Distributional cues and the onset bias in early word segmentation

Mireille Babineau

&

Rushen Shi

Université du Québec à Montréal

Montreal, Canada, H3C 3P8

babineau.mireille@courrier.uqam.ca; shi.rushen@uqam.ca

Date of the last revision: Oct 17, 2014
Abstract

In previous infant studies on statistics-based word segmentation, the unit of statistical computation was always aligned with the syllabic edge, which had a consonant onset. The current study addressed whether the learning system imposes a constraint that favors word forms beginning with a consonant onset over those beginning with an onsetless sub-syllable, by examining infants' segmentation of vowel-initial nonwords in French liaison. French-learning 20- and 24-month-old infants (N=64) were familiarized with sentences containing variable liaison consonants preceding the same vowel-initial nonword (e.g., /n/onche, /z/onche, /t/onche, /t/onche), such that the distributional cues supported the sub-syllabic target (e.g., onche). After familiarization, we tested sub-syllabic statistical segmentation by presenting the vowel-initial target (e.g., onche) versus another non-familiarized vowel-initial word (e.g., èque). Another group of infants were tested with a consonant-initial mis-segmentation of the target (e.g., zonche) versus another non-familiarized consonant-initial word (e.g., zèque). Results showed that 20-month-olds failed to segment the vowel-initial targets, but they mis-segmented the targets as consonant-initial, indicating that the onset bias dominated over sub-syllabic statistics for word segmentation at this age. Twenty-four-month-olds showed ambiguous interpretations, i.e., both vowel-initial segmentation and consonant-initial mis-segmentation, suggesting that the use of statistics to segment sub-syllabic words was emerging while the onset bias continued to have an impact.


Keywords: Statistical learning, constraints, infant word segmentation, liaison, language acquisition.
Distibutional cues and the onset bias in early word segmentation

Segmenting words from continuous speech is essential for building a vocabulary. Infants must locate the beginning and the end of words, which are not clearly marked in running speech. Previous research has shown that infants segment word-like forms using various kinds of information such as statistical cues, prosodic cues, phonotactic cues, etc. (e.g., Curtin, Mintz & Byrd, 2001; Jusczyk, Houston & Newsome, 1999; Mattys & Jusczyk, 2001; Saffran, Aslin & Newport, 1996; Shi & Lepage, 2008; Shi, Cutler, Werker, & Cruickshank, 2006). Few studies have systematically tested the relative weight of different cues to word segmentation. E. K. Johnson and Jusczyk (2001) showed that early word segmentation relies more on prosody than statistics, whereas Thiessen and Saffran (2003) found that statistical cues are more important than prosody for initial word segmentation.

Most infant word segmentation studies have tested consonant-initial words, showing that segmentation emerges from 6 months of age (e.g., E. K. Johnson & Tyler, 2010; Jusczyk & Aslin, 1995). For such words (e.g., *cup*), the word-initial consonant is also the syllable onset, a characteristic that may assist segmentation. A few studies (with English-learning infants) have examined the segmentation of vowel-initial words, and have shown that these forms are harder to segment (Mattys & Jusczyk, 2001; Nazzi, Dilley, Jusczyk, Shattuck-Hufnagel, & Jusczyk, 2005; Seidl & Johnson, 2008). In particular, English-learning infants at 11 months of age can segment vowel-initial words only in utterance-initial and utterance-final positions (Seidl & Johnson, 2008), and they begin segmenting vowel-initial words in utterance-medial positions only from 13.5 months of age (Nazzi, et al., 2005). Thus, infants are delayed in segmenting vowel-initial words relative to consonant-initial words. These findings are consistent with the universal onset constraint proposed in phonological theory (Kager, 1999), according to which word forms that begin with an onsetless syllable are disfavored whereas word forms with a consonant onset are favored. In this study we used French liaison to examine whether this onset bias guides infants’ word segmentation and how it may affect distribution-based word segmentation.

French liaison involves vowel-initial words that can be particularly hard to segment because of a complex kind of resyllabification. Specifically, an underlying final consonant of a liaison-causing word
(that is, liaison consonant) surfaces as the syllable onset of the following vowel-initial word. The liaison-causing words are primarily determiners (e.g., un “a”, les “the”) and pronouns (e.g., ils “they”, on “we”, etc.), as well as adjectives (e.g., gros “big”, bon “good”, dernier “last”, petit “little”) and some other categories of words (e.g., the verb est, the preposition en). Liaison occurs frequently in noun phrases and in Pronoun+Verb sequences. For instance, when the liaison-causing word un ([œ̃ “a”] and the noun air ([ɛʁ] “melody”) are combined, the phrase surfaces as [œ̃.ɛʁ], which is homophonous with un nerf ([œ̃.nɛʁ] “a nerve”), in which the second word is inherently consonant-initial. Liaison consonants never surface elsewhere. The determiner un, with an underlying /n/, is produced as [œ̃] in isolation, in utterance-final positions (e.g., Il en prend un [ilɑ̃pœ̃e] “He takes one.”), and before a consonant-initial word (e.g., un bébé [œ̃.bebe] “a baby”). This characteristic differs from non-liaison cases, where the word-final consonant can be phonetically realized as a coda (e.g., cold hat, that book, chaque table [ʃa.kabl] “every table”). Since liaison consonants rarely surface as the coda of the preceding word in the input, it may be hard for infants to understand that the consonant belongs to the preceding word and not to the following word.

Furthermore, resyllabification is phonetically more complete in liaison than in non-liaison cases. In non-liaison ambiguous cases a resyllabified onset consonant (e.g., /k/ in chaque ours [ʃa.kuRs] “each bear”) is shorter than an inherent onset consonant (e.g., /k/ in chaque course [ʃa.kuRs] “each errand”), whereas the length difference between a liaison consonant (e.g., /l/ in petit /lami, “little friend”) and the inherent onset consonant of the second word (e.g., e.g., petit tamis “little sieve”) is much smaller (Yersin-Besson & Grosjean, 1996). Consequently, it is easier for French adults to identify the second word in non-liaison ambiguous cases than in liaison ambiguous cases (Yersin-Besson & Grosjean, 1996). Even infants detect the acoustic cues to word boundaries in non-liaison resyllabification cases. In Mattys and Jusczyk (2001), for example, although English-learning 8.5-13-month-olds failed to segment ice from
weird ice, they did not mis-segment it as dice. They segmented dice only from two dice. No perceptual study has been conducted on infants’ segmentation of vowel-initial words in liaison.

Studies on children’s liaison production suggest that French-learning 2-3-year-olds mis-segment vowel-initial words as consonant-initial from liaison cases. Errors included substituting the correct liaison consonant with an incorrect liaison consonant (e.g., *un zarbre for un /n/arbre, “a tree”; *petit noeil for petit /t/oil, “little eye”), adding a liaison consonant (*Néléphant, viens ici instead of Éléphant, viens ici, “Elephant, come here!”), etc. (Chevrot, Chabanal, & Dugua, 2007; Chevrot & Fayol, 2001; Dugua, Chevrot, & Fayol, 2006). Hearing a particular word consistently in one liaison context could yield such mis-segmentations. Children might also hear variable liaisons for the same vowel-initial word in the input (e.g., un /n/arbre, “a tree”; les /z/arbres, “the trees”). Chevrot and Fayol (2001) reported that different liaison onsets (e.g., /n/orage, /z/orage) were produced by a child for the same concept (e.g., orage, “thunderstorm”). This consonant-onset bias is consistent with adult perception models that favor syllabic alignment for word segmentation (e.g., Syllable Onset Segmentation Heuristic: Content, Dumay & Frauenfelder, 2000; Possible Word Constraint: Norris, McQueen, Cutler & Butterfield, 1997). It is also consistent with the universal preference for onsets (i.e., the onset constraint) in phonological grammar (Kager, 1999).

On the other hand, 2-3-year-olds also show evidence of correctly segmenting and storing vowel-initial forms from liaison contexts. For example, they sometimes omit liaison consonants incorrectly, e.g., *des ours [de.uRs] instead of des /z/ours [de.zuRs], “some bears” (Chevrot & Fayol, 2001; Dugua, Chevrot & Fayol, 2006). It is interesting to ask how children acquire the correct vowel-initial interpretation.

One mechanism that could account for children’s acquisition of liaison segmentation is distributional analysis. The use of distributional cues such as transitional probabilities for segmenting words has been demonstrated in English-learning infants during the first year of life (e.g., Saffran, Aslin, & Newport, 1996; Thiessen, & Saffran, 2003). Syllables with high transitional probabilities (TPs) (e.g., between two frequently co-occurring syllables) were perceived as a word, and a boundary was interpreted
between syllables that had low TPs. It is unknown whether infants can perform statistical computations at a sub-syllabic level that yield vowel-initial word segmentation. Specifically, when a vowel-initial word is preceded by various liaison contexts (e.g., *un *arbre, *des *arbres, *petit *arbres*), the TP between the vowel-initial word *arbre* and the liaison consonants is lowered, supporting the correct boundary of the vowel-initial word. The statistical cues, however, are in conflict with the onset constraint, which favors syllables with a consonant onset (Kager, 1999). In previous work the relative contributions of statistical cues versus prosodic prominence to infants’ word segmentation were tested with artificial languages (E. K. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003), and the unit of statistical computation was the syllable. In liaison segmentation, the conflict between the statistical cues and the onset constraint (i.e., onset bias) offers an ideal case for testing the relative impact of these two types of cues on word segmentation.

The present study aimed at examining the mechanisms underlying word segmentation. We tested infants’ ability to use sub-syllabic statistical cues for segmenting vowel-initial words in French liaison. We used a perceptual task, which allowed us to examine French-learning infants younger than those tested in production studies. Importantly, to clearly examine distributional learning, we used pseudo-nouns, which were advantageous over familiar words, and we manipulated the liaison environment in our stimuli. Unlike familiar words, for which it would be difficult to determine the liaison environment that each child had experienced prior to our study, pseudo-words were equally new for our participants. Thus, the exposure was controlled for all infants. We presented infants with a vowel-initial target pseudo-noun following varying liaison consonant contexts, thus creating statistical support for vowel-initial segmentation. This design enabled us to test precise predictions. If infants cannot use sub-syllabic statistical cues, and are guided by the universal onset constraint (Kager, 1999), then they should segment consonant-initial forms even after hearing variable liaison contexts (e.g., segmenting *zonche, nonche, tonche, ronche* after hearing */z/onche, */n/onche, */t/onche, */r/onche*). However, if infants can use sub-syllabic statistical cues, they should segment the vowel-initial form (e.g., *onche*). Thus, Experiments 1a
and 2a tested vowel-initial segmentation, and Experiments 1b and 2b tested consonant-initial segmentation.

**Experiment 1a**

**Method**

**Participants.** Sixteen monolingual Quebec-French-learning infants aged 20 months completed this experiment (mean age = 631 days; range = 615-662 days; 6 girls). The data of another 13 infants were excluded from analyses because of parental interference (4), experimental error (3), and fussiness or failure to complete the experiment (6). We were authorized by the Quebec government to obtain birth records of infants in Montreal. The records were randomly drawn by their computer, with no specification of infants’ background such as ethnicity, cultural and family economic status. Given the strong socialized system in Quebec, which covers equal education and health care as well as other social support, the population does not differ widely in their socioeconomic levels. A letter about our laboratory and our study was first sent to the families. The families were then contacted by phone.

**Stimuli.** The stimuli consisted of two target pseudo-nouns (*onche* and *èque*) that conformed to the phonological structure of French. They were each embedded in eight French sentences. In each sentence the target was preceded by a liaison-causing word (e.g., *J’ai trouvé mon onche sur le pavillon*). Four liaison consonants (/n/, /z/, /t/, /r/) surfaced across sentences. For each liaison consonant, two sentences were created, and two different words preceded the pseudo-noun. Each pseudo-noun was preceded by *mon* [mɔ̃] and *un* [œ̃] respectively, both causing the /n/ liaison, e.g., *un [n]onche* (‘an onche’), *mon [n]onche* (‘my onche’). The words *ces* ([œ]) and *gros* ([ɡʁo]) caused the /z/ liaison, e.g., *ces [z]onches* (‘these onches’), *gros [z]onches* (‘big onches’). The words *grand* [ɡʁɑ̃] and *petit* [pəti] caused the /t/ liaison (e.g., *grand [t]onche* ‘tall onche’, *petit [t]onche* ‘little onche’), and *premier* [pʁəmje] and *dernier* ([dəʁnje]) caused the /r/ liaison (e.g., *premier [r]onche* ‘first onche’, *dernier [r]onche* ‘last onche’).

The two pseudo-nouns in isolation and the sentences were recorded (sample rate: 44.1 kHz, bit rate: 16 bit) by a female native speaker of Quebec French in an IAC (Industrial Acoustics Company)
acoustic chamber. The speaker used infant-directed speech style. The final stimuli included two tokens of each sentence (2 x 8 sentences containing onche and 2 x 8 sentences containing èque), and 12 isolated tokens of each pseudo-noun (12 onche and 12 èque). The sentences served as the familiarization materials, and the isolated pseudo-nouns as the test stimuli. We note that nouns in French primarily occur in noun phrases that contain a determiner. Nevertheless, nouns without a determiner occur sometimes, e.g., following certain prepositions or used in isolation (e.g. calling context: Bébé, viens! “baby, come!”). Since the determiner context typically results in resyllabification with the following vowel-initial word, we therefore used isolated pseudo-nouns so as to clearly assess vowel-initial versus consonant-initial word segmentation.

Auditory stimuli also included a water bubble sound for the pre-trial and a cricket call for the attention getter. Visual stimuli included a green abstract picture for all trials and a video of a moving star for the attention getter. The abstract picture was boundless so that infants would not perceive the image as an object.

**Procedure.** Each infant was tested in a visual preferential procedure. In a sound-attenuated room the infant sat on his or her parent’s lap approximately 1.5 m in front of a 42-in monitor, which presented the visual stimuli. Two loudspeakers on each side of the monitor presented the auditory stimuli. The parent wore headphones that delivered masking music. A computer program (Cohen, Atkinson, & Chaput, 2000) was used to run the experiment.

In the adjacent room, a researcher blind to the audio-visual materials observed the infant’s eye movements through a closed-circuit TV. The researcher initiated each trial by pressing down a computer key when the infant looked at the monitor. The trials terminated when the pre-determined trial duration (6.4 secs for familiarization trials and 18.5 secs for test trials) was reached. The attention getter (a moving star accompanied by the cricket call) appeared between trials to attract the infant back to the screen.

During each trial, the green abstract picture and the speech stimuli were simultaneously presented. The water bubble sound was presented in a pre-trial to acquaint the infant with the procedure, and in a post-trial to mark the end of the task.
All experimental sessions were videotaped. Another researcher naive to the experiment and blind to the audio-visual stimuli coded babies’ visual fixations offline at a rate of 30 frames/sec.

**Design.** A familiarization phase preceded a test phase. Half of the infants were familiarized with sentences containing the pseudo-noun *onche*, and the other half with sentences containing the pseudo-noun *èque* (see Table 1). During the test phase all infants heard isolated tokens of *onche* and *èque* that were presented in alternating trials. One test trial type was the “familiarized” pseudo-noun, and the other was the “non-familiarized” pseudo-noun. That is, for infants who heard sentences containing *onche* during the familiarization phase, the *onche* test trials were considered as “familiarized”, and the *èque* test trials as “non-familiarized”. For infants who were familiarized with sentences containing *èque*, the *èque* test trials were “familiarized”, and the *onche* test trials were “non-familiarized”. The presentation of either *onche* or *èque* as the first test trial was counter-balanced across infants.

[Insert Table 1 here]

Each familiarization trial lasted 6.4 secs, and each test trial lasted 18.5 secs. There were eight familiarization trials, each presenting two different sentences with an inter-sentence interval of approximately 998 ms. The sentences were on average 2.39 secs long. Thus, every infant heard each of the 8 sentences twice across trials.

There were ten test trials (five *onche* trials and five *èque* trials), each containing all 12 tokens of either *onche* or *èque*, with an average inter-stimulus interval of 1000 ms. On average the *onche* tokens were 619 ms, and the *èque* tokens were 529 ms.

The variable liaison contexts in the familiarization phase provided sub-syllabic statistical cues that supported a vowel-initial interpretation, since a low TP was present between the various liaison consonants and the target pseudo-noun (*onche* or *èque*). If 20-month-olds can use the sub-syllabic statistical cues to segment the vowel-initial target from the familiarization sentences, they should discriminate between “familiarized” and “non-familiarized” test trials. Discrimination could in principle be either a familiarity preference or a novelty preference (e.g., Cyr & Shi, 2013; Hunter & Ames, 1988; Thiessen & Saffran, 2003). Both directions of preference are observed in infant segmentation studies, and
either would be evidence for successful segmentation. However, since vowel-initial word segmentation is
delayed into the second year of life (Mattys & Jusczyk, 2001; Nazzi, et al., 2005) and since liaison
segmentation is particularly hard, we expected a familiarity preference if infants could segment the
targets.

**Results.** For each infant, the total looking times in secs for the “familiarized” test trials and for
the “non-familiarized” test trials were compiled respectively. The average looking time per trial was then
calculated for each trial type. As shown in Figure 1, looking times to the “familiarized” vowel-initial
pseudo-noun ($M=9.68; SE=0.83$) and to the “non-familiarized” vowel-initial pseudo-noun ($M=9.55;
SE=.74$) was not different, $t(15)=.29$, $p=.776$ (all $t$-tests reported in this paper are two-tailed), 95% CI [-
0.85, 1.12], $d=0.04$, two-tailed, suggesting that 20-month-olds did not show evidence of segmenting the
vowel-initial targets. To examine if this lack of segmentation was comparable for both the group
familiarized with *onche* and the group familiarized with *èque*, we conducted a repeated-measure ANOVA
with Test Item (familiarized versus non-familiarized) as within-subject factor and *onche*/èque
(familiarization groups) as the between-subject factor. There was no main effect. Crucially, no interaction
of the two factors was observed, indicating that neither *onche* nor *èque* was segmented.

We then hypothesized that infants at this age might mis-segment vowel-initial words in liaison
cases as consonant-initial based on the onset bias. Experiment 1b assessed this hypothesis by presenting
isolated consonant-initial test nonwords (*zonche* and *zèque*) after familiarizing infants with the same
sentences of Experiment 1a.

**Experiment 1b**

**Method**

**Participants.** Participants were a new group of 16 monolingual Quebec-French-learning infants
aged 20 months (mean age = 637 days; range = 619-659 days; 8 girls). An additional 21 infants were
tested, whose data were excluded from analyses because of parental interference (3), ceiling effect (1),
and fussiness or failure to complete the experiment (17).
Stimuli. Familiarization stimuli were identical to those of Experiment 1a. The pseudo-nouns *zonche* and *zèque* were used for the test trials. They were recorded by the same speaker as in Experiment 1a. The final test stimuli consisted of 12 isolated tokens for each pseudo-noun, with an average inter-stimulus interval of 974 ms. On average the *zonche* tokens were 656 ms, and the *zèque* tokens were 596 ms.

Procedure and design. The procedure and the design were the same as in Experiment 1a, except that consonant-initial pseudo-nouns (*zonche* and *zèque*) were used as the test stimuli (see Table 2). For each group of infants, one test trial type was “familiarized”, and the other test trial type was “non-familiarized”. The “familiarized” test word matched the /z/ liaison surface form in the familiarization sentences (e.g., … *ces [z]onches* …), although the intended target form in the sentences was vowel-initial. If 20-month-olds segmented the pseudo-nouns based on the onset bias, they should show a consonant-initial interpretation by discriminating between the “familiarized” and “non-familiarized” test trials. Here we may expect a novelty preference since consonant-initial word segmentation is generally easy for infants even before age one (e.g., Jusczyk & Aslin, 1995).

Results. As in Experiment 1a, looking times in secs to the “familiarized” and the “non-familiarized” test trials were calculated for each infant. As shown in Figure 1, infants discriminated between “familiarized” (M=8.54; SE=.55) and non-familiarized trials (M=9.92; SE=.68), t(15)=-2.94, p=.01, 95% CI [-2.37, -0.38], d =-0.56, two-tailed, suggesting a consonant-initial mis-segmentation driven by the onset bias. To examine if this mis-segmentation was present for both the group familiarized with *zonche* and the group familiarized with *zèque*, we conducted the same ANOVA as in Experiment 1a (within-subject factor: Test Item (familiarized versus non-familiarized); between-subject factor: *zonche/zèque* familiarization groups). While a main effect of Test Item was observed, F(1, 14)=9.452, p = .008, η²=0.37; there was no main effect of *zonche/zèque*, and no interaction of the two factors, indicating that both *zonche* and *zèque* were segmented.
Furthermore, a repeated-measure ANOVA, with Test Item (“familiarized” versus “non-familiarized”) as the within-subject factor, and Segmentation (vowel-initial versus consonant-initial, i.e., Experiment 1a versus 1b) as the between-subject factor revealed a significant interaction between Test Item and Segmentation ($F(1, 30)=5.29$, $p=0.029$, $\eta^2=0.14$), indicating that indeed, infants in the two experiments responded differently. The ANOVA showed no significant main effect of Test Item ($F(1, 30)=3.58$, $p=0.068$, $\eta^2=0.09$), nor a main effect of Segmentation ($F(1, 30)=0.17$, $p=0.686$, $\eta^2<0.001$).

**Discussion.** The combined results of Experiments 1a and 1b suggest that at 20 months of age infants could not use the sub-syllabic statistical information to segment vowel-initial words in liaison. Rather, they were guided by the onset bias, leading to consonant-initial mis-segmentation. It is interesting that infants showed a novelty preference in Experiment 1b. Given that a novelty preference may indicate robust abilities (e.g., Cyr & Shi, 2013; Hunter & Ames, 1988; Thiessen & Saffran, 2003), our results suggest that 20-month-olds are strong in consonant-initial segmentation. This is unsurprising since consonant-initial word segmentation begins very early in life (e.g., E. K. Johnson & Tyler, 2010). Our 20-month-olds failed to segment vowel-initial words in liaison, consistent with the general delay in segmenting vowel-initial words shown in other studies (Mattys & Jusczyk, 2001; Nazzi, et al., 2005; Seidl & Johnson, 2008).

In Experiment 2 we tested whether older infants have developed the ability to segment vowel-initial words in liaison using sub-syllabic statistical cues. Experiment 2a presented the stimuli of Experiment 1a. Experiment 2b presented the stimuli of Experiment 1b. The familiarization stimuli were the same across experiments, containing both sub-syllabic statistical cues supporting vowel-initial segmentation and the onset cue supporting consonant-initial segmentation. The two types of cues were conflicting, leading to different predictions. If the onset bias continues to be dominant over the sub-syllabic statistical cues for 24-month-olds, their responses should be the same as the 20-month-olds’, i.e., discrimination of test trials in Experiment 2b (consonant-initial segmentation) but not in Experiment 2a (vowel-initial segmentation). However, infants at 24 months may have learned that the onset bias is
violable given sufficient input evidence against it. If so, the sub-syllabic statistical cues may dominate over the onset cues, and infants should discriminate the test trials in Experiment 2a but not in Experiment 2b. In this case we would expect a familiarity preference in Experiment 2a as it would be an emerging ability. But if the onset bias and sub-syllabic statistical cues have comparable weight, two alternative outcomes are predicted: 1) the 24-month-olds may be confused in their interpretation and thus fail to discriminate the test trials in both Experiments 2a and 2b; or 2) they may interpret both vowel-initial and consonant-initial forms as equally acceptable, i.e., discriminating test trials in both experiments. For the latter outcome, we may expect a familiarity preference in both experiments since attending to additional cues may make the task harder.

**Experiment 2a**

**Method**

**Participants, stimuli, procedure and design.** Sixteen monolingual Quebec-French-learning infants aged 24 months completed this experiment (mean age = 753 days; range = 744-771 days; 10 girls). The data of an additional 17 infants were excluded from analyses because of parental interference (1), experimental error (4), and fussiness or failure to complete the experiment (12). The stimuli, procedure and design were the same as those of Experiment 1a (see Table 1).

**Results.** Looking times in secs to the “familiarized” and the “non-familiarized” test trials were calculated as in previous experiments. Infants looked significantly longer while listening to the “familiarized” vowel-initial pseudo-noun \((M=11.1; \ SE=.81)\) than to the “non-familiarized” vowel-initial pseudo-noun \((M=9.65; \ SE=.84)\), \(t\) (15)=2.86, \(p=.012\), 95\% CI [0.37, 2.53], \(d=0.44\), two-tailed (see Figure 2). To examine if this segmentation was present for both the group familiarized with *onche* and the group familiarized with *èque*, we conducted the same ANOVA as in Experiment 1a with *onche/èque* familiarization groups as the between-subject factor. The main effect of Test Item was significant, \(F(1, 14)=8.215, p = .012, \eta^2=0.35\). There was no main effect of *onche/èque* familiarization groups. No interaction of the two factors was observed, indicating that both *onche* and *èque* were segmented. Thus, by 24 months infants have learned to use sub-syllabic statistical cues to segment vowel-initial words in
liaison. These results contrast with those of 20-month-olds, who lacked this ability. To fully succeed in segmenting the vowel-initial words, infants must also learn to inhibit the onset bias. Experiment 2b tested this question by presenting 24-month-olds with the stimuli of Experiment 1b.

Experiment 2b

Method

Participants, stimuli, procedure and design. Participants were 16 monolingual Quebec-French-learning infants aged 24 months (mean age = 752 days; range = 739-775 days; 11 girls). The data of another 14 infants were excluded because of parental interference (1), experimental error (2), and fussiness or failure to complete the experiment (11). The stimuli, procedure and design were identical to those of Experiment 1b (see Table 2).

Results. Looking times in secs to the “familiarized” and the “non-familiarized” test trials were calculated as in previous experiments. Infants looked significantly longer in “familiarized” ($M=9.89; SE=.65$) than in “non-familiarized” test trials ($M=8.51; SE=.57$), $t(15)=2.75, p=.015, 95\% \text{ CI}[0.31,2.45], d=0.56$, two-tailed (see Figure 2), suggesting that 24-month-olds continue to show a consonant-initial interpretation despite hearing the target with various liaison consonants. Moreover, we conducted the same ANOVA as in Experiment 1b with zonche/zèque familiarization groups as the between-subject factor. The main effect of Test Item was significant, $F(1, 14)=7.215, p=.018, \eta^2=0.34$. There was no main effect of zonche/zèque familiarization groups. No interaction of the two factors was observed, indicating that both zonche and zèque were segmented.

A repeated-measure ANOVA on the data of Experiments 2a and 2b revealed a significant main effect of Test Item (i.e., familiarized versus non-familiarized) ($F(1, 30)=15.76, p<.01, \eta^2=0.34$), no interaction between Test Item and Segmentation (vowel-initial versus consonant-initial, i.e., Experiment 2a versus 2b), $F(1, 30)=.01, p=.92, \eta^2<.01$, and no significant main effect of Segmentation, $F(1, 30)=1.49, p=.231, \eta^2<.001$. Therefore, the 24-month-olds in both experiments showed the same response, for both the correct vowel-initial form (Experiment 2a) and incorrect consonant-initial form (Experiment 2b).
is, both mechanisms (i.e., sub-syllabic statistics and the onset bias) were used, causing ambiguity in infants’ interpretation of the pseudo-nouns in liaison.

[Insert Figure 2 here]

To further compare the responses of younger and older infants, we conducted a repeated-measure ANOVA on the data of all four experiments, with Test Item as the within-subject factor, and Segmentation and Age (20 months versus 24 months) as the between-subject factors. The Test Item x Age interaction was significant, $F(1,60)=17.66, p<.001, \eta^2=.207$. Thus, younger and older infants differed in their responses. There was no other main effect or interaction.

**General discussion**

Statistical learning has been shown to be powerful for early cognitive development, not only for the learning of linguistic patterns, but also for other cognitive processes such as abstract rule formation for visual and tone sequences (S. P. Johnson, et al., 2009; Saffran, Johnson, Aslin, & Newport, 1999). One important question concerns whether certain constraints may limit the power of statistical learning. This study concerns whether statistical word segmentation is constrained by the onset bias, which disfavor syllables without a consonant onset (Kager, 1999). We examined this question by testing infants’ segmentation of French liaison cases.

Our experiments showed that vowel-initial words in liaison are difficult to segment for infants despite statistical cues that support the segmentation. In contrast to consonant-initial words, which begin to be segmented from 6 months of age, infants in our experiments failed to segment vowel-initial words even at 20 months. Infants were instead guided by the onset bias, which led to consonant-initial mis-segmentation. The bias persisted even at 24 months of age, as shown in Experiment 2b. This consonant-initial mis-segmentation can explain French-learning children’s early production errors such as *néléphant* (mis-segmenting un éléphant) (e.g., Dugua, Chevrot, & Fayol, 2006). Thus, early in acquisition the surfaced underlying coda in liaison is interpreted as the onset of the following word, which preserves the syllabic integrity. This bias is consistent with the constraint proposed in adult perception
theories (Content, Dumay & Frauenfelder, 2000; Norris, McQueen, Cutler & Butterfield, 1997) and with the universal onset constraint formulated in Optimality Theory (Kager, 1999).

In our familiarization sentences the boundary of the vowel-initial words was indicated by variable liaison contexts. The statistical cues were on sub-syllabic statistical units. However, the onset bias supported a different word boundary. Given such conflicting cues, our 20-month-olds relied on the onset bias for segmentation and failed to use the statistical cues. In previous infant statistical segmentation studies, statistical cues were always on syllabic units with an onset, and infants as young as 6 months of age succeeded in word segmentation (e.g., E. K. Johnson & Tyler, 2010). Therefore, statistical computation for word segmentation appears to be constrained by the learning system to favor syllabic units over sub-syllabic ones. The responses of our 20-month-olds suggest that the onset bias is an early processing mechanism that plays a dominant role during initial acquisition.

Our 20-month-olds’ results differ from the finding that English-learning 16-month-olds can segment vowel-initial words in a non-liaison context (e.g., ice from cold ice) (Mattys & Jusczyk, 2001). Note that in Mattys and Jusczyk infants aged 8-16 months did not mis-segment the vowel-initial words as consonant-initial, suggesting that they were sensitive to the acoustical cues marking the vowel-initial words in non-liaison cases. Our 20-month-olds, however, mis-segmented the target words as consonant-initial, indicating that there was no reliable acoustic cue in liaison that could assist vowel-initial word segmentation.

The 24-month-old infants in our experiments were able to use sub-syllabic statistical cues to segment vowel-initial words in liaison. The statistics-based segmentation can also explain previous findings that young French-learning children sometimes produce isolated vowel-initial forms (e.g., ours, viens ici!) or omission errors such as *[de.uRs] (des ours). Chevrot, Dugua, and Fayol (2009) proposed that children first segment and store vowel-initial words in liaison by hearing these words in isolation or after a vowel-ending word (e.g., joli arbre, “pretty tree”). However, such non-resyllabified vowel-initial cases occur rarely in French, as nouns are mostly preceded directly by determiners¹, which always result in resyllabification. Our experiments demonstrate that children do not have to rely on such rare cases.
Rather, vowel-initial words can be segmented based on variable liaison contexts (i.e., sub-syllabic statistical cues).

For our 24-month-olds, both consonant-initial and vowel-initial forms (e.g., *zonche, onche*) were acceptable. This was unlike the younger infants, who only accepted the consonant-initial interpretation (e.g., *zonche*). This suggests that following an initial stage of onset-biased mis-segmentation for liaison, children go through an intermediate stage of segmenting both consonant-initial and vowel-initial forms, before eventually learning to correctly segment vowel-initial words in liaison (as adults do). The intermediate stage involves two competing segmentation mechanisms, a persisting universal onset bias and an emerging ability to use sub-syllabic statistical cues.

In our study 20-month-olds showed a novelty preference in the consonant-initial condition (Experiment 1b), whereas 24-month-olds showed a familiarity preference in both the vowel-initial and consonant-initial conditions (Experiments 2a and 2b). Novelty preference and familiarity preference are both observed in studies using preferential looking procedures. Both directions are interpreted as evidence for successful processing since neither preference by a group of participants can be due to chance responses. The direction of preference is believed to be related to task difficulty and infants’ abilities/knowledge (e.g., Cyr & Shi, 2013; Hunter & Ames, 1988; Thiessen & Saffran, 2003). For example, in the word segmentation study of Thiessen and Saffran (2003) using artificial language stimuli, 7-month-olds showed a novelty preference and 9-month-olds a familiarity preference. The authors interpreted the novelty preference in the younger infants as indicating a robust ability to use statistics and the familiarity preference in the older infants as indicating an emerging ability to use stress. That is, attending to the additional stress cues was more difficult for infants. In Seidl and Johnson (2008) natural speech was used for testing word segmentation, and English-learning 11-month-olds showed a novelty preference for vowel-initial target words in the sentence-final position, but a familiarity preference for vowel-initial targets in the sentence-initial position. The authors interpreted the results as suggesting that sentence-final words are easier to segment than sentence-initial words. Our results are coherent with the idea of task difficulty. At 20 months of age infants showed a novelty preference for consonant-initial
segmentation. Given that infants begin to segment consonant-initial words from 6 months of age (e.g., E. K. Johnson & Tyler, 2010; Jusczyk & Aslin, 1995), it is unsurprising that consonant-initial segmentation was easy for 20-month-olds. In contrast, the segmentation of vowel-initial words is harder (e.g., Mattys & Jusczyk, 2001; Nazzi, et al., 2005; Seidl & Johnson, 2008), especially in liaison cases. Accordingly, our 24-month-olds showed a familiarity preference, suggesting an emerging ability. Moreover, the 20-month-olds were only attending to one cue (the onset cue), whereas the 24-month-olds were attending to both cues (statistical and onset cues). The attention to additional cues made the task harder for older infants, so it is reasonable that they would also show a familiarity preference in the consonant-initial condition.

Our study can be considered in the larger context of the debate regarding the nature of children’s phonological representations during lexical development. One view holds that children’s early words are holistic in their phonological representation such that both correct and phonemically deviant forms of a word are accepted (e.g., Metsala & Walley, 1998; Storkel, 2002). Researchers who oppose this view maintain that children’s lexical representation is phonetically well specified (e.g., Bailey & Plunkett, 2002; Swingley & Aslin, 2000). Data supporting both views have been reported in the literature. For example, in Hallé and de Boysson-Bardies (1996) infants recognized familiar words when the consonant onset was correctly pronounced (e.g., bonjour in French) and mispronounced (e.g., vonjou). More recent studies (e.g., Mani & Plunkett, 2007; Nazzi, 2005; Swingley, 2005; White & Morgan, 2008) showed that a mispronounced segment impedes word recognition, which may suggest detailed specification. However, in comprehension tasks infants still accept mispronounced forms of a familiar word even though their recognition for such forms is poorer than for the correct pronunciations. For instance, in White and Morgan (2008) infants looked at a picture of a cup more when its label was correctly pronounced (cup) than when it was mispronounced (tup), but they still looked at the cup picture above chance (but not another novel object) upon hearing tup. The data on children’s lexical processing in liaison are sparse, but generally consistent. French-learning young children produce multiple forms for the same word (e.g., Chevrot & Fayol, 2001: norage, zorage), and even both consonant- and vowel-initial forms (Chevrot et al., 2009; Dugua, Chevrot & Fayol, 2006: éléphant, zéléphant). Our results are consistent with this
production pattern. In our experiments the familiarization phase presented the target pseudo-words in the context of multiple liaison consonants. Thus, the results of consonant-initial segmentation suggest that the 20- and 24-month-olds likely accepted multiple word forms with different onset consonants (e.g., *nonche*, *zonche*, *tonche*, *ronche*). By 24 months, they learn to also accept the vowel-initial form (e.g., *onche*). We expect that with increasing ability to use input cues to counter the onset bias, children should ultimately accept only the vowel-initial form and abandon the consonant-initial variants, as would be the case according to the adult French grammar. This trajectory of development is consistent with proposals in theoretical phonology concerning the course of acquisition of liaison and the related lexical representation (e.g., Côté, 2005). Our study suggests that children’s acceptance of multiple forms for a word during early acquisition is related to phonemic variation (e.g., due to resyllabification) as well as their use of competing segmentation mechanisms.

In sum, using perceptual experiments we revealed the development of liaison segmentation in infants. Furthermore, our experiments enabled us to better understand the mechanisms underlying early word segmentation in general. We found that 20-month-old infants come to the segmentation task with the onset bias, a constraint that guides the earliest stages of learning. When statistical cues for segmentation are in agreement with this bias (i.e., statistical cues on syllabic units rather than sub-syllabic onsetless units), word segmentation is optimal, and even achieved by 6-month-olds (e.g., E. K. Johnson & Tyler, 2010). When statistical cues violate this constraint, onset-biased segmentation dominates, and a significant period of learning is needed before children learn that the onset bias is violable by sub-syllabic statistical cues. Our results thus demonstrated that statistical learning is subject to certain constraints. This study offers us a better understanding of the interaction of universal biases and acquired abilities during early development.
References


Acknowledgement

This research was supported by SSHRC and NSERC grants to the second author. We thank all the parents and toddlers who participated in this study.
Footnotes

1 We analyzed the speech of 7 Quebec-French-speaking mothers to their 8-month-old infants in our lab corpus (Cécyre & Shi, 2005; total word token frequency 9547). In particular, we examined the types of words that immediately preceded vowel-initial nouns in those mothers’ speech, and found that determiners were the most frequent (determiner 135, adjectives 32, numerals 6, prepositions 2, verbs 2; nothing preceding 3). Note that nouns without any preceding word were rare. Importantly, all the prenominal words were consonant-ending or liaison-causing, which would yield resyllabification with the following vowel-initial word.
Table 1

Stimuli and design for Experiment 1a & 2a

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Familiarization</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>J’ai trouvé mon onche sur le pavillon.</td>
<td>Il y avait mon èque sur l’étagère.</td>
<td></td>
</tr>
<tr>
<td>“I found my ‘onche’ on the pavilion.”</td>
<td>“There was my ‘èque’ on the shelf.”</td>
<td></td>
</tr>
<tr>
<td>Il y avait couramment un onche.</td>
<td>On ne réussit guère un èque.</td>
<td></td>
</tr>
<tr>
<td>“There was commonly an ‘onche’.”</td>
<td>“We do not hardly succeed an ‘èque’.”</td>
<td></td>
</tr>
<tr>
<td>C’était un petit onche qui était ravi.</td>
<td>Voici un petit èque que je trouve fou.</td>
<td></td>
</tr>
<tr>
<td>“It was a little ‘onche’ that was delighted.”</td>
<td>“Here is a little ‘èque’ that I consider crazy.”</td>
<td></td>
</tr>
<tr>
<td>Je suis camouflé au grand onche.</td>
<td>Je t’ai fréquenté au grand èque.</td>
<td></td>
</tr>
<tr>
<td>“I am camouflaged at the tall ‘onche’.”</td>
<td>“I have frequented you at the tall ‘èque’.”</td>
<td></td>
</tr>
<tr>
<td>Ça n’englobe pas ces onches.</td>
<td>J’avais déniché ces èques.</td>
<td></td>
</tr>
<tr>
<td>“That does not include these ‘onches’.”</td>
<td>“I have unearthed these ‘èques’.”</td>
<td></td>
</tr>
<tr>
<td>Ces gros onches sont de sales voyous.</td>
<td>Ces gros èques le chagrinent beaucoup.</td>
<td></td>
</tr>
<tr>
<td>“These big ‘onches’ are dirty rascal.”</td>
<td>“These big ‘èques’ sadden him greatly.”</td>
<td></td>
</tr>
<tr>
<td>Ce premier onche veut s’enfuir.</td>
<td>Le premier èque va rugir.</td>
<td></td>
</tr>
<tr>
<td>“This first ‘onche’ wants to decamp.”</td>
<td>“The first ‘èque’ will roar.”</td>
<td></td>
</tr>
<tr>
<td>Il a discuté du dernier onche.</td>
<td>Voilà la plainte du dernier èque.</td>
<td></td>
</tr>
<tr>
<td>“He has discussed this last ‘onche’.”</td>
<td>“Here comes the moan of the last ‘èque’.”</td>
<td></td>
</tr>
</tbody>
</table>

Test trial types

| onche (familiarized) versus onche (non-familiarized) | èque (non-familiarized) versus èque (familiarized) |

Note. Isolated tokens of the two target words were presented in alternating test trials. Under each group the first test trial was either onche or èque, counterbalanced across infants.
Table 2  

<table>
<thead>
<tr>
<th>Stimuli and design for Experiment 1b &amp; 2b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Familiarization</strong></td>
</tr>
<tr>
<td>Group 1</td>
</tr>
<tr>
<td>Same as the stimuli in Experiment 1a and 2a</td>
</tr>
<tr>
<td>Group 2</td>
</tr>
<tr>
<td>Same as the stimuli in Experiment 1a and 2a</td>
</tr>
<tr>
<td><strong>Test trial types</strong></td>
</tr>
<tr>
<td>\textit{zonche} (familiarized) versus \textit{zonche} (non-familiarized)</td>
</tr>
<tr>
<td>\textit{zèque} (non-familiarized)</td>
</tr>
<tr>
<td>\textit{zèque} (familiarized)</td>
</tr>
</tbody>
</table>

\textit{Note.} Isolated tokens of the two target words were presented in alternating test trials. Under each group the first test trial was either \textit{zonche} or \textit{zèques}, counterbalanced across infants.
Figure 1. Results of Experiment 1a and Experiment 1b

Figure 1. 20-month-olds’ mean looking times (and standard errors) to the familiarized and the non-familiarized test trials. Test words were vowel-initial in Experiment 1a and consonant-initial in Experiment 1b.
Figure 2. Results of Experiment 2a and Experiment 2b

![Graph showing mean looking times](image)

- **familiarized**
- **non-familiarized**

*Figure 2.* 24-month-olds’ mean looking times (and standard errors) to the familiarized and the non-familiarized test trials. Test words were vowel-initial in Experiment 2a and consonant-initial in Experiment 2b.