# Exploiting statistical characteristics of word sequences for the efficient coding of speech

**Caroline Lyon** and **Bob Dickerson**
Computer Science Department, University of Hertfordshire, UK *

## Abstract

Characteristics of natural language can be illuminated through the application of well known tools in Information Theory. This paper shows how some of these characteristics can be exploited in the development of automated speech and language processing applications. The explicit representation of discontinuities in a temporal sequence of sounds, such as pauses in speech, can be utilized to improve the transmission of information. Arguments based on comparative entropy measures are used.

## 1 Introduction

Characteristics of natural language can be illuminated through the application of well known tools in Information Theory. This paper shows how some of these characteristics can be exploited in the development of automated speech and language processing applications. We investigate how discontinuities in a temporal sequence of sounds, such as pauses in speech, can be utilized to improve the transmission of information. The Machine Readable Spoken English Corpus (MARSEC), which is prosodically annotated, is used.

The approach taken is to examine certain observed phenomena in speech, and suggest how their exploitation could have conferred an advantage as human language evolved. Many years ago Mandelbrot 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" (Mandelbrot, 1952). By examining how language is produced for humans to decode, we expect to learn efficient ways to represent language for machine processing.

This work developed out of investigations into data representation for automated natural language parsing (Lyon and Frank, 1997; Lyon and Brown, 1997).

* email: C.M.Lyon@herts.ac.uk

## 2 Background: 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 (Lieberman, 1992)). 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 characterise speech: 15 to 25 phonetic elements per second 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 elements, 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, shows why structured language is likely to evolve, and uses this information to develop more efficient methods of representing speech for automated processing.

## 3 Sequence structure and efficient coding

We now investigate how words are grouped together, and why certain modes of segmentation are likely to evolve. We first describe the metrics that will be used, and then illustrate their application.

### 3.1 Entropy and perplexity

This analysis is based on comparative measures of the entropy of sequential data (Cover and Thomas, 1991). 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 same approach to the development of language models has been used in automated speech recognition for many years (Jelinek, 1990). 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.

The new contribution we make is to show that the entropy can also be reduced by modelling discontinuities along with words. The segmentation of a stream of words in this way is not arbitrary: the segments are related to structural components of language (Arnfield, 1994; Fang and Huckvale, 1996; Ostendorf and Vielleux, 1994).

#### 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

$$H(X) = -\sum_{x \in \mathcal{A}} p(x) * log_2\ p(x)$$

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 [1].

### 3.2 Illustrations from letter sequences

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, 1951). He showed that the entropy $H$ of written English, can be reduced as more of the statistics of the language are taken into account. He produced a series of approximations to the entropy $H$ of written English, which successively take more of the context 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 a high probability of occurring while $q$ had a low probability, then the letter $e$ could have a shorter code than $q$. 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 [2] 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. The formula for calculating the entropy of discrete, sequential data is given in (Lyon, 1999).

#### Entropy reduction and sequence 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, $H_2$ and $H_3$ do in fact decline. The space symbol prevents "irregular" letter sequences between words, and this is one way in which unpredictability is reduced.

### 3.3 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 (Gull and Skilling, 1987) report on an experiment with a string of 32,768 zeroes and ones that are

[1] In many practical applications the formula for perplexity is reduced to a special case based on the (questionable0 assumption that language is ergodic

[2] This notation is derived from that used by Shannon (Shannon, 1951). It differs from that used by Bell, Cleary and Witten (Bell et al., 1990).

| Key: || is a pause, \| is a minor discontinuity | |
|---|---|
| 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 |

Table 1: Example of MARSEC corpus with minimal prosodic annotations

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".

### 3.4 The entropy of strings of words

Now, a similar analysis can be employed to see how words are organised into structured constituents. In (Lyon and Brown, 1997) 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 verbal markers for the opening or closing of clauses 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 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 string exclude "unlikely" tag sequences such as (*noun, pronoun*), but include, for instance, (*noun, virtual-tag1*). The entropy, $H_2$ and $H_3$, with virtual tags explicitly marking some constituents is lower than that without the virtual tags.

## 4 Analysis of the MARSEC corpus

In a similar way the words from a speech signal can be segmented into groups, with periodic discontinuities. There is a relationship between prosody and syntax, and the placement of discontinuities provide clues to syntactic structure (Arnfield, 1994; Fang and Huckvale, 1996; Ostendorf and Vielleux, 1994).

We have investigated how the entropy of sequences of words varies when discontinuities are represented. This research was carried out using MARSEC (Machine Readable Spoken English 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). Discontinuities in speech can have many causes: hesitation phenomena perform a number of roles, particularly in spontaneous speech. However, by using this corpus from professional speakers we propose that the dominant cause for discontinuities will be related to the efficient transfer of information.

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

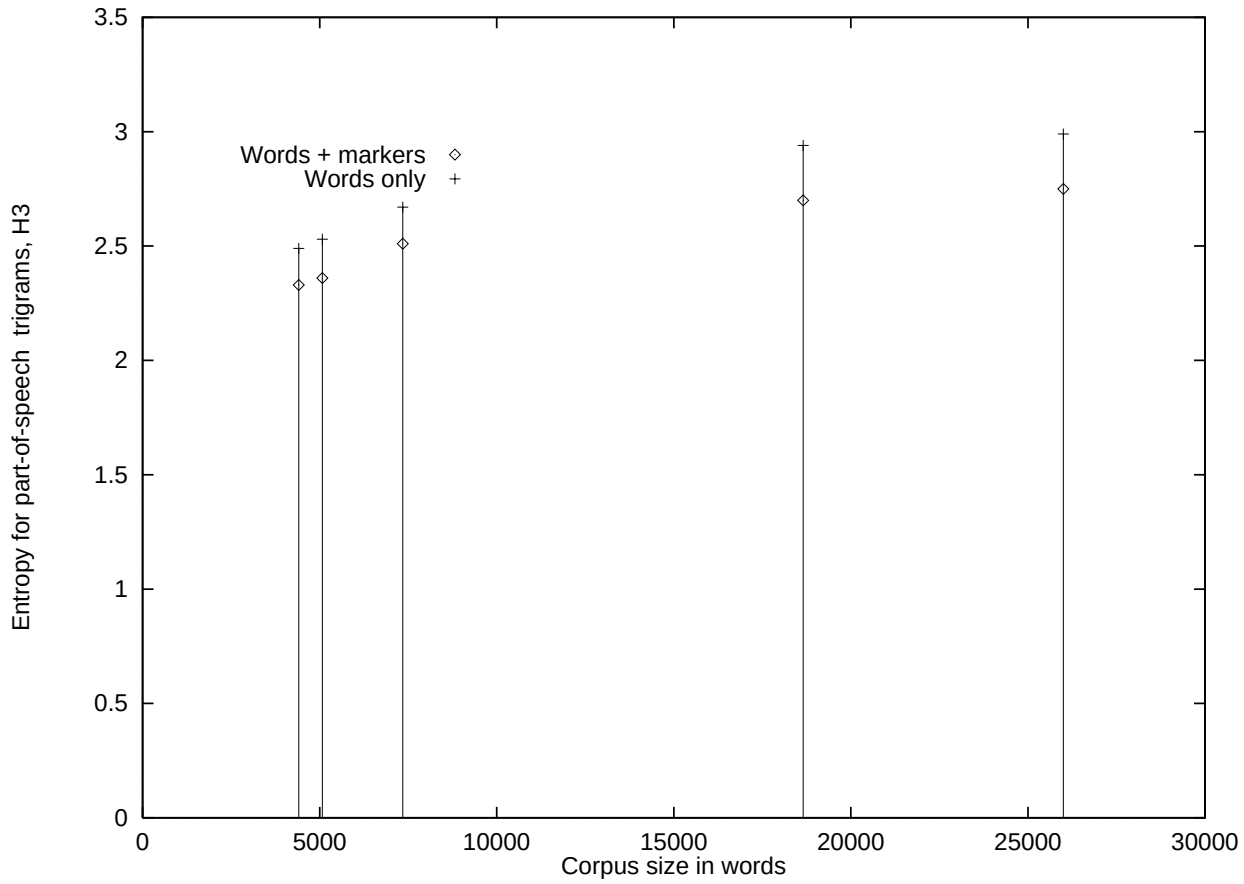


Figure 1: Comparison of trigram part-of-speech entropy for sections of the MARSEC corpus, (i) with both major and minor discontinuities marked (ii) without either. The tagset size is 28 with the discontinuities represented, 26 without them. Table 2 gives the data for the 18655 word corpus as an example

was done by two trained annotators. Most of the corpus was marked up by one or the other of the annotators, but a few sections have been marked up by both. We see that there is a large measure of agreement, but not a total consensus. In Table 1 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.

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

### 4.1 Aim of the investigation

Our purpose is to examine whether the entropy of a corpus is reduced by representing discontinuities as well as words. Since we are interested in syntactic structures we work with parts-of-speech rather than actual words. 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. The tagset used in this work is given in the Appendix. There are 26 classes, 28 when discontinuity markers are also represented.

Care is needed to compare entropy measures for sequences of different alphabet sizes, but we propose to extract information in the following way.

- $H_0$ 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 $H_3$ could be lower or higher depending on
  - the frequency of the new symbol
  - whether the marker captures some of the language structure

We will be looking for cases where $H_1$ rises while $H_2$ and / or $H_3$ decline. This indicates that it is not the frequency of the new symbol that causes the

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

Table 2: Entropy measures for 18655 words of the MARSEC corpus, (sections A, B, C concatenated) with and without major and minor discontinuities. See text on calculation of $H_3$.

| 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 3: 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)

decline, but the capture of some structure. If $H_1$ declines, then a *fall* in $H_2$ and $H_3$ is not informative.

The corpus size increases marginally with the addition of the markers, which should lead to a marginal increase in entropy. So any effect from this will not account for a fall in $H_2$ and $H_3$.

If we worked with words rather than parts-of-speech the discontinuity markers would be significantly more frequent than any words. The entropy would decline but we would not be able to ascertain the cause.

### 4.2 Implementing the investigation

To conduct this investigation the MARSEC corpus was automatically tagged, using a version of the Claws tagger[3]. These tags were mapped onto the smaller tagset (see Appendix). Random inspection indicated about 96% words correctly tagged.

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 2, and in Figure 1.

$H_3$ is calculated in the following way. We assume a dependency does not reach across a major discontinuity (see, for instance Ney et al. (Ney et al., 1997, page 200)). Therefore, we omit any triple that spans a major discontinuity.

Note that we are interested in *comparative* entropies. We do not calculate entropy on unseen test data, since it is not our aim to get a best estimate, but to compare results with and without representing discontinuities.

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 (Farach and et al., 1995). 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.

## 5 Results

Figure 1 shows the results for different size corpora. Table 2 gives the results of this investigation for the corpus of 18655 words. It shows that when major and minor discontinuities are represented, then $H_2$ and $H_3$ decline even though $H_1$ increases. For the representation of words with minor discontinuities alone, results are not conclusive, since $H_1$ declines. For the representation of words with major discontinuities alone $H_3$ declines, though $H_2$ shows a contrary movement.

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 3. $H_2$ 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

## 6 Conclusion

In this paper we have examined mechanisms by which language is encoded in such a way that it can

[3]Claws4, supplied by the University of Lancaster, described by (Garside, 1987)

be decoded as easily as possible. We examined different representations of English speech and saw that it can be more efficiently coded when discontinuities are represented (Section 4). Strings of words with no prosodic boundaries represented are associated with higher levels of entropy, which makes decoding harder. Using information on discontinuities in speech is an aid to producing a more efficient code.

As language has evolved, we would expect selection pressure to encourage the development of segmented modes of representation, where segments correspond to structural elements. The evolution of structured language can be seen as the survival of the fittest in the statistical environment.

In the same way, developers of automated speech and language processing systems can exploit the statistical characteristics of sequences of words. This result is of general interest, and supports the development of improved language models for many applications.

## References


S Arnfield. 1994. *Prosody and Syntax in Corpus Based Analysis of Spoken English*. Ph.D. thesis, University of Leeds.

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

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

Alex Chengyu Fang and Mark Huckvale. 1996. Synchronising syntax with speech signals. In V.Hazan et al., editor, *Speech, Hearing and Language*. University College London.

M Farach and M Noordewier et al. 1995. On the Entropy of DNA. In *Symposium on Discrete Algorithms*.

R Garside. 1987. 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.

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

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

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

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

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

C Lyon. 1999. Exploiting statistical characteristics of word sequences for the efficient coding of speech (extended version). Technical report, Computer Science Department, University of Hertfordshire. Number 333.

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

H Ney, S Martin, and F Wessel. 1997. Statistical language modelling using leaving-one-out. In S Young and G Bloothooft, editors, *Corpus Based Methods in Language and Speech Processing*. Kluwer Academic Publishers.

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

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


### Appendix: The tagset of 26 classes used in the experiments

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