

# The segmentation of speech and its implications for the emergence of language structure<sup>\*</sup>

Caroline Lyon, Bob Dickerson and Chrystopher L. Nehaniv  
University of Hertfordshire, U. K.

This paper reports a phenomenon supporting the hypothesis that the emergence of structure in the evolution of language was a staged process. To develop a grammatical structure it seems necessary to first have discrete constituents which can be the building blocks of a hierarchical system. By analysing observed speech we show that the development of a linear sequence of grammatical constituents has its own advantage, before a possible next stage when constituents are integrated into a hierarchical structure.

A stream of speech sounds has to be segmented to allow for breathing. This segmentation has further developed in a certain way that makes it easier for the hearer to decode than if it were not segmented, or if it were segmented in an arbitrary manner. Well known tools from Information Theory are employed to analyse the ease of decoding speech. Segmentation depends on prosodic discontinuities, such as pauses and intonation marked by tone unit boundaries. These discontinuities usually mark groups of words with some syntactic cohesion, such as phrases and clauses. We show that in a modern corpus of spoken language observed segmentation facilitates the effective transfer of information, while lack of segmentation or arbitrary segmentation imposed on a stream of words makes decoding less efficient. This supports the hypothesis that the necessary constituents of a grammatical structure may have evolved as a consequence of developments favouring more efficient decoding of a linear stream of spoken words.

The source material for this investigation is taken from the prosodically marked up Machine Readable Spoken English Corpus (MARSEC).

---

\* We are grateful to three anonymous reviewers for their helpful suggestions.

## 1. Introduction

This paper reports on a phenomenon suggesting that there is an intermediate stage in the development of structured language with its own distinct evolutionary advantage, before the improvements of the final system are available. It supports the hypothesis that the emergence of syntactic structure could have been a staged process.

A recurring question in the literature on the evolution of language is how and why intermediate forms, if any, could have developed. Our work looks at one aspect of this issue, showing what an intermediate stage could be like, and how its development confers an advantage in the form of improved efficiency in verbal communication.

Among those who have addressed this issue are Bickerton (1995, 1998) and Pinker (1990, 1994). The possibility that there were no intermediate stages is argued by Bickerton in “Catastrophic evolution: the case for a single step from protolanguage to full human language” (Bickerton 1998). He proposes that the computational faculties needed to produce language developed for other purposes, e.g. for sophisticated social intelligence in ancestors of humans, living in a complex social environment (Humphrey, 1976; Byrne & Whiten, 1988; Whiten & Byrne, 1997). This was important in the evolution of language, as argued also, e.g. by Dunbar (1996). In particular, Bickerton proposes that such capacities support a ‘protolanguage’ (unstructured short sequences of meaningful speech sounds, without syntax), and that protolinguistic functional ‘brain units’ were linked in a single evolutionary step with the primate sensitivity to social relations and, most crucially, to thematic relations, thereby resulting in full-blown human syntax. On this view, only a single ‘catastrophic’ event was needed to exapt existing neural processes, resulting in an ability to produce and understand syntactically structured language.

In contrast, others in the field have argued for gradualism in the evolution of language, identifying evolutionary precursors of various competencies (e.g., Pinker and Bloom, 1990; Hauser 1996; Lieberman 2000). Pinker and Bloom argue for an adaptationist account, that no explanation is plausible for the development of a trait with complex design features other than gradual natural selection. They say (page 721) “There must have been a series of steps leading from no language at all to language as we now find it ... Every detail of grammatical competence that we wish to ascribe to selection must have conferred a reproductive advantage on its speakers....[but] There are no conclusive data on any of these issues”.

Other work in this field salient to a Darwinian explanation of segmentation includes Hauser et al.'s (2001) experiments with some non-human primates (cotton-top tamarins). This indicates that they have an ability to respond differentially to sequential phonetic stimuli, identifying word boundaries in a stream of speech sounds, a rudimentary form of the kind of segmentation that occurs in human speech perception. This suggests that a common ancestor of humans and tamarins may have had the ability to segment streams of vocalizations according to statistical regularities.

An important contribution has been made by Lieberman (1992, 2000) whose work over many years illuminates the evolutionary development of the mechanisms of speech production. Though he is at opposite poles to Bickerton, they share the well-founded belief that some neural mechanisms necessary for language also served other functions. Lieberman shows how primitive neural structures are needed in conjunction with higher cognitive processors to produce a 'functional language system', though the move to syntactic competence needs further elaboration.

Lieberman's earlier work (1992) reinforces our approach, which is based on the assumption that more efficient methods of communication would be a driving force in the evolution of structured language. A general evolutionary argument is that the value of mechanisms exapted or adapted to support language can outweigh concomitant disadvantages. Lieberman shows that effective mechanisms for speech production evolved in spite of a physiological cost (described later).

## 2. Context of the investigation

One way of analysing the present structure of language is to see it as a tertiary form. First, there are relationships between adjacent words, which can, to some extent, be modelled by Markov processes, as in speech recognition applications. Secondly, words can be grouped together into constituents and these constituents organised in a hierarchy. 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 analogous to levels in the Chomsky hierarchy: regular grammars, context free grammars and context sensitive grammars.

Though the simplest regular grammars have proved extremely useful in practical engineering applications, particularly in speech recognition, they cannot describe much of language structure well. To do this it is necessary to take a model

that handles constituents, groups of words or phrases, and that can represent their relationships and relationships between elements of different constituents.

Now, for this to be done a stream of words must be segmented into appropriate discrete segments. This is an essential intermediate step, and we suggest that this segmentation process could have its own evolutionary advantage. Whether this evolutionary advantage affected the biological basis of language readiness, or only the structure of early communication systems, is left open. The argument here could be employed as part of either evolutionary scenario, or both.

Speech is of necessity a signal stream, interspersed with periodic pauses for breath, so some sort of segmentation is inevitable. There are also other observed discontinuities, such as intonation marking tone unit boundaries. We show that in a modern corpus of spoken language observed segmentation facilitates the effective transfer of information, while lack of segmentation or arbitrary segmentation imposed on a stream of words makes decoding less efficient. The observed segmentation is usually found only when words are grouped into cohesive syntactic units.

Our work makes a novel application of well known tools in Information Theory, which can illuminate characteristics of natural language, and illustrates how such tools are useful aids in linguistic analysis.

## 2.1 The empirical approach

This work employs an empirical approach to the analysis of language structure, based on an investigation into observed characteristics of speech in the MARSEC corpus (described below). We look at the actual segmentation of spoken language, and see what effect it has on the efficiency of speech as a medium for exchanging information. Thus we start by following Wittgenstein's advice on acquiring knowledge about language "Don't think, but look!" (Wittgenstein, 1953)

Our approach contrasts with a theoretical analysis of a given grammar, which takes an internal, cognitivist position, as opposed to deducing characteristics of language from observed speech. However, these two approaches can be seen as complementary.

As is well known, for decades there has been a divergence between empirical, behaviourist traditions and the rationalist methods particularly associated with Chomsky. But now Chomsky proposes that there is no coherent way to formulate some of the matters at issue, and that they should be set aside; what is needed is unification (Chomsky, 2000, pages vi–xvi). He says:

Currently the best theory is that the initial state of the language faculty incorporates certain general principles of language structure ... the mature state of competence is a generative procedure that interacts with the motor and perceptual system ... A vast range of empirical evidence is relevant in principle (ibid, page 60).

This empirical evidence must include knowledge of neurological bases of human language and of how pre-linguistic capabilities are exploited in the acquisition of language skills. Similarly, external factors such as the statistical realities of speech production and perception can inform language development. Such observable factors can affect the development of inner cognitive processes both in the lifetime of the species and in the lifetime of the individual.

## 2.2 Use of English

Our investigation is constrained by the availability of suitable corpora. There was no reason to do this work in English apart from the fact that the MARSEC corpus was available; we have not found similarly marked up data in other languages.

For the same reason, we are restricted in the type of language we can analyse. Out of the MARSEC corpus we have taken unscripted news reports and lectures, as being the nearest available to spontaneous speech. We omitted poetry and religious readings.

## 2.3 Assumptions in our work

Language structure is a complex and wide ranging area of study. In this work we focus on the issue of finding constituents to act as building blocks in a hierarchical grammatical structure.

We investigate speech as a medium for communicating information. This means we do not consider its use for conveying emotions (Haiman, 1999), nor for performing speech acts (Austen, 1962). If I say "Hello" to a friend I am performing the act of greeting as well as passing on information on my attitude.

We assume that a mapping from a speech signal onto discrete symbols can be done. We leave aside issues concerning the perceptual analysis of speech, and how an utterance is segmented into words (Morgan and Demuth, 1996).

### 3. Related work on prosody and syntax

In this paper we investigate how some aspects of prosody might give rise to syntax. A significant body of work has been concerned with the converse: how syntax of written text can be used to determine appropriate boundaries in synthesised speech. Work in this field includes that done by Arnfield (1994), Ostendorf and Vielleux (1994), Fang and Huckvale (1996), Taylor and Black (1998). For instance, Fang and Huckvale investigated the correspondence of required pause locations with grammatical categories (the sentence, the clause, and the phrase). They describe how the location of most major and minor pauses could be explained on syntactic grounds alone. They also comment that their results are surprisingly similar to those of Crystal (1969) whose work was mainly based on transcribed spontaneous speech. Commercial speech synthesisers use this information to produce natural sounding speech, quite different in quality to the robotic talk of early systems.

The fact that syntactic structure in text maps onto prosodic structure in speech does not imply the converse: that perceptual units are necessarily phrases and clauses. In informal conversation word segment boundaries may correspond to hesitation, part completed utterances and so on. However, by restricting ourselves to unscripted speech of professional speakers, and to some scripted material, we expect to minimise these other factors.

Morgan and Demuth (1996) present evidence that prosody along with other types of probabilistic information indicate syntactic structure. They have investigated the acquisition of syntactic knowledge by children, concluding that there are prosodic clues to syntactic structure, that even infants are sensitive to these clues, and can exploit them in processing speech .

### 4. Selection for efficient and robust communication

There is a biological cost in developing the physiology capable of producing speech. Humans can produce a much wider range of sounds than other primates can, but in order to do this the human anatomy has evolved in a way that has incurred physiological disadvantages. As Darwin first noted, in humans food has to cross the airway to the lungs, with a consequent risk of choking that is absent in other primates. The significance of this is a subject of debate, but the human speech faculty has developed despite concomitant disadvantages.

It is instructive to examine the characteristics of human speech that distinguish it from non-speech sounds. First, as Lieberman (1992) notes, the high transmission rates that characterise speech: 15 or even up to 25 phonetic segments can be produced or recognised per second. The identification of non-speech sounds is much slower: a maximum of 7 to 9 items per second, typically less. Secondly, he notes the larger range of sounds that only humans among primates 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 apparent selection for speed, reliability and range as speech has evolved. In the same way, we expect that other environmental factors that can contribute to efficient communication will be exploited too. This paper examines the statistical environment in which speech operates, and shows that there is a certain inevitability underlying steps towards the development of segmented speech, as a possible forerunner of syntactic structure.

## 5. Measuring the effective transfer of information

Back in 1952 Mandelbrot (1952) applied an information theoretic approach to the analysis of language. 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”. While ignoring the teleological debate about intentionality in speech production, in essence we adopt this approach, proposing that speech has developed for speed, scope, ease of production (encoding) and ease of perception (decoding).

Unfortunately, Mandelbrot’s work in this field was flawed in a number of respects, and he did not endear himself to the traditional linguistic community by claiming ill-advised analogies with thermodynamics. His approach was misunderstood and his work quietly left aside.

The particular issue that we address here is ease of decoding. We have carried out experiments to see whether speech is easier to decode if it is segmented as observed, segmented arbitrarily, or not segmented at all. In order to evaluate ease of decoding we employ entropy metrics, and compare the entropy of sequences of spoken words when different methods of segmentation are employed.

In this work there is no suggestion that humans consciously apply the type of analysis described below - any more than a man usually estimates the size of a distant object by consciously employing laws of optics.

### 5.1 Brief introduction to entropy metrics

Entropy metrics can be applied to sequences of discrete symbols in a message, for instance letters or words. Entropy measures, in a certain sense, the degree of unpredictability of symbols in a sequence. If the entropy can be reduced, the predictability of the next element in an incomplete sequence is increased, and the easier it is for the receiver to decode the message. A sequence represented in a way that lowers the entropy, without reducing its representational or expressive 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 has been used in automated speech recognition for many years (Jelinek,1990). The use of entropy indicators has also been used to support choices of language models for other natural language processing tasks, e.g. Lyon and Brown (1997), Lyon and Frank (1997).

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 speech recognition applications language models generally use trigrams (three consecutive words) instead of single words in order to produce a list of ranked candidate words to integrate with the acoustic recogniser.

The contribution we make here is to show that the entropy can also be reduced by representing the segmentation of 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.

## 6. Description of entropy

The concepts described here were first introduced by Shannon (1948) in his seminal paper "A Mathematical Theory of Communication". A standard reference is Cover and Thomas (1991); for an introductory text see Charniak (1993).

## 6.1 Informal description

Entropy metrics can be applied to any sequence of discrete symbols. In order to review and explain the concepts involved, we first look at a simple example where the symbols are letters. We examine sequences of letters, and how they are put together to create words. The symbol for entropy is  $H$ , and we can calculate the level of  $H$  in different situations. This will indicate how predictable a letter in a sequence will be, how easy it will be to decode the sequence:

- if there is no information on letter probabilities, entropy  $H_0$
- if probabilities of single letter are known, entropy  $H_1$
- if probability of letters, given that the one before is known, entropy  $H_2$
- if probability of letters, given that the two before are known, entropy  $H_3$

With more contextual information predictability increases, ease of decoding increases while entropy declines. Entropy is calculated by combining two factors for each symbol (letters in this example) and adding up the results for all of them. These two factors, very informally, are “how often does this symbol occur” and “how hard is it to predict?” (See Appendix A.)

## 6.2 Mathematical description and illustration

In mathematical terms let  $A$  be an alphabet and  $X$  be a discrete random variable taking values in  $A$ . The probability mass function is then  $p(x)$ , i.e. the probability that  $X$  takes symbol  $x$  as its value, where  $x \in A$ .

$$p(x) = \text{probability}(X=x)$$

In this example the  $x$ 's are the letters of the alphabet. The entropy is defined as

$$H(x) = - \sum p(x) * \log_2(p(x)) \text{ over } x \in A.$$

The log term, very informally, is a measure of the difficulty of predicting the symbol. Since  $p(x) \leq 1$  it will have a negative value, cancelled by the negative sign at the start of the formula. The trivial example below should throw light on this.

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.

### 6.3 Trivial example

Suppose we have an alphabet of just 4 letters,  $A, B, C, D$  with no information on probabilities. Then

$$p(A) = p(B) = p(C) = p(D) = \frac{1}{4}$$

Informally, this is the factor measuring “how often does this symbol occur?”. The second factor is “how difficult is it to predict?”. For each symbol this is  $\left(-\log_2 \frac{1}{4}\right)$

So

$$H_0 = -4 * \left(\frac{1}{4} * \log_2 \frac{1}{4}\right)$$

$$H_0 = 4 * \frac{1}{4} * \log_2(4) = 2$$

Now suppose the probabilities of single letters occurring are:

$$p(A) = \frac{1}{2} \quad p(B) = \frac{1}{4} \quad p(C, D) = \frac{1}{8}$$

$$H_1 = \left(\frac{1}{2} \log 2\right) + \left(\frac{1}{4} \log 4\right) + 2 * \left(\frac{1}{8} \log 8\right)$$

$$H_1 = 1 \frac{3}{4}$$

Entropy declines as we have more statistical information.

### 6.4 Perplexity

Often the related metric of *perplexity* is employed, particularly in speech recognition. 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. For instance, in the trivial example above when we have no information on probabilities there are 4 equally likely choices and

$$P = 2^2 = 4$$

## 7. Shannon's work on English texts

Shannon, the founder of Information Theory, investigated the entropy and prediction of letters in written English (Shannon, 1951). He 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. The  $n$ -gram entropy,  $H_n$ , measures the amount of entropy with information extending over  $n$  adjacent letters of text, and  $H_n \leq H_{n-1}$ . This fact is exploited in games where the contestants have to guess letters in words, such as the "Shannon game" or "Hangman" (Shannon, 1951).

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. 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 symbol will be much more frequent than any other symbol, and this can outweigh the effect of the larger number of symbols. 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  $H_3$  decline. (See Table 1.)

For instance, for the words

COOKING CHOCOLATE

the trigrams "N-G-C" and "G-C-H" will be replaced by "N-G-space", "G-space-C" and "space-C-H" by including the space as a new symbol.

## 8. The entropy of strings of words

Now, a similar analysis can be employed to see how words are organised into structured constituents. Lyon and Brown (1997) have shown how the entropy of text mapped onto part-of-speech tags could be reduced if clauses and phrases

**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.

	$H_0$	$H_1$	$H_2$	$H_3$
Alphabet with 26 letters	4.70	4.14	3.56	3.30
Alphabet with 27 "letters"	4.76	4.03	3.32	3.10

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. Consider a sentence without opening or closing clause boundary markers, like

The shirt he wants is in the wash  
*det noun pronoun verb verb prep det noun*

If this sentence is given part-of-speech tags we can include “virtual tags” to represent the clause boundaries such as symbols “{”, virtual-tag1, and “}”, virtual-tag2, to give:

The shirt he wants is in the wash  
*det noun { pronoun verb } verb prep det noun*

The part-of-speech tags have probabilistic relationships with the virtual tags in the same way that they do with regular tags. The pairs and triples generated by this second string exclude “unlikely” tag sequences, such as (*noun, pronoun*), (*noun, pronoun, verb*) but include for instance (*noun, virtual-tag1, pronoun*)

If we add virtual tags to the tag set then the entropy  $H_0$  will be higher 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 entropy,  $H_2$  and  $H_3$  with virtual tags explicitly marking some constituents, is lower than that without the virtual tags, because certain combinations will not occur.

## 9. Analysis of MARSEC (Machine Readable Spoken English Corpus)

Entropy metrics can be applied to a sequence of words from a speech signal in a similar way to their application to a sequence of letters. A stream of spoken words can be segmented by periodic discontinuities, and we have investigated how the entropy of sequences of words varies with and without representing these discontinuities.

This research was carried out using the MARSEC corpus, organised by Arnfield (1994), which is annotated with prosodic markers. There is an address for the MARSEC web site in the references. The corpus, collected by the BBC,

includes unscripted news commentary, scripted news and lectures, totalling about 26,000 words. This is likely to be “well formed” language, not typical of, for instance, informal conversations. However, as we are examining the use of speech to communicate information, this is an appropriate source of data. Prosody can mark phenomena in speech which we are not considering here — for instance emotional factors. By restricting ourselves to this broadcast material, scripted and unscripted, we focus on the use of speech to convey information.

MARSEC is annotated with a number of prosodic features. We use just the major and minor tone unit boundaries; the major tone unit boundary is a pause. The term “discontinuity” is taken to cover both these features. The prosodic mark up has been done by two trained annotators, who determine the tone unit boundaries. An example is shown in Table 2.

Some sections of the corpus have been marked up by both annotators for consistency checking. Inter-annotator agreement is 94%, if we assume that a major and minor pause differ. If we take the insertion of a major pause by one at the same position as a minor pause by the other as agreement, then inter-annotator agreement is 95%.

**Table 2.** Example of data from the MARSEC corpus. | marks a minor discontinuity, || marks a major one. The two annotators usually agree (see text), but this illustrates a point of disagreement.

annotator 1	annotator 2
we	we
heard	heard
automatic	automatic
fire	fire
a	a
few	few
yards	yards
away	away
we	we
drove	drove
on	on
a	a
jet	jet
appeared	appeared

The issue of providing benchmarks for prosodic annotation presents problems. Automated annotation has not been very successful, partly because it is difficult to interpret acoustic representations of utterances in terms of human auditory perception. Relative differences in acoustic patterns and subtle contextual information are important (Price and Ostendorf, 1996, page 80).

Manually marked corpora, such as MARSEC, are taken as benchmarks for automated prosodic marking. However here we meet a possibly confounding factor. Perceptually based transcriptions of prosodic signals could reflect the annotator's syntactic expectations as well as the acoustic signal. Feedback between prosody and syntax in perception (and hence in annotation) makes it difficult to ascertain to what extent annotated texts faithfully represent phonetic structure of human speech.

## 10. Mapping words onto part-of-speech tags

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. This also mitigates the problem of having a comparatively small corpus: word frequencies would not reflect the frequencies of words in an indefinitely large corpus of actual usage. However, tag frequencies are similar to those in much larger corpora.

There is evidence that humans process acoustic speech signals and in some circumstances map them onto basic part-of-speech categories before lexical access (Morgan, Shi and Allopenna, 1996). Though our decision to use part-of-speech tags was based on practical considerations, this suggests that our method is not unreasonable.

To conduct the investigation the MARSEC corpus was pre-processed to leave the words plus major and minor tone unit boundaries, omitting other prosodic information. Then it was automatically tagged, using a version of the CLAWS tagger (supplied by the University of Lancaster, described by Garside (1987)). These tags were mapped onto a smaller tag set with 26 classes, 28 including the major and minor discontinuities. The tag list is in Appendix B. Random inspection indicated about 96% of words correctly tagged, which accords with other evaluations of this tagger.

## 11. Investigation of entropy measures

We can measure the entropy  $H_0$ ,  $H_1$ ,  $H_2$ , and  $H_3$  for the MARSEC corpus with and without the prosodic markers for discontinuities. The formulae used to calculate the entropies are given in Appendix A. We expect

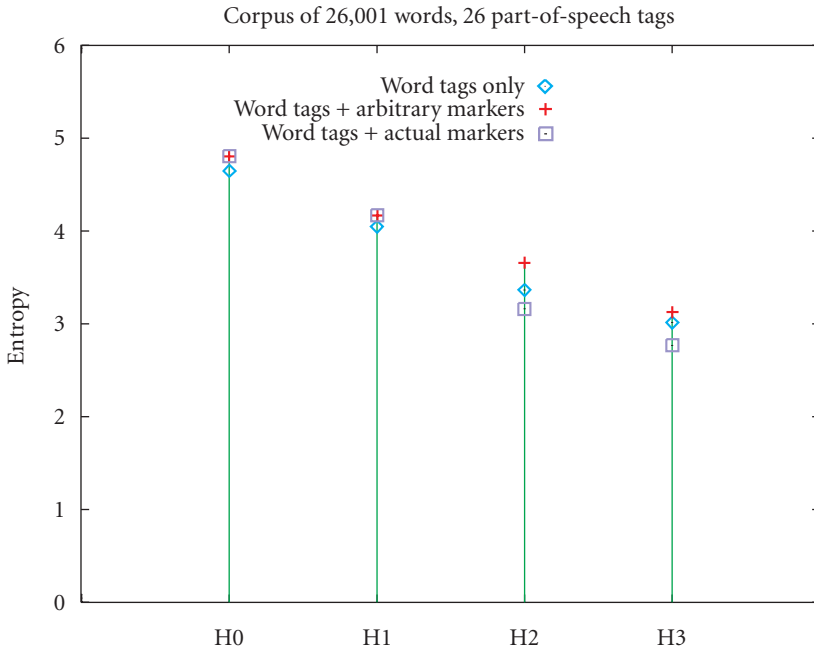
- $H_0$  will be higher with markers than without since the alphabet size increases, there are more symbols to choose from.
- $H_1$  could be lower or higher, depending on the frequency of the new symbols.
- $H_2$  and  $H_3$  should fall if the markers capture some of the structure of the language.

The entropy of part of the corpus was calculated (i) for tagged words only (ii) with minor discontinuities represented, but without major ones (iii) with major discontinuities, pauses, represented but without minor discontinuities, and (iv) with major and minor discontinuities both represented. Results are shown in Table 3, and in Figure 1. To calculate  $H_3$  with major discontinuities, we take the major discontinuity as equivalent to sentence ends, and omit any triple that spans two sentences.

Note that we are interested in *comparative* entropies. The entropy converges to its asymptotic value as the size of the corpus increases. Ignoring this may give misleading results (Farach and Noordewier, 1995). The reason why entropy may be underestimated for small corpora comes from the fact that we estimate probabilities by frequency count, and for small corpora these may be poor approximations. This is why we use part-of-speech tags rather than words themselves. Previous experiments showed that the entropy levels out as the corpus size reaches about 20,000 words (Lyon, 1998, Figure 1).

**Table 3.** Entropy measures for 26001 words of the MARSEC corpus, with and without discontinuities.  $H_3$  measures entropy without triples spanning a major discontinuity.

Speech representation	$H_0$	$H_1$	$H_2$	$H_3$
Tagged words only	4.70	4.10	3.31	2.99
Tagged words + minor discontinuities	4.75	4.09	3.19	2.89
Tagged words + major discontinuities	4.75	4.19	3.34	2.90
Tagged words + both	4.81	4.17	3.18	2.75



**Figure 1.** Entropy of tagged words from MARSEC corpus showing the difference between (i) tagged words alone (ii) tagged words with arbitrary segmentation markers inserted (iii) tagged words with actual prosodic segmentation markers. Derived from data in Tables 3 and 4.

### 11.1 Comparison with arbitrarily placed discontinuities

In another experiment we took the tagged words only and inserted discontinuity markers without regard to the observed placement. In order to have about the same number as those in the real data, major discontinuities were inserted every 19 words and minor ones every 7 words, except where there was a clash with a major one. We call this “arbitrary” placement. 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.

The experiments show that the trigram entropy declines when information on observed discontinuities is explicitly represented. This decline in entropy is associated with greater ease of decoding.

**Table 4.** Entropy measures for the same part of the MARSEC corpus with the same number of discontinuities, but inserted in arbitrary positions: a major discontinuity every 19 words, a minor discontinuity every 7 words (except for clashes with major ones).

Speech representation	$H_0$	$H_1$	$H_2$	$H_3$
Tagged words + discontinuities in arbitrary positions	4.81	4.19	3.64	3.12

## 12. Discussion

In this paper we have explored the idea that speech may have developed so that, among other factors, the hearer would understand as easily as possible. We have investigated the ease of decoding a stream of words by using the measurable concept of entropy, the degree of order or disorder.

We have shown that the segmentation of a linear stream of words as observed facilitates efficient decoding of speech and thus is likely to emerge. Informally speaking, it is easier to understand speech if it is divided up into the right sort of chunks. Experiments in the development of speech synthesisers showed that it is harder to understand a string of words without pauses or intonation, and we have offered a partial explanation of this.

The results of our experiments, given in Tables 3 and 4, illustrated in Figure 1, and described in Section 11, show that the entropy of a string of tagged words is less when observed discontinuities are included than when they are omitted. The entropy rises markedly when the string of tagged words is segmented by arbitrary discontinuity markers. This indicates that spoken English can be more easily decoded when discontinuities as observed are present.

From this standpoint, we then consider how this type of observed segmentation can be characterised. Work described in Section 3, on the relationship between prosody and syntax (Crystal, 1969; Arnfield, 1994; Ostendorf and Vielleux, 1994; Fang and Huckvale, 1996; Taylor and Black, 1998) shows that the stream of linear segments is usually composed of components, such as phrases and clauses, with some syntactic cohesion. On inspection, this is, as expected, also found in the annotated MARSEC corpus.

Now, these components are the building blocks of a hierarchical language structure. Syntactically cohesive elements such as phrases and clauses are the necessary raw material needed for the development of a full hierarchical language structure. Thus this type of segmentation can be seen as an intermediate stage in the development of a full syntactic system.

Segmentation by inspiration may have already existed when speech expressed some sort of protolanguage consisting of unstructured word strings. This segmentation could have been extended for the reasons advanced in this article, opening the way to differentiation of the segments between breaths into structured phrases. These might then have gradually become integrated into a syntactic hierarchy, with its greatly improved power of communication. The semantic meaning of single content words may be the same in protolanguage and full language, but communicative effect is greatly enhanced by the ability to relate words, for instance by using inflections, prepositions, time and place indicators, to produce meaningful phrases and clauses. A great step forward can then be taken by introducing relative pronouns enabling language to link concepts separated by time or space. If one purpose of speech is to convey information then a far reaching advantage is conferred by the ability to use structured language rather than unstructured strings of words.

This paper has described how a step on the way to developing a fully structured language could have evolved.

## References

- Arnfield, S. (1994) *Prosody and Syntax in Corpus Based Analysis of Spoken English*. PhD thesis, University of Leeds.
- Austin, J. (1962) *How to do things with words*. Oxford Clarendon Press.
- Bell, T. C., Cleary, J. G. and Witten, I. H. (1990) *Text compression*. Prentice Hall.
- Bickerton, D. (1995) *Language and Human Behavior*. University of Washington Press
- Bickerton, D. (1998) Catastrophic Evolution: the case for a single step from protolanguage to full human language. In Hurford, J.R., Studdert-Kennedy, M. and Knight, C. *Approaches to the Evolution of Language*. Cambridge University Press.
- Byrne, R.W., and Whiten, A. (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*. Oxford University Press.
- Charniak, E. (1993) *Statistical Language Learning*. MIT Press.
- Chomsky, N. (2000) *New Horizons in the Study of Language and Mind*. Cambridge University Press.
- Cover, T. and Thomas, J. A. (1991) *Elements of Information Theory*. John Wiley and Sons Inc.
- Crystal, D. (1969) *Prosodic systems and intonation in English*. Cambridge University Press.
- Dunbar, R. (1996) *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.
- Fang, A. C. and Huckvale, M. (1996) Synchronising syntax with speech signals. In Hazan, V. et al., editors, *Speech, Hearing and Language*. University College London.
- Farach, M., Noorderwier, M. et al. (1995) et al. On the Entropy of DNA. *Proceedings of the 6th Annual Symposium on Discrete Algorithms (SODA95)*. ACM Press.

- Garside, R. (1987) The CLAWS word-tagging system. In Garside, R., Leech, G. and Sampson, G. editors, *The Computational Analysis of English: a corpus based approach*, Longman.
- Haiman, J. (1999) From doing to saying. *Evolution of Communication*, 3(2): 185–205.
- Hauser, M. D. (1996) *The Evolution of Communication*, MIT Press.
- Hauser, M. D., Newport, E. L. and Aslin, R. N. (2001) Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition* 78: B53-B64.
- Humphrey, N. K. (1976). The social function of intellect. In Bateson P.P.G. and Hinde R. A., Eds. *Growing Points in Ethology*. Cambridge University Press, pp. 303–321.
- Jelinek, F. (1990) Self-organized language modeling for speech recognition. In Waibel, A. and Lee, K. F. editors, *Readings in Speech Recognition*, pages 450–503. Morgan Kaufmann, IBM T.J. Watson Research Centre.
- Lieberman, P. (1992) On the evolution of human language. In Hawkins, J. A. and Gell-Mann, M. editors, *The Evolution of Human Language*, pages 21–47, Addison-Wesley.
- Lieberman, P. (2000) *Human Language and our Reptilian Brain*. Harvard University Press.
- Lyon, C. (1998) *Language evolution: survival of the fittest in the statistical environment*. Technical report, number 322, Computer Science Department, University of Hertfordshire.
- Lyon, C. and Brown, S. (1997) Evaluating Parsing Schemes with Entropy Indicators. In *MOL5, 5th Meeting on the Mathematics of Language*.
- Lyon, C. and Frank, R. (1997) Using Single Layer Networks for Discrete, Sequential Data: an Example from Natural Language Processing. *Neural Computing Applications*, 5(4).
- Mandelbrot, B. (1952) An informational theory of the statistical structure of language. In *Symposium on Applications of Communication Theory*. Butterworth.
- MARSEC *Machine Readable Spoken English Corpus*. <http://www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec/>
- Morgan, J. and Demuth, K. (1996) Signal to syntax: an overview. In Morgan, J. and Demuth, K. editors, *Signal to Syntax*. Lawrence Erlbaum.
- Morgan, J., Shi, R., and Allopenna, P. (1996) Perceptual bases of rudimentary grammatical categories. In Morgan, J. and Demuth, K. editors, *Signal to Syntax*. Lawrence Erlbaum.
- Ostendorf, M. and Vielleux, N. (1994) A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20(1).
- Pinker, S. (1994) *The Language Instinct*. New York : Morrow; London: Penguin.
- Pinker, S. and Bloom, P. (1990) Natural Language and Natural Selection. *Behavioral and Brain Sciences*, 13: 707–784.
- Price, P. and Ostendorf, M. (1996) Combining linguistic with statistical methods in modelling prosody. In Morgan, J. and Demuth, K. editors, *Signal to Syntax*. Lawrence Erlbaum, 1996.
- Shannon, C. E., (1948) A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27 : 379–423, 623–656. Reprinted in Sloane, N. J. A., and Wyner, A. D., editors, (1993) *Shannon: Collected Papers*. IEEE Press.
- Shannon, C. E., (1951) Prediction and entropy of printed English. *Bell Sys. Tech. Journal*, 30: 50–64. Reprinted in Sloane, N. J. A., and Wyner, A. D., editors, (1993) *Shannon: Collected Papers*. IEEE Press

- Taylor, P. and Black, A. (1998) Assigning phrase breaks from part-of-speech sequences. *Proc. Eurospeech '97*.
- Whiten, A., and Byrne, R. W. (1997). *Machiavellian Intelligence II: Evaluations and Extensions*. Cambridge University Press.
- Wittgenstein, L. (1953) *Philosophical Investigations*. Translated by G. Anscombe, Blackwell Publishers.

## Appendix A

### *Derivation of the formula for calculating entropy*

This is derived from Shannon's work on the entropy of symbol sequences (Shannon, 1951). He produced a series of approximations to the entropy  $H$  of written English, taking letters as symbols, which successively take more account of the statistics of the language.

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

$b_i$  is a block of  $n-1$  symbols,  $j$  is an arbitrary symbol following  $b_i$   
 $p(b_i, j)$  is the probability of the  $n$ -gram consisting of  $b_i$  followed by  $j$   
 $p_{b_i}(j)$  is the conditional probability of symbol  $j$  after block  $b_i$ , that is  
 $p(b_i, j) \div p(b_i)$

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

$$\text{since } \sum_{i,j} p(b_i, j) = \sum_i p(b_i)$$

N. B. This notation is derived from that used by Shannon. It differs from, for instance, that used by Bell, Cleary and Witten (1990).

## Appendix B

### *Description of the Tag Set*

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

- article – singular e.g. “a”
- determiner – singular or plural “the”
- predeterminer e.g. “all”
- pronomial determiner e.g. “some”
- pronomial determiner – singular e.g. “this”
- 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

Two extra tag classes are added for the analysis of tone unit boundaries:

- minor discontinuity
- major discontinuity

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.

*Authors' address*

Computer Science Department  
University of Hertfordshire  
Hatfield, Herts, AL10 9AB  
UK  
c.m.lyon@herts.ac.uk

*Accepted:* June 13, 2003