

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

1 Kirill Elin^{1†}, Svetlana Malyutina^{1*}, Oleg Bronov², Ekaterina Stupina¹, Aleksei Marinets², Olga
2 Dragoy^{1,3}

3 ¹Center for Language and Brain, HSE University, Moscow, Russia

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

6 ³Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

7 † The authors contributed equally to this manuscript.

8 * **Correspondence:**

9 Svetlana Malyutina
10 smalyutina@hse.ru

11 **Keywords:** functional magnetic resonance imaging¹, presurgical mapping², language mapping³,
12 language localizer paradigm⁴, test-retest reliability of fMRI⁵

13 Abstract

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

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

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

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

37 **1 Introduction**

38 **1.1 Language mapping in neurosurgical patients**

39 When patients undergo neurosurgical interventions for brain tumors, refractory epilepsy,
40 arteriovenous malformations et cetera, a crucial goal is to remove pathological tissue while sparing
41 eloquent (functionally necessary) areas, so that the respective functions, including cognitive ones
42 (Satoer et al., 2016), are not impaired following neurosurgery (Duffau, 2012). One critical function is
43 language processing: it lies at the core of human communication, and its impairment negatively
44 impacts return to work, social inclusion, and general quality of life (Gabel et al., 2019, Hilari et al.,
45 2003).

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

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

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

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

90 **1.2 Choice of a language task for fMRI mapping**

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

96 Most ‘language localizers’ have used single-word tasks, particularly expressive single-word tasks
97 (Manan et al., 2020). A traditional expressive single-word task adapted from early intraoperative
98 batteries (Ojemann, 1993, Ruge et al., 1999) was number counting. However, counting is a highly
99 automated process that engages linguistic processing only superficially and is no longer considered
100 sufficiently sensitive to identify language networks (Morrison et al., 2016b; Petrovich Brennan et al.,
101 2007). Today, other tasks are used to engage word retrieval, such as picture naming (Petrovich
102 Brennan et al., 2007, Pouratian et al., 2002, Roux et al., 2003, Rutten et al., 2002) and verbal fluency,
103 where the participant has to name as many words as possible from a given semantic category or
104 starting with a given letter (Ruff et al., 2008 Sanjuan et al., 2010). Expressive single-word tasks are
105 intuitive for the participant and are easily timed relative to fMRI scanning. Still, they only engage the
106 sound and word levels of language production, leaving out any grammatical processing and any
107 language comprehension. Thus, they cannot fully identify brain areas that are crucial for sentence-
108 level communication. Indeed, compared to sentence-level tasks, expressive single-word tasks elicit
109 less activation in both anterior and posterior left-hemispheric ‘language regions’ (Połczyńska et al.,
110 2017, Manan et al., 2020). Additionally, picture naming shows a poor lateralizing ability (Bradshaw
111 et al., 2017, Deblaere et al., 2002).

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

123 To more fully activate language networks, a number of fMRI protocols used sentence-level tasks
124 requiring to process not only individual words but also their grammatical and semantic relations.
125 Examples of expressive sentence-level tasks are describing a picture with a sentence (Mauler et al.,
126 2017, Partovi et al., 2012) or generating a sentence with given words (Hakyemez et al., 2016). Such
127 tasks appear to successfully activate both anterior and posterior language areas (Mauler et al., 2017,
128 Partovi et al., 2012, Hakyemez et al., 2016). However, they are taxing for the patient and difficult to
129 time relative to fMRI scanning, especially given interindividual variability in task completion speed
130 across patients, so they are not widely used.

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

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

151 One such comprehensive task is sentence completion, advocated in a recent white paper of the
152 American Society of Functional Neuroradiology (Black et al., 2017) and a metaanalysis by Manan et
153 al. (2020). In this task, the participant has to read aloud a sentence with a missing final word and
154 complete it with a semantically and grammatically appropriate word. This task comprehensively
155 involves many linguistic processes in both comprehension (orthographic processing, word access,
156 grammatical parsing, semantic integration) and production (word search, grammatical inflection in
157 morphologically complex languages such as Russian, phonological encoding and articulation).
158 Empirically, previous works proved sentence completion superior to other tasks in assessing both
159 lateralization and localization of language processing networks (Salek et al., 2017, Połczyńska et al.,
160 2017, Zacà et al., 2012, Barnett et al., 2014, Wilson et al., 2017, Unadkat et al., 2019).

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

167 **1.3 Choice of a baseline task for fMRI mapping**

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

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

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

196 To the best of our knowledge, our study is the first to make an empirical comparison between two
197 different baselines for the sentence completion task. These are a syllable baseline, where the
198 participant has to read aloud a sequence consisting of the same syllable and repeat the syllable once
199 more, and a pseudoword baseline, where the participant has to read aloud a sequence of pseudowords
200 and repeat any of them once. Both baselines are theoretically plausible. In contrast to the
201 experimental condition, they do not consist of real words or resemble grammatical structures, so they
202 do not elicit any linguistic processing. At the same time, they involve the same ‘lower-level’
203 processes as the experimental condition: visual and orthographic processing, motor planning and
204 articulation, and initiation of a response (completion of a sequence). Thus, their subtraction allows to
205 maximally isolate language-related from ‘lower-level’ neural activity. Previous sentence completion
206 paradigms used other baselines that subtracted ‘lower-level’ activity less fully (rest: Połczyńska et al.,
207 2017, passive viewing of nonsense symbols: Barnett et al., 2014, Zacà et al., 2012) or a conjunction
208 analysis approach without an explicit baseline (Wilson et al., 2017). The present study is the first to
209 employ and compare a syllable and pseudoword baseline for the sentence completion task.

210 **1.4 Reliability of fMRI mapping**

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

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

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

237 Previous studies used different metrics to quantify test-retest reliability: Dice coefficient (Fesl et al.,
238 2010, Morrison et al., 2016a, Nettekoven et al., 2018, Wilson et al., 2016), Jaccard index (Morrison
239 et al., 2016a), Euclidean distance (Morrison et al., 2016a, Nettekoven et al., 2018), voxelwise
240 intraclass correlation coefficient (Fernández et al., 2003, Nettekoven et al., 2018, Whalley et al.,
241 2009), correlation of lateralization indices (LIs; Morrison et al., 2016a, Fesl et al., 2010). The present
242 study adopted two of them: between-session Dice coefficient and correlation of lateralization indices.
243 Advantages of the Dice coefficient are that it is widely used in the literature, is straightforward to
244 interpret and, unlike intraclass correlation coefficient, can provide a global whole-brain measure and
245 is calculated individually with no reference to group data (Bennett & Miller, 2010, Wilson et al.,
246 2016). Besides, using a Dice coefficient ensures comparability to Wilson et al. (2016), the only
247 previous study that included a sentence completion task and compared its reliability to other tasks.
248 Additionally, we measured the test-retest correlation of LIs because, just as individual activation
249 maps, they also present a clinically relevant measure that can inform a neurosurgeon’s decision on
250 the necessity of awake surgery.

251

252 **1.5 The present study**

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

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

261 **2 Method**

262 **2.1 Participants**

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

270 **2.2 Task and Stimuli**

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

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

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

287 *Умная соседка прочла*

288 *A clever neighbour read...*

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

290 *Скрипачка сдала сложный ...*

291 *A violinist passed a challenging ...*

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

293 *Грабитель ловко украл ...*

294 *A thief skillfully stole...*

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

302 The sentences were selected from a set of 160 sentences tested in an online pilot study, where 100
303 participants (50 females, age: mean 38.3, SD 11.5, range 18–68 years) read the sentences and
304 finished them aloud with a semantically plausible word within five seconds, matching the task timing
305 in the fMRI study. Their responses were recorded and scored for accuracy by a single rater. A
306 response was considered accurate if it was a semantically plausible single word in a grammatically
307 correct form (direct object). Based on the results, 120 sentences were selected and split into two
308 halves with similar accuracy, to be used in the SYLL and PW paradigm (SYLL: mean accuracy
309 90.8%, SD 5.8%, range 78% – 100%; PW: mean accuracy 89.9%, SD 5.4%, range 80% – 100%;
310 $t(118) = .842, p = .401$).

311 The two final lists were matched for the gender of the subject (31 feminine, 29 masculine), sentence
312 structure (type 1: $n=20$, type 2: $n=19$, type 3: $n=21$), number of present and past verb forms (SYLL:
313 30 present / 30 past forms, PW: 29 present / 31 past forms), length in syllables (SYLL: mean 7.43,
314 SD .76, range 5–8; PW: mean 7.31, SD .74, range 5–8; $t(118) = .843, p = .400$) and overall word
315 frequency (SYLL: mean 39.37, SD 55.18, range .4–424.1; PW: mean 33.75, SD 45.71, range .5 –
316 277; $t(358) = 1.051, p = .293$).

317 In the SYLL baseline condition, stimuli were 60 phonologically legal Russian syllables (consonant +
318 vowel), where the vowel was spelled multiple times in order to match the experimental condition for
319 length in letters (for example, for the syllable “*ϕe*” - “*fe*”, the stimulus was «*ϕeee ϕeeeeee ϕeeeeee*
320 *...*» - «*Feee feeeee feeeee ...*»). Participants were instructed to ignore the exact number of vowel
321 letters and pronounce the syllable duration approximately. In the PW baseline condition, the stimuli
322 were 60 pseudowords phonotactically legal in Russian, constructed to pairwise match the number of
323 syllables in experimental sentences. All pseudowords were constructed as pseudonouns: that is, none
324 of them had inflection typical for other parts of speech.

325 **2.3 MRI data acquisition**

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

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

335 field of view (FoV) 320×320 mm, repetition time (TR) 2200 ms, echo time (TE) 2.43 ms, flip angle
336 8°. Functional blood-oxygenation-level-dependent (BOLD) data were obtained using the following
337 parameters: voxel size 3.0×3.0×3.0 mm, 30 oblique slices in interleaved order, slice thickness 3 mm,
338 FoV 205×205 mm, TR 7000 ms, TE 30 ms, delay in TR 5000 ms (sparse sampling), flip angle 90°,
339 128 volumes per paradigm.

340 **2.4 Procedure**

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

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

354 **FIGURE 1 ABOUT HERE**

355 **2.5 Data analysis**

356 **2.5.1 Behavioral data**

357 The participants' auditory responses from both sessions were transcribed, except for one participant's
358 responses that were lost due to technical error. Response accuracy was assessed independently by
359 two raters. In the experimental condition, a response was considered accurate if it was a
360 grammatically correct and semantically appropriate sentence completion. In the baseline conditions, a
361 response was considered accurate if the participant read and repeated a syllable/pseudoword without
362 any phonological errors. Inter-rater reliability, as assessed using percent agreement and Cohen's
363 kappa (Cohen, 1960), was high (1st session: 98.36% and .76, respectively; 2nd session: 98.62% and
364 .69, respectively). All inconsistencies were resolved by discussion between the two raters.

365 **2.5.2 Activation maps**

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

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

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

399 Anatomical labels for activation clusters were determined based on the Brainnetome atlas (Fan et al.,
400 2016) implemented in the ICN_atlas toolbox (Kozák et al., 2017) for SPM12. The Brainnetome atlas
401 was selected due to detailed parcellation, as it includes 246 regions in total. The activation maps were
402 visualized in MRICroGL 1.2.2 (<https://www.nitrc.org/projects/mricrogl>).

403 **2.5.3 Assessing individual-level activation of ‘key language-related areas’**

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

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

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

425 **2.5.4 Test-retest reliability**

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

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

448 **2.5.5 Lateralization indices**

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

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

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

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

467 **3 Results**

468 **3.1 Behavioral results**

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

474 **3.2 Group-level activation maps**

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

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

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

492 **FIGURE 2 ABOUT HERE**

493 **FIGURE 3 ABOUT HERE**

494 **3.3 Individual-level activation in key 'language-related areas'**

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

498 **FIGURE 4 ABOUT HERE**

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

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

505 Bonferroni-corrected repeated-measures ANOVAs showed that individual-level activation volume
506 was greater with the PW than SYLL baseline in three of ‘key language-related areas’: pars
507 triangularis of the inferior frontal gyrus, $F(1,17) = 25.22, p < .001$, posterior middle temporal gyrus,
508 $F(1,17) = 25.75, p < .001$, and angular gyrus, $F(1,17) = 27.97, p < .001$. In the other five ‘key
509 language-related areas’, individual-level activation volume was not significantly affected by baseline.
510 Expectedly, individual-level activation volume was significantly affected by statistical threshold in
511 six of ‘key language-related areas’: pars opercularis of the inferior frontal gyrus, $F(1.08, 18.29) =$
512 $15.82, p = .001$, pars triangularis of the inferior frontal gyrus, $F(1.11, 18.78) = 21.40, p < .001$,
513 posterior superior temporal gyrus, $F(1.10, 18.78) = 9.16, p = .006$, posterior middle temporal gyrus,
514 $F(1.17, 19.94) = 8.92, p = .005$, angular gyrus, $F(1.06, 18.01) = 10.71, p = .003$, supramarginal gyrus,
515 $F(1.25, 21.19) = 15.05, p < .001$. Post-hoc pairwise comparisons showed that in all six regions,
516 cluster-size correction with a minimum cluster size of $k \geq 200 \text{ mm}^3$ at $p < .001$ yielded fewer
517 significantly activated voxels than the FWE correction for multiple comparisons at $p < .05$ (all $p <$
518 $.001$). No other pairwise comparisons were significant when corrected for multiple comparisons. The
519 interaction between baseline and statistical threshold was not significant in any of ‘key language-
520 related areas’.

521 3.4 Test-retest reliability

522 Mean Dice coefficients quantifying the overlap of individual activation in the first and second
523 scanning session are presented in Table 1. Mean Dice coefficients ranged from .39 to .61, that is,
524 from low-moderate to moderate-high. A Greenhouse-Geisser-corrected repeated-measures ANOVA
525 demonstrated a significant effect of brain region, $F(1.04, 17.74) = 6.72, p = .018$. Post-hoc pairwise
526 comparisons showed that Dice coefficients were higher in the frontal than temporo-parietal ($p = .019$)
527 or frontal-temporo-parietal ($p = .050$) region, and in the frontal-temporo-parietal than temporo-
528 parietal ($p = .016$) region. Dice coefficients were significantly higher with the PW than SYLL
529 baseline, $F(1, 17) = 5.08, p = .038$. Finally, there was a significant effect of statistical threshold,
530 $F(1.41, 23.92) = 6.49, p = .011$. Post-hoc pairwise comparisons showed that Dice coefficients were
531 higher with the most liberal statistical threshold (cluster-size correction with a minimum cluster size
532 of $k \geq 200 \text{ mm}^3$ at $p < .001$) relative to FWE correction for multiple comparisons at $p < .05$ ($p <$
533 $.001$) and relative to the adaptive thresholding as implemented in Gorgolewski et al. (2012) at $p < .05$ ($p =$
534 $.011$). No interactions were significant.

535 TABLE 1 ABOUT HERE

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

540 3.5 Lateralization indices

541 Table 2 presents mean lateralization indices (LI) for three regions (frontal, temporal-parietal and
542 frontal-temporal-parietal) for the two paradigms (SYLL and PW) in the two scanning sessions. Table
543 2 also includes Spearman’s correlation coefficients examining reproducibility of the individual LI
544 values across two sessions.

545

TABLE 2 ABOUT HERE

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

553 A repeated-measures ANOVA showed a significant three-way interaction between baseline, region
554 and session, $F(1.11, 18.93) = 8.92, p = .006$. To address this interaction, separate repeated-measures
555 ANOVAs were performed for the SYLL and PW baseline (there was no main effect of baseline,
556 $F(1,17) = .06, p = .811$). With the SYLL baseline, a main effect of region was significant, $F(1.12,$
557 $18.99) = 15.88, p = .001$, with higher LIs in the frontal than temporo-parietal ($p = .001$) or fronto-
558 temporo-parietal region ($p = .038$) and in the fronto-temporo-parietal than temporo-parietal region (p
559 $< .001$). No other main effects or interactions were significant. With the PW baseline, a two-way
560 interaction of session and region was significant, $F(1.08, 18.31) = 6.71, p = .017$, so separate
561 repeated-measures ANOVAs were performed for the first and second session (there was no main
562 effect of session, $F(1,17) = .26, p = .618$). In the first session with the PW baseline, the main effect of
563 region was significant, $F(1.07, 18.19) = 15.21, p = .001$. Post-hoc pairwise comparisons showed the
564 same hierarchy of LIs as with the SYLL baseline: LIs were higher in the frontal than temporo-
565 parietal ($p = .001$) and frontal-temporo-parietal region ($p = .006$) and in the frontal-temporo-parietal
566 than temporo-parietal region ($p = .001$). In the second session with the PW baseline, the main effect
567 of region was again significant but much greater, $F(1.09, 18.52) = 80.49, p < .001$. Post-hoc pairwise
568 comparisons showed the same hierarchy of LIs as for the first session or for the SYLL baseline: LIs
569 were higher in the frontal than temporo-parietal ($p < .001$) and fronto-temporo-parietal region ($p <$
570 $.006$) and in the frontal-temporo-parietal than temporo-parietal region ($p < .001$). That is, the
571 significant three-way interaction between baseline, region and session was driven by the main effect
572 of region being the greatest in the second session with the PW baseline.

573 **4 Discussion**

574 We presented a new fMRI language localizer for preoperative language mapping in Russian-speaking
575 individuals. Following the world's best practices (Barnett et al., 2014, Black et al., 2017, Salek et al.,
576 2017, Połczyńska et al., 2017, Unadkat et al., 2019, Wilson et al., 2017, Zacà et al., 2012), the
577 paradigm used a sentence completion task that uniquely engages both language production and
578 comprehension at the word and sentence level. The current study validated the localizer paradigm in
579 a control group of neurologically healthy individuals. In this group, the paradigm successfully
580 activated key language-related areas, elicited expected left-hemispheric lateralization and showed
581 test-retest reliability comparable to previous studies. Apart from demonstrating general validity and
582 reliability of the paradigm, we compared two different baseline conditions (SYLL and PW), for the
583 first time for the sentence completion task. All outcomes were reported at three statistical thresholds.

584 **4.1 Activation of key language-related areas**

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

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

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

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

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

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

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

637 On the other hand, most participants did not show significant individual-level activation in the basal
638 temporal area or supramarginal gyrus. The lack of activation in the supramarginal gyrus was
639 surprising, given that activation at the temporo-parietal junction is expected in sentence-level tasks
640 (Połczyńska et al., 2017) and has sometimes also been found with word-level tasks (Roux et al.,
641 2003, Stippich et al., 2007). Still, the majority of participants in the present study showed activation
642 in the adjacent angular gyrus. With regard to the basal temporal language area, it has not been
643 consistently activated by previous fMRI language localizers either (Połczyńska et al., 2017, Barnett
644 et al., 2014, Wilson et al., 2017), despite its long-presumed involvement in lexical retrieval (Krauss
645 et al., 1996, Lüders et al., 1991).

646 **4.2 Lateralization of language processing**

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

658 Interestingly, language-related activity was significantly more strongly left-lateralized in the frontal
659 (and, correspondingly, frontal-temporal-parietal) than temporal-parietal region. This is in line with
660 contemporary models of language processing. For example, the dual-stream model (Hickok &
661 Poeppel, 2007) postulates a bilaterally organized ventral stream, which primarily involves the
662 temporal lobe and connects speech sounds to meanings, and a left-lateralized dorsal stream, which
663 extends to the frontal lobe and maps speech sounds to articulatory networks. In the same vein, Peelle
664 (2012) argues that phonological and lexical information are processed bilaterally in the temporal
665 lobe, whereas sentence processing engages a left-lateralized pathway including the left inferior
666 frontal gyrus. Although the above models (Hickok & Poeppel, 2007, Peelle, 2012) are mainly
667 concerned with auditory speech processing and differ in the specific division of labor between the
668 frontal and temporal regions, our findings converge with them with regard to stronger left-
669 hemispheric lateralization in the frontal than temporal lobe.

670 **4.3 Test-retest reliability**

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

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

684

TABLE 3 ABOUT HERE

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

695 Hemispheric lateralization of language-related activity also showed moderate test-retest reliability.
696 LIs in the first and second scanning session showed either a significant moderate-to-strong
697 correlation (with the PW baseline) or a statistical trend for a moderate correlation (with the SYLL
698 baseline). This held true when LIs were calculated for the frontal region, temporal-parietal region,
699 and combination thereof. The findings on moderate reliability of the paradigm in identifying both
700 localization and hemispheric lateralization of language-related activity contribute to the literature on
701 general test-retest reliability of fMRI (Bennett & Miller, 2010, Elliott et al., 2020, Holiga et al.,
702 2018).

703 **4.4 Comparison of baseline conditions**

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

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

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

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

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

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

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

761 4.5 Comparison of statistical thresholds

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

765 correction for multiple comparisons with a minimum cluster size of $k \geq 200$ mm³ at $\alpha = .001$. Many
766 previous studies have also reported results at multiple statistical thresholds (Dodoo-Schnitko et al.,
767 2012, Morrison et al., 2016a, Nadkarni et al., 2015, Nettekoven et al., 2018 Wilson et al., 2017),
768 since there is no ‘gold standard’ for statistical thresholding in individual or group-level fMRI
769 analysis. Moreover, studies have shown that the statistical threshold may vastly impact metrics
770 induced from fMRI analysis, such as lateralization indices (Nadkarni et al., 2015) or test-retest
771 reliability metrics (Stevens et al., 2016), so reporting results at only one threshold could be
772 misleading.

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

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

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

806 **4.6 Future directions**

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

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

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

823 **5 Conflict of Interest**

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

826 **6 Author Contributions**

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

832 **7 Funding**

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

835 **8 Acknowledgments**

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

838 **9 References**

- 839 Agarwal, S., Sair, H. I., Gujar, S., and Pillai, J. J. (2019). Language mapping with fMRI. *Top. Magn.*
840 *Reson. Imaging.* 28:4, 225–233. doi:10.1097/rmr.0000000000000216
- 841 Almairac, F., Duffau, H., and Herbet, G. (2018). Contralesional macrostructural plasticity of the
842 insular cortex in patients with glioma: A VBM study. *Neurology.* 91:20, e1902-e1908.
843 doi:10.1212/WNL.0000000000006517
- 844 American College of Radiology (2014). ACR–ASNR–SPR practice parameter for the performance of
845 functional magnetic resonance imaging (fMRI) of the brain. Amended 2014 (Resolution 39).
846 American College of Radiology.

- 847 Barnett, A., Marty-Dugas, J., and McAndrews, M. P. (2014). Advantages of sentence-level fMRI
848 language tasks in presurgical language mapping for temporal lobe epilepsy. *Epilepsy Behav.* 32,
849 114–120. doi:10.1016/j.yebeh.2014.01.010
- 850 Bennett, C.M., and Miller, M.B. (2010). How reliable are the results from functional magnetic
851 resonance imaging? *Ann. N. Y. Acad. Sci.* 1191, 133-155. doi:10.1111/j.1749-
852 6632.2010.05446.x
- 853 Benjamin, C. F., Walshaw, P. D., Hale, K., Gaillard, W. D., Baxter, L. C., Berl, M. M., Polczynska,
854 M., Noble, S., Alkawadri, R., and Hirsch, L. J. (2017). Presurgical language fMRI: Mapping of
855 six critical regions. *Hum. Brain Mapp.* 38:8, 4239–4255. doi:10.1002/hbm.23661
- 856 Benjamin, C.F.A., Dhingra, I., Li, A.X., Blumenfeld, H., Alkawadri, R., Bickel, S., Helmstaedter, C.,
857 Meletti, S., Bronen, R.A., Warfield, S.K., Peters, J.M., Reutens, D., Połczyńska, M.M., Hirsch,
858 L.J., and Spencer, D.D. (2018). Presurgical language fMRI: Technical practices in epilepsy
859 surgical planning. *Hum. Brain Mapp.* 39:10, 4032-4042. doi:10.1002/hbm.24229
- 860 Binder J. R. (2015). The Wernicke area: Modern evidence and a reinterpretation. *Neurology.* 85:24,
861 2170–2175. doi:10.1212/WNL.0000000000002219
- 862 Binder, J.R., Swanson, S.J., Hammeke, T.A., Morris, G.L., Mueller, W.M. Fischer, M., Benbadis, S.,
863 Frost, J.A., Rao, S. M., and Haughton, V. M. (1996). Determination of language dominance
864 using functional MRI: A comparison with the Wada test. *Neurology.* 46:4, 978–984.
865 doi:10.1212/wnl.46.4.978
- 866 Binder, J. R., Swanson, S. J., Hammeke, T. A., and Sabsevitz, D. S. (2008). A comparison of five
867 fMRI protocols for mapping speech comprehension systems. *Epilepsia.* 49:12, 1980–1997.
868 doi:10.1111/j.1528-1167.2008.01683.x
- 869 Binder J. R., Desai R. H., Graves W. W., and Conant L. L. (2009). Where is the semantic system? A
870 critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex.* 19:12,
871 2767–2796. doi:10.1093/cercor/bhp055
- 872 Black, D.F., Vachha, B., Mian, A., Faro, S.H., Maheshwari, M., Sair, H.I., Petrella, J.R., Pillai, J.J.,
873 and Welker, K., 2017. American Society of Functional Neuroradiology-recommended fMRI
874 paradigm algorithms for presurgical language assessment. *Am. J. Neuroradiol.* 38:10, E65–E73.
875 doi:10.3174/ajnr.A5345
- 876 Bleich-Cohen, M., Hendler, T., Kotler, M., and Strous, R. D. (2009). Reduced language lateralization
877 in first-episode schizophrenia: An fMRI index of functional asymmetry. *Psychiatry Res.*
878 *Neuroimaging.* 171:2, 82–93. doi:10.1016/j.psychresns.2008.03.002
- 879 Bonilha L., Hillis A. E., Hickok G., Den Ouden D. B., Rorden C., and Fridriksson J. (2017).
880 Temporal lobe networks supporting the comprehension of spoken words. *Brain.* 140:9, 2370–
881 2380. doi:10.1093/brain/awx169
- 882 Bradshaw, A. R., Thompson, P. A., Wilson, A. C., Bishop, D., and Woodhead, Z. (2017). Measuring
883 language lateralisation with different language tasks: a systematic review. *PeerJ.* 5, e3929.
884 doi:10.7717/peerj.3929
- 885 Brennan, N.M.P., Whalen, S., De Moraes Branco, D., O’Shea, J.P., Norton, I.H., and Golby, A.J.
886 (2007). Object naming is a more sensitive measure of speech localization than number counting:
887 converging evidence from direct cortical stimulation and fMRI. *NeuroImage.* 37, S100–S108.
888 doi:10.1016/j.neuroimage.2007.04.052
- 889 Buchsbaum, B. R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal
890 gyrus in phonological processing for speech perception and production. *Cogn. Sci.* 25:5, 663-
891 678. doi:10.1016/S0364-0213(01)00048-9
- 892 Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.
893 doi:10.1177/001316446002000104
- 894 Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., Krieger-
895 Redwood, K., Bernhardt, B. C., Smallwood, J., and Jefferies, E. (2016). Exploring the role of the

- 896 posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with
897 executive processes. *NeuroImage*. 137, 165–177. doi:10.1016/j.neuroimage.2016.05.051
- 898 Deblaere, K., Backes, W.H., Hofman, P., Vandemaele, P., Boon, P.A., Vonck, K., Boon, P., Troost,
899 J., Vermeulen, J., Wilmink, J., Achten, E., and Aldenkamp, A. (2002). Developing a
900 comprehensive presurgical functional MRI protocol for patients with intractable temporal lobe
901 epilepsy: a pilot study. *Neuroradiology*. 44:8, 667–673. doi:10.1007/s00234-002-0800-4
- 902 De Guibert, C., Maumet, C., Ferré, J.-C., Jannin, P., Biraben, A., Allaire, C., Barillot, C., and Le
903 Rumeur, E. (2010). fMRI language mapping in children: a panel of language tasks using visual
904 and auditory stimulation without reading or metalinguistic requirements. *NeuroImage*. 51(2),
905 897-909. doi:10.1016/j.neuroimage.2010.02.054f
- 906 Den Ouden D., Malyutina S., Basilakos A., Bonilha L., Gleichgerrcht E., Yourganov G., Hillis A. E.,
907 Hickok G., Rorden C., and Fridriksson J. (2019). Cortical and structural-connectivity damage
908 correlated with impaired syntactic processing in aphasia. *Hum. Brain Mapp*. 40:7, 2153-2173.
909 doi:10.1002/hbm.24514
- 910 De Witte E., and Mariën P. (2013). The neurolinguistic approach to awake surgery reviewed. *Clin.*
911 *Neurol. Neurosurg*. 115(2), 127–145. doi:10.1016/j.clineuro.2012.09.015
- 912 Dodoo-Schittko, F., Rosengarth, K., Doenitz, C., and Greenlee, M. (2012). Assessing language
913 dominance with functional MRI: The role of control tasks and statistical analysis.
914 *Neuropsychologia*. 50, 2684–2691. doi:10.1016/j.neuropsychologia.2012.07.032
- 915 Dragoy, O., Zyryanov, A., Bronov, O., Gordeyeva, E., Gronskaya, N., Kryuchkova, O., Klyuev, E.,
916 Kopachev, D., Medyanik, I., Mishnyakova, L., Pedyash, N., Pronin, I., Reutov, A., Sitnikov, A.,
917 Stupina, E., Yashin, K., Zhirnova, V., and Zuev, A. (2020). Functional linguistic specificity of
918 the left frontal aslant tract for spontaneous speech fluency: Evidence from intraoperative
919 language mapping. *Brain Lang*. 208, 104836. doi:10.1016/j.bandl.2020.104836
- 920 Duffau, H. (2012). The challenge to remove diffuse low-grade gliomas while preserving brain
921 functions. *Acta Neurochir*. 154, 569-574. doi:10.1007/s00701-012-1275-7
- 922 Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2020). *Ethnologue: Languages of the World*.
923 Twenty-third edition. Dallas, Texas: SIL International. Online version:
924 <http://www.ethnologue.com>.
- 925 Ek L., Almkvist O., Wiberg M.K., Stragliotto G., and Smits A. (2010). Early cognitive impairment in
926 a subset of patients with presumed low-grade glioma. *Neurocase*. 16(6), 503–511.
927 doi:10.1080/13554791003730634
- 928 Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M.L., Moffitt,
929 T.E., Caspi, A., and Hariri, A.R. (2020)/ What is the test-retest reliability of common task-
930 functional MRI measures? New empirical evidence and a meta-analysis. *Psychol. Sci*. 31(7),
931 792-806. doi:10.1177/0956797620916786
- 932 Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox,
933 P. T., Eickhoff, S. B., Yu, C., and Jiang, T. (2016). The Human Brainnetome Atlas: A new brain
934 atlas based on connectional architecture. *Cereb. Cortex*. 26(8), 3508–3526.
935 doi:10.1093/cercor/bhw157
- 936 Fedorenko, E., Hsieh, P.J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010).
937 New method for fMRI investigations of language: defining ROIs functionally in individual
938 subjects. *J. Neurophysiol*. 104:2, 1177-1194. doi:10.1152/jn.00032.2010
- 939 Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J.,
940 Reul, J., and Elger, C. E. (2003). Intrasubject reproducibility of presurgical language
941 lateralization and mapping using fMRI. *Neurology*. 60:6, 969-975.
942 doi:10.1212/01.wnl.0000049934.34209.2e

- 943 Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., and Brueckmann, H. (2010).
944 Sensitivity and reliability of language laterality assessment with a free reversed association task -
945 a fMRI study. *Eur. Radiol.* 20:3, 683-695. doi:10.1007/s00330-009-1602-4
- 946 Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T.,
947 and Crone, N. E. (2015). Redefining the role of Broca's area in speech. *Proc. Natl. Acad. Sci. U.*
948 *S. A.* 112:9, 2871-2875. doi:10.1073/pnas.1414491112
- 949 Gabel, N., Altshuler, D. B., Brezzell, A., Briceño, E. M., Boileau, N. R., Miklja, Z., Kluin, K.,
950 Ferguson, T., McMurray, K., Wang, L., Smith, S. R., Carlozzi, N. E., and Hervey-Jumper, S. L.
951 (2019). Health-related quality of life in adult low and high-grade glioma patients using the
952 National Institutes of Health Patient-Reported Outcomes Measurement Information System
953 (PROMIS) and Neuro-QOL assessments. *Front. Neurol.* 10, 212. doi:10.3389/fneur.2019.00212
- 954 Gorgolewski, K., Storkey, A. J., Bastin, M. E., and Pernet, C. R. (2012). Adaptive thresholding for
955 reliable topological inference in single subject fMRI analysis. *Front. Hum. Neurosci.* 6, 245.
956 doi:10.3389/fnhum.2012.00245
- 957 Graves, W. W., Grabowski, T. J., Mehta, S., and Gupta, P. (2008). The left posterior superior
958 temporal gyrus participates specifically in accessing lexical phonology. *J. Cogn. Neurosci.* 20:9,
959 1698-1710. doi:10.1162/jocn.2008.20113
- 960 Gross, W. L., and Binder, J. R. (2014). Alternative thresholding methods for fMRI data optimized for
961 surgical planning. *NeuroImage.* 84, 554-561. doi:10.1016/j.neuroimage.2013.08.066
- 962 Grummich, P., Nimsky, C., Pauli, E., Buchfelder, M., and Ganslandt, O. (2006). Combining fMRI
963 and MEG increases the reliability of presurgical language localization: a clinical study on the
964 difference between and congruence of both modalities. *NeuroImage.* 32:4, 1793-1803. doi:
965 10.1016/j.neuroimage.2006.05.034
- 966 Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9:9, 416-23.
967 doi:10.1016/j.tics.2005.07.004
- 968 Hakyemez, B., Erdogan, C., Yildirim, N., Bora, I., Bekar, A., and Parlak, M. (2006). Functional MRI
969 in 507 patients with intracranial lesions near language areas. *Neuroradiol. J.* 508:19, 306-312.
970 doi:10.1177/197140090601900306
- 971 Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R.,
972 Gurney, E. M., and Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Hum.*
973 *Brain Mapp.* 7:3, 213-223. doi: 10.1002/(sici)1097-0193(1999)7:3<213::aid-hbm5>3.0.co;2-n
- 974 Hammers, A., Allom, R., Koeppe, M. J., Free, S. L., Myers, R., Lemieux, L., Mitchell, T. N., Brooks,
975 D. J., and Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human
976 brain, with particular reference to the temporal lobe. *Hum. Brain Mapp.* 19:4, 224-247.
977 doi:10.1002/hbm.10123
- 978 Hervey-Jumper, S.L., Li, J., Lau, D., Molinaro, A.M., Perry, D.W., Meng, L., and Berger, M.S.
979 (2015). Awake craniotomy to maximize glioma resection: methods and technical nuances over a
980 27-year period. *J. Neurosurg.* 123(2), 325-339. doi:10.3171/2014.10.JNS141520
- 981 Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev.*
982 *Neurosci.* 8, 393-402. doi:10.1038/nrn2113
- 983 Hilari, K., Wiggins, R., Roy, P., Byng, S., and Smith, S. (2003). Predictors of health-related quality
984 of life (HRQL) in people with chronic aphasia. *Aphasiology.* 17:4, 365-381.
985 doi:10.1080/02687030244000725
- 986 Hillis A. E., Rorden C., and Fridriksson, J. (2017). Brain regions essential for word comprehension:
987 Drawing inferences from patients. *Ann. Neurol.* 81:6, 759-768. doi:10.1002/ana.24941
- 988 Holiga, Š., Sambataro, F., Luzy, C., Greig, G., Sarkar, N., Renken, R. J., Marsman, J.-B. C., Schobel,
989 S. A., Bertolino, A., and Dukart, J. (2018). Test-retest reliability of task-based and resting-state
990 blood oxygen level dependence and cerebral blood flow measures. *PLoS ONE.* 13:11, e0206583.
991 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 Champfleury, N. M., Deverduin, 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 Lehericy, 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

- 1039 Mauler, J., Neuner, I., Neuloh, G., Fimm, B., Boers, F., Wiesmann, M., Clusmann, H., Langen, K.J.,
1040 and Shah, N.J. (2017). Dissociated crossed speech areas in a tumour patient. *Case Rep. Neurol.*
1041 9:2, 131–136. doi:10.1159/000475882
- 1042 Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of
1043 the human brain: Theory and rationale for its development. *Neuroimage.* 2:2, 89–101. doi:
1044 10.1006/nimg.1995.1012
- 1045 Miró, J., Ripollés, P., López-Barroso, D., Vilà-Balló, A., Juncadella, M., de Diego-Balaguer, R.,
1046 Marco-Pallares, J., Rodríguez-Fornells, A., and Falip, M. (2014). Atypical language organization
1047 in temporal lobe epilepsy revealed by a passive semantic paradigm. *BMC Neurol.* 14, 98.
1048 doi:10.1186/1471-2377-14-98
- 1049 Morrison, M. A., Churchill, N. W., Cusimano, M. D., Schweizer, T. A., Das S., and Graham, S. J.
1050 (2016a). Reliability of task-based fMRI for preoperative planning: A test-retest study in brain
1051 tumor patients and healthy controls. *PLoS ONE.* 11(2), e0149547.
1052 doi:10.1371/journal.pone.0149547
- 1053 Morrison, M. A., Tam, F., Garavaglia, M. M., Hare, G. M. T., Cusimano, M. D., Schweizer, T. A.,
1054 Das, S., and Graham, S.J. (2016b). Sources of variation influencing concordance between
1055 functional MRI and direct cortical stimulation in brain tumor surgery. *Front. Neurosci.* 10, 461.
1056 doi:10.3389/fnins.2016.00461
- 1057 Nadkarni T. N., Andreoli M. J., Nair V. A., Yin P., Young B. M., Kundu B., Pankratz J., Radtke A.,
1058 Holdsworth R., Kuo J. S., Field A. S., Baskaya M. K., Moritz C. H., Meyerand M. E., and
1059 Prabhakaran V. (2014). Usage of fMRI for pre-surgical planning in brain tumor and vascular
1060 lesion patients: task and statistical threshold effects on language lateralization. *Neuroimage Clin.*
1061 7, 415-423. doi:10.1016/j.nicl.2014.12.014
- 1062 Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., and Weiß Lucas, C. (2018). Short- and long-
1063 term reliability of language fMRI. *Neuroimage.* 176, 215-225.
1064 doi:10.1016/j.neuroimage.2018.04.050
- 1065 Newman, S. D., Twieg, D. B., and Carpenter, P. A. (2001). Baseline conditions and subtractive logic
1066 in neuroimaging. *Hum. Brain Mapp.* 14:4, 228–235. doi:10.1002/hbm.1055
- 1067 Ojemann, G. A. (1979). Individual variability in cortical localization of language. *J. Neurosurg.* 50:2,
1068 164-169. doi:10.3171/jns.1979.50.2.0164
- 1069 Ojemann, G.A. (1993). Functional mapping of cortical language areas in adults. Intraoperative
1070 approaches. *Adv. Neurol.* 63, 155–163.
- 1071 Partovi, S., Jacobi, B., Rapps, N., Zipp, L., Karimi, S., Rengier, F., Lyo, J. K., and Stippich, C.
1072 (2012). Clinical standardized fMRI reveals altered language lateralization in patients with brain
1073 tumor. *Am. J. Neuroradiol.* 33, 2151–2157. doi:10.3174/ajnr.A3137
- 1074 Patrikelis P., Gatzonis S., Siatouni A., Angelopoulos E., Konstantakopoulos G., Takousi M., Sakas
1075 D. E., and Zalonis I. (2016). Preoperative neuropsychological presentation of patients with
1076 refractory frontal lobe epilepsy. *Acta Neurochir.* 158, 1139–1150. doi:10.1007/s00701-016-
1077 2786-4
- 1078 Pearson, J., Naselaris, T., Holmes, E. A., and Kosslyn, S. M. (2015). Mental imagery: Functional
1079 mechanisms and clinical applications. *Trends Cogn. Sci.* 19:10, 590–602.
1080 doi:10.1016/j.tics.2015.08.003
- 1081 Peelle J. E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is:
1082 A hierarchical perspective. *Front. Hum. Neurosci.* 6, 309. doi:10.3389/fnhum.2012.00309
- 1083 Picht, T., Krieg, S.M., Sollmann, N., Rösler, J., Niraula, B., Neuvonen, T., Savolainen, P., Lioumis,
1084 P., Mäkelä, J.P., Deletis, V., Meyer, B., Vajkoczy, P., and Ringel, F. (2013). A comparison of
1085 language mapping by preoperative navigated transcranial magnetic stimulation and direct
1086 cortical stimulation during awake surgery. *Neurosurgery.* 72:5, 808-819.
1087 doi:10.1227/NEU.0b013e3182889e01

- 1088 Pillai, J. J., and Zaca, D. (2011). Relative utility for hemispheric lateralization of different clinical
1089 fMRI activation tasks within a comprehensive language paradigm battery in brain tumor patients
1090 as assessed by both threshold-dependent and threshold-independent analysis methods.
1091 *NeuroImage*. 54:S1, S136-S145.
- 1092 Połczyńska, M., Japardi, K., Curtiss, S., Moody, T., Benjamin, C., Cho, A., Vigil, C., Kuhn, T., Jones,
1093 M., and Bookheimer, S. (2017). Improving language mapping in clinical fMRI through
1094 assessment of grammar. *Neuroimage Clin*. 15, 415-427. doi:10.1016/j.nicl.2017.05.021
- 1095 Pouratian, N., Bookheimer, S.Y., Rex, D.E., Martin, N.A., and Toga, A.W. (2002). Utility of
1096 preoperative functional magnetic resonance imaging for identifying language cortices in patients
1097 with vascular malformations. *J. Neurosurg*. 97:1, 21–32. doi:10.3171/jns.2002.97.1.0021
- 1098 Racine C.A., Li J., Molinaro A.M., Butowski N., and Berger M.S. (2015). Neurocognitive function in
1099 newly diagnosed low-grade glioma patients undergoing surgical resection with awake mapping
1100 techniques. *Neurosurgery*. 77:3, 371-379. doi:10.1227/NEU.0000000000000779
- 1101 Richardson F. M., Seghier M. L., Leff A. P., Thomas M. S., and Price C. J. (2011). Multiple routes
1102 from occipital to temporal cortices during reading. *J. Neurosci*. 31:22, 8239–8247.
1103 doi:10.1523/JNEUROSCI.6519-10.2011
- 1104 Robinson, G., Shallice, T., Bozzali, M., and Cipolotti, L. (2010). Conceptual proposition selection
1105 and the LIFG: neuropsychological evidence from a focal frontal group. *Neuropsychologia*. 48:6,
1106 1652–1663. doi:10.1016/j.neuropsychologia.2010.02.010
- 1107 Rodd, J. M., Davis, M. H., and Johnsrude, I.S. (2005). The neural mechanisms of speech
1108 comprehension: fMRI studies of semantic ambiguity. *Cereb. Cortex*. 15, 1261–1269.
1109 doi:10.1093/cercor/bhi009
- 1110 Rofes, A., Mandonnet, E., de Aguiar, V., Rapp, B., Tsapkini, K., and Miceli, G. (2019). Language
1111 processing from the perspective of electrical stimulation mapping. *Cogn. Neuropsychol*. 36:3-4,
1112 117-139. doi:10.1080/02643294.2018.1485636
- 1113 Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., Valk, J., and Scheltens, P. (1997).
1114 Test-retest analysis with functional MR of the activated area in the human visual cortex. *Am. J.*
1115 *Neuroradiol*. 18:7, 1317–1322.
- 1116 Roux, F.E., Boulanouar, K., Lotterie, J.A., Mejdoubi, M., LeSage, J. P., and Berry, I. (2003).
1117 Language functional magnetic resonance imaging in preoperative assessment of language areas:
1118 correlation with direct cortical stimulation. *Neurosurgery*. 52:6, 1335–1347.
1119 doi:10.1227/01.neu.0000064803.05077.40
- 1120 Ruge, M.I., Victor, J., Hosain, S., Correa, D.D., Relkin, N.R., Tabar, V., Brennan, C., Gutin, P.H.,
1121 and Hirsch, J. (1999). Concordance between functional magnetic resonance imaging and
1122 intraoperative language mapping. *Stereotact. Funct. Neurosurg*. 72:2–4, 95–102. doi:
1123 10.1159/000029706
- 1124 Rumshiskaya, A., Vlasova, R., Litvinova, L., Pechenkova, E., and Merzhina, E. (2014). Combined
1125 analysis of two tasks improves localization of Wernicke's area in patients with primary brain
1126 tumors. Poster presented at the European Society for Neuroradiology. doi:10.1594/ecr2014/C-
1127 1232
- 1128 Rutten, G.J.M., Ramsey, N.F., Van Rijen, P.C., Alpherts, W.C., and Van Veelen, C.W.M. (2002).
1129 fMRI-determined language lateralization in patients with unilateral or mixed language
1130 dominance according to the Wada test. *NeuroImage*. 17:1, 447–460.
1131 doi:10.1006/nimg.2002.1196
- 1132 Salek, K.E., Hassan, I.S., Kotrotsou, A., Abrol, S., Faro, S. H., Mohamed, F. B., Zinn, P. O., Wei,
1133 W., Li, N., Kumar, A. J., Weinberg, J. S., Wefel, J. S., Kesler, S. R., Liu, H.-L. A., Hou, P.,
1134 Stafford, R. J., Prabhu, S., Sawaya, R., and Colen, R. R. (2017). Silent sentence completion
1135 shows superiority localizing Wernicke's area and activation patterns of distinct language

- 1136 paradigms correlate with genomics: prospective study. *Sci. Rep.* 7, 12054. doi:10.1038/s41598-
1137 017-11192-2
- 1138 Sanjuan, A., Bustamante, J.-C., Forn, C., Ventura-Campos, N., Barros-Loscertales, A., Martinez J-C,
1139 Villanueva V, and Avila C. (2010). Comparison of two fMRI tasks for the evaluation of the
1140 expressive language function. *Neuroradiology* 52:5, 407–415. doi:10.1007/s00234-010-0667-8
- 1141 Satoer, D., Visch-Brink, E., Dirven, C., and Vincent, A. (2016). Glioma surgery in eloquent areas:
1142 can we preserve cognition? *Acta Neurochi.* 158:1, 35–50. doi:10.1007/s00701-015-2601-7
- 1143 Silva, M. A., See, A. P., Essayed, W. I., Golby, A. J., and Tie, Y. (2018). Challenges and techniques
1144 for presurgical brain mapping with functional MRI. *NeuroImage Clin.* 17, 794–803.
1145 doi:10.1016/j.nicl.2017.12.008
- 1146 Sollmann, N., Kubitscheck, A., Maurer, S., Ille, S., Hauck, T., Kirschke, J. S., Ringel, F., Meyer, B.,
1147 and Krieg, S.M. (2016). Preoperative language mapping by repetitive navigated transcranial
1148 magnetic stimulation and diffusion tensor imaging fiber tracking and their comparison to
1149 intraoperative stimulation. *Neuroradiology.* 58, 807–818. doi:10.1007/s00234-016-1685-y
- 1150 Spena G., Nava A., Cassini F., Pepoli A., Bruno M., D'Agata F., Cauda F., Sacco K., Duca S.,
1151 Barletta L., and Versari P. (2010). Preoperative and intraoperative brain mapping for the
1152 resection of eloquent-area tumors. A prospective analysis of methodology, correlation, and
1153 usefulness based on clinical outcomes. *Acta Neurochir.* 152:11, 1835-46. doi:10.1007/s00701-
1154 010-0764-9
- 1155 Springer, J. A., Binder, J. R., Hammeke, T. A., Swanson, S. J., Frost, J. A., Bellgowan, P. S., Brewer,
1156 C. C., Perry, H. M., Morris, G. L., and Mueller, W. M. (1999). Language dominance in
1157 neurologically normal and epilepsy subjects: a functional MRI study. *Brain.* 122:11, 2033-2046.
1158 doi:10.1093/brain/122.11.2033
- 1159 Stevens, M. T. R., D'Arcy, R. C. N., Stroink,, C. D. B., and Beyea, S. D. (2013). Thresholds in fMRI
1160 studies: Reliable for single subjects? *J. Neurosci. Methods.* 219, 312–323.
1161 doi:10.1016/j.jneumeth.2013
- 1162 Stevens, M. T., Clarke, D. B., Stroink, G., Beyea, S. D., and D'Arcy, R. C. (2016). Improving fMRI
1163 reliability in presurgical mapping for brain tumours. *J. Neurol. Neurosurg. Psychiatry.* 87:3, 267-
1164 274. doi:10.1136/jnnp-2015-310307
- 1165 Stippich, C., Rapps, N., Dreyhaupt, J., Durst, A., Kress, B., Nennig, E., Tronnier V. M., and Sartor,
1166 K. (2007). Localizing and lateralizing language in patients with brain tumors: Feasibility of
1167 routine preoperative functional MR imaging in 81 consecutive patients. *Radiology.* 243:3, 828–
1168 836. doi:10.1148/radiol.2433060068
- 1169 Stoppelman, N., Harpaz, T., and Ben-Shachar, M. (2013). Do not throw out the baby with the bath
1170 water: choosing an effective baseline for a functional localizer of speech processing. *Brain*
1171 *Behav.* 3:3, 211-222. doi:10.1002/brb3.129
- 1172 Suarez, R.O., Taimouri, V., Boyer, K., Vega, C., Rotenberg, A., Madsen, J. R., Loddenkemper, T.,
1173 Duffy, F, Prabhu, S., and Warfield, S. K. (2014). Passive fMRI mapping of language function for
1174 pediatric epilepsy surgical planning: Validation using Wada, ECS, and FMAER. *Epilepsy Res.*
1175 108:10, 1874–1888. doi:10.1016/j.eplepsyres.2014.09.016
- 1176 Szaflarski, J.P., Holland, S.K., Jacola, L.M., Lindsell, C., Privitera, M.D., and Szaflarski, M. (2008).
1177 Comprehensive presurgical functional MRI language evaluation in adult patients with epilepsy.
1178 *Epilepsy Behav.* 12, 74–83. doi:10.1016/j.yebeh.2007.07.015
- 1179 Thivard, L., Hombrouck, J., Du Montcel, S.T., Delmaire, C., Cohen, L., Samson, S., Dupont, S.,
1180 Chiras, J., Baulac, M., and Lehericy, S. (2005). Productive and perceptive language
1181 reorganization in temporal lobe epilepsy. *NeuroImage.* 24:3, 841–851.
1182 doi:10.1016/j.neuroimage.2004.10.001.

- 1183 Unadkat, P., Fumagalli, L., Rigolo, L., Vangel, M. G., Young, G. S., Huang, R., Mukundan, S. Jr.,
1184 Golby, A., and Tie, Y. (2019). Functional MRI task comparison for language mapping in
1185 neurosurgical patients. *J. Neuroimaging* 29:3, 348-356. doi:10.1111/jon.12597
- 1186 Van Poppel, M., Wheless, J.W., Clarke, D.F., McGregor, A., McManis, M.H., Perkins, F.F. Jr., Van
1187 Poppel, K., Fulton, S., and Boop, F.A. (2012). Passive language mapping with
1188 magnetoencephalography in pediatric patients with epilepsy. *J. Neurosurg. Pediatr.* 10:2, 96-102.
1189 doi:10.3171/2012.4.PEDS11301
- 1190 Voyvodic J. T., Petrella J. R., and Friedman A. H. (2009). fMRI activation mapping as a percentage
1191 of local excitation: consistent presurgical motor maps without threshold adjustment. *J. Magn.*
1192 *Res. Imaging.* 29:4, 751-759. doi:10.1002/jmri.21716
- 1193 Wada, J., and Rasmussen, T. (1960). Intracarotid injection of sodium amytal for the lateralization of
1194 cerebral speech dominance. *J. Neurosurg.* 17, 266–282.
- 1195 Weng, H.H., Noll, K.R., Johnson, J.M., Prabhu, S.S., Tsai, Y.H., Chang, S.W., Huang, Y.C., Lee,
1196 J.D., Yang, J.T., Yang, C.T., Tsai, Y.H., Yang, C.Y., Hazle, J.D., Schomer, D.F., and Liu, H.L.
1197 (2018). Accuracy of presurgical functional MR imaging for language mapping of brain tumors:
1198 A systematic review and meta-analysis. *Radiology.* 286:2, 512-523.
1199 doi:10.1148/radiol.2017162971.
- 1200 Whalley, H. C., Gountouna, V. E., Hall, J., McIntosh, A. M., Simonotto, E., Job, D. E., Owens, D.
1201 G., Johnstone, E. C., and Lawrie, S. M. (2009). fMRI changes over time and reproducibility in
1202 unmedicated subjects at high genetic risk of schizophrenia. *Psychol. Med.* 39:7, 1189-99.
1203 doi:10.1017/S0033291708004923
- 1204 Wilke, M., and Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-
1205 data. *J. Neurosci. Methods.* 163:1, 128–136. doi:10.1016/j.jneumeth.2007.01.026
- 1206 Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., and Eriksson, D. K. (2016). Validity and
1207 reliability of four language mapping paradigms. *NeuroImage Clin.* 16, 399–408.
1208 doi:10.1016/j.nicl.2016.03.015
- 1209 Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., and Gerig, G. (2006).
1210 User-guided 3D active contour segmentation of anatomical structures: Significantly improved
1211 efficiency and reliability. *NeuroImage* 31:3, 1116–1128. doi:10.1016/j.neuroimage.2006.01.015
- 1212 Zacà, D., Nickerson, J.P., Deib, G., and Pillai, J. J. (2012). Effectiveness of four different clinical
1213 fMRI paradigms for preoperative regional determination of language lateralization in patients
1214 with brain tumors. *Neuroradiology.* 54, 1015–1025. doi:10.1007/s00234-012-1056-2
- 1215 Zhang, N., Xia, M., Qiu, T., Wang, X., Lin, C.P., Guo, Q., Lu, J., Wu, Q., Zhuang, D., Yu, Z., Gong,
1216 F., Farrukh Hameed, N. U., He, Y., Wu, J., and Zhou, L. (2018). Reorganization of cerebro-
1217 cerebellar circuit in patients with left hemispheric gliomas involving language network: A
1218 combined structural and resting-state functional MRI study. *Hum. Brain Mapp.* 39:12, 4802-
1219 4819. doi:10.1002/hbm.24324
- 1220 Zyryanov A., Malyutina S., and Dragoy O. (2020). Left frontal aslant tract and lexical selection:
1221 Evidence from frontal lobe lesions. *Neuropsychologia.* 147, 107385.
1222 doi:10.1016/j.neuropsychologia.2020.107385

1223 **Tables**

1224 Table 1. Dice coefficients in the two paradigms (SYLL: syllable baseline vs. PW: pseudoword
 1225 baseline) in three regions (frontal, temporal-parietal and frontal-temporal-parietal) at three statistical
 1226 thresholds (FWE correction for multiple comparisons at $p < .05$, cluster-size correction with a
 1227 minimum cluster size of $k \geq 200 \text{ mm}^3$ at $p < .001$, AT: adaptive thresholding as implemented in
 1228 Gorgolewski et al. (2012) at $p < .05$).

| Dice coefficients | | | | | | | |
|-------------------------------|------|-----|--|------|-----|--------------|-----|
| Frontal | | | | | | | |
| | FWE | | | AT | | Cluster-Size | |
| | SYLL | PW | | SYLL | PW | SYLL | PW |
| Baseline | | | | | | | |
| 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 | |
| | SYLL | PW | | SYLL | PW | SYLL | PW |
| Baseline | | | | | | | |
| 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 | |
| | SYLL | PW | | SYLL | PW | SYLL | PW |
| Baseline | | | | | | | |
| 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 |

1229

1230 Table 2. Lateralization indices for each paradigm (SYLL: syllable baseline vs. PW: pseudoword
 1231 baseline) in three regions (frontal, temporal-parietal, frontal-temporal-parietal) along with results of
 1232 Spearman's correlations between LIs in the two scanning sessions. Significant correlations ($p < .05$)
 1233 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 | $r = .401, p = .099$ | | $r = .490, p = .038^*$ | |

1234

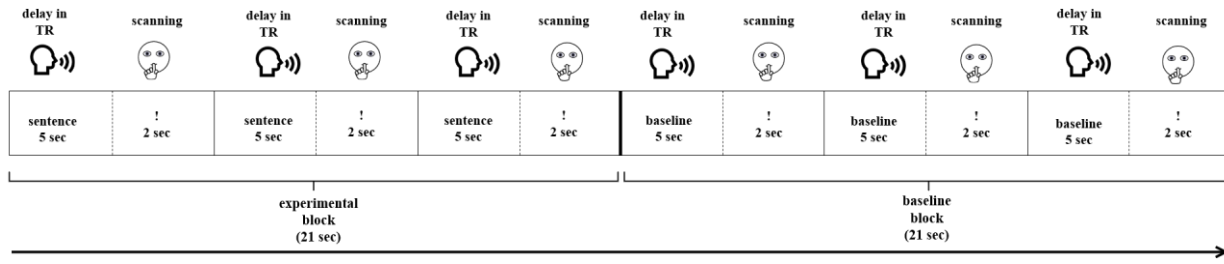
1235 Table 3. Comparison of Dice coefficients to previous fMRI paradigms using language tasks in
 1236 neurologically healthy participants.

| 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 | |
| Morrison et al. (2016a) | Sentence completion | .27 to .47 | Group average, whole-brain |
| | Phonemic fluency | .36 | |
| Nettekoven et al. (2018) | Rhyming | .54 | Group average, depending on statistical threshold |
| | Picture naming | .47 or .60 | |

1237

1238 **Figures**

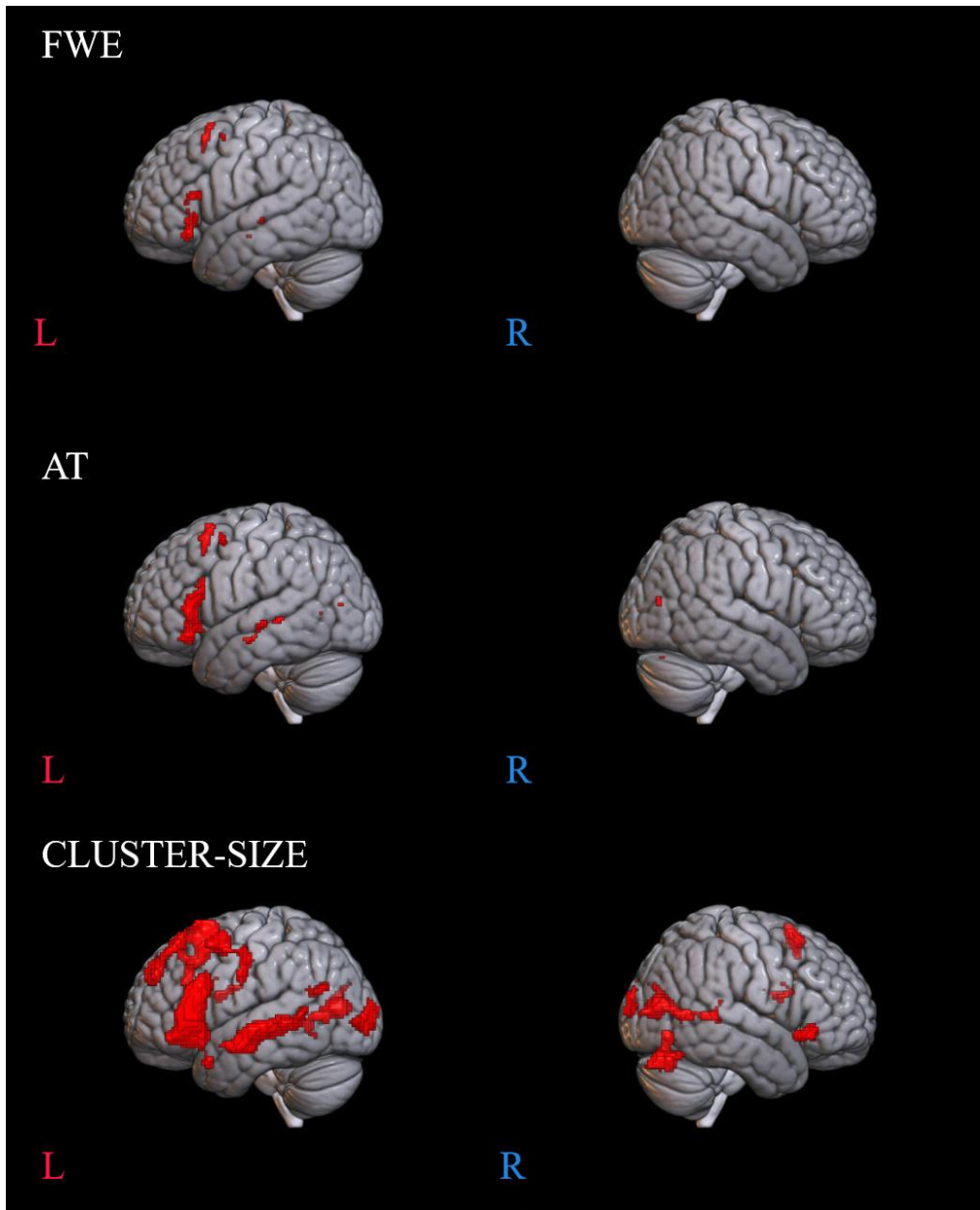
1239 Figure 1. Experimental design of the functional paradigms.



1240

1241

1242 Figure 2. Language-related activation in the SYLL paradigm (paradigm with the syllable baseline).
1243 Top: FWE correction for multiple comparisons at $p < .05$, middle: adaptive thresholding (AT) as
1244 implemented in Gorgolewski et al. (2012) at $p < .05$, bottom: cluster-size correction for multiple
1245 comparisons with a minimum cluster size of $k \geq 200 \text{ mm}^3$ at $p < .001$.

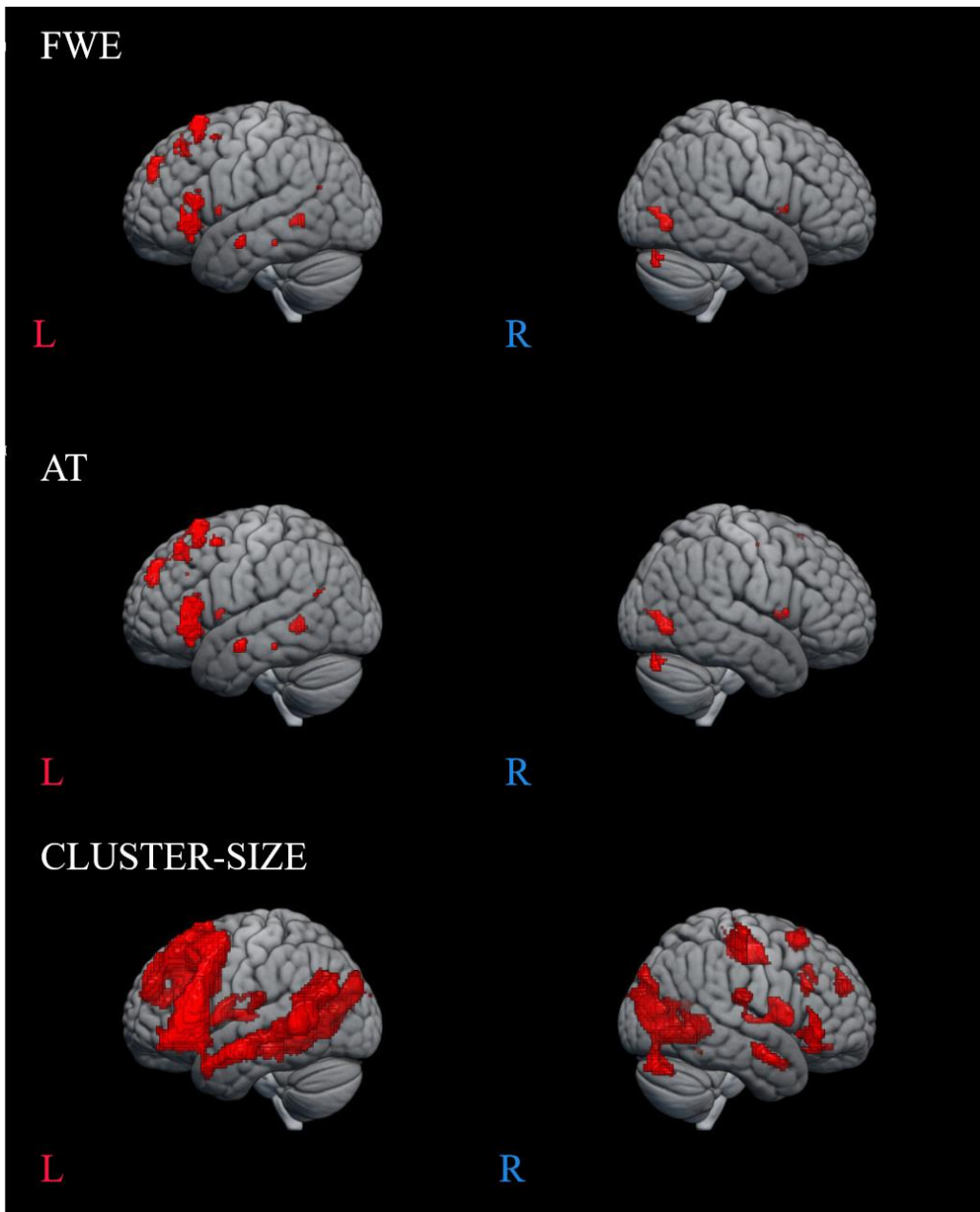


1246

1247

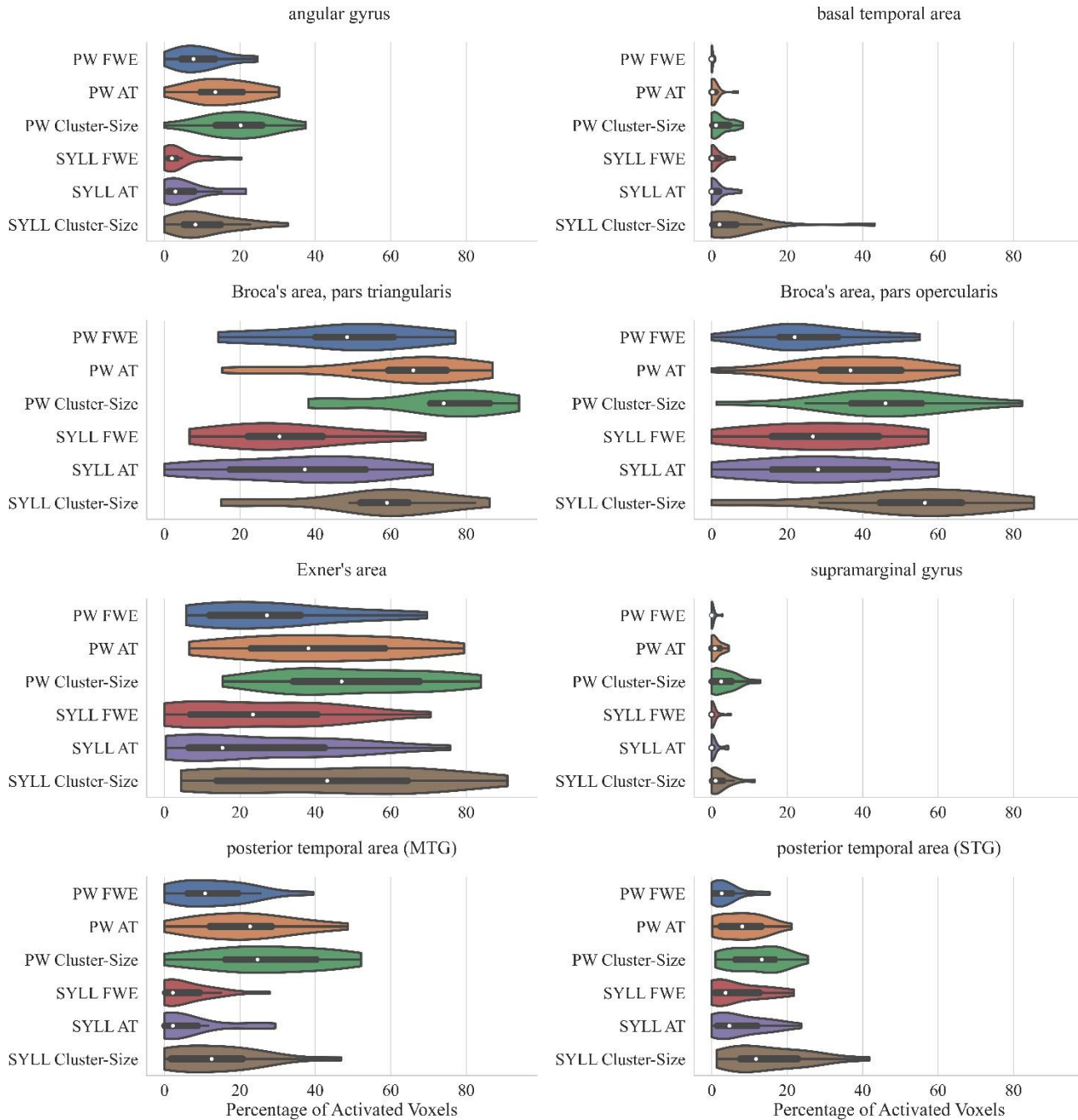
1248

1249 Figure 3. Language-related activation in the PW paradigm (paradigm with the pseudoword baseline).
1250 Top: FWE correction for multiple comparisons at $p < .05$, middle: adaptive thresholding (AT) as
1251 implemented in Gorgolewski et al. (2012) at $p < .05$, bottom : cluster-size correction for multiple
1252 comparisons with a minimum cluster size of $k \geq 200 \text{ mm}^3$ at $p < .001$.



1253
1254

1255 Figure 4. Percentage of significantly activated voxels in 'key language-related areas' adopted from
 1256 Benjamin et al. (2017) depending on the paradigm (PW – pseudoword baseline, SYLL – syllable
 1257 baseline) and threshold (FWE correction for multiple comparisons at $p < .05$, cluster-size correction
 1258 with a minimum cluster size of $k \geq 200 \text{ mm}^3$ at $p < .001$, AT: adaptive thresholding as implemented
 1259 in Gorgolewski et al. (2012) at $p < .05$). The white dot in the middle of each 'violin' represents the
 1260 median value and the thick black bar in the center represents the interquartile range.



1261