Language and Cognition **10** (2018), 391–407. doi:10.1017/langcog.2018.5

Largest-chunk strategy for syllable-based segmentation

LÁSZLÓ DRIENKÓ*

(*Received* 03 *November* 2017 – *Revised* 23 *February* 2018 – *Accepted* 26 *February* 2018 – *First published online* 17 *May* 2018)

ABSTRACT

We apply the largest-chunk segmentation algorithm to texts consisting of syllables as smallest units. The algorithm was proposed in Drienkó (2016, 2017a), where it was used for texts considered to have letters/characters as smallest units. The present study investigates whether the largest chunk segmentation strategy can result in higher precision of boundary inference when syllables are processed rather than characters. The algorithm looks for subsequent largest chunks that occur at least twice in the text, where text means a single sequence of characters, without punctuation or spaces. The results are quantified in terms of four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability. We segment CHILDES texts in four languages: English, Hungarian, Mandarin, and Spanish. The data suggest that syllable-based segmentation enhances inference precision. Thus, our experiments (i) provide further support for the possible role of a cognitive largest-chunk segmentation strategy, and (ii) point to the syllable as a more optimal unit for segmentation than the letter/phoneme/character, (iii) in a cross-linguistic context.

KEYWORDS: cognitive/computer modelling, segmentation, language acquisition.

1. Introduction

The problem of how to segment continuous speech into components dates back at least to Harris (1955). Harris used "successor frequencies", i.e., statistics, to predict boundaries between linguistic units, ideally morphemes. Saffran, Aslin, and Newport (1996) used syllable-based artificial languages to demonstrate that statistical information is indeed available for infants acquiring language. Results in language acquisition research indicate that speech segmentation is affected by various lexical and sublexical linguistic cues (see, e.g., Mattys, White, & Melhorn, 2005). Such cues can readily offer

[*] Address for correspondence: e-mail: adadad@freemail.hu

themselves as the base for various cognitive segmentation strategies. The distribution of strong and weak syllables, for instance, may help the language learner to use a metrical segmentation strategy (Cutler & Carter, 1987; Cutler & Norris, 1988), infants can possibly learn to use stress patterns for segmentation (Thiessen & Saffran, 2007), or they can exploit prosodic cues like lengthening, or rise in fundamental frequency of speech sounds (Bagou, Fougeron, & Frauenfelder, 2002).

Despite the diverse details that are known about the segmentation process (see Sonderegger, 2008, for a review), the question concerning the basic unit of segmentation is still open. Although the linguistic or psycholinguistic status of the syllable is rather complex (e.g., Bell & Hooper, 1978; Cholin, 2011; Livingstone, 2014) and a generally accepted precise definition is still lacking, it is widely assumed that an infant's language acquisition is based on syllables (e.g., Mehler, Dupoux, & Segui, 1990; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Eimas, 1997). Syllable-based segmentation seems to be relevant for artificial languages (Saffran et al., 1996), and for writing skills (Liberman, Shankweiler, Fischer, & Carter, 1974) as well.

Drienkó (2016) proposed an algorithm for inferring boundaries of utterance fragments in relatively small unsegmented texts. The algorithm looks for subsequent largest chunks that occur at least twice in the text. The results were interpreted in terms of four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability. In Drienkó (2017a) the largest-chunk algorithm was used cross-linguistically to segment CHILDES utterances in four languages: English, Hungarian, Mandarin, and Spanish. The author found an Inference Precision range of 53.5%–65.6%, which grew when segments of specified lengths were merged. The unit for segmentation was the letter, i.e., the computer character, which can be regarded as a rough written equivalent of the speech sound. The advantage of the Largest Chunk method over other proposed segmentation strategies is that it allows direct quantitative results based solely on the linguistic structure of the given text without needing further cues like stress or metrical features. The strategy is in line with Peters' (1983) approach to language acquisition, where the learner uses various cognitive heuristics to extract large chunks from the speech stream and the 'ultimate' units of language are formed by segmenting and fusing the relevant chunks.

The present study investigates whether the largest-chunk segmentation strategy can result in higher precision of boundary inference when syllables rather than characters are processed.[1] We do not distinguish between word or utterance boundaries. For the sake of direct comparison, we use the same data

[1] Preliminary results were communicated through Drienkó (2017b).

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

as Drienkó (2017a), i.e., CHILDES texts in four languages: English (Anne, Manchester corpus; Theakston, Lieven, Pine, & Rowland, 2001), Hungarian (Miki, Réger corpus; Réger, 1986; Babarczy, 2006), Mandarin Chinese (Beijing corpus; Tardif, 1993, 1996), and Spanish (Koki, Montes corpus; Montes, 1987, 1992). Additionally, we segment two chapters from *Gulliver's Travels* by Jonathan Swift in order to possibly detect text size effects. The length range of texts is 1,743–10,574 syllables, 5,499–43,433 characters.

After a short description of the algorithm in Section 2, we present our results in Section 3. This will be followed by a discussion and some conclusions in Sections 4 and 5, respectively.

2. Description of the algorithm

Our algorithm is basically identical with that of Drienkó (2016, 2017a) except that there was an additional MERGE component included in that work. The basic, CHUNKER, module of the algorithm looks for subsequent largest syllable sequences that occur more than once in the text. Starting from the first syllable, it concatenates the subsequent syllables, and if a resultant string s_i occurs in the text only once, a boundary is inserted before its last syllable since the previous string, s_{i-1}, is the largest re-occurring one of the i strings. Thus the first boundary corresponds to s_{i-1}, our first tentative speech fragment. The search for the next fragment continues from the position after the last character of s_{i-1}, and so on.

The EVALUATE module computes four precision metrics: Inference Precision, Alignment Precision, Redundancy, and Boundary Variability. Inference Precision (IP) represents the proportion of correctly inferred boundaries (cib) to all inferred boundaries (aib), i.e., IP = cib / aib. The maximum value of IP is 1, even if more boundaries are inferred than all the correct (original) boundaries (acb). Redundancy (R) is computed as the proportion of all the inferred boundaries to all the correct (original) boundaries, i.e., R = aib / acb. R is 1 if as many boundaries are inferred as there are boundaries in the original text, i.e., aib = acb; R is less than 1 if fewer boundaries are inferred than acb; and R is greater than 1 if more boundaries are inferred than optimal. Alignment Precision (AP) is specified as the proportion of correctly inferred boundaries to all the original boundaries, i.e., AP = cib / acb. Naturally, the maximum value for AP is 1. Boundary Variability (BV) designates the average distance (in characters) of an inferred boundary from the nearest correct boundary, i.e., BV = (Σdf_i) / aib. The above measures are not totally independent, since Inference Precision × Redundancy = Alignment Precision, but emphasise different aspects of the segmentation mechanism. Obviously, IP = AP for R = 1. The Largest-Chunk (LCh) segmentation algorithm is outlined in Table 1.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 1. *The Largest-Chunk Segmentation Algorithm*

1. CHUNKER
   input: segmented text *T*, unsegmented sequence *UST* of linguistic symbols (characters) of text T
   For each syllable position p in UST
   {fragment_candidate=""
   while the occurrence of fragment_candidate in UST is > 1
   {fragment_candidate = fragment_candidate + syllable_at_p}
   fragment_candidate → FRAGMENTS
   p → ALL INFERRED BOUNDARIES: p=p+1 }
2. EVALUATE
   For all words *w* in *T*
   {boundary_position of *w* →ALL CORRECT BOUNDARIES
   }
   acb= the number of all correct boundaries
   For all boundaries in ALL INFERRED BOUNDARIES and ALL CORRECT BOUNDARIES
   {
   compute the number of correctly inferred boundaries: cib
   compute boundary_variability, i.e. the average distance (in characters) of an inferred boundary from the nearest correct boundary: bv $=\Sigma df_i/aib$
   }
   Compute:
   Inference Precision: cib / aib
   Redundancy: aib / acb
   Alignment Precision: cib / acb

For some immediate insight, (1) illustrates how the arrangement of the individual elements in a sequence affect largest-chunk segmentation. Spaces correspond to inferred boundaries. Letters *a*, *b*, *c*, *d*, *e*, and *f* can also be seen as symbolising syllables.

(1) a) abcabc → abc abc
    b) abcab → ab c ab
    c) abc → a b c
    d) abccba → a b c c b a
    e) abcdefefcdab → ab cd ef ef cd ab
    f) abcdefabcdef → abcdef abcdef

In (1b), for instance, the algorithm starts from the first *a* element, detects that *a* occurs twice in the sequence *abcab*, so takes the next element, *b*, detects that the corresponding segment, *ab*, occurs twice, proceeds to consider segment *abc*, detects that it occurs only once and infers a boundary after *ab*, the first largest 'chunk'. Segmentation then continues from element *c*. Since *c* has only a single occurrence, a boundary should be inferred before it. However, this boundary has already been detected, so nothing happens and

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

segmentation continues from the next position, *a*. Again, dual occurrence of *a* is detected, so segment *ab* is considered. As *ab* occurs twice, the algorithm should step to the next element. However, since *b* is the last element of the sequence, a boundary is inferred after it. Note that the inference of the last boundary is actually independent of the number of occurrences of the last segment. When an element occurs only once, a boundary is inserted before it and processing continues from the position immediately after it. As a result, any single-occurrence element is treated as a potential meaningful unit. An extreme example is given in (1c), where each element occurs only once. Recall that the LCh algorithm looks for largest RE-OCCURRING sequences, so single-occurrence units constitute a specific case. Arguably, regarding a single-occurrence sequence as a succession of its single-occurrence elements, rather than as a chunk itself, has the practical advantage of not missing any true boundary, and, perhaps more theoretically, it reflects our assumption that the 'atomic' segmentation elements are somehow – explicitly or implicitly – known to the segmenter.

Examples (1d–e) demonstrate a special case where symmetrical arrangement can effect 'minimal largest chunking'. Each element in (1d) occurs twice, but there is no re-occurring combination of at least two elements, so a boundary is inferred at each position. This property of largest chunking can result in optimal segmentation for coupled elements. Suppose we have the 'words' *ab*, *cd*, and *ef*. If the input sequence is such that it contains each word twice, and the arrangement of letters/syllables is symmetrical, as in (1e) – a kind of 'central embedding' – a boundary will be inferred precisely after each word. In contrast, the largest chunks of (1f) conflate the three words.

To see how precision values are calculated, consider two mini-sets of utterances, {*baby is, baby it*} and {*what about, what a boot*}. We provide a character-based analysis here, as summarised in Table 2, which will be contrasted with the syllable-based case in Section 4.

The *baby is baby it* text contains four word boundaries, thus $acb = 4$. The Largest-Chunk algorithm infers four boundaries corresponding to segments *babyi*, *s*, *babyi*, and *t*, which entails that $aib = 4$. Two of the four inferred boundaries are correct, $cib = 2$, resulting in Inference Precision $IP = cib / aib = 2 / 4 = 0.5$ and Alignment Precision $AP = cib / acb = 2 / 4 = 0.5$. Since the number of the inferred boundaries equals the number of the original boundaries, $aib = acb$, Redundancy is 1. The second and the fourth boundaries are correct, so their distance from the respective correct boundaries is zero, i.e., $df_2 = df_4 = 0$. If we shift the first inferred boundary one character to the left, we reach the first correct boundary, following *baby*. If we shift the first boundary one character to the right, we reach the second correct boundary, following *is*. Clearly, then, $df_1 = 1$. Similarly, $df_3 = 1$ as

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 2. *Calculating precision values (characters)*

baby is baby it – 4 boundaries, acb = 4
babyisbabyit →babyi s babyi t
2 correct of 4 inferred boundaries: cib=2, aib=4,
IP=cib/aib=2/4=0.5
2 correctly identified boundaries: AP = cib /acb =2/4=0.5
R=aib/acb=4/4=1
BV=(1+0+1+0)/4=0.5

what about what a boot – 5 boundaries, acb = 5
whataboutwhataboot → whatabo u t whatabo o t
2 correct of 6 inferred boundaries: cib=2, aib=6,
IP= cib/aib=2/6= 0.33
2 correctly identified boundaries: AP= cib /acb =2/5=0.4
R=aib/acb =6/5=1.2 (>1)
BV=(2+1+0 +2+1+0)/6=1

well, since by shifting the third inferred boundary one character either to the left or to the right, we reach the third or the fourth correct boundary, respectively. We compute Boundary Variability as $BV = (df_1 + df_2 + df_3 + df_4) / aib = (1 + 0 + 1 + 0) / 4 = 0.5$. Note that, when the distance of an inferred boundary is different for the left-side correct boundary and the right-side correct boundary, the shorter distance is chosen for *df*. Thus, for example, $df_1 = 2$ for the *what about what a boot* text because the first inferred boundary, corresponding to *whatabo* is three characters away from the first correct left-side boundary, which follows *what*, and two characters away from the first correct right-side boundary, following *about*, so the right-side distance is chosen.

3. The experiments

In our experiments we used data from the CHILDES database (MacWhinney, 2000). All files were converted to simple text format, annotations were removed together with punctuation symbols and spaces. Hyphens were inserted after each syllable, so syllable, word, and utterance boundaries were indicated by hyphens. The segmentation problem consisted in differentiating word or utterance boundaries from word-medial syllable boundaries. Mother and child utterances were not separated, so the dataset for each language constituted an unsegmented (written) stream of 'mother–child language' represented as a single sequence of characters. The length range of the CHILDES texts was 1,743–9,021 syllables, 5,499–40,864 characters. Segmentation into syllables was done with the help of the Lyric Hyphenator (Juicio Brennan <http://juiciobrennan.com/hyphenator/>) for English, manually by the author for

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

Hungarian, and the Spanish syllables were produced by the MARELLO.ORG syllabifier (<https://marello.org/tools/syllabifier/>). In the case of Chinese, syllable boundaries were understood as indicated by tone-marking numbers (1 to 4) and spaces in the pinyin transcript, so boundaries were inserted accordingly.

3.1. EXPERIMENT 1 – ENGLISH

In this experiment the first Anne file, *anne01a.xml*, of the Manchester corpus (Theakston et al., 2001) was analysed. The original text consisted of 374 utterances, 1,826 word tokens (acb), and 2,100 syllables. The unsegmented version of the text consisted of 8,899 characters. The segmentation algorithm inserted 1,133 boundaries (aib), of which 1,072 were correct (cib), thus Inference Precision = cib / aib = 0.946. The other precision values were as follows: Redundancy = 0.62, Alignment Precision = 0.59, Boundary Variability = 0.19. Table 3 contrasts the precision values with those obtained in Drienkó (2017a), where the character was regarded as the primary segmentation unit. The data reveal that both Inference Precision and Alignment Precision are considerably higher for syllables, along with almost identical Redundancy values. The IP value approaching 1 (i.e., cib ≈ aib) entails that Redundancy (R = aib / acb) and Alignment Precision (AP = cib / acb) converge (0.62 vs. 0.59). The reduction of the Boundary Variability value indicates that the inferred boundaries are even closer to the correct ones in the case of syllables: on average, for an inferred boundary a correct boundary can be found within the distance of about 0.19 characters.

3.2. EXPERIMENT 2 – HUNGARIAN

The Hungarian data used in this experiment correspond with the *miki01.xml* file of the Réger corpus (Réger, 1986; Babarczy, 2006). The original text consisted of 589 utterances, 1,541 word tokens (acb), and 2,527 syllables. The unsegmented version of the text consisted of 9,358 characters. The segmentation algorithm inserted 1,324 boundaries (aib), of which 1,020 were correct (cib). The precision values were as follows: Inference Precision = cib / aib = 0.77, Redundancy = 0.86, Alignment Precision = 0.66, Boundary Variability = 0.87. Table 4 contrasts the precision values with those obtained in Drienkó (2017a), where the character was regarded as the primary segmentation unit. The data reveal that both Inference Precision and Alignment Precision are higher for syllables, along with almost identical Redundancy values. Boundary Variability is slightly higher for syllables, although both values remain below 1.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 3. *Precision values for Experiment* 1 (*Anne*)

	IP	R	AP	BV
Characters	0.66	0.62	0.41	0.53
Syllables	0.95	0.62	0.59	0.19

TABLE 4. *Precision values for Experiment* 2 (*Miki*)

	IP	R	AP	BV
Characters	0.53	0.82	0.44	0.85
Syllables	0.77	0.86	0.66	0.87

3.3. EXPERIMENT 3 – MANDARIN CHINESE

In this experiment we segmented Mandarin Chinese text included as *bb1.xml* in the Beijing corpus (Tardif, 1993, 1996). The file contains the pinyin transcription of the utterances. The original text consisted of 2,118 utterances, 7,064 word tokens (acb), and 9,021 syllables. The unsegmented version of the text consisted of 40,864 characters. The segmentation algorithm inserted 4,636 boundaries (aib), of which 4,271 were correct (cib). The precision values were the following: Inference Precision = 0.92, Redundancy = 0.66, Alignment Precision = 0.60, Boundary Variability = 0.34. Table 5 contrasts the precision values with the character-based results. It can be seen that both Inference Precision and Alignment Precision are higher for syllables, whereas Redundancy values are nearly the same. Boundary Variability is reduced by almost 50% for syllables.

3.4. EXPERIMENT 4 – SPANISH

The Spanish data for this segmentation experiment came from the Koki material contained in the *01jul80.cha* file of the Montes corpus (Montes, 1987, 1992). The original text consisted of 398 utterances, 957 word tokens (acb), and 1,743 syllables. The unsegmented version of the text consisted of 5,499 characters. The segmentation algorithm inserted 641 boundaries (aib), of which 521 were correct (cib). The precision values were the following: Inference Precision = 0.81, Redundancy = 0.67, Alignment Precision = 0.54, Boundary Variability = 0.54. Table 6 contrasts the precision values with the character-based results. It can be seen that both Inference Precision and Alignment Precision are higher for syllables, whereas Redundancy values are nearly the same. Boundary Variability is slightly higher for syllables.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 5. *Precision values for Experiment* 3 (*Beijing*)

	IP	R	AP	BV
Characters	0.6	0.62	0.37	0.65
Syllables	0.92	0.66	0.60	0.34

TABLE 6. *Precision values for Experiment* 4 (*Koki*)

	IP	R	AP	BV
Characters	0.64	0.65	0.42	0.51
Syllables	0.81	0.67	0.54	0.54

3.5. EXPERIMENT 5 – GULLIVER

Switching from letters to syllables naturally reduces the number of linguistic units. For instance, the 8,899 characters of the English text in Experiment 1 represented 2,100 syllables. To have some insight on how the change in the number of processing units might affect segmentation precision in the case of the same language, we analysed an English text whose number of syllables is comparable to the number of characters in the Anne text of Experiment 1. We chose Chapters 1 and 2 from *Gulliver's Travels* by Jonathan Swift. The two chapters were merged into a single text containing 7,765 word tokens, 238 utterances, and 10,574 syllables.[2] The unsegmented version of the text consisted of 43,433 characters. The segmentation algorithm inserted 6,078 boundaries (aib), of which 5,125 were correct (cib). The precision values were the following: Inference Precision = 0.84, Redundancy = 0.78, Alignment Precision = 0.66, Boundary Variability = 0.62. Table 7 contrasts the precision values with the character-based results. It can be seen that both Inference Precision and Alignment Precision are higher for syllables, whereas Redundancy values are nearly the same. Boundary Variability is lower for syllables. The quantitative results from all the experiments are summarised in Figure 1.

4. Discussion

Our segmentation experiments allow the following observations:

1. Inference Precision is higher for syllables.
2. Redundancy is almost the same.
3. Alignment Precision is also higher for syllables.

[2] The text contained 31 'Lilliputian' word tokens, which were not hyphenated.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 7. *Precision values for Experiment* 5 (*Gulliver*)

	IP	R	AP	BV
Characters	0.53	0.75	0.4	0.8
Syllables	0.84	0.78	0.66	0.62

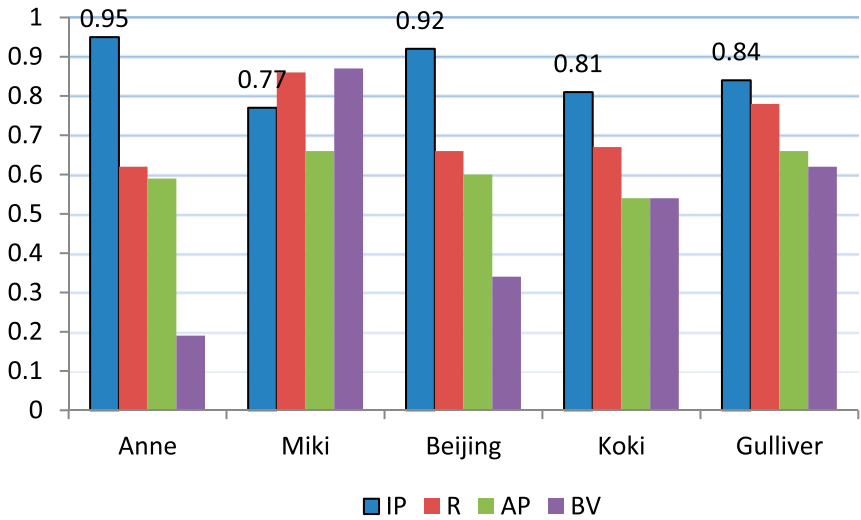

Fig. 1. Precision values for all the texts used in the segmentation experiments. (IP: Inference Precision; R: Redundancy; AP: Alignment Precision; BV: Boundary Variability)

4. When measured in characters, Boundary Variability can be both higher and lower for syllables – on average, it is lower – but the values stay below 1.

Let us consider these observations in more detail below.

1. The Inference Precision values are remarkably higher for syllables. The 53%–66% IP value range, averaging 59%, for characters rose to 77%–95%, averaging 86%, in the case of syllables (cf. Table 8).
2. Redundancy is almost the same for characters and for syllables – although slightly higher for syllables (cf. Table 9). The average R value is 3% higher in the case of syllables. This means that switching to syllables as basic segmentation units does not notably change the proportion of inferred boundaries. It is perhaps worth noting that the R values stay below 1, i.e., fewer boundaries are inserted than would be required by the original segmentation, by the number of words in the original texts.
3. Alignment Precision is higher for syllables (cf. Table 10). The 41% average of AP values for letters became 61% in the case of syllables.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 8. *IP values across texts*

	Anne	Miki	Beijing	Koki	Gulliver	Average
IP (Characters)	0.66	0.53	0.6	0.64	0.53	0.59
IP (Syllables)	0.95	0.77	0.92	0.81	0.84	0.86

TABLE 9. *R values across texts*

	Anne	Miki	Beijing	Koki	Gulliver	Average
R (Characters)	0.62	0.82	0.62	0.65	0.75	0.69
R (Syllables)	0.62	0.86	0.66	0.67	0.78	0.72

TABLE 10. *AP values across texts*

	Anne	Miki	Beijing	Koki	Gulliver	Average
AP (Characters)	0.41	0.44	0.37	0.42	0.4	0.41
AP (Syllables)	0.59	0.66	0.60	0.54	0.66	0.61

This is the consequence of the relation AP = IP × R and the fact that R is about the same for characters and syllables. If IP is greater for syllables, then multiplying by about the same R entails that AP will be greater as well. In other words, if a larger proportion of the inferred word boundaries is correct for syllables than for letters, then a larger proportion of the original boundaries will be detected correctly for syllables if the same percentage of boundaries is inserted as for letters.

4. The Boundary Variability values do not show a consistent pattern. They can be both higher and lower for syllables – on average, they are 16% lower (cf. Table 11). However, the values stay below 1, which means that for any inferred boundary a correct boundary can be found within the average distance of less than one character, i.e., a correct boundary can be reached by shifting the boundary less than one character, on average, to the left or to the right.

Note that we measured Boundary Variability in characters, not in syllables. We believe that this gives a more precise picture of the segmentation process since syllables can vary in length, and they are composed of letters/phonemes anyway. However, Table 12 displays BV values measured in syllables as well (cf. the $BV(syllables)_s$ row). The data show that BV is much lower when measured in syllables: a correct boundary can be reached by shifting the incorrect boundary 0.148 syllables, on average, to the left or to the right.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 11. *BV values across texts*

	Anne	*Miki*	Beijing	*Koki*	Gulliver	Average
BV (Characters)	0.53	*0.85*	0.65	*0.51*	0.8	0.67
BV (Syllables)	0.19	*0.87*	0.34	*0.54*	0.62	0.51

TABLE 12. *Details for measuring BV in syllables*

	Anne	Miki	Beijing	Koki	Gulliver	Average
BV (syllables)	0.19	0.87	0.34	0.54	0.62	0.51
$BV(syllables)_s$	0.054	0.248	0.079	0.192	0.169	0.148
Average syllable length*	4.2	3.7	4.5	3.15	4.1	3.93
$BV(syllables)_s$ (estimated)	0.045	0.23	0.076	0.17	0.15	0.134

[*] Recall that syllable ends are marked by hyphens in the texts to be segmented, so in reality each syllable is 1 character shorter than in our texts. Naturally, this is also true for the averages in the table.

Table 12 also illustrates that a rough estimate for Boundary Variability in terms of syllables can be calculated by dividing the BV value measured in characters (first row of Table 12) by the average syllable length for the given text (third row). See 'Appendix' for an explanation.

To illustrate the basic information gain effected by the transition to syllables from characters, consider the examples of Table 2 (repeated here as Table 13).

The Largest Chunk segmentation algorithm possibly inserts boundaries syllable-medially as, e.g., *babyi* exemplifies. Such errors are naturally ruled out in syllable-based segmentation: no boundary can be inserted into the smallest unit. As a consequence, we see an increase in precision values. In the case of the *ba-by-is-ba-by-it* text, for example, all boundaries are correct, which amounts to 100% Inference Precision, all the other values being optimal (cf. Table 14).

On the other hand, the syllable-based LCh algorithm still may undesirably infer word-medial syllable boundaries. For instance, the first inferred boundary, after *what-a-* in the *what-a-bout-what-a-boot-* text is incorrect since it divides *a-bout-* into two. Such errors reduce the effectiveness of segmentation. This is demonstrated by the *what-a-bout-what-a-boot-* example where Inference Precision cannot reach 100%, i.e., IP = 0.75. Nevertheless, the 0.75 value constitutes a considerable increase from 0.33 in the character-based case. Tables 15 and 16 illustrate that the change in precision metrics due to switching from characters to syllables is fairly similar for our current examples and for our experiments. Besides the IP values there is an increase in Alignment Precision. The BV values become lower for syllables whether measured in characters or in syllables (values in brackets). Recall that in the experiments BV values became

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 13. *Calculating precision values* (*characters*)

baby is baby it – 4 boundaries, acb = 4
babyisbabyit →babyi s babyi t
2 correct of 4 inferred boundaries: cib=2, aib=4,
IP=cib/aib=2/4=0.5
2 correctly identified boundaries: AP = cib /acb =2/4=0.5
R=aib/acb=4/4=1
BV=(1+0+1+0)/4=0.5

what about what a boot – 5 boundaries, acb = 5
whataboutwhataboot → whatabo u t whatabo o t
2 correct of 6 inferred boundaries: cib=2, aib=6,
IP= cib/aib=2/6= 0.33
2 correctly identified boundaries: AP= cib /acb =2/5=0.4
R=aib/acb =6/5=1.2 (>1)
BV=(2+1+0 +2+1+0)/6=1

TABLE 14. *Calculating precision values* (*syllables*)

baby is baby it – 4 boundaries, acb = 4
ba-by-is-ba-by-it- → ba-by- is- ba-by- it-
4 correct of 4 inferred boundaries: cib=4, aib=4,
IP=cib/aib=1
4 correctly identified boundaries: AP = cib /acb=1
R=aib/acb=4/4=1
BV=(0+0+0+0)/4=0

what about what a boot – 5 boundaries, acb = 5
what-a-bout-what-a-boot- → what-a- bout- what-a- boot-
3 correct of 4 inferred boundaries: cib=3, aib=4,
IP=cib/aib =0.75
3 correctly identified boundaries: AP = cib /acb=3/5=0.6
R=aib/acb=4/5=0.8
BV(ch)=(2+0+0+0)/4=0.5 ('a-' two characters→df_1=2)
BV'(Sy)=(1+0+0+0)/4=0.25 ('a-' one syllable → df_1=1)

unambiguously lower only when they were measured in syllables. Redundancy values are the same for the {*baby is baby it*} text but for the {*what about what a boot*} text they show a decrease, contrary to the experimental results. This may underline, on the one hand, that our toy examples are not capable of capturing all aspects of the segmentation mechanism, and/or, on the other hand, that Redundancy is somehow more independent of the sort of information gain which our examples were designed to visualise.

5. Conclusions

The present study examined how various precision metrics are affected by a transition from characters to syllables in applying the Largest-chunk method to text segmentation. The data show an increase in Inference Precision,

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

TABLE 15. *Precision values for the* 'baby is baby it' *example*

	IP	R	AP	BV
Characters	0.5	1	0.5	0.5
Syllables	1	1	1	0 (0)

TABLE 16. *Precision values for the* 'what about what a boot' *example*

	IP	R	AP	BV
Characters	0.33	1.2	0.4	1
Syllables	0.75	0.8	0.6	0.5 (0.25)

as well as in Alignment Precision. Redundancy is almost the same, while a reduction in Boundary Variability can be observed, which is more unambiguously pronounced when measured in syllables. Nevertheless, BV remains below 1 even when measured in characters. Overall, our quantitative results seem to underline the role of the syllable, as opposed to the letter or speech sound, in text segmentation using the Largest-chunk strategy. Conversely, our results indicate that the strategy might serve as an insightful component for a model of speech segmentation.

We did not attempt to explain the differences in precision values for the different texts. That would be an exciting topic for further research. Clearly, on the one hand, segmentation must be affected by typological differences between languages, but, on the other hand, idiosyncratic parameters of a given text, such as length, genre, register, speaker, etc., may also play a role. Research on infant word segmentation suggests that extraction of target words is facilitated when they are aligned with utterance boundaries (Seidl & Johnson, 2006; Johnson, Seidl, & Tyler, 2014). Such findings would make it reasonable to investigate how LCh segmentation would be affected by information on utterance boundaries.

REFERENCES

Babarczy, A. (2006). The development of negation in Hungarian child language. *Lingua*, **116**, 377–392.

Bagou, O., Fougeron, C., & Frauenfelder, U. H. (2002). Contribution of prosody to the segmentation and storage of 'words' in the acquisition of a new mini-language. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11–13 April.

Bell, A., & Hooper, J. B. (1978). Issues and evidence in syllabic phonology. In A. Bell & J. B. Hooper (Eds.), *Syllables and segments* (pp. 3–22). Amsterdam: North-Holland.

Cholin, J. (2011). Do syllables exist? Psycholinguistic evidence for the retrieval of syllabic units in speech production. In Ch. E. Cairns & E Raimy (Eds.), *Handbook of the syllable* (pp. 225–253). Leiden: Koninklijke Brill NV.

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in English vocabulary. *Computer Speech and Language*, **2**, 133–142.
Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113–121.
Drienkó, L. (2016). Discovering utterance fragment boundaries in small unsegmented texts. In A. Takács, V. Varga, & V. Vincze (Eds.), *XII. Magyar Számítógépes Nyelvészeti Konferencia* [12th Hungarian Computational Linguistics Conference] (pp. 273–281). Szeged: University of Szeged. Online: <http://rgai.inf.u-szeged.hu/mszny2016/>.
Drienkó, L. (2017a). Largest chunks as short text segmentation strategy: a cross-linguistic study. In A. Wallington, A. Foltz, & J. Ryan, (Eds.), *Selected papers from the 6th UK Cognitive Linguistics Conference*, (pp. 273–292).
Drienkó, L. (2017b). Syllable-based largest-chunk segmentation. Poster presentation for the Linguistics Beyond and Within (LingBaW) Conference, 18–19 October 2017, Lublin, Poland.
Eimas, P. D. (1997). Infant speech perception: processing characteristics, representational units, and the learning of words. In Robert L. Goldstone, Phillipe G. Scyhns, & Douglas L. Medin (Eds.), *The psychology of learning and motivation*, vol. 36 (pp. 127–169). London: Academic Press.
Harris, Z. S. (1955). From phoneme to morpheme. *Language*, **31**, 190–222.
Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PLoS ONE*, **9**(1), e83546. Online: <https://doi.org/10.1371/journal.pone.0083546>.
Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, **32**(3), 402–420.
Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, **18**(2), 201–212.
Livingstone, J. (2014). Do syllables exist? *The Guardian* 25 June. Online:<https://www.theguardian.com/education/2014/jun/25/english-do-syllables-exist-linguists>.
MacWhinney, B. (2000). *The CHILDES Project: tools for analyzing talk* (3rd ed.) (Vol. 2): *The database*. Mahwah, NJ: Lawrence Erlbaum Associates.
Mattys, S. L, White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, **134**(4), 477–500.
Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: the onsetof word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 236–262). Cambridge, MA: MIT Press.
Montes, R. G. (1987). Secuencias de clarificación en conversaciones con ninos. *Morphe* **3/4**, Universidad Autónoma de Puebla.
Montes, R. G. (1992). Achieving understanding: repair mechanisms in mother–child conversations. Unpublished doctoral dissertation, Georgetown University.
Peters, A. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.
Réger, Z. (1986). The functions of imitation in child language. *Applied Psycholinguistics*, **7**, 323–352.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**(5294), 1926–1928.
Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, **9**, 565–573.
Sonderegger, M. (2008). Infant word segmentation: a basic review. Online:<http://people.cs.uchicago.edu/~morgan/segReview.pdf>.
Swift, J. (n.d.). *Gulliver's Travels*. The Project Gutenberg eBook. Online:<http://www.gutenberg.org/files/829/829-h/829-h.htm>.
Tardif, T. (1993). Adult-to-child speech and language acquisition in Mandarin Chinese. Unpublished doctoral dissertation, Yale University.

Tardif, T. (1996). Nouns are not always learned before verbs: evidence from Mandarin speakers' early vocabularies. *Developmental Psychology*, **32**, 492–504.
Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, **28**(1), 127–152.
Thiessen, E. D., & Saffran, J. R. (2007) Learning to learn: infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, **3**(1), 73–100.

Appendix

We show that for text T, consisting of L characters, k syllables, Boundary Variability in terms of syllables can be estimated by dividing the BV value measured in characters by the average syllable length for the given text. Recall we calculate BV as in (A1), where $n = aib$ (all inferred boundaries) and df_i denotes the distance of the i-th inferred boundary from the nearest correct boundary.

$$\mathrm{BV} = \frac{1}{n}\sum_{i=1}^{n} df_i \tag{A1}$$

We write Δf for the sum, as in (A2).

$$\mathrm{BV} = \frac{1}{n}\sum_{i=1}^{n} df_i = \frac{1}{n}\Delta f \tag{A2}$$

Suppose Δf contains k_Δ syllables with average syllable length s, i.e., (A3) holds.

$$\Delta f = \mathrm{k}_\Delta \times s, \;\; s = \frac{\Delta f}{k_\Delta} \tag{A3}$$

Assume that the distribution of syllable length in sum Δf is the same as for the whole text T, entailing the equality of average syllable length for Δf with the average syllable length for the whole text, as expressed in (A4).

$$s = \frac{\Delta f}{k_\Delta} = \frac{L}{k} \tag{A4}$$

Writing (A4) for s in (A3) we get (A5), and writing (A5) for Δf in (A2) we get (A6).

$$\Delta f = k_\Delta \times s = k_\Delta \frac{\Delta f}{k_\Delta} = k_\Delta \frac{L}{k} \tag{A5}$$

$$\mathrm{BV} = \frac{1}{n}\Delta f = \frac{1}{n} k_\Delta \frac{\Delta f}{k_\Delta} = \frac{1}{n} k_\Delta \frac{L}{k} \tag{A6}$$

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press

We can approximate BV in syllables if we change the 'length scale' by choosing average syllable length $s = \frac{L}{k}$ as unit length, i.e., by dividing (A6) by s. Thus Equation (A7) can be used for approximating BV', Boundary Variability measured in syllables, when BV is known.

$$\mathrm{BV}' = \frac{1}{n} k_{\Delta} = \frac{BV}{s} \tag{A7}$$

https://doi.org/10.1017/langcog.2018.5 Published online by Cambridge University Press