each participant individually, separately for each paradigm in each scanning session.

Given right-handedness of participants in our study, we expected to observe typical left-hemispheric dominance of language-related activation in the majority of participants. This outcome would confirm the validity of the paradigms. Finally, we used a repeated-measures ANOVA to test how LIs were affected by baseline, region, number of session, and interactions thereof. Mauchly's test was

used to check the assumption of sphericity. In case it was violated, Greenhouse-Geisser correction was applied.

3 Results

3.1 Behavioral results

Participants' task performance was at ceiling. In the SYLL paradigm, mean sentence completion accuracy was 96.1% (SD 3.1%, range 91.4% – 100.0%) in the first session and 98.4% (SD 1.6%, range 95.0% – 100.0%) in the second session. In the PW paradigm, the mean accuracy was 96.4% (SD 2.4%, range 93.1% – 100.0%) in the first session and 97.5% (SD 1.8%, range 94.9% – 100.0%) in the second session.

3.2 Group-level activation maps

For illustration purposes, we demonstrate group-level activation maps from the first session produced at three statistical thresholds: with FWE correction, cluster-size correction and AT (Gorgolewski et al., 2012) (see Figure 2 and Figure 3). A full list of activation clusters comprising more than 100 voxels is presented in Supplementary Table S2 (SYLL paradigm) and Supplementary Table S3 (PW paradigm).

At the most conservative statistical threshold (FWE correction; upper panel in Figures 2 and 3), both versions of the paradigm elicited significant group-level activation in the left inferior and superior frontal gyri. Additionally, the SYLL paradigm activated the orbital gyrus and insula and the PW paradigm activated the left middle and superior temporal gyri. With AT (Gorgolewski et al., 2012; middle panel in Figures 2 and 3), significant group-level activation in both paradigms extended to the left middle frontal gyrus. The activation additionally extended to the left middle and superior temporal gyri in the SYLL paradigm and the orbital gyrus, insula and parts of occipital cortex in the PW paradigm. Finally, when using the most liberal cluster-size correction (bottom panel in Figures 2 and 3), activation in both paradigms extended to a wide network of left frontal, left temporal and bilateral occipital regions, particularly extensive with the PW paradigm.

For illustrative purposes, we also provide individual activation maps from three example participants (first session) in Supplementary Figures S1-S3.

FIGURE 2 ABOUT HERE

FIGURE 3 ABOUT HERE

3.3 Individual-level activation in key 'language-related areas'

The violin plots in Figure 4 present the number of activated voxels in 'key language-related areas' adopted from Benjamin et al. (2017) across participants (as a percentage of the total number of voxels in the area). The respective numeric values are presented in Supplementary Table S4.

FIGURE 4 ABOUT HERE

As seen in Figure 4, individual-level activation was the most extensive in pars triangularis (median activation ranging from 17% to 59% depending on the paradigm and threshold) and pars opercularis of the inferior frontal gyrus (median activation ranging from 22% to 56%). Individual-level activation was also extensive in the Exner's area (median activation ranging from 15% to 47%) and

the middle temporal gyrus (median activation ranging from 2% to 25%). In other analyzed regions, individual-level activation was less extensive.

Bonferroni-corrected repeated-measures ANOVAs showed that individual-level activation volume was greater with the PW than SYLL baseline in three of 'key language-related areas': pars triangularis of the inferior frontal gyrus, $F(1,17) = 25.22$, $p < .001$, posterior middle temporal gyrus, $F(1,17) = 25.75$, $p < .001$, and angular gyrus, $F(1,17) = 27.97$, $p < .001$. In the other five 'key language-related areas', individual-level activation volume was not significantly affected by baseline. Expectedly, individual-level activation volume was significantly affected by statistical threshold in six of 'key language-related areas': pars opercularis of the inferior frontal gyrus, $F(1.08, 18.29) = 15.82$, $p = .001$, pars triangularis of the inferior frontal gyrus, $F(1.11, 18.78) = 21.40$, $p < .001$, posterior superior temporal gyrus, $F(1.10, 18.78) = 9.16$, $p = .006$, posterior middle temporal gyrus, $F(1.17, 19.94) = 8.92$, $p = .005$, angular gyrus, $F(1.06, 18.01) = 10.71$, $p = .003$, supramarginal gyrus, $F(1.25, 21.19) = 15.05$, $p < .001$. Post-hoc pairwise comparisons showed than in all six regions, cluster-size correction with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$ yielded fewer significantly activated voxels than the FWE correction for multiple comparisons at $p < .05$ (all $p < .001$). No other pairwise comparisons were significant when corrected for multiple comparisons. The interaction between baseline and statistical threshold was not significant in any of 'key language-related areas'.

3.4 Test-retest reliability

Mean Dice coefficients quantifying the overlap of individual activation in the first and second scanning session are presented in Table 1. Mean Dice coefficients ranged from .39 to .61, that is, from low-moderate to moderate-high. A Greenhouse-Geisser-corrected repeated-measures ANOVA demonstrated a significant effect of brain region, $F(1.04, 17.74) = 6.72$, $p = .018$. Post-hoc pairwise comparisons showed that Dice coefficients were higher in the frontal than temporo-parietal ($p = .019$) or frontal-temporo-parietal ($p = .050$) region, and in the frontal-temporo-parietal than temporo-parietal ($p = .016$) region. Dice coefficients were significantly higher with the PW than SYLL baseline, $F(1, 17) = 5.08$, $p = .038$. Finally, there was a significant effect of statistical threshold, $F(1.41, 23.92) = 6.49$, $p = .011$. Post-hoc pairwise comparisons showed that Dice coefficients were higher with the most liberal statistical threshold (cluster-size correction with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$) relative to FWE correction for multiple comparisons at $p < .05$ ($p < .001$) and relative to the adaptive thresholding as implemented in Gorgolewski et al. (2012) at $p < .05$ ($p = .011$). No interactions were significant.

TABLE 1 ABOUT HERE

The range of Dice coefficients in Table 1, varying between 0 and .79, indicated substantial inter-individual variability. Low Dice coefficients around 0 resulted exclusively following the application of the AT method suggested by Gorgolewski et al. (2012). In some cases, no or almost no voxels survived the statistical threshold established by this method in one of the two participant's sessions.

3.5 Lateralization indices

Table 2 presents mean lateralization indices (LI) for three regions (frontal, temporal-parietal and frontal-temporal-parietal) for the two paradigms (SYLL and PW) in the two scanning sessions. Table 2 also includes Spearman's correlation coefficients examining reproducibility of the individual LI values across two sessions.

TABLE 2 ABOUT HERE

Mean LIs in all regions across two scanning sessions showed left lateralization regardless of the paradigm (SYLL or PW), ranging from .32 to .51. Individual LIs ranged between very strong left-hemispheric dominance (.88) and bilateral organization (-.04). Minimum individual LIs showing bilateral organization were observed mostly in the temporal-parietal region in both sessions and for both paradigms. The Spearman's correlation tests showed statistically significant correlation between LIs in the two experimental sessions for the PW paradigm in all regions (all $p < .05$). For the SYLL paradigm, the correlation in regions remained at the level of a statistical trend.

A repeated-measures ANOVA showed a significant three-way interaction between baseline, region and session, $F(1.11, 18.93) = 8.92$, $p = .006$. To address this interaction, separate repeated-measures ANOVAs were performed for the SYLL and PW baseline (there was no main effect of baseline, $F(1,17) = .06$, $p = .811$). With the SYLL baseline, a main effect of region was significant, $F(1.12, 18.99) = 15.88$, $p = .001$, with higher LIs in the frontal than temporo-parietal ($p = .001$) or fronto-temporo-parietal region ($p = .038$) and in the fronto-temporo-parietal than temporo-parietal region ($p < .001$). No other main effects or interactions were significant. With the PW baseline, a two-way interaction of session and region was significant, $F(1.08, 18.31) = 6.71$, $p = .017$, so separate repeated-measures ANOVAs were performed for the first and second session (there was no main effect of session, $F(1,17) = .26$, $p = .618$). In the first session with the PW baseline, the main effect of region was significant, $F(1.07, 18.19) = 15.21$, $p = .001$. Post-hoc pairwise comparisons showed the same hierarchy of LIs as with the SYLL baseline: LIs were higher in the frontal than temporo-parietal ($p = .001$) and frontal-temporo-parietal region ($p = .006$) and in the frontal-temporo-parietal than temporo-parietal region ($p = .001$). In the second session with the PW baseline, the main effect of region was again significant but much greater, $F(1.09, 18.52) = 80.49$, $p < .001$. Post-hoc pairwise comparisons showed the same hierarchy of LIs as for the first session or for the SYLL baseline: LIs were higher in the frontal than temporo-parietal ($p < .001$) and fronto-temporo-parietal region ($p < .006$) and in the frontal-temporo-parietal than temporo-parietal region ($p < .001$). That is, the significant three-way interaction between baseline, region and session was driven by the main effect of region being the greatest in the second session with the PW baseline.

4 Discussion

We presented a new fMRI language localizer for preoperative language mapping in Russian-speaking individuals. Following the world's best practices (Barnett et al., 2014, Black et al., 2017, Salek et al., 2017, Połczyńska et al., 2017, Unadkat et al., 2019, Wilson et al., 2017, Zacà et al., 2012), the paradigm used a sentence completion task that uniquely engages both language production and comprehension at the word and sentence level. The current study validated the localizer paradigm in a control group of neurologically healthy individuals. In this group, the paradigm successfully activated key language-related areas, elicited expected left-hemispheric lateralization and showed test-retest reliability comparable to previous studies. Apart from demonstrating general validity and reliability of the paradigm, we compared two different baseline conditions (SYLL and PW), for the first time for the sentence completion task. All outcomes were reported at three statistical thresholds.

4.1 Activation of key language-related areas

At the group level, both versions of the localizer (with the SYLL and PW baseline) elicited significant activation in an expected network of language-related areas. At the most stringent statistical threshold (FWE correction at $\alpha = .05$), both versions of the paradigm elicited significant activation in the left posterior inferior and posterior superior frontal gyri (differences in results with

the SYLL and PW baseline are discussed in Section 4.4). The left posterior inferior frontal gyrus has been implicated in many linguistic processes engaged by the sentence completion task: sentence parsing (Hagoort et al., 2005), conceptual and lexical selection of the completing word (Robinson et al., 2010; Zyryanov et al., 2020), morphosyntactic inflection of the completing word (Den Ouden et al., 2019), and articulatory encoding (Flinker et al., 2007). Such multifaceted involvement of the left inferior frontal gyrus in linguistic processes may explain why its activation was the most statistically robust. The posterior superior frontal gyrus (premotor to supplementary motor cortex) was likely activated as part of the dorsal route supporting articulation (Hickok & Poeppel, 2007) or speech initiation (Dragoy et al., 2020, Kinoshita et al., 2014).

At a more liberal statistical threshold (AT at α = .05 as implemented by Gorgolewski et al. (2012)), activation in both versions of the localizer (with the SYLL and PW baseline) extended to the left middle frontal gyrus and to a new cluster of activation encompassing the mid-posterior portions of the left middle temporal gyrus and superior temporal sulcus. More superior and posterior portions of the temporal activation may pertain to phonological processing (Graves et al., 2008, Buchsbaum et al., 2001), whereas activation in the mid part of the middle temporal gyrus may reflect several components of semantic processing, such as storage of heteromodal semantic knowledge (Binder et al., 2009), linkage of word forms to meanings (Bonilha et al., 2017, Hickok & Poeppel, 2007), and semantic control when searching for a completing word (Davey et al., 2016).

Finally, at the most liberal statistical threshold (cluster size correction for multiple comparisons at α = .05), activation extended to a broad left-lateralized frontotemporal and bilateral occipital network. Occipital activation may pertain to reading, including the linkage between visual word processing and phonological word representations (Mano et al., 2013, Richardson et al., 2011): although the baseline condition also required reading, it did not involve any linkage to word representations. Alternatively, the greater occipital activation in the experimental condition may reflect mental imagery of the sentence content (Pearson et al., 2015).

Therefore, group-level results proved that the sentence completion paradigm successfully activated both anterior and posterior areas implicated in language processing. This is in line with previous empirical work (Barnett et al., 2014, Połczyńska et al., 2017, Salek et al., 2017, Unadkat et al., 2019, Wilson et al., 2017, Zacà et al., 2012) and reviews (Black et al., 2017, Manan et al., 2020) promoting sentence-level tasks and particularly sentence completion for eliciting activation in a more comprehensive language network than with word-level tasks. However, significant group-level clusters may result from different individual-level patterns: consistent activation across all/most participants versus strong activation in fewer participants. For clinical use, it is most crucial whether the paradigm consistently elicits significant activation of language-related areas in each tested individual, so that individual maps of language-related areas can be routinely used by a neurosurgeon.

To test this, we analyzed individual-level activation in each participant's first session. We focused on eight 'key language-related areas' adapted with slight modifications from Benjamin et al. (2017). Mirroring the group-level findings, individual-level activation was the most consistent in pars triangularis of the Broca's area, followed by pars opercularis of the Broca's area, followed by the Exner's area. Among these, pars triangularis of the Broca's area was to some extent activated in each participant (with an exception of the AT statistical thresholding with the SYLL baseline). As represented by the interquartile range of activation area, activation spanned one third to two thirds of this area in most participants. Similarly, most participants showed activation in about one third to one half of pars opercularis of the Broca's area and, with somewhat greater individual variability, of the

Exner's area. In the vast majority of participants, significant individual-level activation was also present in the posterior middle temporal gyrus, followed by the posterior superior temporal gyrus and the angular gyrus.

On the other hand, most participants did not show significant individual-level activation in the basal temporal area or supramarginal gyrus. The lack of activation in the supramarginal gyrus was surprising, given that activation at the temporo-parietal junction is expected in sentence-level tasks (Połczyńska et al., 2017) and has sometimes also been found with word-level tasks (Roux et al., 2003, Stippich et al., 2007). Still, the majority of participants in the present study showed activation in the adjacent angular gyrus. With regard to the basal temporal language area, it has not been consistently activated by previous fMRI language localizers either (Połczyńska et al., 2017, Barnett et al., 2014, Wilson et al., 2017), despite its long-presumed involvement in lexical retrieval (Krauss et al., 1996, Lüders et al., 1991).

4.2 Lateralization of language processing

All participants in the present study were right-handed with no history of neurological disorders. Thus, we expected that the paradigm should elicit primarily left-hemispheric lateralization of language processing, with a certain degree of individual variability (Knecht et al., 2000, Springer et al., 1999). Indeed, mean LI values indicated left-hemispheric lateralization of task-related brain activity, with individual values ranging between bilateral organization and very strong left lateralization. Thus, the ability of the paradigm to detect hemispheric lateralization of language processing activity was confirmed. Numerically, the LI values (mean .32 to .51, depending on the brain region and baseline) were comparable to those in previous studies with neurologically healthy right-handed participants. For example, Deblaere et al. (2002) found individual LI values from –.08 to .58 across four language tasks; Dodoo-Schittko et al. (2012) found mean LI values of .44 and .45 in a verb and antonym generation tasks respectively (for review, see Bradshaw et al., 2017).

Interestingly, language-related activity was significantly more strongly left-lateralized in the frontal (and, correspondingly, frontal-temporal-parietal) than temporal-parietal region. This is in line with contemporary models of language processing. For example, the dual-stream model (Hickok & Poeppel, 2007) postulates a bilaterally organized ventral stream, which primarily involves the temporal lobe and connects speech sounds to meanings, and a left-lateralized dorsal stream, which extends to the frontal lobe and maps speech sounds to articulatory networks. In the same vein, Peelle (2012) argues that phonological and lexical information are processed bilaterally in the temporal lobe, whereas sentence processing engages a left-lateralized pathway including the left inferior frontal gyrus. Although the above models (Hickok & Poeppel, 2007, Peelle, 2012) are mainly concerned with auditory speech processing and differ in the specific division of labor between the frontal and temporal regions, our findings converge with them with regard to stronger left-hemispheric lateralization in the frontal than temporal lobe.

4.3 Test-retest reliability

Dice coefficients measuring the spatial overlap of significant activation in the individual's first and second scanning session were in the moderate range: .39 to .61, depending on the region of interest, baseline condition and statistical threshold. In the only previous study measuring Dice coefficients for the sentence completion task (Wilson et al., 2016), the coefficients indicated a smaller spatial overlap, ranging between .06 and .47 depending on the region of interest and statistical threshold. As a more specific example, at the statistical threshold of $\alpha = .001$ with the minimum cluster size of 2 cm2, the mean Dice coefficient in the broadest examined region of interest ('supratentorial region' in

Wilson et al. (2016) and the combination of frontal, temporal and parietal lobe in the present study) was .34 in Wilson et al. (2016) versus .43 or .56 with the SYLL and PW baseline respectively in the present study. The Dice coefficients in the present study were also comparable to those reported in previous studies for other language tasks in neurologically healthy participants, presented in Table 3, and to the mean overlap of .48 across a variety of tasks established in a meta-analysis by Bennett and Miller (2010).

TABLE 3 ABOUT HERE

Test-retest reliability was affected by the brain region and statistical threshold. Dice coefficients were significantly higher in the frontal (and, correspondingly, frontal-temporal-parietal) than temporal-parietal region. This is consistent with higher Dice coefficients in the frontal than temporal or parieto-occipital region of interest in a free reversed association task by Fesl et al. (2010). Conversely, Nettekoven et al. (2018) showed higher Dice coefficients for a picture naming task in the inferior frontal gyrus than the superior temporal gyrus. Possibly, this discrepancy could arise because Nettekoven et al. (2018) used smaller regions of interest than here and in Fesl et al. (2010). With regard to the statistical threshold, Dice coefficients were higher with the most liberal than two more conservative statistical thresholds, in line with previous literature (Nettekoven et al., 2018, Stevens et al., 2013, Wilson et al., 2016).

Hemispheric lateralization of language-related activity also showed moderate test-retest reliability. LIs in the first and second scanning session showed either a significant moderate-to-strong correlation (with the PW baseline) or a statistical trend for a moderate correlation (with the SYLL baseline). This held true when LIs were calculated for the frontal region, temporal-parietal region, and combination thereof. The findings on moderate reliability of the paradigm in identifying both localization and hemispheric lateralization of language-related activity contribute to the literature on general test-retest reliability of fMRI (Bennett & Miller, 2010, Elliott et al., 2020, Holiga et al., 2018).

4.4 Comparison of baseline conditions

Each participant was administered two versions of the paradigm with different baselines: reading a sequence consisting of the same syllable and repeating the syllable once more (SYLL baseline) and reading a sequence of pseudowords and repeating any of them once (PW baseline). Both baselines are theoretically plausible for the sentence completion task, yet no previous studies have empirically investigated how their choice may affect the outcomes.

The SYLL and PW baselines showed a very similar spatial distribution of significantly activated areas at the group level (see Section 4.1). Among minor differences in the spatial distribution, one may highlight somewhat more inferior temporal activation with the PW baseline: it largely extended to the inferior temporal sulcus and gyrus, whereas significant activation with the SYLL baseline encompassed more of the superior temporal sulcus and gyrus. Possibly, this pattern emerged because the PW baseline exactly matched the experimental condition in phonological complexity. Therefore, their comparison yielded significant activations in more inferior temporal areas implicated in lexical-semantic processing (Binder et al., 2009, Davey et al., 2016) but not in more superior temporal areas enabling phonological processing (Graves et al., 2008, Buchsbaum et al., 2001).

With regard to the extent of activation, the area of significant group-level activation was somewhat greater with the PW than SYLL baseline across statistical thresholds. This was unexpected: we hypothesized that subtraction of the PW baseline should have yielded a smaller difference from the

experimental condition because the PW baseline matched it closer in terms of phonological complexity and real-word neighbors and associations. One possible explanation are differences in how individual participants approached the SYLL baseline: for example, how accurately they tried to pronounce the length (number of vowels) in the string, whether they imposed prosody when reading, et cetera. Possibly, such individual variability could introduce noise in the data, reducing the statistical power of comparison to the experimental condition. Another possible account is that the simpler SYLL baseline allowed more time and cognitive resources for non-task-related cognitive activity, which could also introduce noise in the data.

At the individual level, the paradigms with the SYLL and PW baseline also showed a very similar pattern. With both baselines, activation was most robust across individuals in the inferior frontal gyrus and Exner's area, followed by the posterior middle and superior temporal gyri and the angular gyrus, whereas the basal temporal area and the supramarginal gyrus showed no activation in most participants (Section 4.1). Mirroring the group-level results, areas of individual activation were numerically lower with the SYLL than PW baseline in most regions of interest. This difference was significant in pars triangularis of the inferior frontal gyrus, posterior middle temporal gyrus, and angular gyrus. With regard to hemispheric lateralization, it did not significantly differ with the SYLL versus PW baseline.

Finally, test-retest reliability, or spatial overlap between significant activation in the participant's first and second scanning session, was significantly higher with the PW than SYLL baseline. Test-retest reliability of hemispheric lateralization was also higher with the PW baseline. With the PW baseline, the LIs in the first and second session showed a significant moderate-to-high correlation, whereas with the SYLL baseline, they remained at the level of a statistical trend for moderate correlation. This held true for the frontal region, temporal-parietal region and combination thereof.

To summarise, the PW baseline provided more robust activation, as reflected in somewhat more extensive significant activation and higher test-retest reliability. As discussed above, the cognitively simpler SYLL baseline may have allowed more time and cognitive resources for non-task-related cognitive activity or, alternatively, have provoked interindividual variability in the specifics of task performance. Both could introduce noise in the data and reduce statistical power compared to the more cognitively taxing PW baseline. On the other hand, this quantitative difference was not large, and the SYLL baseline appeared to have a qualitative advantage. Namely, posterior temporal activation with the SYLL baseline encompassed more superior areas than with the PW baseline. Damage to posterior superior temporal gyrus impairs phonological processing (Binder, 2015) and possibly word comprehension, although with some potential for neuroplasticity (Hillis et al., 2017), so its mapping is crucial. Apart from that, the SYLL baseline has the advantage of being cognitively simpler and thus more feasible in the clinical population. Here in the control group of neurologically healthy participants, task performance accuracy was at ceiling and did not differ between the two versions of the paradigm. However, in case of preoperative neuropsychological deficits in patients with brain tumors (Ek et al., 2010, Racine et al., 2015) and epilepsy (Patrikelis et al., 2016), lower cognitive complexity and thus greater feasibility may present an important clinical advantage of the SYLL baseline, despite greater robustness of the PW baseline in the control group.

4.5 Comparison of statistical thresholds

We reported all measures and activation maps at three different statistical thresholds. The most conservative was the FWE correction for multiple comparisons at $\alpha = .05$, followed by the AT method proposed by Gorgolewski et al. (2012) at $\alpha = .05$, followed by the most liberal cluster-size

correction for multiple comparisons with a minimum cluster size of $k \geq 200$ mm3 at $\alpha = .001$. Many previous studies have also reported results at multiple statistical thresholds (Dodoo-Schnitko et al., 2012, Morrison et al., 2016a, Nadkarni et al., 2015, Nettekoven et al., 2018 Wilson et al., 2017), since there is no 'gold standard' for statistical thresholding in individual or group-level fMRI analysis. Moreover, studies have shown that the statistical threshold may vastly impact metrics induced from fMRI analysis, such as lateralization indices (Nadkarni et al., 2015) or test-retest reliability metrics (Stevens et al., 2016), so reporting results at only one threshold could be misleading.

In the present study, the spatial distribution of activation both at the group and individual level was expectedly similar across statistical thresholds, although some relevant clusters of activation only emerged at more liberal statistical thresholds. For example, significant group-level activation in the posterior superior temporal gyrus became evident at the two more liberal statistical thresholds, and significant group-level activation in the angular and particularly supramarginal gyrus mainly emerged at the most liberal statistical threshold.

At the individual level, participants highly varied in the extent of activation depending on the statistical threshold: the extent of activation that was present in some participants at the most stringent threshold only appeared in others at more liberal thresholds (Supplementary Table S4). This adds to the evidence for impossibility of using a one-for-all statistical threshold in individual preoperative mapping in clinical practice. Various methods have been proposed in previous literature for individualized statistical thresholding. They have been based, for example, on receiver operating characteristic reliability (Stevens et al., 2016), normalizing statistical maps to the local peak activation amplitude within a brain region (Gross & Binder, 2014, Voyvodic et al., 2009), thresholding based on a fixed percentage of brain activation rather than a statistical threshold (Wilson et al., 2016), and expert judgement by a clinician (American College of Radiology, 2014, Benjamin et al., 2017, 2018). In the present study, we reported the results using one method of individualized thresholding: the AT method by Gorgolewski et al. (2012), which is based on the combination of Gamma-Gaussian mixture modelling with topological FDR thresholding. The AT method did not alleviate individual variability in the extent of activation: the percentage of activation in key language-related areas was not more homogeneous across participants when using the AT method than the two non-adaptive thresholding methods (Figure 4). For clinical practice, this means that the AT method would not solve the issue of largely variable activation strength across individuals that confounds the interpretation of the presence or absence of significant activation in an area. An important research direction, which was beyond the scope of the present study, would be to compare other methods of individualized statistical thresholding.

Test-retest reliability, as measured by Dice coefficients, was in the moderate range across statistical thresholds. Still, Dice coefficients were significantly higher with the most liberal statistical threshold compared to the two more conservative statistical thresholds, in line with previous literature (Nettekoven et al., 2018, Stevens et al., 2013, Wilson et al., 2016). With regard to LIs, these were calculated using adaptive thresholding and taking into account the values of suprathreshold voxels as implemented in the LI Toolbox for SPM (Wilke & Lidzba, 2007), so comparison of different statistical thresholds did not apply to this measure.

4.6 Future directions

The present study validated the fMRI language localizer in a control group of neurologically healthy participants. For full validation of the localizer, the crucial next step is to test it in the clinical group

of presurgical patients with brain tumors and drug-resistant epilepsy. Data from a clinical sample will test the ability of the localizer to elicit activation in critical language-related areas in patients with different etiology and localization of pathological tissue and thus ultimately assess its clinical value. Data from a clinical sample would also provide the best test case for assessing the clinical value of different methods of individualized statistical thresholding (American College of Radiology, 2014, Benjamin et al., 2017, 2018, Gross & Binder, 2014, Stevens et al., 2016, Voyvodic et al., 2009, Wilson et al., 2016), which remained beyond the scope of the present study.

Finally, as a validation against the gold standard, the findings of the fMRI language localizer in the clinical group will need to be compared to the findings from intraoperative mapping using DES. So far, such comparisons between DES and fMRI language localizer protocols have yielded diverging results (Morrison et al., 2016b, Roux et al., 2003, Spena et al., 2010; for review, see De Witte & Mariën, 2013). Thus, it would be informative to validate our particular fMRI language localizer protocol against DES and thereby add to general evidence on the sensitivity and specificity of fMRI language localizer protocols for preoperative language mapping.

5 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6 Author Contributions

KE and SM contributed equally to the manuscript. OD conceptualized, designed and supervised the study. ES and OD created linguistic materials. SM and OB implemented and tested the paradigm. KE, SM, OB and AM collected the data. KE and SM performed data analysis. SM and KE wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version

7 Funding

The research was supported by the Center for Language and Brain NRU Higher School of Economics, RF Government grant, ag. No. 14.641.31.0004.

8 Acknowledgments

We would like to thank Evgenii Kalenkovich, Olga Buivolova and Valeriya Zelenkova for their help with paradigm development, and all study participants for their contribution.

9 References

Agarwal, S., Sair, H. I., Gujar, S., and Pillai, J. J. (2019). Language mapping with fMRI. Top. Magn. Reson. Imaging. 28:4, 225–233. doi:10.1097/rmr.0000000000000216

Almairac, F., Duffau, H., and Herbet, G. (2018). Contralesional macrostructural plasticity of the insular cortex in patients with glioma: A VBM study. Neurology. 91:20, e1902-e1908. doi:10.1212/WNL.0000000000006517

American College of Radiology (2014). ACR–ASNR–SPR practice parameter for the performance of functional magnetic resonance imaging (fMRI) of the brain. Amended 2014 (Resolution 39). American College of Radiology.

Barnett, A., Marty-Dugas, J., and McAndrews, M. P. (2014). Advantages of sentence-level fMRI language tasks in presurgical language mapping for temporal lobe epilepsy. Epilepsy Behav. 32, 114–120. doi:10.1016/j.yebeh.2014.01.010

Bennett, C.M., and Miller, M.B. (2010). How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133-155. doi:10.1111/j.1749-6632.2010.05446.x

Benjamin, C. F., Walshaw, P. D., Hale, K., Gaillard, W. D., Baxter, L. C., Berl, M. M., Polczynska, M., Noble, S., Alkawadri, R., and Hirsch, L. J. (2017). Presurgical language fMRI: Mapping of six critical regions. Hum. Brain Mapp. 38:8, 4239–4255. doi:10.1002/hbm.23661

Benjamin, C.F.A., Dhingra, I., Li, A.X., Blumenfeld, H., Alkawadri, R., Bickel, S., Helmstaedter, C., Meletti, S., Bronen, R.A., Warfield, S.K., Peters, J.M., Reutens, D., Połczyńska, M.M., Hirsch, L.J., and Spencer, D.D. (2018). Presurgical language fMRI: Technical practices in epilepsy surgical planning. Hum. Brain Mapp. 39:10, 4032-4042. doi:10.1002/hbm.24229

Binder J. R. (2015). The Wernicke area: Modern evidence and a reinterpretation. Neurology. 85:24, 2170–2175. doi:10.1212/WNL.0000000000002219

Binder, J.R., Swanson, S.J., Hammeke, T.A., Morris, G.L., Mueller, W.M. Fischer, M., Benbadis, S., Frost, J.A., Rao, S. M., and Haughton, V. M. (1996). Determination of language dominance using functional MRI: A comparison with the Wada test. Neurology. 46:4, 978–984. doi:10.1212/wnl.46.4.978

Binder, J. R., Swanson, S. J., Hammeke, T. A., and Sabsevitz, D. S. (2008). A comparison of five fMRI protocols for mapping speech comprehension systems. Epilepsia. 49:12, 1980–1997. doi:10.1111/j.1528-1167.2008.01683.x

Binder J. R., Desai R. H., Graves W. W., and Conant L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex. 19:12, 2767–2796. doi:10.1093/cercor/bhp055

Black, D.F., Vachha, B., Mian, A., Faro, S.H., Maheshwari, M., Sair, H.I., Petrella, J.R., Pillai, J.J., and Welker, K., 2017. American Society of Functional Neuroradiology-recommended fMRI paradigm algorithms for presurgical language assessment. Am. J. Neuroradiol. 38:10, E65–E73. doi:10.3174/ajnr.A5345

Bleich-Cohen, M., Hendler, T., Kotler, M., and Strous, R. D. (2009). Reduced language lateralization in first-episode schizophrenia: An fMRI index of functional asymmetry. Psychiatry Res. Neuroimaging. 171:2, 82–93. doi:10.1016/j.pscychresns.2008.03.002

Bonilha L., Hillis A. E., Hickok G., Den Ouden D. B., Rorden C., and Fridriksson J. (2017). Temporal lobe networks supporting the comprehension of spoken words. Brain. 140:9, 2370–2380. doi:10.1093/brain/awx169

Bradshaw, A. R., Thompson, P. A., Wilson, A. C., Bishop, D., and Woodhead, Z. (2017). Measuring language lateralisation with different language tasks: a systematic review. PeerJ. 5, e3929. doi:10.7717/peerj.3929

Brennan, N.M.P., Whalen, S., De Morales Branco, D., O'Shea, J.P., Norton, I.H., and Golby, A.J. (2007). Object naming is a more sensitive measure of speech localization than number counting: converging evidence from direct cortical stimulation and fMRI. NeuroImage. 37, S100–S108. doi:10.1016/j.neuroimage.2007.04.052

Buchsbaum, B. R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cogn. Sci. 25:5, 663-678. doi:10.1016/S0364-0213(01)00048-9

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. doi:10.1177/001316446002000104

Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., Krieger-Redwood, K., Bernhardt, B. C., Smallwood, J., and Jefferies, E. (2016). Exploring the role of the

posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. NeuroImage. 137, 165–177. doi:10.1016/j.neuroimage.2016.05.051

Deblaere, K., Backes, W.H., Hofman, P., Vandemaele, P., Boon, P.A., Vonck, K., Boon, P., Troost, J., Vermeulen, J., Wilmink, J., Achten, E., and Aldenkamp, A. (2002). Developing a comprehensive presurgical functional MRI protocol for patients with intractable temporal lobe epilepsy: a pilot study. Neuroradiology. 44:8, 667–673. doi:10.1007/s00234-002-0800-4

De Guibert, C., Maumet, C., Ferré, J.-C., Jannin, P., Biraben, A., Allaire, C., Barillot, C., and Le Rumeur, E. (2010). FMRI language mapping in children: a panel of language tasks using visual and auditory stimulation without reading or metalinguistic requirements. NeuroImage. 51(2), 897-909. doi:10.1016/j.neuroimage.2010.02.054f

Den Ouden D., Malyutina S., Basilakos A., Bonilha L., Gleichgerrcht E., Yourganov G., Hillis A. E., Hickok G., Rorden C., and Fridriksson J. (2019). Cortical and structural-connectivity damage correlated with impaired syntactic processing in aphasia. Hum. Brain Mapp. 40:7, 2153-2173. doi:10.1002/hbm.24514

De Witte E., and Mariën P. (2013). The neurolinguistic approach to awake surgery reviewed. Clin. Neurol. Neurosurg. 115(2), 127–145. doi:10.1016/j.clineuro.2012.09.015

Dodoo-Schittko, F., Rosengarth, K., Doenitz, C., and Greenlee, M. (2012). Assessing language dominance with functional MRI: The role of control tasks and statistical analysis. Neuropsychologia. 50, 2684-2691. doi:10.1016/j.neuropsychologia.2012.07.032

Dragoy, O., Zyryanov, A., Bronov, O., Gordeyeva, E., Gronskaya, N., Kryuchkova, O., Klyuev, E., Kopachev, D., Medyanik, I., Mishnyakova, L., Pedyash, N., Pronin, I., Reutov, A., Sitnikov, A., Stupina, E., Yashin, K., Zhirnova, V., and Zuev, A. (2020). Functional linguistic specificity of the left frontal aslant tract for spontaneous speech fluency: Evidence from intraoperative language mapping. Brain Lang. 208, 104836. doi:10.1016/j.bandl.2020.104836

Duffau, H. (2012). The challenge to remove diffuse low-grade gliomas while preserving brain functions. Acta Neurochir. 154, 569-574. doi:10.1007/s00701-012-1275-7

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2020). Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Ek L., Almkvist O., Wiberg M.K., Stragliotto G., and Smits A. (2010). Early cognitive impairment in a subset of patients with presumed low-grade glioma. Neurocase. 16(6), 503–511. doi:10.1080/13554791003730634

Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt, T.E., Caspi, A., and Hariri, A.R. (2020)/ What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. Psychol. Sci. 31(7), 792-806. doi:10.1177/0956797620916786

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2016). The Human Brainnetome Atlas: A new brain atlas based on connectional architecture. Cereb. Cortex. 26(8), 3508–3526. doi:10.1093/cercor/bhw157

Fedorenko, E., Hsieh, P.J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104:2, 1177-1194. doi:10.1152/jn.00032.2010

Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., and Elger, C. E. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology. 60:6, 969-975. doi:10.1212/01.wnl.0000049934.34209.2e

Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., and Brueckmann, H. (2010). Sensitivity and reliability of language laterality assessment with a free reversed association task - a fMRI study. Eur. Radiol. 20:3, 683-695. doi:10.1007/s00330-009-1602-4

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T., and Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proc. Natl. Acad. Sci. U. S. A. 112:9, 2871–2875. doi:10.1073/pnas.1414491112

Gabel, N., Altshuler, D. B., Brezzell, A., Briceño, E. M., Boileau, N. R., Miklja, Z., Kluin, K., Ferguson, T., McMurray, K., Wang, L., Smith, S. R., Carlozzi, N. E., and Hervey-Jumper, S. L. (2019). Health-related quality of life in adult low and high-grade glioma patients using the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS) and Neuro-QOL assessments. Front. Neurol. 10, 212. doi:10.3389/fneur.2019.00212

Gorgolewski, K., Storkey, A. J., Bastin, M. E., and Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. Front. Hum. Neurosci. 6, 245. doi:10.3389/fnhum.2012.00245

Graves, W. W., Grabowski, T. J., Mehta, S., and Gupta, P. (2008). The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. J. Cogn. Neurosci. 20:9, 1698–1710. doi:10.1162/jocn.2008.20113

Gross, W. L., and Binder, J. R. (2014). Alternative thresholding methods for fMRI data optimized for surgical planning. NeuroImage. 84, 554–561. doi:10.1016/j.neuroimage.2013.08.066

Grummich, P., Nimsky, C., Pauli, E., Buchfelder, M., and Ganslandt, O. (2006). Combining fMRI and MEG increases the reliability of presurgical language localization: a clinical study on the difference between and congruence of both modalities. NeuroImage. 32:4, 1793–1803. doi: 10.1016/j.neuroimage.2006.05.034

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. Trends Cogn. Sci. 9:9, 416-23. doi:10.1016/j.tics.2005.07.004

Hakyemez, B., Erdogan, C., Yildirim, N., Bora, I., Bekar, A., and Parlak, M. (2006). Functional MRI in 507 patients with intracranial lesions near language areas. Neuroradiol. J. 508:19, 306-312. doi:10.1177/197140090601900306

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., Gurney, E. M., and Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7:3, 213-223. doi: 10.1002/(sici)1097-0193(1999)7:3<213::aid-hbm5>3.0.co;2-n

Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks, D. J., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19:4, 224–247. doi:10.1002/hbm.10123

Hervey-Jumper, S.L., Li, J., Lau, D., Molinaro, A.M., Perry, D.W., Meng, L., and Berger, M.S. (2015). Awake craniotomy to maximize glioma resection: methods and technical nuances over a 27-year period. J. Neurosurg. 123(2), 325-339. doi:10.3171/2014.10.JNS141520

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402. doi:10.1038/nrn2113

Hilari, K., Wiggins, R., Roy, P., Byng, S., and Smith, S. (2003). Predictors of health-related quality of life (HRQL) in people with chronic aphasia. Aphasiology. 17:4, 365-381. doi:10.1080/02687030244000725

Hillis A. E., Rorden C., and Fridriksson, J. (2017). Brain regions essential for word comprehension: Drawing inferences from patients. Ann. Neurol. 81:6, 759–768. doi:10.1002/ana.24941

Holiga, Š., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R. J., Marsman, J.-B. C., Schobel, S. A., Bertolino, A., and Dukart, J. (2018). Test-retest reliability of task-based and resting-state blood oxygen level dependence and cerebral blood flow measures. PLoS ONE. 13:11, e0206583. doi:10.1371/journal.pone.0206583

992 Huettel, S., Song, A., and McCarthy, G. (2008). Functional Magnetic Resonance Imaging, Second
993 Edition. Sunderland, MA: Sinauer Associates Inc.
994 Hund-Georgiadis, M., Lex, U., and von Cramon, D. Y. (2001). Language dominance assessment by
995 means of fMRI: Contributions from task design, performance, and stimulus modality. J. Magn.
996 Reson. Imaging. 13:5, 668-675. doi:10.1002/jmri.1094
997 Ille, S., Sollmann, N., Hauck, T., Maurer, S., Tanigawa, N., Obermueller, T., Negwer, C., Droese, D.,
998 Zimmer, C., Meyer, B., Ringel, F., and Krieg, S.M. (2015). Combined noninvasive language
999 mapping by navigated transcranial magnetic stimulation and functional MRI and its comparison
1000 with direct cortical stimulation. J. Neurosurg. 123:1, 212-225. doi:10.3171/2014.9.JNS14929
1001 Jansen, A., Menke, R., Sommer, J., Förster, A.F., Bruchmann, S., Hempleman, J., Weber, B., and
1002 Knecht, S. (2006). The assessment of hemispheric lateralization in functional MRI—Robustness
1003 and reproducibility. NeuroImage. 33:1, 204-217. doi:10.1016/j.neuroimage.2006.06.019
1004 Jones, S.E., Mahmoud, S.Y., and Phillips, M.D. (2011). A practical clinical method to quantify
1005 language lateralization in fMRI using whole-brain analysis. Neuroimage 54:4, 2937-2949.
1006 doi:10.1016/j.neuroimage.2010.10.052
1007 Kinno, R., Ohta, S., Muragaki, Y., Maruyama, T., and Sakai, K. L. (2014). Differential
1008 reorganization of three syntax-related networks induced by a left frontal glioma. Brain. 137:4,
1009 1193–1212. doi:10.1093/brain/awu013
1010 Kinoshita, M., de Champfleur, N. M., Deverdun, J., Moritz-Gasser, S., Herbet, G., and Duffau, H.
1011 (2015). Role of fronto-striatal tract and frontal aslant tract in movement and speech: an axonal
1012 mapping study. Brain Struct. Funct. 220:6, 3399–3412. doi:10.1007/s00429-014-0863-0
1013 Knecht, S., Deppe, M., Dräger, B., Bobe, L., Lohmann, H., Ringelstein, E., and Henningsen, H.
1014 (2000). Language lateralization in healthy right-handers. Brain. 123:1, 74-81.
1015 doi:10.1093/brain/123.1.74
1016 Kozák, L. R., van Graan, L. A., Chaudhary, U. J., Szabó, Á. G., and Lemieux, L. (2017). ICN_Atlas:
1017 Automated description and quantification of functional MRI activation patterns in the framework
1018 of intrinsic connectivity networks. NeuroImage. 163, 319–341.
1019 doi:10.1016/j.neuroimage.2017.09.014
1020 Lehéricy, S., Cohen, L., Bazin, B., Samson, S., Giacomini, E., Rougetet, R., Hertz-Pannier, L., Le
1021 Bihan, D., Marsault, C., and Baulac, M. (2000). Functional MR evaluation of temporal and
1022 frontal language dominance compared with the Wada test. Neurology. 54:8, 1625-1633.
1023 doi:10.1212/WNL.54.8.1625
1024 Litvinova, L., Pechenkova, E., Vlasova, R., Berezutskaya, Y., and Sinitsyn, V. (2012). “Lokalizatsiya
1025 zon, svyazannyh s vospriyatiem rechi: sopostavlenie tryokh prob dlya fMRT na material
1026 russkogo yazyka. [A comparison of three fMRI paradigms for mapping speech perception in
1027 Russian speakers],” in International Symposium on Functional Neuroimaging: Basic Research
1028 and Clinical Applications. Abstracts, ed. S. Novikova (Moscow: MSUPE), 76-79.
1029 Loddenkemper, T., Morris, H.H., and Möddel, G. (2008). Complications during the Wada test.
1030 Epilepsy Behav. 13:3, 551-553. doi:10.1016/j.yebeh.2008.05.014.
1031 Loring, D.W., Meador, K.J., and Lee, G.P. (1992). “Criteria and validity issues in Wada assessment,”
1032 in The neuropsychology of epilepsy, ed. I. Benett (New York: Plenum Press), 233-245.
1033 Manan, H. A., Franz, E.A., and Yahya, N. (2020). Utilization of functional MRI language paradigms
1034 for pre-operative mapping: a systematic review. Neuroradiology. 62, 353–367.
1035 doi:10.1007/s00234-019-02322-w
1036 Mano Q. R., Humphries C., Desai R. H., Seidenberg M. S., Osmon D. C., Stengel B. C., and Binder
1037 J. R. (2013). The role of left occipitotemporal cortex in reading: Reconciling stimulus, task, and
1038 lexicality effects. Cereb. Cortex. 23:4, 988–1001. doi:10.1093/cercor/bhs093

Mauler, J., Neuner, I., Neuloh, G., Fimm, B., Boers, F., Wiesmann, M., Clusmann, H., Langen, K.J., and Shah, N.J. (2017). Dissociated crossed speech areas in a tumour patient. Case Rep. Neurol. 9:2, 131–136. doi:10.1159/000475882

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development. Neuroimage. 2:2, 89–101. doi: 10.1006/nimg.1995.1012

Miró, J., Ripollés, P., López-Barroso, D., Vilà-Balló, A., Juncadella, M., de Diego-Balaguer, R., Marco-Pallares, J., Rodríguez-Fornells, A., and Falip, M. (2014). Atypical language organization in temporal lobe epilepsy revealed by a passive semantic paradigm. BMC Neurol. 14, 98. doi:10.1186/1471-2377-14-98

Morrison, M. A., Churchill, N. W., Cusimano, M. D., Schweizer, T. A., Das S., and Graham, S. J. (2016a). Reliability of task-based fMRI for preoperative planning: A test-retest study in brain tumor patients and healthy controls. PLoS ONE. 11(2), e0149547. doi:10.1371/journal.pone.0149547

Morrison, M. A., Tam, F., Garavaglia, M. M., Hare, G. M. T., Cusimano, M. D., Schweizer, T. A., Das, S., and Graham, S.J. (2016b). Sources of variation influencing concordance between functional MRI and direct cortical stimulation in brain tumor surgery. Front. Neurosci. 10, 461. doi:10.3389/fnins.2016.00461

Nadkarni T. N., Andreoli M. J., Nair V. A., Yin P., Young B. M., Kundu B., Pankratz J., Radtke A., Holdsworth R., Kuo J. S., Field A. S., Baskaya M. K., Moritz C. H., Meyerand M. E., and Prabhakaran V. (2014). Usage of fMRI for pre-surgical planning in brain tumor and vascular lesion patients: task and statistical threshold effects on language lateralization. Neuroimage Clin. 7, 415-423. doi:10.1016/j.nicl.2014.12.014

Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., and Weiß Lucas, C. (2018). Short- and long-term reliability of language fMRI. Neuroimage. 176, 215-225. doi:10.1016/j.neuroimage.2018.04.050

Newman, S. D., Twieg, D. B., and Carpenter, P. A. (2001). Baseline conditions and subtractive logic in neuroimaging. Hum. Brain Mapp. 14:4, 228–235. doi:10.1002/hbm.1055

Ojemann, G. A. (1979). Individual variability in cortical localization of language. J. Neurosurg. 50:2, 164-169. doi:10.3171/jns.1979.50.2.0164

Ojemann, G.A. (1993). Functional mapping of cortical language areas in adults. Intraoperative approaches. Adv. Neurol. 63, 155–163.

Partovi, S., Jacobi, B., Rapps, N., Zipp, L., Karimi, S., Rengier, F., Lyo, J. K., and Stippich, C. (2012). Clinical standardized fMRI reveals altered language lateralization in patients with brain tumor. Am. J. Neuroradiol. 33, 2151–2157. doi:10.3174/ajnr.A3137

Patrikelis P., Gatzonis S., Siatouni A., Angelopoulos E., Konstantakopoulos G., Takousi M., Sakas D. E., and Zalonis I. (2016). Preoperative neuropsychological presentation of patients with refractory frontal lobe epilepsy. Acta Neurochir. 158, 1139–1150. doi:10.1007/s00701-016-2786-4

Pearson, J., Naselaris, T., Holmes, E. A., and Kosslyn, S. M. (2015). Mental imagery: Functional mechanisms and clinical applications. Trends Cogn. Sci. 19:10, 590–602. doi:10.1016/j.tics.2015.08.003

Peelle J. E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is: A hierarchical perspective. Front. Hum. Neurosci. 6, 309. doi:10.3389/fnhum.2012.00309

Picht, T., Krieg, S.M., Sollmann, N., Rösler, J., Niraula, B., Neuvonen, T., Savolainen, P., Lioumis, P., Mäkelä, J.P., Deletis, V., Meyer, B., Vajkoczy, P., and Ringel, F. (2013). A comparison of language mapping by preoperative navigated transcranial magnetic stimulation and direct cortical stimulation during awake surgery. Neurosurgery. 72:5, 808-819. doi:10.1227/NEU.0b013e3182889e01

Pillai, J. J., and Zaca, D. (2011). Relative utility for hemispheric lateralization of different clinical fMRI activation tasks within a comprehensive language paradigm battery in brain tumor patients as assessed by both threshold-dependent and threshold-independent analysis methods. NeuroImage. 54:S1, S136-S145.

Połczyńska, M., Japardi, K., Curtiss, S., Moody, T., Benjamin, C., Cho, A, Vigil, C., Kuhn, T., Jones, M., and Bookheimer, S. (2017). Improving language mapping in clinical fMRI through assessment of grammar. Neuroimage Clin. 15, 415-427. doi:10.1016/j.nicl.2017.05.021

Pouratian, N., Bookheimer, S.Y., Rex, D.E., Martin, N.A., and Toga, A.W. (2002). Utility of preoperative functional magnetic resonance imaging for identifying language cortices in patients with vascular malformations. J. Neurosurg. 97:1, 21–32. doi:10.3171/jns.2002.97.1.0021

Racine C.A., Li J., Molinaro A.M., Butowski N., and Berger M.S. (2015). Neurocognitive function in newly diagnosed low-grade glioma patients undergoing surgical resection with awake mapping techniques. Neurosurgery. 77:3, 371-379. doi:10.1227/NEU.0000000000000779

Richardson F. M., Seghier M. L., Leff A. P., Thomas M. S., and Price C. J. (2011). Multiple routes from occipital to temporal cortices during reading. J. Neurosci. 31:22, 8239–8247. doi:10.1523/JNEUROSCI.6519-10.2011

Robinson, G., Shallice, T., Bozzali, M., and Cipolotti, L. (2010). Conceptual proposition selection and the LIFG: neuropsychological evidence from a focal frontal group. Neuropsychologia. 48:6, 1652–1663. doi:10.1016/j. neuropsychologia.2010.02.010

Rodd, J. M., Davis, M. H., and Johnsrude, I.S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. Cereb. Cortex. 15, 1261–1269. doi:10.1093/cercor/bhi009

Rofes, A., Mandonnet, E., de Aguiar, V., Rapp, B., Tsapkini, K., and Miceli, G. (2019). Language processing from the perspective of electrical stimulation mapping. Cogn. Neuropsychol. 36:3-4, 117-139. doi:10.1080/02643294.2018.1485636

Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., Valk, J., and Scheltens, P. (1997). Test-retest analysis with functional MR of the activated area in the human visual cortex. Am. J. Neuroradiol. 18:7, 1317–1322.

Roux, F.E., Boulanouar, K., Lotterie, J.A., Mejdoubi, M., LeSage, J. P., and Berry, I. (2003). Language functional magnetic resonance imaging in preoperative assessment of language areas: correlation with direct cortical stimulation. Neurosurgery. 52:6, 1335–1347. doi:10.1227/01.neu.0000064803.05077.40

Ruge, M.I., Victor, J., Hosain, S., Correa, D.D., Relkin, N.R., Tabar, V., Brennan, C., Gutin, P.H., and Hirsch, J. (1999). Concordance between functional magnetic resonance imaging and intraoperative language mapping. Stereotact. Funct. Neurosurg. 72:2–4, 95–102. doi: 10.1159/000029706

Rumshiskaya, A., Vlasova, R., Litvinova, L., Pechenkova, E., and Mershina, E. (2014). Combined analysis of two tasks improves localization of Wernicke's area in patients with primary brain tumors. Poster presented at the European Society for Neuroradiology. doi:10.1594/ecr2014/C-1232

Rutten, G.J.M., Ramsey, N.F., Van Rijen, P.C., Alpherts, W.C., and Van Veelen, C.W.M. (2002). fMRI-determined language lateralization in patients with unilateral or mixed language dominance according to the Wada test. NeuroImage. 17:1, 447–460. doi:10.1006/nimg.2002.1196

Salek, K.E., Hassan, I.S., Kotrotsou, A., Abrol, S., Faro, S. H., Mohamed, F. B., Zinn, P. O., Wei, W., Li, N., Kumar, A. J., Weinberg, J. S., Wefel, J. S., Kesler, S. R., Liu, H.-L. A., Hou, P., Stafford, R. J., Prabhu, S., Sawaya, R., and Colen, R. R. (2017). Silent sentence completion shows superiority localizing Wernicke's area and activation patterns of distinct language

paradigms correlate with genomics: prospective study. Sci. Rep. 7, 12054. doi:10.1038/s41598-017-11192-2

Sanjuan, A., Bustamante, J.-C., Forn, C., Ventura-Campos, N., Barros-Loscertales, A., Martinez J-C, Villanueva V, and Avila C. (2010). Comparison of two fMRI tasks for the evaluation of the expressive language function. Neuroradiology 52:5, 407–415. doi:10.1007/s00234-010-0667-8

Satoer, D., Visch-Brink, E., Dirven, C., and Vincent, A. (2016). Glioma surgery in eloquent areas: can we preserve cognition? Acta Neurochi. 158:1, 35–50. doi:10.1007/s00701-015-2601-7

Silva, M. A., See, A. P., Essayed, W. I., Golby, A. J., and Tie, Y. (2018). Challenges and techniques for presurgical brain mapping with functional MRI. NeuroImage Clin. 17, 794–803. doi:10.1016/j.nicl.2017.12.008

Sollmann, N., Kubitscheck, A., Maurer, S., Ille, S., Hauck, T., Kirschke, J. S., Ringel, F., Meyer, B., and Krieg, S.M. (2016). Preoperative language mapping by repetitive navigated transcranial magnetic stimulation and diffusion tensor imaging fiber tracking and their comparison to intraoperative stimulation. Neuroradiology. 58, 807–818. doi:10.1007/s00234-016-1685-y

Spena G., Nava A., Cassini F., Pepoli A., Bruno M., D'Agata F., Cauda F., Sacco K., Duca S., Barletta L., and Versari P. (2010). Preoperative and intraoperative brain mapping for the resection of eloquent-area tumors. A prospective analysis of methodology, correlation, and usefulness based on clinical outcomes. Acta Neurochir. 152:11, 1835-46. doi:10.1007/s00701-010-0764-9

Springer, J. A., Binder, J. R., Hammeke, T. A., Swanson, S. J., Frost, J. A., Bellgowan, P. S., Brewer, C. C., Perry, H. M., Morris, G. L., and Mueller, W. M. (1999). Language dominance in neurologically normal and epilepsy subjects: a functional MRI study. Brain. 122:11, 2033-2046. doi:10.1093/brain/122.11.2033

Stevens, M. T. R., D'Arcy, R. C. N., Stroink,, C. D. B., and Beyea, S. D. (2013). Thresholds in fMRI studies: Reliable for single subjects? J. Neurosci. Methods. 219, 312–323. doi:10.1016/j.jneumeth.2013

Stevens, M. T., Clarke, D. B., Stroink, G., Beyea, S. D., and D'Arcy, R. C. (2016). Improving fMRI reliability in presurgical mapping for brain tumours. J. Neurol. Neurosurg. Psychiatry. 87:3, 267-274. doi:10.1136/jnnp-2015-310307

Stippich, C., Rapps, N., Dreyhaupt, J., Durst, A., Kress, B., Nennig, E., Tronnier V. M., and Sartor, K. (2007). Localizing and lateralizing language in patients with brain tumors: Feasibility of routine preoperative functional MR imaging in 81 consecutive patients. Radiology. 243:3, 828–836. doi:10.1148/radiol.2433060068

Stoppelman, N., Harpaz, T., and Ben-Shachar, M. (2013). Do not throw out the baby with the bath water: choosing an effective baseline for a functional localizer of speech processing. Brain Behav. 3:3, 211-222. doi:10.1002/brb3.129

Suarez, R.O., Taimouri, V., Boyer, K., Vega, C., Rotenberg, A., Madsen, J. R., Loddenkemper, T., Duffy, F, Prabhu, S., and Warfield, S. K. (2014). Passive fMRI mapping of language function for pediatric epilepsy surgical planning: Validation using Wada, ECS, and FMAER. Epilepsy Res. 108:10, 1874–1888. doi:10.1016/j.eplepsyres.2014.09.016

Szaflarski, J.P., Holland, S.K., Jacola, L.M., Lindsell, C., Privitera, M.D., and Szaflarski, M. (2008). Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy. Epilepsy Behav. 12, 74–83. doi:10.1016/j.yebeh.2007.07.015

Thivard, L., Hombrouck, J., Du Montcel, S.T., Delmaire, C., Cohen, L., Samson, S., Dupont, S., Chiras, J., Baulac, M., and Lehericy, S. (2005). Productive and perceptive language reorganization in temporal lobe epilepsy. NeuroImage. 24:3, 841–851. doi:10.1016/j.neuroimage.2004.10.001.

Unadkat, P., Fumagalli, L., Rigolo, L., Vangel, M. G., Young, G. S., Huang, R., Mukundan, S. Jr., Golby, A., and Tie, Y. (2019). Functional MRI task comparison for language mapping in neurosurgical patients. J. Neuroimaging 29:3, 348-356. doi:10.1111/jon.12597

Van Poppel, M., Wheless, J.W., Clarke, D.F., McGregor, A., McManis, M.H., Perkins, F.F. Jr., Van Poppel, K., Fulton, S., and Boop, F.A. (2012). Passive language mapping with magnetoencephalography in pediatric patients with epilepsy. J. Neurosurg. Pediatr. 10:2, 96-102. doi:10.3171/2012.4.PEDS11301

Voyvodic J. T., Petrella J. R., and Friedman A. H. (2009). fMRI activation mapping as a percentage of local excitation: consistent presurgical motor maps without threshold adjustment. J. Magn. Res. Imaging. 29:4, 751-759. doi:10.1002/jmri.21716

Wada, J., and Rasmussen, T. (1960). Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance. J. Neurosurg. 17, 266–282.

Weng, H.H., Noll, K.R., Johnson, J.M., Prabhu, S.S., Tsai, Y.H., Chang, S.W., Huang, Y.C., Lee, J.D., Yang, J.T., Yang, C.T., Tsai, Y.H., Yang, C.Y., Hazle, J.D., Schomer, D.F., and Liu, H.L. (2018). Accuracy of presurgical functional MR imaging for language mapping of brain tumors: A systematic review and meta-analysis. Radiology. 286:2, 512-523. doi:10.1148/radiol.2017162971.

Whalley, H. C., Gountouna, V. E., Hall, J., McIntosh, A. M., Simonotto, E., Job, D. E., Owens, D. G., Johnstone, E. C., and Lawrie, S. M. (2009). fMRI changes over time and reproducibility in unmedicated subjects at high genetic risk of schizophrenia. Psychol. Med. 39:7, 1189-99. doi:10.1017/S0033291708004923

Wilke, M., and Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. J. Neurosci. Methods. 163:1, 128–136. doi:10.1016/j.jneumeth.2007.01.026

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., and Eriksson, D. K. (2016). Validity and reliability of four language mapping paradigms. NeuroImage Clin. 16, 399–408. doi:10.1016/j.nicl.2016.03.015

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31:3, 1116–1128. doi:10.1016/j.neuroimage.2006.01.015

Zacà, D., Nickerson, J.P., Deib, G., and Pillai, J. J. (2012). Effectiveness of four different clinical fMRI paradigms for preoperative regional determination of language lateralization in patients with brain tumors. Neuroradiology. 54, 1015–1025. doi:10.1007/s00234-012-1056-2

Zhang, N., Xia, M., Qiu, T., Wang, X., Lin, C.P., Guo, Q., Lu, J., Wu, Q., Zhuang, D., Yu, Z., Gong, F., Farrukh Hameed, N. U., He, Y., Wu, J., and Zhou, L. (2018). Reorganization of cerebro-cerebellar circuit in patients with left hemispheric gliomas involving language network: A combined structural and resting-state functional MRI study. Hum. Brain Mapp. 39:12, 4802-4819. doi:10.1002/hbm.24324

Zyryanov A., Malyutina S., and Dragoy O. (2020). Left frontal aslant tract and lexical selection: Evidence from frontal lobe lesions. Neuropsychologia. 147, 107385. doi:10.1016/j.neuropsychologia.2020.107385

Tables

Table 1. Dice coefficients in the two paradigms (SYLL: syllable baseline vs. PW: pseudoword baseline) in three regions (frontal, temporal-parietal and frontal-temporal-parietal) at three statistical thresholds (FWE correction for multiple comparisons at $p < .05$, cluster-size correction with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$, AT: adaptive thresholding as implemented in Gorgolewski et al. (2012) at $p < .05$).

Dice coefficients						
Frontal						
	FWE		AT		Cluster-Size	
Baseline	SYLL	PW	SYLL	PW	SYLL	PW
Mean	.49	.56	.43	.59	.56	.61
SD	.11	.12	.20	.15	.09	.12
Min	.23	.31	.02	.20	.42	.34
Max	.67	.76	.69	.78	.69	.79
Temporal + Parietal						
	FWE		AT		Cluster-Size	
Baseline	SYLL	PW	SYLL	PW	SYLL	PW
Mean	.41	.47	.39	.52	.54	.58
SD	.16	.15	.21	.16	.07	.10
Min	.14	.23	.00	.15	.39	.42
Max	.76	.71	.76	.74	.65	.71
Frontal + Temporal + Parietal						
	FWE		AT		Cluster-Size	
Baseline	SYLL	PW	SYLL	PW	SYLL	PW
Mean	.49	.54	.43	.56	.42	.51
SD	.09	.11	.20	.17	.15	.12
Min	.31	.33	.01	.17	.15	.26
Max	.66	.70	.68	.74	.75	.71

Table 2. Lateralization indices for each paradigm (SYLL: syllable baseline vs. PW: pseudoword baseline) in three regions (frontal, temporal-parietal, frontal-temporal-parietal) along with results of Spearman's correlations between LIs in the two scanning sessions. Significant correlations ($p < .05$) are marked with *.

Lateralization indices				
Frontal				
	SYLL		PW	
Session	Session 1	Session 2	Session 1	Session 2
Mean	.46	.49	.50	.51
SD	.18	.20	.16	.14
Min	.13	.26	.25	.32
Max	.80	.85	.88	.85
Spearman's correlation	$r = .447$, $p = .063$		$r = .496$, $p = .036$*	
Temporal + Parietal				
	SYLL		PW	
Session	Session 1	Session 2	Session 1	Session 2
Mean	.33	.40	.38	.32
SD	.17	.20	.19	.16
Min	-.04	.01	.10	.08
Max	.57	.75	.88	.67
Spearman's correlation	$r = .382$, $p = .117$		$r = .603$, $p = .009$*	
Frontal +Temporal + Parietal				
	SYLL		PW	
Session	Session 1	Session 2	Session 1	Session 2
Mean	.42	.46	.45	.44
SD	.17	.18	.17	.14
Min	.08	.20	.22	.26
Max	.70	.84	.84	.77
Spearman's correlation	$r = .401$, $p = .099$		$r = .490$, $p = .038$*	

Table 3. Comparison of Dice coefficients to previous fMRI paradigms using language tasks in neurologically healthy participants.

Study	Task	Dice coefficients	Comment
The present study	Overt sentence completion	.39 to .61	Group averages depending on the region, baseline and statistical threshold
Fesl et al. (2010)	Free reversed association task	.61	Group average in the global defined language network
Wilson et al. (2016)	Picture naming	.38 to .61	Group averages in the 'supratentorial region', depending on statistical threshold
	Naturalistic comprehension	.30 to .51	
	Narrative comprehension	.07 to .37	
	Sentence completion	.27 to .47	
Morrison et al. (2016a)	Phonemic fluency	.36	Group average, whole-brain
	Rhyming	.54	
Nettekoven et al. (2018)	Picture naming	.47 or .60	Group average, depending on statistical threshold

Figures

Figure 1. Experimental design of the functional paradigms.

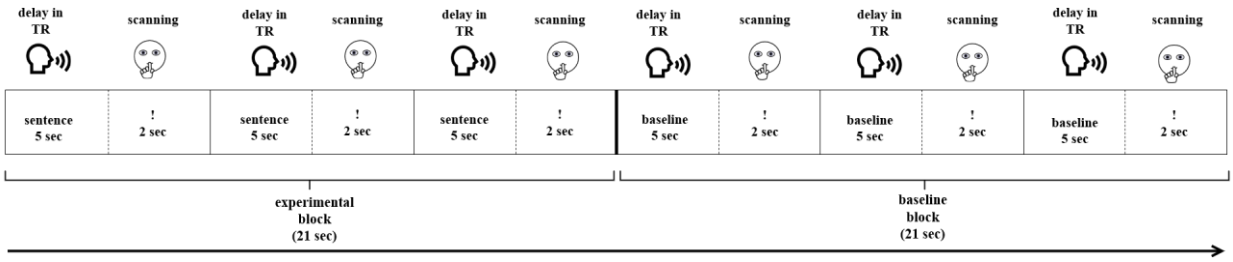

Figure 2. Language-related activation in the SYLL paradigm (paradigm with the syllable baseline). Top: FWE correction for multiple comparisons at $p < .05$, middle: adaptive thresholding (AT) as implemented in Gorgolewski et al. (2012) at $p < .05$, bottom: cluster-size correction for multiple comparisons with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$.

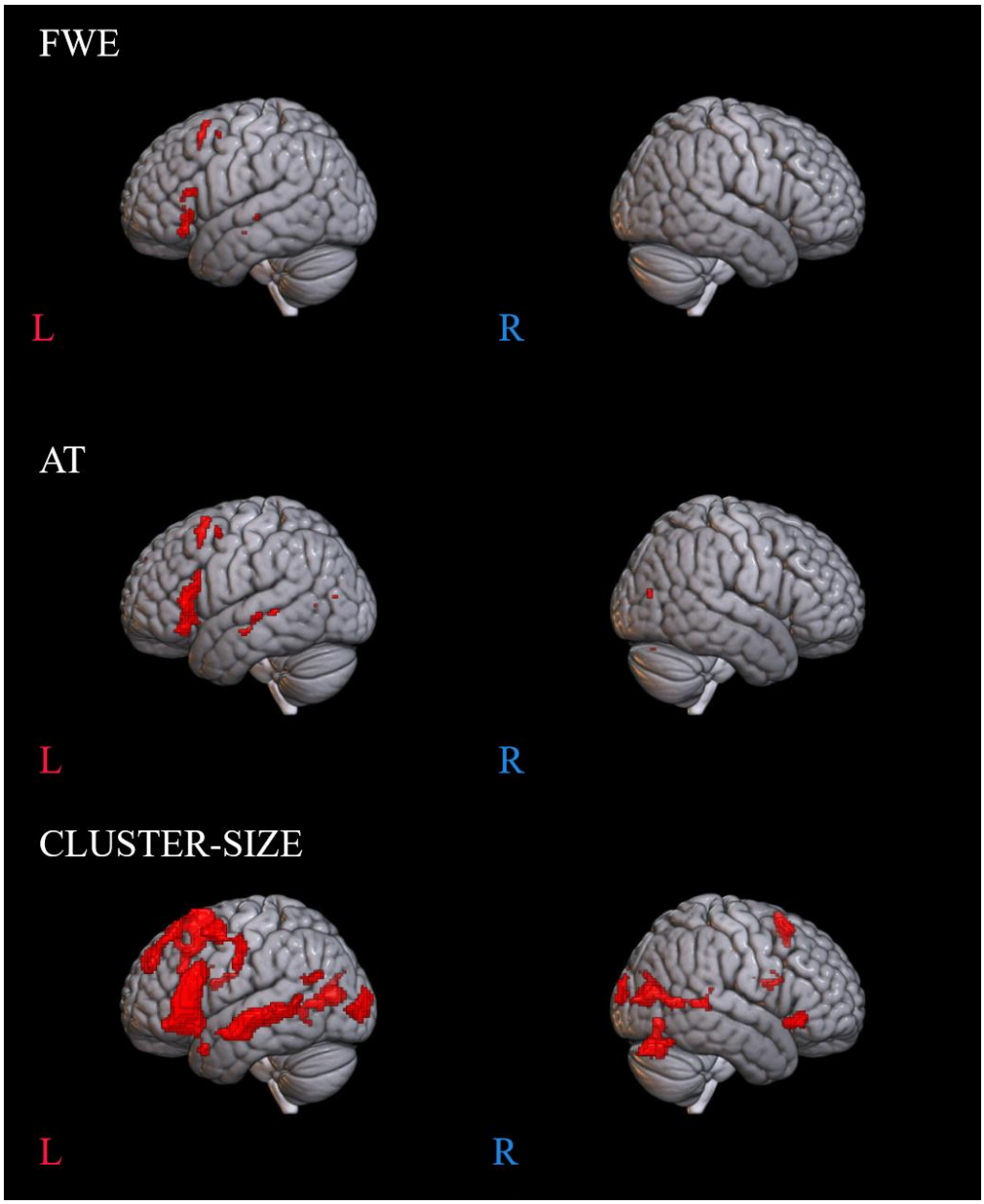

Figure 3. Language-related activation in the PW paradigm (paradigm with the pseudoword baseline). Top: FWE correction for multiple comparisons at $p < .05$, middle: adaptive thresholding (AT) as implemented in Gorgolewski et al. (2012) at $p < .05$, bottom : cluster-size correction for multiple comparisons with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$.

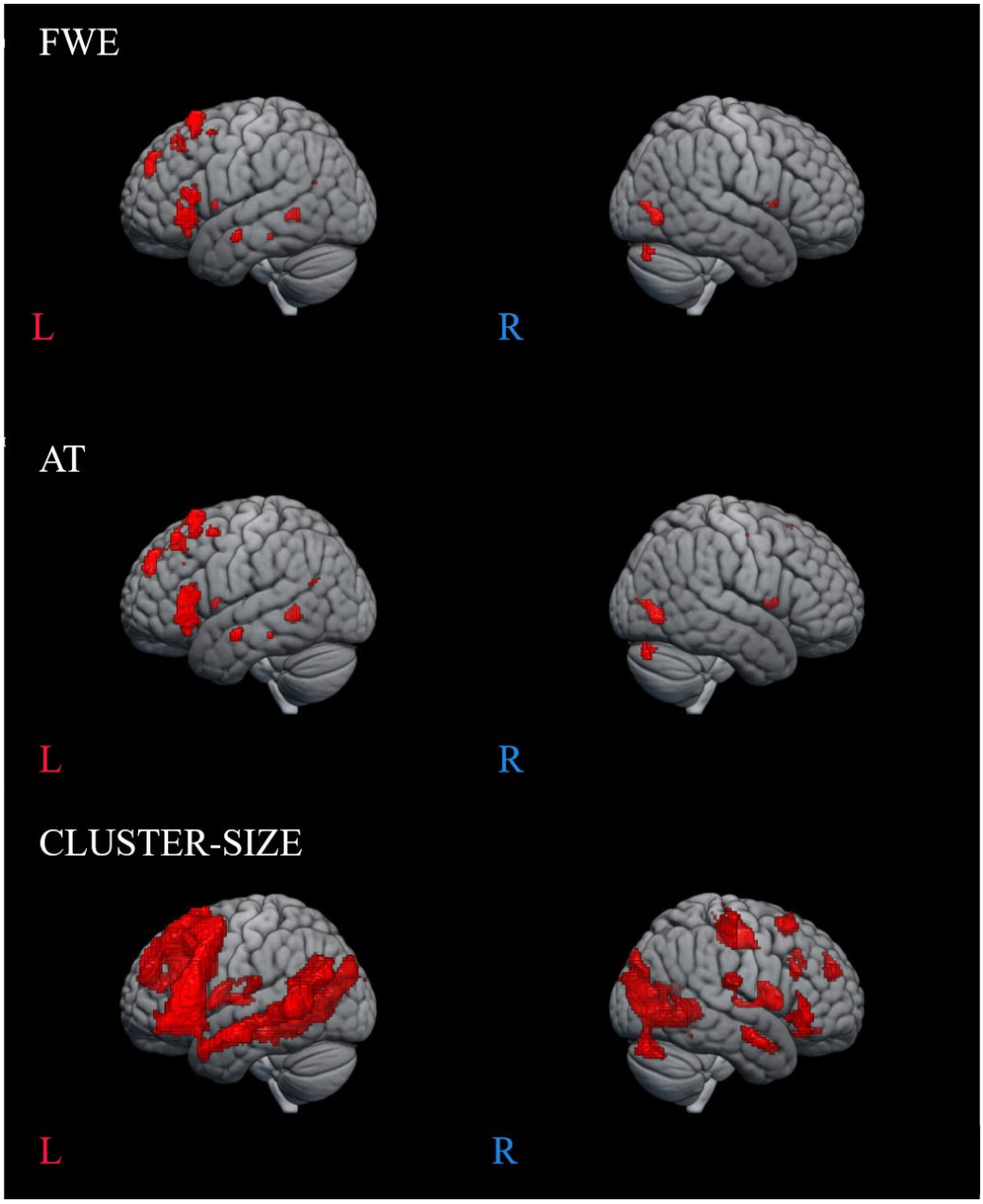

Figure 4. Percentage of significantly activated voxels in 'key language-related areas' adopted from Benjamin et al. (2017) depending on the paradigm (PW – pseudoword baseline, SYLL – syllable baseline) and threshold (FWE correction for multiple comparisons at $p < .05$, cluster-size correction with a minimum cluster size of $k \geq 200$ mm^3 at $p < .001$, AT: adaptive thresholding as implemented in Gorgolewski et al. (2012) at $p < .05$). The white dot in the middle of each 'violin' represents the median value and the thick black bar in the center represents the interquartile range.