Neurophysiological evidence for the interplay of speech segmentation and word-referent mapping during novel word learning

Clément François, Toni Cunillera, Enara Garcia, Matti Laine, Antoni Rodriguez-Fornells

Neurophysiological evidence for the interplay of speech segmentation and word-referent mapping during novel word learning

Clément François[a,b,c], Toni Cunillera[b], Enara Garcia[a,b], Matti Laine[d], Antoni Rodriguez-Fornells[a,b,e]

[a] Cognition and Brain Plasticity Group, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, 08097, Spain

[b] Dept. of Cognition, Development and Educational Science, Institute of Neuroscience, University of Barcelona, L'Hospitalet de Llobregat, Barcelona 08097, Spain.

[c] Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Barcelona, Spain

[d] Department of Psychology, Abo Akademi University, 20500, Turku, Finland

[e] Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain

***Corresponding Authors*:** Department of Basic Psychology, University of Barcelona, Campus de Bellvitge - Pavelló de Govern, 08908 L'Hospitalet de Llobregat, Barcelona, Spain.

E-mail addresses: fclement24@hotmail.com (C. François), antoni.rodriguez@icrea.cat (A. Rodriguez-Fornells).

Abstract

Learning a new language requires the identification of word units from continuous speech (the speech segmentation problem) and mapping them onto conceptual representation (the word to world mapping problem). Recent behavioral studies have revealed that the statistical properties found within and across modalities can serve as cues for both processes. However, segmentation and mapping have been largely studied separately, and thus it remains unclear whether both processes can be accomplished at the same time and if they share common neurophysiological features. To address this question, we recorded EEG of 20 adult participants during both an audio alone speech segmentation task and an audiovisual word-to-picture association task. The participants were tested for both the implicit detection of online mismatches (structural auditory and visual semantic violations) as well as for the explicit recognition of words and word-to-picture associations. The ERP results from the learning phase revealed a delayed learning-related fronto-central negativity (FN400) in the audiovisual condition compared to the audio alone condition. Interestingly, while online structural auditory violations elicited clear MMN/N200 components in the audio alone condition, visual-semantic violations induced meaning-related N400 modulations in the audiovisual condition. The present results support the idea that speech segmentation and meaning mapping can take place in parallel and act in synergy to enhance novel word learning.

Keywords: Speech segmentation, word-referent association, audiovisual statistical learning, Event-Related Brain Potentials, FN400.

1. Introduction

Learning a new language requires the complex task of isolating new auditory word-forms and associating them onto meanings. Several learning mechanisms have been proposed to explain how adults and infants can solve the word-to-world mapping problem (Kuhl, 2004; Davis and Gaskell, 2009; Rodriguez-Fornells et al., 2009). In particular, statistical learning (SL) or the ability to track regularities or patterns of various sorts in the input seems to be one of the core learning mechanisms for language acquisition, allowing human infants and adults to decipher the units contained in continuous speech streams even after a brief exposure (Saffran et al., 1996). Probably based on the idea that infants need to segment the auditory input first to then attribute meaning to the isolated words, the segmentation and mapping processes have been originally studied in a sequential manner. For instance, the first behavioral studies on segmentation and mapping were conducted in both infant and adult participants who were asked to perform a classic segmentation task followed by a consecutive word-picture mapping task (Mirman et al., 2008; Graf et al., 2007; Hay et al., 2011). Results of these studies indicated that the statistical learning mechanism creates possible word candidates ready to be mapped onto meaningful representations. These memory traces stemming from statistical learning may thus constitute a proto-vocabulary that is gradually created before being mapped to the corresponding conceptual units (Fernandes et al., 2009; Rodriguez-Fornells et al., 2009). This mapping could be accomplished by associative or statistical learning mechanisms that would bind conceptual representations onto these proto-vocabulary traces.

A recent article presents an alternative view on the original idea that segmentation and mapping processes operate in a sequential manner (Räsänen and Rasilo, 2015). These authors propose a computational model of joint word segmentation and meaning mapping based on empirical and simulated data in which a shared domain-general statistical learning mechanism may act both within and across modalities. Importantly, these authors propose that the learning process might be facilitated by the simultaneous appearance of the word and its visual referent. Indeed, new words are usually heard in different contexts and in some cases with the simultaneous appearance of the external referent associated to the new auditory word-form. Binding new-words and possible referents across different contexts may rely on a bootstrapping mechanism that helps to fill the language-learning gap. The statistical properties found in the speech signal as

well as the statistical consistency between speech and the world (external context and referents) might be therefore crucial to infer the possible meaning of a new word and also to help the speech segmentation process.

In line with this idea, Cunillera and colleagues (2010a) conducted a behavioral study using an audiovisual learning paradigm in which a word-segmentation task was coupled with a word-picture association task with varying degrees of association consistency and meaningfulness. In contrast to the sequential proposal that meaning mapping would occur after speech segmentation (Graf et al., 2007; Hay et al., 2011), the results of this study clearly showed that segmentation and word-referent mapping could be accomplished after a very short exposure, most likely in a simultaneous fashion. Word learning was more effective when visual referents were meaningful objects and speech segmentation performance increased with the presence of systematic word-picture associations. The benefit of visual cues on auditory segmentation was further confirmed with different types of multisensory cues shown to boost speech segmentation (Thiessen, 2010; Glicksohn and Cohen, 2013). These data are in agreement with the observation that the prosodic structure of child-directed speech offers the possibility to solve the speech segmentation and word-to world mapping problem in parallel (Yurovsky et al., 2012). Thus, several converging evidences and theoretical proposals concur by pointing out that both processes, speech segmentation and meaning mapping, could be functionally active in parallel during learning. However, all these studies based their interpretation on the analysis of the behavioral data collected after the learning has taken place.

In the present study, we went one step further by investigating the neurophysiological mechanisms involved when the two processes of segmenting new words and binding conceptual representations onto these newly isolated words are taking place at the same time. One crucial component of the Event-Related Potentials underlying semantic-conceptual processing is the N400 component (Kutas and Hillyard, 1980). Classical views of the central-parietal N400 associate its amplitude modulations to lexical and semantic retrieval processes (Kutas and Federmeier, 2000). Nonetheless, recent ERP studies have provided converging evidence for a fronto-central N400-like component (FN400) in artificial language learning tasks with increasing amplitude during the initial stages of extracting new words (Cunillera et al., 2009; De Diego-Balaguer et al., 2007; François et al., 2014). Interestingly, the topographical distribution

of the learning-related N400 component suggests the involvement of specific neurophysiological mechanisms indexing speech segmentation that differ from the lexical-semantic retrieval process (see also Dittinger et al. (2016)). Similar frontal N400 components were observed in 14-month-old infants performing a word-picture mapping task (Friedrich and Friederici, 2008; see also for similar ERP results in 17–21 month-old infants, Mills et al. (2005)). Frontally distributed N400 were also observed in language-learning studies using semantic priming with newly learned words (Mestres-Missé et al., 2007). Moreover, a similar FN400 has been recently associated to conceptual/semantic priming processes in several studies of memory recognition (see for a discussion, see Voss and Federmeier (2011), Voss et al. (2010)). Therefore the FN400 seems to be a good candidate for studying speech segmentation and word-picture association processes notably when occurring in parallel. Here, we used this FN400 (i) to explore the neurophysiological mechanisms of speech segmentation and word picture mapping when occurring in parallel during an initial learning phase and, (ii) to collect implicit brain responses of incorrect visual word-picture associations during an implicit test phase.

Besides the FN400, we took advantage of two other ERP components of interest, namely the Mismatch Negativity (MMN) and the N200. The MMN is a sensitive measure of pre-attentive, automatic and implicit auditory change detection, which may reflect the formation of an echoic memory trace within the auditory cortex (Näätänen et al., 2005). The MMN is elicited by rare stimuli presented in a sequence of repeated standard stimuli (Näätänen et al., 2005) and is observed for simple (Näätänen et al., 1978, 2005; Deguchi et al., 2010; Chobert et al., 2012) as well as for more complex auditory patterns of speech and non-speech stimuli (Boh et al., 2010; Herholz et al., 2009; Wang et al., 2012). The N200 component has been observed in artificial grammar learning studies with ungrammatical sequences eliciting larger N200 than grammatical sequences (Carrion and Bly, 2007; Selchenkova et al., 2014). The N200 has been also related to working memory processes underlying template mismatch mechanisms (Sams, Alho and Näätänen, 1983) and may thus index the acquisition of implicit knowledge (Selchenkova et al., 2014). Therefore, after a learning phase we exposed participants to test streams composed of standard tri-syllabic words and infrequent structurally illegal trisyllabic words to collect implicit measures of structural auditory change detection (for a similar procedure, see De Diego-Balaguer et al. (2007)).

In order to explore the neurophysiological mechanisms of speech segmentation and binding new-words onto conceptual representations, we exposed participants to a similar multimodal statistical learning paradigm as used by Cunillera et al. (2010a). Auditory streams composed of statistically concatenated new artificial tri-syllabic words were presented with (Audiovisual condition) or without (Audio alone condition) consistently associated visual information while we recorded EEG (see **Fig. 1** for the experimental design). For both the audio alone and audiovisual (AV) conditions, participants were also exposed to a baseline in which all the syllables appeared randomly, thus providing no statistical cue to word boundaries. Importantly, in the random streams of the AV condition both the syllables and the visual referents were randomly presented. After the learning phases, we collected brain activity during an implicit test in which structural (Audio alone condition) or visual binding mismatches (AV condition) were pseudo-randomly inserted in the streams. Behavioral measures of word recognition and word-picture associations were finally collected. For the learning phases of both conditions, we compared ERPs elicited by structured streams to those elicited by random ones. We hypothesized that the AV streams should provide facilitatory cues in the segmentation process by allowing participants to bind possible meaning representations onto newly segmented words (see **Fig. 1**). The facilitation provided by this additional mapping process should be accompanied by enhanced word recognition performance during the explicit behavioral test. At the electrophysiological level, we expected to observe a similar FN400 in the two conditions if simultaneous segmentation and mapping involve similar neurophysiological mechanisms. On the other hand, if the simultaneous tracking of words and their referents involve an increased cognitive load, we expected to observe larger and/or later FN400 in the AV than in the audio alone condition. In order to test for these hypotheses, we performed a complementary analysis on the FN400 effects observed in the two conditions to provide direct evidence for the involvement of different neurophysiological mechanisms in the two processes. During the implicit test phase of the audio alone condition, we expected auditory structural mismatches to elicit MMN/N2 components (Carrion and Bly, 2007; Selchenkova et al., 2014). On the contrary, we expected the visual binding mismatches to elicit modulation of the lexical-semantic N400 potential during the implicit test phase of the AV condition (Ganis et al., 1996).

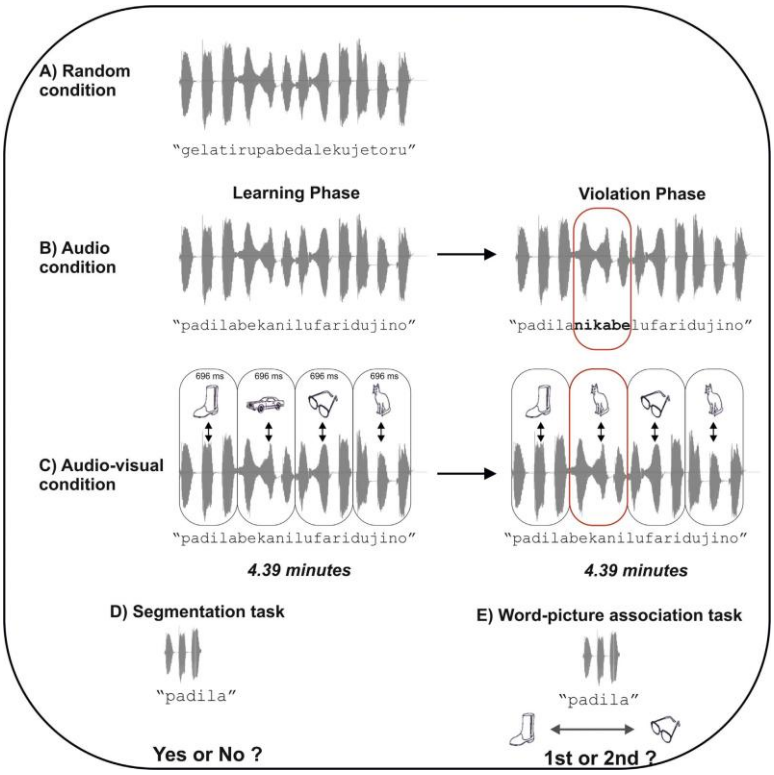

Fig. 1: Illustration of the experimental procedure. Two different languages and a random stream were used in each condition [A: Audio alone and B: Audiovisual (AV) condition]. The learning phases were immediately followed by an implicit test phase. C: Illustration of one trial from the segmentation task. D: Illustration of one trial from the word to picture association task.

2. Methods

2.1 Participants

Twenty-five volunteers participated in the study (11 men, mean age=22.7, SD=1.6). All of them were right-handed. Twenty of them were Spanish-Catalan bilinguals and 5 were Spanish monolinguals. None of them had a history of neurological deficits. Written consent was obtained from each volunteer prior to the experiment and they were paid after the experiment. The experiment was approved by the Ethics Committee of the University of Barcelona. Five participants were discarded from the final analysis due to excessive eye-movements (n=3) or excessive muscle artifacts (n=2) during the electroencephalogram recording

2.2 Stimuli

Forty-eight different CV syllables were used to create four different artificial speech streams with the same structure as in Cunillera and colleagues (2010a). Each language contained four tri-syllabic pseudo-words concatenated in a pseudo-random order with

no immediate repetition of the same pseudo-word. None of the syllables were repeated across the different language streams. The streams were synthetized using the MBROLA speech synthesizer with the Spanish male database (es1). All phonemes had the same duration (116 ms) and pitch (200 Hz). The streams were first created by synthetizing short streams of 100 words each and repeating them four times to reach a total of 400 words. Each word had 696 ms of duration, leading to 4 min and 39 sec streams. Thus, transitional probabilities were the only cue to segment the streams, as no acoustic cues at word boundaries were present. The transitional probability between syllables was 1.0 within words and 0.33 at word boundaries.

In addition to these statistically structured streams, random streams were created by pseudo-randomly mixing the syllables used in the four different language streams. Random streams had the same duration as the structured streams but the transitional probability between all syllables was 0.09. Therefore, there was no way to extract words based on statistical cues, and ERPs from this condition could be used as a baseline.

For each participant, two streams were presented in the audio alone condition and two other streams in the audiovisual condition. In the audio alone condition, the streams were presented with a stable fixation cross on a computer screen. In the audiovisual condition, four different pictures were synchronously presented with the four words of the 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 & Vanderwart, 1980). Importantly, each word was associated with a single picture resulting in four fully consistent word-picture associations.

Four implicit test streams were also created. In the audio alone condition, the implicit test streams consisted of the same language stream previously heard in which illegal non-words were pseudo-randomly inserted without immediate repetition. Illegal non-words were composed of the same syllables of a previously exposed word but in the opposite order: the order of the first and last syllables was reversed (see **Fig. 1**). Each illegal item was repeated 8 times thus leading to 32 illegal items in the implicit test stream. In the audiovisual condition, the violation consisted of the presentation of a picture that was previously associated to another word.

2.3 Procedure

For each stream, the participants were required to listen carefully to the stream with the task of discovering the words of an "alien" language. One random, two structured and two implicit test streams were used in each of the two conditions (audio alone and audiovisual). The order of presentation was counterbalanced across participants. For each language stream, the experiment was composed of four phases: the random, the learning, the implicit test and the explicit behavioral test phase.

After being presented with the random, learning and implicit test phases, performance for word segmentation was assessed with a lexical decision task (LDT). In each trial, a word from the language or a non-word was randomly presented in the auditory modality. Non-words were compiled by mixing the syllables of the words pseudo-randomly. The participants had to decide whether the item was a word from the language or not. Previous studies on speech segmentation have used a two-alternative forced choice (2AFC) test to assess word recognition (Cunillera et al., 2009; François et al., 2011; De Diego Balaguer et al., 2007). However, the 2AFC only provides a single response for each pair of test-items thus leaving it open as to whether the word is correctly accepted and/or the non-word is correctly rejected. Therefore, we chose to assess word recognition with a LDT in order to collect behavioural responses for both types of test items.

In the audiovisual condition only, participants' word to picture association performance was assessed with a 12-trial associative word to picture matching task. On each trial, the participants heard a word of the language while two pictures were displayed on the computer screen. The participants had to choose which picture (left or right) was associated to the word. Importantly, both pictures were contained in the learning and implicit test streams.

2.4 data acquisition and analyses

The EEG signal was recorded from the scalp using tin electrodes mounted in an electrocap (Electro-Cap International) and located in 29 standard positions during Day 1 and Day 5 in *Experiment 1* and during the entire session in *Experiment 2*. Biosignals were re-referenced off-line to the left and right mastoidal electrodes. Vertical eye movements were monitored with an electrode placed at the infraorbital ridge of the right eye. Electrode impedances were kept below 5 kΩ. The electrophysiological signals were digitalized at a rate of 250 Hz. Electrophysiological data were analyzed using ERPLAB 13.5.4b. The EEG was filtered off-line using a 30 Hz low-pass filter only for display

figures. Epoch rejection criteria were individually determined using a simple voltage threshold within a range of +/- 50 for eye electrode and +/- 75 µV for the other channels and forward visually checked for each trial and participant. For the analysis, we focused on the learning phases of the experiment. Epochs of 900 ms were time-locked to both object and word presentation considering a -100 ms pre-stimulus baseline.

2.5 ERP analyses of the learning phases

Mean amplitudes in different time-windows (TWs) encompassing the major ERP components observed during the learning phases were selected based on previous results in the literature (Cunillera et al., 2009; De Diego Balaguer et al., 2007; François et al., 2014) and on visual inspection of the ERP components. Thus, a TW in the 200-350 ms time range was selected for analyzing the audio alone condition and a TW in the 400-550 ms time range was selected for analyzing the AV condition. As a first step, we conducted separate analyses for the two conditions and for each of the TWs. As a second step, the difference waveforms (structured minus random) in the two conditions and the two TWs were directly compared.

For the first step analyses, overall repeated measures ANOVAs were carried out separately for each condition with the mean amplitudes in the selected TWs and including Stream condition (Structured stream vs. Random stream) and Electrode (15 levels: Fz, F7/8, F3/4, Cz, C3/4, T3/4, Pz, P3/4, T5/6) as a within-subject factors. In order to decompose significant Stream condition x Electrode interactions, subsequent topographical analyses were carried out using twelve of the 15 selected electrodes. The ANOVA's design included 12 selected electrodes (F7/8, F3/4, T3/4, C3/4, T5/T6, P3/4) divided according to three topographical factors: Hemisphere [right (F7, F3, T3, C3, T5, P3) vs. left (F8, F4, T4, C4, T6, P4)], Anterior-Posterior [frontal (F7, F3, F8, F4) vs. central (T3, C3, T4, C4) vs. parietal (T5, P3, T6, P4)], and Laterality [lateral (F7, T3, T5, F8, T4, T6) vs. medial (F3, C3, P3, F4, C4, P4)].

For the second step analysis that aimed to directly compare the two conditions in the two time-windows, the difference waveforms of the two conditions (structured minus random) were submitted to two analyses: (i) a main ANOVA with 15 electrodes and (ii) a topographical analysis including 12 electrodes as done for separate analyses.

Finally, the voltage maps were also transformed into reference-free current source density (CSD) estimates for illustrative purpose. These CSD estimates represent the radial current flow entering and leaving the scalp and are proportional to the direction, location, and intensity of current generators that underlie an ERP map (Tenke & Kayser, 2012; Kayser et al., 2012). Moreover, CSD estimates are known to more closely represent the direction, location and intensity of current generators that underlie an ERP topography (Perrin et al., 1989; Kayser & Tenke, 2015; Mitzdorf, 1985; Nicholson, 1973).

2.6 ERP analyses of the implicit test phases

For the implicit test phase, we used the maximum voltage latencies of the MMN/N200 (audio alone condition, structural syllabic violations) and FN400 components (AV condition, visual binding mismatch) located in the difference waveforms (words minus non-words) at Cz electrode for the audio alone condition (see De Diego Balaguer et al., 2007 for structural syllabic violations), and at Pz for the audio-visual condition (see kutas and Federmeier, 2000 for lexical/semantic violations) in order to accurately select the TWs of interest in each condition. Two different TWs of 100 ms were considered for the analyses of the audio alone condition because two structural syllabic violations occurred in the non-words (peaks at Cz localized at 176 ms and 568 ms). In the audiovisual condition, one visual binding violation occurred and thus we considered a single TW in this case (peak at Pz: 312 ms) centered on the peak. Separate analyses were conducted for the audio alone and the audiovisual conditions with the factor Word-type (words vs. non-words) and Electrode (15 levels). As done for the learning phases, twelve of the 15 selected electrodes were used for topographical analysis. This analysis was used to decompose significant interactions in which the electrode factor was involved.

For all statistical effects involving two or more degrees of freedom in the numerator, the Huynh-Feldt epsilon was used to correct for possible violations of the sphericity assumption (Jennings and Wood, 1976). The uncorrected degrees of freedom and adjusted p-values after the correction are reported. For illustrative purposes only, 8 Hz and 16 Hz low-pass filters were applied to the grand-average ERPs for the audio alone and audiovisual conditions, respectively.

3. Results

3.1 Behavioral data

For the speech segmentation task (see results in **Fig. 2**), the results of the repeated measures ANOVA including condition (audio alone vs. audiovisual) and item type (Word vs. Non-words) as within-subjects factors revealed a significant condition x item type interaction [$F(1, 19) = 14.16$; $p = .001$]. Post-hoc analysis revealed that the level of performance for non-words detection was better in the audiovisual (77.5% correct responses) than in the audio alone condition (62%; $p = .003$). This difference was not significant for words ($p = .57$). Comparison of performance against chance level (50%) showed that the participants' level of performance was above chance for both words and non-words and in both conditions (all p's < .008).

For the associative meaning task (see **Fig. 2**), the mean percentage of correct responses was significantly above chance level (91.5%; $t(19) = 14.74$; $p < .001$), indicating that the participants were able to associate the pictures to the words.

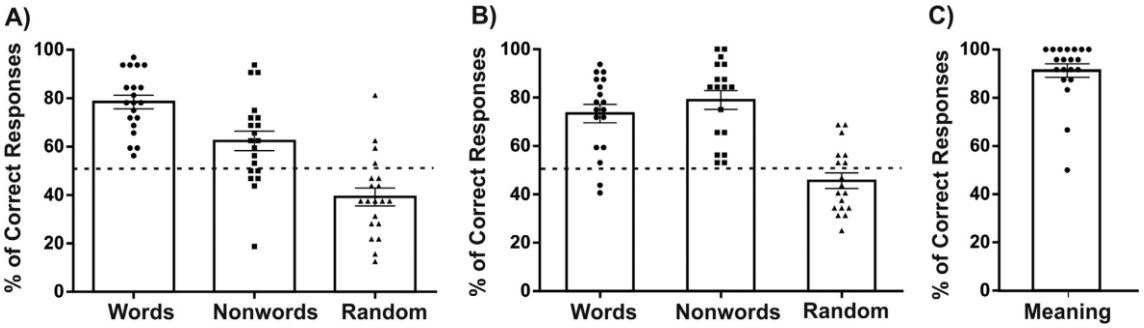

Fig. 2: Behavioral data. Percentage of correct responses in the segmentation task for words and non-words (A: Audio alone, B: Audiovisual) and in the word to picture association task (C). Dots represent individual values and bars correspond to the mean and standard error of the mean (SEM) in each condition. Note that performance for words and non-words were combined in the random condition.

3.2 ERP data

3.2.1 Learning phases

Figures 3A and 3B depict the grand average ERPs elicited by language and random streams in the two conditions for the whole learning phases.

3.2.2 Separate analyses for each condition

In the audio alone condition, a broadly distributed negativity with maximum amplitude over fronto-central electrodes in the 200-350 ms latency range was larger for the structured than for the random streams [main effect of Stream (Language vs. Random): $F(1,19) = 7.02$; $p < .02$; Stream x Electrode (15 levels): $F(14,266) = 3.57$; $p < .02$]. The topographical analysis with 12 electrodes revealed that the Stream x Laterality and Stream x Laterality x Anterior-posterior interactions were also significant [($F(1,19) = 6.99$; $p < .02$) and ($F(2,38) = 3.39$; $p < .05$), respectively]. Additional analyses revealed that the effect was largest at medial fronto-central electrodes [frontal (F3 & F4) vs. central (C3 & C4): $F(1,19) = 2.23$; $p > .1$; frontal vs. posterior (P3 & P4): $F(1,19) = 5.0$; $p < .04$; central vs. posterior: $F(1,19) = 6.1$; $p < .03$].

In the audiovisual condition (**Fig. 3B**), a frontal negative component resembling the FN400 but with a later onset than in the audio alone condition, was larger for the structured than for the random streams [for the TW 400-550 ms: main effect of Stream ($F(1,19) = 7.16$; $p < .01$), Stream x Electrode (15 levels): $F(14,266) = 5.3$; $p < .001$]. The topographical analysis with 12 electrodes revealed that the Stream x Anterior-Posterior interaction was significant ($F(2,38) = 7.08$; $p < .02$). Additional analyses revealed that the effect was largest at medial fronto-central electrodes [frontal (F3 & F4) vs. central (C3 & C4): $F(1,19) = 1.8$; $p > .1$; frontal vs. posterior (P3 & P4): $F(1,19) = 7.77$; $p < .02$; central vs. posterior: $F(1,19) = 16.3$; $p < .001$].

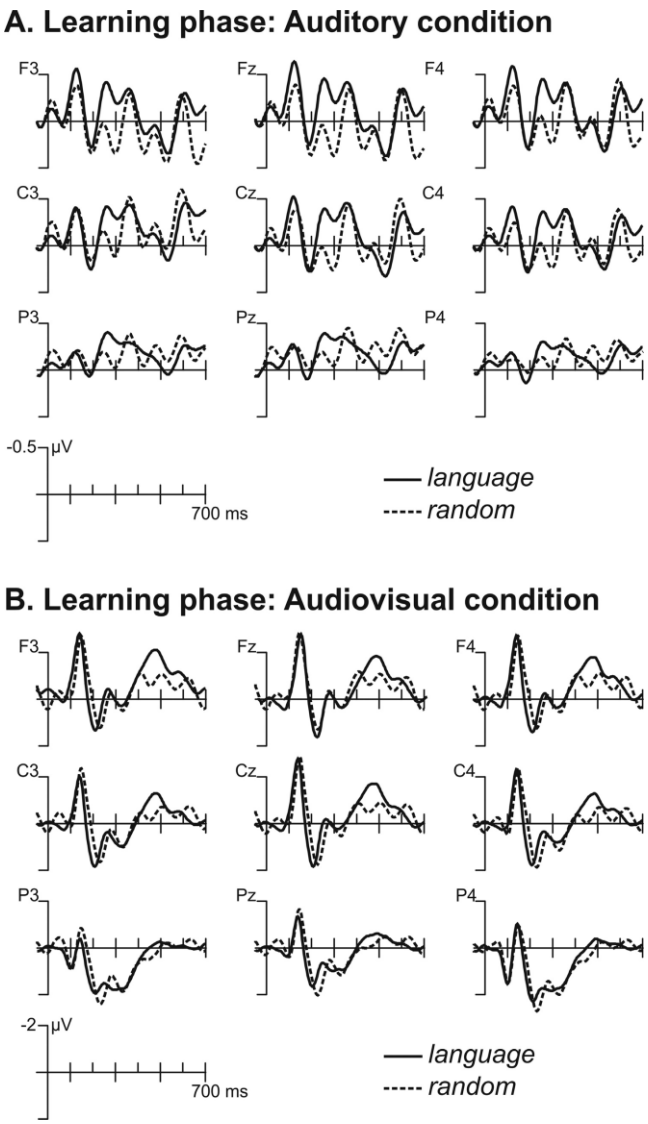

Fig. 3: ERP averages for the learning phases. Grand average across participants over 15 electrodes for the Language (words) and Random (non-words) streams in the Audio alone (A) and Audiovisual condition (B) condition.

3.2.3 Comparison of the two conditions

In Figure 4a we depicted the difference waveforms (structured minus random) for both conditions: audio alone streams and AV streams. As it can be observed, the audio alone streams elicited an early FN400 (between 200-350 ms with a maximum peak latency at 330 ms). In the AV condition, the FN400 was also observed but with an onset at 400 ms (extending until 550 ms with a maximum peak latency at 470 ms). Therefore, as a second step analysis, we directly compared the audio alone and audiovisual conditions at the different time windows (introducing Time range as two-level time-window amplitude measures: early, 200-350 ms and late, 400-550 ms). The results revealed no main effects of condition or Time range, but a significant Condition x Time

range interaction [$F(1,19) = 14.89$; $p = .001$] and a three-way Condition x Time range x Electrode (15 levels) interaction [$F(14,266) = 6.23$; $p < .001$]). The topographical analysis with 12 electrodes revealed that the Condition x Time range $F(1,19) = 14.9$; $p < .001$, Condition x Time range x Laterality $F(1,19) = 9.74$; $p < .005$ and the Condition x Time range x Laterality x Anterior-Posterior $F(2,38) = 10.8$; $p < .0003$) interactions were significant therefore confirming the differences in the onset and in the topographical distribution of the FN400 effects observed between the audiovisual and the audio alone condition.

Moreover, because CSD estimates are known to more closely represent the direction, location and intensity of current generators that underlie a ERP topography, we used current source density (CSD) maps (Kayser & Tenke, 2015) to determine more accurately the voltage source locations of the difference waveforms. The CSD maps suggested differences over central and frontal sites concerning the FN400 components in the audio alone vs. the AV conditions (see the topographic evolution of the CSD maps in **Fig 4c**).

In sum, we found that the structured audio alone streams, in comparison to the audio alone random streams, elicited an early negative fronto-central ERP component (early FN400, 200 to 350 ms). In contrast, the binding of meaning onto new words in the audio-visual streams induced a delayed FN400 in a latency range (400-550 ms) that is more resembling the classical lexical-semantic N400 component (see Kutas and Federmeier, 2000) and showing a different topographical distribution when compared to the audio alone FN400 effect.

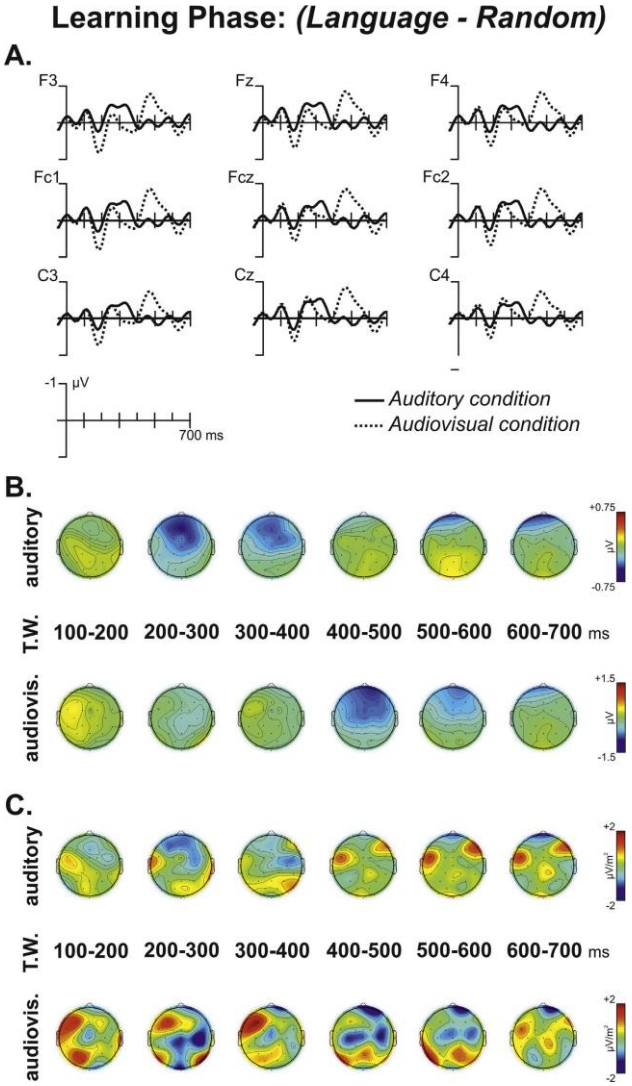

Fig. 4: A. Difference waveforms for the audio alone and audiovisual condition for the comparison between language and random streams. Notice the appearance of a FN400 in the audio alone condition at about 200 ms. In the audiovisual condition, the appearance of this FN400 is delayed until 400 ms. B. Scalp distribution of the difference waveforms (and corresponding Current Source Density maps depicted in C) is presented from 100 ms to 700 ms (100 ms time-window mean amplitude).

3.3 Implicit test phases

3.3.1 Audio alone condition

Figure 5 depicts the grand average ERPs for words and non-words in the online test phase of the audio alone condition. It should be noted that here we tested for structural violations by comparing items with a CBA syllabic structure instead of an ABC structure. Therefore, we expected two fronto-central MMN/N200 components to occur at the first and third syllabic positions of non-words (the syllable at B position remains the same). The averages shown in Figure 5 clearly showed increased fronto-central

negativities (MMN components) during the first and the third syllable (see the voltage distribution at the bottom of **Fig. 5**). The CSD maps of both negativities hinted at similar underlying sources (frontal and central) reiteratively appearing after the first and third syllables.

Statistically, violations at the first syllabic position elicited a significantly larger negativity than standard words (peak of the difference waveform at Cz: 176 ms; main effect of Word-type, TW 126-226 ms: $F(1,19) = 7.55$; $p < .02$; Word-type x Electrode: $F(14,266) = 3.28$; $p < .02$). This effect was largest over medial regions (Word-type x Laterality interaction [$F(1,19) = 15.11$; $p = .001$]). Violations at the third syllabic position elicited again a significantly larger negativity than standard words (peak at Cz: 568; TW 518-618 ms; main effect of Word-type: $F(1,19) = 11.75$; $p < .01$; Word-type x Electrode: $F(14,266) = 3.64$; $p < .02$). This effect was largest over medial and fronto-central regions (Word-type x Laterality x Anterior-Posterior: $F(2,38) = 8.30$; $p = .001$).

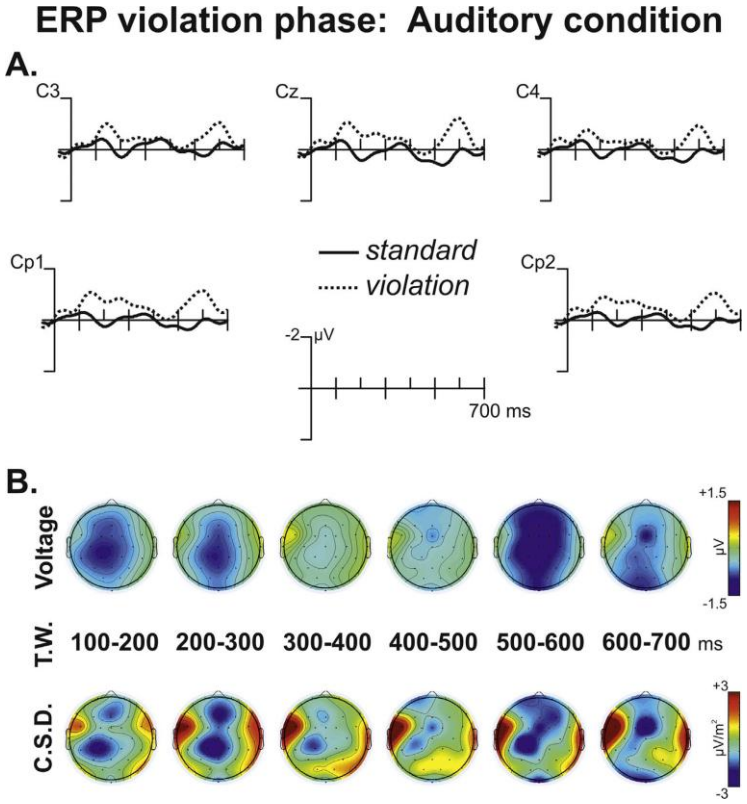

Fig. 5: Grand average across participants over midline and mid-central electrodes for the legal (words) and illegal items (non-words) in the audio alone implicit test phase. Below, the serial topographical maps (and CSD maps) are depicted from 100 ms to 700 ms (100 ms time-window mean amplitude).

3.3.2 Audio-visual condition

Figure 6 shows the grand average ERPs for matching and mismatching word-referent pairs. In this case, incorrectly associated pictures were inserted in the stream, thus allowing the collection of ERP data that reflect the detection of violations in the binding process. We expected an early modulation of the central-parietal semantic N400 component. As can be seen in Figure 6, the violations elicited a larger N400 than standard words (peak of the difference waveform at Pz: 312 ms) (main effect of Word-type, TW 262-362 ms: $F(1,19) = 4.35$; $p = .05$; Word-type x Electrode: $F(14,266) = 4.37$; $p = .001$). The N400 effect was largest over central-parietal regions, being left lateralized [Word-type x Hemisphere: $F(1,19) = 5.8$; $p < .01$; Word-type x Anterior-Posterior: $F(2,38) = 8.0$; $p < .01$; Word-type x Hemisphere x Laterality x Anterior-Posterior: $F(2,38) = 6.02$; $p = .06$]). Interestingly, the corresponding CSD maps showed very clear involvement of right and left parietal sources in the development of the N400 effect. Notice the sharp contrast between the CSD and voltage maps between the FN400 effects in the exposition phase (**Fig. 4b,c**) vs. the N400 component developed in the implicit test phase.

Thus, well before being behaviorally tested, the participants' neurophysiological responses indicated successful detection of online new-word structural mismatches in the audio alone condition. Interestingly, the participants also showed brain responses of visual binding mismatches in the audiovisual condition.

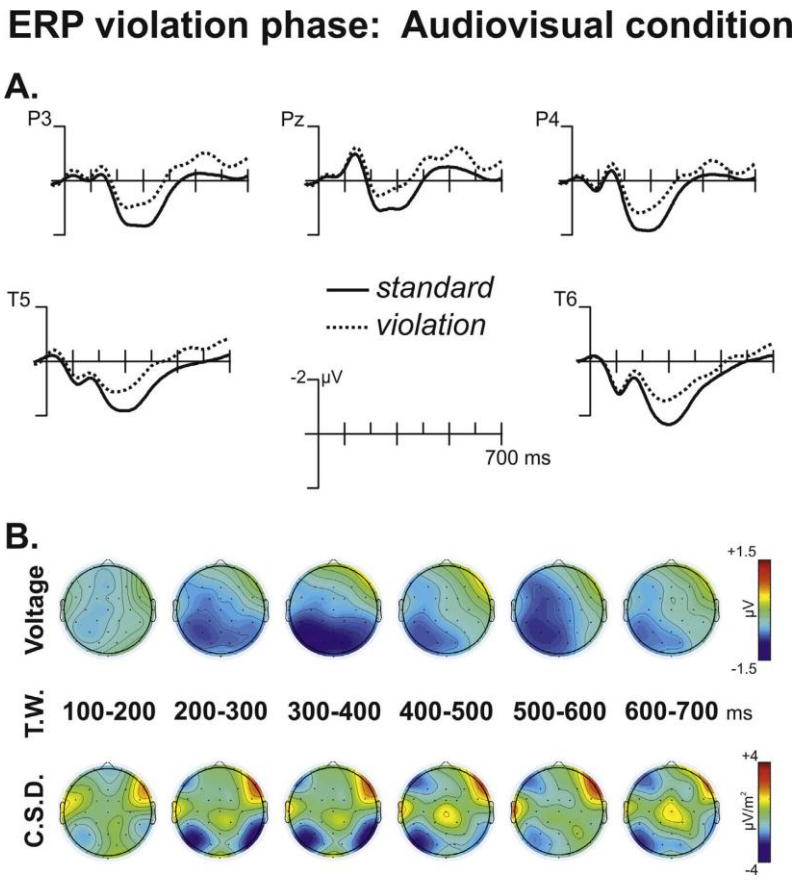

Fig. 6: Grand average across participants over posterior parietal-temporal electrodes for the correctly matching and mismatching (violations) word-referent pairs. Below, the serial topographical maps (and CSD maps) are depicted from 100 ms to 700 ms (100 ms time-window mean amplitude).

4. Discussion

The main goal of the present study was to investigate parallel speech segmentation and meaning mapping processes. We also aimed to determine the implication of the FN400 component in word learning when the binding of a visual referent is possible. With these goals in mind, EEG was recorded during the presentation of continuous streams of artificial words with and without consistently associated visual referents (following Cunillera et al., 2010a). In both conditions, we observed a fronto-central negative component during the learning phases, but with differences in the latency and topographical distributions. During the online implicit test, we found ERP evidence of successful auditory structural and visual semantic mismatch detection. At the behavioural level, the addition of visual referents during the learning phase was accompanied by higher rates of correctly rejected non-words. In sum, we found both

behavioral and electrophysiological evidence of two important milestones of vocabulary development (Saffran, 2014), word segmentation and meaning mapping, taking place in parallel.

4.1 Learning-related ERP modulations

In the audio alone condition, statistically structured streams elicited a larger fronto-central negativity than random streams which is in line with previous ERP studies on speech segmentation and adds further support to the idea that this "N400-like component" or FN400 may be considered as an online index of speech segmentation (Cunillera et al., 2009, De Diego Balaguer et al., 2007; François et al., 2014). In the audiovisual condition, the addition of synchronous visual referents induced a delayed fronto-central negativity or FN400 when comparing structured vs. random streams (see **Fig. 3**). The comparison between the FN400 in the audio alone and the AV conditions clearly showed a delayed onset of the FN400 in the AV condition (see the difference waveforms in **Fig. 4**). Moreover, despite the similarities in the scalp distribution of the FN400 component in the two conditions, the topographical analysis and CSD maps suggested the involvement of different neural sources underlying these two FN400. The differences in the latency and scalp distribution of the FN400 component observed between the audio alone and the audiovisual conditions suggest that the mapping of visual referent induced a cost of processing, probably reflecting the process of mapping newly created auditory word forms onto already existing conceptual information.

Based on Räsänen & Rasilo (2015), an alternative interpretation can be made. The presence of visual referents may qualitatively change the segmentation process by inducing a strategic switch from "bracketing" to "clustering" of the continuous streams of syllables. Indeed, instead of searching for the onsets of the words, the presence of visual referents may trigger a different learning mechanism that helps participants to build and store multisensory memory traces, which may be reflected in the delayed FN400 component we observed. The delayed FN400 component in the AV condition together with the CSD maps may suggest the involvement of different neural resources or cognitive processes. This also agrees with previous ERP findings in which a similar FN400 potential was observed in conjunction with conceptual/semantic priming processes during memory recognition tasks (for a discussion, see Voss and Federmeier, 2011; Voss et al., 2010). Taken together, our results suggest that speech segmentation

and word-to-picture association can occur after a very brief exposure and that different neural sources may underlie the elicitation of the FN400 components during the initial stages of word learning (McLaughlin et al., 2004). These results also converge with recent findings considering language acquisition as a joint inference problem for several components of language that need to be learned at once or in parallel (Johnson et al., 2010; Lim et al., 2015). This view stands in contrast with the seminal sequential approach (see Graf Estes et al., 2007) in which language acquisition is considered as a discrete sequence of inference problems for which learners may use one single source of information to acquire one single component of language which will then be used to facilitate the acquisition of a subsequent language component (Kuhl, 2004). Therefore, the results obtained in the audiovisual condition suggest that the detection of cross-modal regularities can be used in an interactive manner during word learning, yielding a more robust representation of the word forms. Nonetheless, because our study did not use visual material with a statistical structure to be extracted, we cannot make a clear claim on the domain-generality issue. Further studies are needed to investigate more closely and with better spatial resolution the neural regions involved during the speech segmentation process and during simultaneous segmentation and mapping of new words to meaningful referents. These studies will be important to better understand the benefit of audiovisual information on word learning. Moreover, it would be interesting to perform a similar study including a condition where a structured stream is presented with visual referents with varying degrees of consistency of word-referent association (Cunillera et al., 2010a and 2010b). This may provide additional evidence on how the consistency of word-referent mappings may influence word segmentation.

4.2 ERP evidence of online structural and lexical-semantic violations

An important and innovative aspect of the present study is the fact that we collected implicit measures of online violations before explicitly assessing the result of the learning in the final behavioural test. Indeed, several studies of speech segmentation, sequence learning or language acquisition have suggested that implicit measures are often more sensitive than explicit overt behavioural responses (François & Schön, 2010; Cleeremans, 2006; Tremblay et al., 1998; McLaughlin et al., 2004). In the audio alone condition, the violation streams contained illegal words with a CBA structure (in comparison with the ABC structure of the legal word) that were pseudo-randomly inserted throughout the language stream. These illegal words elicited a negative

component over medial and fronto-central regions for violations occurring at the first and last syllable positions (see **Fig. 5**). Previous results on artificial grammar learning of pitch sequences have shown that out of tune target tones appearing during the exposure phase induced a larger N200 than non-target tones (Selchenkova et al., 2014). We interpret these negative components as MMN/N200 components reflecting the detection of a mismatch between the newly acquired word-forms and the incoming syllable (see for similar results, De Diego-Balaguer et al., 2007). Interestingly, the effect size for the last syllable position was larger than for the first syllable position, suggesting that statistical regularities had been correctly extracted during the learning phase.

In contrast, in the audiovisual condition, the implicit test streams contained visual-semantic mismatches that were pseudo-randomly inserted. Visual-semantic violations elicited a larger left-lateralized N400 when compared to correct word-picture associations. Previous studies have reported larger frontal N400 effects for incongruous than congruous pictures in sentences (Kutas and Van Petten, 1990). Similar results have been found for pictures that are presented sequentially, but with larger frontal N400 effect for non-matching object pictures (e.g., fork-ring) than for matching pictures (e.g., fork-spoon; Barrett and Rugg, 1990). In addition, Holcomb and McPherson (1994) showed in an object-decision task a left-dominant frontally distributed N400 effect for unrelated vs. semantically related objects. Ganis et al. (1996) also showed a more frontal distribution of the N400 component for incongruent pictures at the end of written sentences, with an earlier onset (around 150 ms) and larger duration when compared to written endings of the sentences. Taken together, our results provide clear implicit electrophysiological evidence for correct online semantic mismatch detection.

4.3 Effect of audio-visual cues on word and non-word recognition

It is worth mentioning that at the behavioural level, we observed that word recognition was above chance level, indicating that participants were able to segment the streams in both conditions. Besides, performance was high for the recognition of learned word-to-picture associations, indicating that the participants were able to segment the continuous streams and associate visual referents to the newly learned words at the same time. Interestingly, when comparing the word segmentation performance between the two conditions, we found a significant Item type by Condition interaction showing that participants rejected non-words more accurately in the

audiovisual than in the audio alone condition. This suggests that the addition of visual referents may result in preserved memory traces of the correct syllabic patterns. The building of more robust memory trace may then allow participants improving their capacity to correctly reject non-words. Taken together these results confirm the idea that cross-modal associations may act as memory glue that provides more refined representations of the word learned (anchored in their visual referents). This may in turn result in overall better word recognition, in this particular case increasing the capacity to correctly identify non-words (Räsänen and Rasilo, 2015). It is also consistent with the proposal that more elaborate semantic processing during word learning aids subsequent memory (Balass, Nelson, and Perfetti, 2010; Bird, 2012; Cunillera, et al., 2010; Henderson et al., 2013) or does not delay the time-course of lexical integration (Hawkins & Rastle, 2016). Binding newly segmented words with already existing semantic referents might create more differentiated new-word representations that might allow participants to discriminate better between words and non-words during the lexical decision task. The binding between the new-word and their semantic referent might be stored initially into episodic memory and attached with pre-existing lexical-semantic memory-related networks. This process might allow the storage of this new-word into a richer associative network when compared to the new-word from the audio alone condition. The interaction between pre-existing lexical-semantic knowledge could therefore be used to tune up the learning of the new-words. However and considering the present results, it is also possible to think that the encoding of this richer episodic trace might require more cognitive resources (or deeper encoding). In our experiment, the increased demand in the audiovisual condition could not be reflected in better correct recognition of new-words but it clearly affected non-word correct rejection (in the sense that a more refined neural representation of the new-word is encoded in audiovisual condition). Two alternative explanations have to be considered. First the fact that we used a deterministic set of familiar visual stimuli that had to be associated with unfamiliar word forms may explain this effect. Indeed, during the learning phases the participants were exposed to audiovisual streams in which the associations between new word forms and already existing conceptual representations were fully consistent. These newly and fully consistent associations may (i) strengthen the memory traces until a plateau has been reached but also (ii) increase the distinctiveness of non-words composed of similar syllabic patterns but with a low probability (or familiarity). Second, in the present study the participants were presented with several streams back

to back. The repetition of the different phases through the entire experiment may have induced a specific interference effect on new word forms and not on nonwords. Indeed, there is evidence that the learning a first language decreases the capacity to learn the second language (Franco, Cleeremans & Destrebecqz, 2011). In sum, our results refine previous studies showing that multisensory cues enhance speech segmentation (Cunillera et al., 2010b; Thiessen, 2010; Glicksohn and Cohen, 2013; Yurovsky, Yu and Smith, 2012).

5. Conclusion

Our results show that the addition of visual referents in a speech segmentation task modulates the latency of the FN400 component. Additionally, the CSD analyses suggested the involvement of different neural sources when compared to the condition when no word-referent binding has to be done. The presence of a visual referent also aided subsequent non-word rejection by inducing more refined representations of the newly acquired word-forms.

Acknowledgments: We wish to thank the participants for their participation in the study. We are also grateful to David Cucurell for his help in programming the experiment as well as 2 anonymous reviewers for their constructive comments on a previous version. This research has been supported by a FYSSEN post-doctoral grant awarded to CF. ML was supported by grants from the Academy of Finland (project #260276) and the Abo Akademi University Endowment (the BrainTrain project).

Author contributions: Conceived and designed the experiments: CF AR EG ML. Performed the experiments: CF EG. Analyzed the data: TC CF. Wrote the paper: CF AR TC ML.

References

Balass, M., Nelson, J.R., Perfetti, C.A. (2010). Word learning: An ERP investigation of word experience effects on recognition and word processing. *Contemporary Educational Psychology*. 35(2): 126-140.

Barrett, S.E., Rugg, M.D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and Cognition*.14: 201-212.

Boh, B., Herholz, S.C., Lappe, C., Pantev, C. (2010). Processing of complex auditory patterns in musicians and nonmusicians. *Plos One*, 6 (7), e21458.

Bird, S. (2012). Expert knowledge, distinctiveness, and levels of processing in language learning. *Applied Psycholinguistics*, 33(4): 665-689.

Carrión, R.E., Bly, B.M. (2007). Event-related potential markers of expectation violation in an artificial grammar learning task. *Neuroreport*, 18: 191-195.

Chobert, J., François, C., Velay, J.L., Besson, M. (2012). Twelve months of active musical training in 8- to 10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cereb. Cortex,* 24(4): 956-967.

Cleeremans A. (2006). Conscious and unconscious cognition: a graded, dynamic, perspective. In: Jing Q, Rosenzweig MR, d'Ydewalle G, Zhang H, Chen H-C, Zhang K, editors. Progress in psychological science around the world. Vol I. Neural, cognitive and developmental issues. Hove: Psychology Press, pp. 401–418.

Cunillera, T., Camara, E., Toro, J. M., Marco-Pallares, J., Sebastian-Galles, N., Ortiz, H., Pujol, J. RodriguezFornells, A. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *NeuroImage.* 48: 541-553.

Cunillera, T., Laine, M., Cámara, E., Rodriguez-Fornells, A. (2010a). Bridging the gap between speech segmentation and Word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language,* 63: 295-305.

Cunillera, T., Camara, E, Laine, M., Rodriguez-Fornells, A. (2010b). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology,* 63(2): 260-274.

Davis, M.H., Gaskell, M.G. (2009). A complementary systems account of word learning: neural and behavioural evidence. Philos Trans R Soc Lond B Biol Sci, 364(1536): 3773-3800.

De Diego-Balaguer, R., Toro, J.M., Rodriguez-Fornells, A. Bachoud-Levi, A.C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PlosOne,* 2(11): e1175.

Deguchi, C., Chobert, J., Brunellière, A., Nguyen, N., Colombo, L., Besson, M. (2010). Pre-attentive and attentive processing of French vowels. *Brain Res,* 1366, 149-161.

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., Besson, M. (2016). Professional Music Training and Novel Word Learning: From Faster Semantic Encoding to Longer-lasting Word Representations. *Journal of Cognitive Neuroscience, 1-19. [Epub ahead of print]*.

Fernandes, T., Kolinsky, R., Ventura, P. (2009). The metamorphosis of the statistical segmentation output: lexicalization during artificial language learning. *Cognition,* 112(3): 349-366.

Fiser, J. Aslin, R.N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12:* 499-504.

François, C., Jaillet, F., Takerkart, S., Schön, D. (2014). Faster sound stream segmentation in musicians than in nonmusicians. *PlosOne*, 9(7): e101340.

François, C., Schön, D. (2010). Learning of musical and linguistic structures: comparing event-related potentials and behavior. *Neuroreport*, 21(14): 928-932.

Friederici, A.D., Steinhauer, K., Pfeifer, E. (2002). Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. *Procedings of The National Academy of Science of the United States of America,* 99: 529-534.

Friedrich, M. Friederici, A.D. (2008). Neurophysiological correlates of online word learning in 14-month-old infants. *Neuroreport,* 19: 1757-1761.

Ganis, G., Kutas, M., Sereno, M. (1996). The search for common sense: an electrophysiological investigation of the semantic analysis of words and pictures in sentences. *Journal of Cognitive Neuroscience,* 8: 89-106.

Glicksohn, A., Cohen, A. (2013). The role of cross-modal association in statistical learning. *Psychon Bull Rev*, 20(6): 1161-1169.

Graf, E.K., Evans, J.L., Alibali, M.W., Saffran, J.R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science.* 18: 254-260.

Hawkins, E.A., Rastle, K. (2016). How does the provision of semantic information influence the lexicalization of new spoken words? *The Quarterly Journal of Experimental Psychology*, 69(7): 1322-1339.

Hay, J.F., Pelucchi, B., Graf Estes, K.G., Saffran, J.R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology,* 63: 93-106.

Henderson, L., Weighall, A., Gaskell, G. (2013). Learning new vocabulary during childhood: Effects of semantic training on lexical consolidation and integration. *Journal of Experimental Child Psychology*. 116: 572-592.

Herholz, S.C., Lappe, C., Pantev, C. (2009). Looking for a pattern: an MEG study on the abstract mismatch negativity in musicians and nonmusicians. *BMC Neurosci*. 10, 42.

Holcomb, P.J., McPherson, W.B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain and Cognition*, 24: 259-276.

Jennings, J.R., Wood, C.C. (1976). Epsilon-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology,* 13: 277-278.

Johnson, M., Frank, M.C., Demuth, K., Jones, B.K. (2010). Synergies in learning words and their referents. *Adv Neural Inf Process Syst*, 23: 1018-1026.

Kayser, J., Tenke, C.E., Kroppmann, C.J., Alschuler, D.M., Fekri, S., Gil, R., Jarskog, L.F., Harkavy-Friedman, J.M., Bruder, G.E. (2012). A neurophysiological deficit in early visual processing in schizophrenia patients with auditory hallucinations. *Psychophysiology*, 49: 1168-1178.

Kayser, J., Tenke, C.E. (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology. *Int J Psychophysiol*, 97(3): 171-173.

Kuhl, P.K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience,* 5: 831-843.

Kutas, M., Hillyard, S.A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science,* 207(4427): 203-205.

Kutas, M., Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science,* 4: 463-470.

Kutas, M., Van Petten, C. (1990). Electrophysiological perspectives on comprehending written language. *Electroencephalography & Clinical Neurophysiology Supplement*, 41: 155-167.

Lim, S., Lacerda, F., Holt, L.L. (2015). Discovering functional units in continuous Speech. *J. of Exp. Psychol: Human Perception and Performance,* 41: 1139-1152.

McLaughlin, J., Osterhout, L., Kim. A. (2004). Neural correlates of second language word learning: minimal instruction produces rapid change. *Nat. Neurosci*, 7: 703-704.

Mestres-Missé, A., Rodriguez-Fornells, A., Munte, T.F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex*, 17: 1858-1866.

Mills, D.L., Plunkett, K., Prat, C., Schafer, G. (2005). Watching the infant brain learn words: effects of vocabulary size and experience. *Cogn. Dev*, 20: 19-31.

Mirman, D., Magnuson, J.S., Estes, K.G., Dixon, J.A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, 108(1): 271-280.

Mitzdorf, U. (1985). Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and EEG phenomena. *Physiological Review*, 65, 37-100.

Näätänen, R., Gaillard, A.W.K., Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol*. 42 (4), 313-329.

Näätänen, R., Jacobsen, T., Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. *Psychophysiology*, 42(1), 25-32.

Nicholson, C. (1973). Theoretical analysis of field potentials in anisotropic ensembles of neuronal elements. *IEEE Transactions on Biomedical Engineering*, 20, 278-288.

Perrin, F., Pernier, J., Bertrand, O., Echallier, J.F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr Clin Neurophysiol*, 72(2), 184-187.

Räsänen, O., Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, 122(4): 792-829.

Rodriguez-Fornells, A., Cunillera, T., Mestres-Missé, A., de Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philos Trans R Soc Lond B Biol Sci*, 364(1536): 3711-3735.

Saffran, J.R., Aslin, R.N. Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, 27: 1926-1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., Newport, E.L. (1999). Statistical learning of tone sequences by adults and infants. *Cognition, 70:* 27-52.

Saffran, J. (2014). Sounds and meanings working together: Word learning as a collaborative effort. *Language Learning*, 64(Suppl 2): 106-120.

Selchenkova, T., François C., Schön D., Corneyllie A., Perrin F., Tillmann B. (2014). Metrical presentation boosts implicit learning of artificial grammar. *PlosOne*, 9(11): e112233.

Snodgrass J.G., Vanderwart, M. (1980). Standarized set of 260 pictures. Norms for afreement, mage agreement, familiarity and visual complexity. *Journal of Experimental Psychology Human Learning*, 6(2): 174-215.

Tenke, C.E., Kayser, J. (2012). Generator localization by current source density (CSD): implications of volume conduction and field closure at intracranial and scalp resolutions. *Clin Neurophysiol*, 123: 2328–2345.

Thiessen, E.D. (2010). Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning. *Cognitive Science*, 34: 1093-1110.

Tillmann, B., McAdams, S. (2004). Implicit learning of musical timbre sequences: statistical regularities confronted with acoustical (dis)similarities. *J Exp Psychol Learn Mem Cogn*, 30(5): 1131-1142.

Tremblay, K., Kraus, N., McGee, T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech- sound training. *Neuroreport*, 9: 3557-3560.

Voss, J.L., Schendan, H.E., Paller, K.A. (2010). Finding meaning in novel geometric shapes influences electrophysiological correlates of repetition and dissociates perceptual and conceptual priming. *NeuroImage*, 49(3): 2879-2889.

Voss, J.L., Federmeier, K.D. (2011). FN400 potentials are functionally identical to N400 potentials and reflect semantic processing during recognition testing. *Psychophysiology*, 48(4): 532-546.

Wang, X.D., Gu, F., He, K., Chen, L.H., Chen, L. (2012). Preattentive extraction of abstract auditory rules in speech sound stream: a mismatch negativity study using lexical tones. *PlosOne*, 7 (1), e30027.

Yurovsky, D., Yu, C. Smith L.B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. *Frontiers in Psychology*, 3, 374.