From Phoneme to Morpheme: A Computational Model

Sascha Griffiths and Matthew Purver and Geraint Wiggins

Abstract—Zellig Harris proposed a method for grouping phonemes in an utterance into morphemes by simply using counts of each of the phonemes in a corpus relative to their position in sequences contained in the data set. Thus, using an n-gram model, one can model this process and see whether a computational model can actually group representations of phonemes into segments which correspond to morphemes. Here, we use a general n-gram modelling tool created for melodic grouping in music corpora and apply it to a natural language data set. We show that this method which approximates Harris's can indeed find morphemes in a given language corpus by calculating the distributions of phonemes across a corpus.

I. INTRODUCTION

The underlying principles contained in our current approach were first introduced by Harris [1]. Harris described a procedure by which phoneme sequences could be grouped into morphemes. He envisioned a use-case for this method in which one knows about a given alphabet (in the formal sense of the word) but has not worked out what the meaningful segments are. He specifically hypothesised that the distributional properties could be used to determine whether an item in a sequence of phonemes constitutes a morpheme without reference to meaning.

This is the task that the IDyOM framework was developed for. IDyOM [2] stands for Information Dynamics Of Music. However, it was developed for the purpose of finding boundaries in sequences of musical notes. The purpose of the analysis presented here is to see what results one can expect when IDyOM is used on a natural language corpus. In previous work it was also hypothesised that IDyOM performs well at determining morpheme boundaries [3]. In a test with respect to other linguistic units, it performed reasonably well for syllable segmentation and word boundary detection and to a lesser extent regarding phrase boundary detection [4]. Golcher [5] similarly tried to segment text into morphemes, words and multi-word expressions with a related but different approach. Although both methods use the predictability strategy for segmentation [6], the latter approach used text as the input whereas in the current contribution our model is trained on representations of phonemes. Harris [7] stressed that the method he described was intended for analysing sequences of phonemes.

In this contribution we examine the role of morphemes in segmenting a natural language data set comprised of sequences of symbols representing phonemes. Phonemes will group into morpheme segment candidates without reference to “meaning” simply by considering their distributional properties in sequences across the data set.

II. MODELLING GROUPING AND BOUNDARY PERCEPTION USING INFORMATION DYNAMICS

IDyOM calculates the regularities of a corpus using a multidimensional variable-order Markov model. Thus, it is based on n-gram modelling [8, pp. 845–847]. Harris [1] referred to this as predecessor counting. He used a simple counting method to determine rises in frequency for each element both forward counting (successor count) and backward counting (predecessor count). He then determined for every utterance how often a given phoneme would appear in a certain context. His assumption was that a given distribution would show periodicity determined by boundaries which group phonemes into morphemes.

In contrast to raw counts of frequencies of elements in a sequence taking a given position, we propose using information content as a measure of frequency. More precisely, we call this a measure of unexpectedness (sometimes also called 'surprisal', e.g. in [9]). Following MacKay [10], we formalise information content as:

\[ h(e_i|e_1^{i-1}) = \log_2 \frac{1}{p(e_i|e_1^{i-1})}. \] (1)

With elements \( e \) from an alphabet \( E \) being the phonemes in a sequence. For each element \( e_i \) in \( e \) one can calculate its probability given the context – more specifically the preceding context \( e_1^{i-1} \) – which can be defined as:

\[ p(e_i|e_1^{i-1}) \] (2)
as used in (1).

We use information content as the measure of the predictability of boundaries. It has been shown that for music this measure is particularly useful [11] in computational models of boundary detection. However, apart from its usefulness in computational models, it has also been demonstrated to be a useful predictor of segments in experimental research [12].

Our segmentation method assumes that local peaks will indicate a boundary. However, not every rise will be associated with a segment boundary. We assume that there is a parameter \( d \) such that:

\[ h(e_i|e_1^{i-1}) < d \] (3)

will be identified as boundary. A different setting of \( d \) will result in different segmentations. In order of finding a good \( d \) one needs to compare different segmentation results to a ground-truth (such as annotated syllable boundaries [3].
for language or expert judgements for music [13]). Our method for determining an appropriate value for $d$ is further explained below (see Section III).

The model is not as such a direct implementation of the model presented by Harris [1], [7] but similar to the model of Golcher [5] and sources cited therein ([14], [15], [16]), it is inspired by the work of Harris in the sense that it is purely statistical, uses successor or predecessor counts of the elements in strings of language and predicts the next element in the sequence based on these counts. Brent [6] calls this the predictability strategy of text segmentation which contrasts with utterance boundary detection methods (e.g. [17], [18]) and recognition based approaches (e.g. [19], [20], [21]).

III. METHODS

We now discuss what kind of units the segmentation predicts in the corpus with different settings of the parameter $d$ and how these develop as $d$ changes. The corpus we use in this evaluation is the TIMIT corpus [22] which was created for training speech recognition systems. Our processed version of this dataset contains 81,533 phoneme tokens (40 types) which make up 20,756 words and 2,342 utterances; average utterance length is therefore 8.9 words.

The data were presented to the IDyOM system in a total of 5 conditions, as itemised in Tab. I. IDyOM can be used with a Long Term model (LTM), which is exposed to an entire corpus (modelling the learned experience of a listener) and a Short Term model (STM), which is exposed only to the current melody or utterance (modelling a specific listening experience). Also, there is a version of the LTM which is called LTM+ in which the LTM learns from its current stimulus presented to the system. Additionally, both LTM and LTM+ can be combined with the STM to give two further models – Both and Both+. The LTM, LTM+, Both and Both+ models are trained using ten fold cross-validation.

In each condition, the resulting model was used to predict the information content of each phoneme in the corpus, in context of its utterance prefix. The resulting signal was differentiated (see equation (3) above), and values larger than a parameter $d$ were taken as boundaries. $d$ was varied with $d \in [0 : 10]$ at 0.01 increments which yields a 1,000 different possible segmentations. In order of obtaining “good” possible segmentations, we compared all possible segmentations against a ground truth for syllables, words and phrase-chunks. We use these three ground truths as a reference in lieu of a ground truth for morphemes as the TIMIT corpus does not have annotations for morpheme boundaries.

This procedure is an automation of the search for rises in the counts of phonemes at a given position [1]. Harris did not have a threshold value above which a new segment was to be identified. However, he was aware that within a segment the counts do not fall linearly but fall and rise with high rises defining a new segment.

IV. RESULTS

The performance of each of the configurations is shown in Tab. I. The different models actually give different results which is to be expected for language.

The STM’s performance is worse than that of all other configurations. Also, the Both and Both+ performance is worse than the LTM and LTM+ configurations which can be explained by the fact that the STM contributes in a detrimental way to the performance of the latter configurations. In music segmentation the STM actually performs well [23]. An interpretation of this difference is that music is self-referential and much of its “meaning” is therefore emergent from repetition and variation in its local structure (see also discussion in [24], [25]), whereas in language (other than rhyming poetry) the semantics of segments contributes more to their interpretation [3]. More details regarding the performance can be found in Tab. I.

Harris [1] also assumed, that for his method to work, the counts would have to be based on a sufficiently large corpus and could not be derived from an utterance in isolation. Therefore, we now look at the results one can obtain by applying this method considering the best performing model which is the LTM which itself is also the closest approximation to Harris’s method. Tab. II shows the 10 most frequent items produced by the segmentation method for the cases in which $d$ was optimised according to a ground truth relative to (1) syllables, (2) words and (3) phrase-chunks.

One can see that there are segments which seem relatively stable across different kinds values of $d$. The most frequent words the, and you are also frequent in all three segmentations. Our observations are that the five most frequent words in the word segmentation task are also reliably found as individual words in the syllable and phrase segmentation. For example, the is 1st in both the syllable segmentation and the word segmentation but also 3rd in the phrase-chunks segmentation. and on the other hand is not among the ten most frequent items for the phrase-chunk segmentation; however, it is among is the 20th most common item there. As you, and he all appear in the most frequent items in all three lists, in becomes the only exception as it does appear in the most frequent terms of the phrase-chunk list (and doesn’t appear within the most frequent 100 items at all). This may be due to the fact that instances of in have to a large amount been absorbed into larger chunks (see below).

A. Morphemes

A consistent pattern is found with respect to morphemes appearing in the list. Both $[s]$ and $[z]$ which are inflectional morphemes which may indicate plurals of nouns, possessive forms of nouns and 3rd person singular forms of verbs are found among the most frequent ten candidate segments within the lists for syllable, word and phrase-chunk segments. Similarly, $[t]$ and $[d]$ which indicate the past tense forms of verbs are consistently among the most twenty most frequent items. Among, the 20 most frequent items in the best segmentation for syllables, one also finds the items syllable -ing and -ly which are derivational morphemes. In Fig. 1 one can see the appearance of these 6 morpheme candidates plotted. With their frequency plotted on the y-axis and the corresponding value for $d$ on the x-axis. 
TABLE I
SUMMARY OF RESULTS FOR THE TIMIT CORPUS FOR WORDS (LEFT) AND PHRASES (RIGHT) USING ALL FIVE CONFIGURATIONS OF IDyOM. MORE DETAILS CAN BE FOUND IN GRIFFITHS ET AL. [4].

<table>
<thead>
<tr>
<th>Model</th>
<th>(1) SYLLABLES (phonemes)</th>
<th>(2) WORDS (phonemes)</th>
<th>(3) PHRASE-CHUNKS (phonemes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{h} ) ( \kappa ) F1</td>
<td>( \hat{h} ) ( \kappa ) F1</td>
<td>( \hat{h} ) ( \kappa ) F1</td>
</tr>
<tr>
<td>STM</td>
<td>5.46 2.43 0.11 0.26</td>
<td>3.95 0.17 0.24</td>
<td>6.96 0.39 0.42</td>
</tr>
<tr>
<td>LTM</td>
<td>3.55 1.29 0.47 0.65</td>
<td>1.96 0.58 0.69</td>
<td>4.50 0.41 0.47</td>
</tr>
<tr>
<td>LTM+</td>
<td>3.54 1.15 0.47 0.66</td>
<td>1.95 0.56 0.69</td>
<td>4.40 0.41 0.47</td>
</tr>
<tr>
<td>Both</td>
<td>3.68 1.26 0.45 0.64</td>
<td>1.65 0.55 0.67</td>
<td>4.44 0.42 0.48</td>
</tr>
<tr>
<td>Both+</td>
<td>3.67 1.05 0.45 0.65</td>
<td>1.94 0.56 0.69</td>
<td>4.52 0.42 0.48</td>
</tr>
</tbody>
</table>

TABLE II
THE TOP TEN SEGMENTS FOR THE BEST SEGMENTATION WITH RESPECT TO THE GROUND TRUTH FOR (1) SYLLABLES, (2) WORDS AND (3) PHRASE-CHUNKS. THE SEGMENTS ARE SORTED BY FREQUENCY WITH THAT SEGMENTATION AND A POSSIBLE INTERPRETATION IS GIVEN IN BRACKETS.

<table>
<thead>
<tr>
<th>(1) Syllables</th>
<th>(2) Words</th>
<th>(3) Phrase-chunks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[dh ax] (the)</td>
<td>[dh ax] (the)</td>
</tr>
<tr>
<td>2</td>
<td>[s] (noun and verb inflection realized as [s])</td>
<td>[ae n d] (and)</td>
</tr>
<tr>
<td>3</td>
<td>[z] (noun and verb inflection realized as [z])</td>
<td>[y uw] (you)</td>
</tr>
<tr>
<td>4</td>
<td>[h ih y] (the)</td>
<td>[h ih y] (the)</td>
</tr>
<tr>
<td>5</td>
<td>[t] (verb inflection realised as [t])</td>
<td>[ih n] (in)</td>
</tr>
<tr>
<td>6</td>
<td>[ih n] (in)</td>
<td>[z] (noun and verb inflection realized as [z])</td>
</tr>
<tr>
<td>7</td>
<td>[ae n d] (and)</td>
<td>[ah v] (*not in ground truth)</td>
</tr>
<tr>
<td>8</td>
<td>[y uw] (you)</td>
<td>[s] (noun and verb inflection realized as [s])</td>
</tr>
<tr>
<td>9</td>
<td>[d] (verb inflection realised as [d])</td>
<td>[h ih z] (his)</td>
</tr>
<tr>
<td>10</td>
<td>[ah v] (*not in ground truth)</td>
<td>[l iy] (adverbialiser '-ly')</td>
</tr>
</tbody>
</table>

Fig. 1. The frequency of common morpheme segments (noun and verb inflection markers [s] & [z], past tense markers [t] & [d], adverbialiser [ly] and gerund-marker [ing]) are plotted against the parameter \( d \). Thus, on the x-axis one finds \( d \in [0 : 10] \) at 0.01 increments. On the y-axis one finds the frequency of the examined items.

Fig. 2. The frequency of common word segments (the, and, you, he, in) are plotted against the parameter \( d \). Thus, on the x-axis one finds \( d \in [0 : 10] \) at 0.01 increments. On the y-axis one finds the frequency of the examined items.

B. Words
As can be seen in Fig. 2 the shape of the graphs for words displays a much rounder behaviour than the almost linear decent that the derivational morphemes show. The increased roundness as compared to the inflectional morphemes also suggests that the patterns in lexical units becomes clearer with higher values for \( d \).

As the segments become larger with increases in the value of \( d \), sequences of phonemes will be re-analysed and a transition from smaller syllable-like and morpheme-like units to "word-like" units occurs. An example of this would be a re-classification of an and following \( d \) to form the word and.

It is noticeable here that the shape of the curves indicates that there is a difference between the inflectional morphemes and the derivational morphemes. The derivational morphemes -ing (indicating a gerund) and -ly (an adverbialiser) seem to show a more steady behaviour than the inflectional morphemes for [s]/[z] which stand for plurals and possessives on nouns and 3rd person singular on verbs. This is indicated by the sharper drop in the graph of the derivational morphemes compared to the rounder shape of the graph for the inflectional morphemes.
This is most likely the explanation for the graph starting low and then a growing increase in frequency can be noted before it drops again.

C. Phrases

We examined the most frequent multi-word units which Harris [1] also proposed could be identified using his method under certain conditions. It is noticeable that measured against a ground truth which sees phrase-chunks as units such as adjective phrases, noun phrases, verb phrases, etc. the selected units, he had, this is, it was, has been, through the will be incorrectly classified. However, it is interesting that at higher \( d \) such units do appear. As can be seen in Fig. 3, the persistence of these segments is short lived. Despite being scattered they are more frequent in the regions of higher values of \( d \). Though, none of them is very frequent overall with it was being the most frequent and being found 24 times in one early segmentation.

D. Comparison

Three things are observable from the shape of the graphs in Fig. 4. The behaviour of morpheme-like segment candidates, word-like segment candidates and multi-word segment candidates is quite distinct judging from these graphs. First, the derivational morphemes show a more constant drop than all other units. They will be absorbed into large word-like segment candidates to a large extent before the segmentation reaches its best result with respect to the word ground truth. However, the inflectional morphemes are more persistent. Second, word-like segment candidates show higher frequencies than all other units from a certain value of \( d \) onwards. Third, the frequency of multi-word segment candidates is dwarfed by the frequency of word-like and morpheme-like segment candidates. The graphs are barely visible in comparison and they appear very late overall, at high \( d \) values. This is to a large extent after a majority of derivational morphemes has been absorbed into word-like units.

V. DISCUSSION

In the following section, we will discuss the results with respect to a few examples drawn from the segmentation results. As discussed elsewhere [4] the phrase-chunk segmentation did not perform as well as expected and thus these will only briefly be discussed in the text and not in Fig. 5 which shows a few sentences drawn from the TIMIT corpus.

In this contribution, we specifically wanted to address the question of whether there is a tendency to favour morpheme segmentation over possible other linguistic units even when parameter \( d \) is chosen for a specific type of unit such as syllables, words and phrase-chunks. In the previous sections quantitative measures were explored and the results are promising (see below in section VI). Additionally, we discussed the lexicon with respect to the frequency at which inflectional and derivational morphemes appear. In Fig. 5 three examples of sentences segmented with respect to both the syllable and word ground truth are shown in order to be able to discuss these results further.

In example (1), the false positives and false negatives for the first two words are particularly interesting and align with our argument. In the syllable segmentation task, the method sees the word “only” as one unit although the syllable segmentation would be two units, “on” and “ly” whereas the method splits “incomplete” into “in” and “plete” which would be “in” “com” and “plete” in a true syllable segmentation. Correspondingly, in the word discovery task there should two lexical items: “only” and “incomplete”. However, again, one finds the segmentation into “only” and “in” and “complete”. Also, independent of the task (i.e. the ground truth chosen for which \( d \) is chosen), “things” gets split into “thing” and the plural morpheme [z].

In example (2), one can see that “catastrophic” (containing the syllables “cata”, “stro” and “phic”) is split up into “catastrophe” and [k] in the syllable task. This is despite
the fact that the word “catastrophe” does not appear in the
corpus on its own. However, for a larger parameter $d$ in
the word discovery task this segmentation can no longer be
found. Overall, example (2) shows a poorer performance than
the other two examples. Although, the final two segments
“the” and “poor”, which are both mono-syllabic and mono-

inflectional markers [s] & [z] and past tense markers [t] & [d]
will remain frequent even when the segment candidates become
large enough to allow the segmentation to include multi-

word segments. Hence, even at high values of parameter $d$
inflectional morphemes are still regarded “unexpected”

enough to be segmented on their own.

Similar to the proposal of Harris [1] we can show that
morpheme boundaries can be detected in a continuous stream
of phonemes without reference to meaning just by using the
distributional properties of the events in sequences in a given
corpus. Further, it is not even required as Harris postulated
that one knows about the existence of morphemes to find
these using a distributional method.

VII. 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 Programme for Research
of the European Commission, under FET grant number
611733. The authors would like to thank the two anonymous
reviewers for their constructive comments.

REFERENCES

[1] Z. S. Harris, “From phoneme to morpheme,” Language,
http://www.jstor.org/stable/411036

in music information retrieval: New methods and applications,” in
Advances in music information retrieval. Berlin: Springer, 2010,
pp. 364–388.

cognitive model of musical learning,” in Statistical Learning and Lan-
guage Acquisition, P. Rebuschat and J. Williams, Eds. Amsterdam,

Wiggins, “Information-theoretic segmentation of natural language,” in
International Workshop on Artificial Intelligence and Cognition (MIC
2015), A. Lieto, C. Battaglini, and M. Sanguinetti, Eds. Turin, Italy:
AI*IA, 2015.


tional perspective,” Trends in Cognitive Sciences, vol. 3, no. 8,

[7] Z. Harris, “Morpheme boundaries within words: Report on a computer
test,” in Papers in Structural and Transformational Linguistics, ser.


591, 2013.

[10] D. J. C. MacKay, Information Theory, Inference, and Learning Algo-

expectation and probabilistic learning in auditory boundary perception:
A model comparison,” Perception, vol. 39, no. 10, pp. 1365–1389,
2010.


abilistic models of expectation violation predict psychophysiological
emotional responses to live concert music,” Cognitive, Affective, &
Fig. 5. An example segmentation produced by the best-performing system, the LTM. | denotes correctly predicted boundaries, : denotes false positives and . denotes false negatives in correspondence to the TIMIT annotations. The phoneme representation corresponds to the TIMIT format. The upper segmentation is with the best \( d \) for syllables (in red) and the lower segmentation is with the best \( d \) for words (in blue).


