1.4 Research questions

1.4.1 Expected discoveries

Replication of EEG/fMRI results with MEG? Which cortical regions are involved in the conditional effect? Is semantic content relevant? How long does each processing stage take? In which order do the steps take place? Can we see a different pathway in kids?

1.4.2 Hypotheses

Condition effect mainly in pSTG, BA44 (adults), BA45 (kids) Kids: worse performance than adults Kids: less involvement of pSTG Adults vs kids: Dorsal II vs. Ventral II

1.5 Choice of measurement methods

1.5.1 Acquisition

Neuroanatomic principles Neuronal activity creates a combination of electrical and magnetic fields. On a cellular level, activating a neuron causes depolarization, which in turn creates a weak electric field. Most neurons are equipped with a long axonal fibre that transmits a relatively strong postsynaptic signal. Especially in pyramidal cells which are responsible for long-distance transmissions, this axon can span several centimeters in length. Since signals along axons travel by a complex combination of transmitter binding, ion flux and electric fields, neuron-to-neuron data transmission exhibits three important restrictions. First, a successful signal transmission requires a short refractory period until the next transmission is possible again. This leads to the phenomenon that transmissions can only travel one way, and overlapping signals on the same fiber are impossible. Second, axonal transmissions can only be binary. If the minimum threshold voltage is reached, the signal will be transmitted at maximum speed along the entire fibre. Below the threshold, no transmission can occur. Third, maximum transmission speed is relatively low at approximately $100\frac{m}{s}$. Other commonly used transmission fibers, for comparison, achieve speeds

of $2 \cdot 10^8 \frac{m}{s}$ (copper wire) or even $3 \cdot 10^8 \frac{m}{s}$ (glass fibre). Additionally, transmision speed depends on the thickness of the fiber, with the thinnest fibers transmitting as slowly as 0.1m/s. This property causes a considerable delay when transmitting signals over macroscopic distances. Signal delay in technical applications is generally small enough to be disregarded or considered an inescapable nuisance. In the brain however, two interconnected neurons at opposite sides of the head can generate output simultaneously, but their signals will be offset by several miliseconds by the time they arrive. For any neural network, delayed information is therefore an widely expected issue and must be properly factored into signal processing.

Available acquisition methods Due to the long axon and the low transmission speed, transmitted signals will create a electrical field that moves along the axon during a non-trivial time window. Aggregated electric fields from thousands of similar signals can be acquired with an electroencephalograph (EEG). This process involves measuring the voltage at two or more arbitrary points in the brain; typically, on the surface of the head or the cortex (which is called intracranial EEG, or iEEG).

Every electric signal that travels along a conducting wire also induces a magnetic field. Since an axon is no exception to this rule, any transmitted neuron-to-neuron signal can also be acquired with a magnetic sensor. A magnetoenecephalograph (MEG) consists of several dozens of these sensors distributed around the head. MEG and EEG are currently the only non-invasive acquisition methods that measure brain activity with a milisecond resolution [1.4.MEG.a][1.4.MEG.b][1.4.MEG.c]. The focus of this project is on cognitive processes that require timespans in the single-digit second range to complete. Considering these small time frames, a good temporal resolution is integral for yielding a sufficient amount of data from every processing step. This is the reason why I decided to use a EEG/MEG method, while using the highest available temporal resolution.

Choice of acquisition methods Compared to EEG, the MEG-based measuring strategy has a few advantages and drawbacks.

The first difference between MEG and EEG is due to the sensor technology. While both methods rely on strong amplification of very small input signals, only MEG uses superconducting sensors. Contemporary superconductors require cooling with liquid helium. Magnetic fields from neurons are also weaker than environmental magnetic noise by several magnitudes. To limit measurements to neural activity, the MEG device need to be shielded with large quantities of highly magnetically permeable material (most commonly, an nickel-iron alloy). These requirements makes MEG much less portable, and equally more expensive, than EEG. Passive shielding already reduces environmental noise levels by 25-60db [1.4.SNR]. In addition to that, it is possible to dampen noise levels by another 60db with adaptive noise reduction [1.4.SNR]. Good adaptive noise reduction requires large external coils that can counteract outside magnetic fields. The use of superconducting MEG sensors is necessary due to the extreme weakness of neural magnetic fields. With the large amplification factors involved in both methods, amplifiers contribute a large amount of noise to the signal. But since the amplifier noise depends directly on their operating temperature, suspending the amplifiers in liquid helium drastically lowers the noise level [1.4.MEG.a]. Generally, EEG and MEG are considered equally sensitive [1.4.MEG.a][1.4.MEG.c]. However, the addition of magnetic shielding can elevate signal-to-noise ratios in MEG measurements above equivalent acquisitions from EEG [1.4.SNR].

The second difference between MEG and EEG is the drastically different distortion from surrounding tissue. For an EEG to be able to measure a potential difference on the skin surface, an electrical current needs to pass the tissues surrounding the cortical surface [1.4.tissues.b]. Some of these tissue layers, like blood vessels or cerebreal spinal fluid (CSF), are 5 times more conductive than gray matter; so they smooth and diffuse electric fields [1.4.tissues.a][1.4.tissues.b]. Other tissue layers, especially the compact bone, conduct electricity 78 times worse than gray matter; so they distort and attenuate every passing signal [1.4.tissues.a]. After these tissues have been passed, the original signal has substantially decreased in intensity, and changed drastically in shape and location. In contrast to electrical fields, the same tissues are highly permeable to magnetic fields. The magnetic

permeability of water, which most human tissue is based on, differs from the magnetic permeability of vacuum only by 0.0008%. As a general rule, only metal-based materials are substantially less permeable than water. Since the human head doesn't contain metal in considerable quantities, it practically allows magnetic fields to pass without distortion. [1.4.tissues.a]

This circumstance does not imply, unfortunately, that every neural transmission arrives at the magnetic sensors with equal strength. There are three main reasons for that.

First, the main source of electrical and magnetical activity are the pyramidal cells in the cortical tissue on the brain surface. The activity of a single neuron, besides being highly unlikely in vivo, doesn't create a field with sufficient strength to elict a response in contemporary EEG or MEG sensors. Only the combined and synchronous activity of larger clusters of neurons can cross the detection threshold.

Second, the human cortex is folded into gyri and sulci. When two neural groups at opposite cortical walls produce identical activity, the two created electric and magnetic fields are directly opposed to each other. Opposing fields cancel each other out, so the original signal will be systematically underestimated by surrounding sensors.

Third, their basic physical properties imply that electric and magnetic fields are orthagonal to each other. Signals that travel along fibers orthogonally to the head surface create the strongest magnetic activation in surrounding sensors, but the weakest voltage in surface electrodes. Fibers that lie parallel to the head surface, in contrast, create strong electric but weak magnetic activation. The implication on measurements of neural activity is that EEG is most sensitive for gyri and sulci, and MEG is most sensitive for the radial walls inbetween. To counteract this particular measurement bias, EEG and MEG data would have to be acquired simultaneously.

Although the simultaneous acqusisition of EEG and MEG data provides theoretical benefits to the signal quality, I ultimately decided against this strategy. MEG acquisition consists of three preparation steps: Getting written consent, applying the HPI coils and ocular electrodes, and digitizing the head. This preparation typically requires 25 minutes

for children and 15 minutes for experienced adults. Including EEG acquisition would have added several lengthy steps to this procedure: Fitting the gel electrode cap, ensuring a good connection for each electrode, and plugging in each of the 63 cables individually. These additional steps would have extended the preparation time to at least 60 minutes. Since patience is not a strong trait in children[1], I wanted to minimize any idle waiting times between their arrival and the experiment. Therefore, I only acquired MEG data during this study.

1.5.2 Preprocessing

Once the signals were acquired by the MEG, there were two necessary decisions for preprocessing.

DC bias removal The first decision concerns the removal of DC bias that is often caused by slow sensor drift. Traditionally, for the exploration of effects in ERP and ERF, a subtractive baseline correction is applied to every trial before averaging. For this purpose, the average is computed from roughly 100 to 500 (typically 200) miliseconds of activity before the conditional cue. This initial interval is assumed to originate from brain activity unrelated to the post-cue task. The computed average value is then subtracted from activity data in the corresponding trial. This procedure assures that different DC components from long-term trends (for either technical or cognitive reasons) don't disturb the trigger-dependent effect. However, there is a fundamental issue with the baseline correction. Because the stimuli are spoken sentences in this study, there is no silent time window around the critical words. Therefore, the pre-cue interval reflects electric fields from unrelated brain activity. By computing the average from unrelated activity, I effectively introduces a random DC error into the correction procedure. This issue makes a subtractive baseline correction a tradeoff between the original DC error and a random DC error. A process to remove DC components without this tradeoff is to use a highpass filter [1.4.highpass]. Since my longest expected

[1] Implementing details with the goal to prevent boredom was a common theme in this study, as explained in more detail in chapter 3.

evoked field, the ELAN, has a base frequency of 5Hz, I choose a much lower value of 0.4Hz for the high-pass.

Artifact removal The second decision concerns the removal of various measuring artifacts. There are four major types of artifacts during the acquisition of electric or magnetic fields.

The first type of artifact is caused from cardiac activity. Cardiac muscles create an almost continuous, very regular field with low frequency and medium strength.

The second type of artifact is caused by ocular movements. Since the eyeballs are electrically charged, all eye movements are associated with a continuously changing field of low frequency and medium strength.

The third type of artifact is caused by muscle movements. Muscle activity creates relatively long and strong distortions in a wide frequency band.

The final type of artifact is caused by oversaturation in MEG sensors. Oversaturation happens randomly when no high pass is in use, and reduces the sensitivity of the affected channel to zero. This condition is remedied with an automatic reset, which in turn produces a single very short and very large jump in amplitude.

The first two artifacts can be eliminated with the help of three additional acquisition channels. With electrodes attached to the chest and to the eye sockets, electric fields from ocular and cardiac activity are measured directly. MEG data is then deconstructed into independent data components with an independent component analysis. The artifact channels are used to identify artifact components in the measured MEG data. If the extracted data component is similar enough to one of the measured artifact channels, it is removed. The remaining components are then assembled to a data composition, ideally containing no cardiac or ocular artifacts.

The last two artifacts can be removed with a simple threshold detection. Their high amplitude make it possible to set a manual amplitude threshold, and reject segments that

exceed this threshold in any channel. I determined the threshold manually after visual artifact inspection, and decided to reject entire trials if this threshold is exceeded.

1.5.3 Timewindow estimation

The acquired signals need to be explored for the impact of the conditional effect. This effect is usually spatially and temporally limited. For establishing time intervals (TOI) and spatial regions of interest (ROI), there are two possible approaches.

First, existing literature can be consulted for activity effects from syntax contrasts in similar experiments.

Second, a bootstrapping approach can be used. For this approach, the measured activity is compared between conditions. The TOI and ROI that involve considerable contrast between conditions can then be selected for the comparison of mean activity. The drawback to this approach is both spurious contrast and activity from different cognitive processes are considered as condition effect. The statistical testing will therefore systematically overestimate the condition effect. This issue is known as “double dipping“ [1.4.Kriegeskorte]. However, for exploratory analysis, bootstrapping is a valuable tool that can uncover previously unknown TOI and ROI.

I decided to use both approaches with different purposes: First, comparisons within previously discovered ROI and TOI provide results that can be compared well to the findings of earlier studies. Second, a bootstrapped comparison allows for the exploration of spatial and temporal properties of the syntactic effect.

1.5.4 Source localization

Motivation Magnetic fields, when induced by the brain, arrive at MEG sensors only as mixture of many cortical sources. Demixing these signals is a fundamentally flawed process. The problem of discerning these signal sources is equivalent to finding the location and intensity of all the flames in a hot air balloon, while only looking at the outside of the hull. There are infinitely many possible configurations of light sources that can generate the

same brightness pattern on the outer skin. The same issue is valid for localizing magnetic signal sources in the human head. This task involves creating a bidirectional map between the curved plane of MEG sensors and the threedimensional human head. Because of the different dimensionality, the task of creating this map is an underdefined problem. This means that there are infinitely many possible locations and intensities for magnetic fields that can generate the exact same signal pattern in the MEG sensors. This multitude of possible solutions needs to be constrained to make the results meaningful. One popular set of constraints is the use of a source model.

Choice of source model A source model assumes that there is a limited number of discrete current sources distributed throughout the brain. Usually, these current sources are assumed to be generated by neuronal tissue. There are three popular types of source modelling.

The first type of source modelling uses spatial filtering. A popular spatial filtering strategy is the single-core beamformer method (Barnes and Hillebrand, 2003; Gross and Ioannides, 1999; Gross et al., 2001; Hillebrand and Barnes, 2003; Robinson and Vrba, 1999; Sekihara et al., 2001; Van Veen et al., 1997). Its main weakness is the assumption that data from different sources is completely uncorrelated. This assumption is especially detrimental to the analysis of cortical signals, since neuronal-level synchronizity is one of the fundamental principles behind attention and learning (Kandel et al., 2000).

In the focal source model, neural current flow is represented by a limited set of point-shaped current dipoles. There are are three subcategories to this model: an unconstrained variant (the moving dipole model), dipoles with a fixed position (the rotating dipole model) and dipoles with a fixed position and rotation (the fixed dipole model). Popular applications of this approach include “multiple signal classification“ (MUSIC) [1.4.music] and “multi-start spatio-temporal multiple-dipole modeling“ [1.4.simplex] Dipole models have had limited success with representing neuronal responses for two main reasons. First, reducing extended neuroanatomical structures to a point current source introduces a system-

atic model error. Second, the number and location of dipoles has a strong influence on the localized results, yet is hard to estimate in advance (Huang et al., 1998).

The third type of source model is the distributed source model. For these approaches, a dense grid of dipoles is derived from a cortical layer. The goal is to place dipoles in homogenous density at every location that is able to produce currents. Typically, the continuous cortex surface is extracted from anatomical data and populated with several thousands of (roughly) equally spaced dipoles. For determining localized activity, moments are computed for all dipoles. The dipole moments are then used to simulate activity in the MEG sensors. Simulated activity then is optimized so that the error to the reference sensor activity is minimal. Because every sensor activity pattern can be created by infinitely many source configurations, the localization process is facilitated with two processes. First, dipole activity is spatially regularized with a predefined factor. Second, the best pattern of localized sources is selected by minimizing the norm over all dipoles. The most popular norm today is the L2-norm [1.4.L2], and implementations are widely available. Alternatively, the L1-norm [1.4.L1a, 1.4.L1b] can result in a more focally reconstructed activity. This process is usually computed separately for every temporal sample. Popular implementations include dSPM (Dale et al., 2000), MNE (Hammalaien, 2005) and sLORETA (Pascual-Marqui, 2002)). I opted for this approach because the spatial filtering approaches aren't recommended for localizing cortical activity, and the quality of results from the dipole fit models depend too strongly on the inital parameters.

L2-norm-based solutions have two major drawbacks. First, the solution has a relatively low spatial resolution. This issue leads to spatially distributed activity clusters even if the real sources are very focal. If the sources are in close proximity, unintended mixing of reconstructed source activity can occur as well. Second, generic L2-normal solutions contain a mandatory systematical spatial bias. The sLORETA algorithm, by contrast, has been designed to create solutions with zero bias. Since this algorithm has minimal drawbacks out of all readiliy available software, it became the method of choice for my source localization purposes.

1.5.5 Information transfer

After localization, activity in my selected cortical regions is ready for the analysis of transfered information. According to the Wiener principle [1.4.information], information processing consists of three separate tasks: information storage, information manipulation and information transfer. In comparison to true causality, which involves all three tasks, the analysis of transfered information is feasible on discrete electrophysiological data. There are two fundamentally different approaches to this task.

Chapter 2

Pilot Study

Experimental paradigm The pilot study was designed with two goals: First, to find any strong inherent bias in the stimuli material. Especially the choice of animal pairings (predators, prey and insects mixed with each other) was a potential confounding effect. It was also unclear if the different types of actions (predatory and antropomorphic/social) could introduce a response bias.

Second, to determine whether 10-year-old children are the correct age bracket for examining our research questions. Especially important was to test the ability of our subjects to answer questions in object-relative clause correctly. I expected that our subjects could solve the task correctly, but with a drastically less-than-perfect performace.

I employed a repeated measures factorial design with one factor: syntax order. The two conditions of this factor differed in the use of either subject-relative or object-relative clauses.

Participants 21 children ([9 female]) were recruited from the internal participants database. Subjects were selected if they spoke German as native language, if their language development was unremarkable, if their handedness score was above 70 and if their medical history was free of cognitive abnormalities. Children were aged between [10y0m] and [10y11m] and described as right-handed by their parents. Parents gave written informed

consent and were compensated with 7,50€. Children agreed to participate in the study and were compensated with a toy at an approximate value of 12€. All experimental procedures were approved by the University of Leipzig Ethical Review Board.

Task and Stimuli The session consisted of two sections: a tutorial and the main experiment. The tutorial section introduced the physical interface and 36 practice trials. The main section consisted of 18 clusters with 12 trials each (216 total). All trials were randomized at the beginning of the experiment. There was an exception to the random order: no two identical images were presented after another. Subjects were shown a feedback screen at the end of each cluster.

Each trial started by showing a set of visual stimuli. 200ms later, a spoken sentence started playing. Possible responses included pressing the left, right or the skip button. The trial ended with an auditory and visual feedback.

A set of visual stimuli consisted of a two side-by-side images on black background. Each image depicted two cartoon animals on a white background. In one picture, one of the animals performed a social action on the other animal. In the other picture, the roles were reversed. Subjects needed to identify the picture that answered the question correctly.

The spoken sentences were always posed as questions in order to elicit an immediate response. Each question was posed in a right-branching structure (see table 2.1 for an example). Object-relative clauses and subject-relative clauses were presented in random order at a ratio of 2.5:1. Questions were voiced in a natural, child-directed tone by a professional native speaker. The question consisted of an actor, a recipient and the verb that described their interaction. The performed action was either painting, pushing, combing, washing, pulling or catching. The actor and recipient were selected randomly from 12 different species: Lion, rabbit, wolf, bird, fox, hedgehog, dog, tiger, ape, ladybug, bear and frog. No two identical species appeared in the same picture.

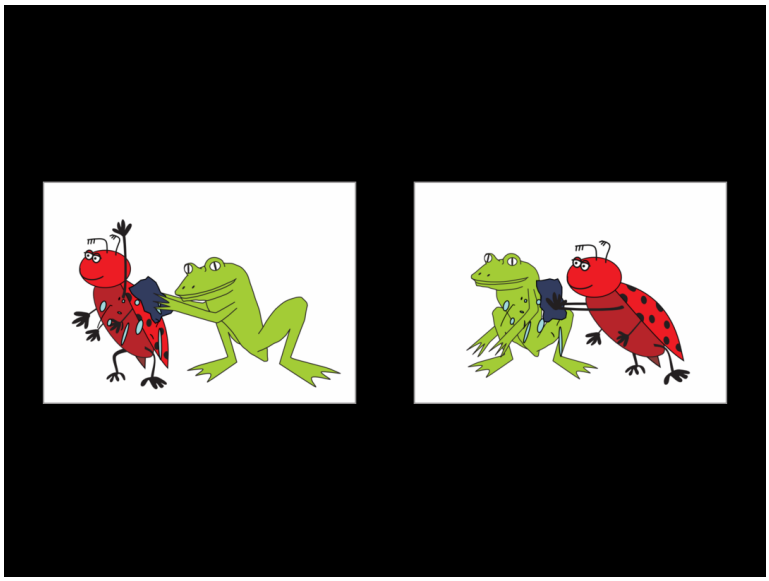

Figure 2.1: Example of a typical visual stimulus before the response

Original	Wo	ist	der	Käfer,	den	der	Frosch	wäscht?
Translated	Where	is	the	bug_{OBJ},	who_{ACC}	the	frog_{SUBJ}	washes?
Word index	1	2	3	4	5	6	7	8

Table 2.1: Example stimulus sentence. Top: original spelling in German. Middle: Literal translation in English. Bottom: Word index within the sentence

Immediately after each response, an icon appeared below the two animal pairs. A green checkmark, a diagonal red cross and a yellow skip symbol signified a correct response, an incorrect response and an invalid trial, respectively. The trial feedback screen was presented for a random interval between 400ms and 800ms.

This experiment created a tradeoff between speed and accuracy. To encourage a high level of attention and a high number of usable trials, a feedback screen was displayed at the end of each cluster. On this screen, two bar graphs visualized response speed and accuracy during the preceding cluster.

Visual and auditory stimuli were produced by a computer running the software package Presentation (Neurobehavioral Systems, Inc., version [14.6]). Video signal was displayed

by a 17-inch TFT display at a distance of approximately 80cm. Sound was played with a pair of semi-open headphones.

Analysis Behavioral data were analyzed with Matlab (version 2014a). Response accuracy was evaluated for the case that subjects responded randomly. For this purpose, accuracy was compared to the outcome of a random sequence of binary events. I established the alpha = 0.01 confidence interval for the percentage of correct trials ($\frac{k}{n}$) that could be answered correctly purely by chance. This calculation was implemented using the binomial fit method in Matlab: binofit(k, n, alpha). If the upper confidence interval of this calculation exceeded the subject-specific accuracy, the subject was removed from further analysis. Two subjects failed to exceed chance level performance, leaving 19 subjects for the analysis.

Two types of behavioral data were analyzed for group and condition effects: response time (RT) and response accuracy (RA). Response time was measured at the condition onset, i.e. at the "d" sound of the sixth word. Trials were omitted when the subject skipped or answered them incorrectly, or responded earlier than the cue. Accuracy was calculated by dividing the amount of correct trials by the amount of total trials for each subject.

A Shapiro-Wilk test was used to test for normal-distributed residuals. Accuracy passed this test at a p = 0.01 significance level. The impact of the syntactic condition on response accuracy was determined with a T-test.

Any unintended bias on response time was determined by splitting RT from each subject into two groups along one of five impact factors. The two groups were then compared with a T-test. The five grouping factors were response side (left / right), condition (object-relative / subject-relative), the race of the actor (12) and recipient (12) and the peformed action.

Results Subjects responded after a median delay of 2.0s (quickest 5%: 1.0s, slowest 5%: 5.2s). The median accuracy was 70% (worst 5%: 62%, best 5%: 94%). The syntactic condition had a highly significant effect on accuracy ($p < 0.001$, $t(18) = -14.0$). No factor had a significant effect on response time ($p > 0.1$, $F < 1.5$).

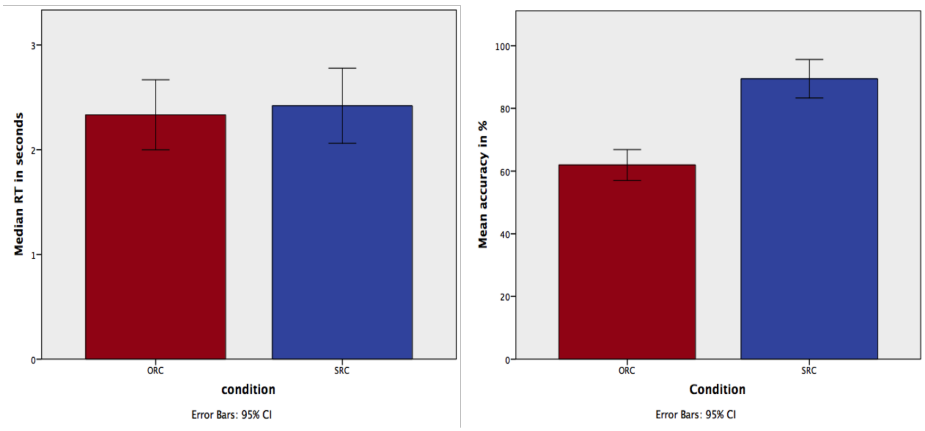

Figure 2.2: Chart of the grand average performances for subject-relative (blue) and object-relative (red) clauses. Left: Median response time from conditional onset. Right: Average response accuracy.

Post-hoc tests were conducted to reveal condition-specific accuracy values. The average accuracy for responses to subject-relative clauses was much higher than to object-relative clauses (93% and 64%, respectively). Due to the proximity to a 50% chance level, I repeated the binomial fit test of individual accuracy, exclusively with responses to object-relative clauses. 8 of 19 subjects failed this test at a $p = 0.01$ significance level. After removing these subjects from the analysis, selected tests were repeated.

Median response times remained unchanged for both conditions. Overall accuracy in the remaining subjects improved slightly (95% for subject-relative clauses and 68% for object-relative clauses). The difference in accuracy between conditions weakened slightly ($p < 0.001$, $t(10) = -9.4$).

Discussion The first goal of this pilot study was to establish an unbiased test paradigm. The variation in grammatical elements had no undesired impact on response times. Syntax conditions produced a strong effect in both performance metrics. These findings support the current stimulus setup for use during MEG measurements.

The second goal of the pilot study was to determine if 10-year old children were suitable subjects for the designed task. Accuracy levels were comparable with the findings of similar experiments. Due to different trigger points and sentence lengths, comparisons of response times couldn't be made directly and were instead limited to comparisons of effect size. I start by comparing our results to our spiritual predecessor, the study by [2.1]. They found no significant conditional impact on response time in the 9-10-year age bracket. In stark contrast to our results (93% and 64% for subject- and object-relative clauses), their subjects were not influenced by a condition effect, with accuracy levels of 94% in both conditions. The good performance was met with surprise and speculation that semantic cues may have helped with sentence comprehension. This speculation was supported by their fMRI findings, which indicated that children relied heavily on semantic-centric processing areas, rather than on pure syntax-related processing areas that adults use. Our setup didn't include semantic cues, which puts our results more in line with other infant studies.

[2.2], for instance, measured repetition performance in 3- and 4-year-old children. The study presented German subject-relative and object-relative clauses with two interacting people in third person, similar to our setup. Their subjects performed with an accuracy of just 5% (3 years) and 26% (4 years) for object-relative clauses. Subject-relative clauses were performed with 13% and 31% accuracy, respectably. Note that these ratings represent accurate verbal repetition, and don't need to be corrected for chance level performance.

[2.3] found more accurate responses to portuguese right-branching subject- and object-relative clauses. Subject-relative clauses were correctly answered 23%, 50%, 83% and 80% of the time (in 3-, 4-, 5- and 6-year old children, respectably). Object-relative clauses reached only slightly lower accuracy levels of 18%, 40%, 75% and 68%, respectably. They employed a more user-friendly approach by requiring the children to act out the posed sentences with toy animals. Because the authors deliberately removed as many processing constraints as possible, these accuracy levels can be considered upper performance limits on these age brackets.

6 to 8 year old children solved syntactic problems with higher complexity in the study by [2.4]. The experimental design varied two syntactical factors and one contextual factor. One syntactical factor, “question“, varied English object- and subject-relative clauses, coinciding with our setup. Object-first and subject-first clauses were responded with an accuracy of 64% and 83%, respectably.

Response time and accuracy performance indicates that the 10-year age bracket was successful in selecting subjects with an incomplete dorsal tract II. Compared to typical adults, our subjects performed considerably worse in both accuracy and response time. Compared to younger children, our subjects performed condsiderably better in subject-relative clauses.

42% of our subjects failed to perform better than chance level when the task required a fully-developed AF. I have no sufficient reason to assume that these subjects were using a random-button strategy. If anything, the weakened condition effect indicates that these performances were due to honest mistakes. Hence, there is no reason to exclude subjects with chance-level performances to object-relative clauses.

Finally, the pilot study revealed a few design flaws as well.

First, sentences differed systematically between conditions even before the intended condition cue. The syntactical condition reverses the order of pronouns, creating an opposition between “der den“ and “den der“. To prevent confounding effects from different content, the initial sentence fragment needs to be identical at least across conditions, and, preferably, across trials as well. Ideally, the distance between the end of this identical sentence fragment and the conditional cue point should be as short and invariant as possible.

Second, our sentence structure contained a theoretical loophole. With sufficient time and wit, subjects could develop an alternative strategy that doesn’t require syntactic processing of the whole sentence. The alternative strategy exploits the fact that the minimum information for a correct decision is already available at the fifth word (see table **??**). Subjects only needed to complete three sequential steps: First, attending only to the left of the two pictures. Second, waiting until the fourth word is spoken. If the mentioned animal was displayed as actor, the left button would be correct and vice versa. Third, using the fifth

word to execute the previous or the reverse button mapping. If the mentioned word was a “der“, the previously correct button remained correct and could be pressed immediately. If the mentioned word was a “den“, the previously wrong button had to be pressed for a correct result. This strategy could potentially reduce the task into a simple series of motor preparation and pattern matching. Complex syntactic processing and, presumably, the use of pSTS and dorsal pathway II would be circumvented. Employing this strategy could create suspicious behavioral results in the form of systematically reduced and less varied response times. This flaw was the main reason for the redesign of stimuli for the main study.

Chapter 3

Methods

3.1 Participants and stimuli

3.1.1 Participants

18 children and 22 adults were recruited from the internal participants database. Subjects were selected if they spoke German as native language, if their language development was unremarkable, if their handedness score was above 70, if they fulfilled the prerequisites for MRI scans, and if their medical history was free of cognitive abnormalities. 4 children and 4 adults dropped out inbetween sessions of the study. Two children were excluded from the analysis because their behavioral performance was at chance level. 12 children (5 female) and 18 adults (9 female) were left for the subsequent analysis. Children were aged between 9y11m and 10y9m and described as right-handed by their parents. Adults were aged between 22 and 33 years and scored between 73 to 100 (median: 95) on the laterality quotient test (?, ?). The point of reference for these ages is the time of the MEG session. Parents gave written informed consent and were compensated with 40€ for the MEG session and 7,50€ for the MRI session. Children agreed to participate in the study and were compensated with a 10€ gift voucher for each session. Adult participants were compensated with 20€. All experimental procedures were approved by the University of Leipzig Ethical Review Board.

3.1.2 Task

The study consisted of two sessions: an anatomical MRI acquisition (duration: 50 minutes) and an interactive magnetoencephalographic measurement (typical duration: 90 minutes). Since they took place in two different locations, there was a delay (median: 98 days, maximum: 243 days) between the two sessions.

The MRI session is described in detail in section 3.2.2.

The MEG session consisted of two sections: a tutorial section and a main section.

MEG tutorial section First, the tutorial section described the usage of the interface.

Second, subjects needed to respond to an example stimulus with the spoken sentence written out below the screen.

Third, three example trials followed without the written sentence.

Fourth, an artificially incomprehensible sentence was presented together with otherwise innocuous visual stimuli.

When subjects pressed either response button instead of skipping the trial, they were instructed with the skip function.

Finally, a series of randomized tutorial trials followed. When subjects showed behavioral proficiency of the task, the tutorial ended prematurely. Two thresholds for proficiency were possible: either an average response time below 3000ms and an accuracy score above 80%, or an accuracy score above 88%. Either threshold could only be reached after completing at least 5 or 8 trials, respectively. When none of these thresholds were met, the tutorial ended after 36 trials.

MEG main section The main section was used for MEG acquisition and consisted of 304 trials grouped in two blocks. There was a scheduled break between the blocks (usually 1-2 minutes) which included interaction with the research assistant. Subject-specific trial randomization was performed before the task. All stimuli-related randomization tasks were implemented with a time-seeded Mersenne-Twister approach in Python 2.7. Randomization

contained two exceptions: neither the same image nor the same sentence could be played twice in a row. Each block consisted of 8 clusters. Subjects were shown a feedback screen at the end of each cluster, summarizing their performance throughout the recent cluster. Since manual intervention was necessary to proceed to the next cluster, subjects frequently used this opportunity for a tiny break (typically 5-20 seconds). Each cluster consisted of 19 trials.

Structure of a single trial Each trial started by showing two pictures side-by-side. In one picture, one of the animals performs a social action on the other animal. In the other picture, the roles are reversed. 10ms later, the spoken question started playing. The subject could respond by pressing one of the direction buttons or the skip button. There were two direction buttons, left or right, signifying that the left or right image contained the answer to the question. The skip button was used to mark the trial as invalid for further analysis, and excluded the trial from performance feedback. This response was the correct choice when the subject was distracted or failed to comprehend the question immediately. This opened a minor pitfall: subjects could have gotten perfect scores by just pressing the skip button each time. Fortunately, none of the subjects discovered this opportunity. The trial ended with an auditory and visual feedback.

3.1.3 Visual stimuli

Character motivation A set of visual stimuli consisted of a two side-by-side images on black background. Each image depicted two different animals on a white background. I selected selected social activities that were only plausible for antropomorphised characters, not for their real animal counterparts. Antropomorphization includes the use of their front limbs for object manipulation and standing on their hindlegs. These measures are introduced to prevent associations with real-world animalistic behavior. For example, a lion “catching“ a monkey could resemble predatory behavior. This association with chasing and killing would introduce a semantic bias against the reverse interaction: a real-life mon-

key “catching“ a lion is much more implausible than the reverse. To further detract from a naturalistic view, the animals were represented in a cartoon style. To prevent unnecessary stress on this interpretation, I only selected animals whose real-life counterparts were approximately equally sized.

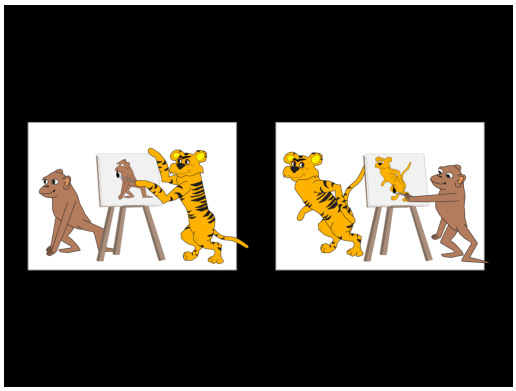

Figure 3.1: A typical visual stimulus, featuring two pairs of animals.

Image components were adapted with permission and kind advice from (?, ?). Modifications were performed with Inkscape.

Figure 3.2: Illustrations of all five animals performing their social activities. From left to right: catching, combing, pushing, painting and washing

Trial feedback Immediately after each response, an icon appeared below one of the two displayed animal pairs. The presented side was determined by the subject’s response. In the case of the skip button, the icon appeared at the same height as the others, but in the middle of the screen. A green checkmark, a diagonal red cross and a yellow skip symbol signified a

correct response, an incorrect response and an invalid trial, respectively. The trial feedback screen was presented for a random interval between 400ms and 800ms.

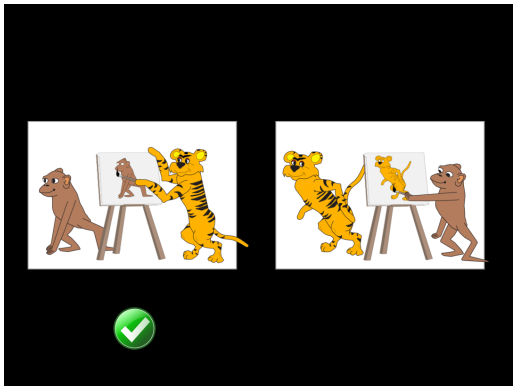

Figure 3.3: The visual feedback to a correct response.

Cluster feedback In this experiment, there was an obvious tradeoff between speed and accuracy. To encourage a high level of attention and a high number of usable trials, two bar graphs visualized performance speed and accuracy (see Fig. 3.4). In order to maximize the amount of usable trials for further analysis, the visualization valued accuracy much more than response time (see Fig. 3.5).

3.1.4 Auditory stimuli

Sentence content Each pair of images was presented with a spoken question. The question format fit well with the stimulus-response paradigm, and allowed the sentences to be identical until the conditional article (“den“ or “der“) appeared. Syntactically, the sentences used an equal number of subject-relative and object-relative clauses. In order to minimize confounding effects, these two conditions were designed to show as little auditory distinction as possible. The structure of the final sentences is displayed in table 3.1.

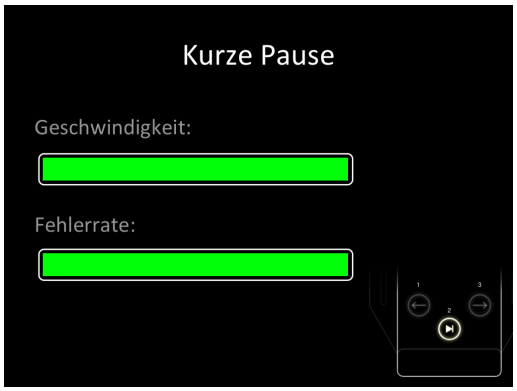

Figure 3.4: An ideal cluster feedback screen that appears after completing 19 trials. Upper bar: speed, lower bar: accuracy. Bottom right: indicator to press the skip button to advance

Original	Wo	ist	das	Tier,	das	der	Tiger	malt?
Translated	Where	is	the	animal$_{OBJ}$,	which	the$_{NOM}$	tiger$_{SUBJ}$	paints?
Word index	1	2	3	4	5	6	7	8

Table 3.1: Example stimulus sentence. Top: original spelling in German. Middle: Literal translation in English. Bottom: Word index within the sentence. NOM: Nominative case.

Tutorial sentences During the pilot study, children often assumed that the every sentence was a subject-relative construction, miscategorizing "den" for "der". Tutorial sentences made the two animal nouns explicit, so that all sentences were structured in the format "Where is the monkey that is caught by the dog?". While this setup was creating strong auditory differences, it was easier to comprehend. If the children didn't notice the difference by the eighth tutorial trial, the research assistant repeated the question with an exaggerated "den" pronounciation.

Audio format Sentences were spoken by a professional female native speaker in an unisonous and moderately child-directed prosody. Recording and playback was performed at a sampling rate of 44100Hz with one channel. Loudness of each sentence was normalized. Overall loudness was adjusted to 50db above each subject's individual hearing threshold.

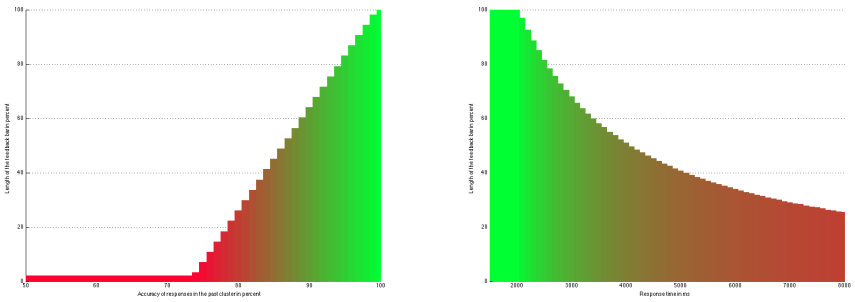

Figure 3.5: Relation between performance and displayed feedback bars. Performance (left: RA, right: RT) is drawn along the X-axis, and length of the bars in % is drawn along the Y-Axis.

Trial feedback Immediately after each response, one of two short sounds played. The sounds were extracted from Microsoft Windows XP. The “external device plugged in“ icon (two bell sounds in ascending tone) and the “external device removed“ icon (two bell sounds in descending tone) represented correct and incorrect responses, respectively. No sound was played after the skip button.

3.1.5 Experimental setup

The participants were seated on a comfortable chair, inside a shielded, dimly-lit cabin. Visual and auditory stimuli were produced by a computer running Presentation (version 14.0) at 60Hz refresh rate and 1024x768 resolution. Video signal was routed through a video splitter MSV1235 into a Panasonic PT-D7700 projector. Audio signals were generated by a Soundblaster Audigy 2 ZS [SB0350]. An audio amplifier (Compumedics, Hamburg, Germany) drove a pair of TIP-300 loudspeakers (Nicolet, Biomedical Madison, WI, U.S.A.). Sound was routed through a pair of plastic tubes (50cm length, approx. delay of 1.6ms) Sound arrived in the subjects’ ears via ER3-14A/B earplugs (Etymotic Research Inc., Elk Grove Village IL, U.S.A.).

3.2 Data acquisition

3.2.1 MEG

MEG data were collected with an Elekta Neuromag VectorView® MEG scanner in Bennewitz, at the development unit for Magnetoencephalography and Cortical Networks, Institute for Cognitive and Brain sciences, Leipzig, Germany. The scanner comprised 306 MEG-channel sensors (102 magnetometers, 204 planar gradiometers). Sensors were tuned prior to each MEG recording session to limit noise levels to approximately $2.5\frac{fT\sqrt{Hz}}{cm}$. Sensors that became very noisy during a recording block would be individually re-tuned at the next inter-block break, either by using the fine-tuning options or the selective heating function. Continuous MEG data were recorded at 1000 Hz sampling rate (330 Hz lowpass filter).

Prior to data acquisition, all metal and other potential sources of electromagnetic interference were removed from participants. Quality of recording was confirmed by visual inspection of a live view of MEG recording before each session without the subject present. Electro-oculogram (EOG) and electrocardiogram (ECG) time-series were recorded simultaneously with MEG to track potential noise sources and artifacts. Five head position indicator (HPI) coils were attached to the participant's forehead and a Polhemus stylus and digitizer device were used to record the locations of fiducial points (right and left pre-auricular points (RPA, LPA) and nasion), the HPI coils, and between 150 and 200 extra digitizer points on the head surface. Prior to the recording of each stimulus block, head location in the scanner was measured with an automatic process that detected the coils. Continuous HPI recorded any head movements during data acquisition.

3.2.2 MRI

Anatomical magnetic resonance imaging (aMRI) data were collected with a 3.0 Tesla TIM Trio scanner, located at the Max-Planck-Institute for Cognitive and Brain sciences. Two scans were acquired from each participant in one session: A T1-weighted scan and a

T2-weighted scan. The T1-weighted scan used the magnetization-prepared rapid gradient echo (MPRAGE, (?, ?)]) sequence (flip angle = 9°, TR/TE/TI = $2300ms/2.96ms/900ms$). This scan was oriented transverse (176 slices) with an isotropic resolution of 1mm. The T2-weighted scan used the SPACE sequence by (?, ?) (flip angle = 120°, TR/TE = $3200ms/402ms$). This scan was oriented transverse (176 slices at 1mm) with an inplane resolution of 0.5mm x 0.5mm. All scans used a 32-channel head coil for the acquisition.

3.3 Data analysis

Data were preprocessed with the three software packages: Elekta Neuromag® MaxFilter (version 2.2, (?, ?)), Matlab (version 2014a) and MNE-Python (version 0.8.6, (?, ?)).

3.3.1 Behavioral data

Two types of behavioral data were analyzed for group and condition effects: response time (RT) and response accuracy (RA). Response time was measured at the condition onset, i.e. at the “d“ sound of “den“ or “der“ (in the subject-relative clause or the object-relative clause, respectively). Trials were omitted when the subject skipped or answered them incorrectly. Trials were also omitted if the response took longer than 4000ms. This procedure removed 11.1% of the childrens’ trials, and 2.5% of the adults’ trials.

RT and RA were determined for each subject separately from the remaining trials. Both metrics were tested for the requirements for an analysis of variance (ANOVA). Normality of the residuals was tested with a Shapiro-Wilk test (?, ?), implemented in Matlab. Equality of variances was tested with a Levene test(?, ?), implemented in SPSS. RA data failed the normality test. To include RA data in the following analysis, they were transformed to fit a normal distribution. This transformation was accomplished with the inverted sigmoid function:

$$\hat{a} = -log(\frac{1}{a} - 1)$$

All results from the ANOVA were transformed back into milisecond space with the sigmoid function:

$$r = \frac{1}{1+e^{-\hat{r}}}$$

3.3.2 Sensor-space activity

Preprocessing and HPI correction Signal-space separation (?, ?) was used to reduce noise in the data by suppressing magnetic interference coming from outside and inside the sensory array. MEG recordings were corrected for HPI movements, and co-registered across blocks to the inital head position for each individual. All of these steps were computed with MaxFilter. Data were then subjected to a 0.4Hz FIR highpass filter (Hamming window design, 4367 coefficients, -130db suppression at 0Hz, -3db at 0.4Hz,processing in Matlab) to remove slow trends.

Artifact removal MEG channels with abnormally high noise levels as identified by visual inspection were rejected from further analysis. A median of 1 channel (maximum: 3 channels) was removed. The resulting pre-processed data contained major artifacts from spontaneous channel jumps, electrocardiographic (ECG) activity and electrooculographic (EOG) activity. Jump amplitudes were detected by selecting peaks in the z-transformed continuous data that exceeded a threshold of 12 standard deviations. Segments of 2 seconds in the pre-processed continuous data were rejected if any magnitude channel exceeded an amplitude of $6 \cdot 10^{-12} T$ (gradiometer channels: $4 \cdot 10^{-12} \frac{T}{cm}$). Continuous data were then decomposed into independent components (ICA) that explained 99% of the variance. Components that correlated with EOG or ECG channels were removed with the MNE functions *preprocessing.ica_find_ecg_events()* and *preprocessing.ica_find_eog_events()*, respectively. ICA-based correction removed an average of 2.1 components per subject and block (minimum: 1, maximum: 4). The remaining ICA components were used to reconstruct continuous data.

Epoching The main trigger was set at the condition onset (described in section 3.1.4). Epochs were created between 1000ms before and 4000ms after the main trigger. An epoch

was rejected if the trial was skipped, or answered too slow (more than 4000ms) or answered incorrectly. This procedure yielded an average of [] trials in children and [] trials in adults. Data were filtered before epoching with a 45Hz FIR lowpass (using the MNE function *raw.filter()*) exclusively for the following two steps.

Cluster analysis The condition effect was used to determine suitable time windows. Selecting data purely based on contrast will include spurious differences as well as the condition-based differences in activity. Since there the subsequent statistical analysis compares the same contrast, it is prone to overestimate the condition effect. This problem was resolved with a cluster-level permutation comparison. Spurious differences in activity should vary randomly between trials and subjects, while the condition contrast is expected with a roughly equal delay and duration.

One pair of trials was pooled for each syntax condition over all trials in a group. Each trial consisted of mean activity from one of each of three sensor groups (from parietal, temporal and frontal locations). Clusters were computed by running the MNE function *stats.permutation_cluster_test()* (?, ?) over the pair of trials. The function was run with 2500 permutations, and an t-threshold of 2.0. Since the group had a strong impact on RT (see [4.1.1]), effective time windows were estimated separately for children and adults.

Interval analysis Additionally, a blind comparison was performed for sensor activity in a series of time intervals. 10 time intervals were established from 0ms to 2200ms[1] after onset, spanning 200ms each. The mean sensor activity was computed for each sensor group (3), hemisphere (2), time interval (10), yielding 60 activity values for each subject and condition. The corresponding values were pooled over all subjects within each of the two groups, and compared between syntax conditions with a paired Student's T-test. Results from each hemisphere and sensor group were adjusted with the false discovery rate correction (10 comparisons).

[1] As determined in the following analysis of response times, these intervals cover 95% of childrens' trials completely, and 98% of the adults'.

For visualization purposes, grand average activity was also calculated for each sensor group and condition, separately for children and adults.

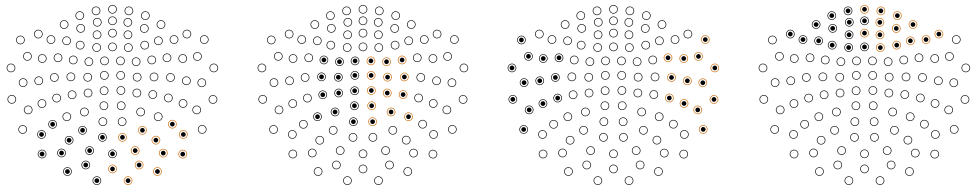

Figure 3.6: Selected channels for each sensor location. From left to right: occipital, parietal, temporal, frontal. The right hemisphere (in red) is also depicted on the right side in each illustration.

3.3.3 Source space activity

Anatomical preprocessing Cortical reconstruction and volumetric segmentation was performed with the Freesurfer image analysis suite, which is documented and freely available for download online (*http://surfer.nmr.mgh.harvard.edu/*). The technical details of these procedures are described in prior publications (Dale et al., 1999; Dale and Sereno, 1993; Fischl and Dale, 2000; Fischl et al., 2001; Fischl et al., 2002; Fischl et al., 2004a; Fischl et al., 1999a; Fischl et al., 1999b; Fischl et al., 2004b; Han et al., 2006; Jovicich et al., 2006; Segonne et al., 2004, Reuter et al. 2010, Reuter et al. 2012). I followed the recommended processing pipeline ("recon-all"), with three optional functions.

First, the option "-nuintensitycor-3T" improved brain segmentation accuracy by optimizing the bias field correction (?, ?).

Second, by invoking "-notal-check", I skipped the Talairach registration checks. Talairach registration was prone to failure especially in the infant subjects, and uneccessary for our further processing steps.

Third, I supplied and included T2-weighted MRI datasets with the options "-T2" and "-T2pial". The combination of T1- and T2-weighted images improves tissue differentiation

especially around the pia mater, yielding a more accurate cortex segmentation. This pipeline yielded a continuous, antomically plausible cortical surface in MRI space.

Forward and inverse operator For the forward operator, three components were necessary: a source model, a BEM model and a coregistration file.

The cortical surface from Freesurfer was used to construct the source model. Sources were generated by the MNE package *mne_setup_bem*. The result were 20484 sources (10242 per hemisphere), distributed with approximately equal density over the cortical surface.

The head surface from Freesurfer was used to extract a scalp surface layer. The BEM was constructed from this scalp layer with the MNE package *mne_surf2bem*, using the default options. This function sampled down the original surface to the 4th subdivision of an icosahedron. The finished BEM consisted of 5140 nodes.

Finally, a coregistration file provided the transformation between MRI space and MEG space. This coregistration attempted to minimize the distance between digitized head surface points and the head surface extracted from the MRI. It was performed for each subject individually using the MNE package *mne_analyze*. The initial fit was done manually, with visual error feedback. The following fine adjustment was performed automatically. This process was repeated until the average spatial error was less than 2mm. These three components were assembled into a forward operator by the method *mne_do_forward_solution()*.

For the inverse operator, three components were necessary: the forward model, a noise covariance matrix, and a regularization factor. Each component was calculated individually for each subject.

The first component, the forward model, was supplied by the previous step.

For the second component, the noise covariance matrix, the 1000ms after visual onset were extracted from each trial. Then, the covariance matrix was computed from this data with the function *mne.compute_covariance()*.

The third component, the regularization factor was determined from this noise covariance matrix. First, only coefficients from gradiometer channels were selected. Second, these coefficients were transformed with a singular value decomposition. Third, the upper cutoff was defined as the first value of the transformed coefficients. Fourth, the index at which the transformed coefficients performed the steepest drop in logarithmic value was determined. Fifth, this index was defined as the maximum amount of usable dimensions. Sixth, the lower cutoff was defined as the value at this index, plus 15%. Seventh, the regularization factor was computed by dividing the lower cutoff by the higher cutoff.

The inverse operator was computed from these three components by the method *mne_do_inverse_operator()*. The regularization factor was supplied with the option "–megreg".

Inverse solution For determining regional cortical activity, 8 regions needed to be defined: the primary auditory cortex (PAC), the anterior and posterior parts of the superior temporal sulcus and gyrus (aSTS, pSTS, aSTG and pSTG), Brodmann area 45 (BA45), Brodmann area 44 (BA44) and the ventral Brodmann area 6 (BA6v). These regions were spatially defined manually on the cortex of the reference subject. Freesurfer provided the aparc.a2009s segmentation, which became the basis for this regional selection. The final regions of interest on the reference brain are visualized in Fig. 3.7. Regions were then mapped from the reference cortex onto the cortices of all other subjects during the next step.

The inverse operator was then used to calculate inverse solutions from MEG sensor data. Inverse solutions were calculated for each time point, region, trial and subject individually. The process was performed by the function *mne.minimum_norm.apply_inverse_epochs()*, with sLORETA as the inverse method. The option "pick_ori=normal" designated currents leaving and entering the cortex as positive and negative, respectively. Due to the combination of passive and active noise reduction and artifact suppression, I assumed a fairly high signal-to-noise-ratio (SNR) of 50:1 for each individual source. The regularization factor

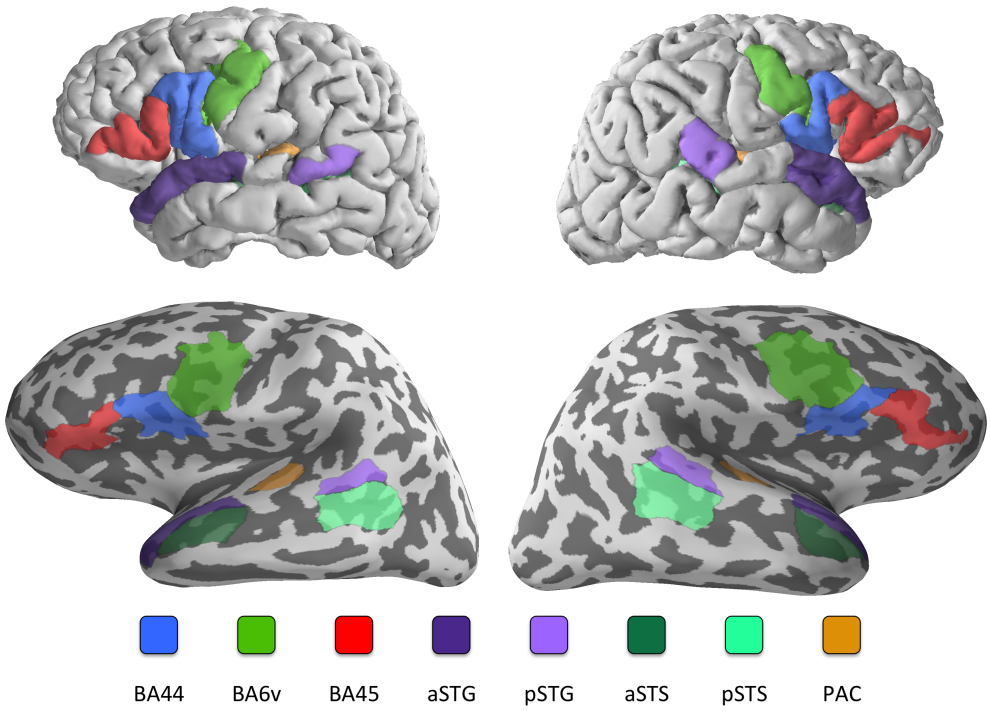

Figure 3.7: Selected regions of interest on the reference brain. Top: selected regions on the folded cortex. Bottom: selected regions on the inflated cortex.

was estimated by $\frac{1}{SNR} = 2.5 * 10^{-3}$. The result was a series of activation patterns within each region. Finally, the mean of regional node activity was calculated for each time point, region, trial and subject.

The resulting localized activity was subjected to a cluster analysis. Extracted trials were pooled over all subjects within a group. Within each group, two sets of trials were created from the two syntax conditions. Each trial contained mean activity from eight regions (PAC, aSTS, aSTG, pSTS, pSTG, BA44, BA45 and BA6v). Clusters were determined by running the MNE function $stats.permutation\_cluster\_test()$ (?, ?) over the two sets of trials. The function was run with 2500 permutations, and an t-threshold of 2.0. The results from each group was evaluated separately.

Additionally, a blind comparison was performed for sensor activity in a series of time intervals. 10 time intervals were established from 0ms to 2200ms after onset, spanning 200ms each. The mean activity was computed for each region (8), hemisphere (2), and time interval (10), yielding 160 activity values for each subject and condition. The corresponding values were pooled over all subjects within each of the two groups, and compared between syntax conditions with a paired Student's T-test. Results from each hemisphere and region were adjusted with the false discovery rate correction (10 comparisons).

For visualization purposes, grand average activity was calculated for each cortical region, group and condition.

3.3.4 Interaction analysis

The TRENtool software (version 3.3.1, February 2015) was used for exploring entropy transfers between cortical areas. For this purpuse, data was prepared for each subject individually. The procedure took place in three sections.

During the first section, the input datasets were prepared. The preparation required three components: regional activity from single trials, a list of regional comparisons and a set of parameters.

For the first component, two sets of single trials were selected from each subject. Mean regional activity were then extracted from each trial. Cortical regions were selected if they showed syntax effects either in the cluster analysis or the interval analysis. This process yielded six relevant cortical regions: PAC, aSTG, pSTS, BA44, BA45 and BA6v.

For the second component, I selected only the most relevant functional connections between cortical regions. There are 720 unique possiblities to compare these regions, which would overwhelm[2] my computational capacities, I limited the comparisons to regions connected along an axonal fiber bundle.

[2]The comparison over one pair of cortical regions (in forward and reverse directions, over two conditions) required a computation time of 4 hours per subject and timewindow. Calculating transfer entropy between 36 pairs of regions took a total of 70 hours. The task of computing 720 comparisons would take two months - a sufficient reason to switch a GPU-based algorithm(?, ?).

Comparisons between these regions were performed for all regions along three pathways (see Fig. 3.8). The regions associated to each pathway are derived from (?, ?). First, PAC and pSTG showed effects across both hemispheres. Therefore, I included comparisons across hemispheres between the mirrored regions (two connections: $PAC_{lh} \rightarrow PAC_rh$ and $pSTG_{lh} \rightarrow pSTG_rh$). Whenever PAC or pSTG were involved in a pathway, I included the regions from both hemispheres. The primary dorsal pathway connects PAC and BA6v, yielding two connections ($PAC_{lh} \rightarrow BA6v_{lh}$ and $PAC_{rh} \rightarrow BA6v_{lh}$). The secondary dorsal pathway connects PAC, pSTG and BA44, yielding six unique connections ($PAC_{lh} \rightarrow pSTG_{lh}$, $PAC_{lh} \rightarrow pSTG_{rh}$, $PAC_{rh} \rightarrow pSTG_{lh}$, $PAC_{rh} \rightarrow pSTG_{rh}$, $pSTG_{lh} \rightarrow BA44_{lh}$ and $pSTG_{rh} \rightarrow BA44_{lh}$). The ventral pathway connects PAC, aSTG and BA45, yielding three unique connections ($PAC_{lh} \rightarrow aSTG_{lh}$, $PAC_{rh} - aSTG_{lh}$ and $aSTG_{lh} \rightarrow BA45_{lh}$). Since BA44 and BA45 are direct neighbors, their connections ($BA44_lh \rightarrow BA45_lh$) was included too. Finally, the end points of secondary dorsal and ventral tract were added to the comparison, yielding four connections in total ($PAC_{lh} \rightarrow BA44_{lh}$, $PAC_{rh} \rightarrow BA44_{lh}$, $PAC_{lh} \rightarrow BA45_{lh}$ and $PAC_{rh} \rightarrow BA45_{lh}$). The reverse direction was added for each of these 18 comparisons, yielding 36 pair-wise comparisons in total.

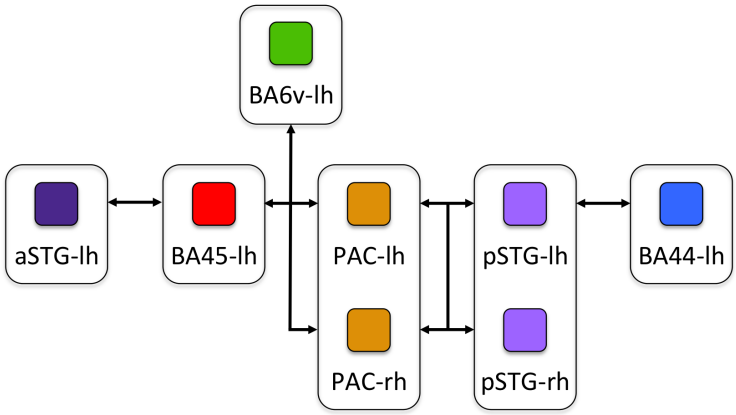

Figure 3.8: Connections of selected cortical regions for the computation of transfered entropy. Left of PAC: ventral pathway; top left of PAC: primary dorsal pathway; right of PAC: secondary dorsal pathway.

The third component concerned a set of parameters for the data preparation.

The interval analysis yielded (see chapter 4.3) three distinct time intervals with syntax effects: 0ms - 300ms, 400ms - 800ms and 1000ms - 1600ms. I supplied those intervals for the estimation of transfer entropy with the parameter *cfg.toi* in three subsequent analyses.

The transfer entropy estimation fundamentally relies on the comparison between two probability density functions. These mathematical constructs can only be approximated by using scalar time series. For this purpose, time series from individual trials were combined into a multidimensional state space by TrenTOOL. This multidimensional space was then reduced into probability density function in a process called embedding. For optimizing the embedding procedure, I selected the Ragwitz criterion (*cfg.optimizemethod = 'ragwitz'*). The parameters were supplied as ranges: the relative embedding delay *cfg.ragtaurange = 0.15:0.3* and the embedding dimension *cfg.ragdim = 2:8*.

For the embedding process, a considerable section of data is necessary to estimate the baseline entropy before the interaction time cue occurs. This section of data is referred to as embedding delay, and won't be included in the analysis of transfer entropy. The required embedding delay for this combination of data and embedding parameters was 753ms. I extended the lower time limit of each analysis time window by the same amount. The Ragwitz criterion then optimized the embedding parameters for each set of time series (once per subject, time window and condition). The most common optimal embedding dimension was 8, and the mean optimal relative embedding delay was 0.2.

Finally, the parameter *cfg.repPred* indicated the amount of samples that would be used for the interaction analysis. I maximized this value by calculating *cfg.repPred = sampleLength - embeddingDelay - 1*, with *sampleLength* as the amount of samples in the selected time window and *embeddingDelay*[3] the amount of samples in the absolute embedding de-

[3]Unfortunately, TrenTOOL doesn't display the embedding delay. The formula I used to calculate it is, in Matlab notation, *(max(cfg.ragdim)-1) · max(cfg.ragtaurange) + cfg.predicttimemax_u + max(ACT)*. *ACT* (the delay for the first minimum of the autocorrelation), in turn, is only computed during the preprocessing phase. Due to time constraints, I opted for a manual approach. An error message printed the current *max(ACT)* if *cfg.repPred* was too small, so I could supply this value into the configuration variable and re-run the preprocessing phase. After a few iterations, the absolute maximum ACT was found, allowing me to adjust the time windows and finish the preprocessing phase.

lay. This is the largest possible value that doesn't violate the assumptions behind the transfer analysis algorithm.

During the second section, TrenTOOL performed a interaction shift test. This procedure (performed with the function *InteractionDelayReconstruction_calculate()*) counteracts bias introduced by noise, for example the common false positive detection of an interaction from a less noisy to a more noisy data set. The shift test establishes the optimal delay of transfered entropy for each comparison of activity. Transfer entropy was computed for a time range of possible interaction delays: 1ms to 40ms in steps of 5ms. TrenTOOL then selected the interaction delay with the biggest associated transfer entropy as the most likely representation of the signal delay caused by cognitive processes and anatomic constraints. To alleviate issues with volume conduction, I supplied the option *cfg.extracond = 'Faes_method'*. Unfortunately, using this method prevents the determination of the precise signal delay. Since the signal delay is not relevant to my research goals, this trade-off was trivial to solve. The optimal dimension for each data set was considered with the option *cfg.optdimusage = 'indivdim'*. For significance testing, TrenTOOL creates surrogate data. This data was created by shuffling existing trials (*cfg.surrogatetype = 'trialshuffling'*). The significance level for this procedure was *cfg.alpha = 0.05*. I used t-values to represent the statistical results of the shift test (*cfg.permstatstype = 'indepsamplesT'*). These results indicate the likelihood if an entropy transfer has occurred between the selected pair of regional activity.

The third section evaluated the impact of the syntax condition on the whole group. For this purpose, a comparison was conducted between conditions over all subjects within a group.

Chapter 4

Results

4.1 Behavioral results

If not noted otherwise, two comma-separated values in brackets describe the upper and lower values of a 95% confidence interval.

4.1.1 Response times

A Shapiro-Wilk test was conducted to determine if individual response times were normal distributed. All subjects failed this test ($p < 0.001$), indicating a strong deviation from normality. Therefore, I represented individual response times by their median.

Children needed a median time of 1.91s (1.64s, 2.19s) to respond to object-relative clauses. For subject-relative clauses, they needed 1.97s (1.69s, 2.25s). Adults needed a median time of 1.51s (1.28s, 1.73s) to respond to object-relative clauses. For subject-relative clauses, they needed 1.60s (1.37s, 1.82s).

A Shapiro-Wilk test determined that response time data was normal distributed with a probability between 1.5% and 27%. A Levene's test determined that the probability of median response times being normal distributed was ?%. Supported by these findings, the response times were included into the ANOVA.

4.1.2 Response accuracy

For the analysis of variance (ANOVA), all data must be normal distributed with equal variance. A Shapiro-Wilk test determined that the probability of accuracy data being normal distributed was between 2.0 and 20.3%. A Levene's test yielded that the probability that accuracy data were distributed with equal variance was less than p = 0.1%.

The ANOVA is known to be robust for considerable deviations from the normal distribution. However, it is highly vulnerable to violation the assumption of equal variances. To meet this requirement, I transformed the accuracy data with the inverse sigmoid function. This procedure, however, created singularities in some extreme cases, i.e., when a subject performed with a 100% accuracy rate. To prevent this issue, I added a single incorrect trial to every subject's performance for the following analysis.

After the transformation, the same tests as before were conducted. The probability for transformed accuracy data being normal distributed was between 0.5% and 27%. The probability for transformed accuracy data being distributed with equal variance was p = 72.2%. Supported by these findings, the transformed accuracy data was included in the ANOVA.

4.1.3 Analysis of combined performance data

Two ANOVA were conducted with the transformed accuracy data and the median response times. Each subject provided one data point for each metric. Data were analyzed with a group x condition design. Accuracy estimates were transformed back with the sigmoid function $r = \frac{1}{1+e^{-\hat{r}}}$.

Children responded 0.39s slower than adults (1.94s vs. 1.55s). This difference was significant ($F_{56} = 9.4$, $p = 0.3\%$).

Children responded with an average accuracy of 93.8% (92.2%, 95.0%). Adults performed much better, with an average accuracy of 97.9% (97.5%, 98.3%). This difference was highly significant ($F_{56} = 52$, $p = 1.6 * 10^{-9}$).

Sentence condition had no impact on median response times ($F = 0.33$, $p = 57\%$) or on response accuracy ($F = 1.3$, $p = 26\%$). There was no interaction effect between group and sentence condition ($F < 0.1$, $p > 80\%$).

4.2 Sensor-space activity

We computed average event-related fields (ERF) for each subject and sensor region. Activity from these ERF was selected with two different types of time windows.

A positive effect indicates that activity evoked by object-relative clauses was more positive than activity evoked by subject-relative clauses. In the case of gradiometers, “more positive“ means a higher regional RMS. With localized data, “more positive“ means a higher z-score from the sLORETA source reconstruction.

4.2.1 Interval analysis

For this analysis, sensor activity from separate regions and hemispheres was compared blindly between 0 and 2200ms after onset in 200ms intervals.

After FDR-correction for 10 comparisons, no sensor region showed any non-spurious effect in either group ($p > 3\%$).

4.2.2 Cluster analysis

For this analysis, activity was compared between conditions using a temporal cluster analysis.

For children, significant differences were observed in the following sensor groups: Right parietal magnetometers showed a negative effect between 159ms and 374ms ($p = 1.4\%$). Left frontal magnetometers showed a negative effect between 375ms and 666ms ($p = 1.8\%$). Left temporal magnetometers showed a negative effect between 349ms and 627ms ($p = 0.8\%$). Right temporal gradiometers showed a positive effect between 1384ms and 1663ms ($p = 3.8\%$).

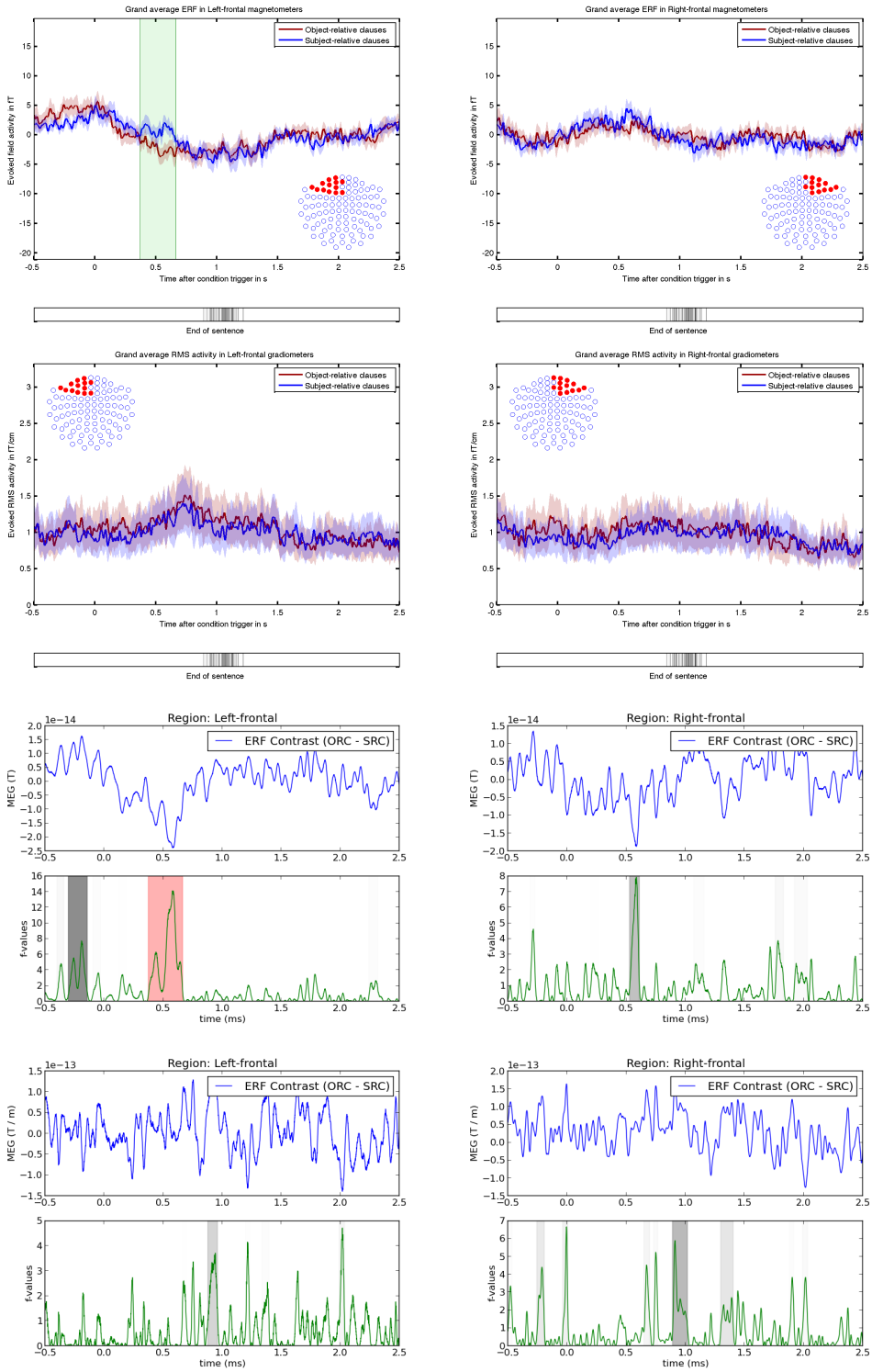

Figure 4.1: Combined frontal sensor activity from children in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

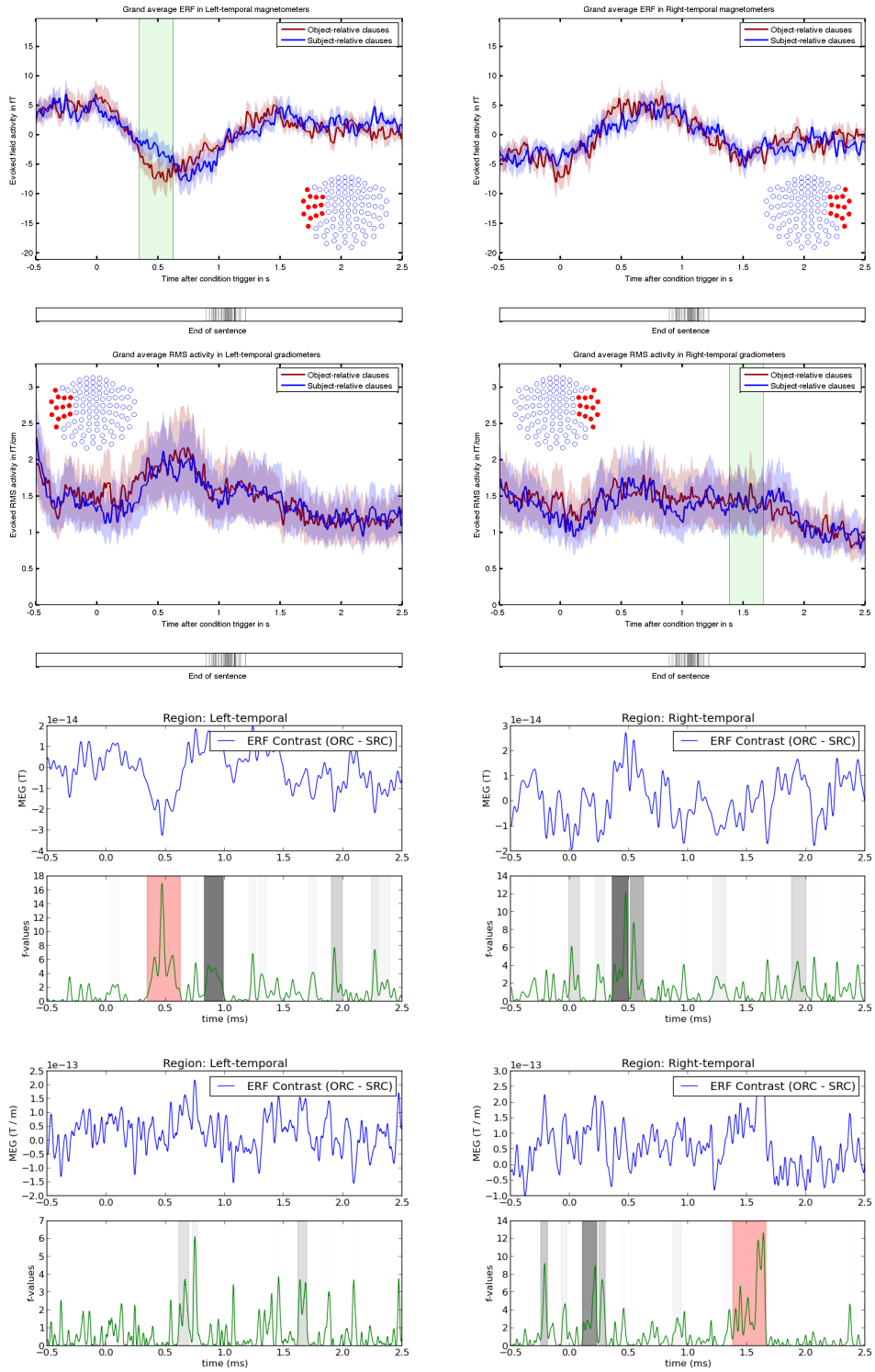

Figure 4.2: Combined temporal sensor activity from children in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

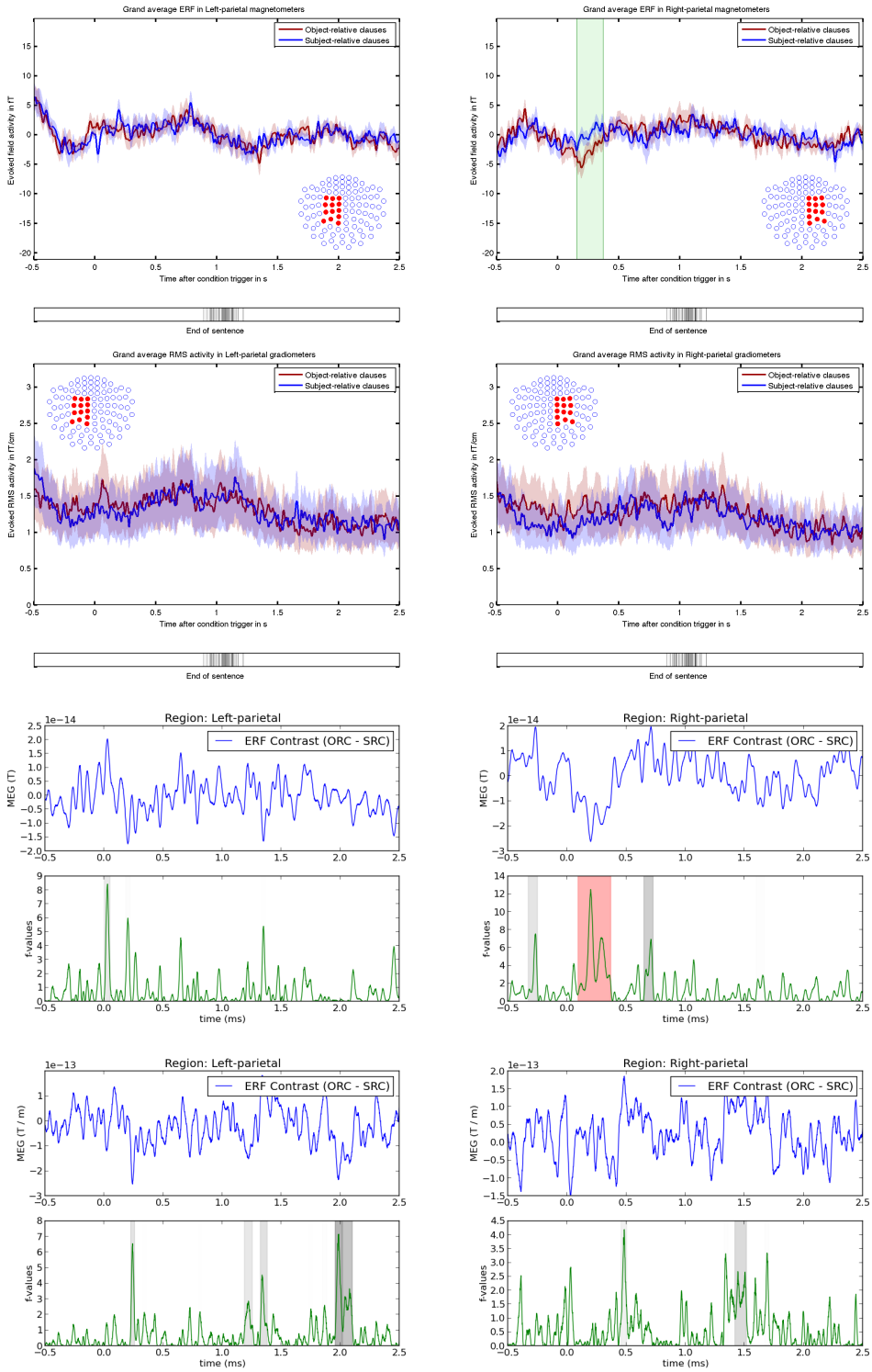

Figure 4.3: Combined parietal sensor activity from children in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

For adults, clusters of significant differences were observed by left-temporal gradiometers and left-parietal magnetometers. Left frontal gradiometers showed a weak positive effect between 131ms and 284ms ($p = 6.3\%$). Left temporal gradiometers showed a positive effect between 257ms and 480ms ($p = 2.1\%$). Left parietal magnetometers showed a positive effect between 618ms and 765ms ($p = 0.64\%$). Left temporal magnetometers showed a weak negative effect between 1351 and 1491ms ($p = 8.4\%$).

The generally lower significance levels in adults imply an overall weaker impact of syntactic condition on sensor activity. Syntactic effects in the left fronto-parietal region occurred earlier in adults (131-480ms) than in children (349-666ms). Effects were much more lateralized in children, with a weak but distinct support from right parietal and temporal regions. The strong effect in adults' left parietal regions between 618ms and 765ms is unparalleled in children. These differences seem to promise a group effect, and will be resolved more accurately in the next section.

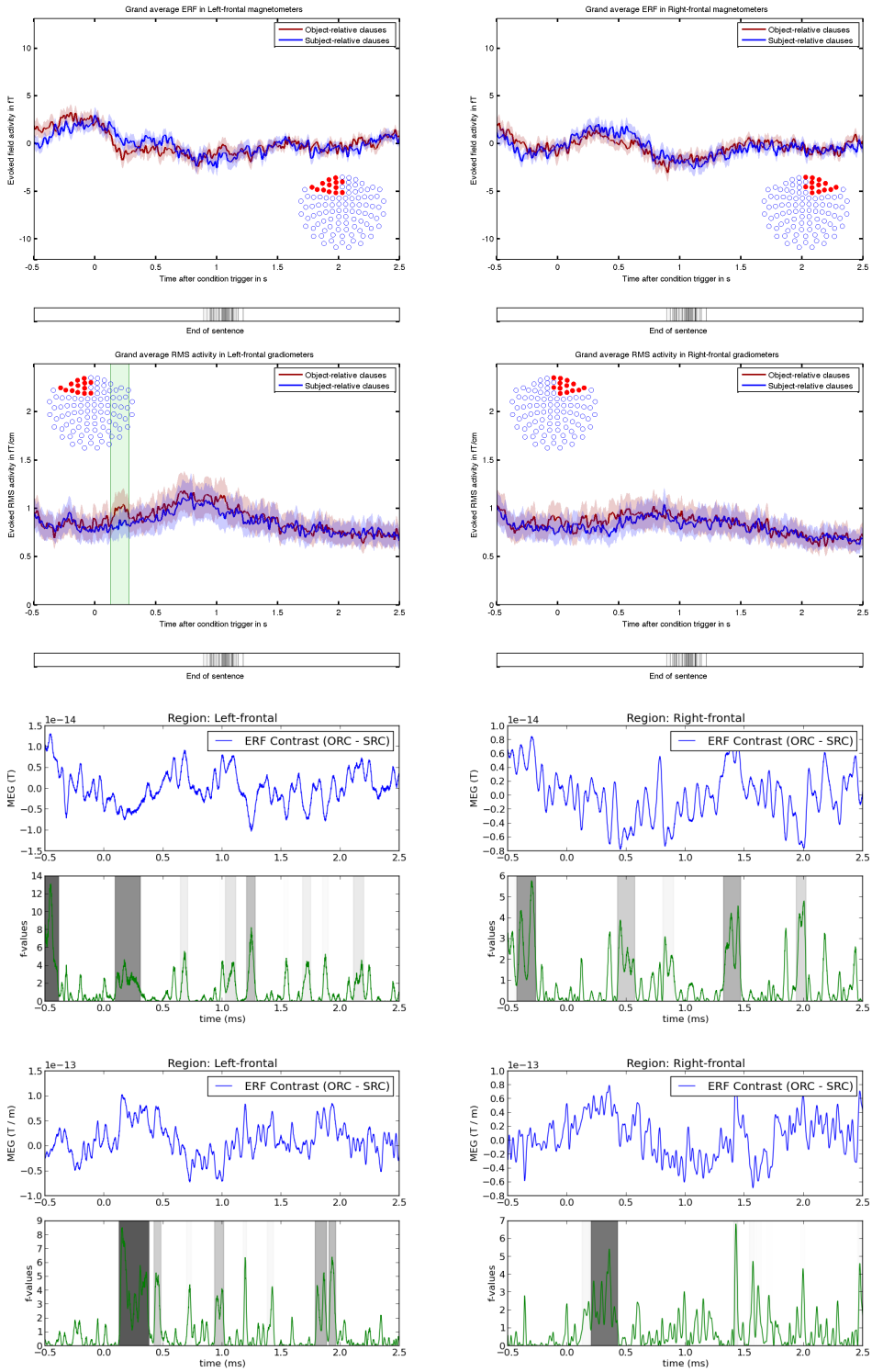

Figure 4.4: Combined frontal sensor activity from adults in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

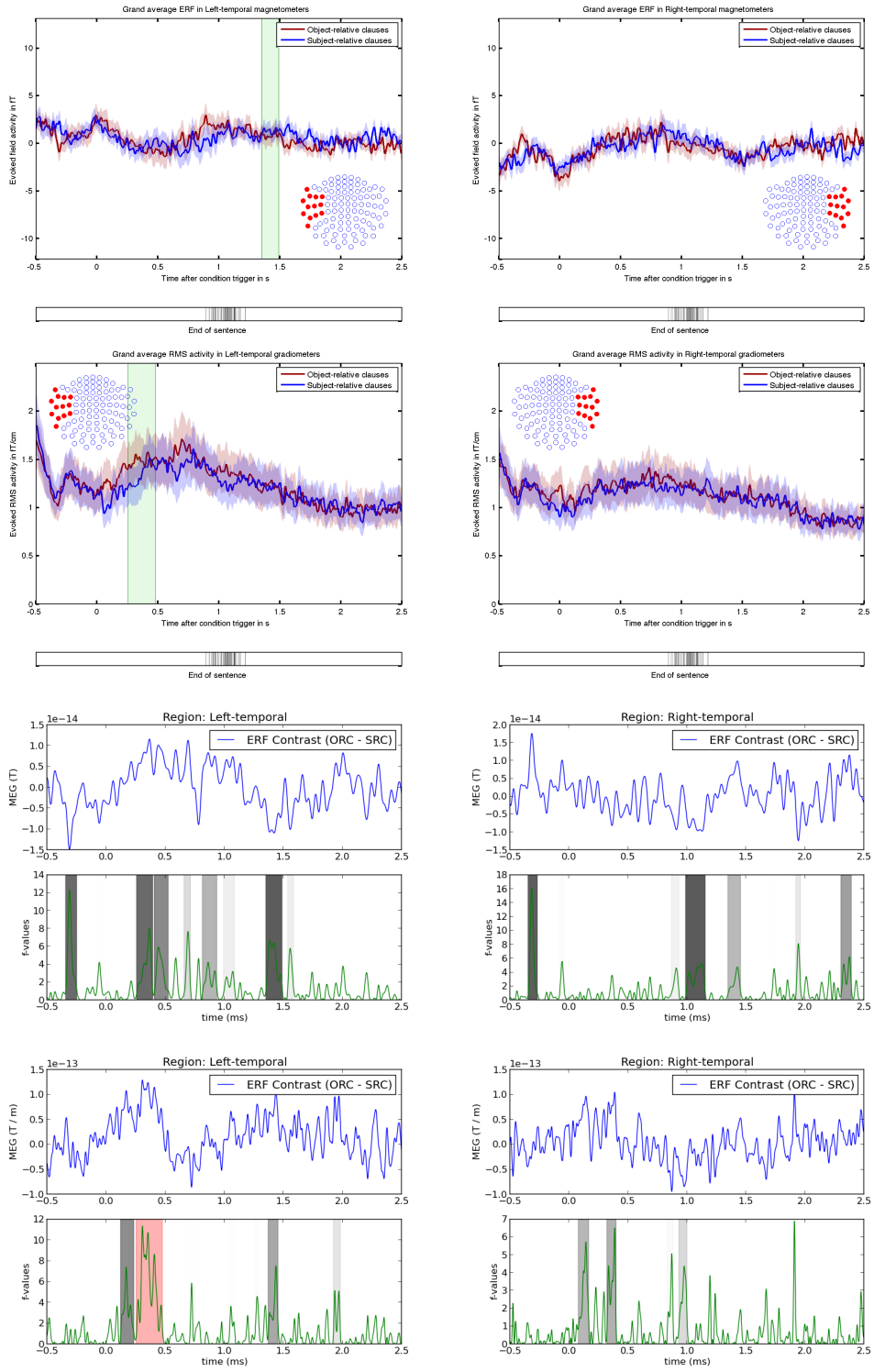

Figure 4.5: Combined temporal sensor activity from adults in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

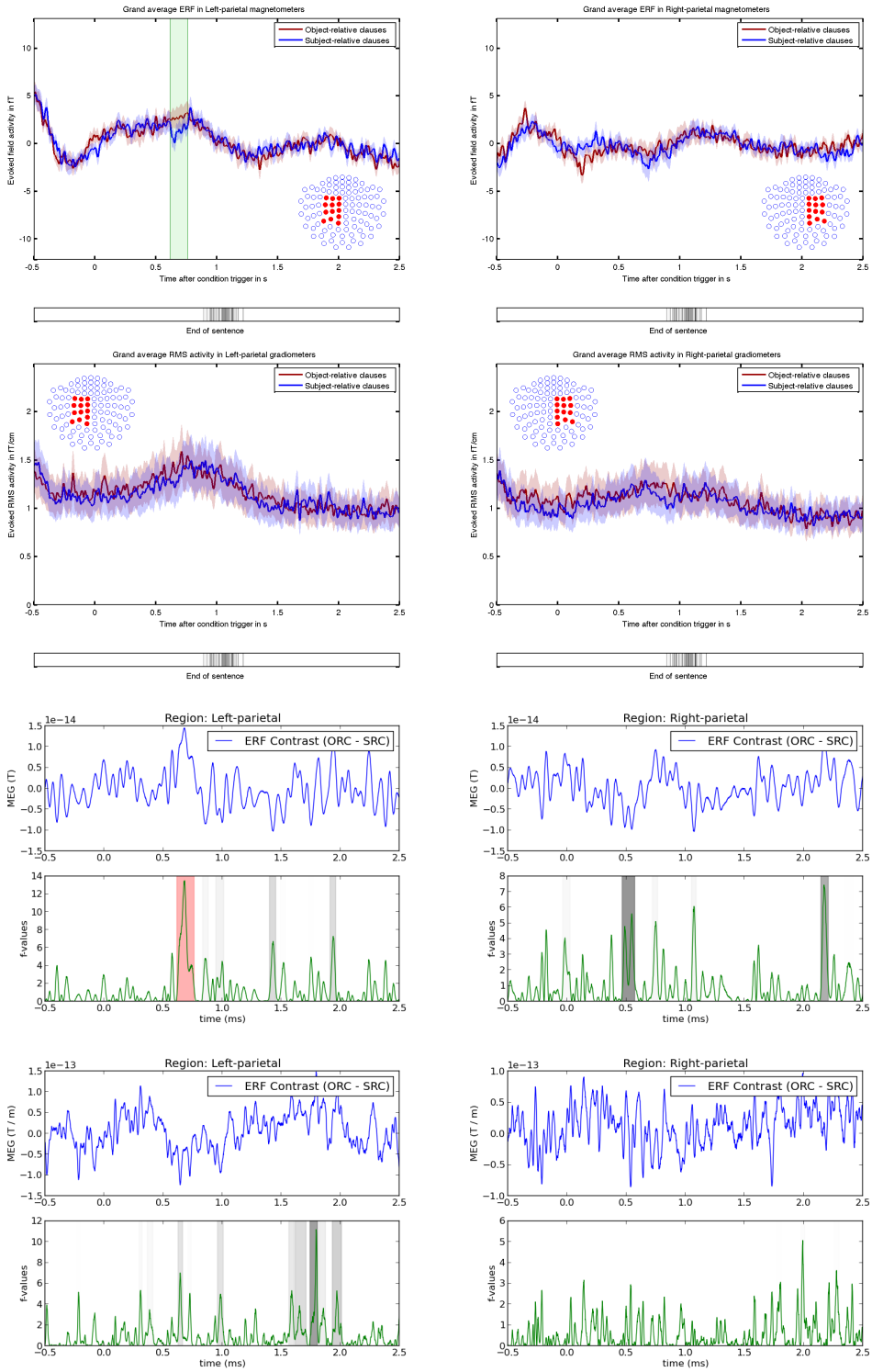

Figure 4.6: Combined parietal sensor activity from adults in separate sensor groups. Top row: magnetometer activity; middle row: gradiometer activity; bottom two rows: equivalent results from the cluster analysis. Charts on the left side depict activity from the left hemisphere and vice versa.

4.3 Source-space activity

4.3.1 Comparison of means

Localized activity from eight regions of interest (PAC, aSTS, aSTG, pSTS, pSTG, BA44, BA45 and BA6v) were examined for a syntactic effect. Two different types of analysis were conducted: a cluster-based comparison and an interval-based comparison.

Cluster analysis The cluster comparison yielded only spurious effects in both groups ($p > 7\%$).

Interval analysis The interval comparison (0ms to 2200ms) yielded three distinct effect clusters in adults.

During the first time cluster, 0ms to 200ms, especially the right PAC showed a strong negative effect ($p = 0.5\%$, $t_{17} = -3.7$). Other negative effects were visible in the left PAC ($p = 3.0\%$, $t_{17} = -2.9$) and the left BA45 ($p = 5.1\%$, $t_{17} = -2.6$).

The second time cluster, 400ms to 800ms, was signified by simultaneous effects in seven regions. Both left and right PAC showed a positive effect ($p = 3.0\%$ and $p < 0.1\%$, $t_{17} = 2.9$ and $t_{17} = 4.8$). The right pSTG showed negative effects, which grew more noticable during the progression of the time cluster. In the first half of the cluster (400ms to 600ms), the condition effect started very weakly ($p = 7.9\%$, $t_{17} = -2.4$). During the second half of the cluster (600ms to 800ms), the effect strengthened noticably ($p = 1.0\%$, $t_{17} = -3.9$). The left aSTG showed a similar pattern: weak negative effect between 400ms and 600ms ($p = 2.1\%$, $t_{17} = -3.2$), which became noticably more distinct between 600ms and 800ms ($p = 0.4\%$, $t_{17} = -4.3$). The left BA45 showed a constant and strong positive activation effect ($p = 0.6\%$, $t_{17} = 3.9$). Otherwise, very weak positive effects ($p < 7\%$, $t_{17} = 3.0$) could be observed in the left BA6v (400ms to 600ms) and the left BA44 (600ms to 800ms).

The third time cluster, 1000ms to 1600ms, only showed effects in three regions. The right PAC showed a very late, distinctly negative effect ($p = 0.9\%$, $t_{17} = -3.3$, between 1400ms and 1600ms). Very weak positive effects ($p < 8\%$, $t_{17} = 2.4$) were observed in the

right pSTG (1000ms to 1200ms) and the left aSTG (1400ms to 1600ms). At the same time (1400ms to 1600ms), the left aSTS showed a very weak effect as well, but with opposite polarity ($p = 6.6\%$, $t_{17} = -3.0$).

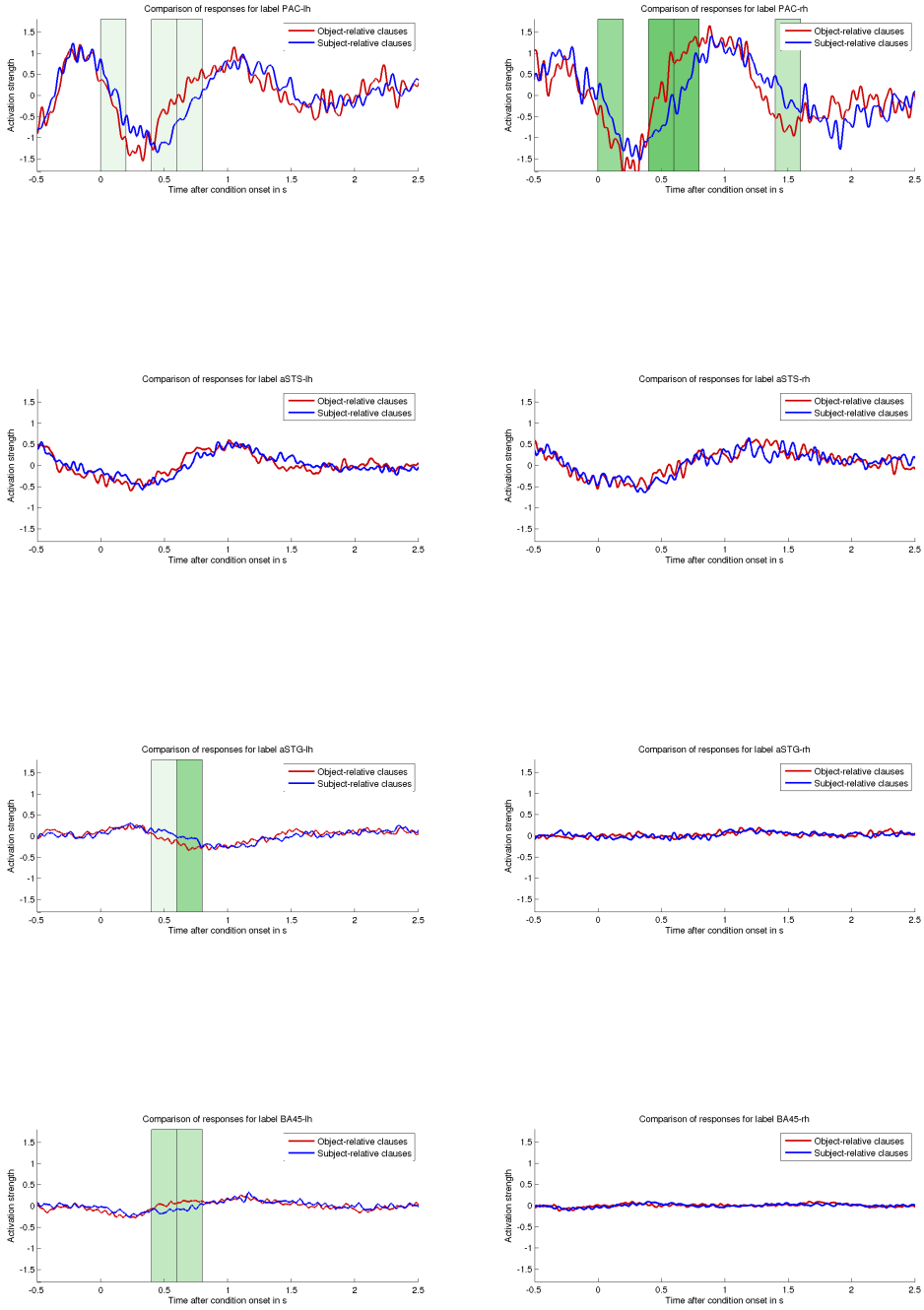

Figure 4.7: Combined activity from adults in separate cortical regions. Top row: PAC; second row: aSTS; third row: aSTG; bottom row: BA45. Charts on the left side depict activity from the left hemisphere and vice versa.

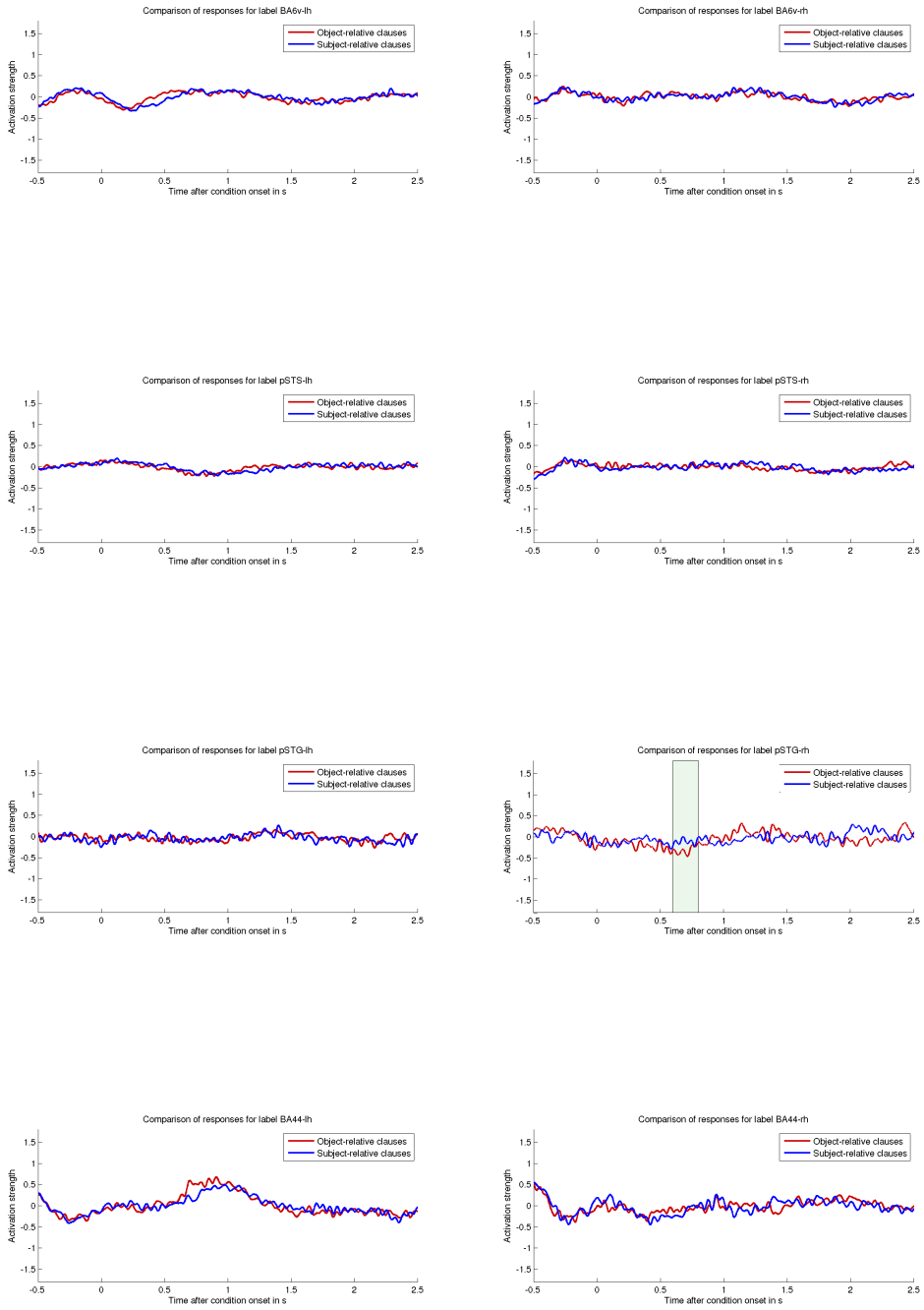

Figure 4.8: Combined activity from adults in separate cortical regions. Top row: BA6v; second row: pSTS; third row: pSTG; bottom row: BA44. Charts on the left side depict activity from the left hemisphere and vice versa.

The interval analysis yielded no non-spurious effects ($p > 5\%$) from children. This fact prompted a more extensive investigation.

4.3.2 Post-hoc analysis

Compared to the sensor-space analysis, source-space data yielded much more pronounced effects in adults. Since this effect differentiation was the main goal of the localization process, these results were in line with my expectations. However, localization of cortical activity in children failed to yield a similar improvement.

There were two major possibilities for these unexpected results.

Possible explanations First, the localization process may have been produced drastically worse inverse solutions for children than for adults. Due to the hands-off approach of the generation of cortical surfaces, there are no parameters for the transfer of regional boundaries between the (adult) reference brain and each infant brain. This automated process could have resulted in a systematically higher spatial error between my regional definitions and the actual functional regions in infants than in adults. Less realistic regional definitions in children than in adults can result in unintended overlap between functional regions, and lead to a diminished experimental effect in each region. It is also be possible that the automated segmentation process, that Freesurfer uses for extracting the cortex surface, is not optimized for infant brains. An imprecise definition of the cortical surface would result in a skewed spatial location of the source dipoles, again causing regional overlapping and diminishing the experimental effect.

Second, it is possible that the cognitive processes of my young subjects were much more diverse than those of my adults. Both cluster analysis and interval analysis are based on the assumption that all subjects in one group use the same cognitive strategy. If this assumption is true for adults, but not for children, different processing strategies could cancel each other out and produce no effect. Both the average localized activity and pooled single trials would be vulnerable to this violation.

If the localization performed drastically worse for children, they would have to be removed from the following signal transfer analysis.

Testing the explanations To decide between these two possibilities, I conducted two tests:

First, if the localization was drastically worse for children than for adults, the computed L2-norm would be significantly lower. Since sLORETA is based on the premise that the lowest L2-norm is equivalent to the best inverse solution, a worse solution would be equivalent to worse localization accuracy. To test this hypothesis, I selected the inverse solution of an average trial from each subject. The basis for this trial was the sensor activity elicted by a subject-relative clause. I computed the L2-norm over all vertices and pooled the results into the two groups, children and adults. These two groups were then compared with a Mann-Whitney-Wilcoxon rank sum test. If the set of L2-norms from childrens' results was higher, the first hypothesis would be supported.

Second, if the cognitive processes were much more varied in children than in adults, this circumstance should be reflected in a more homogenuous evoked activity within the adults. To test this hypothesis, I pooled the average reconstructed cortical activity for all subjects in one group. The basis for the time series was the regional activity elicted by a subject-relative clause. I calculated the variance for all time points over the subjects in a group. The set of variances were compared in a Mann-Whitney-Wilcoxon rank sum test, once for each region. If the adults' variances were significantly and consistently lower than the childrens', the second hypothesis would be supported.

Test results