The informativeness of linguistic unit boundaries

Jeroen Geertzen[1], James P. Blevins[1] & Petar Milin[2]

[1] *Language Technology Lab, University of Cambridge* <jg532@cam.ac.uk>, <jpb39@cam.ac.uk>
[2] *Department of Linguistics, Eberhard Karls Universität Tübingen* <petar.milin@uni-tuebingen.de>

Contemporary models of structural analysis tend to operate with discrete units at different linguistic levels. There is, however, considerable debate regarding the choice of units and the validity of the cues that guide their demarcation. At the level of grammatical analysis, this debate focuses largely on the status of words *vs* sub-word units and on the generality of the linguistic properties that mark each type of unit. This paper suggests that the status of a unit type can be evaluated in terms of its informativity. A measure of informativity is obtained by assessing the influence that different unit boundary types have on text compressibility. The results obtained from this initial study support a pair of general conclusions. The first is that unit boundaries primarily reflect a statistical structure, and that the typological variability of linguistic cues reflects the fact that they serve a secondary reinforcing function. The second is that word boundaries are the most informative boundary type, and that the demarcation of words provides the most informative description of the regular patterns in a language.

KEYWORDS: linguistic units, words, abstractive perspective, information theory, Shannon entropy, Kolmogorov complexity

1. Introduction

To a large extent, modern linguistic approaches operate with a descriptive vocabulary inherited from earlier traditions. Familiar grammatical categories, word classes, construction types and a wide range of other classificatory notions survive mostly intact from classical models. In some cases, these terms preserve sources of unclarity that were present in the classical tradition, as Matthews (1972: 160ff.) notes. However, the more general problem lurking behind the use of this vocabulary is that descriptions are stated in terms for which there are no generally accepted definitions. This problem is not confined to traditional nomenclature; a modern student will struggle to find an explicit contemporary definition of the notion of 'morpheme', and many of the labels attached to syntactic descriptions are defined ostensively.

What can be called 'the problem of linguistic commensurability'

(received January 2016)

arises within a single language when descriptions attempt to classify different expressions as occurrences of the same units, categories or constructions, though the problem takes an especially acute form in cross-linguistic descriptions. Even in cases where definitions are available, they tend to refer to specific properties of individual languages and, consequently, fail to generalize in any useful way to other languages or language types. Although this issue has been broached in a general form in the typological literature (Haspelmath 2010; Croft & Van Lier 2012), the problem of commensurability has been most intensively investigated in connection with linguistic units.

1.1. The status of words and other units

The status of words has been a particular focus. Matthews (2002: 266) summarizes the collection of typological studies in Dixon & Aikhenvald (2002) by observing that they "make clear not just that criteria conflict, but that different linguists may resolve some kinds of conflict very differently". Haspelmath (2011: 70) echoes this bleak assessment in acknowledging that "'Words' as language-specific units are often unproblematic […] but the criteria employed in different languages are often very different". Following a comprehensive cross-linguistic survey of prosodic domains, Schiering *et al.* (2010: 657) likewise conclude that "the 'word' has no privileged or universal status in phonology, but only emerges through frequent reference of sound patterns to a given construction type in a given language". The incommensurability of language-specific definitions of words has provoked a chorus of pessimistic conclusions about the indeterminacy of words, the status of word-based generalizations and approaches and even the viability of the morphology/syntax split. However, the existing literature mainly provides an indictment of current linguistic methodology. As in nearly all domains of linguistics, the description of grammatical units employs familiar terms for which there are no adequate cross-linguistic definitions.

This paper suggests that the inadequacy of current linguistic definitions is fully compatible with one traditional conception of words. On this view, words and other linguistic units are not independent components from which larger expressions are constructed but are, instead, abstracted from larger utterances. The idea that words are abstracted from utterances, and sub-word units are in turn abstracted from words is stated with particular clarity by Bloomfield (1914: 65):

> it has long been recognized that the first and original datum of language is the sentence, — that the individual word is the product

> of a theoretical reflection which ought not to be taken for granted, and, further, that the grouping of derived and inflected words into paradigms, and the abstraction of roots, stems, affixes, or other formative processes, is again the result of an even more refined analysis.

A similar perspective is expressed, within a different tradition, by Robins (1959: 128):

> On the other hand words anchored, as it were, in the paradigms of which they form a part usually bear a consistent, relatively simple and statable grammatical function. The word is a more stable and solid focus of grammatical relations than the component morpheme by itself. Put another way, grammatical statements are abstractions, but they are more profitably abstracted from words as wholes than from individual morphemes.

The units abstracted from larger forms will be 'emergent' in essentially the sense of Schiering *et al.* (2010: 657), reflecting "frequent reference […] to a given construction type in a given language". What factors could then guide the frequency-driven abstraction of emergent units? The linguistic literature on word demarcation suggests that the abstraction of units in different languages cannot be guided by linguistic cues of the sort enumerated in Haspelmath (2011). The most obvious problem is that these cues vary across languages in ways that reflect cross-linguistic differences in sound patterns and grammatical structure. It is in fact hard to imagine how the situation could be otherwise, or why one would expect languages of different types to employ uniform strategies for marking units. The cues that are available for marking unit boundaries in a language (e.g. stress, pitch-accent, harmony, boundary lengthening, etc.) will depend on the phonological properties of a language, and even shared cues may perform distinct functions in different languages. Given the variation in the sound systems of the world, no single cue will be universally applicable. A more fundamental challenge (though one which has received somewhat less attention in the linguistic literature) is that there is no natural generalization of language-specific cues. If one language appears to use stress to mark boundaries and another contains domains over which harmony applies, what is the basis for identifying these as units of the same type? Parallel problems confront accounts that invoke clusters of cues, which must be assumed to pick out the same units.

1.2 Statistical abstraction

From an abstractive perspective, the cross-linguistic variability of linguistic cues does not establish the intrinsic indeterminacy of units such as words. Instead, variability reflects the fact that these cues, being language-specific, can serve at most a secondary function, reinforcing a structure that is exhibited by languages, irrespective of variation at the level of sound patterns and grammatical systems. What might this structure be? The hypothesis explored in this paper is that the units that emerge in different languages are abstracted on the basis of recurrent statistical patterns, specifically patterns of syntagmatic and paradigmatic interpredictability.

It has long been known that languages exhibit statistical regularities. Over a century ago, Markov (1913) outlined how the theory of probability could be extended to account for chains of linked events (since known as 'Markov chains'). Applying this analysis to a text of Pushkin's novel *Eugene Onegin*, Markov showed how the probability of a letter being a vowel depended on the preceding vowel or consonant. Shannon (1948) subsequently introduced a measure, which he termed "entropy", to quantify the amount of information in discrete communication. In Shannon (1951), he used an entropy measure to provide the first rigorous statistical analysis of character and word sequences in English text, inspiring a tradition of statistical analysis.

Although statistical patterns play no role in the linguistic literature cited above, they provide the basis for a parallel psycholinguistic and computational literature on word segmentation and recognition. The approaches within this broad tradition subsume the work on word recognition in Marslen-Wilson & Welsh (1978), Marslen-Wilson & Tyler (1980) and Balling & Baayen (2012), neural network-based predictive models (Elman 1990), and statistical models of word segmentation (Goldwater *et al.* 2009). Of particular importance to these diverse approaches is the predictive structure of language. The general approach to word recognition and segmentation developed in this tradition exploits the fact that entropy (roughly, uncertainty about the segments that follow) varies systematically across an utterance.

Entropy is relatively high at the beginning and the end of a word because there is much uncertainty on what may follow. Entropy generally decreases as more of a word is processed but may increase at morph boundaries, typically where sequences of derivational or inflectional material could follow. 'Words' can thus be abstracted from this predictive structure as sequences with the following statistical properties:

> Observations about predictability at word boundaries are consistent with two different kinds of assumptions about what constitutes a word: either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units (but to a lesser degree than the beginning of a word predicts its end). (Goldwater *et al.* 2009: 22)

Traditional 'words' correspond to sequences with high predictive value. In a syntagmatic expansion, as noted above, they occur between peaks of uncertainty in an utterance (Pléh & Juhász 1995), and reduce uncertainty about following words (Hale 2001, 2003, 2006; Levy 2008; Dye *et al.* 2016). As shown in the literature on inflectional entropy (Kostić 1991, 1995; Kostić *et al.* 2003; Moscoso del Prado Martín *et al.* 2004; Milin *et al.* 2009b; Ackerman & Malouf 2013; Blevins 2016), sequences with high syntagmatic predictive value also reduce uncertainty about paradigmatic variants that occur within other utterances in a language (or corpus). Furthermore, the paradigmatic entropy effect reported by Milin *et al.* (2009a) suggests that the distribution of word-sized sequences reduces speakers' uncertainty not only about the existence but also about the distribution of corresponding forms of other items.

The idea that sequences corresponding to words serve a mainly predictive function fits within a broadly implicational conception of language that has come into sharper focus over the past decade, partially due to a convergence of information-theoretic and discriminative learning perspectives. From this perspective, the central organizational principles of a language relate more to the reduction of uncertainty than to the signalling of meaning. The studies of paradigmatic entropy cited above provide various means of measuring the degree to which members of inflectional paradigms (and, to a lesser extent, derivational families) reduce uncertainty about other members of a paradigm (or family). The literature on syntagmatic entropy provides corresponding measures (often formulated in terms of 'surprisal') that express how sequences reduce uncertainty about following elements. Taken together, these independently-derived measures provide a cohesive conception of language structure in terms of uncertainty-reducing patterns of interpredictability.

There are other general implications of this type of approach. The first is that the absence of invariant linguistic cues does not call into question the status of the units, categories or constructions that are cued. Rather, it is the role that these cues play in defining units that requires reassessment. A second implication relates to the source of the intractability of what Spencer (2012) terms “The Segmentation Problem”. Given that uncertainty is continuous, not discrete, it is only by smoothing that we obtain boundaries and units. Generalizing over these units lends a measure of support to the traditional claim that words are “a more stable and solid focus of grammatical relations than the component morpheme” (Robins 1959:128). But this does not necessarily determine a unique segmentation of the speech stream, least of all at the sub-word level. Although much of the recent literature probing the status of units focusses on discrepancies between the cues that demarcate words in different languages, these discrepancies arise precisely because there are phonetic and other cues which, with varying degrees of reliability, provide secondary marking of predictive, statistically-independent, sequences (Goldwater *et al.* 2009: 22) or their boundaries.

It is sometimes suggested that the lack of universal cues for word demarcation supports analyses in terms of even smaller units. However, this type of proposal tends to rest on a peculiar kind of double standard. Although grammatical words may be imperfectly demarcated, sub-word units – including, significantly, roots – are rarely if ever cued at all by phonetic properties.[1] There is, for example, no discrepancy between the ‘grammatical morpheme’ and the ‘phonological morpheme’ in any given language for the simple reason that there is no such thing as a ‘phonological morpheme’. The absence of language-internal cues immediately precludes the possibility of cross-linguistic discrepancies. So there is no general problem of cross-linguistic commensurability for morphemes, not because there are invariant cross-linguistic cues, but because there are not even reliable language-specific cues. Hence the objection that grammatical words are not reliably and invariantly cued in the speech stream provides no motivation for shifting the focus of morphological analysis onto units smaller than the word (such as stems, roots or morphemes), since these units require an even greater degree of abstraction from the speech signal.

Nevertheless, an abstractive perspective suggests a way of breaking the mold of a word-based/morpheme-based dichotomy and capturing some of the basic intuitions about linguistic patterning expressed in the models of morphemic analysis in Harris (1942) and Hockett (1947; 1954). Uncertainty will be distributed over an utterance, and not concentrated solely at word boundaries. Although uncertainty is expected to be high at word boundaries, there will also be spikes at larger unit boundaries, corresponding to sentence boundaries, along with smaller peaks at the boundaries of recurrent sub-word sequences. The following sections now test these expectations by comparing the informativeness of different boundary types.

2. Measures of informativeness

The approach outlined below falls within a general probabilistic conception of language structure and use (summarized in Bod *et al.* 2003). There is as yet no broad consensus within this tradition regarding the exact principles that underlie language comprehension and production. However, it is now much more widely accepted that the communicative function of linguistic systems is the contingent outcome of complex processes that can best be modelled in probabilistic terms. Approaching language as a probabilistic system immediately raises questions about the status of linguistic units such as sentences, words, and morphs, the nature of their boundaries, and the types of statistical and linguistic cues that guide their abstraction.

The indeterminacy of unit boundaries in general, and of word boundaries in particular, gives rise to questions regarding the amount of information carried by boundaries. Intuitively, the importance of units – and their boundaries – depends on how predictable they are within a language. The more predictable a cue is from the context in which it occurs, the less important it is for the successful communication of a message. At one extreme, cues that can be fully predicted from a context are fully uninformative, and formally redundant. This determines a general relation between predictability and compressibility. The more predictable the cues are in a message, the more the message can be compressed relative to its original form.

2.1 Shannon entropy

Information theory provides two related measures of the information associated with linguistic units: the predictability of new units in terms of Shannon entropy (Shannon 1948), and the minimal length of the description of the unit, known as Kolmogorov complexity (Kolmogorov 1965). As noted in Section 1.2 above, Shannon entropy provides a measure of information content that is equivalent to the average unpredictability (the less probable an event is, the more information it contains). Shannon (1951) shows that the entropy of written English is rather low: a few letters often suffice to predict what is coming next. Following on from Shannon's initial estimation of the entropy of English, a large body of subsequent work has been devoted to calculating probability distributions over sequences of linguistic elements. These distributional models, often known as 'language models', have been applied to characters and other orthographic elements of increasing complexity in order to estimate entropy as accurately as possible (Teahan & Cleary 1996; Brown *et al.* 1992; Cover & King 1978). Information theory has also been applied to the task of identifying boundaries in a separate linguistic literature that includes the early proposals of Harris (1954) and the more recent work of Goldsmith (2001a,b).

Yet standard entropy measures presuppose that objects to be encoded are outcomes of a known random source, and only the characteristics of that random source (its probability distribution) determine the encoding. The characteristics of the objects that are the outcomes of this source are not relevant, as reflected in Shannon's idealization of communication as "select[ing] from a set of possible messages" in the following passage. As Shannon explains below, the properties of individual messages – including their meaning – are "irrelevant to the engineering problem" of message selection:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon 1948: 22)

Estimating the probability distribution of a random source of unit boundaries quickly leads to a number of practical challenges, which also arise in measuring the entropy of a text (discussed in Shannon 1951). One set of issues arises in evaluating the different language models (i.e. different probabilistic descriptions) of the language under consideration. These models can be based on character or word sequences but could involve any kind of grammatical constituent. Assessing the different formalisms that could provide the basis for these descriptions raises further complications.

2.2 Kolmogorov complexity

One practical advantage of using Kolmogorov complexity is that we are not required to make probabilistic assumptions about a source, and we can consider the information content of an object in isolation.

Kolmogorov stresses this point in his explanation of the relationship between Shannon entropy and Kolmogorov complexity:

> our definition of the quantity of information has the advantage that it refers to individual objects and not to objects treated as members of a set of objects with a probability distribution given on it. (Kolmogorov 1983: 37)

Grünwald & Vitányi (2004: §2.3) show that there is a correspondence between Shannon entropy, H(*X*) (where *X* is a random variable with a set of outcomes), and the expected Kolmogorov complexity, K(*x*) (where *x* ranges over the outcomes associated with *X*). Hence, these measures largely coincide in the present study, given that the study always operates at the level of average (i.e. expected) Kolmogorov complexity, and that this value closely corresponds to Shannon entropy.

Nevertheless, the use of Kolmogorov complexity does not avoid all of the types of practical problems that arise in applying Shannon entropy. In particular, the standard formulation of Kolmogorov complexity in terms of "minimal description length" (MDL; Rissanen 1978, 2007) requires the choice of a description language, just as Shannon entropy requires the choice of a language model. One possibility, explored in Sagot & Walther (2013), involves adopting a fixed description language and using an MDL metric to evaluate alternatives expressed within that language.

Yet a distinctive practical advantage of an MDL approach is that it can also use general-purpose data compression algorithms to approach the redundancy-free ideal formalized by Kolmogorov complexity. The attractiveness of this kind of theory-free measure has led a community of researchers to investigate the compressibility of linguistic utterances or structures as a measure of complexity. Previous work in this tradition includes Juola (1998, 2007); Bane (2008); Sadeniemi *et al.* (2008) and Moscoso del Prado Martín (2011). In the approach outlined below, it is especially useful to be able to apply a naive measure directly to unannotated corpora.

3. Measuring the informativeness of unit boundaries

The study outlined below uses an approach based on Kolmogorov complexity to explore the amount of information that categories, as structural entities, express in a language by means of their boundaries. As in earlier studies, we operationalize Kolmogorov complexity by means of general-purpose data compression algorithms. However, rather than assuming that there is a single normative segmentation to be discovered, we investigate a range of alternative boundary segmentations.

3.1 Method

To explore the information that unit boundaries add to language, we use the size of compressed linguistic data as a measure of information that data contains. We quantify the information that category boundaries carry by controlling the presence of category boundaries in the data, and by comparing the resulting changes in compressed data size. Significantly, we use lossless compression methods to ensure that no linguistic data are discarded for the sake of achieving high compression ratios.

The Lempel-Ziv compression algorithms (Ziv & Lempel 1977, 1978), known as LZ77 and LZ78, are among the most popular lossless compression algorithms.[2] These algorithms have been implemented in the standard Unix tools *gzip* and compress, which were used in this study. Of particular importance for the generality of this approach, the algorithms have been shown to converge towards source entropy (Wyner & Ziv 1994).

Both algorithms achieve compression by replacing repeated occurrences of data by references to an earlier occurrence. A reference contains the starting position and length of the earlier occurrence. One algorithmic difference between the two variants is that while passing through all data, LZ77 looks back for earlier occurrences in a preceding buffer whereas LZ78 updates a dictionary. Hence, LZ77 is often said to use a 'sliding window' whereas LZ78 is characterized as 'dictionary-based'.

For language data, we draw on the *Europarl* corpus (Koehn 2005), containing the proceedings of the European Parliament. This is a particularly useful and attractive source. The availability of parallel texts allows us to control semantic and pragmatic content. The range of languages represented also permits follow-up studies that incorporate additional languages and language families. The analyses below are applied to an initial set of texts in English, Estonian, Finnish and Hungarian.

The method for measuring boundary informativeness proceeds in three steps. We begin by performing a sanity check to test whether compression with common algorithms allows us to differentiate between linguistically-motivated and randomly-distributed unit boundaries. We first create multiple versions of a matched text version in which boundaries are placed randomly with the same probability of occurrence as real word boundaries, and then compare the mean compression ratio to the ratio based on data with linguistically-motivated boundaries. The compression ratio of the version with linguistically-motivated boundaries is expected to be considerably higher than the ratio of the version with randomly-placed boundaries.

The second step explores the relation between boundaries of various unit types and the amount of information that each type conveys. We investigate this relation by randomly removing an increasing proportion of boundaries. The compression ratio is then plotted as a function of the number of removed boundaries. Without any particular a priori assumptions, we would expect this relation to be linear.

The final step attempts to control for extrinsic distributional differences. In assessing the difference in information content between boundaries of different category types, we need to disregard variation in the probability of occurrence due to irrelevant factors such as the fact that there are far fewer sentences than words. We do that by normalizing the compression rates by the number of boundaries.

3.2 Data

The *Europarl* data from the parallel corpora used here comprise 1,924,942 sentences in English, containing 47,460,063 words, 651,746 sentences in Estonian, containing 11,214,221 words, 1,924,942 sentences in Finnish, containing 32,266,343 words, and 624,815 sentences in Hungarian, containing 12,420,578 words. The *Europarl* data naturally contain sentence and word boundaries, but not morph boundaries. In fact, no large corpora are available that provide manual or semi-manual annotations of morph boundaries. One way of addressing this problem is to make use of the results of unsupervised morphological segmentation techniques (Hammarström & Borin 2011). For instance, Bane (2008) measures morphological complexity using *Linguistica* (Goldsmith 2006), a system also based on an MDL principle. The implemented algorithm, however, permits at most a single morph boundary in a word.

More importantly, state-of-the art segmentation algorithms achieve between a 75% and 80% F-score for English (Dasgupta & Ng 2006). As impressive as these results appear, they still leave a considerable gap with gold-standard segmentations. For this reason, we use gold-standard surface-level morphological segmentations of the 120,000 English words in CELEX (Baayen *et al.* 1995) compiled by Creutz & Lindén (2004) to tag words in the English part of our dataset. This approach allowed us to cover an encouraging 96.1% of the words in the English part of *Europarl*, though comparable rates were not attainable for the other languages in our study.

3.3 Informativeness of sentence and word boundaries

We start with a text *T* that contains all category boundaries (i.e. sentence, word, and morph boundaries). We then generate a version of *T* without any boundaries, *Tn*, which will serve as a baseline. Boundaries of different categories are subsequently added to *Tn*, to show what compression reveals about the information content of each of the boundary types. Thus, we end up with the following text variants:[3]

1. *Tn*: no boundaries;
2. *Ts*: sentence boundaries;
3. *Tw*: word boundaries;

Compression allows us to assess the information content of a type of category boundary. For example, the difference in size between a compressed *Tn* and a compressed *Tw* will represent the information content of the word boundaries in *Tw*. To understand how this size difference depends on the number of boundaries that are added, we start with *Tn* for each category and see how the difference in size develops as successively larger numbers of boundaries are added at valid positions.

As mentioned in Section 3.1 above, we also test whether compression allows us to differentiate between linguistically-motivated and random unit boundaries by creating matched text versions in which boundaries are placed randomly with the same probability of occurrence as the linguistically-motivated boundaries. The only constraint we put on this placement is that any two boundaries should be at least two characters apart. We expect that compression of a text with linguistically-motivated unit boundaries will be significantly different from compression of the same text with boundaries placed at random intervals. The addition of boundaries will also add information to the language data, leading to the expectation that compression ratios will decrease as more boundaries are added, both in general, and as compared to *Tn*. Finally, if compression reveals differences between linguistically-motivated and randomly-placed boundaries, we expect that random boundaries will result in progressively lower compression ratios as more boundaries are added.

These comparative analyses are presented from two perspectives below. To provide a measure of 'absolute' complexity, we compare the decrease (in %) in the compressed size of files with and files without boundaries. To provide a measure of 'relative' complexity, we compare the compression ratio (normalizing the increase in compression relative to file size), again in files with and files without boundaries.

Figures 1 and 2 below now illustrate the effect that sentence boundaries have on compressibility, viewed in terms of 'absolute' and

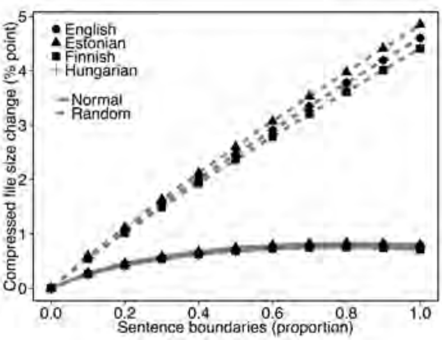

Figure 1. Relation between sentence boundaries and file size change.

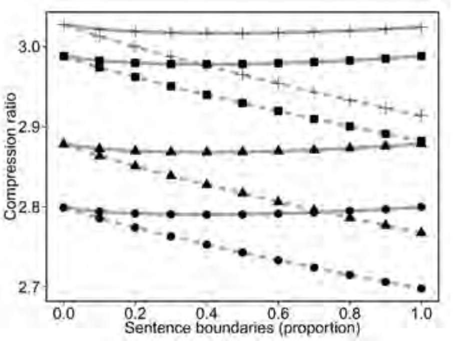

Figure 2. Relation between sentence boundaries and compression ratio.

'relative' complexity. For each language, we plot the difference in compressed size between *Tn* and *Ts* (Figure 1) and the compression ratio of *Ts* (Figure 2) as a function of the proportion of category boundaries present in the text. In the same figures we plot the random boundaries.

Figure 1 shows that the addition of randomly placed boundaries with the probability of occurrence of sentence boundaries leads to a linear increase in added information. Figure 2 shows that these boundaries produce a correspondingly linear decrease in compression ratio. Both effects are as expected. Figure 1 also indicates that the difference in compressed size between *Tn* and *Ts* increases as more sentence boundaries are added. However, this increase is not linear and even begins to asymptote once approximately 60% of the sentence boundaries are present. This pattern is again mirrored by the stabilization and slight increase in compression rates in Figure 2 at the point where 60% of the sentence boundaries are present.

We next consider word boundaries, and plot the same functions for each language in Figures 3 and 4. The patterns determined by linguistically-motivated and random boundary placement again differ significantly.

Interestingly, the largest change in the compressed size difference again appears when 60% of word boundaries are introduced. After this point, the difference again decreases. This development can be best understood by considering the compression ratio in Figure 4, which initially drops as boundaries are added to the data, but then increases once 30% of boundaries have been introduced. This pattern suggests that, as the presence of word boundaries increases, the compression algorithm becomes increasingly successful in generalizing

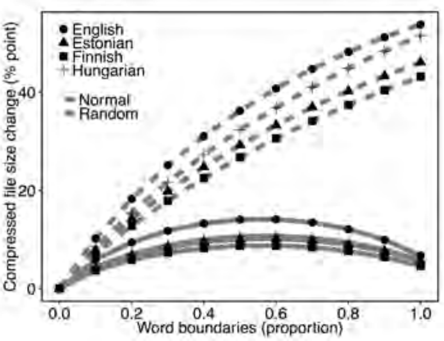

Figure 3. Relation between word boundaries and file size change.

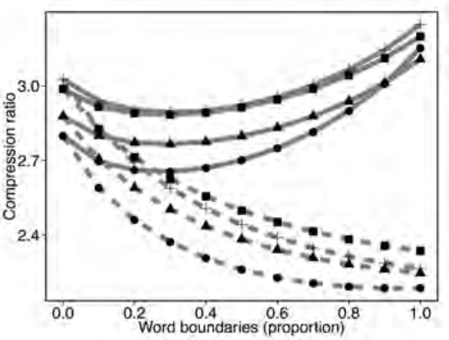

Figure 4. Relation between word boundaries and compression ratio.

the systematic regularities that they mark in the data.

Figures 5 and 6 now zoom in on the data for linguistically motivated boundaries from Figures 3 and 4 in order to clarify the informativeness of boundaries at the word level. When we closely examine the patterns determined by linguistically motivated boundary placement in Figures 5 and 6, we see that Finnish shows the smallest increase in compression file size and English shows the greatest increase, with Estonian and Hungarian in the middle. Interestingly, the growth curve for English shows a steeper increase than those of the other languages, implying that word boundaries carry more information for English than for the other languages in the sample. The higher infor-

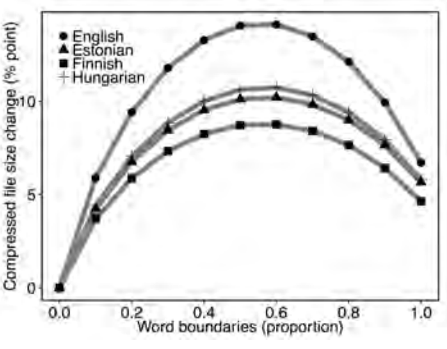

Figure 5. Relation between boundaries and file size change.

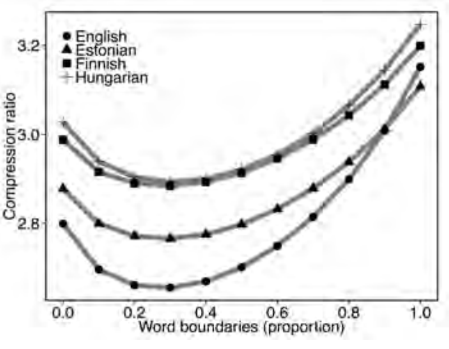

Figure 6. Relation between boundaries and compression ratio.

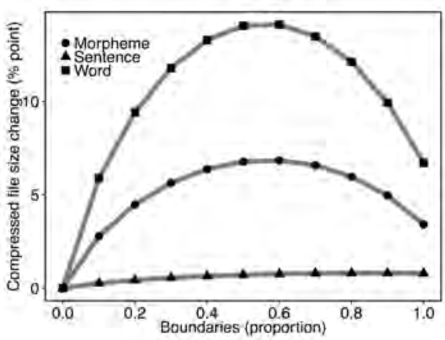

Figure 7. Relation between boundaries and file size change in English.

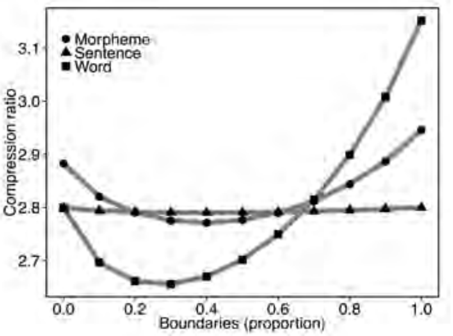

Figure 8. Relation between boundaries and compression ratio in English.

mation load carried by word boundaries in English accords with the observation that the morphological structure of English is not only poorer but also less transparently segmentable than that of the other languages.

3.4 Informativeness of morph boundaries

Having considered the information content of sentence and word boundaries, we now consider the content of morph boundaries in the language, English, for which we have gold-standard-level morphological segmentations. Figures 7 and 8 plot the compressed file size change and compression ratios for each type of boundary.

The graph in Figure 7 shows that word boundaries add the most information, followed by morph and sentence boundaries. This pattern is again mirrored by the graph in Figure 8 which shows that the addition of word boundaries achieves the highest compression ratio, followed by morph and sentence boundaries.

4. Discussion

The graphs of compressed size change exhibit a strikingly consistent decrease after about 60% of linguistically motivated boundaries have been added. This pattern seems to be driven by a compression ratio that initially drops but starts rising rapidly. This means that the LZ78 compression algorithm gets increasingly better at compressing the same data as the number of boundaries increase. The data with randomly placed boundaries does not show such development, which confirms that the effect is driven by the nature of linguistically motivated boundaries

4.1 Non-linear increases in compression differences and ratios

One explanation for the U-shaped compression ratio may come from the distribution of units defined by linguistically motivated boundaries. That is, words and morphs are known to exhibit a Zipfian distribution (Zipf 1935), in which the frequency rank of a word or morph tends to be inversely proportional to its actual frequency. This distributional bias makes it likely that newly added boundaries will mostly demarcate higher frequency words, boosting the algorithm's capacity to achieve better compression and therefore better compression ratios.

To test whether the Zipfian distribution of linguistic units explains the observed non-linearity, we created two variants in which boundaries are added at linguistically valid positions separating either low-frequency or high-frequency units. The variants were generated by sampling from the lower and higher quartile of the Zipfian distribution, keeping the number of boundaries equal. Figures 9 and 10 show the results for the compression size difference with word boundaries in English.

Figure 9 shows that when word boundaries are added in low frequency contexts, the amount of information measured by compression keeps increasing. This contrasts with the pattern in high frequency contexts in Figure 10.

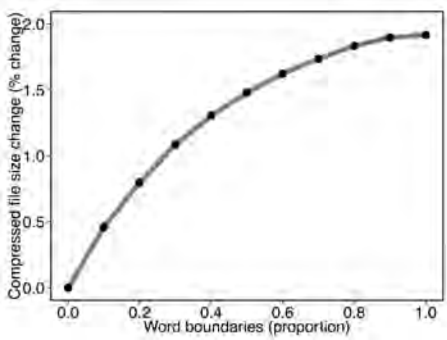

Figure 9. File size difference for boundaries added at low frequency contexts.

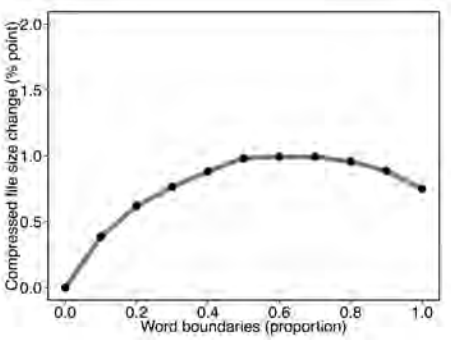

Figure 10. File size difference for boundaries added at high frequency contexts.

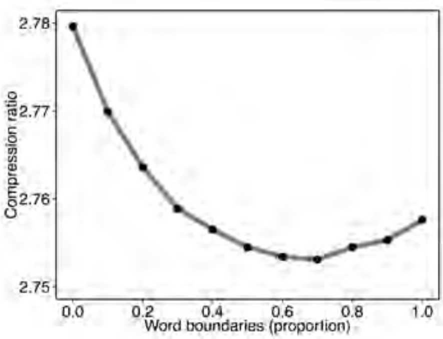

Figure 11. Compression ratios for boundaries added at low frequency contexts.

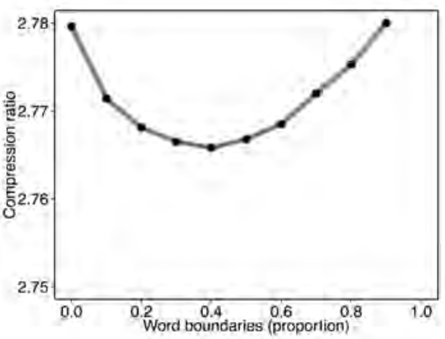

Figure 12. Compression ratios for boundaries added at high frequency contexts.

The contrast between compression size differences is again mirrored by the change in compression ratio when adding boundaries in Figures 11 and 12.

4.2 General conclusions

Although the studies reported above are largely exploratory in nature, the results obtained support a number of tentative conclusions. The increase in the compression ratio for randomly placed boundaries at the word level is considerably lower than that of motivated word boundaries, confirming the general validity of the approach. The results obtained are also fully consistent across algorithm implementations. The comparisons of file size change and compression ratios also show some suggestive patterns. The largest change in the compressed file size appears to occur when approximately 60% of word boundaries are introduced. This pattern occurs in each of the languages, as exhibited in Figure 5. In contrast, compression ratios appear to exhibit a different threshold. Ratios initially drop as boundaries are added to the data, but then increase once 30% of boundaries have been introduced, as shown in Figure 6.

We interpret these patterns as reflecting a process by which the compression algorithm becomes increasingly successful in generalizing the systematic regularities that word boundaries mark in the data. However, our initial hypotheses contrasted types of boundaries and made no specific assumptions about numerical thresholds. Hence we merely observe here that the different thresholds observed for file size change and compression ratios invite further investigation with different languages and corpora, and we conclude with a summary of the variation across boundary types.

The analyses presented above suggest that the informativeness of boundaries differs across types and, in varying degrees, across languages. In all of the languages sampled, sentence boundaries add relatively little information. Word boundaries are the most informative, dividing corpora into subsequences that are maximally amenable to compression. Morph boundaries lie between these extremes. The modest information added by sentence boundaries is roughly comparable across our languages. In contrast, the information added by word boundaries appears to be language dependent. As noted in the discussion of Figures 5 and Figure 6, word boundaries add more information in English than they do in the other languages in the sample. This contrast is plausibly attributed to general differences in morphological structure. More information is encoded morphologically in Estonian, Finnish and Hungarian than in English.[4] This entails that there are fewer words per sentence in the texts of these languages.

Moreover, the form in which information is expressed morphologically appears relevant to the information load of words. English approaches the isolating ('uninflected') ideal, retaining few exponents from the earlier fusional stages of the language. Hungarian and Finnish – at least in its standardized written form (Karlsson 1999: §22) – are both conventionally described as agglutinating ('beads on a string') languages. Modern Estonian falls between these types, having, as Dressler (2003: 468) notes, "changed from an agglutinating type to a predominantly inflecting-fusional language". The contrast between isolating, fusional and agglutinating types in part reflects the transparency of morphological structure. Intriguingly, this transparency is directly mirrored by the variation in the informativeness of word boundaries measured by the difference in compression ratios in Figure 6.

As noted earlier, the addition of word boundaries has the greatest effect on compression ratio in English. In the other languages, the effect correlates with degree of agglutination. The curves for Finnish and Hungarian largely converge, showing the least effect, with Estonian falling between the isolating and agglutinating types. Yet the neatness of this correlation is again disrupted by a contrast between compression ratio and compressed file size change. In Figure 5, Finnish and Hungarian again cluster together, but in this case they occur between English and Estonian.

4.3 Summary

Taken together, these initial results lend a measure of support to the traditional view that words are optimal-sized units for describing the regularities in at least the languages considered above. The limitations of sentences reflect the fact that they are individually too large and collectively too sparse. Morphs are more informative but exhibit less reliable patterns of syntagmatic and paradigmatic interpredictability than the larger units from which they are abstracted. It is the maximally predictive word-sized units that serve as primary focus of grammatical abstraction for the speaker. The usefulness of words for describing regularities enhances their role in a speaker's predictive language model. This in turn facilitates the innovation and preservation of language-specific cues that reinforce words or their boundaries.

Notes

[1] It has often been observed that there are restrictions on the composition of roots and other sub-word sequences. These restrictions have been described in terms of "morpheme structure constraints" (Stanley 1967) and by other mechanisms in the subsequent phonological literature. Significantly, these types of restrictions do not mark units or boundaries so much as perform a predictive function, by facilitating the identification of items or by reducing uncertainty about following sequences.

[2] A comparison of results obtained from other compression algorithms, notably Prediction by Partial Match (PPM), would strengthen the approach methodologically. But this comparison would not be expected to provide any insight beyond what can be obtained from LZ78.

[3] The boundary types considered below are defined positionally (as whitespace) and are thus not independent (as they would be if distinct boundary symbols, such as '+', '#', etc., were introduced). However, this does not affect the boundary counts, given that unit boundaries overlap systematically in a text. Every sentence boundary is also a word boundary, and every word boundary is a morph boundary. By extension, every sentence boundary is likewise also a morph boundary.

[4] Obvious examples include case forms that correspond to prepositional phrases in English, verb forms that express pronominal dependents or correspond to periphrastic constructions, etc.

Bibliographical references

Ackerman, Farrell & Malouf, Robert 2013. Morphological organization: The Low Conditional Entropy Conjecture. *Language* 89. 429-464.

Baayen, R. Harald; Piepenbrock, Richard & van Rijn, Hedderick 1995. *The CELEX Lexical Database* (Release 2, CD-ROM edn). Linguistic Data Consortium. University of Pennsylvania, Philadelphia, PA.

Balling, Laura & Baayen, R. Harald 2012. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125. 80-106.

Bane, Max 2008. Quantifying and measuring morphological complexity. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*. 69-76.

Blevins, James P. 2016. *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Bloomfield, Leonard 1914. Sentence and word. Transactions of the American

Philological Society 45. 65-75. Reprinted in Hockett 1970. 38-46.

Bod, Rens; Hay, Jennifer & Jannedy, Stefanie (eds.) 2003. *Probabilistic Linguistics*. Cambridge: Cambridge University Press.

Brown, Peter F.; Pietra, Vincent J. D.; Mercer, Robert L.; Pietra, Stephen A. D. & Lai, Jennifer C. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics* 18. 31-40.

Cover, Thomas M. & King, Roger 1978. A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory* 24. 413-421.

Creutz, Mathias & Lindén, Krister 2004. Morpheme segmentation gold standards for Finnish and English. *Publications in Computer and Information Science*. Report A 77.

Croft, William & Van Lier, Eva 2012. Language universals without universal categories. *Theoretical Linguistics* 38. 57-72.

Dasgupta, Sajib & Ng, Vincent 2006. Unsupervised morphological parsing of Bengali. *Language Resources and Evaluation* 40. 311-330.

Dixon, Robert M. W. & Aikhenvald, Alexandra Y. (eds.) 2002. *Word: A Cross-Linguistic Typology*. Cambridge: Cambridge University Press.

Dressler, Wolfgang U. 2003. Naturalness and morphological change. In Joseph, Brian & Janda, Richard (eds.), *The Handbook of Historical Linguistics*. Oxford: Blackwell. 461-471.

Dye, Melody; Milin, Petar; Futrell, Richard & Ramscar, Michael 2016. A functional theory of gender paradigms. In Kiefer, Ference; Blevins, James P & Bartos, Huba (eds.), *Perspectives on Morphological Structure: Data and Analyses*. (in press). Leiden: Brill.

Elman, Jeffrey L. 1990. Finding structure in time. Cognitive Science 14. 179-211.

Gammon, Edward Roy 1966. The statistical determination of linguistic units. Technical Report No. 99, Stanford University.

Goldsmith, John A. 2001a. On information theory, entropy and phonology in the 20th century. *Folia Linguistica* 34. 85-100.

Goldsmith, John A. 2001b. The unsupervised learning of natural language morphology. *Computational Linguistics* 27. 153-198.

Goldsmith, John A. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering* 12. 353-371.

Goldwater, Sharon; Griffiths, Thomas L. & Johnson, Mark 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112. 21-54.

Grünwald, Peter & Vitányi, Paul 2004. Shannon information and Kolmogorov complexity. arXiv preprint cs/0410002.

Hale, John 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. 159-166.

Hale, John 2003. The information conveyed by words in sentences. *Journal of Psychological Research*. 32. 101-123.

Hale, John 2006. Uncertainty about the rest of the sentence. *Cognitive Science*. 30. 643-672.

Hammarström, Harald & Borin, Lars 2011. Unsupervised learning of morphology. *Computational Linguistics*. 37. 309-350.

Harris, Zellig S. 1942. Morpheme alternants in linguistic analysis. *Language* 18. 169-180.

Harris, Zellig S. 1954. Distributional structure. *Word* 10. 146-162.

Haspelmath, Martin 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86. 663-687.

Haspelmath, Martin 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45. 31-80.

Hockett, Charles F. 1947. Problems of morphemic analysis. *Language* 23. 321-343.

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210- 231.

Hockett, Charles F. (ed.) 1970. *A Leonard Bloomfield Anthology*. Chicago: University of Chicago Press.

Joos, Martin (ed.) 1957. *Readings in Linguistics I*. Chicago: University of Chicago Press.

Juola, Patrick 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5. 206-213.

Juola, Patrick 2007. Assessing linguistic complexity. In Miestamo, Matti; Sinnemäki, Kaius & Karlsson, Fred (eds.), *Language Complexity: Typology, Contact, Change*. John Benjamins. 89-108.

Karlsson, Fred 1999. *Finnish: An Essential Grammar*. London: Routledge.

Koehn, Philipp 2005 Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, Volume 5. 79-86.

Kolmogorov, Andrei N. 1965. Three approaches to the quantitative definition of 'information'. *Problems of information Transmission* 1. 1-7.

Kolmogorov, Andrei N. 1983. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys* 38. 29-40.

Kostić, Aleksandar 1991. Informational approach to processing inflectional morphology: Standard data reconsidered. *Psychological research* 53. 62-70.

Kostić, Aleksandar 1995. Informational load constraints on processing inflectional morphology. In Feldman, Laurie (ed.), *Morphological Aspects of*

Language Processing. Hillsdate, NJ: Lawrence Erlbaum. 317-344.

Kostić, Aleksandar; Marković, Tanja; & Baucal, Aleksandar 2003. Inflectional morphology and word meaning: Orthogonal or co-implicative domains? In Baayen, Herald R. & Schreuder, Robert (eds.), *Morphological Structure in Language Processing*. Berlin: Mouton de Gruyter. 1-44.

Levy, Roger 2008. Expectation-based syntactic comprehension. *Cognition* 106, 1126-1177.

Markov, Andrei A. 1913. Primer statisticheskogo issledovaniya nad tekstom "Evgeniya Onegina", illyustriruyuschij svyaz ispytanij v cep. *Izvestiya Akademii Nauk*, Serija 6, 153-162.

Marslen-Wilson, William D. & Tyler, Lorraine K. 1980. The temporal structure of spoken language understanding. *Cognition* 8. 1-71.

Marslen-Wilson, William D. & Welsh, Alan 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10. 29-63.

Matthews, Peter H. 1972. *Inflectional Morphology: A Theoretical Study based on Aspects of Latin Verb Conjugation*. Cambridge: Cambridge University Press.

Matthews, Peter H. 2002. What can we conclude? In Dixon, Robert M.W. & Aikhenvald, Alexandra Y. (eds.), *Word: A Cross-Linguistic Typology*. Oxford: Oxford University Press. 266-281.

Milin, Petar; Filipovi 18.1. Inquadramento della nullità di protezione Đurđevi 18.1. Inquadramento della nullità di protezione, Dušica & Moscoso del Prado Martín, Fermín 2009a. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60. 50-64.

Milin, Petar; Kuperman, Victor; Kosti 18.1. Inquadramento della nullità di protezione, Aleksandar; & Baayen, R. Harald 2009b. Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In Blevins, James P. & Blevins, Juliette (eds.), *Analogy in Grammar: Form and Acquisition*. Oxford: Oxford University Press. 214-253.

Moscoso del Prado Martín, Fermín 2011. The myth of morphological complexity. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. 3524-3529.

Moscoso del Prado Martín, Fermín; Kosti 18.1. Inquadramento della nullità di protezione, Aleksandar & Baayen, R. Harald 2004. Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition* 94. 1-18.

Pléh, Csaba & Juhász, Levente 1995. Processing of multimorphemic words in Hungarian. *Acta Linguistica Hungarica* 43. 211-230.

Rissanen, Jorma 1978. Modeling by shortest data description. *Automatica* 14. 465-471.

Rissanen, Jorma 2007. *Information and Complexity in Statistical Modeling*. Springer, Berlin.

Robins, Robert H. 1959. In defence of WP. *Transactions of the Philological Society* 58. 116-144. Reprinted in *Transactions of the Philological Society* 2001. 99. 116-144.

Sadeniemi, Markus; Kettunen, Kimmo; Lindh-Knuutila, Tiina & Honkela, Timo 2008. Complexity of European Union languages: A comparative approach. *Journal of Quantitative Linguistics* 15. 185-211.

Sagot, Benoît & Walther, Géraldine 2013. Implementing a formal model of inflectional morphology. In *Proceedings of SFCM 2013, Communications in Computer and Information Science*. Berlin: Springer. 115-134.

Schiering, René; Bickel, Balthasar & Hildebrandt, Kristine A. 2010. The prosodic word is not universal, but emergent. *Journal of Linguistics* 46. 657-709.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27. 379-423. 623-656.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal* 30. 50-64.

Spencer, Andrew J. 2012. Identifying stems. *Word Structure* 5. 88-108.

Stanley, Richard 1967. Redundancy rules in phonology. *Language* 43. 393-436.

Teahan, W. & Cleary, J. G. 1996. The entropy of English using PPM-based models. In *Proceedings of Data Compression Conference 1996*. IEEE. 53-62.

Wyner, Aaron D. & Ziv, Jacob 1994. The sliding-window Lempel-Ziv algorithm is asymptotically optimal. *IEEE Transactions on Information Theory* 82. 872- 877.

Zipf, George K. 1935. *The Psychobiology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.