The effects of lexical phonotactics, saturation, and frequency skew on segmentability

Benjamin Börschinger
Macquarie University
Heidelberg University

Robert Daland
Linguistics
UCLA

Abdellah Fourtassi
LCSP
ENS/EHESS/CNRS

Emmanuel Dupoux
LCSP
ENS/EHESS/CNRS

benjamin.borschinger@mq.edu.au
{r.daland|abdellah.fourtassi|emmanuel.dupoux}@gmail.com

Abstract

Previous works have proposed that the 'segmentability' of a language depends on its phonotactic structure and can be measured as the entropy of licit segmentations of a corpus sample. These proposals are tested here by generating artificial languages and measuring their segmentability. Maximally permissive and restrictive grammars (Pseudo-Berber, Pseudo-Senufo) were used to generate corpus samples in which the lexical saturation and frequency distributions were varied parametrically. The results show heretofore unsuspected nuances of the relationship between phonotactic complexity, word length, and word segmentation.

1 Introduction

Word segmentation is the perceptual process by which infant and adult listeners parse the continuous speech signal into a sequence of discrete, word-like units. The acquisition of word segmentation has been the subject of intense computational scrutiny, where it is typically operationalized as the unsupervised partitioning of a phonemically transcribed, child-directed corpus, and assessed against an orthographically derived gold standard (Goldwater et al., 2009; Daland and Pierrehumbert, 2011; Pearl et al., 2010).

A number of modeling studies indicates that across a typologically diverse range of languages (e.g. Arabic, Japanese, Korean, Russian, and Spanish), and across a range of models (lexical, phonotactic, and hybrid), better segmentation is unfailingly found for English than for other languages (Fleck, 2008; Daland, 2009; Daland and Pierrehumbert, 2011; Fourtassi et al., 2013; Daland and Zuraw, 2013). The empirical data bearing on this point is sparse, and occasionally conflicting (Nazzi et al., 2006; Nazzi et al., 2014). Therefore, while current models may be insufficient, it seems worth considering the possibility that languages genuinely differ in their intrinsic ease of segmentation. Along those lines, Daland and Zuraw (2013) investigated the segmentability of Korean using a phonotactic segmentation model (Daland and Pierrehumbert, 2011), finding that Korean's many edge-sensitive phonological processes did not help word segmentation. Fourtassi et al. (2013) compared the segmentability of Japanese and English using lexical segmentation models (Goldwater et al., 2009; Johnson and Goldwater, 2009) and argued that the comparatively poor segmentability of Japanese was strongly predicted by the 'normalized segmentation entropy' of a corpus, the average (per-character) entropy over all licit segmentations of a corpus given the gold standard lexicon. Both Daland and Zuraw and Fourtassi et al. speculated that the poor segmentability of these languages bears close connection to their restrictive phonotactics.

The present study seeks to test this proposal by applying a popular Bayesian word segmentation model (Brent, 1999; Goldwater, 2007; Goldwater et al., 2009) to artificially generated corpora. Artificial corpora are used here for the same reason that psycholinguistic experiments use carefully controlled stimuli: natural language corpora variation arises from complex and potentially not understood relationships between interacting language properties. For example, it stands to reason that the word frequency distribution might differ between, e.g., Japanese and English for essentially syntactic (and morphological) reasons. By generating artificial corpora, it is possible to vary one property (phonotactics, frequency distributions, word lengths) while holding other factors constant, controlling (to a certain extent) the impact of these additional factors. In fact, this is exactly what the present paper does.

2 Methods

Two artificial grammars were created, representing extremes of phonotactic permissiveness and restrictiveness. Pseudo-Senufo is a strict CV grammar – words must begin with a consonant and end with a vowel. Pseudo-Berber has no phonotactic constraints whatsoever – every possible sequence of consonants and vowels is grammatical. Artificial lexicons were generated for each grammar as a sample from the space of phonotactically licit forms; average word length (and lexical saturation) were manipulated using a gradient penalty (see below). From a generated lexicon, a corpus was generated by imposing a word frequency distribution. The word frequency distribution can be manipulated (see below); thus, this generation dissociates phonotactic properties from word frequency distribution properties.

2.1 Generating lexicons from phonotactic grammars

A *phonotactic grammar* assigns a probability distribution over possible lexical items ($\mu : \Sigma^* \rightarrow [0, 1]$). The alphabet Σ used here defined a small but standard segmental inventory: [ptk—fsx—mnŋ—lr—wj—aeiou]. Phonotactic grammars were defined using the maximum entropy harmonic grammar framework (Hayes and Wilson, 2008): 'features' are constraints penalizing particular dimensions of ill-formedness (e.g. the constraint ONSET penalizes words which begin with a vowel). Each constraint is associated with a weight, and the 'harmony' of a string is the weighted sum of its constraint violations. The log-odds of two forms is the difference in their harmonies, which is sufficient to determine a log-linear model. Thus, a collection of constraints and their weights serves to define a phonotactic grammar. A *lexicon* is generated as a sample (without replacement) from the probability distribution defined by a phonotactic grammar, i.e. a list of n distinct word types. For tractability, a hard maximum string length of 6 was imposed on lexical forms (i.e. $\Sigma^* \rightarrow \Sigma^{\leq 6}$). The Pseudo-Senufo grammar included 4 constraints favoring a CV syllable structure (ONSET, *CODA, *HIATUS, *COMPLEX). These constraints were given a sufficiently high weight (-25) as to categorically enforce a CV structure in sample lexicons. It also includes a constraint, *STRUCT, which penalizes words according to their length, gradiently favoring shorter words. The weight of *STRUCT was manipulated, yielding varying degrees of pressure for Pseudo-Senufo grammars to 'saturate' the space of phonotactically licit short forms. The Pseudo-Berber grammar included only *STRUCT, allowing any sequence of segments as a licit word-form ('phoneme salad').

2.2 Generating a corpus from a lexicon

The distribution of patterns in their input plays a crucial role for the outcome of learning for statistical models. Side-stepping the complex interactions between semantics, syntax and morphology that give rise to word distributions in real languages we generate artificial corpora mimicking, to varying degrees, the token distribution of natural language. As a blue-print, we take the type and token statistics from the Brent-Bernstein-Ratner corpus (Brent, 1999), containing a specific distribution for 1365 distinct types and 33347 tokens. From this, we can generate artificial corpora that match this distribution by first sampling 1365 distinct types from the phonotactic grammars described above and then generating a random permutation of 33347 tokens that match the original distribution perfectly.[1] We refer to corpora that follow this distribution as BBR, and also experiment with a distribution in which all types occur with (roughly) the same frequency (FLAT) to see in how far the frequency distribution has an impact on segmentability.

2.3 Unigram word segmentation

The Bayesian model for word segmentation we use is a generative model for sequences of words, built on the non-parametric Dirichlet Process (DP) prior. For reasons of space, we refer the reader to (Goldwater, 2007) for details, providing only a brief explanation. The goal of applying the model to unsegmented input is to identify a segmentation that is compatible with a compact probabilistic lexicon, assigning most of the probability mass to a small number of reusable words. The model assumes a pre-specified prior distribution over admissible words, and here we follow previous work (Brent, 1999; Goldwater, 2007) in using a simple lexical generator: it assigns geometrically decaying probabilities to arbitrary se-

[1] To generate utterance boundaries, we assume a stopping probability of $\frac{1}{3}$, resulting in corpora with identical numbers of types and tokens but slightly differing numbers of utterances.

quences of any length and allows any segment to occur anywhere in a word with equal probability. We deliberately opt for such a naïve prior to focus on the information conveyed by the input, rather than any linguistically motivated prior biases learners might bring to the task as studied in, among others, (Johnson and Goldwater, 2009) and (Börschinger et al., 2012). Crucially, the DP prior favors lexicons that exhibit a Zipfian rich-get-richer dynamic with relatively few high-frequency items and a long tail of low-frequency items, penalizing degenerate solutions in which every utterance counts as its own word-type. Also, the model assumes that words within an utterance are independent, hence it can be seen as a Unigram language model over the infinite space of admissible words. The goal of learning on this view is identifying the finite number of words that provide a compact explanation of the unsegmented input. While the Unigram assumption has been shown to result in undersegmentation of frequent collocations, we chose this model for its simplicity and because the questions of this paper are by and large independent of this known shortcoming. In particular, our artificial languages fit the Unigram assumption by construction, allowing us to focus on phonotactics and word-frequency related issues.

3 Results

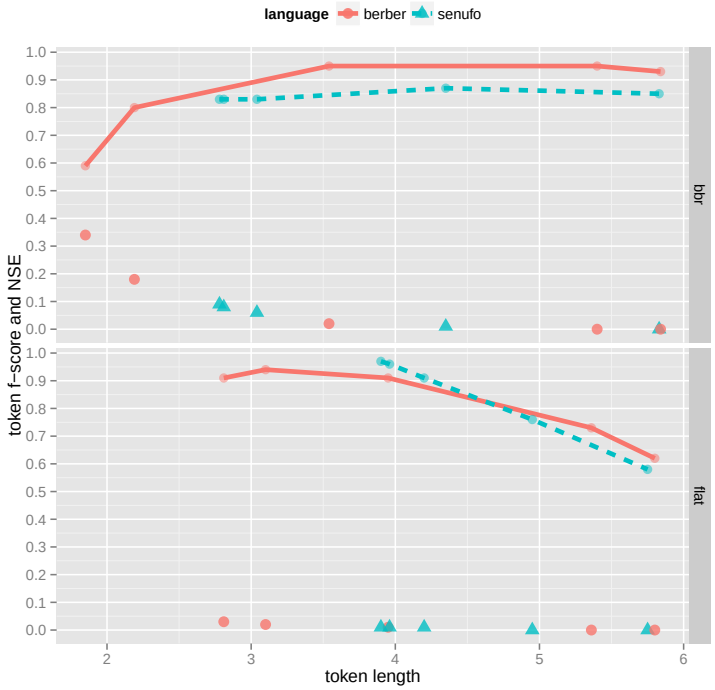
Our experimental paradigm follows closely that of (Johnson and Goldwater, 2009). For each of our artificial corpora we run four independent Markov Chains for the Unigram model for 1000 iterations using the Adaptor Grammar software (Johnson et al., 2007), using hyper-parameter sampling and calculating a single marginal maximum a posteriori segmentation for each utterance from the samples collected across all 4 chains during the last 200 iterations. We evaluate segmentation accuracy by calculating the token f-score according to the gold boundaries in the artificial corpora. Token f-score is the harmonic mean of token precision, i.e. the fraction of correctly identified tokens in the posited segmentation over all posited tokens, and token recall, i.e. the fraction of correctly recovered tokens over all tokens in the gold standard.

The results are given in Figure 1. Overall, pseudo-Berber seems to be segmented better than pseudo-Senufo, with its best token f-scores for BBR being 95%, compared to 87% for pseudo-Senufo. In the unnatural FLAT condition, pseudo-Senufo peaks with 97%, with pseudo-Berber also reaching 94%. There also is a clear impact of the *STRUCT constraint that governs the average word lengths, although the importance of this constraint varies considerably. For BBR, pseudo-Senufo yields a token f-score of 80%+x for all values of *STRUCT whereas pseudo-Berber drops to 59% for *STRUCT=5 but yields a considerably higher token f-score than pseudo-Senufo for *STRUCT < 4. In contrast, FLAT yields the overall best token f-scores for both pseudo-Berber and pseudo-Senufo for high values of *STRUCT and results in worse performance for smaller values of *STRUCT.

4 Discussion

The results raise several questions. First, it is surprising that the phonotactically much less constrained pseudo-Berber is, on average, more easy to segment, yielding considerably higher token f-scores for the natural condition than pseudo-Senufo except for *STRUCT=5. Even if the segmentation model does not explicitly consider phonotactic cues, a natural expectation would have been that regular phonotactics yield a more regular pattern of lexical forms that would make identification of repeating units easier. There is also considerable variability of segmentability as a function of the frequency distribution and the average word length, with the overall best segmentability achieved for a highly unnatural flat distribution with predominantly short words. This is especially curious, given that the flat distribution violates the prior expectation about of the segmentation model about word frequencies much stronger than the natural one.

The first observation lends itself to an explanation along the lines of (Fourtassi et al., 2013): we calculate the normalized segmentation entropy (NSE) for the artificial corpora, a measure of how much ambiguity with respect to word segmentation remains if the gold lexicon is provided to the learner, 0 indicating no uncertainty at all and 1.0 indicating complete uncertainty; we refer the reader to (Fourtassi et al., 2013) for a more detailed explanation. For both BBR and ZIPF, NSE provides a satisfying explanation for the variation in segmentability. Note that for pseudo-Berber, NSE tends to be lower than for pseudo-Senufo for *STRUCT-values up to roughly 3, and these are exactly the conditions where pseudo-Berber attains higher token f-scores. The trend



S	L	LA	SAT
1	pseudo-Senufo	.03	0.01
	pseudo-Berber	.0	0.00
2	pseudo-Senufo	1.03	0.33
	pseudo-Berber	.05	0.00
3	pseudo-Senufo	.99	0.88
	pseudo-Berber	1.59	0.36
4	pseudo-Senufo	.96	0.98
	pseudo-Berber	1.77	0.90
5	pseudo-Senufo	.95	1.00
	pseudo-Berber	1.75	0.99

Figure 1: On the left, token f-score (lines) and normalized segmentation entropy (dots) for the different settings, plotted against average token-length. The right table lists lexicon ambiguity (LA) and saturation (SAT) for pseudo-Berber (b) and pseudo-Senufo (s), grouped by *STRUCT. Small values of *STRUCT in the table correspond to large average token-length in the figure, and high values to short average token-length.

reverses for *STRUCT>3, again as would be predicted by looking at NSE. In sum, NSE and token f-score are highly correlated ($R = 0.81 \pm .13$). To understand why SE varies the way it does, note that for any given word-length, there are strictly less possible word types for pseudo-Senufo than for pseudo-Berber. For example, at the length of 2, there are 65 CV sequences licit for pseudo-Senufo, while there are 18^2 licit XX sequences for pseudo-Berber. Even in the absence of strong penalties on word length, it will in general be necessary for Pseudo-Senufo to recycle shorter words as sub-parts of longer words. This picture reverses, however, if the use of long words is heavily penalized. For *STRUCT>3, pseudo-Berber is effectively limited to words of length 3 to 4 phonemes because longer sequences are penalized so strongly. Now, the absence of any constraint on possible word-forms turns into a disadvantage, due to the high probability of high-frequency single-segment "words". Single-segment words cause high segmentation ambiguity because they might be segmented out (incorrectly) whenever the segment occurs. This issue is ameliorated in pseudo-Senufo, where the phonotactic requirements categorically disallow single-segment words. This is apparent for *STRUCT=4 and *STRUCT=5: pseudo-Senufo shows only an increase in SE from 0.08 to 0.09, while pseudo-Berber's SE changes from 0.18 to 0.34, and this huge increase in ambiguity is directly reflected in the low token f-score of 59%, compared to 83% for pseudo-Senufo. This difference between pseudo-Senufo and pseudo-Berber can also be quantified through lexical saturation (SAT) and lexicon ambiguity (LA), also given in Figure 1. For a given set of word types, its lexical saturation with respect to the language it belongs to is defined as the probability of a randomly sampled word from that language to already be included in this set. We approximate these values for pseudo-Berber and pseudo-Senufo by simulation. A second measure is (within-)lexicon ambiguity: we use a dynamic program to calculate the average number of parses of each lexicon type given the other words in the lexicon for the artificial corpora we generated. These additional measures allow us to refine the explanation provided by SE alone. Note, for example, that even though for both pseudo-Berber and pseudo-Senufo NSE is essentially 0 for *STRUCT<3, LA and SAT are already considerably higher for pseudo-Senufo than for pseudo-Berber, due to the former's CV-constraint on possible words. While this higher ambiguity is not reflected in NSE because knowledge of the gold-lexicon and frequen-

cies is assumed, the actual segmentation is performed fully unsupervised and is consequently affected by it, accounting for the difference between pseudo-Berber and pseudo-Senufo.

This explanation does not only pertains to the artificial setting of this paper, it also provides strong support to the previous suggestions by (Daland and Zuraw, 2013) and (Fourtassi et al., 2013) that restrictive phonotactics can have a *negative* impact on segmentability. While the languages used here are highly simplified as compared to natural languages, the problematic phenomena do occur in natural languages. For example, the Slavic languages have numerous single-segment prepositions (e.g. Russian: /s/ 'with', /k/ 'to (dative)', /v/ 'in'), and the segmentation of these items is an open research problem.

There is a final outstanding puzzle in these results, pertaining to the FLAT distribution. Recall that in this distribution, each type was generated with an equal frequency (about 21 tokens); this 'unnatural' frequency distribution violates the model's assumptions about the frequency distribution much more strongly than BBR, and yet the result is comparable or actually better segmentation for *STRUCT $>$ 2, while it is significantly worse for *STRUCT $\leq$ 2, for both pseudo-Senufo and pseudo-Berber. These results do not lend themselves to an explanation through NSE which is close to zero throughout this distribution. Instead, we speculate that this pattern arises from an interaction between (i) the balanced evidence for each wordform, and (ii) the model's assumptions about word lengths. Recall that its prior on admissible words assumes a geometric distribution over word lengths. Following previous work, which reported that segmentation on natural language data is largely invariant to variation of this (Goldwater et al., 2009), we set $Pr(\text{EOW}) = \frac{1}{2}$, yielding an expected average length of 2. When *STRUCT is large enough, words appear to be 'close enough' to the expected length. We speculate that the superior performance in these cases arises from an abundance of evidence about each word type; in other words the absence of expected ultra-high-frequency types does not hurt segmentation as much as the absence of ultra-low-frequency items helps it. However when *STRUCT is too low, words are quite long (the average is not far below the ceiling of 6), incurring a large penalty from the geometric prior. In this case, we speculate that the absence of high-frequency words turns into a severe disadvantage. We suspect segmentation to be better for BRB because high-frequency items are picked out reliably by statistical learners, allowing them to overcome the larger penalty arising from a mismatch between the geometric prior and the data. Indeed, we performed additional experiments for the FLAT distribution and found that changing the expected word-length yields considerably better scores for *STRUCT$<$3, as expected.The 'moral' from this is that high-frequency elements provide an 'anchor' for statistical learning mechanisms, allowing them overcome moderate mismatches between the data and expectations, a point which is consistent with recent experimental work (Kurumada et al., 2013).

5 Conclusion

We used artificial language corpora to test the idea that more restrictive phonotactics negatively impact statistical word segmentation models. Three variables were manipulated: categorical phontoactics (strict CV versus 'phoneme salad'), the pressure for words to be short (*STRUCT), and the word frequency distribution. The results show a complex interaction between these three factors, which may be summarized as follows: when the pressure for words to be short is very high, the language is forced to recycle existing words as subparts of longer words, yielding intrinsic segmentation abiguity. In this case, restrictive phonotactics provide a modest benefit, by placing a lower limit on the sub-parts that can be recycled. However when there is less pressure for words to be short, restrictive phonotactic simply reduce the space of licit forms, encouraging the reuse of sub-parts which causes a modest decrement. Natural language distributions appear to facilitate word segmentation by providing a small number of high-frequency words, which serve as 'anchors' to facilitate segmentation, even in the case of mismatches between phonotactic expectations and the observable language data. This work has caused us to look at the relationship between word shape, word length, and word frequency in a way that we did not before; it illustrates how artificial languages can tease out subtle predictions of computational models, address specific hypotheses about the design of language and the cognitive tools we bring to learning it.

References

Benjamin Börschinger, Katherine Demuth, and Mark Johnson. 2012. Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 325–340, Mumbai, India. Coling 2012 Organizing Committee.

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

Robert Daland and Janet B. Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Robert Daland and Kie Zuraw. 2013. Does korean defeat phonotactic word segmentation? In *Proceedings of the 51st ACL in Sofia, Bulgaria*.

Robert Daland. 2009. *Word segmentation, word recognition, and word learning: a computational model of first language acquisition*. Ph.D. thesis, Northwestern University.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio, June. Association for Computational Linguistics.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. whyisenglishsoeasytosegment? In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics*.

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

Sharon Goldwater. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, June. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Chigusa Kurumada, Stephan C Meylan, and Michael C Frank. 2013. Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127(3):439–453.

Thierry Nazzi, Galina Iakimova, Josiane Bertoncini, Séverine Frédonie, and Carmela Alcantara. 2006. Early segmentation of fluent speech by infants acquiring french: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54(3):283–299.

Thierry Nazzi, Karima Mersad, Megha Sundara, Galina Iakimova, and Linda Polka. 2014. Early word segmentation in infants acquiring Parisian French: task-dependent and dialect-specific aspects. *Journal of Child Language*.

Lisa Pearl, Sharon Goldwater, and Mark Steyvers. 2010. Online learning mechanisms for Bayesian models of word segmentation. *Research on Language and Computation*, 8(2):107–132.