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

Benjamin Börschinger
Department of Computing
Macquarie 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; **?**). 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—mn$n$—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) and an artificial power-law distribution that exhibits a long-tail distribution similar yet different to that of an actual natural language sample (ZIPF).

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 pre-

[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.

vious work (Brent, 1999; Goldwater, 2007) in using a simple lexical generator: it assigns geometrically decaying probabilities to arbitrary sequences 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 Table 1. Overall, pseudo-Berber seems to be segmented better than pseudo-Senufo, with its best token f-scores for BBR and ZIPF being 95% and 92% respectively, compared to 87% and 83% for pseudo-Senufo. Only in the highly unnatural FLAT condition, pseudo-Senufo peaks with 97%, with pseudo-Berber only 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. The ZIPF condition seems to favour lower values of *STRUCT equally for both pseudo-Berber and pseudo-Senufo; again, pseudo-Berber attains the worst token f-score of 45% (compared to 67% for pseudo-Senufo) for *STRUCT=5. In contrast, FLAT looks like a mirror of ZIPF, yielding the overall best token f-scores for both pseudo-Berber and pseudo-Senufo for high values of *STRUCT and bad performance for smaller ones.

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 the highest token f-score for both the natural and the artificial Zipfian condition. 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. Curiously, the flat distribution violates the prior expectation of the DP based Unigram model much stronger than either the natural or the artificial Zipfian condition, nevertheless resulting in the overall highest token f-scores for both pseudo-Berber and pseudo-Senufo.

The first observation lends itself to an explanation along the lines of (Fourtassi et al., 2013): we calculate the normalized segmentation entropy (SE) for the artificial corpora, a measure of how much ambiguity with respect to word segmenta-

		BBR			ZIPF			FLAT		
S	L	TF	SE	TL	TF	SE	TL	TF	SE	TL
1	s	.85	.0	5.83	.83	.0	5.87	.58	.0	5.75
	b	.93	.0	5.84	.91	.0	5.86	.62	.0	5.80
2	s	.87	.01	4.35	.83	.0	4.94	.76	.0	4.95
	b	**.95**	.0	5.40	**.92**	.0	5.48	.73	.0	5.36
3	s	.83	.06	3.04	.71	.04	3.21	.91	.01	4.20
	b	**.95**	.02	3.54	.87	.03	2.97	.91	.01	3.95
4	s	.83	.08	2.81	.68	.08	2.62	.96	.01	3.96
	b	.80	.18	2.19	.62	.24	1.96	.94	.02	3.10
5	s	.83	.09	2.78	.67	.08	2.60	**.97**	.01	3.90
	b	.59	.34	1.85	.45	.39	1.67	.91	.03	2.81

Table 1: Experimental results, grouped by different values of the *STRUCT constraint (S column) and language (L column, with "s" for pseudo-Senufo and "b" for pseudo-Berber). "TF" stands for token f-score, "SE" for normalized segmentation entropy and "TL" for the average gold token length for that condition.

tion 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, SE provides a satisfying explanation for the variation in segmentability. Note that for pseudo-Berber, SE 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 reverses for *STRUCT>3, again as would be predicted by looking at the SE. According to a Pearson's product-moment correlation, the 95% confidence interval for the correlation between SE and TF is $(-0.94, -0.67)$.

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, making the reuse of short words as subparts of longer words necessary to a much larger extent than for pseudo-Berber. In particular, there is a relatively small set of CV-sequences (65 for the inventory used here) that, by construction, have to occur in every possible word and might very well be words themselves, thus leading to inherent ambiguity. In contrast, pseudo-Berber can choose from arbitrary sequences of any length, leading to less ambiguity precisely because word-forms are unconstrained.

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 to 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". High-frequency single segment "words" lead to high segmentation ambiguity because they might be segmented out (incorrectly) whenever the segment occures. In contrast, pseudo-Senufo's requirements on minimal words to be CVs caps this ambiguity, as is particularly apparent for *STRUCT=4 and *STRUCT=5. Whereas pseudo-Senufo shows only an increase in SE from 0.08 to 0.09, pseudo-Berber's SE changes from 0.18 to 0.34, and this huge increase in ambiguity is directly reflected in the low TF of 59%, compared to 83% for pseudo-Senufo.

While this explanation pertains directly 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, of course abstracting away from other possible confounding factors that are impossible to control for when working with natural languages and assuming the kind of distributional model employed here.

The pattern for the FLAT condition doesn't lend itself to an explanation through SE which, for all values of *STRUCT, is virtually 0. We suspect the exceptionally high token f-score for *STRUCT¿2

to be explained by (i) a rough match between the model's expectations about word lengths and (ii) given this match, the balanced evidence for each individual word type. As for (i), specifying the segmentation model requires a setting an expected word-length for the lexical generator. Somewhat arbitrarily, we set this to 2, assuming that the probability of generating an additional segment for a word is equal to that of generating the end-of-word symbol. As for (ii), note that in a flat distribution there are no high- and no low-frequency items and that every word type occurs with (roughly) the same frequency. As long as the true words do not deviate too much from the expected length of 2, this is an ideal setting as there is an equal amount of evidence for every type, allowing a distributional learner to pick these types out with high reliability and, consequently, yield exceptionally high segmentation scores. In particular, there is no problem of high-frequency one or two segment long words, simply because there are no high-frequency words to begin with.

Once the expected word length deviates too much from that of the tokens in the input, however, the absence of high-frequency words turns into a severe disadvantage. The wrong expectation of an average word-length of 2 clearly poses no problem for either BRB nor ZIPF, and we suspect that this is mainly because high-frequency items are picked out reliably by statistical learners, allowing them to overcome this wrong expectation by using these high-frequency items as a cue to segment out further words.

In contrast, the FLAT distribution lacks these kinds of cues and once words tend to get "too long", their prevents the model to ever properly break into the gold segmentation. This also explains why FLAT looks like a mirror-image of BRB and ZIPF. While for high values of *STRUCT, the latter fall prey to the kind of ambiguity that high-frequency one or two segment words give rise to in data that exhibits a Zipfian distribution of types, they can accommodate the deviation from their prior expectation about word lengths due to the existence of several high-frequency items for low values of *STRUCT. Whereas FLAT would be an ideal distribution if the model's prior expectations are matched closely, it doesn't provide sufficient cues for a model to recover from wrong prior expectations.

On this note, it's also worth pointing out that across the three kinds of frequency distributions studied here, the "natural" Brent distribution is the most stable across varying values of STRUCT, in particular for pseudo-Senufo. This indicates that actual natural language frequency distributions do indeed facilitate tasks such as word segmentations and are, in a sense to be more precise in future work, more robust to variations in word lengths than artificial distributions, consistent with recent experimental work (Kurumada et al., 2013). Our observations about the disadvantage of a flat distribution when a model's prior expectation do not match also might explain why natural languages do not exhibit these kinds of distributions — learning would be close to impossible whenever a prior expectation is violated.

5 Conclusion

We used artificial languages to support the idea that and provide an explanation of why more rigid phonotactics negatively impact statistical word segmentation models. We also found some evidence that natural frequency distributions are more robust to variations in word lengths than artificial created ones. Our use of artificial languages instead of real languages allows us to at least partially control for many of the complex interaction natural language exhibits, thus directly addressing the role played by phonotactics. While ultimately, we want to understand how infants segment real languages, we believe to have shown how artificial languages can help in better understanding our computational models and addressing specific hypotheses, steps that need to be taken to design more adequate models in future research.

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.

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.