Units in segmentation: a computational investigation

Çağrı Çöltekin
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

This study investigates the use of syllables and phone(me)s in computational models of segmentation in early language acquisition. We present results of experiments with both syllables and phonemes as the basic unit using a standard state-of-the-art segmentation model. We evaluate the model output based on both word- and morpheme-segmented gold standards on child-directed speech corpora from two typologically different languages. Our results do not indicate a clear advantage for one unit or the other. We argue that the computational advantage for the syllable suggested in earlier research may be an artifact of the particular language and/or segmentation strategy used in these studies.

1 Introduction

Segmentation is a prevalent problem in language processing. We process linguistic input as a combination of linguistic units such as words. However, spoken language does not include reliable cues to word boundaries that are found in many writing systems. The hearer needs to extract words, or lexical units, from a continuous stream of sounds using the information available in the input signal as well as his/her/its (implicit) linguistic knowledge. This makes segmentation a particularly challenging task for the early learners, since they need to discover the lexical units in the input without a lexicon and without much insight into the workings of the input language. The question of how early learners may accomplish this task has been an active area of research.

The problem have been studied extensively, through both psycholinguistic experiments and computational modeling. Experimental studies are mainly focused on particular cues that could help adults or children to solve the segmentation problem. Just to name a few, these cues include predictability statistics (Saffran, Aslin, and Newport, 1996), lexical stress (Cutler and Butterfield, 1992; Jusczyk, Houston, and Newsome, 1999), phonotactics (Jusczyk, Cutler, and Redanz, 1993), allophonic differences (Jusczyk, Hohne, and Bauman, 1999), *vowel harmony* (Kampen et al., 2008; Suomi, McQueen, and Cutler, 1997) and coarticulation (E. K. Johnson and Jusczyk, 2001). Computational models offer a complementary method to the psycholinguistic experiments. There have been an increasing number of computational models of segmentation in the literature, particularly within the last two decades (just to exemplify a few, Elman, 1990, Aslin, 1993, Cairns et al., 1994, Christiansen, Allen, and Seidenberg, 1998, Fleck, 2008, Brent and Cartwright, 1996, Brent, 1999, Venkataraman, 2001, Xanthos, 2004, Goldwater, Griffiths, and M. Johnson, 2009, M. Johnson and Goldwater, 2009, Monaghan and Christiansen, 2010, Çöltekin and Nerbonne, 2014).

In this paper, we investigate a recurring issue in the segmentation literature: the use of syllable or phoneme as the basic input unit in computational models of segmentation.[1] Most psycholinguistic research is based on syllable as the basic unit. The likely reason behind this choice is the early research pointing to syllable as a salient perceptual unit for adults (Cutler, Mehler, et al., 1986; Mehler et al., 1981; Savin and Bever, 1970), and infants (Eimas, 1999). However, these findings do not necessarily mean that infants are not sensitive to, and do not use, sub-syllabic units in speech segmentation. Although it is known that infants do not form adult-like phonetic categories until late

[1] Since the corpora used in majority of the computational studies of segmentation lack phonetic variation, the input unit in these models are effectively phonemes. We acknowledge that the input to the children exhibit phonetic variation, but this is not directly relevant to our results since the same applies to both units we compare.

in the first year in life (Kuhl, 2004), they are sensitive to sub-syllabic changes in the input (Jusczyk and Derrah, 1987; Werker and Tees, 1984). Besides potential constraints due to young age, a logical reason for early learners not to have adult-like phonetic categories is the fact that learning these categories is largely mediated by their use in distinguishing lexical units from each other. For the purposes of segmentation, what really matters is not that infants are capable of classifying relevant phonetic segments into adult-like categories, but being able to detect the differences (and similarities) between such segments. Furthermore, it is also unrealistic to expect infants, who did not form phonetic categories, to perceive syllables categorically. Hence, whether the syllable or the phoneme is an earlier or better perceptual unit is still open to debate, and reality seems to be more complex than choosing one over the other (Dumay and Content, 2012; Foss and Swinney, 1973; Healy and Cutting, 1976; Morais and Kolinsky, 1994; Pallier, 1997).

A few exceptions aside (Gambell and Yang, 2006; Lignos and Yang, 2010; Phillips and Pearl, 2014; Swingley, 2005), most of the computational models in the literature take phonetic segments as the basic unit. For some of the models, the syllable is a natural choice as the basic unit because they are based on information associated with syllables rather than sub-syllabic units. For example both lexical stress (Gambell and Yang, 2006; Swingley, 2005), and vowel harmony (Ketrez, 2013) operates at the level of syllable. Even when such information, e.g., lexical stress, is used in phoneme-based models (e.g., by Christiansen, Allen, and Seidenberg, 1998, Çöltekin, 2011), the lexical stress is marked on all phonemes that span the stressed syllables, effectively informing the model about the syllable boundaries. For other models, the choice of basic unit does not alter the computations involved. However, the performance of the model may be affected by the choice of the basic unit.

Assuming syllables are the basic units, and evaluating the models based on gold-standard segmentation of words eases the learning task in general. However, syllabification of the input is not necessarily straightforward. In fluent speech, words are not uttered in isolation, hence, perceived syllables are likely to straddle lexical unit boundaries. For example the utterance [gɛt ɪt] will be syllabified as [gɛt.ɪt] if the word boundaries are given. However, the likely syllabification will be [gɛ.tɪt] when we do not assume word boundaries. Another problem with assuming that the syllable is an indivisible unit for lexical segmentation comes from the fact that some morphemes that learners eventually learn to extract out of continuous speech and use it productively are sub-syllabic. Hence, not only that the syllable is not the *only* unit of perception in early language acquisition, but it is also not necessarily the best basic unit for segmenting natural speech since some lexical unit boundaries may be syllable-internal.

This study contrasts the use of phoneme and syllable as the basic units in speech segmentation. To this end, we use a simple state-of-the-art segmentation model, and run a set of simulations on two typologically different languages, English and Turkish. We evaluate the results based on word- and morpheme-segmented gold standards.

The next section describes the model and the data used in this study, Section 3 presents results from a series of computational simulations, we discuss the results in Section 4 and conclude in Section 5.

2 Method and the data

2.1 Data

For the experiments reported in this paper, we use corpora of child-directed speech from English and Turkish. Both corpora used parts of the CHILDES (MacWhinney and Snow, 1985).

For English, we use the *de facto* standard corpus collected by Bernstein Ratner (1987) and processed by Brent (1999). The age range of children in our English data (the BR corpus) is between 0;6 and 0;11.29. Unlike earlier studies, we do not make use of phonemic transcriptions by Brent (1999) in our main experiments. Instead, we convert the orthographic transcriptions to transcriptions based on Carnegie Mellon University pronouncing dictionary (version 7b, Carnegie Mellon University, 2014). The main motivation for using an alternative (but more conventional) transcription has been to be able to apply the standard syllabification methods. The new transcription also avoids some of the arbitrary choices in phonemic transcriptions of Brent (1999).

Turkish child-directed corpus was formed by taking all child-directed utterances from the Aksu corpus (Slobin, 1982). The Aksu corpus contains 53 files (one for each recording session) with 33 target children between ages 2;0–4;4. Although

	English	Turkish
Utterances	9 790	10 767
MLU (word)	3.41	3.42
MLU (morph)	3.89	5.82
MLU (syl.)	4.00	7.45
MLU (phon.)	10.81	17.44
Word tokens	33 377	36 789
Word types	1 380	4 808
Word TTR	0.041 35	0.130 69
Morph tokens	38 081	62 612
Morph types	1 024	1 802
Morph TTR	0.026 89	0.037 89
Syllable tokens	39 150	80 178
Syllable types	1 165	1 044
Syllable TTR	0.029 76	0.013 02
Phone tokens	105 801	187 738
Phone types	37	29
Phone TTR	0.000 35	0.000 16

Table 1: General statistics about the corpora used. Besides type and token counts of each unit, type/token ratio (TTR) and mean length utterance (MLU) measured in different units are given.
Note: the number of utterances was mistakenly reported as 10 206 in the version printed in the proceedings. The number of utterances, and the other numbers affected from this change has been corrected in this table.

the age range is not similar to the BR corpus, this corpus is currently the best option available for Turkish. We order the files by the age of the target child, and take all child-directed utterances. Similar to Brent (1999), onomatopoeia, interjections and disfluencies are removed. Turkish corpus was not converted to a phonetic/phonemic transcription as Turkish orthography follows the standard pronunciation rather closely (this practice is common in the literature, e.g., Göksel and Kerslake, 2005; Ketrez, 2013).

Table 1 presents some basic statistics about the corpora used. Although our corpora are similar in number of utterances, there are important differences due to differences between languages, and potentially due to the age of the target children.

2.1.1 Gold-standard syllabification and morpheme segmentation

Both corpora are syllabified and marked for morpheme boundaries for some of the experiments reported below. Most of the earlier studies rely on dictionaries or human judgments in syllabification of English. Since we do not only syllabify words, but also utterances, we do not use a dictionary-based method. For English, we use a freely available syllabification software that implements a few additional sub-regularities over the maximum-onset principle. English morpheme segmentation is done manually (Gorman, 2013). The morpheme boundaries are determined for each word type, and the same morpheme segmentation is used for all tokens of the same word. For syllabification and morpheme segmentation of Turkish, we use another set of freely available tools (Çöltekin, 2010, 2014).

Some statistics regarding morpheme-segmented and syllabified corpora are given in Table 1. Additionally, we note that the ratios of multi-syllabic word tokens are 16 % and 56 % in our English and Turkish input, respectively.

2.2 Evaluation

As with other models of language acquisition, evaluating models of segmentation is non-trivial. Not only we do not know our target, the early child lexicon, well, but it is also likely to differ substantially based on age, language and even the individual child. Furthermore, the linguistic units used by linguists may not necessarily match the units in a typical human lexicon. For the lack of a better method, we evaluate our model based on gold-standard word and morpheme segmentations. We acknowledge that early learners' lexicon is likely to contain multi-word units. To avoid arbitrary and corpus dependent decisions, however, we do not quantitatively evaluate the model's output based on a selection of multi-word expressions.

As in earlier studies, we report three types of F_1-*scores* (or F-scores). *Boundary* F-score (BF), measures the success of the model in finding boundaries. *Word*, or token, F-score requires both boundaries of a word to be found. Hence, discovering only one of the boundaries of a word does not indicate success for this measure. *Lexicon*, or type, F-score similar to token scores, however, the comparisons are done over the word types the model proposed and word types in the gold standard. F-score is the harmonic mean of precision and recall, and these three types of F-scores (also precision and recall) have conventional measures of success reported in the field (see e.g., Goldwater, Griffiths, and M. Johnson, 2009, for precise defintions).

Besides F-scores, we present oversegmentation

(EO) and undersegmentation (EU) rates with the following definitions.

$$\mathrm{EO} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \qquad \mathrm{EU} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

where TP, FP, FN and TN stands for true positive, false positives, false negatives and true negatives, respectively. The error rates defined above are related to boundary precision and recall. Especially, the undersegmentation rate is equal to $1 -$ recall. The difference between the information conveyed by EO and boundary precision is more subtle. Unlike precision which measures the rate of the correct decisions over all boundary decisions made by the model, EO ranges over the word-internal positions in the gold-standard segmentation. For example, if the model admits one correct and one incorrect boundary, the precision will be 0.5. However, the EO depends on the number of word-internal positions in the gold standard. The smaller the number of potential false positives, the higher the EO will be for the same number oversegmentation errors. As a result, the error measures defined above give a more direct indication of how much room is left for improvement.

Similar to the earlier literature, we do not split our data as test and training set since we are using an unsupervised learning method.

2.3 The segmentation model

For the experiments reported below, we implement and use a well-known segmentation model.[2] The model assigns probabilities to possible segmentations as described in Equations 1 and 2.

$$P(s) = \prod_{i=1}^{n} P(w_i) \tag{1}$$

$$P(w) = \begin{cases} (1-\alpha) f(w) & \text{if } w \text{ is known} \\ \alpha \prod_{j=1}^{m} f(a_j) & \text{if } w \text{ is unknown} \end{cases} \tag{2}$$

where s is a sequence of phonemes (e.g., an utterance or a corpus), w_i is the i^{th} word in the sequence, a_j is the j^{th} basic unit in the word, $f(w_i)$ and $f(a_j)$ are the relative frequencies of word w_i or basic unit a_i respectively, n is the number of words in the utterance, m is the length of the word in input units, and $0 \leq \alpha \leq 1$ is the only parameter of the model. The parameter α can be interpreted as the probability of admitting novel lexical items, and it also affects how eager or the conservative the model is in inserting boundaries. In the simulations reported in this paper, we fix α at 0.5, and adopt an incremental learning method where learner processes the input utterance by utterance. Each utterance is segmented using the current model parameters (phoneme and word frequencies), and parameters are updated based on the segmented utterance before proceeding to the next.

[2] The source code of the implementation, the data files and utilities used in preprocessing the data are publicly available at `http://doi.org/10.5281/zenodo.27433`.

model	BF	WF	LF
Brent, 1999	82.3	68.2	52.4
Venkataraman, 2001	82.1	68.3	55.7
Goldwater, Griffiths, and M. Johnson (2009)	85.2	72.3	59.1
Blanchard, Heinz, and Golinkoff (2010)	81.9	66.1	56.3
Current model (incremental)	83.4	71.6	55.3
Current model (final)	86.6	76.3	70.7

Table 2: Performance scores of the present model in comparison to some of the models in the literature that are tested on the BR corpus.

One way to view this model is as an instance of minimum description length (MDL) principle (Rissanen, 1978). (Creutz and Lagus, 2007; Goldsmith, 2001; Marcken, 1996; Rissanen, 1978). Equation 1 imposes a preference for short utterances (in number of words). Assuming each word is represented by an index or pointer in the lexicon, this leads to a preference towards a representation that minimizes the corpus length. Let alone, this preference would result in no segmentation, and corpus size would be equal to the number of utterance types. Despite small corpus representation, this would lead to a large lexicon containing all the utterance types. The second part of Equation 2, on the other hand, imposes a preference for short words and, since shorter strings result in fewer word types, a shorter lexicon. In its limiting case, this preference would result in a lexicon containing the basic units. Resulting in a large corpus representation despite a very small lexicon. As a result, learning for this model is about finding a compromise (hopefully the best) between these two extremes.

The model as described above can also be seen as a generative model. At each step, the model either decides to produce a novel word with probability α, or pick a word from the lexicon with probability $1 - \alpha$. The probability of words from the lexicon is proportional to their empirical prob-

ability (relative frequency with which they are observed). If the model decides to generate a novel word, it produces a series of basic units. Choice of basic units is, again, proportional to their probability of occurrence (for completeness, one needs to either introduce a special end-of-word unit which terminates the sequence) With this description, the model is similar to the model suggested by Brent (1999), Venkataraman (2001, although he does not formulate his model as a generative model), and the unigram model of Goldwater, Griffiths, and M. Johnson (2009).

For simplicity, we use a fixed α and we do not consider word context (e.g., word bigrams). Despite these simplifications, the performance of the present model is competitive with the state-of-the-art models in the literature. To enable a rough comparison, we provide the performance scores of some of the similar models evaluated on the same corpus, together with result obtained using the present model in Table 2. Unlike the rest of the experiments reported in this paper, to increase the comparability of the results with the earlier literature, the result presented here are obtained using the phonemic transcription of the original BR corpus (the version transcribed by Brent, 1999). The row marked 'incremental' reflects the scores obtained by evaluating the segmentations on the whole corpus during a single pass. Although it is the common method of evaluating the incremental models in the literature, this leads to an unfair disadvantage when the model is compared with a batch model which would have already made many passes over the complete corpus at the time of evaluation. The row marked 'final' in Table 2 reports the final evaluation metrics obtained while they were calculated for each 1 000-utterance block. Hence, the 'final' results are obtained from a more 'learned state' of the model, providing a better comparison with batch models. In the rest of this paper, we present only the 'incremental' version of the performance score.

Although the results in Table 2 indicate that the model is comparable to (and better than on some counts) the state of the art, we note that our aim in this work is not to introduce another segmentation model, but compare two basic units using a model that shares many features with the earlier state-of-the-art models.

	BF	WF	LF	EO	EU
En (words)	89.1	77.6	55.1	3.2	0.0
En (uttr.)	71.9	56.7	42.4	5.8	19.3
Tr (words)	54.5	17.4	5.8	25.8	0.0
Tr (uttr.)	48.6	16.0	3.4	27.5	11.0

Table 3: Scores of 'syllable as word' baseline.

3 Experiments and results

3.1 'Syllable as word' baseline

For languages like (child-directed) English where most words are mono-syllabic, a potentially deceiving aspect of using syllable as the basic unit for segmentation is that the learner may learn single input units, syllables, as words. Hence, an interesting baseline can be obtained by segmenting at every syllable boundary. We segment both corpora trivially at syllable boundaries, and evaluate against the gold-standard word segmentation of these corpora. To approximate a possible syllabification when word boundaries are not given, we also preset results where syllabification algorithm is applied without marking the word boundaries. The evaluation results for both languages are presented in Table 3.

Not surprisingly, when syllabification is done at word boundaries, the model recovers all word boundaries, hence EU is 0 for both languages. The oversegmentation errors in this setting is the upper bound for EO when syllables are used as the basic unit. The F-scores of the syllable baseline on the BR corpus, where the words are predominantly monosyllabic, is similar to the state-of-the-art models presented in Table 2, while the numbers are substantially lower for Turkish.

To contrast with this 'syllable as word' baseline, it is also instructive to consider a 'phoneme as word' strategy. If one would segment at every phoneme, the error rates go up to $62.1\,\%$ and $89.7\,\%$ for English and Turkish respectively. This results in $0.4\,\%$ and $0.04\,\%$ lexical F-scores for English and Turkish. Clearly, the models considering syllable as the basic unit starts with a great advantage for English. While helpful, the results for Turkish is far from what we observe for English.

As expected, when syllabification is done without word boundaries, error rates increase for both languages. Undersegmentations are caused by syllables straddling the word boundaries, and oversegmentations increase because of increased number of word-internal syllable boundaries. However, the effect is not as drastic as the differences

	BF	WF	LF	EO	EU
En (phon)	80.9	68.2	51.0	5.7	20.3
En (syl/w)	48.5	29.4	23.5	0.01	67.9
En (syl/u)	55.5	36.1	25.0	0.2	61.3
Tr (phon)	65.7	42.1	29.0	9.4	24.3
Tr (syl/w)	69.8	50.7	39.1	2.6	38.5
Tr (syl/u)	68.5	49.6	38.2	2.8	39.3

Table 4: Segmentation scores using phonemes and syllables.

between the two languages in the same setting.

3.2 Syllables vs. phonemes

Table 4 presents segmentation performance of models that use phonemes or syllables as basic input units. We present results for syllabification with and without restricting syllable boundaries at word boundaries. We first note that the phonemic transcription we use seems to be harder to segment than the transcription by Brent (1999). The F-scores presented on the first row of Table 4 are all lower than the corresponding F-scores in Table 2.

For English, we observe an overall decrease in performance scores when the basic units are syllables. Despite the fact that model makes very few oversegmentation errors, the undersegmentation rate is even worse than a process that inserts boundaries at random. Given the overall conservative segmentation tendency of the model, this is not surprising. Surprisingly, however, when the syllabification is done based on whole utterances, the model seems to perform better. The decrease in EU seems to result in an improvement in all conventional F-scores.

The phoneme-based segmentation scores for Turkish is lower than English. This is in-line with earlier studies that compared English with other languages. As in English, the EO decreases, and the EU increases when syllables (rather than phonemes) are used as the basic units. However, unlike English, the effect of this is positive on all F-score measures. The surprising positive effect of syllabification of full utterances does not persist on the Turkish corpus. The utterance-based syllabification causes an increase on both EU and EO, resulting in a slight drop in all F-scores.

Although the overall performance/error scores presented are informative, the pattern of learning for the model is also important. To show how learning proceeds for both models, we plot over- and under-segmentation rates progressively for both languages, both for phoneme and syllable as basic units in Figure 1. As the description of the model in Section 2.3 indicate, all models start with a preference of undersegmentation. In the process, the EO increases, and EU decreases. In general, the models learn quickly. After a short initial period of the increase in EO and decrease in EU, the changes are rather small.

With syllables, the decrease in EU is very small, particularly for English. We observe a quicker drop of errors for phoneme-based models in general, and the expected trend of higher EU lower EO of the syllable-based model in comparison to phoneme based models holds in all settings. With respect to the differences between the languages, the undersegmentation curves for phoneme-based models are very similar, leading to similar error rates at the end of the learning. However, for Turkish we observe a higher rate of oversegmentation errors. The peak in EO for the phoneme-based segmentation just before the 1 000th utterance for English seems to be due to the particular ordering of the BR corpus. Multiple experiments with shuffling the sentences produce similar curves without abrupt changes.

3.3 Words vs. morphemes

Next, we use the same input described in Section 3.2, but evaluate on the morpheme-segmented gold standard corpora. The scores are presented in Table 5. Compared to the scores based on word segmentation in Table 4, we observe a slight increase in the performance scores in segmenting the BR corpus, since fewer of the model's segmentations are now marked as oversegmentation errors. The undersegmentation, on the other hand, increases slightly. The positive effect of reduced oversegmentation errors are more pronounced for Turkish. However, segmentation performance for Turkish with phonemes as the basic unit is still much lower than English. For both languages, the performance with syllables as basic unit is lower when tested against morpheme-segmented gold standard. Most of the morphemes being formed by sub-syllabic units, this is the expected result for English. However, syllabification does not help the model to find morphemes for Turkish either.

4 General discussion

Our main motivation in this study has been to gain further insight into usefulness of syllables or phonemes as the basic input units. We presented

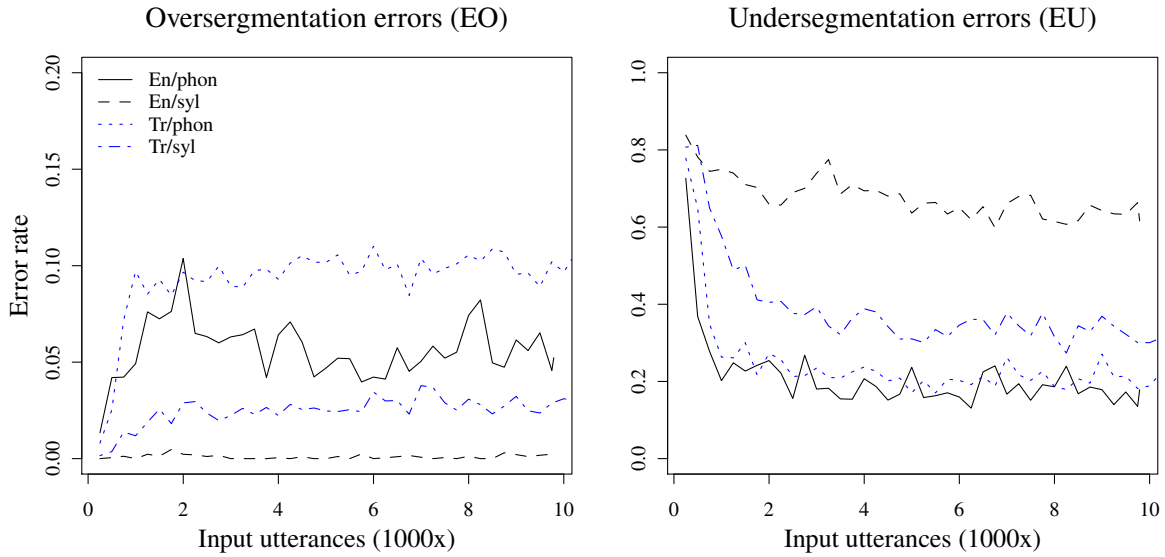

Figure 1: Oversegmentation (left) and undersegmentation (right) rate plotted incrementally during learning. Note that the y-axis ranges are not the same.

	BF	WF	LF	EO	EU
En (phon)	82.7	70.5	51.3	2.6	25.2
En (syl/w)	47.7	24.1	20.8	0.1	68.6
En (syl/u)	54.6	30.4	22.5	0.1	62.4
Tr (phon)	68.4	44.1	33.6	3.6	43.4
Tr (syl/w)	55.3	25.4	21.9	0.6	61.2
Tr (syl/u)	56.1	27.4	23.3	0.7	60.3

Table 5: Segmentation scores using phonemes and syllables with morph-segmentation as gold-standard segmentation.

results from experiments from two typologically different languages and two different settings for the gold-standard, one considering written words as the lexical units as in earlier studies, and the other with morphemes as lexical units.

Unlike earlier studies (e.g., Gambell and Yang, 2006; Phillips and Pearl, 2014), our results do not suggest a direct indication of the usefulness of the syllable (or the phoneme) as the basic input representation for segmentation. The syllable-based model performs worse than phoneme-based model on English, while it improves the segmentation performance on our Turkish corpus. For both languages, the invariant trend is that syllable-based models make fewer oversegmentation mistakes with the cost of higher undersegmentation rate. For English, where the words are rather short, the undersegmentation is severe, and syllable-based segmentation causes F-scores to drop drastically. For Turkish, since the average word length is much larger (see Table 1), the undersegmentation is less severe, and we see increase in the F-scores for segmentation.

The low oversegmentation is expected from the syllable-based models, simply because the models are restricted to insert boundaries in fewer locations. As the 'syllable as word' baseline results presented in Table 3 suggests, most of these locations are true word boundaries. If we allow a more eager segmentation strategy (through a different model, or different parameter settings), syllable-based models are expected to yield good segmentation scores for English. The success of the most eager segmentation strategy 'syllable as word' baseline is a clear example of this case. If such an eager strategy is constrained in the right direction, it is not surprising that one can get really good segmentation performance from a syllable-based segmentation model. This, probably, is also the reason for high segmentation scores of stress-based segmentation strategy presented by Gambell and Yang (2006). Since their model is restricted to insert word boundaries only at syllable boundaries and include some linguistically-informed constraints, the high segmentation F-socre is expected as the 'syllable as word' baseline already achieves a boundary F-score of 89 % (Table 3).

Besides the fact that syllables constrain the locations that one can insert boundaries, the success of syllable-based models are also related to some of

the fine details of the model definition. As an example, consider the boundary decision involving a known word w consisting of basic units $a_1 \dots a_k$, and an adjacent unknown string s. With the model defined in Equations 1 and 2, the decision to insert a boundary between w and s in string ws (or sw) requires

$$(1-\alpha)P(w)\alpha P(s) > \alpha P(a_1)\dots P(a_k)P(s)$$
$$(1-\alpha)P(w) > P(a_1)\dots P(a_k)$$

In this setting, the probability of inserting a boundary decreases with the length of the known word. Since syllables reduce the lengths of lexical units, the model becomes more conservative.[3] This partially explains the low scores we obtain using syllable as the basic unit. A potential reason for the model to segment more eagerly (hence better) is high lexical word probabilities. Probably, this is part of the explanation for the better segmentation performance reported by Phillips and Pearl (2014) for syllable-based models only with bigram word probabilities. The probabilities of (real) words conditioned on the previous word will be higher if the words tend to cooccur. Hence, the model tends to segment more eagerly around the frequent bigrams, counteracting the conservative segmentation tendency introduced by using syllable as the basic unit.

Unlike our results on English, syllable-based model improves word segmentation of Turkish. Contrary to our expectations, however, the scores go down when evaluated on morpheme-segmented gold standard. There are at least three reasons for expecting the results to be even better with the syllable-based models when evaluated on the morpheme-based gold standard. First, on average, Turkish words are formed by longer sequences of morphemes. Second, Turkish morphemes are syllabic, our Turkish corpora does not contain any morpheme boundaries that are not syllable boundaries. Third, similar to the English function words, frequent affixes are more frequent than frequent roots/stems. Hence they should be more likely to be picked as lexical units. However, for both languages, syllable based model performs worse when evaluated against morpheme-based gold standard. Looking closer to the errors suggests that the syllable-based models exhibit a similar behavior on morpheme-based gold standard as the English syllable-based model evaluated on word-based gold standard. The model is precise, but misses many of the boundaries.

Besides missing the morphemes that may be formed by sub-syllabic sequences, another potential problem with the syllable-based models when evaluated against morpheme segmented gold-standard is that the syllables perceived from fluent speech may straddle word boundaries. As a result, we expect worse segmentation scores if the syllabification does not consider word boundaries as absolute syllable boundaries. However, the results are surprising for English, at least. It seems syllabification of complete utterances causes a decrease in undersegmentation errors. Despite a small increase in oversegmentation errors, the overall effect of this on the F-scores reported in Table 4 is positive. The reason for this seems to be the change in the syllable distribution resulting in smaller syllable probabilities on average, and hence, more eager segmentation. In general, it seems both problems mentioned above regarding syllable-based models do not cause serious difficulties. However, in general, we did not find a clear computational benefit of one unit or the other as the basic unit for both languages.

5 Conclusions

In this paper, we compared the effects of syllables or phonemes as the basic unit for segmentation using child-directed speech corpora from two typologically different languages. The simulations reported in this paper do not favor one unit over another. In different settings, the success of models based on syllables or phonemes seems to differ. A reasonable explanation for these differences is the relative lengths of lexical and basic units, and their distributions. In other words, the differences observed are likely to be an artifact of the modeling practice. This is not necessarily a disadvantage if the model in question matches the way humans perform the task. Otherwise, the conclusions that may be drawn from these models regarding whether syllable or phoneme is a better choice as the basic unit for early segmentation may be misleading.

In this paper, we investigated the behavior of a single family of models. It would be interesting to observe the difference between syllables and

[3] Also note the model's unintuitive preference for low-probability basic unit sequences as known lexical units. If word length is fixed, right side of the inequality will be higher if the probabilities of the basic units forming the word are higher.

phonemes in other modeling approaches, such as the ones that use local cues, possibly using more distributed representations for the basic units. Although our aim here was to contrast these two potential basic units, it is likely that humans make use of multiple units at different levels. Hence, another interesting question for the future work is whether these units play complementary roles in segmentation.

References

Richard N. Aslin 1993. Segmentation of fluent speech into words: Learning models and the role of maternal input. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*. Ed. by B. De Boysson-Bardies et al. Kluwer Academic Publishers pp. 305–315.

Nan Bernstein Ratner 1987. The phonology of parent-child speech. In *Children's language*. Ed. by K. Nelson and A. van Kleeck. Vol. 6. Hillsdale, NJ: Erlbaum pp. 159–174.

Daniel Blanchard, Jeffrey Heinz, and Roberta Golinkoff 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language* 37(Special Issue 03):487–511.

Michael R. Brent 1999. An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning* 34(1-3):71–105.

Michael R. Brent and Timothy A. Cartwright 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61(1-2):93–125.

Paul Cairns et al. 1994. Modelling the acquisition of lexical segmentation. In *Proceedings of the 26th Child Language Research Forum*. University of Chicago Press.

Carnegie Mellon University 2014. *CMU pronouncing dictionary version 7b*. URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict (visited on 04/01/2015).

Morten H. Christiansen, Joseph Allen, and Mark S. Seidenberg 1998. Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes* 13(2):221–268.

Çağrı Çöltekin 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta pages 820–827.

Çağrı Çöltekin 2011. *Catching Words in a Stream of Speech: Computational simulations of segmenting transcribed child-directed speech*. PhD thesis. University of Groningen.

Çağrı Çöltekin 2014. A set of open source tools for Turkish natural language processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Çağrı Çöltekin and John Nerbonne 2014. An explicit statistical model of learning lexical segmentation using multiple cues. In *Proceedings of EACL 2014 Workshop on Cognitive Aspects of Computational Language Learning*.

Mathias Creutz and Krista Lagus 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.* 4(1):3.

Anne Cutler and Sally Butterfield 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language* 31(2):218–236.

Anne Cutler, Jacques Mehler, et al. 1986. The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25(4):385–400.

Nicolas Dumay and Alain Content 2012. Searching for syllabic coding units in speech perception. *Journal of Memory and Language* 66(4):680–694.

Peter D. Eimas 1999. Segmental and syllabic representations in the perception of speech by young infants. *The Journal of the Acoustical Society of America* 105(3):1901–1911.

Jeffrey L. Elman 1990. Finding structure in time. *Cognitive Science* 14:179–211.

Margaret M. Fleck 2008. Lexicalized phonotactic word segmentation. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL-08)* pages 130–138.

Donald J. Foss and David A. Swinney 1973. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior* 12(3):246–257.

Timothy Gambell and Charles Yang 2006. *Word segmentation: Quick but not dirty*. Unpublished manuscript.

Aslı Göksel and Celia Kerslake 2005. *Turkish: A Comprehensive Grammar*. London: Routledge.

John Goldsmith 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2):153–198.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1):21–54.

Kyle Gorman 2013. *syllabify.py: Automated English syllabification*.

Alice F. Healy and James E. Cutting 1976. Units of speech perception: Phoneme and syllable. *Journal of Verbal Learning and Verbal Behavior* 15(1):73–83.

Elizabeth K. Johnson and Peter W. Jusczyk 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language* 44(4):548–567.

Mark Johnson and Sharon Goldwater 2009. Improving nonparameteric Bayesian inference: experiments

on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* pages 317–325.

Peter W. Jusczyk, Anne Cutler, and Nancy J. Redanz 1993. Infants' preference for the predominant stress patterns of English words. *Child Development* 64(3):675–687.

Peter W. Jusczyk and Carolyn Derrah 1987. Representation of speech sounds by young infants. *Developmental Psychology* 23(5):648–654.

Peter W. Jusczyk, Elizabeth A. Hohne, and Angela Bauman 1999. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics* 61(8):1465–1476.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome 1999. The Beginnings of Word Segmentation in English-Learning Infants. *Cognitive Psychology* 39:159–207.

Anja van Kampen et al. 2008. Metrical and statistical cues for word segmentation: The use of vowel harmony and word stress as cues to word boundaries by 6- and 9-month-old Turkish learners. In *Language Acquisition and Development: Proceedings of GALA 2007*. Ed. by Anna Gavarro and M. Joao Freitas pages 313–324.

F. Nihan Ketrez 2013. Harmonic cues for speech segmentation: a cross-linguistic corpus study on child-directed speech. *Journal of Child Language* 41:1–23.

Patricia K. Kuhl 2004. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience* 5(11):831–843.

Constantine Lignos and Charles Yang 2010. Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning* pages 88–97.

Brian MacWhinney and Catherine Snow 1985. The child language data exchange system. *Journal of Child Language* 12(2):271–269.

Carl de Marcken 1996. Linguistic structure as composition and perturbation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Santa Cruz, California: Association for Computational Linguistics pages 335–341.

Jacques Mehler et al. 1981. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* 20(3):298–305.

Padraic Monaghan and Morten H. Christiansen 2010. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language* 37(Special Issue 03):545–564.

José Morais and Régine Kolinsky 1994. Perception and awareness in phonological processing: the case of the phoneme. *Cognition* 50(1–3):287–297.

Christophe Pallier 1997. Phonemes and Syllables in Speech Perception: size of the attentional focus in French. In *Proceedings of Eurospeech '97*. Vol. 4 pages 2159–2162.

Lawrence Phillips and Lisa Pearl 2014. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Quebec City, CA: Cognitive Science Society pages 2775–2780.

Jorma Rissanen 1978. Modeling by shortest data description. *Automatica* 14(5):465–471.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport 1996. Statistical learning by 8-month old infants. *Science* 274(5294):1926–1928.

H.B. Savin and Thomas G. Bever 1970. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* 9(3):295–302.

Dan I. Slobin 1982. Universal and particular in the acquisition of language. In *Language acquisition: the state of the art*. Ed. by Eric Wanner and Lila R. Gleitman. Cambridge University Press. Chap. 5 pp. 128–170.

Kari Suomi, James M. McQueen, and Anne Cutler 1997. Vowel Harmony and Speech Segmentation in Finnish. *Journal of Memory and Language* 36(3):422–444.

Daniel Swingley 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology* 50(1):86–132.

Anand Venkataraman 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics* 27(3):351–372.

Janet F. Werker and Richard C. Tees 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development* 7(1):49–63.

Aris Xanthos 2004. An incremental implementation of the utterance-boundary approach to speech segmentation. In *Proceedings of Computational Linguistics in the Netherlands (CLIN) 2003* pages 171–180.