Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation

Lawrence Phillips & Lisa Pearl
University of California, Irvine
3151 Social Sciences Plaza
Irvine, CA 92697 USA
[lawphill, lpearl]@uci.edu

Abstract

Models of language acquisition are typically evaluated against a "gold standard" meant to represent adult linguistic knowledge, such as orthographic words for the task of speech segmentation. Yet adult knowledge is rarely the target knowledge for the stage of acquisition being modeled, making the gold standard an imperfect evaluation metric. To supplement the gold standard evaluation metric, we propose an alternative utility-based metric that measures whether the acquired knowledge facilitates future learning. We take the task of speech segmentation as a case study, assessing previously proposed models of segmentation on their ability to generate output that (i) enables creation of language-specific segmentation cues that rely on stress patterns, and (ii) assists the subsequent acquisition task of learning word meanings. We find that behavior that maximizes gold standard performance does not necessarily maximize the utility of the acquired knowledge, highlighting the benefit of multiple evaluation metrics.

1 The problem with model evaluation

Over the past decades, computational modeling has become an increasingly useful tool for studying the ways children acquire their native language. Modeling allows researchers to explicitly evaluate learning strategies by whether these strategies would enable acquisition success. But how do researchers determine if a particular learning strategy is successful? Traditionally, models have been evaluated against adult linguistic knowledge, typically captured in an explicit "gold standard". If the modeled learner succeeds at acquiring this adult linguistic knowledge, then it is said to have succeeded and the learning strategy is held up as a viable option for how the acquisition process might work.

Gold standard evaluation has two key benefits. First, it provides a uniform measure of evaluation, especially when gold standards are relatively similar across corpora (e.g. orthographic segmentation for speech). Second, this kind of evaluation is typically straightforward to implement for labeled corpora, and so is easy to use for model comparison.

Still, there are several potential disadvantages to gold standard evaluation. First, the choice of an appropriate gold standard is non-trivial for many linguistic tasks since there is disagreement about what the adult knowledge actually is (e.g., speech segmentation, grammatical categorization, syntactic parsing). Second, implementation may require a large amount of time-consuming manual annotation (e.g. visual scene labeling for word-object mapping). Third, and perhaps most importantly, it is unclear that adult knowledge is the appropriate output for some modeled learning strategies, particularly those that are meant to occur early in acquisition.

For example, consider the early stages of speech segmentation that rely only on probabilistic cues. The earliest evidence of speech segmentation comes at six months (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) and it appears that probabilistic cues to segmentation, which are language-independent because their implementation does not depend on the specific language being acquired, give way to language-dependent cues between eight and nine

Proceedings of CMCL 2015, pages 68–78,
Denver, Colorado, June 4, 2015.

months (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). So, accurate models of this early stage of speech segmentation should output the knowledge that a nine-month-old has, and this may differ quite significantly from the knowledge an adult has about how to segment speech.

Unfortunately, addressing this last issue with gold standard evaluation is non-trivial. One strategy might be to create a gold standard representing age-appropriate knowledge. However, without empirical data that can identify exactly what children's knowledge at a particular age is, this is difficult. Because of this, few (if any) age-specific gold standards exist for the many acquisition tasks that we wish to evaluate learning strategies for. An alternative is to compare model results against qualitative patterns that have been reported in the developmental literature. For instance, Lignos (2012) compares his segmentation model results against qualitative patterns of over- and undersegmentation reported in diary data (Brown, 1973; Peters, 1983). Still, such comparisons are often difficult to make since the behavioral data may come from children of different ages than the modeled learners (e.g., the segmentation patterns mentioned above come from two- and three-year-olds while the modeled learners are at most nine months old).

So, the essence of the evaluation problem is this: the true target for model output is potentially unknown, but we still wish to evaluate different models. Fortunately for language acquisition modelers, this is exactly the problem faced in computer science when unsupervised learning algorithms are applied and a gold standard does not exist. There are two main ways a model without a gold standard can be explicitly evaluated (Theodoridis & Koutroubas, 1999; von Luxburg, Williamson, & Guyon, 2011):

1. Apply real-world, expert knowledge to determine if the output is reasonable.

2. Measure the "utility" of the output.

Adding these two evaluation approaches to a language acquisition modeler's toolbox can help alleviate the issues surrounding gold standards. Still, the first option of applying expert knowledge is often time intensive, since this typically involves querying human knowledge. Moreover, given the key concern about what the output of language acquisition models ought to look like anyway, it is unclear that querying linguistic experts is appropriate. Given this, we focus on measuring the utility of the model's output (Mercier, 1912; von Luxburg et al., 2011) to supplement a gold standard analysis.

This means we must be more precise about "utility". Because children acquire linguistic knowledge and then apply that acquired knowledge to learn more of their native language system (Landau & Gleitman, 1985; Morgan & Demuth, 1996), one definition of utility for language acquisition is for the model output to facilitate further knowledge acquisition. Importantly, determining what future knowledge is acquired is often much easier than determining the exact state of that knowledge, as with a gold standard. This is because we often have empirical data about the order in which linguistic knowledge is acquired (e.g., language-independent cues to speech segmentation are used to identify language-dependent cues, which are then used to facilitate further segmentation). We can use these empirical data to identify what a model's output should be used *for*, and assess if the acquired knowledge helps the learner acquire the appropriate additional knowledge. Then, if a modeled strategy yields this kind of useful knowledge, the modeled strategy should be counted as successful; in contrast, if the acquired knowledge isn't useful (or is actively harmful), then this is a mark of failure. Under this view, a strategy's utility is equivalent to its ability to prepare the learner for subsequent acquisition tasks.

As we will see when we apply this utility-based evaluation to speech segmentation strategies, we may still encounter some familiar evaluation issues. In particular, to evaluate whether a model's output prepares a learner for subsequent acquisition tasks, we must have some idea as to what counts as "good enough" preparation for those subsequent tasks. The simplest answer seems to be that "good enough" for the subsequent task means that the output for that task is "good enough" for the next task after that. In some sense then, the best indicator of utility would be that the modeled strategy yields adult level knowledge once the entire acquisition process is complete. However, it is currently impractical to model the entire language acquisition process. Instead, we have to restrict ourselves to smaller segments of the entire process – here, two sequential stages. Given the available empirical data, it may be that we have a better idea about what children's knowledge is for the second stage than we do for the first stage. That is, an age-appropriate gold standard may be available for the subsequent acquisition task. For both utility evaluations we do here, we have something like this for each subsequent task, though it is likely still an imperfect approximation of young children's knowledge.

We note that this utility-based approach differs from a joint inference approach, where two tasks occur simultaneously and information from one task helpfully informs the other (Jones, Johnson, & Frank, 2010; Feldman, Griffiths, Goldwater, & Morgan, 2013; Dillon, Dunbar, & Idsardi, 2013; Doyle & Levy, 2013; Börschinger & Johnson, 2014). Joint inference is appropriate when we have empirical evidence that children accomplish both tasks at the same time. In contrast, the utility-based evaluation approach is appropriate when empirical evidence suggests children accomplish tasks sequentially.

In this paper, we consider the task of speech segmentation and investigate different ways of assessing the utility of previously proposed strategies. Notably, these strategies have generally succeeded when evaluated against some version of a gold standard (Phillips & Pearl, in press, 2014a, 2014b). We first briefly review speech segmentation in infants, and then describe the segmentation strategies previously investigated: a Bayesian segmentation strategy (Goldwater, Griffiths, & Johnson, 2009; Pearl, Goldwater, & Steyvers, 2011) and a subtractive segmentation strategy (Lignos, 2011, 2012). We then evaluate each modeled strategy on two utility measures relating to (i) the creation of language-dependent segmentation cues relying on stress, and (ii) the subsequent acquisition task of learning word meanings.

We find that the strategies differ significantly in their ability to identify stress segmentation cues and facilitate word meaning acquisition, with the Bayesian strategy yielding more useful output than the subtractive segmentation strategy. We discuss how these utility results relate to other qualitative patterns, such as oversegmentation, noting that behavior that maximizes performance against a gold standard does not necessarily maximize the utility of the acquired knowledge for subsequent learning.

2 Speech segmentation strategies

One of the first acquisition tasks infants solve is identifying useful units in fluent speech, and the useful units are typically thought of as words. While word boundaries are inconsistently marked by pauses (Cole & Jakimik, 1980), there are several linguistic cues that infants can leverage (Morgan & Saffran, 1995; Jusczyk, Houston, & Newsome, 1999; Mattys, Jusczyk, & Luce, 1999; Jusczyk, Hohne, & Baumann, 1999; Johnson & Jusczyk, 2001). However, many of these cues are specific to the language being acquired (e.g., whether words of the language generally begin or end with a stressed syllable), and so require infants to identify some words in the language before the language-specific cue can be instantiated. Fortunately, experimental evidence suggests that infants can leverage language-independent probabilistic cues to identify that initial seed pool of words (Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998; Thiessen & Saffran, 2003; Pelucchi, Hay, & Saffran, 2009). This had led to significant interest in the early probabilistic segmentation strategies infants use (Brent, 1999; Batchelder, 2002; Goldwater et al., 2009; Blanchard, Heinz, & Golinkoff, 2010; Pearl et al., 2011; Lignos, 2011).

The two strategies we examine here, a Bayesian strategy (Goldwater et al., 2009; Pearl et al., 2011; Phillips & Pearl, 2014a, 2014b, in press) and a subtractive segmentation strategy (Lignos, 2011, 2012), have two attractive properties. First, they can be implemented so that the modeled learner perceives the input as a sequence of syllables, in accord with the infant speech perception experimental literature (Jusczyk and Derrah (1987); Bertonicini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler (1988); Bijeljac-Babic, Bertoncini, and Mehler (1993); Eimas (1999) and see Phillips and Pearl (in press) for more detailed discussion). Second, their syllable-based implementations perform well on English child-directed speech when compared against a gold standard (Phillips & Pearl, in press; Lignos, 2011).

2.1 Bayesian segmentation

The Bayesian strategy[1] has two variants, using either a unigram or bigram generative assumption for how words are generated in fluent speech. The model assumes utterances are produced via a Dirichlet process (Ferguson, 1973). In the unigram case, the identity of the i^{th} word is chosen according to (1):

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

where n_{i-1} is the number of times w appears in the previous $i-1$ words, α is a free parameter, and P_0 is a base distribution specifying the probability that a novel word will consist of the perceptual units $x_1 \dots x_m$ (which are syllables here):

$$P_0(w = x_1 \dots x_m) = \prod_j P(x_j) \quad (2)$$

In the bigram case, the model assumes a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006) and additionally tracks the frequencies of two-word sequences:

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

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

where $n_{i-1}(w', w)$ is the number of times the bigram (w', w) has occurred in the first $i-1$ words, $n(w')$ is the number of bigrams beginning with word w', $b_{i-1}(w)$ is the number of times w has occurred as the second word of a bigram, b is the total number of bigrams, and β and γ are free parameters.[2]

In both the unigram and bigram variants, this generative model implicitly incorporates preferences for smaller lexicons by preferring words that appear frequently (due to equations 1, 3, and 4) and preferring shorter words in the lexicon (due to equation 2). These can be thought of as domain-general parsimony biases.

The ideal (**Batch**) learner for this model is taken from Goldwater et al. (2009) and utilizes Gibbs sampling (Geman & Geman, 1984) to batch process the entire input corpus, sampling every potential word boundary 20,000 times. This represents the most idealized learner, since Gibbs sampling is guaranteed to converge on the segmentation which best fits the underlying generative model. Because this learner does not include cognitive processing or memory constraints, we also implement one of the constrained learners developed by Pearl et al. (2011) that better approximates actual human inference. In addition, that constrained learner was shown to be very successful on English (Phillips & Pearl, in press).

The constrained (**Online**) learner processes data incrementally, but uses a Decayed Markov Chain Monte Carlo algorithm (Marthi, Pasula, Russell, & Peres, 2002) to implement a kind of limited short-term memory. This learner is similar to the Batch learner in that it uses something like Gibbs sampling. However, the Online learner does not sample all potential boundaries; instead, it samples s previous boundaries using the decay function b^{-d} to select the boundary to sample, where b is the number of potential boundary locations between the boundary under consideration b_c and the end of the current utterance, while d is the decay rate. Thus, the further b_c is from the end of the current utterance, the less likely it is to be sampled. Larger values of d indicate a stricter memory constraint. All results presented here use a set, non-optimized value for d of 1.5, which was chosen to implement a heavy memory constraint (e.g., 90% of samples come from the current utterance, while 96% are in the current or previous utterance). Having sampled a set of boundaries[3], the learner can then update its beliefs about those boundaries and subsequently update its lexicon before moving on to the next utterance.

[1] Called DPSEG by Goldwater et al. (2009).

[2] α, β, and γ for all modeled learners were chosen, as in previous work, to maximize the gold standard word token F-score of the unigram and bigram Batch learner: $\alpha = 1, \beta = 1, \gamma = 90$.

[3] The Online learner samples $s = 20,000$ boundaries per utterance. For a syllable-based learner, this works out to approximately 74% less processing than the Batch learner (Phillips & Pearl, in press).

2.2 Subtractive segmentation

The subtractive segmentation strategy (Lignos, 2011) processes the corpus one utterance a time. It begins by assuming that every utterance is a single word and then, as it adds vocabulary to its lexicon, it segments out those words when possible. The specific variant we investigate is the beam search subtractive segmenter without stress information, which is allowed the same segmentation cues as the Bayesian strategy.

In cases where there is ambiguity with respect to a particular word boundary, the model considers the two possible segmentations (the one with the boundary and the one without) and chooses the one with the higher score. A segmentation's score is the geometric mean of the score of each word in the potential segmentation. A word's score is determined by two factors: (i) its frequency in previous inferred segmentations, and (ii) how often it has been part of potential segmentation that was previously rejected.

2.3 Baseline comparison: Random oracle

We additionally examine a random oracle baseline (Lignos, 2012). This strategy makes guesses about word boundaries as a series of Bernoulli trials, where the probability of a boundary p_b is set to the true probability according to the gold standard. Although this is unrealistic as an actual strategy infants use because it assumes knowledge of word boundary frequency, this strategy serves as a best-case scenario for what random guessing might achieve.

3 Previous results with the gold standard

These strategies were evaluated against a gold standard in English by using the UCI Brent Syllables corpus of English child-directed speech (Phillips & Pearl, in press) available through CHILDES (MacWhinney, 2000), which contains 28,391 utterances of speech directed to American English children between six and nine months old. Word token F-scores (shown in Table 1) provide a convenient summary statistic for segmentation model evaluation, where the F-score is the harmonic mean of precision and recall. So, the F-score balances how accurate the set of identified words is (precision = $\frac{\#\ correctly\ identified}{\#\ identified}$) with how complete the set of identified words is (recall = $\frac{\#correctly\ identified}{\#\ true}$).

Word Token F-scores			
Batch (Uni)	0.531	Online (Uni)	0.551
Batch (Bi)	0.771	Online (Bi)	0.863
Subtractive Seg	0.879	Random	0.588

Table 1: Word token F-score results on the UCI Brent Syllables corpus as reported by Phillips and Pearl (in press) for the Bayesian learners (Batch vs. Online, Unigram vs. Bigram), the subtractive segmenter, and the random oracle baseline.

Based on this evaluation metric, the subtractive segmenter performs the best, though the Bayesian Online bigram learner does nearly as well. Notably, the Bayesian unigram learners suffer significantly in comparison, doing worse than even the random oracle baseline. This suggests the unigram assumption is harmful if the goal is to generate the adult knowledge represented in the gold standard.

4 Stress cue identification

A language-dependent segmentation cue that infants use fairly early is their native language's predominant stress pattern (Jusczyk, Houston, & Newsome, 1999; Morgan & Saffran, 1995). In particular, while seven-month-olds rely more on probabilistic cues, nine-month-olds rely more on stress-based cues (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). So, while probabilistic cues and stress-based cues may be used jointly (Lignos, 2012; Doyle & Levy, 2013), infants likely use probabilistic cues only until enough evidence has been accumulated to identify the language-dependent stress cue. In particular, infants want to identify whether words tend to begin with stressed syllables or end with stressed syllables, since that can provide a convenient heuristic for identifying word boundaries. For example, if words begin with stressed syllables, then a stressed syllable signals that the previous word has ended and a new word has begun.

Given this, a measure of the utility of a segmentation strategy's output is whether the generated lexicon yields the appropriate stress cue. To determine this, we must first identify where stressed syllables are in the English child-directed data. Because the UCI Brent Syllables corpus does not mark stress, we make use of the English Callhome Lexicon (Kingsbury, Strassel, McLemore, & MacIntyre,

1997) to identify the main stress in words. For child-register words not found in standard dictionaries (like *moosha*), we manually coded the stress when the words were familiar enough to us to deduce the stress pattern. If a word was not familiar enough for us to be confident about its stress pattern (e.g., *bonino*), we ignored it for the purposes of this analysis. All words in the analyses presented below were given their dictionary stress patterns. In order to better approximate the stress of actual utterances, monosyllabic words were left unstressed.

Table 2 presents the stress pattern of the bisyllabic word types in each learner's lexicon.[4] Our corpus of English child-directed speech has 1344 unique bisyllabic words with 89.9% beginning with a stressed syllable (SW: *báby*) and 10.1% ending with a stressed syllable (WS: *ballóon*), as shown by the *Adult Seg* row. For the learner to correctly infer that English words tend to be stress-initial, the inferred lexicon should have more words with the stress-initial pattern. This serves as an approximate age-appropriate gold standard, since the goal is to match the qualitative stress distribution pattern that would yield the stress cue English nine-month-olds use (i.e., stressed syllables begin words).

	SW	**WS**
Adult Seg	**89.9%**	**10.1%**
Batch (Uni)	80.0%	20.0%
Online (Uni)	80.8%	19.2%
Batch (Bi)	80.4%	19.6%
Online (Bi)	79.6%	20.4%
Subtractive Seg	59.4%	40.6%
Random	68.5%	31.5%

Table 2: Stress pattern results for all learners on bisyllabic word types. Percentages are calculated out of all bisyllabic words identified by the model.

All Bayesian learners capture the qualitative stress pattern, and come fairly close to capturing the quantitative distribution, with 79.6% - 80.8% of the bisyllabic word types having word-initial stress (SW). The subtractive segmenter weakly shows the same pattern, identifying more bisyllabic word types with word-initial stress (59.4% SW). The random oracle baseline actually produces a stronger word-initial bias than the subtractive segmenter (68.5% SW). This suggests an advantage for the Bayesian strategy when it comes to inferring the English stress segmentation cue from the bisyllabic words in the inferred lexicon.

When we turn to trisyllabic words, however, the Bayesian strategy no longer does better – both strategies fail to capture the qualitative stress pattern (as does the random oracle baseline). Table 3 shows the results across the 345 trisyllabic word types. The qualitative pattern in the true distribution is similar to the bisyllabic words (though the distribution is less pronounced), with the majority (69.2%) having initial stress. However, all strategies yield a preference for word-medial stress in trisyllabic words (37.4% - 50.1%). Interestingly, if a learner was attempting to infer a segmentation cue, word-medial stress actually doesn't yield an obvious cue – there is no word boundary either immediately before or immediately after the stressed syllable. So, even if the inferred stress pattern is incorrect for trisyllabic words, it may not actually harm a learner who is looking for segmentation cues – it just fails to help.

	SWW	**WSW**	**WWS**
Adult Seg	**69.2%**	**2.2%**	**28.6%**
Batch (Uni)	22.7%	50.1%	27.2%
Online (Uni)	22.8%	49.2%	28.0%
Batch (Bi)	22.0%	46.6%	31.4%
Online (Bi)	23.7%	47.7%	28.6%
Subtractive Seg	19.1%	48.7%	32.2%
Random	28.6%	37.4%	34.0%

Table 3: Stress pattern results for all learners on trisyllabic word types. Percentages are calculated out of all trisyllabic words identified by the model.

More generally, these results suggest that the word token F-score is not necessarily correlated with knowledge utility, at least when it comes to inferring language-dependent stress-based cues to segmentation. For instance, the Online Bayesian bigram learner and the subtractive segmenter have similar word token F-scores (0.863 vs. 0.879), but generate quantitatively different predictions for the English stress-based segmentation cue. Similarly, the Bayesian unigram learners have far lower word to-

[4] We note that we calculate this over word types rather than word tokens, since learners may ignore frequency when deciding how far to extend generalizations (Yang, 2005; Perfors, Ransom, & Navarro, 2014).

ken F-scores (0.531-0.551), yet yield correct predictions for the English stress cue, based on bisyllabic word types.

If any of these strategies are the ones infants use, then we would predict that infants in the early stages of segmentation have different expectations about the prevalent stress pattern for bisyllabic vs. trisyllabic words in English. This is something that can be verified experimentally. However, we do note that the current analyses leading to this prediction are based on particular assumptions about how accurately infants perceive stress in their input (here, perfectly accurately), and so future analyses should consider other cognitively plausible instantiations of infant stress perception. In addition, while this stress analysis was only applied to English here, it is worthwhile to do so for other languages that vary in how their stress system operates.

5 Word meaning

A task that infants tackle after they are somewhat able to segment the speech stream is learning word meaning. In particular, word meaning learning begins as early as six months (Tincoff & Jusczyk, 1999, 2012; Bergelson & Swingley, 2012), focusing on concrete items in the learner's environment like *apple* and *hand*. So, another test of a segmentation strategy's utility is whether the lexicon it generates facilitates this kind of early word-object mapping.

5.1 A model of early word-object mapping

Drawing on the intuition that early word-object mapping could leverage cross-situational learning, Frank, Goodman, and Tenenbaum (2009) developed a Bayesian learning strategy for early word-object mapping. The modeled learner infers a referential lexicon of word-object mappings based on the utterances spoken and the set of objects visually salient in the environment. In the generative model shown in the plate diagram in Figure 1, the learner assumes there are some objects (O) in the environment, and the speaker intends to refer to some subset of them (I) using words. The speaker draws words from the referential lexicon (L) to refer to those intended objects, with non-referential words also occurring in the utterances with some probability. So, based on a set of situations (S) containing observable utterances comprised of words (W) and sets of visually salient objects (O), the modeled learner can infer the referential lexicon L of word-object mappings as well as the specific intended objects (I).

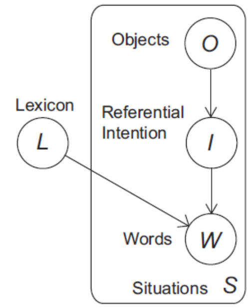

Figure 1: Plate diagram of the Frank et al. (2009) word-object mapping generative model.

This model vastly outperformed other word-object mapping strategies on a sample of English child-directed speech, yielding a referential lexicon that was significantly more accurate (higher precision) compared to other strategies. High lexicon precision is likely more important than high lexical recall for early word-object mapping because this is only the first stage of word meaning learning. So, it is better to have a small set of reliable word-object mappings than a large set of unreliable word-object mappings if the learner is using these mappings to bootstrap future word meaning acquisition.

Notably, the model assumes the utterances are already segmented into words. So, a natural evaluation measure for segmentation strategies is to use the inferred segmentation of the utterances, rather than the adult orthographic segmentation used in the original Frank et al. (2009) demonstration. We can then see if the mapping strategy is still able to identify a reliable referential lexicon. As with the previous utility evaluation, the desired output (a lexicon of word-object mappings) is a gold standard, in this case based on how adults construct word-object mappings. However, because the inferred mappings focus on concrete objects infants are known to learn mappings for, we believe it is at least an approximation of an age-appropriate gold standard.

5.2 Segmentation strategy evaluation

Originally, the word-object model was evaluated on a small subset of 700 utterances from the Rollins corpus from CHILDES (MacWhinney, 2000) which was labeled with visually salient objects (O in the Figure 1). We used this corpus to evaluate the segmentation strategies. We first trained the segmentation strategies on the 28,391 utterances of the UCI Brent Syllables corpus (Phillips & Pearl, in press) so that the modeled learners using those strategies could infer a lexicon of word forms with associated probabilities of occurrence. We then applied the resulting knowledge to the Rollins corpus subset, letting each strategy segment those utterances as best it could, given the knowledge it had inferred from the training set. The word-object mapping model was then applied with the inferred segmentations as part of the observed input (W). Due to the stochastic nature of the inference process, we repeated this process five times and present averaged results.

We present lexical precision scores due to the importance of inferring high quality mappings during early word meaning learning. However, to measure precision we need to identify what constitutes a "correct" mapping. Frank et al. (2009) created a gold standard referential lexicon by hand and we follow their basic guidelines in creating our own.

One consideration when dealing with non-adult segmentation is the possibility of legitimate mappings between non-words and objects. For instance, the undersegmentation *abunny* might reasonably be mapped onto the object BUNNY. Our gold standard referential lexicon allows these combinations of determiners and content words as legitimate "words" for an object to be mapped to, unlike the original Frank et al. (2009) study. In contrast, an oversgementation like *du* or *ckie* for *duckie* was not allowed as a correct "word" for the object DUCK. This is because neither unit (*du* or *ckie*) captures the true word form. For instance, it isn't good if the child thinks every instance of /ki/ – *key*, *ckie*, etc. – refers to DUCK. Given this, oversegmention errors are worse than undersegmentations, since they damage the ability to form a reasonable word-object mapping.

5.3 Results

Table 4 presents the evaluation results for all modeled learners, including the segmentation word token F-scores, the rate of oversegmentation errors, and the referential lexicon precision scores. We additionally show the word-object mapping results based on the adult orthographic segmentation as an upper-bound comparison.

	Segmentation		**Mapping**
	F-score	Overseg.	Lex. prec.
Adult Seg	**1.000**	**0.0%**	**0.583**
Batch (Uni)	0.514	1.7%	0.427
Online (Uni)	0.524	9.0%	0.458
Batch (Bi)	0.746	13.8%	0.544
Online (Bi)	0.813	44.8%	0.347
Subtractive Seg	0.833	90.7%	0.336
Random	0.576	53.2%	0.406

Table 4: Average results over five runs from all modeled learners, showing word token F-score segmentation performance, the rate of oversegmentation errors, and the precision of the inferred referential lexicon.

First, we can see that using the adult segmentation yields a referential lexicon with precision 0.583. While this may not seem very high, it is far more precise than other competing word-object mapping strategies investigated by Frank et al. (2009), which had precision scores between 0.06-0.15.

When we turn to the segmentation performance of the learners, we see similar results on the Rollins corpus as we found before. The Bayesian unigram learners have F-scores around 50% (0.514-0.524), which is worse than the random oracle guesser (0.576). In contrast, the Bayesian bigram learners fare much better (0.746-0.813), with almost as good token F-score performance as the subtractive segmenter (0.833).

Interestingly, we see vast differences in the rate of oversegmentation errors. The subtractive segmenter's errors are nearly always oversegmentations (90.7%). The Online Bayesian bigram learner and the random oracle guesser have about half their errors as oversegmentations (44.8%, 53.2%), while the remaining Bayesian learners have very few oversegmentation errors (1.7%-13.8%). Given how damaging oversegmentation errors can be for word-object mapping, we might expect high oversegmentation rates to take their toll despite highly "accurate" word segmentation.

This is precisely what we find for the subtractive segmenter: it has the highest token F-score for segmentation but the worst lexical precision for word-object mappings (0.336). The Online Bayesian bigram learner suffers in lexical precision for a similar reason (0.347), though its oversegmentation bias is lower. Notably, both these learners generate referential lexicons that are worse than what can be achieved by best-case random guessing (0.406). In contrast, the Bayesian learners with very few oversegmentations fare better (0.427-0.544). Given that the best possible performance for lexical precision was 0.583, 0.544 seems quite respectable.

When we examine the mapping errors made by each modeled learner (samples shown in Table 5), the detrimental impact of oversegmentation is more apparent. Notably, many words in English child-directed speech are made up of two syllables (e.g. *birdie*, *bunny*, *piggy*). If these words are oversegmented, the model cannot create a lexical mapping from *birdie* to its object and instead tends to map both *bir* and *die* to the same object. The Bayesian unigram learners never produce these types of oversegmentations for the concrete nouns which the model is attempting to learn (they do, however, produce oversegmentations such as *hip-hop* segmented as *hip* and *hop*). In contrast, the Bayesian bigram learners, the subtractive segmenter, and random oracle learner generate these errors for words that otherwise might have been learned correctly (between 6.4% - 10.2% of all inferred mappings).

	Word	Object	% Over Err
Batch (Bi)	**bu**(nnies)	RABBIT	6.4%
	(bu)**nnies**	RABBIT	
Online (Bi)	(bir)**die**	DUCK	10.2%
	bir(die)	DUCK	
Subtr. Seg	**bu**(nnies)	RABBIT	8.1%
	bir(die)	DUCK	
Random	**pi**(ggy)	PIG	8.5%
	bir(die)	DUCK	

Table 5: Example oversegmentation errors from the four learners that make them for items in the referential lexicon. Oversegmented lexical items are shown in bold with the remainder of the correct word in parentheses. The percentage of all lexical mappings that were incorrect because of oversegmentation is also given.

More generally, similar to the stress utility evaluation, this word-object mapping utility evaluation reveals that segmentations which are more "correct" are not necessarily more useful. In particular, having a non-detrimental segmentation error pattern (i.e., preferring undersegmentation to oversegmentation) may matter more than having a more accurate segmentation for the early stages of both speech segmentation and word-object mapping. However, these results do not necessarily indicate that the online bigram Bayesian or subtractive segmentation strategies are not used by infants. It simply means that if they are, oversegmentations may need to be corrected before word-object mapping can successfully get off the ground. We note that the particular parameters used for the Bayesian strategy can influence the rates of over- and undersegmentation. Because we selected parameters that optimized word token F-score performance, it may be that parameters can be optimized for word-object mapping (and also stress cue induction).

6 Conclusion

We have presented two concrete suggestions for evaluating the utility of speech segmentation strategies, capitalizing on the bootstrapping nature of language acquisition. This utility-focused evaluation approach demonstrates that a more accurate segmentation when compared to a gold standard does not equate to a more useful segmentation for subsequent language acquisition processes. Notably, the types of errors made may significantly impact the utility of the inferred lexicon, so it is worthwhile to analyze not just what is right about a model's output but also exactly what is wrong. This is a specific demonstration of a larger methodological point about how to evaluate unsupervised models of language acquisition. While gold standard evaluation can tell us whether a strategy reproduces adult knowledge, measuring model output utility can indicate what strategies are actually useful for learners.

Acknowledgments

We would like to thank Michael Frank, Ulrike von Luxburg, Sharon Goldwater, Stella Frank, and the members of the Computation of Language Laboratory at UCI for their helpful discussion and comments. This work was supported by a Jean-Claude Falmagne Research Award to the first author from the department of Cognitive Sciences at UCI.

References

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167–206.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bertonicini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology*, *117*(1), 21–33.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of child language*, *37*, 487–511.

Börschinger, B., & Johnson, M. (2014). Exploring the role of stress in Bayesian word segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, *2*(1), 93-104.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Cole, R., & Jakimik, J. (1980). Perception and production of fluent speech. In R. Cole (Ed.), (pp. 133–163). Hillsdale, NJ: Erlbaum.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, *37*(2), 344–377.

Doyle, G., & Levy, R. (2013). Combining multiple information types in bayesian word segmentation. In *Proceedings of naacl-hlt 2013* (pp. 117–126).

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

Feldman, N., Griffiths, T., Goldwater, S., & Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 579–585.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *6*, 721–741.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A bayesian framework for word segmentation. *Cognition*, *112*(1), 21–54.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Jones, B., Johnson, M., & Frank, M. (2010). Learning words and their meanings from unsegmented child-directed speech. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 501–509).

Jusczyk, P., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*(5), 648–654.

Jusczyk, P., Hohne, E., & Baumann, A. (1999). Infants' sensitivity to allphonic cues for word segmentation. *Perception and Psychophysics*, *61*, 1465–1476.

Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kingsbury, P., Strassel, S., McLemore, C., & MacIntyre, R. (1997). *Callhome american english lexicon (pronlex)*. Linguistic Data Consortium.

Landau, B., & Gleitman, L. (1985). *Language and experience*. Cambridge, MA: Harvard University Press.

Lignos, C. (2011). Modeling infant word segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 29–38).

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the 30th west coast conference on formal linguistics* (pp. 237–247).

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marthi, B., Pasula, H., Russell, S., & Peres, Y. (2002). Decayed mcmc filtering. In *Proceedings of 18th uai* (pp. 319–326).

Mattys, S., Jusczyk, P., & Luce, P. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

Mercier, C. (1912). *A new logic*. London: William Heineman.

Morgan, J., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Lawrence Erlbaum Associates, Inc.

Morgan, J., & Saffran, J. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*(4), 911–936.

Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, *8*(2), 107–132. (special issue on computational models of language acquisition)

Pelucchi, B., Hay, J., & Saffran, J. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Perfors, A., Ransom, K., & Navarro, D. (2014). People ignore token frequency when deciding how widely to generalize. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2759–2764).

Peters, A. (1983). *The units of language acquisition*. New York: Cambridge University Press.

Phillips, L., & Pearl, L. (2014a). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the Computational and Cognitive Models of Language Acquisition and Language Processing Workshop.*

Phillips, L., & Pearl, L. (2014b). Bayesian inference as a viable cross-linguistic word segmentation strategy: It's about what's useful. In *Proceedings of the 36th annual conference of the cognitive science society.*

Phillips, L., & Pearl, L. (in press). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*.

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

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Heirarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Theodoridis, S., & Koutroubas, K. (1999). *Pattern recognition*. Academic Press.

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

von Luxburg, U., Williamson, R., & Guyon, I. (2011). Clustering: Science or art? In *JMLR Workshop and Conference Proceedings 27* (pp. 65–79). (Workshop on Unsupervised Learning and Transfer Learning)

Yang, C. (2005). On productivity. *Linguistic variation yearbook*, *5*(1), 265–302.