TECHNICAL REPORT

COMPUTER SCIENCE

Evolution of language structure: survival of the fittest in a statistical environment

Caroline Lyon

Report No 322

Dec 1998

Evolution of language structure: survival of the fittest in a statistical environment

Caroline Lyon

December 18, 1998

Abstract

The structure and other characteristics of natural language can be illuminated by an investigation into the real world conditions within which language functions. Language is taken as a medium in which messages are produced, and we investigate how it is best adapted to efficient coding. We examine naturally occurring phenomena and suggest how their exploitation would confer an evolutionary advantage.

The main focus of this paper is approached by first examining a related phenomenon: the expected distribution of some linguistic entities. It is shown how this is relevant to a language representation that can facilitate the efficient transfer of information.

This leads to an investigation into how naturally occurring events, such as pauses in speech, can be utilized to improve the transmission of information. Using arguments based on comparative entropy measures we show that there seems a certain inevitability to the development of structured language. This is a novel application of well known tools from Information Theory, and demonstrates how a relationship between prosodic information and the organization of language might be expected to develop.

1 Introduction

The aim of this paper is to show that an examination of the statistical nature of spoken and written communication will help explain why language is likely to evolve with a structure. We examine naturally occurring phenomena, suggest how their exploitation would confer an evolutionary advantage, and explain how this supports the development of structured language. This is based on a novel application of well known tools in Information Theory, which can illuminate many characteristics of natural language.

The transfer of information is taken as the principal role that language fulfills, and any consideration of other "speech acts" are left aside. Language is taken as a variable medium in which messages are produced, and we investigate how it is best adapted so that an acoustic speech signal can be efficiently coded.

In this paper some characterisitics of human speech that distinguish it from non-speech sounds are briefly reviewed (Section 2). Then we look at two aspects of natural language where efficient coding is supported. We first revisit a subject examined in the past by Mandelbrot [1] (Section 3). He proposed that a general statistical structure, independent of meaning, underlies human languages, and that language is "intentionally if not consciously produced in order to be decoded word-by-word in the easiest possible fashion". He supported his hypothesis by investigating the distribution of English words, and it is then shown how this is relevant to a language representation that can facilitate the efficient transfer of information.

This leads into the main topic. We report on an investigation into how naturally occurring events, such as pauses in speech, can be utilized to improve the transmission of information (Section 4). Using arguments based on comparative entropy measures we show that there seems a certain inevitability in the development of structured language.

This work developed out of investigations into data representation for automated natural language parsing [2, 15].

2 Selection for efficient and robust communication

There is a high biological cost in developing the physiology capable of producing speech. Humans can produce a much wider range of sounds than other species can, but in order to do this the human anatomy has evolved in a way that has incurred significant physiological disadvantages (Lieberman [3]). In spite of this, the human speech faculty has developed: presumably the ability to communicate by speech greatly outweighs the concomitant disadvantages.

It is instructive to examine the characteristics of human speech that distinguish it from non-speech sounds. First, Lieberman notes the high transmission rates that characterize speech: 15 to 25 phonetic segments per second, which map very roughly onto letters, can be produced or recognized. The identification of non-speech sounds is much slower: a maximum of 7 to 9 items per second. Secondly, he notes the larger range of sounds that only humans have the anatomy to produce. These include vowels like [i] and [u] which are less susceptible to perceptual confusion than some other phonetic segments, and more easily combined with other sounds.

Observing these characteristics of human speech, we see selection for speed, reliability and wider scope as language has evolved. Now, if speech has evolved to meet these requirements for efficient communication at some biological cost, we expect that other empirical factors will be exploited too. This paper examines the statistical environment in which speech operates, and shows why structured language is likely to evolve.

2.1 Representing speech as discrete words

The primary form of language is speech, an analog signal. The perceptual analysis of speech, and an understanding of how an utterance is segmented into words, is an active area of research [4]. However, for this paper we take as a starting point the assumption that a mapping from a speech signal onto discrete symbols can be done.

3 The distribution of words and efficient coding

This section introduces the principle that efficient coding is likely to be a factor in the organization of language. Before addressing the main subject of language structure in the following section we investigate how the distribution of different words can facilitate the efficient transfer of information.

As speed of production is a characteristic of human speech sounds, we expect faster methods of communication to be favoured rather than slower ones. Therefore, we expect that more common words will typically be shorter than less common ones.

3.1 The distribution of English words

The LOB corpus is a collection of English texts from different fields, totalling about 1 million words [5]. In it the most common words, are, in rank order (see page 19): [1]

the of and to a in that is was

it for he as with be on I his

The first 50 are typically short, closed class function words, as opposed to content words.

We find a characteristic distribution of words in English and other languages to which Shannon drew attention in his classic paper [6]. Informally stated, it models the fact that a small number of words are very common, but a significant number are used infrequently. For example, words that have a frequency of less than 1 in 50,000 make up about 20-30% of typical English language news-wire reports [7]. In the Brown corpus of about 1 million words 40% of word forms are *hapax legomena*: occurring only once [8].

A common way of displaying this distribution is to rank words into frequency order, and then show rank against actual frequency. If word forms are ranked in order of frequency, and r denotes rank, then there is an empirical relationship between the probability of the word at rank r occurring, $p(r)$, and r itself, known as Zipf's Law:

$$p(r) * r = constant$$

With the constant taken as about 0.1 this gives a surprisingly good approximation to word probabilities in English and other languages, and indicates the extent to which a significant number of words occur infrequently [6].

Relevance to automated speech and language processing

This typical distribution of words has a significant effect on speech and natural language processing systems. Constructing a vocabulary that will cover ordinary topics, or speech in limited domains, is no trivial task. Even with very large training sets, new words will occur later in unseen test sets.

3.2 Theoretical distribution of 'words'

Before we look further at the actual frequency distribution of words, let us examine a more general case. Consider a sequence of discrete elements, of which one special element serves to divide the sequence into sections.

This is analogous to a sequence of spoken words. Individual sounds of speech correspond roughly to letters which make up words, and strings of spoken words can be mapped onto written sequences of letters, separated by spaces.

Suppose there are n different elements in an alphabet $\mathcal{A}$ and that the special element that divides up the sequence is the space symbol x_s. Let the probability of the space symbol occurring be p_s.

Assume initially that (i) the other letters all have equal probability of occurring p_q, and (ii) that all letter combinations are legal.

Whenever x_s occurs a letter sequence, or 'word' is completed.

[1] Compare for interest gender differences in frequency and case: 'he', nominative, rank 12, 'his', genitive, rank 18, contrasts with 'her', accusative and genitive, at rank 29, 'she', nominative, at rank 30, 'him', accusative, at rank 49.

Now, consider the probability of 'words' of various lengths occurring. Suppose a 'word' is about to start.

Probability of 'word' length zero $= p_s$
Probability of any 'word' length 1 $= p_q * p_s$
Probability of any 'word' length $\lambda = (p_q)^\lambda * p_s$

The probability of longer 'words' occurring is less than that of shorter 'words'.

Now consider how many 'words' there are within each probability band. There are more possible 'words', or letter combinations, as λ increases: for 'words' of length λ there are n^λ different 'words'. Therefore,

Probability of particular 'word' length λ

$$= \frac{(p_q)^\lambda * p_s}{n^\lambda}$$

As λ increases there is both (i) less likelihood of any 'word' of that length occurring and (ii) many more 'words' of that length to choose from. Hence we have the initial pattern of distribution, where a few short strings occur frequently, while many longer ones occur rarely.

3.3 Exploiting a statistical phenomen

When Zipf described the distribution that bears his name he explained it by postulating a "Principle of Least Effort", where the shortest words are typically the most frequently used [9]. Bell, Cleary and Witten [10] point out that there is no need to invoke such a principle, since the Zipfian distribution will emerge without it, as described above. They support their view with a report on experiments with 100 specially trained monkeys (computer simulated ones), which typed random characters, constrained by observed frequencies.

However, the two positions complement each other: we see that a naturally occurring phenomenon can be exploited to confer an evolutionary advantage. By representing frequent words with shorter codes we approach a more efficient coding system. Compact codes will facilitate fast communication, which is one of the notable features of human speech. If a zipfian type distribution of words is likely to occur, we might expect that this would be used to advantage.

Mandelbrot's work

Mandelbrot drew attention to this phenomenon, and investigated the way in which the statistical structure of language is best adapted to word coding. [1]. However, he did not endear himself to the traditional Linguistic community by claiming ill-advised analogies with thermodynamics. His basic theory was misunderstood and his work quietly left aside.

4 Sequence structure and efficient coding

The first subject examined in this paper looked at the way that speech sounds were likely to be combined into words. We analyzed sequences of a limited number of elements: the characters of the alphabet.

Now we consider sequences with a much larger number of elements: an indefinite number of words that can be created out of these phonetic elements. We investigate how words are grouped together, and why statistical realities make certain modes of segmentation likely to evolve.

In this section we first describe the metrics that are used, and then illustrate their application in a related, but simpler, field.

4.1 Entropy and perplexity

This analysis is based on comparative measures of the entropy of sequential data. [2]

Entropy is a measure, in a certain sense, of the degree of uncertainty. If the entropy can be reduced, the predictability of the next element in an incomplete sequence is increased. A sequence represented in a way that lowers the entropy without reducing its representational power is a more efficient message carrier. Therefore, we would expect language to evolve so that it enabled lower entropy coding of a sequence of words.

This approach to the evaluation of language models has been used in automated speech recognition (ASR) for many years [13]. Typically, entropy is reduced by taking more of the context into account. If we know preceding words there is reduced uncertainty about the next word. In ASR applications language models generally use bigrams or trigrams instead of single words.

The new contribution we make is to show that the entropy can also be reduced by imposing a structure on a sequence of words. If the segmentation of speech is modelled along with the words, then the entropy declines. We conclude that it is likely that this phenomenon will have been exploited as language has evolved.

Definitions

Let $\mathcal{A}$ be an alphabet, and X be a discrete random variable. The probability mass function is then $p(x)$, such that

$$p(x) = probability(X = x), x \in \mathcal{A}$$

If we consider letter sequences the x's could be the 26 letters of the standard alphabet.

The entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in \mathcal{A}} p(x) * log \quad p(x)$$

Very informally, this combines information on how difficult it is to predict a letter together with its probability of occurring. If logs to base 2 are used, the entropy measures the minimum number of bits needed on average to represent X: the more limited the choice the less bits will be needed to describe it.

We talk loosely of the entropy of a sequence, but more precisely consider a sequence of symbols X_i which are outputs of a stochastic process. We estimate the entropy of the distribution of which the observed outcome is typical.

Often the related metric of *perplexity* is employed. If P represents perplexity and H entropy, then

$$P = 2^H$$

and P can be seen as a measure of the branching factor, or number of choices.

[2] For reference see Cover [11, chapter 2], or, for an introduction, Charniak [12, chapter 2].

4.2 Illustrations from letter sequences

Shannon's work on English texts

Though we are investigating groups of words, the subject is introduced by recalling Shannon's well known work on the entropy of letter sequences.

Shannon showed that the entropy H of written English, can be reduced in two ways. First, it declines as more of the statistics of the language are taken into account. H_n, the n-gram entropy, measures the amount of entropy with information extending over n adjacent letters of text, and $H_n <= H_{n-1}$.

Secondly, the entropy can be reduced if an extra character representing a space between words is introduced. The introduction of the space captures some of the structure of the letter sequence.

Shannon produced a series of approximations to the entropy H of written English, which successively take more of the statistics of the language into account. H_0 represents the average number of bits required to determine a letter with no statistical information. Thus, for an alphabet of 16 symbols $H_0 = 4.0$.

H_1 is calculated with information on single letter probabilities. If we knew, for example, that letter e had probability of 20% of occurring while k had 1%, then the letter e could have a shorter code than k. Messages using this alphabet could be coded with fewer bits than could be done without this information. H_1 would be lower than H_0.

H_2 uses information on the probability of 2 letters occurring together; H_n, called the n-gram entropy, measures the amount of entropy with information extending over n adjacent letters of text [3] and $H_n \leq H(n-1)$. As n increases the n-gram entropy declines: the degree of predictability is increased as information from more adjacent letters is taken into account. This fact is exploited in games where the contestants have to guess letters in words, such as the "Shannon game" or "Hangman" [13].

The formula for calculating entropy is given in Appendix 1.

Entropy and structure

The entropy can also be reduced if some of the structure of the letter strings is captured. As Shannon says "a word is a cohesive group of letters with strong internal statistical influences" so the introduction of the space character to separate words will lower the entropy H_2 and H_3.

With an extra symbol in the alphabet H_0 will rise: there will be more choice, less predictability. H_1 may go down because the space will be much more frequent than any other symbol, and this can outweigh the effect of the larger number of symbols. However, as there will be more potential pairs and triples, H_2 and H_3 could rise. But in practice the space symbol will prevent "irregular" letter sequences between words, and thus reduce the unpredictability. H_2 and H_3 do in fact decline. For instance, for the words

COOKINGCHOCOLATE

If a space is inserted the trigrams "N-G-C" and "G-C-H" will be replaced by "N-G-space", "G-space-C" and "space-C-H"..

The significance of boundary marking for ASCII data

For other representations too, the insertion of boundary markers that capture the structure of a sequence will reduce the entropy. Gull and Skilling [14] report on an experiment with a string

[3] This notation is derived from that used by Shannon. It differs from that used by Bell, Cleary and Witten [10].

	H_0	H_1	H_2	H_3
26 letter	4.70	4.14	3.56	3.3
27 letter	4.76	4.03	3.32	3.1

Table 1: From Shannon's work on letter sequences: a comparison of entropy for different n-grams, with and without representing the space between words

of 32,768 zeroes and ones that are known to be ASCII data organised in patterns of 8 as bytes, but with the byte boundary marker missing. By comparing the entropy of the sequence with the marker in different positions the boundary of the data is "determined to a quite astronomical significance level".

4.3 The entropy of strings of words

Now, a similar analysis can be employed to see how words are organised into structured constituents. In [15] Lyon and Brown showed how the entropy of text mapped onto part-of-speech tags could be reduced if clauses and phrases were explicitly marked. Syntactic markers can be considered analogous to spaces between words, or to virtual punctuation marks.

Consider, for example, how subordinate clauses are discerned. There may be an explicit opening marker, such as a 'wh' word, but often there is no mark to show the end of the clause. If markers are inserted and treated as virtual punctuation some of the structure is captured and the entropy declines. A sentence without opening or closing clause boundary markers, like

The shirt he wants is in the wash.

can be represented as

The shirt { he wants } is in the wash.

If this sentence is given part-of-speech tags the symbols '{' and '}' will represent two classes in the tagset. We call them virtual-tag1 and virtual-tag2. The entropy H_0 will be higher for the second representation because there are more tags from which to choose. The change in H_1 will depend on the frequency with which the new tags occur. However, H_2 and H_3 will decline if some of the structure is captured.

The part-of-speech tags have probabilistic relationships with the virtual tags in the same way that they do with each other. The pairs and triples generated by this second string exclude "unlikely" tags sequences such as (*noun, pronoun*), (*noun, pronoun, verb*) but include, for instance, (*noun, virtual-tag1*), (*noun, virtual-tag1, pronoun*) The entropy, H_2 and H_3, with virtual tags explicitly marking some constituents is lower than that without the virtual tags.

4.4 Analysis of MARSEC (Machine Readable Spoken English Corpus)

In a similar way the words from a speech signal can be segmented into groups, with periodic discontinuities. It has been shown that there is a relationship between prosody and syntactic structure, and that the placement of pauses and other discontinuities provide clues to syntactic structure [16, 17, 18, 19].

An example given by Ostendorf is

Mary was amazed Ann Dewey was angry.

which was produced as

Mary was amazed || Ann Dewey was angry.

where || represents a pause.

We have investigated how the entropy of sequences of words varies when discontinuities are represented. This research was carried out using the MARSEC corpus, which is annotated with prosodic markers.

The corpus has been mainly collected from the BBC, and is available free on the web. We have used part of the corpus, just over 26,000 words, comprising the 4 categories of news commentary (A), news broadcasts (B), lectures aimed at a general audience (C) and lectures aimed at a restricted audience (D).

The prosodic markers in MARSEC which we retain are the major and minor tone unit boundaries. The term "discontinuity" is taken to cover both these features. The major tone unit boundary can also be labelled as a pause.

The prosodic markup was done by two trained annotators. Some sections of the corpus have been marked up by both, and we see that there is a large measure of agreement, but not a total consensus. Prosodic perception is an inexact science. However, there are general standards that are widely accepted, and agreement between several annotators is one way of establishing norms for evaluation. In Table 2 we show some sample data as we used it, in which only the major and minor tone unit boundaries are retained. When passages were marked up twice, we chose one in an arbitrary way, so that each annotator was chosen about equally.

<table>
<tr><td>Key:</td><td></td></tr>
<tr><td>|| is a pause,</td><td>| is a minor discontinuity</td></tr>
<tr><td>annotator 1</td><td>annotator 2</td></tr>
<tr><td>we</td><td>we</td></tr>
<tr><td>heard</td><td>heard</td></tr>
<tr><td>automatic</td><td>automatic</td></tr>
<tr><td>fire</td><td>fire</td></tr>
<tr><td>|</td><td>|</td></tr>
<tr><td>a</td><td>a</td></tr>
<tr><td>few</td><td>few</td></tr>
<tr><td>yards</td><td>yards</td></tr>
<tr><td>away</td><td>away</td></tr>
<tr><td>|</td><td>||</td></tr>
<tr><td>we</td><td>we</td></tr>
<tr><td>drove</td><td>drove</td></tr>
<tr><td>on</td><td>on</td></tr>
<tr><td>||</td><td>||</td></tr>
<tr><td>a</td><td>a</td></tr>
<tr><td>jet</td><td>jet</td></tr>
<tr><td>appeared</td><td>appeared</td></tr>
</table>

Table 2: Example of MARSEC corpus with minimal prosodic annotations

Taking the discontinuities as virtual words, we find that the minor discontinuities have a probability of approximately 0.15, pauses 0.04, jointly 0.19.

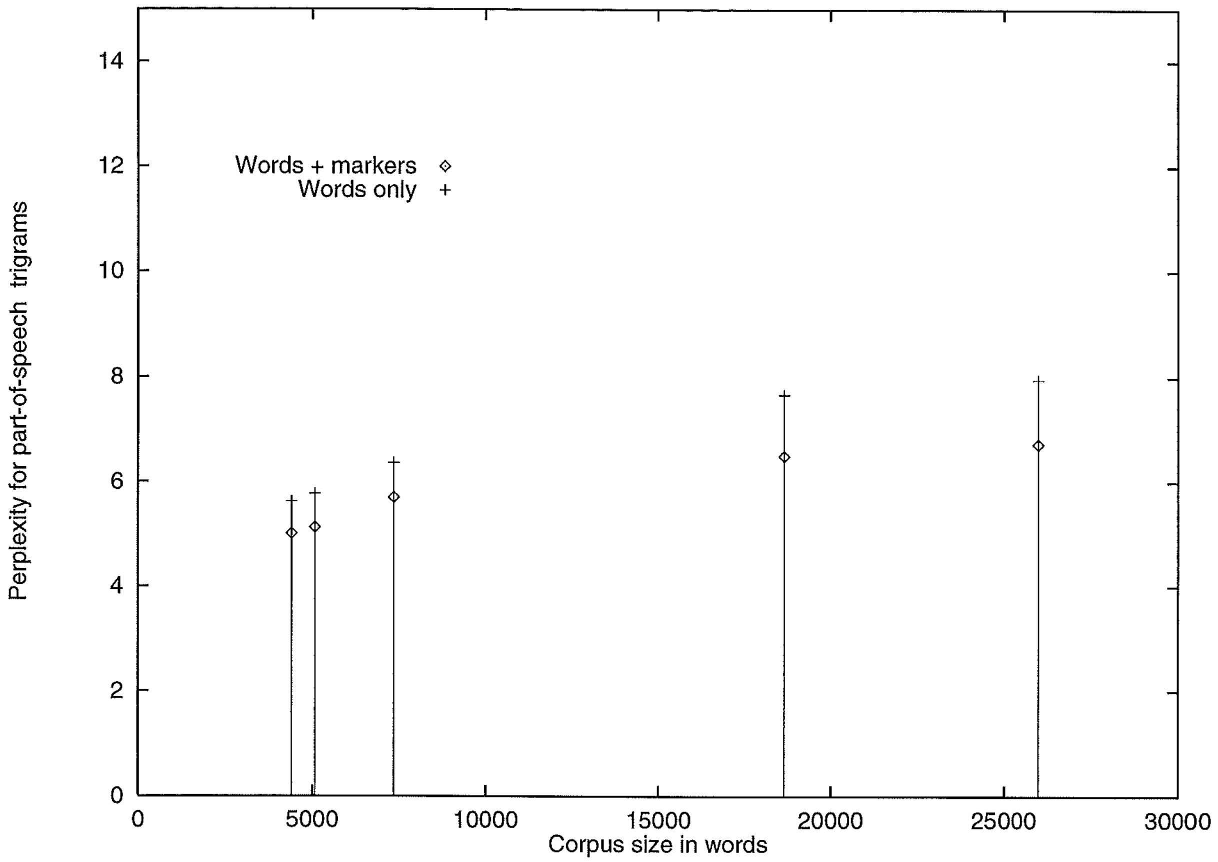

Figure 1: Comparison of trigram part-of-speech perplexity for sections of the MARSEC corpus, (i) with both major and minor discontinuities marked (ii) without either. The tagset size is 28 with the discontinuities represented, 26 without them. The data is taken from Table 3 and converted to perplexity from entropy. The entropy is calculated with trigrams spanning a major discontinuity (pause) omitted, as in Table 3 column H_3 (2).

4.5 Investigation of entropy measures

We can measure the entropy H_0, H_1, H_2 and H_3 for the corpus with and without prosodic markers for major and minor discontinuities. However, rather than use words themselves we map them onto part-of-speech tags. This reduces an indefinite number of words to a limited number of tags, and makes the investigation computationally feasible. The tagset used in this work is given in Appendix 2. We expect

- H_0 will be higher with markers, since the alphabet size increases.
- H_1 could be lower or higher depending on the frequency of the new symbols.
- H_2 and H_3 will be inclined to fall if the marker captures some of the language structure.

To conduct this investigation the MARSEC corpus was taken off the web, and pre-processed to leave the words plus major and minor tone unit boundaries. Then it was automatically tagged, using a version of the Claws tagger[4]. These tags were mapped onto a smaller tagset

[4] Claws4, supplied by the University of Lancaster, described by [20]

Speech representation	Number of minor discontinuities	Number of major discontinuities	H_0	H_1	H_2	H_3 (1)	H_3 (2)
Words only	0	0	4.70	4.11	3.29	2.94	2.94
Words + minor	3454	0	4.75	4.09	3.18	2.84	2.84
Words + major	0	1029	4.75	4.19	3.32	2.94	2.84
Words + both	3454	1029	4.81	4.17	3.16	2.82	2.70

Table 3: Entropy measures for 18655 words of the MARSEC corpus, (sections A, B, C concatenated) with and without major and minor discontinuities. H_3 (2) measures entropy without triples spanning a major pause (see text).

Speech representation	Number of minor discontinuities	Number of major discontinuities	H_0	H_1	H_2	H_3
Words + discontinuities in arbitrary positions	3109	1209	4.81	4.19	3.63	3.05

Table 4: Entropy measures for same part of MARSEC corpus with discontinuities in arbitrary positions : major discontinuity every 19 words, minor discontinuity every 7 words (except for clashes with major)

with 26 classes, 28 including the major and minor discontinuities. Random inspection indicated about 96% words correctly tagged. Then the entropy of part of the corpus was calculated (i) for words only (ii) with minor discontinuities represented (iii) with major discontinuities, pauses, represented and (iv) with major and minor discontinuities represented. Results are shown in Table 3, and in Figure 1. H_3 is calculated in two different ways. First, the sequence of tags is taken as an uninterrupted string (column H_3 (1) in Table 3). Secondly, we take the major discontinuities, pauses, as points of segmentation, and omit any triple that spans 2 segments (column H_3 (2)).

An alternative method of representation would be to have one symbol for a discontinuity, and mark a major discontinuity by a double occurrence. This would avoid the triple spanning a pause. However, it would exaggerate a decline in entropy for pairs and triples, since the discontinuity marker would often be followed by itself. The introduction of a frequently occurring pair would affect the entropy measure.

The formula used to calculate entropy is given in Appendix 1.

This experiment shows that the entropy declines when information on discontinuities is explicitly represented. Though there is not a transparent mapping from prosody to structure, there is a relationship between them which can be exploited. These experiments indicate that English language can be coded more efficiently when structural markers are represented.

Note that we are interested in *comparative* entropies. The entropy converges slowly to its asymptotic value as the size of the corpora increases, and this is an upper bound on entropy values for smaller corpora. Ignoring this may give misleading results [21]. The reason why entropy may be underestimated for small corpora comes from the fact that we approximate probabilities by frequency counts, and for small corpora these may be poor approximations.

Comparison with arbitrary segmentation

Compare these results to those of another experiment where the corpora of words only were taken and discontinuities inserted in an arbitrary manner. Major discontinuities were inserted every 19 words, minor ones every 7 words, except where there is a clash with a major one. The numbers of major and minor discontinuities are comparable to those in the real data. Results are shown in Table 4. H_2 and H_3 are higher than the comparable entropy levels for speech with discontinuities inserted as they were actually spoken.

Moreover, the entropy levels are higher than for speech without any discontinuities: the arbitrary insertion has disrupted the underlying structure, and raised the unpredictability of the sequence.

5 Conclusion

In this paper we have examined two mechanisms by which language is encoded in such a way that it can be decoded as easily as possible. We first saw how a zipfian distribution of word frequencies was compatible with this, and was therefore likely to evolve (Section 3).

Secondly, we examined different representations of English speech and saw that it can be more efficiently coded when discontinuities such as pauses are represented (Section 4. They act as structural markers. Unstructured strings of words are associated with higher levels of entropy, which makes decoding harder. So as language has evolved, we would expect selection pressure to encourage the development of structured modes of representation. Once an embryonic language structure began to emerge this may have been reinforced by the real world conditions of speech transmission.

It is hard to resist the temptation to conjecture about how initial moves towards structured language may have begun. Consider, for instance, a hunting scenario in the distant past. Suppose the leader of one tribe can say things like "Hide in the tree where I waited with your father while I chase the deer past you down the gully". Compare this to the utterance of the leader of another tribe that can only string words together in an unstructured way: "Hide tree wait father chase deer gully". If the aim of speech is to transfer information, then an advantage is conferred by the ability to use structured language rather than unstructured strings of words.

Research into the acquisiton of syntactic knowledge by children investigates the mapping from prosody to language structure [4]. There is a strong case that there are prosodic cues to syntactic structure, that infants are sensitive to these cues and can exploit them in speech processing. Whether this provides a base on which a grammar is constructed or a means by which innate grammar is discovered is a subject of ongoing research.

The emergence of structured language and statistical realities

One way of analysing the structure of language is to see it as a tertiary form. First, there are relationships between adjacent words, a structure that can be modelled by Markov processes. Then words can be grouped together into constituents and these constituents are organized in a secondary structure. Thirdly, there are relationships between elements of constituents, such as the agreement between the head of a subject and the main verb. These three levels are compatible with levels in the Chomsky hierarchy.

In this paper we consider the development of secondary structure, the organization of groups of words into constituents. Speech is of necessity a signal interspersed with periodic pauses for breath, and these pauses can be utilised to impose structure on a stream of speech. Other discontinuities, minor tone unit boundaries, also occur. By comparing the entropy of the sequence of spoken words with and without discontinuities represented we find that the entropy is reduced by including them. The representation of discontinuities aids the efficient decoding of speech.

Thus, it seems, that by developing structured modes of language representation, advantages of efficient transmission are conferred.

References

[1] B Mandelbrot. An informational theory of the statistical structure of language. In *Symposium on Applications of Communication Theory.* Butterworth, 1952.

[2] C Lyon and R Frank. Using Single Layer Networks for Discrete, Sequential Data: an Example from Natural Language Processing. *Neural Computing Applications,* 5 (4), 1997.

[3] P Lieberman. On the evolution of human language. In J A Hawkins and M Gell-Mann, editors, *The Evolution of Human Language.* 1992.

[4] J Morgan and K Demuth. *Signal to Syntax.* Lawrence Erlbaum, 1996.

[5] S. Johansson and K. Hofland. *Frequency analysis of English vocabulary and grammar.* Clarendon, 1989.

[6] C E Shannon. Prediction and Entropy of Printed English. *Bell System Technical Journal,* pages 50–64, 1951.

[7] T Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics,* pages 61–73, 1993.

[8] J Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language,* 1992.

[9] G K Zipf. *Human Behaviour and the Principle of Least Effort.* Addison Wesley, 1949.

[10] T C Bell, J G Cleary, and I H Witten. *Text Compression.* Prentice Hall, 1990.

[11] T M Cover and J A Thomas. *Elements of Information Theory.* John Wiley and Sons Inc., 1991.

[12] E Charniak. *Statistical Language Learning.* MIT Press, 1993.

[13] F Jelinek. Self-organized language modeling for speech recognition. In A Waibel and K F Lee, editors, *Readings in Speech Recognition,* pages 450–503. Morgan Kaufmann, 1990. IBM T.J.Watson Research Centre.

[14] S Gull and J Skilling. Recent developments at cambridge. In C Ray Smith and Gary Erickson, editors, *Maximum -Entropy and Bayesian Spectral Analysis and Estimation Problems,* 1987.

[15] C Lyon and S Brown. Evaluating Parsing Schemes with Entropy Indicators. In *MOL5, 5th Meeting on the Mathematics of Language,* 1997.

[16] M Ostendorf and N Vielleux. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics,* 20(1), 1994.

[17] Alex Chengyu Fang and Mark Huckvale. Synchronising syntax with speech signals. In V.Hazan, M.Holland, and S.Rosen, editors, *Speech, Hearing and Language.* University College London, 1996.

[18] P Taylor and A Black. Assigning phrase breaks from part-of-speech sequences. 1998.

[19] S Arnfield. *Prosody and Syntax in Corpus Based Analysis of Spoken English.* PhD thesis, University of Leeds, 1994.

[20] R Garside. The CLAWS word-tagging system. In R Garside, G Leech, and G Sampson, editors, *The Computational Analysis of English: a corpus based approach*, pages 30–41. Longman, 1987.

[21] M Farach and M Noordewier et al. On the entropy of dna. In *Symposium on Discrete Algorithms*, 1995.

Appendix 1

Formula for calculating the entropy of discrete, sequential data

The following formula is based on Shannon's work [6], on the entropy of letter sequences. He produced a series of approximations to the entropy H of written English, which successively take more of the statistics of the language into account

H_0 represents the average number of bits required to determine a letter with no statistical information. H_1 is calculated with information on single letter frequencies; H_2 uses information on the probability of 2 letters occurring together; H_n, called the n-gram entropy, measures the amount of entropy with information extending over n adjacent letters of text.[5] As n increases from 0 to 3, the n-gram entropy declines: the degree of predictability is increased as information from more adjacent letters is taken into account. If $n-1$ letters are known, H_n is the conditional entropy of the next letter, and is defined as follows.

b_i is a block of $n-1$ letters, j is an arbitrary letter following b_i

$p(b_i, j)$ is the probability of the n-gram b_i, j

$p_{b_i}(j)$ is the conditional probability of letter j after block b_i, that is $p(b_i, j) \div p(b_i)$

$$\begin{aligned}
H_n &= -\sum_{i,j} p(b_i, j) * log_2 p_{b_i}(j) \\
&= -\sum_{i,j} p(b_i, j) * log_2 p(b_i, j) + \sum_{i,j} p(b_i, j) * log_2 p(b_i) \\
&= -\sum_{i,j} p(b_i, j) * log_2 p(b_i, j) + \sum_{i} p(b_i) * log_2 p(b_i)
\end{aligned}$$

since $\sum_{i,j} p(b_i, j) = \sum_i p(b_i)$
An account of this process can also be found in [11].

[5] This notation is derived from that used by Shannon. It differs from that used by Bell, Cleary and Witten [10].

Appendix 2

Description of the Tagset

The tagset used in these experiments is derived from CLAWS4, mapped onto a smaller set of classes. They are as follows

- article or determiner - singular
- article or determiner - plural
- predeterminer e.g. "all"
- pronomial determiner e.g. "some"
- pronomial determiner - singular
- proper noun
- noun - singular
- noun - plural
- pronoun - singular
- pronoun - plural
- relative pronoun
- possessive pronoun
- verb - singular
- verb - plural
- auxiliary verb - singular
- auxiliary verb - plural
- existential "here" or "there"
- present participle
- past participle
- infinitive "to"
- preposition
- conjunction
- adjective
- singular number "one"
- adverb
- exceptions

The tagging process includes the identification of common phrases or idioms, which are then treated as single lexical items. For instance, "of course" is tagged as an adverb.