Information-Theoretic Segmentation of Natural Language

Sascha S. Griffiths, Mariano Mora McGinity, Jamie Forth, Matthew Purver, and Geraint A. Wiggins

Cognitive Science Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, 10 Godward Square, London E1 4FZ United Kingdom
sascha.griffiths@qmul.ac.uk,
WWW home page: http://compling.eecs.qmul.ac.uk/

Abstract. We present computational experiments on language segmentation using a general information-theoretic cognitive model. We present a method which uses the statistical regularities of language to segment a continuous stream of symbols into "meaningful units" at a range of levels. Given a string of symbols—in the present approach, textual representations of phonemes—we attempt to find the syllables such as *grea* and *sy* (in the word *greasy*); words such as *in*, *greasy*, *wash*, and *water*; and phrases such as *in greasy wash water*. The approach is entirely information-theoretic, and requires no knowledge of the units themselves; it is thus assumed to require only general cognitive abilities, and has previously been applied to music. We tested our approach on two spoken language corpora, and we discuss our results in the context of learning as a statistical processes.

Keywords: Artificial Intelligence; Language Acquisition; Learning; Language Segmentation; Information Content

1 Introduction

The question which we address in this paper is whether language learning can be considered to be a statistical process. This has been an ongoing and fundamentally dividing issue in fields which consider language learning their subject matter.

We assume that language has several layers of structure. At the bottom we find the smallest units; in this paper, we start with phonemes, though our method is not restricted to this level of granularity. These smallest units build larger items of language structure:

- Phonemes build larger units such as syllables and morphemes.
- Morphemes and syllables build words. Words build phrases and phrases are the building blocks of sentences (or spoken utterances).
- These sentences or utterances go in turn to make up larger units such as paragraphs in text or speaker turns in speech.

We must assume some smallest unit such as phonemes in speech or graphemes in text as an entry point into the language system. The question which arises is how one can tell where one such unit above the phoneme or grapheme ends and another one begins. In natural language processing this task is generally called text segmentation for written language and speech segmentation for spoken language.

In the current paper, we assume that the phonemes are presented as one continuous stream - roughly equivalent to removing the white space from sentences in written text - and define our task as determining where a word or other linguistic unit begins or ends. This is similar to the task infants face when learning their first language, itself an open research question. Taking the title as an example, we need to identify that the word *segmentation* is composed of syllables, which are *seg*, *men*, *ta*, and *tion*, and morphemes, which are *segment* and *ation*. From there larger units need to be distinguished such as the words *segmentation*, *of* and *natural*. Longer utterances need to be split up into phrases, perhaps at various levels of granularity e.g. *natural language* and *segmentation of natural language*.

We present a computational approach to the segmentation problem [1] in which we rely entirely on the *information content* of a symbol within a language dataset. Our prediction is that information content will rise at the beginning of a segment and fall at the end of a segment. A similar assumption was used by Harris [2] for finding morpheme boundaries. We assume that this assumption should hold for segments at all levels of the linguistic hierarchy. However, for each level the nature and extent of this fall and rise will vary; but parameters of the model will vary predictably with levels of segmentation. In the following, we will present computational experiments which test this prediction on two datasets of natural language.

Our approach to the cognitive task of language processing therefore places emphasis on domain-independent principles, rather than taking a domain-specific approach as has been argued as appropriate for the case of language.

We outline our information-theoretic approach in the next section; we then present the methods used in this paper in detail. Our results and the discussion of these for computational experiments on language segmentation are presented in the following sections. In our conclusion we return to the question outlined above.

2 Information-Theoretic Speech Segmentation

Applying machine learning and pattern recognition methods to natural language has become a rich source of insights into language structure and theoretical issues of linguistics [3] and the learning of language [4]. While many linguists take the view that natural language requires domain-specific, innate structures (see e.g.[5]), this is debated; one of the tenets of cognitive linguistics is that language processing by humans is domain general and not domain specific. As Geoffrey Sampson puts it, language learning depends on 'general human intelligence and

abilities' [6, p. IX]. Using information theory as a framework for such an approach has been argued to be both cognitively and biologically plausible [7, 8].

It has often been proposed that language and music share certain properties. One can ask the same questions regarding the structure and processing of language as one can ask about music [9]. Indeed, there a number of objective similarities and differences between these two domains [10]. There seem to be shared resources in structural processing of language and music [10], in addition to the conceptual similarity which is that the building blocks of the structures are "cognitive objects" – i.e. percepts . In this paper, we assume that percepts in general can be processed via their statistical regularities in a given corpus. The computational model used here was created for purposes of melodic grouping [11].

The current research is situated within the wider context of the IDyOM and IDyOT frameworks. IDyOM [12] stands for Information Dynamics Of Music. It was developed on the basis of natural language processing methods and can divide melodies into perceptually correct segments using the statistical regularities in a corpus. Generally, however, we argue here that the framework can also still be used to segment a corpus of natural language data into syllables, words and phrases.

In previous work [13, 14] a cognitive architecture called IDyOT is outlined which builds on the principles of IDyOM. IDyOT [14] stands for Information Dynamics Of Thinking. The premise of both of these different incarnations of the underlying research framework is that grouping and boundary perception are central to cognitive science [15] and that the most cognitively plausible way of approaching this task is using Shannon's [16] information theory. Especially, we employ information content as introduced by MacKay [17].

IDyOT is based on the Global Workspace Theory [18]. In the long term it is predicted that IDyOT provides the basis for modelling creativity and eventually aspects of consciousness [14, 19]. In order of testing certain claims about the domain generality of the information dynamics approach embodied by IDyOM and IDyOT, we look at language segmentation to see whether the approach shown to be useful in music segmentation can be transferred back to language.

The IDyOM model [20] and corresponding software[1] were developed for the statistical modelling of music in the context of music perception and cognition research. However, one of the central features of the model is that it can also be used for other types of sequential data, as the principles on which it is based are cognitively inspired and meant to be general rather than domain specific [13]. The model presented in IDyOM relies on a pattern recognition theory of mind [21] which suggests that languages are learned by processing the underlying statistics of the positive data contained in stimuli. Although, at the present moment only representations of auditory stimuli have been studied, our conjecture is that any kind of perceptual data can be processed in this way. Wiggins [22]

[1] The software can be found at https://code.soundsoftware.ac.uk/projects/idyom-project/files.

gives initial indications that the model can extend to language segmentation, and we pursue that idea in more depth here.

As aforementioned, we take an information-theoretic approach here [16]. Predicting the next element in a sequence given the previous element is often called a Shannon Game [23, p. 191]. Here, we assume that both music and language can be modelled as a sequence of elements e from an alphabet $\mathcal{E}$. For each element e_i in e one can calculate its probability given the context – more specifically the preceding context e_1^{i-1}:

$$p(e_i|e_1^{i-1}) \tag{1}$$

There is good evidence that children use transition probabilities during language acquisition [24, p. 33], and this probability can be calculated by approximating on the basis of a context subsequence of finite length n, i.e. by using an n-gram model [25, pp. 845–847].

2.1 The IDyOM Model

IDyOM is a multidimensional variable-order Markov model. The multidimensionality within IDyOM is formalised as a multiple viewpoints system [26], where viewpoints can be either given basic types, or derived and combined from existing viewpoints to form new viewpoints revealing more abstract levels of structure. Predictions from individual variable-order viewpoint models are combined using an entropy-weighting strategy [20].

Two basic information-theoretic measures are central to IDyOM. *Information content* is the measure of unexpectedness—or surprisal to use the terminology of [27]—and *entropy* a measure of uncertainty.

1. information content (h) is *a measure of how unpredictable a [given unit] is given its context* [27];
2. entropy (H) is *the expected information content of an unseen event in a given context.*

More formally, in IDyOM these concepts are modelled as (1) and (2) below:

$$h(e_i|e_1^{i-1}) = log_2 \frac{1}{p(e_i|e_1^{i-1})} \tag{2}$$

$$H(e_1^{i-1}) = \sum_{e \in \mathcal{E}} p(e_i|e_1^{i-1}) h(e_i|e_1^{i-1}) \tag{3}$$

Entropy-based models such as these have been used in natural language learning in the past [28, pp. 21–37]. Given an n-gram model of $p(e_i|e_1^{i-1})$ which characterises the dataset in question, we can calculate h and H at all points in a sequence, and thereby find local falls and rises. Such falls and rises have previously been shown to correlate with the ending and beginning of structural units in music [11, 12] and language [22]. For the case of music it has also been demonstrated that this model outperforms rule-based approaches [29].

2.2 Segmentation of Natural Language

Segmentation of natural language has been a topic for computational psycholinguistics at least since 1990 [30]. However, it can still be regarded as a current problem in computational approaches to language learning (see for example [31–34]). Brent [35] classifies a number of approaches to natural language segmentation into three types of strategies. These are the *utterance-boundary strategy*, the *predictability strategy* and the *recognition strategy*. Our approach employs elements of the predictability strategy: we attempt to detect boundaries based on changes in the information-theoretic properties of the symbol sequences in question. In this way it is similar to, but simpler and more general than, methods such as that of Cohen and Adams [36], who use *boundary entropy* but combine it with other frequency measures via voting experts to segment words in a range of languages, or Sun, Shen and Tsou [37], who use *mutual information* but combine it with other statistical measures to segment Chinese characters into words.

This contrasts with approaches in which one tries to build grammars (or probabilistic models) of likely segment sequences (the predictability strategy), (e.g. for Finnish morphemes [38]), and with those in which one matches patterns of known words against the stream (the recognition strategy); in those approaches one needs to build up a lexicon first, either from external knowledge (e.g. [39]) or from incremental clustering (e.g. [40]). Our boundary detection strategy needs no knowledge of the lexicon or even of the fact that there are such concepts as syllables, words or phrases.

3 Methodology

Our experimental method requires two steps: firstly, building a statistical n-gram (IDyOM) model on the basis of which to calculate information content (entropy is left for future work); secondly, hypothesising boundaries based on local drops and rises in information content.

3.1 Calculating the Information Content

IDyOM has a range of model configurations intended to simulate different aspects of musical listening behaviour. The basic distinction concerns the data used to train an n-gram model: a model can be trained from a large dataset, modelling the learned experience of a listener and termed the Long Term model (LTM) in IDyOM's terminology; or from only the current sequence under consideration, trained incrementally for each utterance being predicted [26, 20, 41], modelling a listening experience in a specific context, and termed the Short Term model (STM). However, variations are possible: the LTM approach can be made dynamic by adjusting its probabilities based on the current sequence as it is observed (termed LTM+); and the STM and LTM models can be combined. This results in a total of five models:

STM model trained on stimuli only in a local context (i.e. notes of the melody or phonemes in the utterance currently being predicted);

LTM model trained on a large training corpus;
LTM+ as LTM, but model also learns from the current example;
Both combination of STM and LTM;
Both+ combination of STM and LTM+.

3.2 Segmentation

Our overall approach is to look for characteristic local contours in information content [22]—what Pearce et al. [15] call 'peak picking'. Rises in information content are signals of unexpectedness, and Wiggins [14] hypothesises that these should correlate with the beginnings of new segments; conversely, falls in information content are signals of predictability, which we expect to correlate with the endings of segments.

Our current method is extremely simple, checking only for a simple rise between successive data-points: the value at the current symbol e_i must exceed that at its immediate predecessor e_{i-1} by some specified amount. This amount is our only parameter, d; thus, a new segment begins if $h(e_i) - h(e_{i-1}) > d$. We evaluate performance using the κ statistic [42, 43], and set d to give the maximal value for κ (for a specific segment type) by testing all d over the interval $[0, 10]$.

3.3 Data

We test this method on two language corpora. The first dataset is a derived corpus[2] of the CHILDES corpus of child-directed adult English speech [44], collated and transcribed at the phoneme level for word segmentation experiments [45]. It contains 93,555 phoneme tokens which make up 33,377 words and 9,790 utterances; average utterance length is 3.4 words. A single viewpoint with phonemes as observed variables, denoted {phonemes}, is used as the basic IDyOM representation.

The second corpus is the TIMIT transcriptions [46], a dataset of spoken English sentences obtained for the purpose of automatic speech recognition model training, and transcribed at the level of sentences, words and phonemes. It contains 81,533 phoneme tokens which make up 20,756 words and 2,342 utterances; average utterance length is therefore 8.9 words. Again we use a simple phoneme viewpoint; as TIMIT also contains stress annotations (represented as primary, secondary, and no stress), this also allows us to construct a linked viewpoint formed of the cross-product of phonemes and stress {phonemes $\otimes$ stress}, and a two-viewpoint system combining both viewpoints {phonemes, phonemes $\otimes$ stress}.

To evaluate phrase-level segmentation, we used the Pattern parser [47] – neither TIMIT nor CHILDES contains phrase structure information. Automatic parses are noisy: we excluded cases where Pattern produced a parse which could not be mapped back onto the phonetic form of the utterance. Thus, our analysis on the phrase level only considers approximately half of the data for both corpora.

[2] `http://www.ling.ohio-state.edu/~melsner/resources/acl12data.tgz`

4 Results

We evaluate our segmentation model in terms of accuracy of boundary placement against the ground truth for each level—syllables, words and phrases—with accuracy assessed via both Kappa values (κ) and the F1-score (harmonic mean of precision and recall). Both κ and F1 are calculated on the basis of individual phoneme tokens, with the gold-standard annotation classifying only the first token in each segment as a boundary. We also examine the mean information content ($\bar{h}$), and optimum value of our segmentation parameter (d). $\bar{h}$ is the same in across all, as it is a property of the (phoneme-based) corpus and model and not of the evaluation.

4.1 CHILDES

The results for the CHILDES corpus segmentation into words and phrases is summarised in Table 1. Lower $\bar{h}$ values mean better predictability, as high $\bar{h}$ signifies more "surprisal" by a new element.

Table 1. Results for the CHILDES corpus for words (left) and phrases (right) using all five IDyOM configurations.

CHILDES							
		WORDS			PHRASES		
Model	$\bar{h}$	{phonemes}			{phonemes}		
		d	κ	F1	d	κ	F1
STM	5.74	5.39	0.39	0.46	6.06	0.52	0.57
LTM	3.42	1.59	0.58	0.71	2.87	0.54	0.63
LTM+	3.42	1.57	0.58	0.71	2.87	0.54	0.63
Both	3.67	1.21	0.54	0.7	3.02	0.55	0.65
Both+	3.66	1.8	0.54	0.7	3.05	0.56	0.65

Performance is reasonable at word level, with F1 around 0.7 and κ approaching 0.6. The performance of the STM is considerably lower than other models, as might be expected; we note that $\bar{h}$ is considerably higher for the STM, indicating worse fit. The d parameter is therefore also correspondingly higher – and takes longer to find – for the STM. The lowest d for words is found in the Both model and LTM and LTM+ for the phrase segmentation.

In terms of both F1 and κ, the LTM and LTM+ are the best models for the word discovery task. In the phrase segmentation task, we find that the Both and Both+ models do marginally better than in the word segmentation task with respect to κ, but with worse performance with respect to F1. The apparent improvement of results for the STM may be due to the lower number of segments which need to be predicted. The improvement in performance by the short term model also leads to improvements in the Both and Both+ configuration as these are combinations of STM and LTM.

In comparison to previous work on the same dataset, our best configuration (LTM) still performs slightly worse with respect to F1-scores in the word segmentation task. While Elsner et al. [45] obtained an F1-score of 0.8, our best F1 score was 0.71. We also checked the baselines with respect to a random segmentation, a segmentation which assumes every symbol is a boundary and a segmentation which assumes no boundaries. In each case, the κ will be 0 as expected with low F1-scores.

4.2 TIMIT

Table 2 shows the results for the TIMIT corpus. The results for the syllable segmentation task are very comparable for all measures with those reported by Wiggins [22]. As with the CHILDES dataset, the STM shows higher values for $\bar{h}$, with the LTM and Both models showing better performance. κ and F1 are almost the same for the STM for all configurations. Therefore, the STM, for which the $\bar{h}$ is determined based on isolated utterances, seems not to be a good model for this task.

Table 2. Summary of results for the TIMIT corpus for words (left) and phrases (right) using all five configurations of IDyOM.

TIMIT											
		SYLLABLES			WORDS			Phrases			
Model	$\bar{h}$	{phonemes} d	κ	F1	{phonemes} d	κ	F1	{phonemes} d	κ	F1	
STM	5.46	2.43	0.11	0.26	3.95	0.17	0.24	6.96	0.39	0.42	
LTM	3.55	1.29	0.47	0.65	1.96	0.58	0.69	4.50	0.41	0.47	
LTM+	3.54	1.15	0.47	0.66	1.95	0.56	0.69	4.40	0.41	0.47	
Both	3.68	1.26	0.45	0.64	1.65	0.55	0.67	4.44	0.42	0.48	
Both+	3.67	1.05	0.45	0.65	1.94	0.56	0.69	4.52	0.42	0.48	
Model	$\bar{h}$	{phonemes $\otimes$ stress} d	κ	F1	{phonemes $\otimes$ stress} d	κ	F1	{phonemes $\otimes$ stress} d	κ	F1	
STM	6.04	3.09	0.11	0.22	3.72	0.18	0.26	7.36	0.39	0.42	
LTM	3.73	1.08	0.48	0.67	2.17	0.60	0.70	4.48	0.42	0.48	
LTM+	3.72	1.10	0.48	0.67	2.05	0.60	0.71	4.84	0.42	0.48	
Both	3.85	1.2	0.46	0.66	2.16	0.58	0.68	4.11	0.42	0.49	
Both+	3.84	1.27	0.47	0.65	2.15	0.58	0.69	4.09	0.42	0.49	
Model	$\bar{h}$	{phonemes, phonemes $\otimes$ stress} d	κ	F1	{phonemes, phonemes $\otimes$ stress} d	κ	F1	{phonemes, phonemes $\otimes$ stress} d	κ	F1	
STM	6.01	2.96	0.11	0.23	3.64	0.18	0.26	7.03	0.39	0.42	
LTM	3.72	1.10	0.49	0.67	2.12	0.61	0.71	3.93	0.42	0.49	
LTM+	3.71	1.14	0.49	0.67	2.13	0.61	0.71	4.49	0.42	0.48	
Both	3.85	1.07	0.47	0.66	2.01	0.58	0.69	4.12	0.43	0.49	
Both+	3.84	1.10	0.47	0.65	1.92	0.58	0.69	4.02	0.42	0.49	

Generally, one can see a trend that the LTM, LTM+, Both and Both+ models perform better if they receive more information, i.e. in the {phonemes $\otimes$ stress} and {phonemes, phonemes $\otimes$ stress} show marginally better performance. The smallest values for d are found in the viewpoints {phonemes$\otimes$ stress}.

There seems no improvement in performance when moving from the LTM to LTM+ or the Both model variants (in all configurations except the {phonemes} condition, where the LTM+ shows slightly higher F1-score). Overall, the best configuration for the word segmentation task is the LTM in the {phonemes, phonemes $\otimes$ stress} condition.

Segmentation at Different Levels For the word segmentation task, the optimal setting of d is higher than that for the syllable segmentation task, for all configurations. This corresponds with intuitive expectation, as one needs to predict fewer segments. In all configurations, the LTM and LTM+ still show better performance than the STM, Both and Both+; overall accuracy is slightly improved over the syllable segmentation task, with F1 scores over 0.7. The LTM+ is the best configuration overall in the {phonemes, phonemes$\otimes$ stress} condition. The STM perhaps also shows some improvement here, with κ values marginally higher.

In the phrase segmentation task, again, the optimum d increases relative to word and syllable tasks, as even fewer segments need to be predicted. Performance in terms of κ and F1-scores is, however, much lower for phrase discovery than for syllables and words. Thus, with regard to this measure the performance on the TIMIT data is less effective.

The κ values and F1 scores for the STM, however, are considerably higher for this task. The STM does, however, not benefit from the additional information which it gets in the {phonemes $\otimes$ stress} and {phonemes, phonemes $\otimes$ stress} conditions.

In all configurations, the LTM, LTM+, Both and Both+ models show worse performance in the phrase segmentation task with respect to κ and F1. Also, there is little difference in the performance of these four models. The Both is the best configuration overall in the {phonemes, phonemes $\otimes$ stress} condition with respect to the κ value.

As expected our results are very similar to those reported in Wiggins [22] for syllable segmentation. We also checked the baselines with respect to a random segmentation, a segmentation which assumes every symbol is a boundary and a segmentation which assumes no boundaries. In each case, the κ will be 0 as expected with low F1-scores.

4.3 Overall Segmentation Performance

Figures 1 and 2 show the variation of κ with the information content threshold parameter d for each corpus, illustrating the process of determining optimum d values.

The LTM and Both model variants show a general pattern for syllables and words: a gradual improvement leading up to a peak in performance (defining

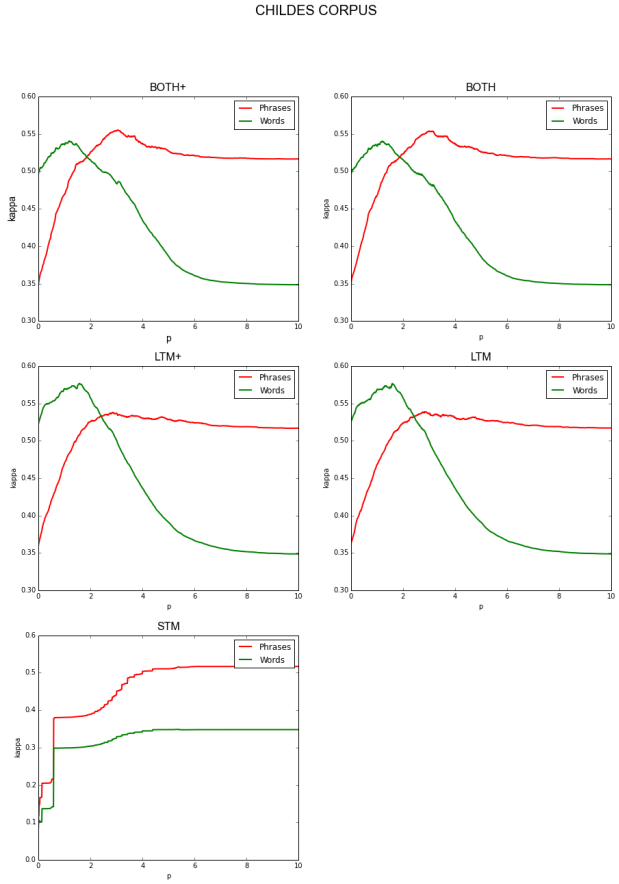

Fig. 1. CHILDES corpus κ vs parameter d for words and phrases.

optimum d), after which performance drops off again. This optimum value of d is higher as we move to longer, higher-level segments (from syllables to words, and from words to phrases): larger changes in information content correspond to segment boundaries at different levels. However, the phrase segmentation curve shows less of a peak: performance reaches a level at which it stays. This suggests that as long as d is large enough one finds segments which have a high probability of coinciding with phrase boundaries. The plateau in the curve after the peak can be explained as an effect of our segmentation method coding the beginning of an utterance as a given start symbol. This is similar to the approach of Elsner et al [45]. Thus, once the method stops oversegmenting at low d it finds the optimum d and afterwards continues to agree on those given symbols at higher d values.

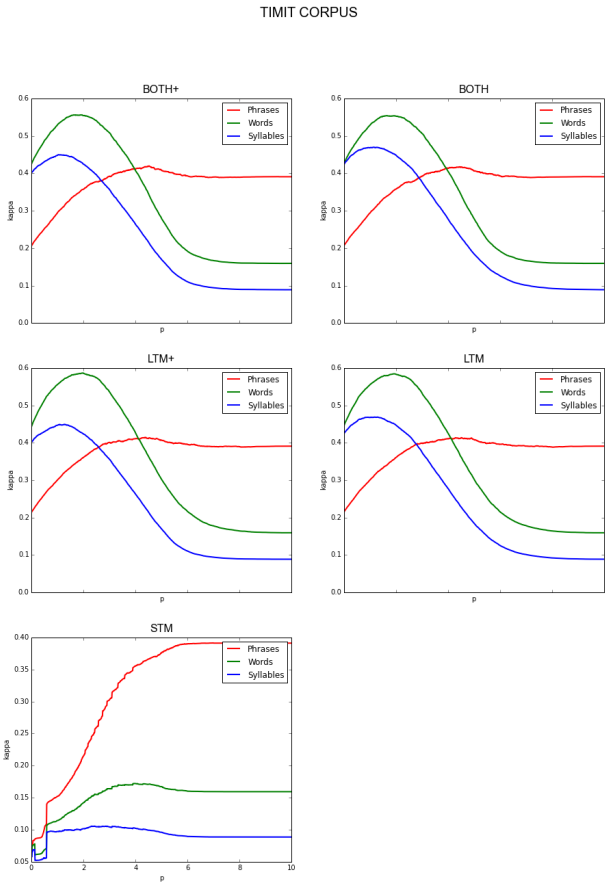

Fig. 2. TIMIT corpus κ vs parameter d for syllables, words and phrases.

The STM shows particularly bad performance initially but then the plots show a sudden leap in performance. This is true for all configurations on both corpora. Thus, short term segmentation seems to require a certain threshold to show any noticeable segmentation performance. Exposure to isolated utterances is insufficient to learn the distributional regularities of language.

5 Discussion & Conclusion

Landis and Koch [48] characterise a $\kappa \in [0.4, 0.6]$ as "moderate". Thus, most of the results reported here show a moderate success. The results reported for the CHILDES corpus with respect to phrases is slightly higher and thus falls into the "substantial" category. However, one has to again note, that results regarding the syntactic units are to be taken with caution. There is less to predict and less to

agree on. Therefore, one would expect a higher agreement between ground-truth and segmentation.

The long term model shows better performance than the short term model. In effect, these two model a listeners knowledge of language (LTM) and a current listening experience (STM). It is to be expected that there is little result to be expected from learning from a single listening experience. Thus, the results with respect to the differences in LTM and STM show that a long term learning from raw stimulus is possible.

The TIMIT data also indicates that learning is improved if stress information can be included. Though, the differences are small, the inclusion of stress in the viewpoints selected for predicting the next phoneme do improve the results. The differences reported here are minor, though.

The present contribution took a strong view of statistical language learning. We claimed that it would be possible to predict syllable, word and phrase boundaries from a raw stimulus without having explicit information about these units encoded in the method. We succeeded in the sense that our results indicate that this is indeed possible. In future work, we plan to explore further in what way the inclusion of different viewpoints improves the results.

6 Acknowledgments

The research reported in this is supported by ConCreTe: the project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

References

1. Elman, J.L.: Computational approaches to language acquisition. In Brown, K., ed.: Encyclopedia of Language and Linguistics. Volume 2. Second edn. Elsevier, Oxford (2006)
2. Harris, Z.S.: From phoneme to morpheme. Language **31**(2) (1955) pp. 190–222
3. Lappin, S., Shieber, S.M.: Machine learning theory and practice as a source of insight into universal grammar. Journal of Linguistics **43**(2) (2007) 393–427
4. Clark, A., Lappin, S.: Linguistic Nativism and the Poverty of the Stimulus. Wiley-Blackwell, Oxford (2011)
5. Chomsky, N.: New Horizons in the Study of Language and Mind. Cambridge University Press, Cambridge (2000)
6. Sampson, G.R.: The Language Instinct Debate. Continuum, London (2005)
7. Wallace, R.: Cognition and biology: perspectives from information theory. Cognitive Processing **15**(1) (February 2014) 1–12
8. Rohrmeier, M., Zuidema, W., Wiggins, G.A., Scharff, C.: Principles of structure building in music, language and animal song. Philosophical Transactions of the Royal Society of London B: Biological Sciences **370**(1664) (2015) 20140097
9. Jackendoff, R., Lerdahl, F.: The capacity for music: What is it, and what's special about it? Cognition **100**(1) (2006) 33–72

10. Patel, A.D.: Music, Language, and the Brain. Oxford University Press, Oxford (2008)
11. Pearce, M.T., Wiggins, G.A.: The information dynamics of melodic boundary detection. In: Proceedings of the Ninth International Conference on Music Perception and Cognition, Bologna (2006) 860–867
12. Pearce, M.T., Müllensiefen, D., Wiggins, G.A.: Melodic grouping in music information retrieval: New methods and applications. In: Advances in music information retrieval. Springer, Berlin (2010) 364–388
13. Wiggins, G.A.: The mind's chorus: creativity before consciousness. Cognitive Computation **4**(3) (2012) 306–319
14. Wiggins, G.A., Forth, J.: IDyOT: A computational theory of creativity as everyday reasoning from learned information. In Besold, T.R., Schorlemmer, M., Smaill, A., eds.: Computational Creativity Research: Towards Creative Machines. Volume 7 of Atlantis Thinking Machines. Atlantis Press (2015) 127–148
15. Pearce, M.T., Müllensiefen, D., Wiggins, G.A.: The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. Perception **39**(10) (2010) 1365–1389
16. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review **5**(1) (1948) 3–55
17. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)
18. Baars, B.J.: A cognitive theory of consciousness. Cambridge University Press, Cambridge (1993)
19. Wiggins, G.A., Tyack, P., Scharff, C., Rohrmeier, M.: The evolutionary roots of creativity: Mechanisms and motivations. Philosophical Transactions of the Royal Society of London B: Biological Sciences **370**(1664) (2015) 20140099
20. Pearce, M.T.: The construction and evaluation of statistical models of melodic structure in music perception and composition. PhD thesis, City University London (2005)
21. Kurzweil, R.: How to create a mind: The secret of human thought revealed. Penguin, London (2012)
22. Wiggins, G.A.: "I let the music speak": Cross-domain application of a cognitive model of musical learning. In Rebuschat, P., Williams, J., eds.: Statistical Learning and Language Acquisition. Mouton de Gruyter, Amsterdam, NL (2012) 463 – 494
23. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA (1999)
24. Ambridge, B., Lieven, E.V.M.: Child Language Acquisition: Contrasting Theoretical Approaches. Cambridge University Press, Cambridge (2011)
25. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Third edn. Prentice Hall International (2013)
26. Conklin, D., Witten, I.H.: Multiple viewpoint systems for music prediction. Journal of New Music Research **24**(1) (March 1995) 51–73
27. Mahowald, K., Fedorenko, E., Piantadosi, S.T., Gibson, E.: Info/information theory: Speakers choose shorter words in predictive contexts. Cognition **126**(2) (2013) 313–318
28. Charniak, E.: Statistical Language Learning. MIT Press, Cambridge, MA (1996)
29. Pearce, M.T., Wiggins, G.A.: Expectation in Melody: The Influence of Context and Learning. Music Perception **23**(5) (2006) 377–405
30. Elman, J.L.: Finding structure in time. Cognitive Science **14**(2) (June 1990) 179–211

31. Elsner, M., Goldwater, S., Feldman, N., Wood, F.: A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2013)
32. Fourtassi, A., Börschinger, B., Johnson, M., Dupoux, E.: Why is English so easy to segment? In: Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL), Sofia, Bulgaria, Association for Computational Linguistics (August 2013) 1–10
33. Pate, J.K., Johnson, M.: Syllable weight encodes mostly the same information for english word segmentation as dictionary stress. In: EMNLP, Doha, Qatar (2010) 844–853
34. Çöltekin, Ç.: Units in segmentation: A computational investigation. In: Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, Lisbon, Portugal, Association for Computational Linguistics (September 2015) 55–64
35. Brent, M.R.: Speech segmentation and word discovery: a computational perspective. Trends in Cognitive Sciences **3**(8) (August 1999) 294–301
36. Cohen, P., Adams, N.: An algorithm for segmenting categorical time series into meaningful episodes. In: Advances in Intelligent Data Analysis, 4th International Conference, Cascais, Portugal (2001) 198–207
37. Sun, M., Shen, D., Tsou, B.K.: Chinese word segmentation without using lexicon and hand-crafted training data. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. (1998) 1265–1271
38. Virpioja, S., Turunen, V.T., Spiegler, S., Kohonen, O., Kurimo, M.: Empirical comparison of evaluation methods for unsupervised learning of morphology. Traitement Automatique des Langues **52**(2) (2011) 45–90
39. Sproat, R., Shih, C., Gale, W., Chang, N.: A stochastic finite-state word-segmentation algorithm for Chinese. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. (1994) 66–73
40. Gold, K., Scassellati, B.: Audio speech segmentation without language-specific knowledge. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society, Vancouver (2006) 1370–1375
41. Wiggins, G.A., Pearce, M.T., Müllensiefen, D.: Computational modelling of music cognition and musical creativity. In Dean, R., ed.: The Oxford Handbook of Computer Music. Oxford University Press (2009) 383–420
42. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1) (April 1960) 37–46
43. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics **22**(2) (1996) 249–254
44. MacWhinney, B.: CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Lawrence Erlbaum Associates, Mahwah, NJ (2000)
45. Elsner, M., Goldwater, S., Eisenstein, J.: Bootstrapping a unified model of lexical and phonetic acquisition. In: Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics. (2012)
46. Zue, V., Seneff, S., Glass, J.: Speech database development at MIT: TIMIT and beyond. Speech Communication **9** (1990) 351–356
47. De Smedt, T., Daelemans, W.: Pattern for Python. Journal of Machine Learning Research **13**(1) (2012) 2063–2067
48. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1) (1977) 159–174