Testing the Robustness of Online Word Segmentation: Effects of Linguistic Diversity and Phonetic Variation

Luc Boruta, Sharon Peperkamp, Benoît Crabbé, Emmanuel Dupoux

Testing the Robustness of Online Word Segmentation: Effects of Linguistic Diversity and Phonetic Variation

Luc Boruta[1,2], **Sharon Peperkamp**[2], **Benoît Crabbé**[1], and **Emmanuel Dupoux**[2]
[1] Univ. Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA, F-75205, Paris, France
[2] LSCP–DEC, École des Hautes Études en Sciences Sociales, École Normale Supérieure, Centre National de la Recherche Scientifique, F-75005, Paris, France
luc.boruta@inria.fr, peperkamp@ens.fr, benoit.crabbe@inria.fr, emmanuel.dupoux@gmail.com

Abstract

Models of the acquisition of word segmentation are typically evaluated using phonemically transcribed corpora. Accordingly, they implicitly assume that children know how to undo phonetic variation when they learn to extract words from speech. Moreover, whereas models of language acquisition should perform similarly across languages, evaluation is often limited to English samples. Using child-directed corpora of English, French and Japanese, we evaluate the performance of state-of-the-art statistical models given inputs where phonetic variation has not been reduced. To do so, we measure segmentation robustness across different levels of segmental variation, simulating systematic allophonic variation or errors in phoneme recognition. We show that these models do not resist an increase in such variations and do not generalize to typologically different languages. From the perspective of early language acquisition, the results strengthen the hypothesis according to which phonological knowledge is acquired in large part before the construction of a lexicon.

1 Introduction

Speech contains very few explicit boundaries between linguistic units: silent pauses often mark utterance boundaries, but boundaries between smaller units (e.g. words) are absent most of the time. Procedures by which infants could develop word segmentation strategies have been discussed at length, from both a psycholinguistic and a computational point of view. Many models relying on statistical information have been proposed, and some of them exhibit satisfactory performance: MBDP-1 (Brent, 1999), NGS-u (Venkataraman, 2001) and DP (Goldwater, Griffiths and Johnson, 2009) can be considered state-of-the-art. Though there is evidence that prosodic, phonotactic and coarticulation cues may count more than statistics (Johnson and Jusczyk, 2001), it is still a matter of interest to know how much can be learned without linguistic cues. To use Venkataraman's words, we are interested in "the performance of *bare-bones* statistical models."

The aforementioned computational simulations have two major downsides. First, all models of language acquisition should generalize to typologically different languages; however, the word segmentation experiments mentioned above have never been carried out on phonemically transcribed, child-directed speech in languages other than English. Second, these experiments use phonemically transcribed corpora as the input and, as such, make the implicit simplifying assumption that, when children learn to segment speech into words, they have already learned phonological rules and know how to reduce the inherent variability in speech to a finite (and rather small) number of abstract categories: the phonemes. Rytting, Brew and Fosler-Lussier (2010) addressed this issue and replaced the usual phonemic input with probability vectors over a finite set of symbols. Still, this set of symbols is limited to the phonemic inventory of the language: the reduction of phonetic variation is taken for granted. In other words, previous simulations evaluated the performance of the models given idealized input but offered no guarantee as to the performance of the mod-

els on realistic input.

We present a comparative survey that evaluates the extent to which state-of-the-art statistical models of word segmentation resist segmental variation. To do so, we designed a parametric benchmark where more and more variation was gradually introduced into phonemic corpora of child-directed speech. Phonetic variation was simulated applying context-dependent allophonic rules to phonemic corpora. Other corpora in which noise was created by random phoneme substitutions were used as controls. Furthermore, to draw language-independent conclusions, we used corpora from three typologically different languages: English, French and Japanese.

2 Robustness benchmark

2.1 Word segmentation models

The segmentation task can be summarized as follows: given a corpus of utterances in which word boundaries have been deleted, the model has to put them back. Though we did not challenge the usual idealization that children are able to segment speech into discrete, phoneme-sized units, modeling language acquisition imposes significant constraints on the models (Brent, 1999; Gambell and Yang, 2004): they must generalize to different (if not all) languages, start without any knowledge specific to a particular language, learn in an unsupervised manner and, most importantly, operate incrementally.

Online learning is a sound desideratum for any model of language acquisition: indeed, human language-processors do not wait, in Brent's words, "until the corpus of all utterances they will ever hear becomes available". Therefore, we favored an 'infant-plausible' setting and only considered online word segmentation models, namely MBDP-1 (Brent, 1999) and NGS-u (Venkataraman, 2001). Even if DP (Goldwater et al., 2009) was shown to be more flexible than both MBDP-1 and NGS-u, we did not include Goldwater et al.'s batch model, nor recent online variants by Pearl et al. (in press), in the benchmark. All aforementioned models rely on word n-grams statistics and have similar performance, but MBDP-1 and NGS-u are minimally sufficient in providing an quantitative evaluation of how cross-linguistic and/or segmental variation impact the models' performance. We added two random segmentation models as baselines. The four models are described below.

2.1.1 MBDP-1

The first model is Heinz's implementation of Brent's MBDP-1 (Brent, 1999; Heinz, 2006). The general idea is that the best segmentation of an utterance can be inferred from the best segmentation of the whole corpus. However, explicitly searching the space of all possible segmentations of the corpus dramatically increases the model's computational complexity. The implementation thus uses an incremental approach: when the ith utterance is processed, the model computes the best segmentation of the corpus up to the ith utterance included, assuming the segmentation of the first $i-1$ utterances is fixed.

2.1.2 NGS-u

This unigram model was described and implemented by Venkataraman (2001). MBDP-1's problems of complexity were circumvented using an intrinsically incremental n-gram approach. The strategy is to find the most probable word sequence for each utterance, according to information acquired while processing previous utterances. In the end, the segmentation of the entire corpus is the concatenation of each utterance's best segmentation. It is worth noting that NGS-u satisfies all three constraints proposed by Brent: strict incrementality, non-supervision and universality.

2.1.3 Random

This dummy model rewrites its input, uniformly choosing after each segment whether to insert a word boundary or not. It defines a chance line at and below which models can be considered inefficient. The only constraint is that no empty word is allowed, hence no consecutive boundaries.

2.1.4 Random$^+$

The second baseline is weakly supervised: though each utterance is segmented at uniformly-chosen random locations, the correct number of word boundaries is given. This differs from Brent's baseline, which was given the correct number of boundaries to insert in the entire corpus. As before, consecutive boundaries are forbidden.

	English		French		Japanese	
	Tokens	Types	Tokens	Types	Tokens	Types
U	9,790	5,921	10,000	7,660	10,000	6,315
W	33,399	1,321	51,069	1,893	26,609	4,112
P	95,809	50	121,486	35	102,997	49

Table 1: Elementary corpus statistics, including number of utterances (U), words (W) and phonemes (P).

2.2 Corpora

The three corpora we used were derived from transcribed adult-child verbal interactions collected in the CHILDES database (MacWhinney, 2000). For each sample, elementary textual statistics are presented in Table 1. The English corpus contains 9790 utterances from the *Bernstein–Ratner* corpus that were automatically transcribed and manually corrected by Brent and Cartwright (1996). It has been used in many word segmentation experiments (Brent, 1999; Venkataraman, 2001; Batchelder, 2002; Fleck, 2008; Goldwater et al., 2009; among others) and can be considered a *de facto* standard. The French and the Japanese corpora were both made by Le Calvez (2007), the former by automatically transcribing the *Champaud*, *Leveillé* and *Rondal* corpora, the latter by automatically transcribing the *Ishii* and *Noji* corpora from rōmaji to phonemes. To get samples comparable in size to the English corpus, 10,000 utterances were selected at random in each of Le Calvez's corpora. All transcription choices made by the authors in terms of phonemic inventory and word segmentation were respected.[1]

2.3 Variation sources

The main effect of the transformations we applied to the phonemic corpora was the increase in the average number of word forms per word. We refer to this quantity, similar to a type-token ratio, as the corpora's *lexical complexity*. As allophonic variation is context-dependent, the increase in lexical complexity is, in this condition, limited by the phonotactic constraints of the language: the fewer contexts a phoneme appears in, the fewer contextual allophones it can have. By contrast, the upper limit is much higher in the control condition, as phoneme substitutions are context-free.

From a computational point of view, the application of allophonic rules increases both the number of symbols in the alphabet and, as a byproduct, the lexical complexity. Obviously, when any kind of noise or variation is added, there is less information in the data to learn from. We can therefore presume that the probability mass will be scattered, and that as a consequence, statistical models relying on word n-grams statistics will do worse than with phonemic inputs. Yet, we are interested in quantifying how such interference impacts the models' performance.

2.3.1 Allophonic variation

In this experiment, we were interested in the performance of online segmentation models given rich phonetic transcriptions, i.e. the input children process before the acquisition of allophonic rules. Consider the following rule that applies in French:

$$/\text{ʁ}/ \rightarrow \begin{cases} [\chi] & \text{before a voiceless consonant} \\ [\text{ʁ}] & \text{otherwise} \end{cases}$$

The application of this rule creates two contextual variants for /kanaʁ/ (*canard*, 'duck'): [kanaʁ‿ʒon] (*canard jaune*, 'yellow duck') and [kanaχ‿flotɑ̃] (*canard flottant*, 'floating duck'). Before learning the rule, children have to store both [kanaʁ] and [kanaχ] in their emerging lexicon as they are not yet able to undo allophonic variation and construct a single lexical entry: /kanaʁ/.

Daland and Pierrehumbert (2010) compared the performance of a phonotactic segmentation model using canonical phonemic transcripts and transcripts implementing conversational reduction processes. They found that incorporating pronunciation variation has a mild negative impact on performance. However, they used adult-directed speech. Even if, as they argue, reduced adult-directed speech may present a worst-case scenario for infants (compared to hyperarticulated child-direct speech), it offers no quantitative evaluation of the models' performance using child-directed speech.

Because of the lack of phonetically transcribed child-directed speech data, we emulated rich transcriptions applying allophonic rules to the phonemic corpora. To do so, we represented the internal structure of the phonemes in terms of articulatory features and used the algorithm described by Boruta

[1] Some transcription choices made by Brent and Cartwright are questionable (Blanchard and Heinz, 2008). Yet, we used the canonical version of the corpus for the sake of comparability.

(2011) to create artificial allophonic grammars of different sizes containing assimilatory rules whose application contexts span phonologically similar contexts of the target phoneme. Compared to Daland and Pierrehumbert's manual inspection of the transcripts, this automatic approach gives us a finer control on the degree of pronunciation variation. The rules were then applied to our phonemic corpora, thus systematizing coarticulation between adjacent segments. We made two simplifying assumptions about the nature of the rules. First, all allophonic rules we generated are of the type $p \rightarrow a \ / \ \_ \ c$ where a phoneme p is realized as its allophone a before context c. Thus, we did not model rules with left-hand or bilateral contexts. Second, we ensured that no two allophonic rules introduced the same allophone (as in English flapping, where both /t/ and /d/ have an allophone [ɾ]), using parent annotation: each phone is marked by the phoneme it is derived from (e.g. $[ɾ]^{/t/}$ and $[ɾ]^{/d/}$). This was done to avoid probability mass derived from different phonemes merging onto common symbols.

The amount of variation in the corpora is determined by the average number of allophones per phoneme. We refer to this quantity as the corpora's *allophonic complexity*. Thus, at minimal allophonic complexity, each phoneme has only one possible realization (i.e. phonemic transcription), whereas at maximal allophonic complexity, each phoneme has as many realizations as attested contexts. For each language, the range of attested lexical and allophonic complexities obtained using Boruta's (2011) algorithm are reported in Figure 1.

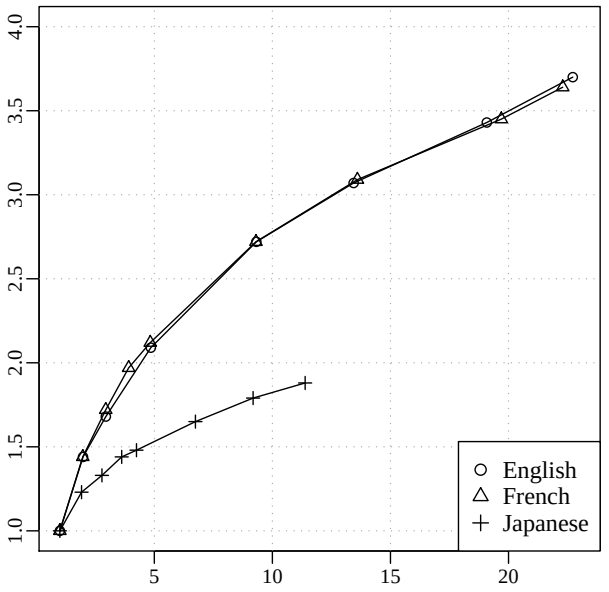

Figure 1: Lexical complexity (the average number of word forms per word) as a function of allophonic complexity (the average number of allophones per phoneme).

2.3.2 Phoneme substitutions

Allophonic variation is not the only type of variation that may interfere with word segmentation. Indeed, the aforementioned simulations assumed that all phonemes are recognized with 100% accuracy, but —due to factors such as noise or speech rate— human processors may mishear words. In this control condition, we examined the models' performance on corpora in which some phonemes were replaced by others. Thus, substitutions increase the corpus' lexical complexity without increasing the number of symbols: phoneme misrecognitions give a straightforward baseline against which to compare the models' performance when allophonic variation has not been reduced. Such corpora can be considered the output of a hypothetical imperfect speech-to-phoneme system or a winner-take-all scalar reduction of Rytting et al.'s (2010) probability vectors.

We used a straightforward model of phoneme misrecognition: substitutions are based neither on a confusion matrix (Nakadai et al., 2007) nor on phoneme similarity. Starting from the phonemic corpus, we generated 10 additional corpora controlling the proportion of misrecognized phonemes, ranging from 0 (perfect recognition) to 1 (constant error) in increments of 0.1. A noise intensity of n means that each phoneme has probability n of being rewritten by another phoneme. The random choice of the substitution phoneme is weighted by the relative frequencies of the phonemes in the corpus. The probability $P(p \rightarrow x)$ that a phoneme x rewrites a phoneme p is defined as

$$P(p \rightarrow x) = \begin{cases} 1 - n & \text{if } p = x \\ n\left(f(x) + \dfrac{f(p)}{|\mathcal{P}| - 1}\right) & \text{otherwise} \end{cases}$$

where n is the noise intensity, $f(x)$ the relative frequency of phoneme x in the corpus and $\mathcal{P}$ the phonemic inventory of the language.

2.4 Evaluation

We used Venkataraman's (2001) implementation of the now-standard evaluation protocol proposed by Brent (1999) and then extended by Goldwater et al. (2009). Obviously, orthographic words are not the optimal target for a model of language acquisition. Yet, in line with previously reported experiments, we used the orthographic segmentation as the standard of correct segmentation.

2.4.1 Scoring

For each model, we report (as percentages) the following scores as functions of the lexical complexity of the corpus:

- P_s, R_s, F_s: precision, recall and F-score on word segmentation as defined by Brent;
- P_l, R_l, F_l: precision, recall and F-score on the induced lexicon of word types: let L be the standard lexicon and L' the one discovered by the algorithm, we define $P_l = |L \cap L'|/|L'|$, $R_l = |L \cap L'|/|L|$ and $F_l = 2 \cdot P_l \cdot R_l/(P_l + R_l)$.

The difference between scoring the segmentation and the lexicon can be exemplified considering the utterance [əwʊdʧʌkwʊdʧʌkwʊd] (*a woodchuck would chuck wood*). If it is segmented as [ə␣wʊdʧʌk␣wʊd␣ʧʌk␣wʊd], both the segmentation and the induced lexicon are correct. By contrast, if it is segmented as [ə␣wʊd␣ʧʌk␣wʊdʧʌk␣wʊd], the lexicon is still accurate while the word segmentation is incorrect. A good segmentation inevitably yields a good lexicon, but the reverse is not necessarily true.

2.4.2 *k*-shuffle cross-validation

As the segmental variation procedures and the segmentation baselines are non-deterministic processes, all scores were averaged over multiple simulations. Moreover, as MBDP-1 and NGS-u operate incrementally, their output is conditioned by the order in which utterances are processed. To lessen the influence of the utterance order, we shuffled the corpora for each simulation. Testing all permutations of the corpora for each combination of parameter values is computationally intractable. Thus, scores reported below were averaged over three distinct simulations with shuffled corpora.

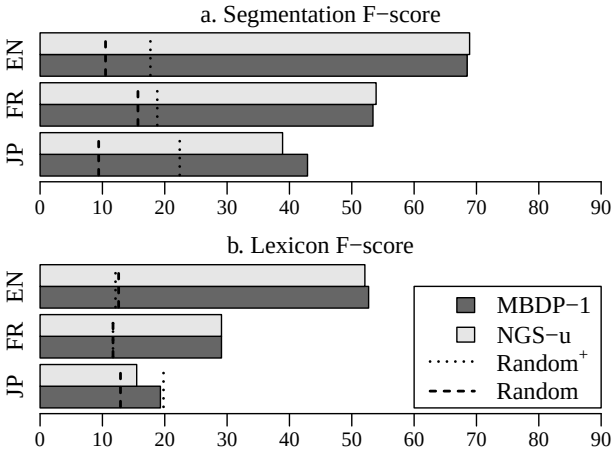

Figure 2: Cross-linguistic performance of MBDP-1 and NGS-u on child-directed phonemic corpora in English (EN), French (FR) and Japanese (JP).

3 Results and discussion

3.1 Cross-linguistic evaluation

Performance of the segmentation models[2] on phonemic corpora is presented in Figure 1 in terms of F_s- and F_l-score (upper and lower panel, respectively). We were able to replicate previous results on English by Brent and Venkataraman almost exactly; the small difference, less than one percent, was probably caused by the use of different implementations.

From a cross-linguistic point of view, the main observation is that these models do not seem to generalize to typologically different languages. Whereas MBDP-1 and NGS-u's F_s value is 69% for English, it is only 54% for French and 41% for Japanese. Similar observations can be made for F_l. Purely statistical strategies seem to be particularly ineffective on our Japanese sample: inserting word boundaries at random yields a better lexicon than using probabilistic models.

A crude way to determine whether a word segmentation model tends to break words apart (oversegmentation) or to cluster various words in a single chunk (under-segmentation) is to compare the average word length (AWL) in its output to the AWL in the standard segmentation. If the output's AWL is greater than the standard's, then the output is undersegmented, and *vice versa*. Even if NGS-u produces

[2]The full table of scores for each language, variation source, and segmentation model was not included due to space limitations. It is available upon request from the first author.

shorter words than MBDP-1, both models exhibit, once again, similar within-language behaviors. English was slightly under-segmented by MBDP-1 and over-segmented by NGS-u: ouputs' AWL are respectively 3.1 and 2.7, while the standard is 2.9. Our results are consistent with what Goldwater et al. (2009) observed for DP: error analysis shows that both MBDP-1 and NGS-u also break off frequent English morphological affixes, namely /ɪŋ/ (*-ing*) and /s,z/ (*-s*). As for French, AWL values suggest the corpus was under-segmented: 3.1 for MBDP-1's output and 2.9 for NGS-u's, while the standard is 2.4. On the contrary, Japanese was heavily over-segmented: many monophonemic words emerged and, whereas the standard AWL is 3.9, the ouputs' AWL is 2.7 for both models.

Over-segmentation may be correlated to the number of syllable types in the language: English and French phonotactics allow consonantal clusters, bringing the number of syllable types to a few thousands. By contrast, Japanese has a much simpler syllabic structure and less syllable types which, as a consequence, are often repeated and may (incorrectly) be considered as words by statistical models. The fact that the models do worse for French and Japanese is not especially surprising: both languages have many more affixal morphemes than English. Consider French, where the lexical autonomy of clitics is questionable: whereas /s/ (*s'* or *c'*) or /k/ (*qu'*) are highly frequent words in our orthographic standard, many errors are due to the agglutination of these clitics to the following word. These are counted as segmentation errors, but should they?

Furthermore, none of the segmentation models we benchmarked exhibit similar performance across languages: invariably, they perform better on English. There may be a correlation between the performance of segmentation models and the percentage of word hapaxes, i.e. words which occur only once in the corpus: the English, French and Japanese corpora contain 31.7%, 37.1% and 60.7% of word hapaxes, respectively. The more words tend to occur only once, the less MBDP-1, NGS-u and DP perform on segmentation. This is consistent with the usual assumption that infants use familiar words to find new ones. It may also be the case that these models are not implicitly tuned to English, but that the contribution of statistical cues to word segmentation differs across languages. In French, for example, stress invariably marks the end of a word (although the end of a word is not necessarily marked by stress). By contrast, there are languages like English or Spanish where stress is less predictable: children cannot rely solely on this cue to extract words and may thus have to give more weight to statistics.

3.2 Robustness to segmental variation

The performance of MBDP-1, NGS-u and the two baselines on inputs altered by segmental variation is presented in Figure 2.[3] The first general observation is that, as predicted, MBDP-1 and NGS-u do not seem to resist an increase in lexical complexity. In the case of allophonic variation, their performance is inversely related to the corpora's allophonic complexity. However, as suggested by the change in the graphs' slope, performance for English seems to stabilize at 2 word forms per word. Similar observations can be made for French and Japanese on which the performance of the models is even worse: F_l values are below chance at 1.7 and 3 variants per word for Japanese and French, respectively; likewise, F_s is below chance at 1.5 for Japanese and 2.5 for French. Phoneme substitutions also impede the performance of MBDP-1 and NGS-u: the more phonemes are substituted, the more difficult it becomes for the algorithms to learn how to insert word boundaries. Furthemore, F_l is below chance for complexities greater than 4 for French, and approximately 2.5 for Japanese. It is worth noting that, in both conditions, the models exhibit similar within-language performance as the complexity increases.

The potential lexicon that can be built by combining segments into words may account for the discrepancy between the two conditions, as it is in fact the models' search space. In the control condition, substituting phonemes does not increase its size. However, the likelihood of a given phoneme in a given word being replaced by the same substitution phoneme decreases as words get longer. Thus, the proportion of hapax increases, making statistical segmentation harder to achieve. By contrast, the

[3] For the control condition, we did not graph scores for noise intensities greater than 0.2: 80% accuracy is comparable to the error rates of state-of-the-art systems in speaker-independent, continuous speech recognition (Makhoul and Schwartz, 1995).

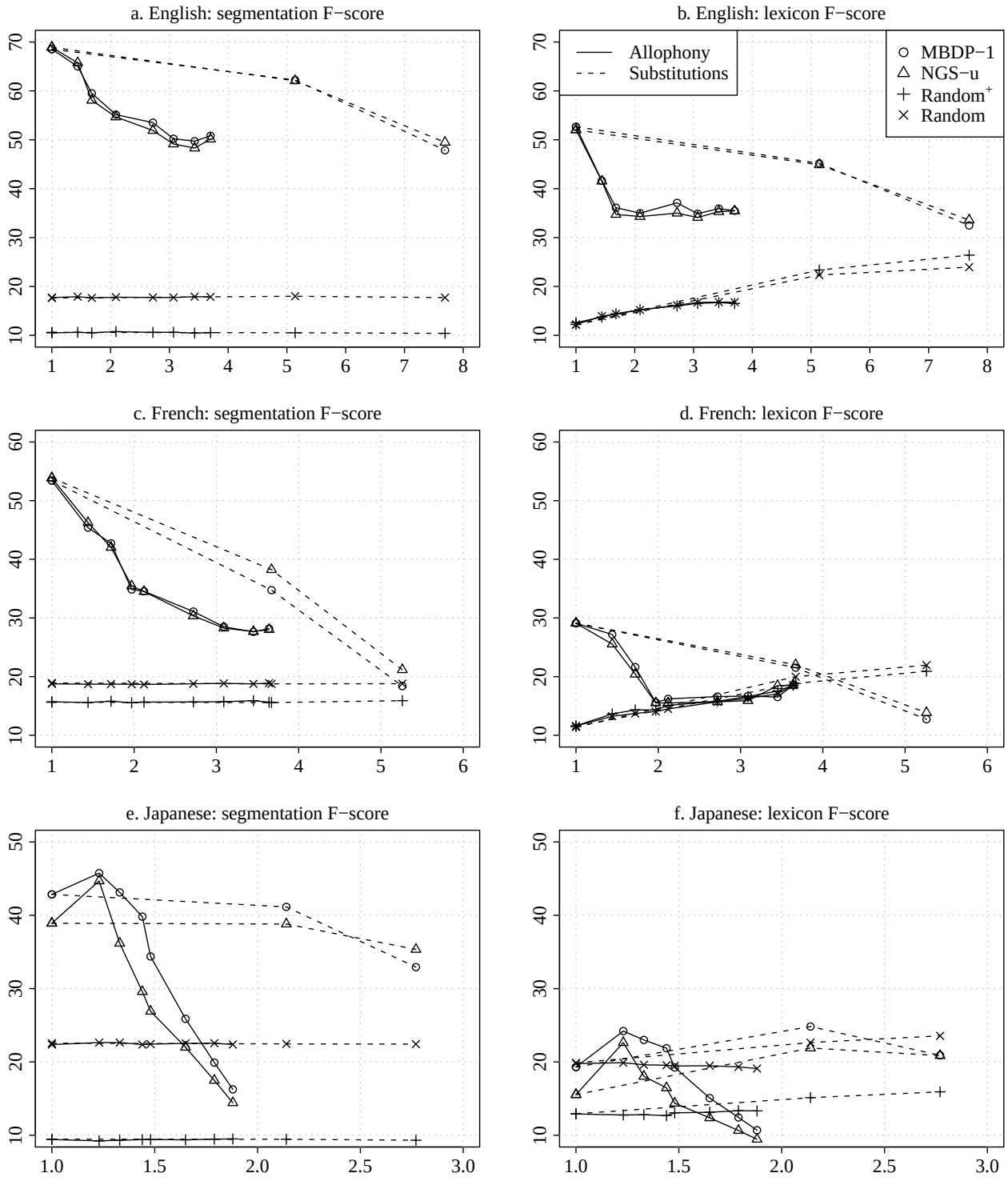

Figure 3: F_s-score (left column) and F_l-score (right column) as functions of the lexical complexity, i.e. the number of word forms per word, in the English (top row), French (middle row) and Japanese (bottom row) corpora.

application of allophonic rules increases the number of objects to build words with; as a consequence, the size of the potential lexicon explodes.

As neither MBDP-1 nor NGS-u is designed to handle noise, the results are unsurprising. Indeed, any word form found by these models will be incorporated in the lexicon: if [læŋgwɪʧ] and [læŋgwɪʤ] are both found in the corpus, these variants will be included as is in the lexicon. There is no mechanism for 'explaining away' data that appear to have been generated by systematic variation or random noise. It is an open issue for future research to create robust models of word segmentation that can handle segmental variation.

4 Conclusions

We have shown, first, that online statistical models of word segmentation that rely on word n-gram statistics do not generalize to typologically different languages. As opposed to French and Japanese, English seems to be easier to segment using only statistical information. Such differences in performance from one language to another emphasize the relevance of cross-linguistic studies: any conclusion drawn from the monolingual evaluation of a model of language acquisition should be considered with all proper reservations. Second, our results quantify how imperfect, though realistic, inputs impact MBDP-1's and NGS-u's performance. Indeed, both models become less and less efficient in discovering words in transcribed child-directed speech as the number of variants per word increases: though the performance drop we observed is not surprising, it is worth noting that both models are less efficient than random procedures at about twenty allophones per phoneme. However, the number of context-dependent allophones we introduced is far less than what is used by state-of-the-art models of speech recognition (Makhoul and Schwartz, 1995).

To our knowledge, there is no computational model of word segmentation that both respects the constraints imposed on a human learner and accommodates noise. This highlights the complexity of early language acquisition: while no accurate lexicon can be learned without a good segmentation strategy, state-of-the-art models fail to deliver good segmentations in non-idealized settings. Our results also emphasize the importance of other cues for word segmentation: statistical learning may be helpful or necessary for word segmentation, but it is unlikely that it is sufficient.

The mediocre performance of the models strengthens the hypotheses that phonological knowledge is acquired in large part before the construction of a lexicon (Jusczyk, 1997), or that allophonic rules and word segmentations could be acquired jointly (so that neither is a prerequisite for the other): children cannot extract words from fluent speech without knowing how to undo at least part of contextual variation. Thus, the knowledge of allophonic rules seems to be a prerequisite for accurate segmentation. Recent simulations of word segmentation and lexical induction suggest that using phonological knowledge (Venkataraman, 2001; Blanchard and Heinz, 2008), modeling morphophonological structure (Johnson, 2008) or preserving subsegmental variation (Rytting et al., 2010) invariably increases performance. *Vice versa*, Martin et al. (submitted) have shown that the algorithm proposed by Peperkamp et al. (2006) for undoing allophonic variation crashes in the face of realistic input (i.e. many allophones), and that it can be saved if it has approximate knowledge of word boundaries. Further research is needed, at both an experimental and a computational level, to explore the performance and suitability of an online model that combines the acquisition of allophonic variation with that of word segmentation.

References

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

D. Blanchard and J. Heinz. 2008. Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of the Conference on Natural Language Learning*, pages 65–72.

L. Boruta. 2011. A note on the generation of allophonic rules. Technical Report 0401, INRIA.

M. R. Brent and T. A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

M. R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1–3):71–105.

R. Daland and J. B. Pierrehumbert. 2010. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-2008*, pages 130–138.

T. Gambell and C. Yang. 2004. Statistics learning and universal grammar: Modeling word segmentation. In *Proceedings of the 20th International Conference on Computational Linguistics*.

S. Goldwater, T. L. Griffiths, and M. Johnson. 2009. A bayesian framework for word segmentation: exploring the effects of context. *Cognition*, 112(1):21–54.

J. Heinz. 2006. MBDP-1, OCaml implementation. Retrieved from http://phonology.cogsci.udel.edu/∼heinz/ on January 26, 2009.

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

M. Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the 10th Meeting of ACL SIGMORPHON*, pages 20–27.

P. Jusczyk. 1997. *The Discovery of Spoken Language*. MIT Press.

R. Le Calvez. 2007. *Approche computationnelle de l'acquisition précoce des phonèmes*. Ph.D. thesis, UPMC.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Elbraum Associates.

J. Makhoul and R. Schwartz. 1995. State of the art in continuous speech recognition. *PNAS*, 92:9956–9963.

A. Martin, S. Peperkamp, and E. Dupoux. Submitted. Learning phonemes with a pseudo-lexicon.

K. Nakadai, R. Sumiya, M. Nakano, K. Ichige, Y. Hirose, and H. Tsujino. 2007. The design of phoneme grouping for coarse phoneme recognition. In *IEA/AIE*, pages 905–914.

L. Pearl, Sh. Goldwater, and M. Steyvers. In press. Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*.

S. Peperkamp, R. Le Calvez, J. P. Nadal, and E. Dupoux. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

C. A. Rytting, C. Brew, and E. Fosler-Lussier. 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, 37:513–543.

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