Oscillatory activity and EEG phase synchrony of concurrent word segmentation and meaning-mapping in 9-year-old children

Neus Ramos-Escobar\textsuperscript{a,b}, Emma Segura\textsuperscript{a,b}, Guillem Olíve\textsuperscript{a,b}, Antoni Rodriguez-Fornells\textsuperscript{a,b,c,*,1}, Clément François\textsuperscript{d,*,1}

\textsuperscript{a} Dept. of Cognition, Development and Educational Science, Institute of Neuroscience, University of Barcelona, L’Hospitalet de Llobregat, Barcelona, 08097, Spain
\textsuperscript{b} Cognition and Brain Plasticity Group, Bellvitge Biomedical Research Institute (IDIBELL), L’Hospitalet de Llobregat, Barcelona, 08097, Spain
\textsuperscript{c} Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain
\textsuperscript{d} Aix Marseille Univ, CNRS, LPL, 13100, Aix-en-Provence, France

ARTICLE INFO

Keywords:
Children
Speech segmentation
Meaning-mapping
EEG
Frequency-tagging
Scalp-based phase synchrony

ABSTRACT

When learning a new language, one must segment words from continuous speech and associate them with meanings. These complex processes can be boosted by attentional mechanisms triggered by multi-sensory information. Previous electrophysiological studies suggest that brain oscillations are sensitive to different hierarchical complexity levels of the input, making them a plausible neural substrate for speech parsing. Here, we investigated the functional role of brain oscillations during concurrent speech segmentation and meaning acquisition in sixty 9-year-old children. We collected EEG data during an audio-visual statistical learning task during which children were exposed to a learning condition with consistent word-picture associations and a random condition with inconsistent word-picture associations before being tested on their ability to recall words and word-picture associations. We capitalized on the brain dynamics to align neural activity to the same rate as an external rhythmic stimulus to explore modulations of neural synchronization and phase synchronization between electrodes during multi-sensory word learning. Results showed enhanced power at both word- and syllabic-rate and increased EEG phase synchronization between frontal and occipital regions in the learning compared to the random condition. These findings suggest that multi-sensory cueing and attentional mechanisms play an essential role in children’s successful word learning.

1. Introduction

When learning a new language, one must segment words from continuous speech and associate them with possible meanings. Recent cognitive models suggest that these processes may rely on statistical learning (SL), a domain-general mechanism that allows the computation of statistical regularities within and across sensory modalities (Frost et al., 2015; Räisänen and Rasilo, 2015). In the case of speech, infants, children, and adults can implicitly segment continuous streams of pseudowords by relying on the transitional probabilities between consecutive syllables (Cumiller et al., 2009; François et al., 2013, 2017a; Safran et al., 1996). Redundant multi-sensory cues can facilitate speech segmentation (Cumiller et al., 2010a, b; Glicksohn and Cohen, 2013; Thiessen, 2010) and word-referent associations (Yu and Smith, 2007; Shukla et al., 2011). Interestingly, these redundant cues may capture learners’ attention by highlighting temporally contiguous target words and referents (Axelsson et al., 2012; Horst and Samuelson, 2008). These results align well with previous studies proposing that the combination of attentional and associative processes may facilitate multi-modal word learning (Plunkett, 1997; Smith, 2000). The attention toward a visual event may trigger the association of the visual referent with the spoken word. This attentional capture may occur through cue facilitation, where the learner uses the referent’s perceptual saliency as a cue for the binding process. At the same time, this process could rely on associative mechanisms that allow the binding of temporally contiguous multi-sensory events (Smith, 2000). In any case, enhancing visual
attention to a referent by perceptual saliency or temporal contiguity may facilitate word learning. Overall, current empirical evidence and theoretical proposals converge that redundant cues might facilitate language learning.

Apart from the mechanisms mentioned above, other cognitive and linguistic factors have been proposed to explain children’s ability to learn new words, such as implicit and perceptual learning (Reber, 1967; Gibson and Gibson, 1955), social context (Tomasetto, 1995; Bloom, 2000) and pre-existing linguistic constraints (Markman, 1990; Carey and Bartlett, 1978). Hollich et al. (2000) proposed a compiled model of word acquisition positing that children’s word-learning ability emerges from combining some of these factors and that the weight of those factors changes through development. During the initial stages of word learning, children would rely on domain-general cues (e.g., social and attentional cues). During later learning stages, they would rely on language-specific cues instead (e.g., phonotactic or grammatical rules).

In this line, some authors suggest there exists a shift in the way children learn novel words during middle school years, from direct, explicit teaching to increased learning from context (Dickinson, 1984; Chall, 1983). For instance, 8- to 9-year-old children benefit more from multiple word-encounters within a story than 6–7-year-olds (Wilkinson and Houston-Price, 2013). In the third and fourth grades, children begin to distinguish between pseudowords and nonsense letter strings in an adult-like manner (Henderson and Chard, 1980; Lefon and Spragins, 1974). Children of this age also show evidence of adult-like semantic processing (Coch, 2015). Interestingly, implicit learning abilities also change during development and follow a U-shaped trajectory across the lifespan (Juhasz et al., 2019; Janacsek et al., 2012). Indeed, 7- to 9-year-old children show superior general skills than younger children (Juhasz et al., 2019), reaching a plateau at around 12 (Janacsek et al., 2012). Finally, orienting attention abilities, which are crucial for multi-sensory tasks, seem to develop between middle and late childhood (Pozuelos et al., 2014). Overall, several converging pieces of evidence suggest a knee point in language learning mechanisms during middle childhood.

Previous studies have used Electroencephalography (EEG) to deci-

The presence of this visual attentional cue might be crucial for the multimodal presentation of words and referents may facilitate word learning. Overall, current empirical evidence and theoretical proposals converge that redundant cues might facilitate language learning.

Here, we explored concurrent speech segmentation and meaning-mapping in a large group of sixty 9-year-old children. Even though infants and adults often use environmental information to aid the extraction and acquisition of novel word forms (Smith et al., 2002; Johns et al., 2016), these two processes have traditionally been explored separately in the literature (Graf-Estes et al., 2007; Hay et al., 2011; Räsänen and Rasilõ, 2015). Besides, language acquisition can be considered a joint inference problem for multiple linguistic elements that need to be learned in parallel (Johnson and Tyler, 2010; Lin et al., 2015). To our knowledge, only two studies in adults have shown that the statistical properties found in continuous speech together with the statistical consistency between speech and visual referents might be crucial to infer the possible meaning of a new word (Cuñillera et al., 2010a, b; François et al., 2017b). However, the neurophysiological mechanisms involved when the two processes of segmenting new words and binding conceptual representations to these newly isolated words are taking place at the same time are poorly understood in children.

With this purpose in mind, we recorded EEG during an audio-visual SL task. We applied frequency-tagging and scalp-based phase synchronization analyses to track the online modulations in neural synchronization and long-distance connectivity during learning. EEG phase synchronization analyses at the sensor level used the weighted Phase-Lag Index (wPLI), a measure of phase synchronization between two signals that avoids the problem of volume conduction (Vinck et al., 2011). Specifically, children were presented with a learning stream with consistent word-picture associations and a random stream with inconsistent ones. Based on previous results in adults (François et al., 2017b), we predicted a higher connection between words and referents for the consistent word segmentation and the binding to its referent. The presence of this visual referent should induce attentional capture and facilitate the emergence of new multi-sensory memory traces. Therefore, we expected larger neural responses at both syllable and word frequencies for learning than for random streams over frontal regions as observed in previous ERP studies on speech segmentation based on SL (Cuñillera et al., 2006, 2009; De Diego-Balaguer, 2007; François et al., 2014). We also expected an enhanced word-frequency response over occipital electrodes.
reflecting the perception of the pictures at the same frequency as words. Finally, we predicted a larger cross-talk between occipital (visual) and fronto-central electrodes (speech-segmentation) with stronger scalp-based connectivity between fronto-occipital regions in the learning than in the random condition.

2. Methods

2.1. Participants

A total of seventy-two children attending the fourth grade in two elementary schools in Barcelona participated in the study. Two children did not finish the task, and ten more were excluded due to noise in the recordings due to excessive movements (4), drowsiness (3), or background noise (3). Hence, a total of sixty children (mean age: 9 years and five months; 34 girls) were included in the study. All children had similar socioeconomic status determined from the parents’ incomes (middle/upper class). They were Catalan-Spanish bilinguals and right-handed, except for four left-handed. A written informed consent was obtained from all the children’s parents before the start of the study. School directors and teachers were also informed and collaborated in the organization of the experiment. The study was conducted following the local norms and guidelines for protecting human subjects and was approved by the local ethics committee.

2.2. Stimuli

We used the same audio-visual artificial language streams as those used in François et al. (2017b). Briefly, we created a synthetic speech stream containing four trisyllabic pseudowords ("tapiru", "bagoli", "gukido", "tokuda"; from now on, called words) that were concatenated in a pseudo-random order with no immediate repetition (see Fig. 1). We used the MBROLA speech synthesizer with the Spanish male database (es1) to create the stream. All phonemes had the same duration (116 ms) and pitch (200 Hz). Each word had a duration of 696 ms and was repeated 25 times in the stream. The language stream was presented four times with a short pause between each presentation for a total of 400 words and a duration of 4 min and 39 s. The transitional probability between syllables was 1.0 within words and .33 between words.

In addition to the statistically structured stream, a random stream was created by pseudo-randomly mixing the syllables. The random stream had the same duration as the structured one, but the transitional probabilities between syllables were .09. Therefore, EEG signals from this condition were used as a baseline.

Four different pictures were synchronously presented with the four words of the structured speech stream. The visual stimuli consisted of four 20 × 120 mm black-and-white drawings belonging to four different semantic categories (animals, vegetables, vehicles, and accessories; Snodgrass and Vanderwart, 1980) while controlling for name agreement (>80 %), familiarity (M = 3.4), and complexity (M = 3.1). Significantly, each word was associated with a single picture resulting in four fully consistent word-picture associations. In the random condition, four
different pictures [name agreement (>80 %), familiarity (M = 3.4), and complexity (M = 2.9)] were pseudo-randomly presented on the screen every 696 ms. However, because the random stream did not contain words, no association could be created.

2.3. Experimental procedure

Children were tested individually in a quiet room of their school in 2 separate sessions on two different days that included neuropsychological assessments and electrophysiological tests, respectively. The electrophysiological session lasted one hour and a half, and the neuropsychological session lasted 30 min.

In the neuropsychological session, children were evaluated with sub-tests of the NEPSY-II battery (Korkman et al., 2007) and the WISC-IV (Wechsler and Corral, 2015). Specifically, language functions were assessed using the verbal fluency and nonword repetition tasks from the NEPSY-II battery. Non-verbal problem solving was evaluated with the matrix-reasoning subtest and working memory with the digit span subtest from the WISC-IV. Overall, they performed within the normal range (Digit span: M = 15.56, SD = 2.89, mean normal score: 13–14; Matrix reasoning: M = 17.9, SD = 3.71, mean normal score 18–19; Non-word repetition: M = 41.18, SD = 2.65, expected range: 34–42; Verbal fluency: M = 33.28; SD = 6.33, expected range: 25–35; Letter fluency: M = 15.02, SD = 6.05, expected range: 11–20).

During the electrophysiological session, children were seated comfortably in a chair and performed different computer-based tasks while we EEG was recorded. Before the learning phase, children were explained to carefully listen to discover the words of an “alien” language. First, the random stream was presented four times (Random Condition), followed by four presentations of the structured stream that contained the novel words (learning condition). Between each stream, a short pause was done, and children were encouraged to continue learning. This pause was especially important in the random condition to avoid children becoming discouraged due to the impossibility of learning. After the learning phase, children were assessed with a speech segmentation and a word-to-picture association task (Fig. 1).

The level of performance for word segmentation was assessed with a lexical decision task. On each trial, children heard one of the four words from the language or a nonword. Each word was presented eight times, thus, leading to 32 trials. The nonwords were compiled by mixing the syllables of the words pseudo-randomly. Children had to decide whether the item was a word from the language or not. We chose to assess word recognition with a lexical decision task (LDT) to collect behavioral responses for both types of test items as was done in adults (François et al., 2017b).

Children’s word to picture association performance was assessed with an associative word-picture matching task. On each trial, children heard one of the four words while two pictures were displayed on the computer screen. Each picture was presented three times, thus leading to 12 trials. They had to choose which picture (left or right) was associated with the word. Notably, both pictures appeared in the continuous trials. They had to choose which picture (left or right) was associated with an associative word-picture matching task. On each trial, children heard one of the four words while two pictures were displayed on the computer screen. Each picture was presented three times, thus leading to 32 trials. They had to choose which picture (left or right) was associated with the word. Notably, both pictures appeared in the continuous trials. They had to choose which picture (left or right) was associated with an associative word-picture matching task. On each trial, children heard one of the four words while two pictures were displayed on the computer screen. Each picture was presented three times, thus leading to 32 trials. They had to choose which picture (left or right) was associated with the word. Notably, both pictures appeared in the continuous trials. They had to choose which picture (left or right) was associated with a discrete Fourier transformation (Matlab; Natick, MA). Finally, the power at both target frequencies was obtained for each child and electrode in the random and learning conditions. All the scripts were manually prepared for this analysis. For the statistical analysis, we used a repeated-measures ANOVA with four within-subject factors: Frequency (Syllable Frequency, Word Frequency), Condition (Learning, Random), region of interest (ROI; Frontal, Central, Parietal, Occipital), and Lateralization (Left, Central, Right).

2.6. Scalp-based phase synchronization analysis – weighted Phase-Lag Index

To assess the phase relationship between different brain regions during speech segmentation, we used the wPLI. We selected this measure because it allows to avoid volume conduction since zero-lag synchronization is left out from the analysis, and thus, the reported measure is not based on shared sources (Bastos and Schoffelen, 2016; Stam et al., 2007). We performed the same pre-processing as in the frequency-tagging analysis, computed the Fourier transform, and finally obtained the wPLI using Fieldtrip Toolbox (Oostenveld et al., 2011) run in Matlab. For the statistical analysis, we defined four different regions of interest (ROI): frontal (F3, F4, Fz), central (C3, C4, Cz), parietal (P3, P4, Pz), and occipital (O1, O2, Oz) to perform a repeated-measures ANOVA with three within-subjects factors: Frequency (Syllable Frequency, Word Frequency), Condition (Learning, Random), and ROI (Fronto-Occipital, Fronto-Central, Fronto-Parietal, Occipito-Central, Occipito-Parietal, Parietal-Central). Post-hoc t-tests were conducted to determine the source of significant interactions and were corrected for multiple comparisons using the Bonferroni-Holm FDR correction method to avoid false positives (Benjamini and Hochberg, 1995). Finally, and for display purpose only, the surviving significant t-values were plotted separately for each target frequency using the BrainNet Viewer (Xia et al., 2013).

3. Results

3.1. Behavioral results

Fig. 2 shows the behavioral data obtained in the speech segmentation and word-picture association tasks. For the speech segmentation task (Fig. 2A), the comparison of performance against chance level (50 %) showed that children performed significantly above chance level for both words (M = 63.4, SEM = 2.15) and non-words (M = 62.2, SEM = 1.78, both P’s < .001). This result indicates that they were able to recall newly segmented words and to reject nonwords correctly. The level of
performance for words and non-words was not significantly different \(t(59) = 0.35; P = .72\). For the word-picture association task, the mean percentage of correct responses \(M = 70.49, \text{SEM} = 2.77\) was significantly above chance level \(t(59) = 7.39; p < .001\), indicating that children were also able to associate the pictures to the segmented words.

We explored the role of age by computing Pearson correlations between the age at test (in months) and the overall performance in both the speech segmentation and word-picture association tasks. Results showed that age was positively associated with the overall performance in the segmentation task \(r = .264; p = .041\); see Fig. 2C but not in the word-picture association task \(r = .075; p = .564\) even though these results did not survive the correction for multiple comparisons. Separated correlations for words and non-words in the segmentation task showed that age was positively associated with word recall \(r = .339; p = .008\); see Fig. 2D) while this was not the case for nonword rejection \(r = .063; p = .627\).

### 3.2. Electrophysiological results

#### 3.2.1. Frequency tagging analysis

Fig. 3A and B display the frequency-tagging analysis results showing clear peaks at word and syllable frequencies. The peak to word frequency was higher than for syllable frequency [significant main effect of Frequency: \(F(1,59) = 11.49; p = .001\)]. As predicted, considering the rhythmic presentation of pictures (matching word length), we observed a clear peak at the word frequency over occipital ROI [significant Frequency x ROI interaction \(F(1,177) = 22.09; p < .001\)]. Interestingly, some ROIs showed larger differences between Learning and Random conditions than others at both word and syllable frequency [evidenced by a non-significant main effect of Condition \(F(1,59) = 2.94; p = .092\) but a significant Condition x ROI interaction \(F(3,177) = 9.18; p < .001\) and a significant Condition x Frequency x ROI x Lateralization interaction: \(F(6,354) = 4.65; p = .007\); Fig. 3A & B]. To further explore these effects, we performed separate repeated-measures ANOVAs for both word and syllable frequency.

At the word frequency, a clear peak was found over all ROIs with a strong response over occipital regions [main effect of ROI: \(F(3,177) = 18.14; p < .001\)]. When looking at differences between conditions across ROIs [significant Condition x ROI interaction: \(F(3,177) = 3.54; p = .029\)] we found that those differences occurred over frontal regions only \([t(59) = −2.31; p = 0.025]\) with an enhanced power in the learning compared to the random condition. Interestingly, despite the clear peak at occipital ROI, there were no differences between conditions in this region \([t(59) = 1.17; p = 0.247]\). The Condition by ROI by Lateralization interaction was also significant \([F(6,354) = 3.74; p = .016]\). Further post-hoc analysis at the electrode level showed significantly enhanced power at word frequency compared to syllable frequency.
frequency in the learning compared to the random conditions at F4, Fz, F3 and P3 electrodes (F4: $t_{59} = -2.2; p = .032$; Fz: $t_{59} = -2.01; p = .048$; F3: $t_{59} = -2.45; p = .017$; P3: $t_{59} = -2.18; p = .033$; see Fig. 3C). However, none of these post-hoc comparisons survived the FDR correction for multiple comparison.

At the syllable frequency, we found clear peaks in all ROIs with a stronger response over frontal regions [main effect of ROI: $F(3,177) = 6.51; p = .002$]. Results also revealed an overall enhanced power in the learning compared to the random condition [main effect of Condition: $F(1,59) = 6.81; p = .01$]. Interestingly, this effect was differently distributed across ROIs [significant Condition x ROI interaction: $F(3,117) = 15.99; p < .001$] with enhanced power in the learning compared to the random condition over frontal ($t_{59} = -4.59; p < .001$) and central regions ($t_{59} = -3.44; p = .01$). No differences between the learning and random conditions were found over parietal ($t_{59} = -.64; p = .526$) and occipital ($t_{59} = 1.17; p = .247$) regions. The Condition by ROI by Lateralization interaction was also significant [$F(6,354) = 3.41; p = .011$]. Further post-hoc analysis at electrode level showed that the power at the syllable frequency was larger in the learning condition than in the random one (FDR correction, $p < .004$) at F4, Fz, F3, C3, Cz, and C4 electrodes (F4: $t_{59} = -4.66; p < .001$; Fz: $t_{59} = -4.11; p < .001$; F3: $t_{59} = -4.68; p < .001$; C3: $t_{59} = -3.77; p < .001$; Cz: $t_{59} = -3.11; p = .003$; C4: $t(59) = -3.02; p = .004$, see Fig. 3D).

3.2.2. **Scalp-based phase synchronization analysis**

Based on the frequency-tagging results, we expected a larger cross-talk between occipital (visual) and fronto-central electrodes or ROIs (speech segmentation) in the learning than in random condition, reflecting the involvement of visual cues in word binding and integration. Specifically, we expected stronger EEG phase synchronization between fronto-occipital electrodes at word frequency. Fig. 4A & B show the significant connections between electrodes at word and syllable frequency, respectively. The wPLI values were stronger at the syllable than at the word frequency [significant main effect of Frequency: $F(1,59) = 72.91; p < .001$]. Fig. 4C depicts the phase synchronization EEG spectra across ROIs, pointing to a higher phase synchronization strength in the learning compared to the random condition [main effect of Condition: $F(1,59) = 15.69; p < .001$]. As can be seen in Fig. 4D & E, this effect was differently distributed between ROIs [significant Frequency x Condition x ROI interaction: $F(1,59) = 3.26; p = .021$]. Further post-hoc $t$-tests conducted at the word frequency showed stronger EEG phase synchronization in the learning compared to the random condition in fronto-occipital ($t_{59} = 2.91; p = .005$), fronto-central ($t_{59} = 2.77; p = .008$), fronto-parietal ($t_{59} = 4.38; p < .001$) and parieto-central pairs of ROIs ($t_{59} = 2.91; p = .005$; FDR correction, $p < .008$; see Fig. 4D). For the syllable frequency we found significantly stronger phase synchronization in the learning compared to random condition in fronto-occipital ($t_{59} = 3.38; p = .001$) and occipito-central pairs of ROIs ($t_{59} = 2.98; p = .004$; FDR correction, $p < .004$; see Fig. 4E).

For the frequency-tagging and the phase synchronization analyses, we also included the factor age (in months) as a covariate. No significant main effect of age was found in these analyses [Frequency tagging: $F(12,47) = 0.73; p = .734$; Phase synchronization: $F(12,47) = 1.29; p = .254$] and none of the interaction survived the correction for multiple comparisons.
The present study investigated the functional role of brain oscillations during concurrent speech segmentation and meaning acquisition in sixty 9-year-old children. Specifically, we used frequency tagging and EEG phase synchronization analyses to assess neural tracking and long-distance scalp-based phase synchronization during an audio-visual SL task. Behavioural results showed that children successfully learned the four novel word-forms and their associated meaning. Frequency-tagging analysis of scalp EEG data revealed enhanced power at word and syllable frequencies in the learning compared to the random condition. Furthermore, we found evidence of enhanced long-distance EEG phase synchronization between frontal and parieto-occipital ROIs at both target frequencies. These converging results suggest that attentional mechanisms modulate word segmentation and meaning-mapping during children’s multimodal word learning induced by consistent visual cues and thus, facilitate learning (Hollich et al., 2000).

At the behavioural level, the results showed that children’s performance was above chance level for both the speech segmentation and the word-to-picture association tasks. Even though the children only learned four words, these results are the first to show that 9-year-old children can successfully perform speech segmentation and meaning-mapping concurrently. This finding is in line with previous reports showing that redundant cues such as prosody can trigger word-object associations in infants (Plunkett, 1997; Smith, 2000; Shukla et al., 2011). The present results also replicate our previous study in adults in which the presence of visual referents induced refined memory traces of the correct syllabic patterns (François et al., 2017b). Nonetheless, the level of performance in the word-picture association task may appear relatively low, with only 70% of correctly recognized associations. Several factors may account for this surprisingly low performance in the association task. First, it is important to keep in mind that the task is not as easy as it seems due to the relatively high syllabic rate (4.4 Hz). Second, children were not explicitly instructed to learn the meaning of the words nor that the pictures corresponded to the word onset and offset. Instead, they were told to listen carefully to the stream to discover the words of an "alien" language. Thus, the instructions were directed toward segmentation and not meaning-mapping leaving children to discover the relation between the pictures displayed on the screen and the auditory words by themselves. Finally, the behavioural tasks (lexical decision task or 2AFC) used in SL experiments are known to generate interferences created by the presentation of the illegal items (François et al., 2012). Further research may use implicit reaction time tasks as recently described in studies with adults (Batterink et al., 2015; François et al., 2012). Overall, our results suggest that 9-year-old children and adults exhibit relatively similar behavioural patterns of explicit recall.

Despite the relatively low variability in age, we explored the association between age and behavioural measures of word segmentation and meaning-mapping. We observed that age in months predicted the overall performance in the speech segmentation task with older children exhibiting higher performance than their younger peers. Interestingly, implicit learning mechanisms, which contribute to SL, seem to follow a U-shaped developmental trajectory across the lifespan, with the knee point observed around nine years of age (Juhasz et al., 2019; Janacsek et al., 2019).
revealed maximal N400 effects over fronto-central regions (Cunillera et al., 2010). Nonetheless, we observed that age predicted word recognition and that facilitated the segmentation of possible words. Therefore, it is difficult to conclude to which extent the observed correlations indicate an improvement of SL with age, or whether they are due to an increasing capacity to use multiple cues when learning new words (Cunillera et al., 2010a). Nonetheless, we observed that age predicted word recognition but did not predict nonword rejection in the speech segmentation task. This last result suggests that the recall of novel word forms is more dependent on age compared to nonwords rejection. Again, these results should be interpreted carefully due to the limited range of age explored as we tested only one single group of children. Such a narrow range of age limits the possibility to elaborate on the changes in cue-weighting that supports language acquisition across development. Further studies are needed to appropriately track the developmental trajectories of concurrent speech segmentation and meaning-mapping as well as its neural signatures.

The frequency tagging analysis revealed clear peaks at both syllable and word frequencies during exposure to the streams. At both target frequencies, we found enhanced power in the learning compared to the random condition. However, the differences survived FDR correction only for the syllable frequency suggesting that differences between conditions were higher when tracking the syllables than the words. The synchronized activity for the syllable frequency showed a frontal distribution, confirming previous ERP data obtained in adults (François et al., 2017b). Using a similar experimental design, François et al. (2017b) found that audio-visual streams elicited a larger fronto-central ERP component (FN400) in the learning than in the random condition. Previous electrophysiological studies of speech segmentation in adults revealed maximal N400 effects over fronto-central regions (Cunillera et al., 2006; François et al., 2014; Diego Balaguer et al., 2007). Importantly, the level of neural alignment to the speech streams was significantly stronger at the word than at the syllable frequency. This result confirms previous findings in adults that show temporal alignment of brain activity with speech and musical stimuli (Luo et al., 2010; Luo and Poeppel, 2007; Nozaradan et al., 2012). Besides, we focused on EEG raw power and observed enhanced activity in the learning condition compared to the random at the word- and syllable frequency. While some studies have also focused on raw power (Buiatti et al., 2009; Nozaradan, 2014), others have used inter-trial phase coherence (ITC) where the phase is also considered (Buiatti et al., 2009; Batterink and Palier, 2017; Hennt et al., 2021). Results based on ITC analyses showed higher coherence in the learning than in the random condition at word frequency but the opposite pattern was observed at syllable frequency with higher ITC values in the random than in the learning condition. Several reasons may account for these differences between previous studies and the present one.

First, methodological differences may explain the different patterns of results. A recent study used both measures to explore the oscillatory activity during an auditory SL task and found similar results (Ordin et al., 2020). While ITC provides information about the synchronization precision across trials, raw power informs on the intensity of the brain response. Therefore, although ITC and power measures are related, they provide different, not overlapping information (Wößmann et al., 2017). Second, our study is the first to explore oscillatory activity in children during speech segmentation. Thus, maturational brain differences between children and adults may also explain the divergent patterns of results mentioned above. For instance, children’s neuronal alignment to speech differs from that of adults. A recent study found that cortical tracking of speech at the syllable rate (4–8 Hz) predicted speech-in-noise discrimination ability in 6- to 9-year-old children. Notably, the detrimental effect of increasing auditory noise on the level of neural tracking was stronger in children than in adults suggesting that children may exhibit an immature cortical tracking of speech (Vander Ghinste et al., 2019). Third, Hollich et al. (2000) proposed that the relative weight of linguistic and non-linguistic cues in word learning facilitation changes over the course of development. Specifically, while children might be biased to rely on domain-general cues during the initial stages of word-learning, they are likely to use language-specific information (e.g., grammatical rules) during later learning stages. Thus, developmental changes in speech brain-tracking and attentional factors such as the reliance on and weighting of different cues may account for the differences between adults and children (Ríos-López et al., 2020; Moore et al., 1999). In the present study, visual cues may have facilitated the discovery of words by prompting participants to focus on the first syllable of the words (synchronized with the picture presentation), which might, in turn, modulate sensory processing (Parasuraman, 1980; Waldorff et al., 1993). The high response to word-frequency over occipital regions further supports this idea (see Fig. 4B). Because the pictures were presented at the same frequency as the words, this neuronal response over occipital regions presumably reflects the neural alignment to pictures. Steady-state visual evoked potentials studies consistently showed that attention toward the stimulus could modulate the neuronal response to that stimulus (Hillyard et al., 1998; Keitel et al., 2013; Morgan et al., 1996; Kim et al., 2007). Therefore, the attentional capture by visual referent may enhance sensory processing and facilitate the overall word learning process.

The EEG phase synchronization analyses based on the wPLI revealed enhanced long-distance phase synchronization strength between frontal-occipital and frontal-parietal ROIs at word frequency in the learning compared to the random condition. This result suggests that attentional modulation induced by consistent visual cues may facilitate learning. This idea fits well with recent results obtained during an audio-visual movie visualization task and showing increased frontal-occipital phase synchrony in the sensor space when children listened to well-formed sentences compared to semantically incongruent sentences (Panda et al., 2020). Previous studies using similar methodologies have also shown that the neural tracking of a specific talker in higher-order auditory regions differs between attended and non-attended speech input (Ding and Simon, 2012; Lakatos et al., 2013). Moreover, endogenous fluctuations of brain excitability may drive differences in behavior during an attentional task (Helfrich et al., 2018). These reports suggest that voluntary attention to a specific stimulus modulates neural tracking and, therefore, stimulus encoding. Studies on infant’s gaze-following showed that word learning differences arise from different object-looking behaviors (Yu and Smith, 2012). Besides, multiple forms of visual attention may support statistical word-referent learning in different ways (see Talsma et al., 2010 for a review; Smith, 2000). Therefore, the fronto-occipital phase connectivity patterns found in the present study may reflect attentional modulations arising from visual cues and targeting frontal and temporal auditory speech brain regions. These attentional modulations may be crucial for optimizing multimodal word-learning mechanisms (Smith, 2000; Plunket, 1997; Hollich et al., 2006). However, the spatial resolution of the 16-electrode EEG recorded in the present study is somewhat limited. Further studies with a better spatial resolution must be conducted to shed more light on the brain mechanisms involved in multi-sensory word learning.

Interestingly, the frequency tagging and the phase connectivity analyses revealed unexpected results. We observed an enhanced syllable frequency power in the learning compared to the random condition and stronger long-distance scalp-based connectivity values for the syllable than for the word frequency. Although these results may seem surprising, Buiatti et al. (2009) reported similar findings in adults. In their
study, EEG frequency tagging results obtained during an auditory speech segmentation task showed larger differences between random and learning conditions in the syllable than word frequency. Here, children may have initially tried to generate predictions about possible word candidates based on the visual cues during early exposure to the random streams. However, the inconsistencies between the visual referents signaling trisyllabic words and the random auditory stream may have interfered with these predictions by continuously decreasing the level of attention. For the syllable frequency, we observed a reduction in power in the random compared to the learning condition. We also found a decreased EEG phase synchronization strength between frontal-occipital and occipito-central pairs of ROIs in the random condition. An interesting question for future research is whether attentional modulation during SL is due to participants’ awareness of impossible learning or if it is internally self-regulated without active decision-making processes. Indeed, intrinsic self-regulated learning mechanisms that select rich learning environments in which the best sources of information exist (see Begus et al., 2016) might explain why infants and children seem to be sensitive to the reliability of information during learning, paying particular attention to reliable speakers during language acquisition (Jaswal and Neely, 2006; Koenig and Harris, 2005; Sabbagh and Baldwin, 2001). For instance, 17-month-old infants show a decreased attention when presented with an unlearnable compared to a learnable artificial grammar (Gerken et al., 2011). Similarly, 4-year-old but not 3-year-old children can predict the reliability of an informant during novel word learning based on previous experience (Koenig and Harris, 2005). In our paradigm, the presence of multiple conflicting cues (unreliable association between visual and auditory information) may have automatically turned off their learning mode by decreasing their attentional resources during the random condition. However, future studies are needed to explore this hypothesis further.

An important concern about neural oscillations is whether the external stimuli directly induce them by phase-resetting or whether they merely reflect highly phase-locked brain responses to the repeated stimulus (see Olesler and Kayser, 2019 for a discussion). This debate is especially relevant as neural oscillations have been proposed to play a mechanistic role in language processing and speech segmentation (Giraud and Poeppel, 2012). Some researchers tried to discern the issues using indirect measures such as sub-threshold stimulation (Oever et al., 2017) and found that oscillation activity still arises in this condition. Others studied the oscillation activity after the stimulus presentation based on the idea that an oscillator would continue after the presentation of the stimulus (Lakatos et al., 2013). Recently, Doelling et al. (2019) confirmed this prediction by showing that the phase lag between the stimulus and the neural response was more variable in an evoked model than in an oscillatory model. Taken together, these findings support the hypothesis that neural oscillations reflect the coupling of an intrinsic oscillator and external stimuli rather than a succession of evoked responses by the stimuli.

Some methodological limitations are worth considering. First, volume conduction may introduce spurious correlations in the phase synchronization of the stimulus (Lakatos et al., 2013). Recently, Doelling et al. (2019) and found that oscillation activity still arises in this condition. Others studied the oscillation activity after the stimulus presentation based on the idea that an oscillator would continue after the presentation of the stimulus (Lakatos et al., 2013). Recently, Doelling et al. (2019) confirmed this prediction by showing that the phase lag between the stimulus and the neural response was more variable in an evoked model than in an oscillatory model. Taken together, these findings support the hypothesis that neural oscillations reflect the coupling of an intrinsic oscillator and external stimuli rather than a succession of evoked responses by the stimuli.

Some methodological limitations are worth considering. First, volume conduction may introduce spurious correlations in the phase synchronization of the stimulus (Lakatos et al., 2013). Recently, Doelling et al. (2019) confirmed this prediction by showing that the phase lag between the stimulus and the neural response was more variable in an evoked model than in an oscillatory model. Taken together, these findings support the hypothesis that neural oscillations reflect the coupling of an intrinsic oscillator and external stimuli rather than a succession of evoked responses by the stimuli.

5. Conclusion

To our knowledge, this is the first study investigating the functional role of brain oscillations and long-distance scalp-based phase synchronization during multimodal SL in children. Our results show (i) that 9-year-old children can simultaneously segment words and associate them onto meaning, (ii) that the presence of multimodal cues triggers speech segmentation and meaning-mapping process in children, (iii) that long-distance EEG phase synchronization between frontal and occipital pairs of electrodes contribute to multimodal word-learning. These results favor an integrative view of language learning, where multi-sensory cues and attentional mechanisms interact to enhance learning.

Data statement

The data that support the findings of this study are available on reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all the children and their families for participating in the study. We also thank the directors and teachers of the two schools where the children were tested and the two anonymous reviewers for their constructive comments on a previous version of the manuscript. We thank D. Cucurell, G. Cardon, and H. Tajadura for their help in different steps of the study. This research has been supported by a Spanish MINECO project (PSI2015-69132P) to CF. Finally, this work, carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

