# Assessing segmentation processes by click detection: online measure of statistical learning, or simple interference?

Ana Franco · Vinciane Gaillard · Axel Cleeremans · Arnaud Destrebecqz

**Abstract** Statistical learning can be used to extract the words from continuous speech. Gómez, Bion, and Mehler (*Language and Cognitive Processes*, 26, 212–223, 2011) proposed an online measure of statistical learning: They superimposed auditory clicks on a continuous artificial speech stream made up of a random succession of trisyllabic nonwords. Participants were instructed to detect these clicks, which could be located either within or between words. The results showed that, over the length of exposure, reaction times (RTs) increased more for within-word than for between-word clicks. This result has been accounted for by means of statistical learning of the between-word boundaries. However, even though statistical learning occurs without an intention to learn, it nevertheless requires attentional resources. Therefore, this process could be affected by a concurrent task such as click detection. In the present study, we evaluated the extent to which the click detection task indeed reflects successful statistical learning. Our results suggest that the emergence of RT differences between within- and between-word click detection is neither systematic nor related to the successful segmentation of the artificial language. Therefore, instead of being an online measure of learning, the click detection task seems to interfere with the extraction of statistical regularities.

**Keywords** Statistical learning · Speech segmentation · Online measures

A. Franco · V. Gaillard · A. Cleeremans · A. Destrebecqz
Cognition, Consciousness and Computation Group, Université Libre
de Bruxelles, Brussels, Belgium

A. Franco (✉) · V. Gaillard · A. Cleeremans · A. Destrebecqz
Center for Research in Cognition & Neurosciences, Université Libre
de Bruxelles, 50 Avenue F.-D. Roosevelt, 1050 Brussels, Belgium
e-mail: afranco@ulb.ac.be

Over the past 15 years, statistical-learning research has shown that adults, children, and infants are able to track statistical patterns in their environment (Aslin, Saffran, & Newport, 1998; Fiser & Aslin, 2002; Saffran, Aslin, & Newport, 1996). This ability seems to involve incidental (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), automatic (Turk-Browne, Jungé, & Scholl, 2005), and domain-general (Kirkham, Slemmer, & Johnson, 2002) learning mechanisms. Importantly, statistical learning may play an important role in the early stages of language acquisition, such as speech segmentation (see Romberg & Saffran, 2010, for a review). Indeed, within a language sample, pairs of syllables or sounds are in general more correlated with each other within a word than when they occur at the boundary between two consecutive words. Accordingly, a series of studies have confirmed that the transitional probabilities (TPs) between sounds can be tracked in order to discover word boundaries (e.g., Aslin et al., 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Thiessen & Saffran, 2003).

Traditionally, statistical learning of artificial languages is measured through offline measures. After exposure to a continuous speech stream, participants perform a two-alternative forced choice (2AFC) task or a recognition task in which they have to distinguish between the words and novel arrangements of the same syllables (Abla, Katahira, & Okanoya, 2008; Saffran et al., 1997). Other techniques have been developed in order to measure statistical learning in infants. Most studies have used the head turn preference procedure (Kemler Nelson et al., 1995; see Gerken & Aslin, 2005, for a review). In this paradigm, infants are seated between two speakers with mounted lights and are free to turn their heads. After exposure, in a test phase, either the right or the left light flashes as an auditory test stimulus is emitted from the corresponding speaker. The time during which the infant turns her head toward the emitting speaker is used as a preference measure for the corresponding stimulus. Noise detection

(Morgan, 1994; see also Morgan & Saffran, 1995) is another technique, which consists in training 9-month-old infants to turn their heads in response to short buzzes. In a first phase, these buzzes are presented during the time intervals between multisyllabic strings. In a second phase, buzzes are presented within these strings, between any two syllable pairs. The differences in response latencies between the two phases have been interpreted as an indication of infants' ability to perceptually organize the stimulus strings.

Although these tasks inform us about the nature and amount of the acquired knowledge, they say little about the temporal dynamics of statistical learning. To address this issue, Gómez, Bion, and Mehler (2011) used a click detection task to study statistical learning online. This procedure was inspired by an experimental paradigm developed in the '60s to study online syntactic processing (Bever, Lackner, & Stolz, 1969; Fodor & Bever, 1965; Fodor, Bever, & Garrett, 1974). Fodor and Bever's original click location task consisted in the presentation of auditory sentences in which clicks were inserted at specific spots. Participants' instructions were to locate the clicks by pressing a key as fast as possible. Back then, the results showed that participants made more errors when the clicks occurred at the boundaries between clauses. When the click co-occurred with the first word of a highly redundant two-word sequence—that is, a sequence with a high TP—participants subjectively perceived that the click occurred later, in the middle of the sequence. When the click co-occurred with the second word of a low-TP two-word sequence, participants subjectively perceived that the click occurred after the sequence. The authors concluded that the TPs between words modulate click localization.

On the basis of Fodor and collaborators' findings, Gómez et al. (2011) proposed that the TPs between syllables in continuous speech could likewise modulate the detection of clicks located between or within the words. They presented adult participants with a continuous speech stream consisting of the randomized repetition of four trisyllabic nonsense words. The speech stream was produced by a speech synthesizer, so that no other segmentation cues but the TPs were present. Clicks were superimposed on the stream, and could occur at two different positions: either at the boundaries of two words (between words) or between the first and second syllables of a word (within words). Participants were instructed to listen to the speech stream and to press a key as fast as possible each time they detected a click. After 2 min of exposure, participants were faster in detecting the clicks located between than within words. The authors proposed that the evolution of reaction time (RT) differences between the two types of click locations reflects the extraction of TPs and the emergence of word candidates. In their view, participants progressively built different expectations about future events when processing the stream of syllables. As a matter of fact, the stream was built in such a way that the TP between two syllables was 33 % across word boundaries and 100 % within words. The rationale was that during learning, participants form representations of word candidates and develop stronger expectations about the next syllable when it is part of a within- rather than a between-word transition. These expectations would in turn modulate participants' tolerances for disruption in the syllables stream, making them less likely to integrate extraneous elements within a word, so that clicks would tend to be perceived at word boundaries rather than within words.

Gómez et al. (2011) concluded that the click detection task could provide an online measure of word segmentation based on statistical learning. However, just like offline measures of statistical learning such as 2AFC tasks, their results could also be explained by an emerging sensitivity to the regularities that was not necessarily accompanied by the extraction and memorization of the words. In other words (no pun intended), although click detection makes it possible to explore how we learn, it does not provide a measure of what, or how much, we learn. Gómez and colleagues themselves acknowledged that the click detection method and a classical offline measure do not necessarily correlate. Actually, since their study did not include such an offline, direct test of word knowledge, the question remains open as to whether participants correctly segmented the words. It might be the case that participants only focused on detecting the clicks, and did not attend (or only partially attended) to the speech stream. In that case, participants would have promoted one of the two concurrent tasks, instead of equally sharing their attentional resources between the tasks. As a consequence, in dual-task situations such as the click detection task, performance should be monitored in both tasks, in order to ensure a full understanding of the mechanisms that subtend performance. A similar critique has been raised to Saffran and colleagues' (1997) study, reporting successful statistical learning when participants were exposed to a speech stream during a simple concurrent task (free drawing). Toro, Sinnett, and Soto-Faraco (2005) argued that there was "no actual guarantee that participants did not occasionally direct their attention to the irrelevant speech stream while they were performing the free drawing task" (p. B26). Indeed, there is evidence that dividing attention during statistical learning leads to poor performance in offline tests (Turk-Browne et al., 2005). There is even an additional cost when the stream of information in the statistical-learning task and the concurrent task share the same modality (Toro et al., 2005).

In summary, it is a safe bet that online click detection will have a negative impact on performance at the offline task, because the click detection can be considered a secondary task recruiting the same sensory modality as the primary learning material. The aim of our study was to investigate the impact of a click detection task on statistical learning. In Experiment 1, we replicated the method used by Gómez et al. (2011) and added an offline measure in order to evaluate the link between these two measures. In Experiment 2, we measured the impact of the click detection task by isolating two factors: the mere

presence of clicks in the continuous speech stream and the need for participants to process those clicks or not. Finally, in Experiment 3, we examined the impact of the clicks' location on statistical learning. Overall, our results showed that the clicks act as extraneous auditory elements interfering with word segmentation and should, therefore, be used with caution.

## Experiment 1

### Method

*Participants* Twenty-eight French-speaking undergraduate psychology students (18 women, 10 men) were included in this study and received course credits for their participation. None of the participants had previous experience with the artificial languages presented in this experiment. All reported no hearing problems.

*Material* Two artificial speech streams were generated using the MBROLA speech synthesiser (Dutoit, 1997) with the French male diphone database fr1 with a sampling frequency of 16 kHz. The streams consisted of the continuous presentation of four nonsense trisyllabic words (*bamoti*, *bikochu*, *lumake*, *telicha*) without pauses. Each syllable lasted 200 ms. Each word was presented 90 to 100 times. Each speech stream lasted for 4 min 10 s, with the words being presented in a pseudorandom order: The same word never occurred twice in succession. Thus, both speech streams were identical and only differed in word order presentation. A set of clicks was inserted into the speech stream with the software Praat (Boersma, 2001). We used the same procedure that was presented in the Gómez et al. (2011) study: The clicks corresponded to five consecutive samples of the audio waveform clipped together. Each click could occur either between two words or within words, between the first and second syllables (within1_2). Figure 1 illustrates the placement of the clicks in the speech stream. Each minute, eight between- and eight within-word clicks were inserted, resulting in a total of 64 clicks, with an average

interval of 3.8 s between any two consecutive clicks. During the 4-min exposure, the 32 within-word clicks were equally distributed between the four words (eight clicks per word), and the 32 between-word clicks were evenly distributed, in order to control the probability of occurrence of a click across the different between-word transitions. The probability of one syllable being preceded by a click was 8.4 % for both the first and the second syllable of each word (between- and within-word clicks, respectively), and 0 % for the third syllable of each word. The click positions were similarly distributed in the two different versions of the speech stream. Participants were randomly assigned to one of the two streams (14 per condition).

*Procedure* Participants were tested individually in a quiet room. They were instructed to pay attention to the speech stream spoken in an "unknown language," to try to extract the words from the speech, and to press a key as fast as possible each time they heard a click. The speech stream was presented binaurally through headphones. Immediately after the 4-min exposure, participants performed a 2AFC task in which they were presented with two trisyllabic sequences on each trial. One sequence was a word of the artificial language, and the other sequence was a nonword, made up of the same syllables but with null TPs between them (i.e., any two successive syllables had never been presented in succession in the exposure phase). Participants were instructed to decide which sequence of each pair sounded more like the unknown language that they had just heard. Four nonwords were used (*baluti*, *chubima*, *liteko*, and *mokecha*). Each word was paired with each nonword twice—either as the first or as the second element of the pair—resulting in a total of 32 trials. The experiment was run on a Mac Mini 1.33-GHz PowerPC G4 using Psyscope X B53 (Cohen, MacWhinney, Flatt, & Provost, 1993) and a Psyscope USB button box to record the RTs of each click response.

### Results and discussion

*Time course of RTs for both types of clicks* As in the original study (Gómez et al., 2011), nonresponded clicks and RTs



**Fig. 1** Example waveform from one of the artificial speech streams. The arrows indicate a click located within a word (bi!kochu) and a click located between two words (bamoti!telicha)

longer than 1,000 ms or shorter than 100 ms were excluded from all of the analyses. These criteria resulted in excluding 3.8 % of the trials from the analysis (on average, 2.5 clicks out of 64). Among these, 0.7 % were missed clicks, 0.1 % were RTs shorter than 100 ms, and 2.9 % were RTs longer than 1, 000 ms. The overall mean RT was 318 ms ($SD$ = 48). We computed the mean RTs by minute (one mean RT by minute) and by location (within or between words) for each participant, resulting in a total of eight mean RTs for each participant. A repeated measures analysis of variance (ANOVA) with Minute (four levels) and Location (two levels) as within-subjects factors showed a significant effect of minute, $F(3, 81) = 6.466$, $p = .001$, $\eta_{\mathrm{p}}^2 = .193$. However, we failed to find a significant effect of location, $F(1, 27) = 2.009$, $p > .1$. The Minute × Location interaction was also nonsignificant, $F < 1$ (Fig. 2a).

In Gómez et al. (2011), 24 out of the 28 participants showed a positive difference between within- and between-

word RTs as a function of time. Four of them showed the reverse interaction: They responded faster to clicks emitted within than between words. One possible explanation for our results could be that in our study, a larger number of participants showed the reverse interaction, thus resulting in null differences when considering all participants together. To ascertain whether this was the case, we computed, for each participant, the mean RT difference between within- and between-word clicks at each minute. Next, we computed the number of participants for whom the mean difference was positive (i.e., RTs for within-word clicks exceeded those for between-word clicks). Following the analysis conducted by Gómez et al., we considered chance level to be 50 %—that is, 14 out of 28 participants. As is shown in Fig. 2b, at Minutes 1 and 2, 12 out of 28 participants showed a positive difference between mean RTs. At Minute 3, however, only six participants showed a positive difference. Finally, at Minute 4, ten showed a positive difference, and 18 participants showed a
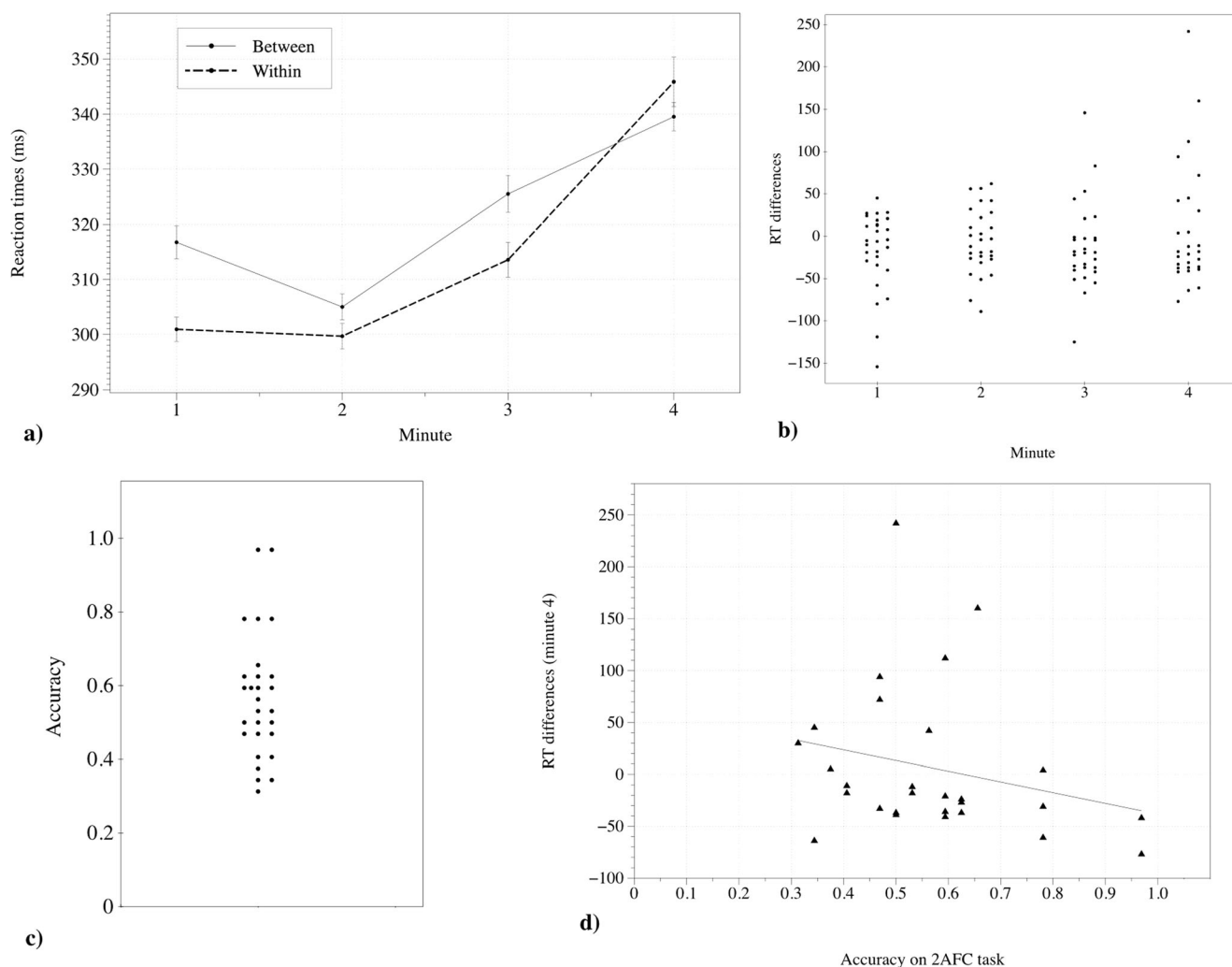


**Fig. 2** Results of Experiment 1. (**a**) Average RTs for both click locations, pooling RTs separately for each minute. (**b**) Mean differences between RTs to clicks within words and to clicks between words over exposure.

Each participant (out of 28) is represented by a dot. (**c**) Distribution of scores for the two-alternative forced choice (2AFC) task. (**d**) Scatterplots of RT differences on Minute 4 and accuracy on the 2AFC task

negative difference. Thus, in contrast with Gómez et al.'s results, participants' tendency to respond faster to between-word clicks when the word candidates emerged seems to have been less marked in our study.

To sum up, as in the original study, we observed a gradual increase of the mean RTs during exposure. With respect to the time course of RTs for the different click locations, however, our results differed from those of Gómez et al. (2011). Although Gómez et al. reported an increase of RTs after the second minute (mostly for clicks located within words), our results suggest that this click-dependent RT difference may not be systematically observed. Indeed, in their study, Gómez and collaborators found that 24 out of 28 participants showed this tendency, whereas four participants showed the reverse effect. In our study, we found a larger number of participants showing the opposite trend (18 out of 28).

*2AFC task* The overall forced choice task performance was 56.81 %, which was above chance level (50 %), $t(27) = 2.123$, $p < .05$, bilateral. Although on average the participants performed above chance, the mean performance was quite low, and the performance of half of the participants ($N = 14$) was at chance. The raw distribution of participants' scores is presented in Fig. 2c.

*Are the RT time course and the mean performance at the 2AFC task related?* The rationale of Gomez et al.'s (2011) study was that the extraction of the TPs present in the stream (high or low within or between words) would produce stronger expectations of the next syllable within words, making participants less likely to expect extraneous elements such as a click. This would be reflected by slower responses to clicks located within than between words. If this was the case, one should expect a positive between-word minus within-word RT difference at the fourth minute in the click detection task. Participants should also have been successful in the 2AFC task. By contrast, a null—or negative—difference should be associated with a low performance level in the offline test. We performed a regression analysis to test the hypothesis that there was a relationship between the sign of the RT difference and accuracy in the 2AFC task. As is shown in Fig. 2d, the RT difference at the fourth minute was not a significant predictor of accuracy in the 2AFC task, $\beta = -.242$, $p > .1$. We did not find a reliable association between RT and forced choice performance in this study.

Finally, we asked whether a specific pattern of RT evolution was associated with successful performance in the 2AFC task. In other words, did those participants who successfully completed the 2AFC task show a specific RT evolution in the click detection task? To address this question, we divided participants into two groups based on their performance in the 2AFC task. A binomial test showed that scores exceeding 22 word or nonword identifications would constitute a

statistically significant deviation from 50 % at a .05 alpha level. According to that criterion, 23 participants performed at chance (mean = 50.55 %), and only five of them performed above the chance level (mean = 85.62 %). We analyzed the RT patterns in the latter group. A Wilcoxon matched-pairs signed rank test showed that the RTs for between- and within-word clicks during the fourth minute (369 and 328, respectively) did not statistically differ from each other ($Z = -1.753$, $p = .080$). This result showed that even those participants who performed above chance level in the 2AFC task did not exhibit a positive RT difference, possibly reflecting word extraction from the speech stream.

Taken together, these results raise two additional questions: (1) Does the click detection task negatively influence correct word extraction, and, as a consequence, does it result in poor performance in the 2AFC task? and (2) Does the RT evolution in the click detection task only predict the success or failure of speech segmentation, or is it also "contaminated" by the click detection task? In the latter case, even if the RTs pattern reflects statistical learning, its interpretation becomes challenging. In Experiment 2, we examined the impact of the click detection task on the extraction of the words embedded in the speech stream by means of two distinct conditions: one in which participants were exposed to the exact same speech stream superimposed with clicks, as in Experiment 1, but were not instructed to respond to the clicks, and another in which they were exposed to the same speech stream without the clicks.
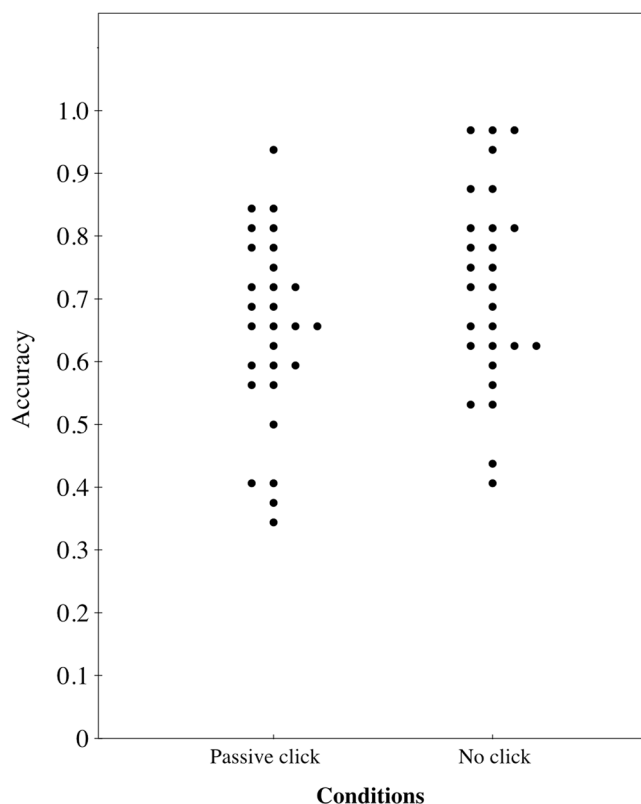
## Experiment 2

### Method

*Participants* Fifty-six French-speaking undergraduate psychology students (42 women, 14 men) were included in this study and received course credits for participation. None of the participants had previous experience with the artificial languages presented in this experiment. All reported no hearing problems.

*Material* Here we used four different speech streams. One was taken from Experiment 1 (Language A). A second speech stream (Language B) was created by systematically replacing each word by a nonword (*baluti*, *chubima*, *liteko*, and *mokecha*). This between-subjects counterbalanced design allowed us to rule out the possibility that any observed preference for a word or a nonword across both conditions was due to any general preference for certain syllable strings. Two other speech streams were created that consisted of the same two speech streams without the clicks. The stimuli used in the 2AFC task were identical to those in Experiment 1.

*Procedure* The procedure was identical to that in Experiment 1, except that participants were randomly assigned to one of the two languages (28 participants assigned to Language A or B) and one of two conditions (28 participants in each condition). In the "passive-click" condition, participants were instructed to pay attention to the speech stream spoken in an "unknown language," trying to extract the words from the speech. They were also told that they would occasionally hear a click, but that they did not have to react to it. In the "no-click" condition, participants were instructed to pay attention to the speech stream and to try to extract the words from the speech.

Results and discussion

Regarding the 2AFC task, no difference was found between the mean performances of participants who were exposed to Language A (69.19 %) or to Language B (67.85 %), $t(54) = 0.321$, $p > .5$, bilateral. Therefore both groups were pooled together in the following analysis. Mean performance in the 2AFC task was significantly above chance level (68.52 %), $t(55) = 8.956$, $p < .001$, bilateral. An independent-samples $t$ test revealed that the mean percentages of correct responses differed between the no-click (72.87 %) and passive-click (64.17 %) conditions, $t(54) = -2.174$, $p < .05$, bilateral (Fig. 3). We also compared these results to the mean



**Fig. 3** Distribution of scores from the two-alternative forced choice task for both the passive-click and no-click conditions in Experiment 2

performance obtained in Experiment 1. A univariate ANOVA with condition (Exp. 1, passive-click, and no-click condition) as a between-subjects variable showed a main effect of condition, $F(2, 81) = 7,367$, $p = .001$, $\eta_p^2 = .154$. The effect was due to a significant difference between the no-click condition and Experiment 1 (Bonferroni-corrected $p = .01$). This result suggests that the mere presence of clicks impairs performance in the 2AFC, but it is not clear whether an extra impairment follows from adding an additional detection task.

Taken together, these results suggest that the click detection task has a detrimental effect on performance in the 2AFC test. Considering that the 2AFC test constitutes a valid measure of statistical learning, these results are in line with the previous studies showing that a decrease in attentional resources impairs word segmentation based on the extraction of statistical regularities (Toro et al., 2005).

In addition, these results suggest that the process of segmentation is disrupted even when participants are not required to respond to the clicks. One possibility could be that the position of the clicks affects segmentation, as participants could use them as benchmarks to chunk the speech stream. In that case, clicks placed between the words would lead to successful segmentation of the artificial language, whereas those placed within the words would induce incorrect boundaries, and therefore result in poor performance in the 2AFC task.

To test this assumption, in a third experiment, we compared one condition in which all clicks were placed between words, and another in which they were all placed within words. If clicks are indeed used as anchors during segmentation, we should observe better performance in the 2AFC task in the "Between" than in the "Within" condition. If clicks act as mere interference with the segmentation process, no difference would be predicted in 2AFC performance between the two conditions. This would suggest that the poor performance observed in the "passive-click" condition cannot be explained by the use of clicks as a cue to word boundaries.

**Experiment 3**

Method

*Participants* Fifty-six French-speaking undergraduate psychology students (39 women, 17 men) were included in this study and received course credits for participation. None of the participants had previous experience with the artificial languages presented in this experiment. All reported no hearing problems. One participant was excluded from the data because he made too many errors in the click detection task.

*Material* As in Experiment 2, two streams were used: Language A and Language B. Two versions of each of these two

streams were created: the "Between" and the "Within" conditions. In the Between condition, 64 clicks were placed at word boundaries (16 clicks per minute equally distributed among the four words—four clicks preceding each of them). In the Within condition, clicks occurred between the first and second syllables of each word (16 clicks per minute equally distributed among the four words, with four clicks occurring within each of them). In both conditions, the probability that one syllable would be preceded by a click was 8.4 %. Two similar versions of these speech streams were used, in which the words in one stream corresponded to nonwords in the other one, and vice versa. This was done to counterbalance the syllable strings to which participants were exposed. The stimuli used in the 2AFC task were identical those in Experiment 1.

*Procedure* The procedure was identical to that in Experiment 1. Participants were randomly assigned to either Language A or Language B (28 participants in each condition) and to either the Between or the Within condition (28 participants in each condition).

Results and discussion

*Time course of reaction times* We excluded 4.9 % of the data from the analysis, corresponding to missed clicks and RT outliers (i.e., on average 3.1 clicks out of 64). A repeated measures ANOVA with Minute (four levels) as a within-subjects and Condition (two levels: inter- and intraword clicks) as a between-subjects factor was performed on the mean RTs. The analysis showed a main effect of minute, $F(3, 162) = 11.237$, $p < .001$, $\eta_p^2 = .172$. As is indicated in Fig. 4, the mean RTs increased significantly from Minute 2 to Minute 3 for both the Between and Within conditions. The Minute × Condition interaction was not significant, $F < 1$.

However, a main effect of condition was found, $F(1, 54) = 6.494$, $p < .05$, $\eta_p^2 = .107$, indicating that mean RTs were faster in the Between condition (338 ms) than in the Within condition (371 ms).

We also compared the evolution of RTs in Experiment 1 with that in the Between and Within conditions of Experiment 3. Two repeated measures ANOVAs were conducted: the first to compare RTs in the between-word condition of Experiment 1 and the Between condition of Experiment 3, and the second on RTs in the within-word condition of Experiment 1 and the Within condition of Experiment 3.

The comparison of between-word click detection in the two experiments revealed a main effect of minute, $F(3, 162) = 8.860$, $p < .001$, $\eta_p^2 = .141$. The effect of experiment (Exp. 1 vs. Between condition) was not significant ($p > .1$). The Minute × Experiment interaction was also not significant, $F < 1$, $p > .5$.

The comparison of within-word click detection in the two experiments showed a main effect of minute, $F(3, 162) = 12.278$, $p = .001$, $\eta_p^2 = .185$. The Minute × Experiment interaction was not significant, $F < 1$. However, a significant main effect of experiment was found, $F(1, 54) = 14.169$, $p < .001$, $\eta_p^2 = .208$: The RTs in Experiment 1 were significantly faster than those in the Within condition of Experiment 3.

*2AFC* The mean performance of participants who were exposed to Language A (58.25 %) was compared to the performance of those who were exposed to Language B (62.94 %). No difference between the two languages was found, $t(54) = -1.107$, $p > .1$, bilateral. Therefore, the two groups were pooled in the following analysis. The mean performance in the 2AFC task was above chance (60.60 %), $t(55) = 4.998$, $p < .001$, bilateral, and was not significantly influenced by the condition [Between (58.48 %) vs. Within (62.72 %); $t(54) = 1.000$, $p = .322$, bilateral]. In order to statistically exclude the hypothesis that clicks placed between words would constitute an aid to
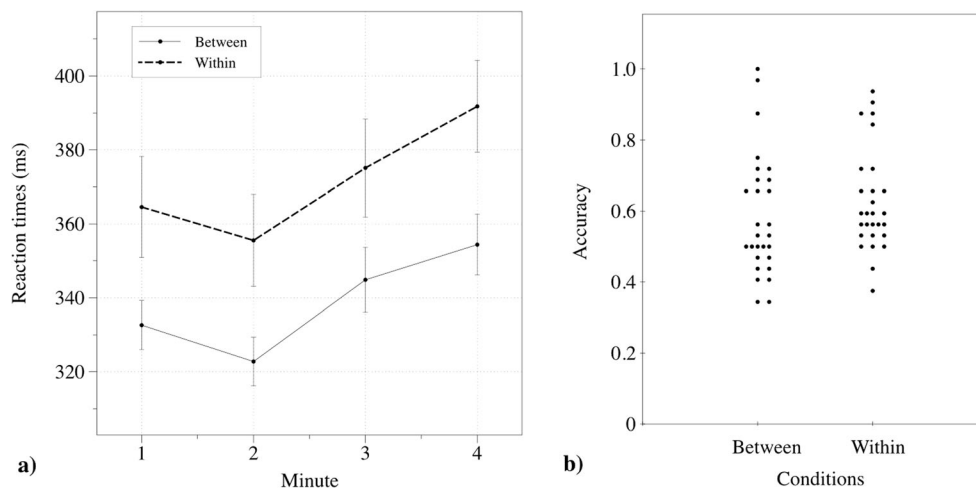


**Fig. 4** Results of Experiment 3. (**a**) Average RTs for both conditions (between- and within-word clicks), pooling RTs separately for each minute in Experiment 3. (**b**) Distribution of scores for the two-alternative forced choice task

speech segmentation, we used a Bayes factor analysis (BF). This analysis allows one to make sure that an observed null difference truly indicates an absence of difference between two conditions by establishing the strength of the evidence for a theory predicting an effect over the null hypothesis, or vice versa (Dienes, 2008, 2011). The BF varies from 0 to infinity. Any value less than a 1/3 is strong evidence for the null hypothesis over the theory, over 3 is strong evidence for the theory over the null hypothesis, and a value between 1/3 and 3 indicates data insensitivity. Thus, a BF can provide what $p$ values cannot, in distinguishing between evidence for the null (BF < 1/3) and no evidence either way because the data are insensitive (1/3 < BF < 3).

Following Dienes (2011), we modeled the predictions of a difference as a half-normal with a standard deviation equal to 0.5. For a mean difference between the Between and Within conditions of 0.042 and a standard error of 0.021, the BF (using the free online calculator from the website for Dienes, 2008) was 0.32—that is, evidence for the null hypothesis over the theory that there was a difference between the Between and Within conditions. This is positive evidence for the null hypothesis, not just insensitive data. This result allowed us to exclude the hypothesis that clicks placed between words constituted an aid to speech segmentation.

Finally, if we consider all conditions together, a univariate ANOVA with condition (Exp. 1, passive click, no click, Between, and Within) as a between-subjects variable showed a reliable main effect, $F(4, 135) = 4.447$, $p < .005$, $\eta_p^2 = .116$. The effect was due to the significant difference between the no-click condition and Experiment 1, on the one hand, and the Between condition on the other hand (Bonferroni-corrected $p$s = .02 and .008, respectively).

The purpose of this experiment was to test the hypothesis that auditory clicks could guide the segmentation process of the artificial language. Because in the original experiment, half of the clicks experienced by each participant were located between and half within words, clicks may have cued the participants to both correct and incorrect word boundaries. Speech segmentation should then be improved in a situation in which all clicks were located at word boundaries, as compared to a situation in which all of them were located within words. In the Between condition, clicks should prompt the correct chunking, whereas they should induce incorrect segmentation in the Within condition. Performance in the 2AFC task should therefore be better in the former than in the latter condition.

Another possibility is that click detection acts as a secondary task and disrupts statistical learning. The results of Experiment 3 support this second alternative. Indeed, we observed that the location of the clicks had no impact on performance in the 2AFC task. Moreover, performance was lower than in the no-click condition of Experiment 2, for both the Between and Within conditions. In addition, the patterns of RTs were similar in both conditions. More importantly, both conditions

exhibited evolutions of RTs similar to those in Experiment 1, in which both types of clicks were presented to each participant.

## General discussion

Nearly 20 years of research on statistical learning have clearly indicated that humans at all stages of development, nonhuman primates (see Conway & Christiansen, 2001, for a review), and even rats (Toro & Trobalón, 2005) are able to process statistical information. Research has focused essentially on what and how much we are able to learn, but little is known about the temporal trajectory of statistical learning. Gómez et al. (2011) proposed a method to explore this issue: the click detection task. In order to test their method and the extent to which it can be used as an online measure of statistical learning, we conducted three experiments in which we used the click detection task combined with a frequently used 2AFC offline task. The underlying idea was that the evolution—and more importantly, the emergence—of a difference in RTs between clicks located within and between words would indicate a sensitivity to the TPs, providing that a specific pattern of RTs can reflect successful statistical learning. Our results showed that (1) the within–between RT differences were not as systematic as had been found in the original study, (2) the RT differences in the click detection task were not related to performance on the 2AFC task, and (3) the click detection task seemed to have a negative impact on the offline task. In the following sections, we discuss the implications of each of these points.

### Within–between RT differences were not as systematic as in the original study

In the original study, 24 out of 28 participants showed a positive difference between the two types of click locations at the fourth minute of exposure. Gómez and collaborators (2011) claimed that this difference reflects the extraction of TPs. In other words, it reflects the participants' expectations regarding the sequence of syllables. In our study, only 10 out of 28 participants showed such a pattern of results, and the remaining participants showed the opposite pattern by the end of training. This is problematic, insofar as the online measure of statistical learning consists in the progressive increase of a positive RT difference between clicks located between and those located within words. The discrepancy between ours and the results reported by Gómez et al. could be due to a smaller proportion of participants who successfully extracted the words in our study, but neither Gómez et al.'s nor our own results can confirm this hypothesis.

RT differences were not linked to performance in an offline task

Experiment 1 revealed that the sign—or the size—of the RT difference did not predict performance in the 2AFC task, suggesting that successful word segmentation (as evidenced by the 2AFC task) is not necessarily predicted by slower RTs for clicks located within rather than between words. Several, not mutually exclusive, reasons could explain this absence of a relationship between RTs and 2AFC performance. First, it is very likely that the two tasks do not measure the exact same knowledge base. Although the 2AFC task is sensitive to knowledge about the words of the artificial language, the click detection task could provide a more subtle measure of the development of the sensitivity to the TPs. In that case, a given participant could be sensitive enough to the TPs to show a specific evolution of RT differences, but not to explicitly recognize the word-like units, leading to a failure in the direct task. However, in that case, we should have seen an RT difference on the fourth minute for at least those participants who performed above chance in the offline task. Only a few participants reached that criterion, and they did not show a specific pattern of RT differences.

Another possibility is that the click detection task impairs performance in the 2AFC task. As was already mentioned by Gómez and collaborators (2011) in their study, as a demanding concurrent task, click detection could interfere with the segmentation process and, therefore, impair performance in a subsequent offline task. This hypothesis, however, challenges the idea that statistical learning occurs automatically (Fiser & Aslin, 2001, 2002; Saffran, Aslin, & Newport, 1996; Turk-Browne et al., 2005). We discuss this issue in the next section.

The click detection task had a detrimental effect on 2AFC performance

In Experiment 2, we compared two conditions: one in which participants were exposed to the exact same speech stream superimposed with the clicks, but were not instructed to respond to the clicks, and another condition in which participants were exposed to the same speech stream without clicks. The comparison between the no-click condition and Experiment 1 clearly showed that the click detection task had a deleterious effect on performance in the 2AFC task. One possibility could be that the position of the clicks affected segmentation. Indeed, participants could use them as benchmarks to chunk the speech stream. Experiment 3 did not support that hypothesis: Participants who were exposed only to a speech stream with between-word clicks obtained 2AFC scores similar to those who were exposed only to a speech stream with within-word clicks. Another possibility could be that participants adopted different strategies during the exposure task. In fact, since they were instructed to react to the

clicks and to identify the words at the same time, participants may have focused on the first task, on the second, or on both successively. In this case, however, participants who focused more on the click detection task should have respond faster overall in that task than did those who focused more on word identification. Their performance should also have been worse in the 2AFC task. The results of Experiment 3 do not support this hypothesis: Participants who were exposed to the stream with within-word clicks only were faster and did not perform worse in the offline test than did those who were exposed to the stream with between-word clicks, or than the participants from Experiment 1.

Gómez et al. (2011) claimed that RTs reflect sensitivity to statistical computations. Our results support an alternative hypothesis: The click detection task impairs statistical learning. Click detection could actually be used as a valuable tool for online statistical learning if and only if its detrimental impact can be fully measured. Toward this aim, two conditions have to be met. First, in order to use the RT evolution as an online measure of statistical learning, a specific pattern of RTs reflecting statistical learning would have to be clearly identified. The present study shows that different patterns of RTs can be observed in this task, and that none of them can be unequivocally associated with another independent measure of knowledge. This could be either because of the negative impact of the click detection task on the 2AFC performance, or because of the limitations of the latter task. In this case, using another offline measure could offer a solution. Second, the influences of the mere presence of the clicks and of the detection instructions have to be disentangled. Indeed, whereas our results clearly show a negative impact of the click detection task on the offline measure, it is not clear whether the presence of the clicks within the speech stream or the additional detection instructions is what provides an additional impairment. Performance in the passive-click condition did not differ from that in either the no-click condition or Experiment 1. There is no obvious interpretation of the results found in the no-click condition. Considering that participants were informed about the presence of the clicks in the stream but were asked to ignore them at the same time, it remains possible that that these somewhat contradictory instructions created an attention suppression situation that could affect learning. If this were the case, however, one would have expected lower performance in the 2AFC task in the passive-click condition than in the standard condition of Experiment 1, in which participants were explicitly instructed to react to the clicks. However, performance in the 2AFC task did not differ significantly between these two conditions.

Do our results suggest that statistical learning is not as robust as generally thought? Indeed, the deleterious impact of the click detection task on statistical learning raises questions regarding its role in naturalistic settings. The question involves several levels of complexity, since word

segmentation in natural language is based on lexical, sublexical, phonetic, phonotactic, and prosodic cues (Mattys, Jusczyk, Luce, & Morgan, 1999). Segmentation cues are hierarchically integrated: Research on natural speech has indicated that lower-level, signal-contingent cues will be more prone to influence segmentation when there is a lack of contextual and lexical information or in the presence of white noise (see Mattys, White, & Melhorn, 2005, for adult participants and Morgan & Saffran, 1995, for infants). The click detection task might be akin to such ambiguous situations. Understanding whether the presence and/or the detection of the clicks impairs performance in the 2AFC task is therefore an important issue for future research.

Finally, another question concerns the general increase of RTs to clicks over exposure. Indeed, in both Experiments 1 and 3, an overall progressive increase of RTs was observed. This result is somewhat surprising, since in sequential-learning paradigms such as the serial reaction time task the reverse pattern of results is systematically observed (Cleeremans & McClelland, 1991; Destrebecqz & Cleeremans, 2001; Nissen & Bullemer, 1987; Perruchet & Amorim, 1992; Perruchet, Gallego, & Savy, 1990; Reber & Squire, 1994). In this paradigm, a stimulus appears at one of several locations on a computer screen, and participants are asked to press on the corresponding key as quickly and accurately as possible. Unknown to them, the sequence of successive stimuli follows a repetitive pattern. Typically, RTs tend to decrease progressively during practice and learning of the sequence. More importantly, RTs decrease even when participants are exposed to random sequences (Frensch & Miner, 1994). Therefore, since the click detection task also requires motor responses from the participants, one would expect to observe a gradual decrease or, at least, no significant increase in RTs.

What could explain the progressive global increase of RTs? We can reasonably exclude fatigue effects, since the task lasted less than 5 min. Serial reaction time tasks are indeed usually longer than the click detection task, and an RT increase is never reported. An alternative interpretation would be that the emergence of word candidates interferes with the click detection task, as was proposed by Gómez et al. (2011). Our study, however, indicates that clicks occurring between words are not systematically detected faster, as would be predicted if click detection were an indirect measure of learning. Our results, therefore, suggest that clicks are not used as a clue to segment the speech stream. The overall increase in RTs in click detection remains, therefore, an open issue for further research.

To conclude, we believe that the click detection task is a promising way to assess the time course of statistical learning, provided that the exact influence of the task on learning processes is clearly determined. Indeed, this online measure could bring new insights to the debate regarding the nature of the representations involved in statistical learning. According to one perspective, statistical learning is based on the development of associations between the temporal contexts in which the successive elements occur and their possible successors (as in the simple recurrent network connectionist model [SRN]; Cleeremans, 1993; Cleeremans, & McClelland, 1991; Elman, 1990). Over training, the network learns to make the best prediction of the next target in a given context. Its predictions are based on learning the TPs between sequence elements. Another perspective is based on the notion that statistical learning is an attention-based chunking process that results in the formation of distinctive, unitary, rigid representations (as in the PARSER model; Perruchet & Vinter, 1998). In contrast with the SRN, PARSER finds and stores the most frequent sequences in memory files or mental lexicon. The method proposed by Gómez and collaborators (2011) seems to be a good candidate to explore this question, in that it makes it possible to measure learning during the exposure phase, as participants become sensitive to the statistical regularities. However, our results challenge the idea that the emergence of RT differences as a function of click locations reflects statistical learning and the emergence of word candidates. In our study, the emergence of RT differences was neither systematic nor related to success or failure in the 2AFC task. Our results, rather, suggest that the click detection task exerts a deleterious effect on speech segmentation. Understanding what are the mechanisms involved in the learning process and how click detection interacts with these mechanisms and with the attentional processes involved in statistical learning would allow us to better understand RT patterns and to improve the use of click detection as an online measure of statistical learning. In the absence of a more detailed model of the statistical-learning mechanisms, click detection should be used with caution.

# References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience, 20,* 952–964. doi:10.1162/jocn.2008.20058

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324. doi:10.1111/1467-9280.00063

Bever, T. G., Lackner, J. R., & Stolz, W. (1969). Transitional probability is not a general mechanism for the segmentation of speech. *Journal of Experimental Psychology, 79,* 387–394.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. Cambridge, MA: MIT Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120,* 235–253. doi:10.1037/0096-3445.120.3.235

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, & Computers, 25,* 257–271. doi:10.3758/BF03204507

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences, 5,* 529–546.

Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review, 8,* 343–350. doi:10.3758/BF03196171

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York, NY: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Dordrecht, The Netherlands: Kluwer.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

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

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 458–467. doi:10.1037/0278-7393.28.3.458

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior, 4,* 414–420.

Fodor, J., Bever, T. G., & Garrett, M. (1974). *The psychology of language*. New York, NY: McGraw-Hill.

Frensch, P. A., & Miner, C. S. (1994). Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory & Cognition, 22*(1), 95–110.

Gerken, L. A., & Aslin, R. N. (2005). Thirty years of research on infant speech perception: The legacy of Peter W. Jusczyk. *Language Learning and Development, 1,* 5–21.

Gómez, D. M., Bion, R. A. H., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes, 26,* 212–223. doi:10.1080/01690965.2010.482451

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The headturn preference procedure for testing auditory perception. *Infant Behavior and Development, 18,* 111–116.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134,* 477–500. doi:10.1037/0096-3445.134.4.477

Morgan, J. L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development, 17,* 389–403.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development, 66,* 911–936.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirement of learning: Evidence from performance measures. *Cognitive Psychology, 19,* 1–32. doi:10.1016/0010-0285(87)90002-8

Perruchet, P., & Amorim, M.-A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 785–800. doi:10.1037/0278-7393.18.4.785

Perruchet, P., Gallego, J., & Savy, I. (1990). A critical reappraisal of the evidence for unconscious abstraction of deterministic rules in complex experimental situations. *Cognitive Psychology, 22,* 493–516.

Perruchet, P., & Vinter, A. (1998). PArser: A model for word segmentation. *Journal of Memory and Language, 39,* 246–263.

Reber, P. J., & Squire, L. R. (1994). Parallel brain systems for learning with and without awareness. *Learning & Memory, 1,* 217–229. doi:10.1101/lm.1.4.217

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 906–914. doi:10.1002/wcs.78

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month old infants. *Science, 274,* 1926–1928. doi:10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your hear. *Psychological Science, 8,* 101–105.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606–621. doi:10.1006/jmla.1996.0032

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97,* B25–B34.

Toro, J. M., & Trobalón, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics, 67,* 867–875. doi:10.3758/BF03193539

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experiment Psychology: General, 134,* 552–564. doi:10.1037/0096-3445.134.4.552