Establishing Cross-linguistic Validity for Unsupervised Word Segmentation Models: A Look at Italian and Farsi

Alicia Yu
Computation of Language Laboratory
University of California, Irvine

Abstract

For early word segmentation, statistical learning strategies using Bayesian models have been offered as alternatives to strategies reliant on language-specific cues. Bayesian word segmentation strategies have been found to be successful for English (Goldwater et al. 2009), but it remains to be seen if this persists in other languages. We evaluate this strategy on child-directed speech across Italian and Farsi to test its cross-linguistic validity. The results of the modeling suggest that this statistical learning strategy is a viable method of word segmentation that is robust cross-linguistically.

Keywords: language acquisition; Bayesian modeling; word segmentation; statistical learning; cross-linguistic

1. Introduction

Word segmentation, the process of identifying word forms in fluent speech, is important in language acquisition because it is a base from which later language learning can occur. Many aspects of language learning such as syntactic structure, grammatical categories, and phonological processes depend on knowledge of word forms.

Experimental studies show that infants can begin to segment speech and identify word boundaries as early as six months (Bortfeld, H. et. al, 2005). Multiple theories have been offered as possible ways children identify words, including phonotactics (Mattys et al. 1999), stress patterns (Morgan et al. 1995), and allophonic variation (Jusczyk at al. 1999). However, these strategies rely on cues whose exact instantiation differs depending on the specific language being learned. For example, in English there is a particular stress pattern where most English words place stress on the first syllable (stress-initial: **PI**/rate). Other languages have various stress patterns; Polish, for example, has a penultimate stress pattern (stress-second to last syllable: uniwersy**TE**tu). This would imply a child would need to know the stress cue of the specific language to begin with before learning word segmentation, since a stressed syllable does not always indicate a word boundary. Statistical learning procedures have been offered as an alternative model as to how children identify words in fluent speech as it provides a language-independent method of learning word segmentation. This way a learner can use this method without any prior knowledge of the specific language.

Experimental research has shown that infants can track statistical information in a number of different ways. As infants have been shown to keep track of conditional probabilities (Saffran et al. 1996, Smith & Yu, 2008) and appear capable of some type of

Bayesian inference (Xu & Tenenbaum 2007, Dewar & Xu 2010), a statistical learning strategy that uses Bayesian inference provides a possible universal method of early word segmentation that does not require any previous knowledge of language. Therefore, a statistical approach is suitable for the initial stages of word segmentation when a child does not know many words or rules of the language.

One current Bayesian inference approach (Goldwater et al 2009, Pearl et al. 2011, Phillips & Pearl 2012) has done rather well with the English language, but it is still unclear how well it will do cross-linguistically. If it is intended as a universal segmentation strategy, it should work well for other languages besides English. This paper will examine performance on Italian and Farsi, languages which have significantly different linguistic qualities than English.

2.1 Implementing Statistical Learning with Bayesian Inference

The underlying Bayesian generative model represents the assumptions the learner brings to the word segmentation problem. This model describes how the observable data are generated, from the learner's perspective. There are two assumptions the model makes: a unigram assumption and a bigram assumption, each with its own set of equations.

The unigram model assumes independence between words – the learner supposes words are randomly generated and have no relation to each other. To encode this into the model, the unigram model uses a *Dirichlet Process* (Ferguson 1973), which assumes that an observed sequence of words $w_1 \ldots w_n$ is generated sequentially using a probabilistic

generative process. In the unigram case, the identity of the *i*th word is chosen according to (1):

(1) $P(w_i = w | w_1 \dots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i - 1 + \alpha}$

where $n_{i-1}(w)$ is the number of times w appears in the previous $i - 1$ words, α is a free parameter of the model which encodes how likely the learner is to encounter a novel word, and P_0 is a *base distribution* (2) specifying the probability that a novel word consists of the particular units (in our case, syllables) that it does $x_1 \dots x_m$. P_0 can be seen as a simplicity bias since the model has a preference for shorter words; the more units that comprise a word, the smaller the probability of that word is. α can be seen as influencing the bias for the number of unique lexical words in the corpus since α controls the probability of creating a new word in the lexicon. Therefore if α is small, the learner is less likely to hypothesize new words to explain observable data, and so the learner prefers fewer unique words in the lexicon.

(2) $P_0(w = x_1 \dots x_m) = \prod_{j=1}^{m} P(x_j)$

The bigram model makes a different, slightly more complex assumption about the relationship between words. Learners using this model believe a word is related to the previous word in an utterance – i.e. a word is generated based on the identity of the word immediately preceding it. To encode this assumption, the bigram model uses a *hierarchical Dirichlet Process* (Teh et al. 2006). This model tracks the frequencies of two-word sequences and is defined in (3-4):

(3) $P(w_i = w | w_{i-1} = w', w_1 \dots w_{i-2}) = \frac{n_{i-1}(w', w) + \beta P_1(w)}{n_{i-1}(w') + \beta}$

(4) $P_1(w_i = w) = \frac{b_{i-1}(w) + \gamma P_0(w)}{b_{i-1} + \gamma}$

where $n_{i-1}(w',w)$ is the number of times the bigram (w',w) has occurred in the first $i-1$ words, $b_{i-1}(w)$ is the number of times w has occurred as the second word of a bigram, b_{i-1} is the total number of bigrams, P_0 is defined as in (2), and β and γ are free model parameters. Both the β and γ parameters, similar to the α parameter described above, control the bias towards fewer unique bigrams (β) and towards fewer unique lexical words (γ).

Both unigram and bigram generative models implicitly incorporate preferences for smaller lexicons by preferring words that appear frequently (due to (1), (3), and (4)) as well as shorter words in the lexicon (due to (2)). These assumptions are a greatly simplified view of how words are actually generated, but they provide reasonable uninformed hypothesis an infant learner may make.

Because Bayesian models often cannot be solved exactly, there are a number of different ways to perform learning. We separate our learners between "ideal" and "constrained" versions, where the constrained learners incorporate specific cognitive limitations into the learning procedure. Ideal learners operate with unlimited processing and memory resources. However, this is clearly not a realistic type of learner as people, in particular children, do not have such unlimited resources when learning a language. To mimic a more realistic learning method, the constrained learner operates with limited processing and memory resources, much like a child must learn language with limited cognitive processes. By creating different learners with different types of restrictions, multiple possibilities of word segmentation are tested which can then be compared to the correct adult word segmentation, otherwise known as the "gold standard." Comparing

these different learners provides a possible outlook of strategies children use when acquiring language and learning to segment words.

Of the different modeled learners, one is an ideal learner and the other three are constrained. These four learners are the batch optimal learner (BatchOpt), the online optimal learner (OnlineOpt), the online suboptimal (OnlineSubOpt), and the online memory learner (OnlineMem).

As indicated by its name, the BatchOpt represents the "optimal" or ideal learner. The BatchOpt (Goldwater et al. 2009) makes decisions using Gibbs sampling, and is allowed unlimited computational resources to remember all the data seen previously. Gibbs sampling is a procedure that is guaranteed to converge on the optimal segmentation, therefore mimicking the ideal situation desired for the ideal learners. This particular algorithm operates by iterating through the corpus multiple times (as many as 20,000 iterations), going through boundary by boundary and deciding based on other decisions it has made, whether or not there should be a boundary at that point the corpus. Gibbs sampling begins essentially as noise, but over time it bootstraps its decisions and eventually converges on the optimal segmentation. This is crucial for the BatchOpt learner as it receives the data all at once or in a "batch." As noted previously, while optimal learners should outperform the constrained learners, they are not a realistic model for modeling children word segmentation as a child does not have unlimited cognitive resources. Liang & Klein (2009) also show that some online learners have properties which make them better suited to particular tasks. They can converge on the right answer more quickly and can have the ability to avoid local minima (a solution that's much

better than any solution near it but which isn't better than all the solutions in the hypothesis space, which the BatchOpt learner might get "stuck" in.)

The OnlineOpt learner, though not optimal like the BatchOpt, performs in a very similar manner but does not receive all the data in a batch; rather the learner receives data one utterance at a time. Using the Viterbi algorithm, the OnlineOpt learner computes an efficient segmentation of each utterance (Brent 1999). Once a decision is made, it continues to do so with each subsequent utterance, using previous decisions to aid its current decision. While it processes each utterance at a time, the OnlineOpt learner still performs much like the BatchOpt in that it selects what it perceives as the best segmentation. While the OnlineOpt learner uses optimal inference, the fact that it only does it over the local information it has makes it a constrained learner versus an ideal learner.

The OnlineSubOpt learner like the OnlineOpt learner operates on one utterance at a time. However, instead of continually opting for the best segmentation, the OnlineSubOpt learner will choose the best segmentation for a majority of decisions, but will have a small chance of also choosing smaller probability segmentation possibilities (like a distribution). For example, a learner could be presented with the utterance *goodbye*. There would be two difference options of segmenting this word – either it is segmented as one full word, *goodbye,* or two words, *good* and *bye.* Now perhaps the boundary *goodbye* has a 75% chance of being true while the boundary *good bye* has a 25% being true. Given this, the OnlineOpt learner would always choose the segmentation *goodbye* since it has the highest percent chance of being true while the OnlineSubOpt

would have a 75% chance of choosing the segmentation *goodbye* and a 25% chance of choosing the segmentation *good bye.*

The OnlineMem learner attempts to incorporate incremental processing as well as a form of short-term memory, pulling from its recent memory of past utterances to help current word segmentation decisions. The OnlineMem learner uses Gibbs sampling just as the BatchOpt learner does, but it performs differently in that it does not go through every boundary possibility as the BatchOpt does. Instead, the OnlineMem learner makes decisions on boundaries one utterance at a time and has the ability to change decisions on word boundaries from past utterances. Which boundaries are chosen to be updated is based on a decaying function, where boundaries further from the end of the current utterance are exponentially less likely to be chosen. This ensures that the vast majority of sampled boundaries are within the current or previous utterance. Limitations of the constrained learners create a much more realistic model for child language acquisition. These learners incorporate constraints much like a child would have when first learning a language.

2.2 Previous Research

These different modeled Bayesian learners have been found to work fairly well for English, although the bigram model (Pearl et. al, 2011) typically outperforms the unigram model.

The Pearl et. al study (2011) along with most other previous studies focus on phonemes being the basic unit of input and learning, such that the learners receive a stream of phonemes as input and must decide word boundaries from that. However,

evidence shows that syllables are a more likely basic unit as infants are aware of syllables as early as 3 months (Eimas, 1999) while they only become aware of a language's full set of phonemes by around 10 months (Werker & Tees, 1984). Following research shifted to focus on syllables than phonemes as a result. Using the syllable as a basic unit, Phillips & Pearl (2012) found an even greater "less is more" effect (discussed in more detail below) than that found in some cases in the Pearl et al. study (2011), with constrained learners in both unigram and bigram models outperforming unconstrained learners.

Table 1 provides a comparison between the resulting F-scores of phonemes versus those of syllables for English from Phillips & Pearl (2012). The F-score (F) is the harmonic mean of precision (p) and recall (r):

$$\mathbf{F = 2pr/(p+r)}$$

Precision represents the percent of identified word tokens which were correct (# correct / # identified). Word tokens refer to distinct individual words that appear multiple times throughout the corpus. Recall, though similar, represents the percent of the true tokens in the corpus which were correctly identified (# correct / # true). For example, a learner that analyzes a corpus of 10 words may detect a total of 9 words. If the learner correctly identifies 7 words, it would have a precision of 7/9 and a recall of 7/10. The F-score provides a way of presenting this accuracy with a single number, in which a higher number indicates greater segmentation accuracy.

Table 1: F-scores of English results with phonemes as base unit vs syllables as base unit. Bolded scores indicate the higher score.

		Phoneme	Syllable
Unigram	Batch-Opt	**54.8**	53.12
	Online-Opt	**65.9**	58.76

	OnlineSub-Opt	58.5	**63.68**
	OnlineMem	**67.8**	55.12
Bigram	Batch-Opt	71.5	**77.06**
	Online-Opt	69.4	**75.08**
	OnlineSub-Opt	39.8	**77.77**
	OnlineMem	73.0	**86.26**

A look at Table 1 shows using syllables as a basic unit of input instead of phonemes presents a slightly better result for the bigram learners. As seen in both the phoneme and syllable column, not only does the bigram model consistently outperform the unigram, but the constrained learners (especially the OnlineMem learner) typically perform better than the ideal. This suggests less knowledge and cognitive processing is more helpful in learning word segmentation.

Limitations in cognitive processing found in both types of studies can help rather than impede language acquisition (Pearl et. al, 2011). The constrained learners outperforming the ideal learners (in the unigram) represent a similar effect to the "less is more" hypothesis (Newport, 1990). Newport argues that some cognitive limitations may explain why children are better at acquiring language than adults. Although the "less is more" hypothesis is traditionally thought of in terms of morphosyntax, that more constrained learners outperform their ideal counterparts fits with a general interpretation of the hypothesis, namely that some constraints help rather than hurt learning. This "less is more" effect was seen in the unigram models of the Pearl et al. study (2011) from the undersegmentation errors (explained at length later on) due to frequent sequences of words. Because ideal learners had an unlimited memory, common sequences of words such as "*in the"* would be mistaken as one word, "*inthe",* and would be undersegmented

as one word rather than segmented into two words. In contrast, constrained learners, with their restricted memory, did not have nearly as many undersegmentation errors as they did not leverage the frequency of a sequence of words as well as ideal learners.

3.1 Cross-Linguistic Word Segmentation

If this particular Bayesian inference strategy is a possible universal method of word segmentation, it must work across multiple languages. Therefore this strategy must be tested and shown to be successful on other languages besides English.

Before testing other languages, an extensive look into the language's grammar must be done first. Languages vary in terms of their morphological properties in ways that affect the word segmentation process. A morpheme is considered the smallest piece of a word that is meaningful; this includes root words, affixes, parts of speech, and so on. Some languages have rich morphology - such that a word might include many morphemes – and may be best segmented at the morphological rather than word level. It may be the case then that English is simply the type of language which this learning strategy is best used for and that other languages will see poorer performance because the model segments units smaller than words. In addition, function words vary across differently languages as well. For instance, Italian has many regular prepositions followed by determiner phrases. A reasonable learner might group these words together because they occur together so regularly while this may not happen with other languages due to different sentence structures. We attempt to identify errors that the modeled learners make and label them as "reasonable" if they meet certain standards. In order to

do this, however, we need to create a list of common morphology and function words in each language.

3.2 Reasonable Errors

Undersegmentation is a type of error the model makes when it does not segment a word where it should and therefore creates a word combining two words. For example, instead of segmenting properly into two words like "*did you,"* an undersegmentation errors would create "*didyou."* Undersegmentation errors seem to be the most common type of error across all the different learners, though we will later see that undersegmentation errors are more evident in the English language in comparison to other languages.

Oversegmentation is a type of error the model makes when it segments a word more than it should, creating multiple words when it should just be one word. For example, instead of segmenting something into a whole word like "*helpful,"* the model might oversegment the word into "*help"* and "*full."* We will see that oversegmentation errors are common with Italian and Farsi.

The other category consists of models that do not fit under undersegmentation or oversegmentation but instead segment words into other type of words. An example of this is when the model may segment "*playful dog"* into "*play fuldog."*

One might think that with these three different reasonable errors that it is all too simple for the learners to perform well. However to prevent this from happening, some precautionary methods are taken. A common error the learners make regards prefixes and suffixes. For example, consider the morpheme *re-*. This particular morpheme is a prefix

in the English language. Now suppose the learner hears *very* and segments it into *ve* and *–ry.*. Normally this would be counted as a reasonable error as *re* is indeed a morpheme. However, the way *re* is segmented in this particular situation segments it as a suffix instead of the prefix that it is. Instead of counting this as a reasonable error, we note this is incorrect word segmentation. In addition, we do not count errors as reasonable errors unless the learner has made it 10+ times. This way, we ensure we do not inflate the reasonable error F-scores of the learners by including nonsense words uttered only once or other accidental utterances.

But how do these different errors fit into a language's morphemes and function words? As seen from the undersegmentation and oversegmentation error examples above, these errors sometimes produce real words. Knowing this, how many errors does the model make that are actually harmful and how many are actually reasonable errors that produce a true word or perhaps a morphological unit?

Depending on the type of segmentation errors the learner makes, it can produce a real word, a morpheme, a function word, or a general mis-segmentation. For example, if the learner mis-segmented the utterance *running* into *run* and *ing*, this would count as both a real word reasonable error and a morpheme reasonable error. While the learner segmented incorrectly, it still produced a real word, *run*, and a morpheme, -*ing*. If a child were to segment "*running"* into "*run"* and "*ing,"* it would not be too harmful an error given that *–ing* is an important morphological unit in the English language.

Function words, in particular, are important for errors when function words are combined together. A function word mis-segmentation may segment *at the* into *atthe* instead since many function words appear in the same order often and the learner may

assume it is one word instead of two. These “stock phrases” of combined function words are common mistakes children make when learning to segment words. These errors also seem reasonable because they are useful, regular units in the language. Therefore to have a complete cross-linguistic analysis, the morphemes and function words of other languages needs to be known in order to account for these reasonable errors.

3.3 Italian and Farsi

The Italian and Farsi language vary in multiple ways compared to the English language. Besides a difference in phonetics and syntax, these two languages vary in morphology. In the spectrum of types of languages, language be analytic, synthetic, or polysynthetic. On one side of the spectrum are analytic languages like Chinese and English, in which one morpheme is typically one word. On the other side of the spectrum are polysynthetic languages like Inuit languages in which there is a high ratio of morphemes per word. Synthetic languages have a lower ratio of morphemes per word than polysynthetic languages but more than that of analytic languages. Both Italian and Farsi fall under the synthetic language category.

The Italian language falls under a sub-category of synthetic languages known as fusional or inflectional languages. These languages have a greater morpheme to word ratio than analytic languages but many morphemes have multiple meanings. For example, the morpheme *–i* in Italian can serve as a plural morpheme, a gender indicator, or a tense indicator. Farsi, in comparison to Italian, falls under a sub-category of synthetic languages known as agglutinative languages. These languages have more regular morphology, which would be identified by the model more easily than the less regular

morphology of Italian and English. This will likely lead to more reasonable morpheme errors for Farsi.

Italian and Farsi also differ from English in regards to average word length. English has an average word length of 4.16 syllables, Farsi of 6.98 syllables, and Italian of 8.78 syllables. Both Farsi and Italian have a longer average word length than English, allowing for greater room for error with ways to mis-segment utterances.

Cross-linguistically, English has been shown to be consistently easier to segment than other languages. A possible explanation for this is the ambiguity of the languages (Fourtassi et. al, 2013). Given that a learner knows all of the words of a language and how many times a word appears in a corpus, a learner should be able to identify and segment words easily. However, there is some ambiguity in a language that may still cause some errors despite this. For example, in English, a learner may segment *goodbye* to *good* and *bye*. Both *good* and *bye* are real words though it would still be an error as *goodbye* was incorrectly segmented. English has been found to be a less ambiguous language than other languages such as Japanese. Though no comparison has been made between English and Italian and Farsi, it still provides an additional explanation of different results between English and Italian and Farsi.

3.4 Italian and Farsi Results

Table 2 shows a comparison between the F-score of word tokens of English, Italian, and Farsi.

Table 2: F-scores of different languages across the different learners including Italian and Farsi (with new F-scores taking reasonable errors into account in bold)

		English	**Italian**	**Farsi**
Unigram	Batch-Opt	53.12	61.85	66.63

		55.70	**70.48**	**72.48**
	Online-Opt	58.76 **60.71**	59.94 **65.05**	67.77 **75.66**
	Online-SubOpt	63.68 **65.76**	60.23 **66.48**	65.93 **74.89**
	Online-Mem	55.12 **58.68**	58.58 **66.77**	59.57 **67.31**
Bigram	Batch-Opt	77.06 **80.19**	71.25 **79.36**	69.63 **76.01**
	Online-Opt	75.08 **78.09**	67.14 **75.78**	69.83 **79.23**
	Online-SubOpt	77.77 **80.44**	61.25 **73.59**	55.34 **67.54**
	Online-Mem	86.26 **89.58**	60.87 **74.08**	62.46 **73.98**

In general, languages such as English perform well against the gold standard, or the standard of correct adult segmentation. However, Italian and Farsi perform noticeably less so. We have determined that an F-score of 70 or better is doing well for our learners, given previous segmentation results for this learning strategy (Goldwater et al. 2009, Pearl et al. 2011, Phillips & Pearl 2012). English achieves around 77, though Italian and Farsi fall short of the 70 mark. However, when taking reasonable errors into account, most of these languages receive a significant boost in regards to their F-score and all languages achieve better performance. Once reasonable errors are accounted for, Italian and Farsi do much better, receiving a 10 point boost in their F-score. In comparison, English received a maximum 4 point boost in its F-score when taking reasonable errors into account.

One major reason for the higher boost is a ceiling effect for the English language. English was already performing well without the reasonable error boost leaving less room for improvement in comparison to Italian and Farsi. This is also due to the

undersegmentation bias of the English language versus the oversegmentation bias of Italian and Farsi. Given that English is more of a monosyllabic language, this leaves very little room for oversegmentation (since you can't segment one syllable any further). As most common errors that get caught by the reasonable error analysis are oversegmentation errors, Italian and Farsi get a bigger boost in their F-score than English.

As seen in Table 2, the constrained learners outperform the ideal learners in English, suggesting a "less is more" effect. This effect, however, does not appear in Italian and Farsi. As seen in Table 2, the ideal learner of Italian and Farsi continually has a higher F-score, although it is important to note that it is not *that* much higher than the F-score of the constrained learners. This suggests that the Bayesian strategy does not necessarily produce a "less is more effect" cross-linguistically, but including cognitive constraints also does not significantly decrease performance.

Table 3 provides a closer look at the specific types of errors the modeled learners made with Italian and Farsi. As mentioned previously, the types of segmentation errors the learner can make include a real word, a morpheme, a function word, or a general mis-segmentation word. Of course, the learner may just segment an utterance into a complete nonsensical word such as segmenting *pirateking* into *pir ateking.*

Table 3: % of types of words produced during mis-segmentation out of total errors made – real words, morphemes, and function words

		Unigram				**Bigram**			
		Batch-Opt	Online-Opt	Online-SubOpt	Online-Mem	Batch-Opt	Online-Opt	Online-SubOpt	Online-Mem
English	**Real**	0.77	2.39	3.42	2.31	4.52	7.31	9.63	16.91
	Morph	0.13	0.48	0.46	0.31	0.71	0.89	2.09	3.19
	Func	4.40	3.15	3.35	5.02	6.32	4.83	2.84	3.61

Italian	**Real**	16.18	22.69	23.16	17.18	19.99	28.24	30.52	26.87
	Morph	1.13	0.17	0.65	1.36	1.60	0.80	1.02	1.07
	Func	3.69	0.70	0.77	2.87	3.05	1.24	0.43	0.32
Farsi	**Real**	12.57	25.26	25.02	14.07	14.38	26.61	17.52	20.14
	Morph	1.58	4.23	2.78	2.26	2.92	3.82	4.89	5.06
	Func	2.24	0.22	0.10	1.36	1.80	0.07	0.10	0.05

Table 4: Examples of reasonable errors in Italian and Farsi

	True Word	**Model Output**
Real Word	mano	ma no
	‘hand’ Italian	‘but’ ‘no’
	hala	‘ha’ ‘la’
	‘now’ Farsi	‘ha’ ‘la’
Morphology	devi	dev i
	‘you must’ Italian	‘must’ PL
	miduni	mi dun i
	‘you know’ Farsi	PRES ‘know’ 2 Singular
Function Word	a me	ame
	‘to me’ Italian	‘tome’
	mæn hæm	mænhæm
	‘me too’ Farsi	‘metoo’

As seen in Table 3, many of the types of words incorrectly segmented produced real words. These errors are most likely quite common as the model aims to segment words, with a preference for words it has already seen. This may not be true of English because English is monosyllabic and the learner cannot oversegment monosyllabic words. If a learner is segmenting bigger words, it’s less likely to produce a real word. Since Italian and Farsi learners oversegment, they’re going to produce real words more

regularly. Italian and Farsi may oversegment more just because words are longer in those languages than in English.

Of interest is Farsi's more frequent mis-segmentation of morphemes than Italian. As mentioned earlier, Farsi falls under agglutinative synthetic languages while Italian falls under fusional synthetic languages. Since the perceptual unit of the Bayesian inference model is syllables, it picks up syllabic morphology which agglutinative languages such as Farsi have more of while Italian has relatively fewer errors in this category. Many of the most common Italian morphemes instead are sub-syllabic, which means the syllable-based learners here can't identify them.

4. Conclusion

While the modeled learner perform differently on the languages examined here with respect to the types and frequency of errors made, many of those errors are due to specific properties of those languages, such as being a more syllabic language or a more morphologically rich language. The "less is more" effect was not found in other languages tested besides English, but it is still possible that this effect exists among languages with similar properties to that of English. Despite the tentative nature of the "less is more" effect, the Bayesian inference model performs well cross-linguistically and is a sound strategy for learning word segmentation in these languages once reasonable errors are counted as correct. This model provides a good foundation that children can later use as a base to learn other language-specific segmentation cues.

References

Bortfeld, H., Morgan, J.L., Golinkoff, R.M. & Rathbun, K. (2005). Mommy and me. Psychological Science, 16(4), 298-304

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*,*21*(12), 1871-1877.

Eimas, P.D. (1999). Segmental and syllabic representations in the perception of speech by young infants. Journal of the Acoustical Society of America, 105(3), 1901-1911.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics, I, 209 230.

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment?. *CMCL 2013*, 1.

Goldwater, S., Griffiths, T. & Johnson, M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. Cognition 112(1), 21-54.

Jusczyk, P., Hohne, E., & Baumann, A. (1999a). Infants‘ sensitivity to allophonic cues for word segmentation. Perception and Psychophysics, 61, 1465-1476.

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. Cognitive Pscyhology, 38, 465-494.

Morgan, J., Bonamo, K., & Travis, L. (1995). Negative evidence on negative evidence. Developmental Psychology, 31, 180-197.

Newport, E. 1990. Maturational constraints on language learning. Cognitive Science, 14, 11-28.

Pearl, L., Goldwater, S., & Steyvers, M. 2011. Online Learning Mechanisms for Bayesian Models of Word Segmentation, Research on Language and Computation, special issue on computational models of language acquisition. DOI 10.1007/s11168-011-9074-5.

Phillips, L. & Pearl, L. (2012). 'Less is More' in Bayesian word segmentation: When cognitively plausible learners outperform the ideal, In N. Miyake, D. Peebles, & R. Cooper (eds), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-Month-Old Infants. Science, 274, 1926-1928.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Werker, J.F. & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior & Development, 7, 49-63.

Xu, F. & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. Psychological Review, 114(2), 245-272