openjournals.nl

Dutch Journal of Applied Linguistics
https://www.dujal.nl | E-ISSN 2211-7253

Measuring (online) word segmentation in adults and children

Iris Broedelet[1], Paul Boersma[1] and Judith Rispens[1]
[1] *University of Amsterdam*

Abstract Since Saffran, Aslin and Newport (1996) showed that infants were sensitive to transitional probabilities between syllables after being exposed to a few minutes of fluent speech, there has been ample research on statistical learning. Word segmentation studies usually test learning by making use of "offline methods" such as forced-choice tasks. However, cognitive factors besides statistical learning possibly influence performance on those tasks. The goal of the present study was to improve a method for measuring word segmentation online. Click sounds were added to the speech stream, both between words and within words. Stronger expectations for the next syllable within words as opposed to between words were expected to result in slower detection of clicks within words, revealing sensitivity to word boundaries. Unexpectedly, we did not find evidence for learning in multiple groups of adults and child participants. We discuss possible methodological factors that could have influenced our results.

Keywords psycholinguistics, statistical learning, word segmentation task, click detection task, online measure

Article history
Received: June 2, 2020
Accepted: April 21, 2021
Online: October 26, 2021

Corresponding author
Iris Broedelet, I.R.L.Broedelet@uva.nl

Author contributions
Iris Broedelet, conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, visualization, project administration, funding acquisition; Paul Boersma, conceptualization, methodology, formal analysis, writing – review and editing, supervision; Judith Rispens, conceptualization, methodology, writing – review and editing, supervision

Funding information
This work was supported by NWO (Netherlands organization for scientific research) under Grant 322-89-014.

Conflicting interests
No potential conflict of interest was reported by the authors.

Supporting information
Data is available on FigShare: https://figshare.com/collections/_/4739162.

1 Introduction

Language is full of patterns and regularities. In the last decades, there has been great interest in the role of *statistical learning* in language acquisition. Statistical learning is a cognitive ability that underlies the implicit discovery of statistical patterns and sequences in sensory input (Siegelman et al., 2017) and has been hypothesized to contribute to different areas of language acquisition (for a review, see Romberg & Saffran, 2010). One of the first demonstrations of statistical learning was the seminal study of Saffran, Aslin and Newport (1996). As word boundaries are not (consistently) marked by pauses or other

How to cite Broedelet, I., Boersma, P. and Rispens, J. (2021). 'Measuring (online) word segmentation in adults and children', *Dutch Journal of Applied Linguistics* 10: DOI 10.51751/dujal9607

prosodic cues in natural speech (Cole, 1980), the authors aimed to investigate whether statistical learning plays a role in learning to recognize separate words in a stream of speech sounds. Eight-month-old infants with English-speaking parents were exposed to a two-minute synthesized stream of uninterrupted syllables. The speech stream consisted of four pseudo-words (*bidaku*, *padoti*, *golabu* and *tupiro*) that were repeated in a random order. The authors wanted to test whether infants were able to recognize these pseudo-words after exposure to the stream, despite the absence of any prosodic cues for word boundaries. Results from a head-turn preference procedure administered after familiarization show that infants listen longer to "part-words" that span word boundaries, such as *ku-pado*, than to target words. This novelty preference indicates that infants learn to recognize target words and are thus sensitive to the statistical probabilities of the input: the transitional probabilities (TPs) between syllables. For example, the probability that *da* followed *bi* in the stream was 1.0, while the probability that *pa* followed *ku* was only 0.333. As there were no pauses or other prosodic cues for word boundaries,[4] infants' learning could only have happened based on these TP values.

The degree of learning in statistical learning tasks is usually inferred from participants' performance on an "offline" task which they undergo after the familiarization phase, during which they have to choose between target words and foils. However, performance on such tasks could be strongly influenced by cognitive processes other than statistical learning, such as encoding and memory capacities, meta-linguistic skills and decision-making biases (Siegelman et al., 2017). Specifically for children, meta-linguistic questions such as "which word sounds better?" are difficult to process and answer, which could lead to underestimation of their (implicit) knowledge. Importantly, while statistical learning is a continuous process, offline measures provide information about behavior at only a single point in time. Online methods, on the other hand, can provide more insight into the trajectory of statistical learning by measuring learning throughout the familiarization phase. It has thus been argued that in future statistical learning studies, especially those focusing on children, it is important to develop sensitive online measures of statistical learning (Lammertink et al., 2019; Siegelman et al., 2017).

Based on the idea that reaction time reflects processing time, stimulus detection tasks have been used as online measures of sentence processing (e.g. Fodor & Bever, 1965; Foss & Lynch, 1969; Cohen & Mehler, 1996), and this paradigm has also been applied to word segmentation tasks. Gómez et al. (2011) added click sounds to the speech stream in their word segmentation task. Italian-speaking adults were asked to listen to a stream of speech sounds, consisting of four pseudo-words (*pabuda*, *gifoto*, *venola* and *minaro*) for four minutes and to push a button as fast as possible when they heard a click sound. Crucially, the clicks occurred either *between* two pseudo-words or *within* a pseudo-word (compare *pabuda*!*gifoto* to *pa*!*buda*, where ! indicates a click). The authors hypothesized that participants who had learned to recognize words in the stream should have stronger expectations for the next syllable when hearing the first syllable of a word compared to when they hear the final syllable of a word. This, in turn, should lead to a larger sur-

prise effect (and thus a slower reaction time) when detecting clicks occurring within words compared to clicks between words. Results showed that after two minutes of exposure, people are indeed slower when detecting clicks within words as opposed to clicks between words, indicating sensitivity to word boundaries that develops over time.

Franco et al. (2015) aimed to replicate these results and tested French-speaking adults on a similar task. As opposed to Gómez et al. (2011), the researchers did not find evidence for a difference in response times to clicks between words and clicks within words. Ten out of 28 participants showed the expected pattern while the other 18 showed the opposite pattern. In their second experiment, Franco et al. (2015) compared performance on two versions of the task: a "passive" word segmentation task with clicks to which participants did not have to respond ("passive-click"); and a word segmentation task without any clicks ("no-click"). They found that performance on the offline test phase of the passive-click version was significantly lower than performance on the no-click condition, indicating that the statistical learning process might have suffered from the addition of clicks to the stream. Hearing the clicks might have diverted attention from the syllable structure in the input, as participants might have focused more on detecting the clicks than on the artificial language. Another possibility is that the clicks might have given participants false cues to word boundaries, as the clicks were the only "prosodic" elements in the speech stream. The click detection paradigm has the potential to reveal the word segmentation process minute by minute,[5] but the finding of mixed results might indicate that an adaptation of the paradigm is called for.

2 The current study

Our aim was to find a method for measuring word segmentation online that would be suitable for adults as well as for children. As Gómez et al. (2011) and Franco et al. (2015) found mixed results, we decided to adapt the click detection by extending the familiarization phase to eight minutes to facilitate learning. The first and final two minutes contained only a few click sounds and were added to provide the participants with more "clean" input (without potential distraction from clicks) to facilitate learning of word boundaries. Based on previous studies, we hypothesized that participants could use statistical information to segment words from uninterrupted speech and that our adaptations to the task would result in a learning effect: slower reaction times for clicks within words compared to clicks between words. We constructed an artificial language based on the study of Haebig et al. (2017), as they tested a similar participant group as we intend to test for our future studies (school-aged typically developing children and children with developmental language disorder; DLD). We conducted three separate experiments. In our first experiment we tested online word segmentation using the click detection task. As we did not find evidence for learning on either the click detection nor the offline task, we conducted a second experiment in which we removed the click

sounds to test whether participants (adults and children) would show learning on the offline task. Finally, as we did not find evidence for offline learning in Experiment 2, we conducted Experiment 3 in which we used non-words (TP = 0) as foils instead of part-words (TP = 0.333), to test whether adults would learn to distinguish words from non-words.

3 Experiment 1

3.1 Methods and materials

3.1.1 *Participants*

Thirty-one adults (21 females and 10 males) participated in the study. Their ages varied between 19;8 (years; months) and 35;11 (M = 28;4, SD = 6;4). All participants were native speakers of Dutch and had been brought up monolingually. The participants reported that they did not have any hearing difficulty, serious visual problems, developmental dyslexia or any other language-based disorders, ADHD, ASD or learning difficulties. People who (had) studied linguistics or had taken courses in linguistics were excluded from participation. Ethical approval for the experiment was obtained from the Ethical Committee of the faculty of Humanities of the University of Amsterdam. All participants filled in an informed consent form.

3.1.2 *Stimuli and design*

3.1.2.1 Familiarization phase

We constructed a speech stream from recorded and modified speech. Two sets of four bisyllabic words were constructed to control for order effects: /kiba/, /moti/, /dalu/, /χido/ (language A) and /bamo/, /tida/, /luχi/, /doki/ (language B).[6] There was no significant difference in mean phonotactic frequency in Dutch[7] between the words of language A (M = 1.425, SD = 0.174) and the words of language B (M = 1.385, SD = 0.189): $t[3] = 0.738$, $p = 0.37$). All syllables were recorded by a female native speaker of Dutch in a soundproof room. To ensure natural co-articulation between all syllables in the stream, three-syllable sequences were recorded of which the middle syllable was used to construct the stream (see Table 1). For example, to construct part of the stream *lukiba*, we recorded *da**luki**, lu**kiba** and ki**bamo*** and used the middle syllables (see Graf Estes, 2012). All sound editing was done using the software Praat (Boersma & Weenink, 2019).

A unique 8-minute pseudo-random sequence of the four words was generated for each participant, with the restriction that a word could not occur twice in a row. Transitional probabilities between syllables were high within a word (TP = 1). For example, /ba/ always followed /ki/ in language A. Across word boundaries, transitional probabilities were lower, as for example /ba/ could be followed by either /mo/, /da/ or /χi/ (TP = 0.333) in language A. The stream was constructed such that there were no pauses or

Table 1 Three-syllable sequences that were recorded for language A and language B. The bolded letters represent the syllables that were used to construct the stream

Language A				**Language B**			
ki	ti**ki**ba	lu**ki**ba	do**ki**ba	*ba*	da**ba**mo	χi**ba**mo	ki**ba**mo
ba	ki**ba**mo	ki**ba**da	ki**ba**χi	*mo*	ba**mo**ti	ba**mo**lu	ba**mo**do
mo	ba**mo**ti	lu**mo**ti	χi**mo**ti	*ti*	mo**ti**da	χi**ti**da	ki**ti**da
ti	mo**ti**ki	mo**ti**da	mo**ti**χi	*da*	**da**ba	**da**lu	**da**do
da	ba**da**lu	ti**da**lu	do**da**lu	*lu*	mo**lu**χi	da**lu**χi	ki**lu**χi
lu	da**lu**ki	da**lu**mo	da**lu**χi	*χi*	lu**χi**ba	lu**χi**ti	lu**χi**do
χi	ba**χi**do	ti**χi**do	lu**χi**do	*do*	mo**do**ki	da**do**ki	**χi**do**ki**
do	xi**do**ki	**χi**do**mo**	**χi**do**da**	*ki*	do**ki**ba	do**ki**ti	do**ki**lu

Table 2 The structure of the familiarization phase of the word segmentation task

	Practice	**Part 1**	**Part 2**	**Part 3**
Duration of block	30 s	2 min	4 min	2 min
Total nr. of clicks	5–6	10	72	10
Clicks per minute	10	5	18	5
Percentage clicks	20%	4%	16%	4%

other prosodic cues for word boundaries: speech was monotone and all syllables were equally long (consonants 118 ms and vowels 160 ms). The syllable rate was 216 syllables (108 words) per minute, resulting in a total of 864 words per participant, with each of the four words occurring 216 times. The stream started with the second syllable of a word and ended with the first syllable of a word, so that the stream did not start or end with a word boundary.

High-pitched 20 ms click sounds (created in Praat) were inserted at random positions in the stream for each participant. There were always at least four syllables in between two clicks, to make sure participants had enough time to respond to every click. Importantly, half of the clicks occurred between two words (for example *kiba*!*dalu*) while the other half were placed within a word (for example *da*!*lu*). The clicks were 1.6 times louder compared to the speech sounds to facilitate the detection of the clicks. The first and final parts of the familiarization phase (2 minutes each) contained 10 clicks,[8] while the middle part (4 minutes) contained 72 clicks. The practice block (30 s), which was included to get the participants used to the click detection task, contained 5 to 6 clicks (see Table 2).

3.1.2.2 Offline test phase

The offline test phase consisted of 16 two-alternative forced-choice items, in which the targets were words (for example *kiba* for language A) and foils were part-words (syllable combinations spanning word boundaries, for example *bamo* for language A). The four targets were combined with each of the four foils to construct 16 test items. The targets of language A were used as foils for language B and vice versa. The test items were recorded in a citation form by the same female speaker who recorded the stimuli for the familiarization phase and were edited in the same way as the sounds used in the familiarization phase.[9] The order of the test items was randomized for each participant with the restriction that stimuli (either as target or foil) could not appear in two test items in a row.

3.1.3 *Procedure*

The experiment was carried out in a quiet room in the speech lab of the University of Amsterdam. The experiment was executed in E-Prime 2.0 software (Schneider et al., 2002). Participants sat behind a laptop computer screen wearing headphones and holding a response box. Test version was counterbalanced across participants. Pre-recorded child-directed[10] instructions told them to carefully listen to "a weird language", to press the button as fast as possible when they heard a click sound, and to pay attention as there would be questions at the end. Participants first practiced the click detection task for 30 seconds and proceeded on to the familiarization phase when confirmed they understood the task. As visual feedback, a hashtag (#) appeared on the screen when the button was pushed. In the test phase, participants heard two sequences for every test item, and were asked to choose which one sounded the most like the language they had just heard. There was one practice item. The numbers 1 and 2 appeared on the screen and the participants had to use the two corresponding buttons on the response box. It was possible to repeat test items once. All participants did another statistical learning task as well, the results of which are not discussed in this paper. Testing took approximately 30 minutes per subject and everyone was compensated with 5 euros for their participation.

3.1.4 *Analysis*

Data was analyzed using the free software R (R Core Team, 2020). For the offline measure, the practice test item was excluded from further analysis. To compute accuracy, test items were scored as correct when the participant chose the target word, and as incorrect when the participant chose the foil. For the online measure, only responses to the clicks from the second block were taken into account (72 clicks per participant). A response was considered valid when it occurred within 2 seconds after a click. Missed clicks and extraneous responses were removed from the data (1.64%). One participant was excluded from analysis due to too many missed clicks (38) and extraneous responses (27). This resulted in data suitable for analysis from 30 participants. As the RT data were not normally distributed (see Figure 1), they were normalized for further analysis to meet

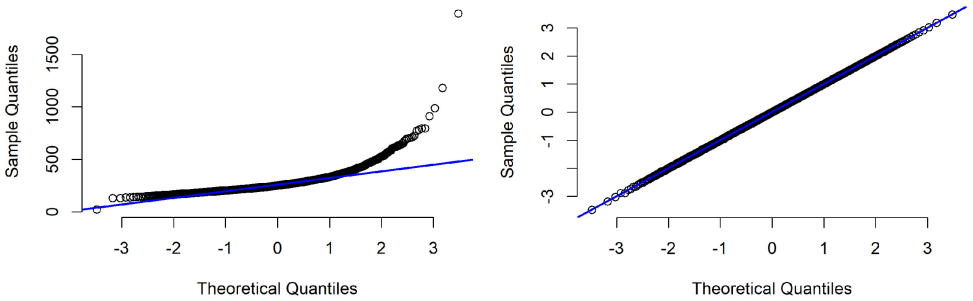

Figure 1 Distribution of the RT data before and after normalization

the normality assumption of mixed effect models: the response times were first ranked from 1 to N (where N is the total number of observations) and then normalized using $qnorm((rank - 0.5)/N)$ in R.

3.2 Results

3.2.1 *Offline test phase*

The average accuracy on the offline test phase was 0.45 (SD = 0.17). A generalized logistic linear mixed effects model (from the package *lme4*; Bates et al., 2015) was constructed to test whether participants performed above chance level (0.50). The dependent variable was Accuracy (a 1 or 0 value for every item). Between-participant predictors were Version (A/B) and TargetOrder (first/second; meaning whether the target was heard first or second during a particular test item). The different levels of the predictors were coded into sum-to-zero orthogonal contrasts (Kraemer & Blasey, 2004): Version was coded as $-\frac{1}{2}$ for A and $+\frac{1}{2}$ for B, and TargetOrder was coded as $-\frac{1}{2}$ for first and $+\frac{1}{2}$ for second. We implemented random intercepts by Participant and by Item, as well as by-participant random slopes for TargetOrder and by-item random slopes for Version.

The estimate for the intercept (converted into probability) was 0.42 (95% CI: 0.34 … 0.50). This performance is significantly *below* chance level (z = −2.016, p = 0.044), from which we might conclude that Dutch adults prefer part-words over words in the offline test phase of the current word segmentation task. This result is contrary to our expectations and, being one of our exploratory results, may be a chance finding. The effects of Version and TargetOrder on response times were not significant. See Figure 2 for the descriptive accuracy data and Table 3 for the results of the model.

3.2.2 *Click detection task*

A linear mixed effects model was conducted to test whether the position of the clicks (ClickPosition) influenced their processing time. The dependent variable was normalized RT. Within-participant predictors were ClickPosition (within words/between words) and Block (the middle part of the familiarization phase was divided in four blocks of 1

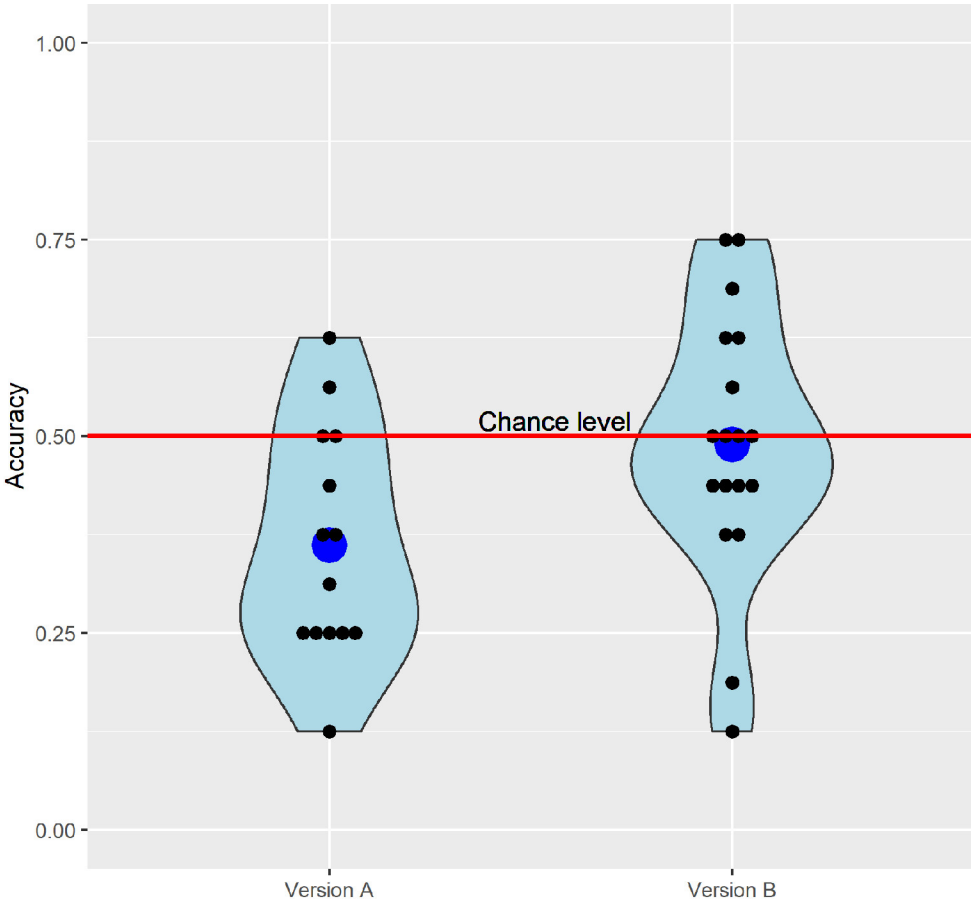

Figure 2 Descriptive plot of participants' accuracy in version A and B of Experiment 1

Table 3 Results from the linear mixed effect model

	Intercept	Version	TargetOrder
Estimate	Probability: 0.42	Odds: 1.72	Odds: 1.40
95 % CI	0.34 … 0.50	0.97 … 3.07	0.80 … 2.60
z	−2.016	1.94	1.013
p	0.044	0.052	0.311

Table 4 Descriptive data: raw and normalized response times for the click detection task

Raw	Block 1	Block 2	Block 3	Block 4	Overall
Overall RT	275 ms	280 ms	277 ms	290 ms	280 ms
RT between words	271 ms	276 ms	275 ms	293 ms	278 ms
RT within words	278 ms	285 ms	278 ms	288 ms	282 ms

Normalized	Block 1	Block 2	Block 3	Block 4	Overall
Overall RT	−0.0096	−0.0073	−0.0647	0.0817	0
RT between words	−0.0022	−0.5109	−0.0289	0.0916	0.0012
RT within words	−0.1681	0.0351	−0.1055	0.0726	−0.0012

minute, each containing 18 clicks). Version (A/B) was a between-participant predictor. We implemented random intercepts by Participant and by Item, as well as by-participant random slopes for ClickPosition and Block and by-item random slopes for Version. The different levels of the predictors ClickPosition and Version were coded into sum-to-zero orthogonal contrasts: ClickPosition was coded as $-\frac{1}{2}$ for between words and as $+\frac{1}{2}$ for within words, and Version was coded as $-\frac{1}{2}$ for A and $+\frac{1}{2}$ for B. The factor Block (1–4) was centered by subtracting 2.5 (the mean), resulting in the following numbers for the four blocks: –1.5, –0.5, +0.5 and +1.5. We expected that an online learning effect should surface as a main effect of ClickPosition and/or an interaction between ClickPosition and Block, the latter meaning that RTs are influenced by their context and that this difference is influenced by the amount of exposure to the stream of speech sounds. For the descriptive data, see Table 4 and Figure 3.

The main effect of ClickPosition (estimated Δz = 0.001, 95% CI −0.087 … +0.084) was not significant: t = −0.031, p = 0.98. Neither was the main effect of Block (estimated Δz = 0.024, 95% CI −0.025 … +0.072): t = 0.917, p = 0.35. The interaction between ClickPosition and Block (estimated $\Delta\Delta z$ = −0.04, 95% CI −0.11 … +0.03) also was not significant: t = −1.090, p = 0.26. On the basis of these results we cannot conclude whether the position of a click (between words or within a word) influenced their processing time, i.e. whether the click detection task revealed sensitivity to word boundaries in the word segmentation task.

There was a significant three-way interaction between ClickPosition, Block and Version (estimated $\Delta\Delta\Delta z$ = −0.16, 95% CI −0.30 … −0.18): t = −2.163, p = 0.036, indicating that the effect of ClickPosition is modified by Block and Version. This is illustrated in

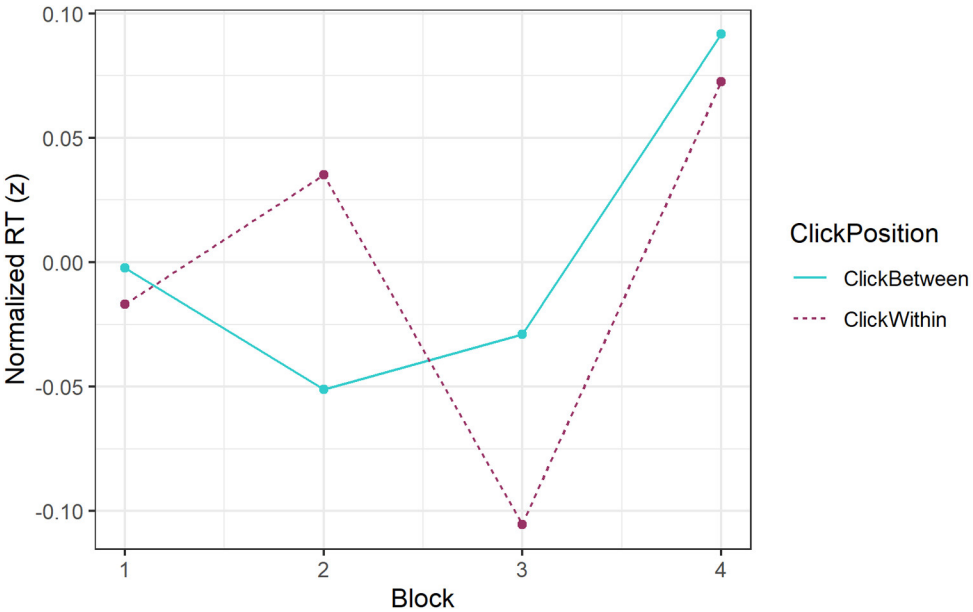

Figure 3 Normalized RT data (*z* scores) Experiment 1

Figure 4. For version A, the effect of ClickPosition developed as expected from Block 2 onwards and increased over time. For version B, however, the effect reversed in the third block. Individual data (Figure 5) shows that there was a large amount of variation in the effect of ClickPosition between participants. Some participants showed a difference in the expected direction, while others showed (almost) no difference or even a difference in the opposite direction.

3.3 Discussion

The aim of Experiment 1 was to adapt the click detection paradigm such that it would be a suitable method to measure word segmentation online. As we did not find evidence for or against an online learning effect, our results do not support the findings of Gómez et al. (2011). The extension of the familiarization phase does not seem to have been helpful for improving the click detection task as an online measure of statistical learning. Exploratorily, we found an unexpected difference between the two test versions: the RT's of the participants who did version A of the task showed the course that we expected, but participants who did version B showed a different pattern.[11] Moreover, similar to Franco et al. (2015), we observed a large amount of individual variation between participants. It could be the case that the words of version A are somehow "easier" to learn, but it might also be true that the click detection task as an online measure of learning works for some people (in the way we expect), but not for all.

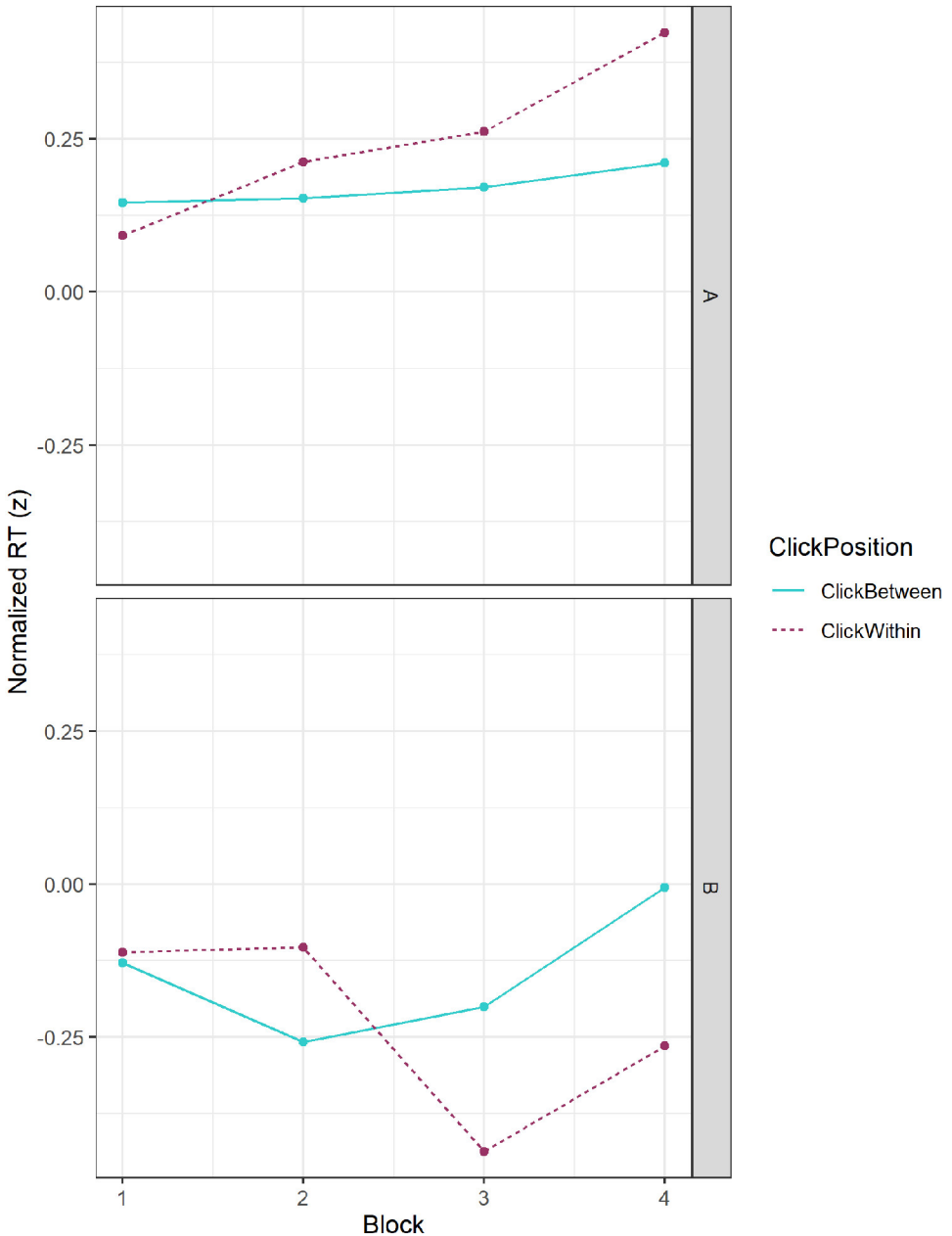

Figure 4 Normalized RT data Experiment 1: version A vs. version B

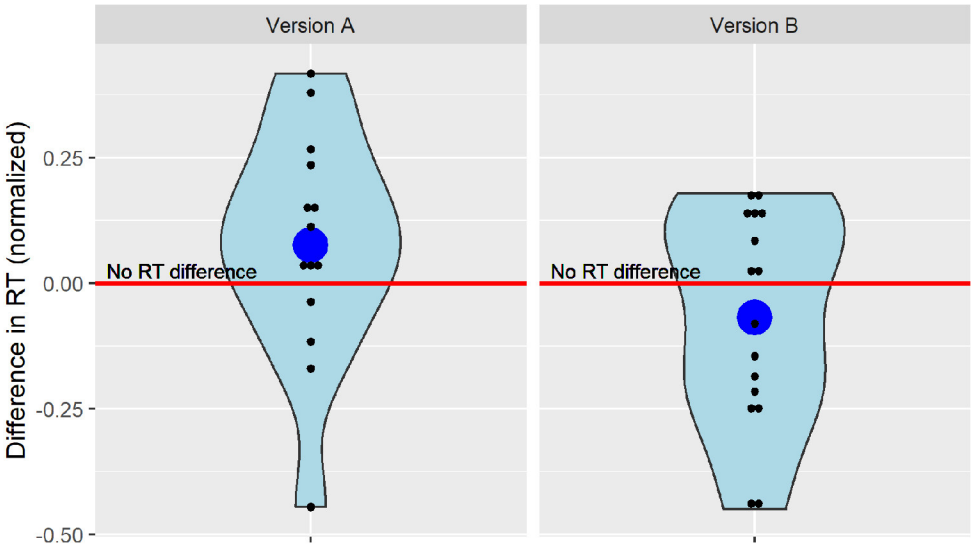

Figure 5 Individual RT data Experiment 1: Mean difference in RT for clicks between words and clicks within words. A positive value implies a learning effect

The fact that we did not find evidence for (or against) learning may reflect that click detection is not suitable as an online method, or that it actually negatively influences statistical learning. As Franco et al. (2015) found that performance on an offline test phase was better when participants listened to a stream without click sounds than when they listened to a stream with click sounds, the authors suggested that the addition of the click detection task, or even just the click sounds, might have hampered statistical learning. Our result of below-chance performance on the offline test phase could be a chance finding, but it could also be the case that the addition of an online measure negatively affected performance on the offline test phase (Toro et al., 2005). Another explanation might be that the click sounds gave participants false cues for word boundaries. As we cannot draw any conclusions on the basis of only this experiment, we conducted another experiment in which we tested two new groups of participants on the same word segmentation task without the addition of the click detection task. As our intended participant group for future studies on (online) word segmentation are school-aged children with and without developmental language disorder (DLD), we included a group of adults and a group of school-aged children in our second experiment.

4 Experiment 2

4.1 Methods and materials

4.1.1 *Participants*

Thirty adults (22 female, 8 male) participated in the study. Their ages varied between 18;0 and 26;10 (M = 20;6, SD = 2;4). Moreover, 30 children (20 girls, 10 boys) between the ages of 7;10 and 10;0 (M = 8;5, SD = 0;7) participated in the study. The data of three children (one girl, two boys) were excluded because of a diagnosis of developmental dyslexia (N=2) or not speaking Dutch as a native language (N=1), resulting in 27 child participants. The caretakers of the child participants gave active written consent for their participation.

4.1.2 *Design*

The familiarization phase and offline test phase were the same as in Experiment 1, except that the click sounds were not inserted into the stream.

4.1.3 *Procedure*

The experiment was carried out in a quiet room in the speech lab of the University of Amsterdam (adults) or a quiet room in the school of the children. The procedure was the same as in the previous experiment, except that during the familiarization phase, participants were asked to color a mandala (adults) or a coloring page (children), similar to the study by Saffran et al. (1997).[12] Testing took approximately 15 minutes. Participants received 5 euros (adults) or sticker sheets (children) as compensation for their participation.

4.2 Results

The mean accuracy on the offline test phase was 0.49 (SD = 0.20) for adults and 0.51 (SD = 0.13) for children. See Figure 6 for the descriptive data. Two generalized logistic linear mixed effect models (see section 3.2.1) were conducted. For the adults, the estimate for the intercept (converted into probability) was 0.50 (95% CI: 0.41 … 0.59), which is not significantly different from chance level (z = −0.123, p = 0.90). The main effect of Version was not significant. There was a significant effect of TargetOrder: the odds that an item in which the target was played first was answered correctly were 1.66 (95% CI: 1.02 … 2.46) times higher compared to an item in which the foil was played first: z = 2.210, p = 0.027. For the children, the estimate for the intercept (converted into probability) was 0.51 (95% CI: 0.45 … 0.56), which is not significantly different from chance level (z = 0.240, p = 0.81). The main effects of Version and TargetOrder were not significant. See Table 5 for the results of the model. Based on these null results, we cannot conclude whether adults and/or children do or do not distinguish words from part-words, indicating knowledge of word boundaries, in the offline test phase of the word segmentation task without the click detection task.

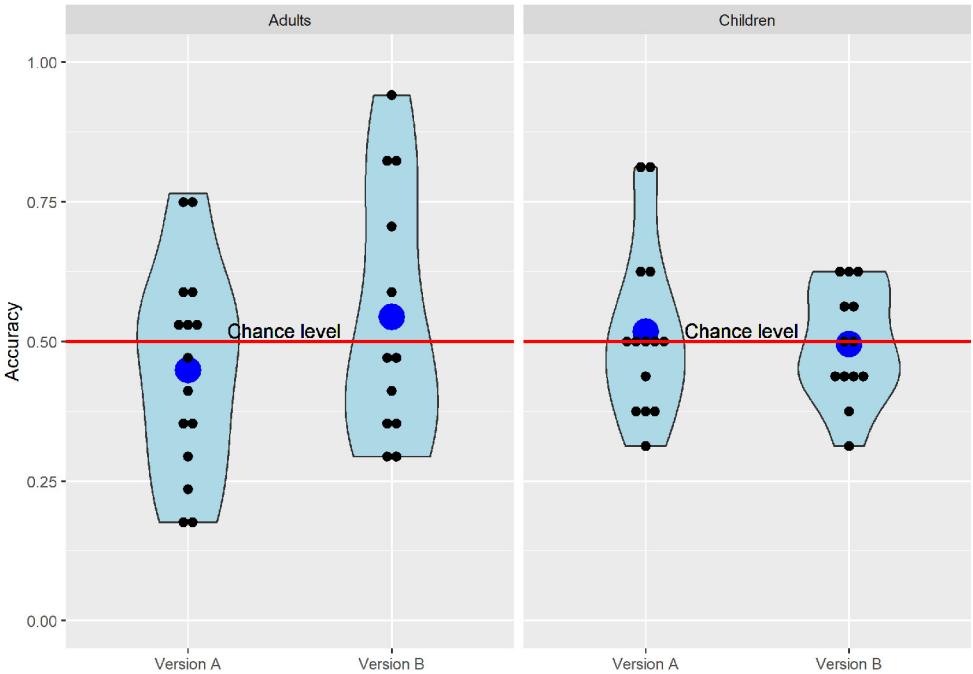

Figure 6 Descriptive plot of adults' and childrens' accuracy in version A and B of Experiment 2

Table 5 Results from the linear mixed effect model

	Intercept		**Version**		**TargetOrder**	
	Adults	***Children***	***Adults***	***Children***	***Adults***	***Children***
Estimate	Probability: 0.50	Probability: 0.51	Odds: 1.52	Odds: 0.91	Odds: 1.66	Odds: 1.22
95 % CI	0.41 … 0.59	0.45 … 0.56	0.52 … 4.51	0.43 … 1.90	1.02 … 2.46	0.75 … 2.01
z	−0.123	0.240	0.810	−0.273	2.210	0.839
p	0.90	0.81	0.42	0.79	0.027	0.40

4.3 Discussion

We did not find evidence for (or against) sensitivity to word boundaries in a group of adults and a group of school-aged children in our second experiment, in which we removed the click detection task from the word segmentation task, let alone that we could have anything to say about whether the null result in Experiment 1 was due to interference of the click detection task. For the adults there was a significant effect of the order of the targets and foils in the test items, indicating that it is easier to recognize

a target when it is played first in a test item. This factor should be considered when analyzing two-alternative forced choice data.

Our result is unexpected, as previous studies on word segmentation tasks (Batterink & Paller, 2019; Evans et al., 2009; Finn et al., 2018; Franco et al., 2015; Haebig et al., 2017; Mainela-Arnold & Evans, 2014; Mirman et al., 2008; Saffran et al., 1997; Saffran et al., 1996a; Saffran et al., 1996b; Toro et al., 2005) show that both adults and children perform above chance level in the offline test phase (but please note that previously found null results may not have been published, as is mentioned by Black and Bergmann (2017)). However, often non-words (combinations of syllables that had never occurred as such in the familiarization phase, TP = 0) are used as foils in the test phase. In the current study we used part-words as foils (TP = 0.333), which did occur in the familiarization phase, just less often than the words. Although discrimination between words and part-words has been shown in infant, child and adult studies on word segmentation (Batterink & Paller, 2019; Johnson & Tyler, 2010; Saffran et al., 1996a; Saffran et al., 1996b; Thiessen et al., 2005), in our third experiment we wanted to investigate whether sensitivity to word boundaries would be revealed as a preference for words over non-words (instead of part-words).

5 Experiment 3

5.1 Methods and materials

5.1.1 *Participants*

Forty-six adults (35 female, 11 male) between the age of 18;4 and 35;7 (M = 22;7, SD = 3;5) participated in the study. One participant was excluded because of the use of medicines, leaving 45 participants for data analysis. All remaining participants met the conditions as described in Experiment 1.

5.1.2 *Design*

The familiarization phase of the word segmentation task was identical to Experiment 2. However, the test phase was changed. Instead of part-words, combinations of syllables that had never occurred in the familiarization phase (non-words, TP = 0) were used as foils. For language A the non-words *kido*, *moba*, *dati* and *gilu* were constructed and for language B the non-words *bagi*, *timo*, *luda* and *doba*. The foils were constructed by combining the first syllable of a word with the second syllable of another word. Sequences with a double vowel (e.g. *daba*) or that only differed from a target word in one sound (e.g. *kida*) were avoided, and we aimed to construct a set of foils that contained all the syllables from the language. The new foils met the same conditions as the test stimuli in Experiment 1 and 2, and the test phase was constructed the same way.

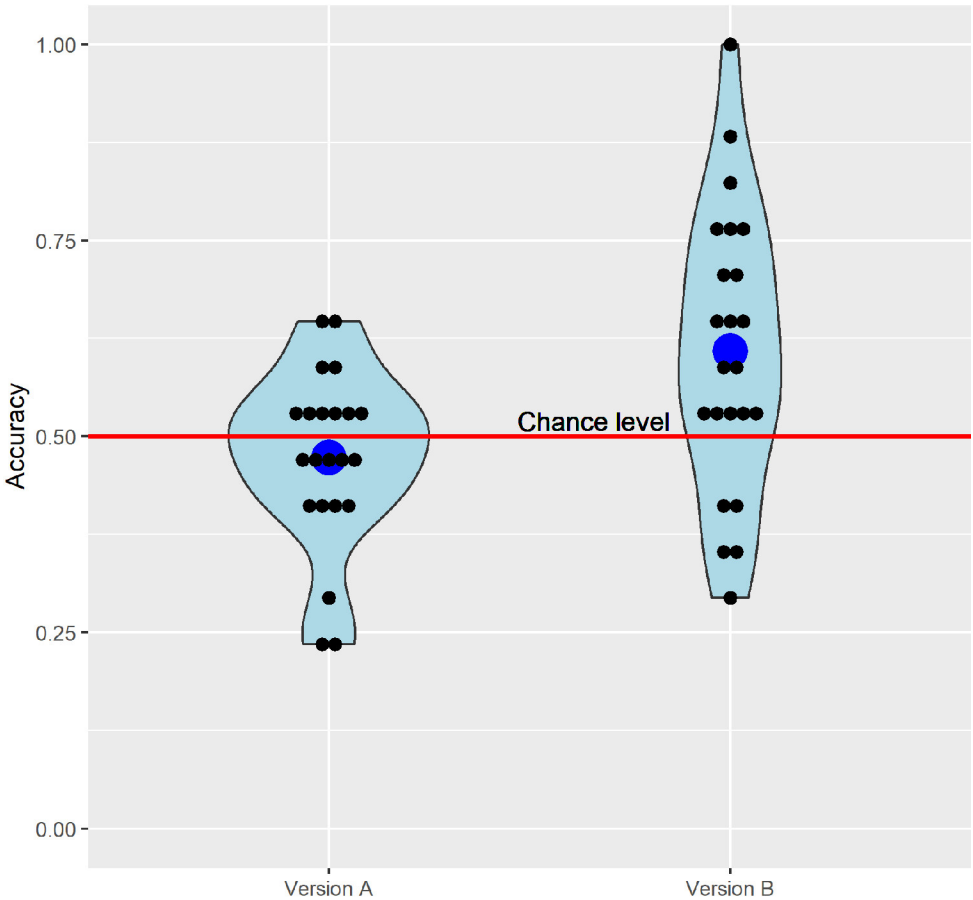

Figure 7 Descriptive plot of participants' accuracy in version A and B of Experiment 3

5.1.3 *Procedure*

The procedure was identical to that of Experiment 2.

5.2 Results

The mean accuracy was 0.54 (*SD* = 0.16). See Figure 7 for the accuracy data. A generalized logistic mixed effect model was conducted (see section 3.2.1). The estimate for the intercept (converted into probability) was 0.55 (95% CI: 0.47 … 0.63), which is not significantly different from chance level: $z = 1.147$, $p = 0.25$. There was a significant effect of Version: the odds that participants who did version B of the task chose the correct answer were 1.85 times (CI: 1.23 … 2.83) higher than for the participants who did version A ($z = 2.974$, $p = 0.0029$). The effect of TargetOrder was not significant. See Table 6 for the results of the model. On the basis of these null-results we cannot say whether adults distinguish

Table 6 Results from the linear mixed effect model

	Intercept	Version	TargetOrder
Estimate	Probability: 0.55	Odds: 1.85	Odds: 1.66
95 % CI	0.47 … 0.63	1.23 … 2.83	0.88 … 3.22
z	1.147	2.974	1.673
p	0.25	0.0029	0.094

words from non-words, which would have indicated knowledge of word boundaries, in the offline test phase of the current word segmentation task.

5.3 Discussion

As in our first two experiments, we did not find evidence that participants were sensitive (or not sensitive) to the statistical regularities in the input of the word segmentation task, even while we may suppose that the discrimination task was easier than in Experiment 1 and 2. There was a significant effect of test version which suggests that word boundaries in language B are somehow easier to learn than those of language A. However, while there was no evidence for lack of balance between the phonotactic probability of the foils and targets in the first two experiments, there was such evidence for Experiment 3. The average phonotactic frequency of the foils of language B (M = 0.769, SD = 0.021) was significantly lower than that of the targets of language B (M = 1.385, SD = 0.036): $t(6)$ = 5.172, p = 0.0021, while there was no significant difference between the foils of language A (M = 1.179, SD = 0.341) and the targets of language A (M = 1.425, SD = 0.174): $t(6)$ = 1.285, p = 0.246).[13] Thus, for language B it could be easier to choose targets over foils, because the targets sound more "Dutch-like" than the foils. It might be the case that the higher performance in version B of Experiment 3 does not (only) reflect learning of the word boundaries, but a bias that is inherent to the stimuli of the test phase.

6 General discussion

In the present set of studies, we aimed to find a method for measuring word segmentation online that was suitable for testing school-aged children, but we encountered some unexpected outcomes. Firstly, the click detection task does not seem to be a reliable measure of online word segmentation. There was a large amount of individual variation in the data, and test version may have influenced the results. We cannot state that responding to click sounds in a stream of interrupted syllables consistently revealed an effect of word boundary knowledge. The fact that our artificial language consisted of bisyllabic words

means that the click sounds occurring within words always divided the words in two single syllables. This is not the case for a language containing trisyllabic words (Franco et al., 2015; Gómez et al., 2011), as within-word clicks in that case still leave two syllables of that word uninterrupted. This repetition of high transitional probabilities (between syllable 1 and 2 and between syllable 2 and 3) might counter the effect of clicks being perceived as cues for word boundaries. Future studies might adapt our paradigm (adding blocks with only a few click sounds to the familiarization phase) using trisyllabic words to test whether the click detection task would reveal online learning then. A general difficulty with the click detection paradigm might be the so-called "auditory streaming effect" or "auditory stream segregation" (Micheyl et al., 2010; Van Noorden, 1975): participants might perceive the syllables and the click sounds as two separate sound streams. If this is the case (for some listeners), this might be the reason that the position of the click sounds does not influence their reaction time to them. Participants reported that they found it very hard to pinpoint in which specific syllable a click occurred. Moreover, the addition of click sounds to the stream could have distracted the listeners' attention away from the to-be-learned word boundaries, as suggested by Franco et al. (2015) and Toro et al. (2005). The below-chance performance on the offline test phase in our first experiment seemed to point in that direction. We wanted to investigate this by conducting Experiment 2, in which we tested the same word segmentation task without the addition of the click detection task.

Contrary to our expectations, we did not find evidence for or against sensitivity to words boundaries in adults or children in Experiment 2. This was also the case for our third experiment, in which the foils in the test phase were not part-words (TP = 0.333) but non-words (TP = 0). There are multiple factors that could have influenced our results. First, it could be the case that two-syllable words are somehow too short to "trigger" a statistical learning mechanism, although Graf Estes and Lew-Williams (2015) and Haebig et al. (2017), who also used bisyllabic words, did find a learning effect. However, differently from Haebig et al. (2017), our participants listened to the language for 8 minutes instead of 4.75 minutes. It is possible that the high amount of exposure to the syllables had given the participants the impression that the language consisted of monosyllabic words instead of bisyllabic words, resulting in less sensitivity to syllable combinations. Second, we used natural modified speech instead of synthesized speech. In a meta-analysis, Black and Bergmann (2017) found that infants' word segmentation ability was stronger in experiments that use synthesized speech. Natural speech contains more information than synthesized speech which could make the processing of the stream and consequently learning of word boundaries more difficult. Regarding the test phase, almost all participants stated that they found it difficult and often also reported that the test became more difficult as it progressed. This might be due to the repetition of the targets and foils in the test phase, which could overwrite the (weak) representations that might have been built during the familiarization phase (Siegelman et al., 2017).

Influence of prior linguistic knowledge could also have played a role (Finn & Kam, 2008; Siegelman et al., 2018; Van Hedger et al., 2020). Participants, especially adults or older children, who are subjected to an artificial language in a word segmentation task are not blank slates but already have linguistic knowledge and thus expectations about sounds and sound combinations. This knowledge might influence the learning process. As this influence is hard to predict correctly, the particular words that are chosen in an experiment might impact participants' performance. This is for example illustrated by findings of Erickson et al. (2016), who tested participants on two word segmentation tasks with different sets of words. Performance on one task did not reliably predict performance on the other task. Siegelman et al. (2018) suggest that the influence of prior linguistic knowledge (or "entrenchment effects") plays a very important factor in the large differences in effect sizes and reliability that is found between statistical learning studies. The influence of prior linguistic knowledge might in some cases be stronger than the influence of the statistical properties of the input that participants are briefly subjected to. Entrenchment effects could have led to the null results and the unexpected version effects in Experiment 1 and 3 of the current study and possibly more studies that have ended up in drawers. Future research should investigate this phenomenon in depth.

In sum, measuring word segmentation ability reliably might be more sensitive to methodological choices than assumed (see also Black and Bergmann, 2017). We would like to emphasize that studies that fail to find a significant effect may not be published (the "file drawer effect"). Access to null results is essential for reliable meta-analyses, which are an important source of empirical evidence. Therefore, it is important to report the results of our current study. Future research should systematically investigate what constraints the word segmentation ability and how (online) learning can be detected reliably (Siegelman et al., 2017). For example, Lukács et al. (2019) tested out an a method of measuring word segmentation online with a syllable detection task, and Kidd et al. (2020) used a serial recall task to measure offline learning more reliably.

Acknowledgements

Many thanks to our lab technician Dirk-Jan Vet for his help with the development of Praat and E-Prime scripts and providing the test equipment. Moreover, we want to thank Jeremy Bos, Lia Morrissey, Sonja Helder and Maayke Sterk for their help in recruiting and testing participants.

Notes

1 Prosody does play an important role in word segmentation (see for example Endress & Hauser, 2010).

2 While neurophysiological measures like EEG offer an excellent temporal resolution (see for example Kooijman et al., 2005), these methods are costly and more difficult to carry out with children compared to behavioral methods.

3 The novel words are loosely based on the study of Haebig et al. (2017), who constructed the sets *time*, *mano*, *dobu*, *piga* and *nome*, *mati*, *gabu*, *pido*. We aimed to use a set of distinct tense Dutch vowels. After consultation with some native speakers of Dutch, we excluded the /e/, as they agreed that a two-syllable word ending in /e/ (for example *time*) sounded "unnatural" in Dutch. As for the consonants, we aimed to use Dutch voiced and voiceless obstruents (plosives and fricatives) as well as sonorants (nasals and approximants). We decided to use /χ/ instead of /g/, as /g/ only occurs in English loanwords in Dutch. Syllables that have a meaning in Dutch (for example *χa* means 'go') were excluded. Finally, we aimed to balance the mean Dutch phonotactic frequency of the novel words (between 1.197 and 1.607).

4 The phonotactic frequency (logTP) of the novel words was computed using the Dutch Phonotactic Frequency Database (Adriaans, 2006).

5 In a pilot study, the familiarization phase was constructed such that there were no clicks at all in part 1 and 3. However, participants reported that it was confusing that it took 2 minutes to hear the first click and that the clicks stopped after 4 minutes. Therefore, we decided to include a few clicks throughout the whole familiarization phase to make the click detection task more consistent.

6 We did not use the same stimuli as in the familiarization phase as these were recorded in co-articulation contexts, and the test items needed to be articulated as separate words.

7 The task was developed in such a way that it should be suitable for child participants as well, as we intended to test online word segmentation in children in a later stage of the project.

8 A reviewer put forward the interesting suggestion that the different performance we found for the two versions might be an item effect, as it could be the case that participants responded faster to clicks occurring in syllables that are more regular in Dutch. If we compare the mean phonotactic probability of the syllable sets *ba*, *ti*, *lu*, *do* ($M = 1.96$) vs *ki*, *mo*, *da*, *χi* ($M = 1.656$), which either contain between-word clicks or within-word clicks depending on the test version, we do not find a significant difference: $t = -1.295$, $p = 0.243$. Although the difference is not significant, the direction of the difference does correspond to the pattern that responses to clicks across blocks in language A developed more according to our expectations compared to language B. For language A, syllables that contained between-word clicks words had a higher phonotactic probability in Dutch than syllables that contained within-word clicks while it was the other way around for language B. As we hypothesized that between-word clicks should have been detected faster, this difference could have contributed to the finding that the results for language A were more as we expected than the results for language B.

9 In a pilot study, participants did the familiarization phase without any other task but listening

to the language. Participants reported that 8 minutes seemed very long and that they felt uneasy.

10 Admittedly, the difference between the *p*-value of 0.0021 and the *p*-value of 0.246 was not significant ($p = 0.13$), so we cannot really say that the disbalance between the targets and foils of language B was greater than that of language A.

References

Adriaans, F. (2006). *PhonotacTools* (Test version) [Computer program]. Utrecht Institute of Linguistics OTS.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Batterink, L.J., & Paller, K.A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. *Cortex*, *115*, 56–71. https://doi.org/10.1016/j.cortex.2019.01.013

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. *Proceedings of the 39th Annual Meeting of the Cognitive Science*, 124–129.

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer* Version 6.1.05 [Computer program]. http://www.praat.org/.

Cohen, L., & Mehler, J. (1996). Click monitoring revisited: An on-line study of sentence comprehension. *Memory & Cognition, 24*(1), 94–102.

Cole, R. (1980). *Perception and production of fluent speech*. Lawrence Erlbaum Associates.

Endress, A.D., & Hauser, M.D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199. https://doi.org/10.1016/j.cogpsych.2010.05.001

Erickson, L., Kaschak, M., Thiessen, E., & Berry, C. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra: Psychology*, *2*(1). https://doi.org/10.1525/collabra.41

Evans, J.L., Saffran, J.R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research: JSLHR*, *52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189)

Finn, A.S., & Kam, C.L.H. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Finn, A.S., Kharitonova, M., Holtby, N., & Sheridan, M.A. (2018). Prefrontal and hippocampal structure predict statistical learning ability in early childhood. *Journal of Cognitive Neuroscience*, *31*(1), 1–12. https://doi.org/10.1162/jocn_a_01342

Fodor, J.A., & Bever, T.G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, *4*(5), 414–420.

Foss, D.J., & Lynch Jr., R.H. (1969). Decision processes during sentence comprehension: Effects of surface structure on decision times. *Perception & Psychophysics, 5*(3), 145–148.

Franco, A., Gaillard, V., Cleeremans, A., & Destrebecqz, A. (2015). Assessing segmentation processes

by click detection: Online measure of statistical learning, or simple interference? *Behavior Research Methods*, *47*(4), 1393–1403. https://doi.org/10.3758/s13428-014-0548-x

Gómez, D.M., Bion, R.A., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes*, *26*(2), 212–223. https://doi.org/10.1080/01690965.2010.482451

Graf Estes, K. (2012). Infants generalize representations of statistically segmented words. *Frontiers in Psychology*, *3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00447

Graf Estes, K., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1517–1528. https://doi.org/10.1037/a0039725

Haebig, E., Saffran, J.R., & Ellis Weismer, S. (2017). Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, *58*(11), 1251–1263. https://doi.org/10.1111/jcpp.12734

Johnson, E.K., & Tyler, M.D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x

Kidd, E., Arciuli, J., Christiansen, M.H., Isbilen, E.S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, *200*, 104964.

Kooijman, V., Hagoort, P., & Cutler, A. (2005). Electrophysiological evidence for prelinguistic infants' word recognition in continuous speech. *Cognitive Brain Research*, *24*(1), 109–116. https://doi.org/10.1016/j.cogbrainres.2004.12.009

Kraemer, H.C., & Blasey, C.M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, *13*(3), 141–151.

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Children with developmental language disorder have an auditory verbal statistical learning deficit: Evidence from an online measure. *Language Learning*. https://doi.org/10.1111/lang.12373

Lukács, A., Dobó, D., Szóllósi, A., Németh, K., & Lukics, K. (2019). Domain general statistical learning impairment in dyslexia: Sensitivity of online and offline measures across modalities and domains. Poster session presented at *Interdisciplinary Advances on Statistical learning 2019*, June 27–29, San Sebastián, Spain.

Mainela-Arnold, E., & Evans, J.L. (2014). Do statistical segmentation abilities predict lexical-phonological and lexical-semantic abilities in children with and without SLI? *Journal of Child Language*, *41*(2), 327–351. https://doi.org/10.1017/S0305000912000736

Micheyl, C., Hunter, C., & Oxenham, A.J. (2010). Auditory stream segregation and the perception of across-frequency synchrony. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1029–1039. https://doi.org/10.1037/a0017601

Mirman, D., Magnuson, J.S., Graf Estes, K., & Dixon, J.A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*(1), 271–280. https://doi.org/10.1016/j.cognition.2008.02.003

R Core Team (2020). *A language and environment for statistical computing* [Computer program]. https://www.r-project.org/.

Romberg, A.R., & Saffran, J.R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

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 ear. *Psychological Science*, *8*(2), 101–105.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime (Version 2.0). [Computer software and manual]. Psychology Software Tools Inc.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, *177*, 198–213. https://doi.org/10.1016/j.cognition.2018.04.011

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2017). Redefining "learning" in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, 1–36. https://doi.org/10.1111/cogs.12556

Thiessen, E.D., Hill, E.A., & Saffran, J.R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71.

Toro, J.M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34. https://doi.org/10.1016/j.cognition.2005.01.006

Van Hedger, S.C., Johnsrude, I., & Batterink, L. (2020). *Prior real-world experience influences nonlinguistic statistical learning*. PsyArxiv. https://doi.org/10.31234/osf.io/yscn8

Van Noorden, L.S. (1975). *Temporal coherence in the perception of tone sequences*. PhD thesis. Eindhoven University of Technology.