Entropy Indicators for Investigating Early Language Processes

Caroline Lyon⋆, Chrystopher L. Nehaniv⋆, and Bob Dickerson⋆

⋆School of Computer Science
University of Hertfordshire
College Lane
Hatfield, Hertfordshire AL10 9AB
United Kingdom

`C.M.Lyon@herts.ac.uk`, `C.L.Nehaniv@herts.ac.uk`, `R.G.Dickerson@herts.ac.uk`

Abstract

We examine evidence for the hypothesis that language could have passed through a stage when words were combined in structured linear segments and these linear segments could later have become the building blocks for a full hierarchical grammar. Experiments were carried out on the British National Corpus, consisting of about 100 million words of text from different domains and transcribed speech. This work extends and supports the results of our previous work based on a smaller corpus reported previously. Measuring the entropy of the texts we find that entropy declines as words are taken in groups of 2, 3 and 4, indicating that it is easier to decode words taken in short sequences rather than individually. Entropy further declines when punctuation is represented, showing that appropriate segmentation captures some of the language structure. Further support for the hypothesis that local sequential processing underlies the production and perception of speech comes from neurobiological evidence. The observation that homophones are apparently ubiquitous and used without confusion also suggests that language processing may be largely based on local context.

1 Introduction

Hypotheses on the evolution of language can sometimes be supported, or undermined, by an investigation into underlying characteristics of present day language. Information theory provides some effective tools for carrying out such investigations, and is employed here as a tool for examining the hypothesis that the underpinnings of modern human language may lie in sequential processing phenomena, (though we also find that simple observations of every day speech can also be illuminating.)

1.1 Overview of the investigations

The core of the work described in this paper is an investigation into the statistical characteristics of spoken and written language which can help explain why language was likely to evolve with a certain structure. We take a large corpus of written text and transcribed speech and see whether the efficiency of encoding and decoding the stream of language is improved by processing a short sequence of words rather than individual words. To do this we measure the entropy of the word sequence, comparing values when we take single words, pairs, triples and quads. A decline in entropy indicates an increase in predictability, facilitating an improvement in decoding efficiency.

We also measure the entropy with and without punctuation, to see whether communication is more efficient if the stream of words is broken into segments that usually correspond to syntactic components. Our experiments (reported below) show that entropy does indeed decline as word sequences up to length three are processed, and thus supports the hypothesis that local sequential processing underpins communication through language. Entropy also declines further with the inclusion of punctuation. As there is a strong correlation between punctuation and prosodic markers in speech (Fang and Huckvale (1996); Taylor and Black (1998)) this decline indicates that there is an advantage in taking language in the segments that prosodic markers provide, since it is then easier to decode.

This suggests that there could be an intermediate stage in the development of a full hierarchical grammar. Processing a linear stream of words that is appropriately segmented is more efficient for the decoder than taking unsegmented, continuous strings of words. Such segments can then be the components of a hierarchical grammar.

Experiments have been carried out with the British National Corpus, BNC, about 100 million words of text and transcribed speech from many different domains (BNC).

1.2 Related work

We point to recent work on the "small world" phenomenon that investigates possible universal patterns of organization in complex systems (i Cancho and Sole (2001)). This effect, which is evident in natural language, picks up on the dominance of local dependencies, and research is going on into how robust complex systems can emerge, Section 5.

We also draw attention to other work that supports our hypotheses: neurobiological, computer modelling, and simple observation of everyday speech.

2 Background to this work

2.1 Co-operative communication

A number of scenarios have been used to introduce hypotheses on the evolution of language, and methods of communication between different animal species in different situations have been studied extensively. This has included a range of possibilities such as "gossip, deceit, alliance building, or other social purposes" (Bickerton (2002)). The work described here is based on those scenarios where producers and receivers are co-operating, sharing information. In the past little work in behavioural ecology had been done to make systematic comparisons of co-operative and non-cooperative signals (Krebs and Davies (1993)). A typical scenario for co-operative communication would be in group hunting or fishing situations, where deceit would be counter-productive. Even with manipulative communication a degree of co-operation is required to enable understanding. We look at modes of communication that are most efficient for producers and receivers. To investigate this we take a large corpus of spoken and written language and apply an analytic tool from information theory, the entropy measure, to help determine which possible characteristics of communication can make it more or less efficient.

2.2 Entropy indicators

The original concept of entropy was introduced by Shannon (1993)[1951]. Informally, it is related to predictability: the lower the entropy the better the predictability of a sequence of symbols. Shannon showed that the entropy of a sequence of letters declined as more information about adjacent letters is taken into account; it is easier to predict a letter if the previous ones are known. Entropy is represented as H, and we measure

- H_0 : entropy with no statistical information, symbols equi-probable.

- H_1 : entropy from information on the probability of single symbols occurring.

- H_2 : entropy from information on the probability of 2 symbols occurring consecutively.

- H_n : entropy from information on the probability of n symbols occurring consecutively.

More precisely, H_n measures the uncertainty of a symbol, conditional on its $n - 1$ predecessors. (For $n > 0$, this is called the conditional entropy.)

For an introductory explanation of the concept of entropy, see (Lyon et al., 2003, page 170). The derivation of the formula for calculating entropy is in Appendix B. For many years Automated Speech Recognition developers have used entropy metrics to measure performance (Jelinek (1990)).

2.3 Using real language

A significant amount of language analysis in this field has not been done with real language. Well known examples include Elman's experiments with recurrent nets (Elman (1991)), which use a 23 word vocabulary: 12 verbs 10 nouns and a relative pronoun. Sentences like *boy sees boy* are considered grammatical, because there is number agreement between the subject and verb, though this sentence would be considered ungrammatical in real language with determiners missing. Elman himself is careful to say that this language is artificial, but this is not the case with many of his followers, who claim it is is a subset of natural language.

In fact many, sometimes most, of the words most people utter are function words. Though in any model we have to abstract out the features we consider most significant, we suggest that the common focus on content words introduces distortions. For example, to jump from words to syntactic combinations of nouns

and verbs without considering the intermediate stage of phrase development leads to unrealistic conclusions. In our work we need to take language as it is.

3   The British National Corpus

Other recent work in this field has been done on a comparatively small corpus of 26,000 words of transcribed speech, annotated with prosodic markers (Lyon et al. (2003, 2004)). However, using the large BNC corpus enables us to confirm those results, and extend them.

The BNC corpus is composed of a representative collection of English texts; about 10% of the total is transcribed speech. As we want to investigate the processing of running language, headlines, titles, captions and lists are excluded from our experiments. Then adding in punctuation marks leads to a corpus of about 107 million symbols.

In order to carry out an analysis on strings of words it is necessary to reduce an unlimited number of words to a smaller set of symbols, and so words are mapped onto parts of speech tags. As well as making the project computationally feasible this approach is justified by evidence that implicit allocation of parts of speech occurs very early in language acquisition by infants, even before lexical access to word meanings (Morgan et al. (1996)).

The BNC corpus has been tagged, with a tagset of 57 parts of speech and 4 punctuation markers. We have mapped these tags onto our own tagset of 32 classes, of which one class represents any punctuation mark (Appendix A). Tag sets can vary in size but our underlying aim is to group together words that function in a similar way, have similar neighbours. Thus, for example, lexical verbs can usually have the same type of predecessors and successors whether they are in the present or past tense:

We like swimming / We liked swimming

so in our tagset they are in one class. This maintains a good degree of discriminability while moving to a smaller, fairly natural tagset. Moreover, another reason for mapping the BNC tagset onto our smaller set is that the entropy measures are more pronounced for the smaller set, while a larger tagset would require even larger corpora to avoid undersampling errors in entropy estimates.

4   Experiments

We have run the following experiments. First, we have processed the whole corpus of 107 million parts of speech tags, with punctuation, and found H_1, H_2, H_3, and H_4 as shown in Table 1. We also ran experiments over each of the 10 directories in which the corpus material is placed to see if there was much variation. In fact, variations between the directories is small: the results cluster round a central tendency shown by the measure for the whole corpus. An example is shown in Table 1.

We also process a comparable set of randomly generated numbers, in order to ensure that distortions do not occur because of undersampling. With 32 tags the number of possible sequences of length 5 are 33,554,432. If too small a sample is used the entropy appears lower than it should, since, e.g. not all the infrequent cases have occurred. A simple empirical test on sample size is through a random number sequence check. For a random sequence, the entropy should not decline as more of the information over preceding numbers is taken into account, since they are generated independently. Thus H for a sequence of random numbers in the range 0 to 31 should stay at 5.0. Sequences of random numbers are produced by the Unix random number generator. The results show that for the whole corpus we can be fully confident up to the H_4 figure, but H_5 should be treated with caution. For the 10 subdirectories, H_4 should be treated with caution, and H_5 is omitted.

Secondly, we process the whole BNC corpus, but omitting punctuation marks, as shown in Table 2. This time there will be 31 tags, as the punctuation symbol is omitted. The number of words is reduced, as punctuation marks are counted as words.

4.1   Analysis of results

The results in Table 1 show that entropy declines as processing is extended over the 1, then 2 and then 3 preceding consecutive parts of speech tags. There is a small further decline when 4 consecutive tags are taken. The results for 5 consecutive tags are not considered fully reliable, in view of the random sequence check for 107 million symbols.

Compare these results with those in Table 2. This time there is one less tag symbol, so we expect unpredictability to decrease compared to that for the corpus tagged with 32 symbols, and entropy to be less. This is what we find for H_0 and for H_1. However, as we take words 2, 3 and 4 at a time we find that entropy is slightly greater than in the first case. This

Corpus	H_0	H_1	H_2	H_3	H_4	H_5
107 million words + punctuation 32 tags	5.0	4.19	3.27	2.94	2.84	(2.75)
107 million random words 32 tags	5.0	5.0	5.0	5.0	5.0	4.8
10 million words, subdirectory F 32 tags	5.0	4.18	3.25	2.91	(2.79)	
10 million random words 32 tags	5.0	5.0	5.0	5.0	4.93	3.05

Table 1: Entropy measures for the BNC corpus, mapped onto 32 parts of speech tags. 3-grams, 4-grams and 5-grams that span a punctuation mark are omitted. Figures in brackets are to be treated with caution.

Corpus	H_0	H_1	H_2	H_3	H_4	H_5
94 million words, no punctuation 31 tags	4.95	4.16	3.29	3.14	3.07	(3.01)
94 million random words, 31 tags	4.95	4.95	4.95	4.95	4.95	4.72

Table 2: Entropy measures for the BNC corpus, mapped onto 31 parts of speech tags, omitting punctuation. The figure in brackets should be treated with caution.

indicates that punctuation captures some of the structure of language, allowing the next parts of speech tag to be be better predicted, and that by removing punctuation (corresponding to prosodic marking in speech) we increase the uncertainty. Paraphrasing Shannon we can say that a string of words between punctuation marks is a cohesive group with internal statistical influences, and consequently the n-grams within such phrases, clauses or sentences are more restricted than those which bridge punctuation ((Shannon, 1993, page 197)).

These results indicate that a stream of language is easier to decode if words are taken in short sequences rather than as individual items, and supports the hypothesis that local sequential processing underlies communication through language.

5 Other evidence for local processes

5.1 Computer modelling and the "small world" effect

In consider local processing, it is instructive to look at syntactic models based on dependency grammar and related concepts. Dependency grammar assumes that syntactic structure consists of lexical nodes (words) and binary relations (dependencies) linking them. Though these models are word based, phrase structure emerges. An online practical example is the Link Parser (Sleator et al. (2005)) where you can parse your own texts and see how the constituent tree emerges. Now, it is reported (i Cancho (2004)) that, in experiments in Czech, German and Romanian with a related system, about 70% of dependencies are between neighbouring words, 17% at a distance of 2.

This is one of the characteristics of the small world effect. A significant amount of syntactic knowledge is available from local information, even before our grammatical capability is enhanced by the addition of long range dependencies associated with phrase structure hierarchies.

From this one could also suggest that an intermediate stage in the development of a fully fledged grammar could have been based on local syntactic constraints.

Returning to another computer model, Elman's recurrent networks, we note that they could have a useful role to play in modelling short phrasal strings, but there are inherent obstacles to modelling longer dependencies (Bengio (1996); Hochreiter et al. (2001)).

5.2 Neurobiological evidence

Our hypothesis is also supported by the fact that primitive sequential processors in the basal ganglia play an essential role in language processing (Lieberman (2000, 2002)). The neural substrate that regulates motor control includes the control of articulatory acts, and this part of the brain seems to have extended its role to manage the sequencing of linguistic elements. An overview of the evidence that language and motor abilities are connected is given in a special edition of *Science* (Holden (2004)).

5.3 Simple observations of everyday speech

Any hypothesis on the evolution of language needs to explain why all languages seem to have homophones (Lyon et al. (2004)). In English some of the most frequently used words have more than one meaning such as *to / too / two*. Even young children seem able to disambiguate them without difficulty. In an agglutinative language such as Finnish they are rarely used by children, but occur in adult speech (Warren (2001)).

Their prevalence undermines the theories based on the assumptions that words in evolutionarily advanced language have a single meaning, that "the evolutionary optimum is reached if every word is associated with exactly one signal" ((Nowak et al., 1999, page 151)) and that there is a "loss of communicative capacity that arises if individual sounds are linked to more than one meaning" ((Nowak et al., 2002, page 613)). While such theories and models may appear to be logically attractive, they do not represent real language.

However, if we accept the hypothesis that local sequential processing underlies our language capability then there is not a problem accounting for the homophone phenomenon: homophones can be disambiguated by the local context.

6 Conclusion

When we look for clues to the evolution of language we can examine the state humans are in now and reason about how we could have arrived at the present position. This may take the form of brain studies, but it can also include the sort of analysis of language that we are doing. Chomsky once famously claimed that "One's ability to produce and recognize grammatical structures is not based on notions of statistical approximation and the like" (Chomsky (1957)). However, statistics can illuminate the way in which language processing has been carried out, and investigations on large corpora can now be done that were not possible a few decades back.

Our experiments suggest that utterances are processed in segments of a few words. We go on to hypothesize that these segments could be the elements out of which a hierarchical grammar is built.

References

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

Yoshua Bengio. *Neural Networks for Speech and Sequence Recognition*. ITP, 1996.

Derek Bickerton. Foraging versus social intelligence in the evolution of protolanguage. In Alison Wray, editor, *The Transition to Language*, pages 207–225. OUP, 2002.

The British National Corpus. The BNC Consortium, http://www.hcu.ox.ac.uk/BNC.

N Chomsky. *Syntactic Structures*. The Hague: Mouton, 1957.

J L Elman. Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning*, pages 195–223, 1991.

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

S Hochreiter, Y Bengio, P Frasconi, and J Schmidhuber. Gradient flow in recurrent nets: the difficuly of learning long term dependencies. In S C Kremer and J F Kolen, editors, *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Constance Holden. The origin of speech. *Science*, 303:1316 –1319, 2004.

R Ferrer i Cancho. Patterns in syntactic dependency networks. *Physical Review E*, 69:051915, 2004.

Ramon Ferrer i Cancho and Ricard V. Sole. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, 2001.

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.

J R Krebs and N B Davies. *An Introduction to Behavioural Ecology*. Blackwell, 1993.

P Lieberman. *Human Language and our Reptilian Brain*. Harvard University Press, 2000.

P Lieberman. On the nature and evolution of the neural bases of human language. *Yearbook of Physical Anthropology*, 2002.

C Lyon, B Dickerson, and C L Nehaniv. The segmentation of speech and its implications for the emergence of language structure. *Evolution of Communication*, 4, no.2:161–182, 2003.

C Lyon, C L Nehaniv, S Warren, and J Baillie. Evolutionary fitness, homophony and disambiguation through sequential processes. In *Proceedings of First International Workshop on Emergence and Evolution of Linguistic Communication, JSAI*, 2004.

J Morgan, R Shi, and P Allopenna. Perceptual bases of rudimentary grammatical categories. In J Morgan and K Demuth, editors, *Signal to Syntax*. Lawrence Erlbaum, 1996.

M A Nowak, N L Komaraova, and P Niyogi. Computational and evolutionary aspects of language. *Nature*, 417:611 – 617, 2002.

Martin A Nowak, Joshua B Plotkin, and David C Krakauer. The evolutionary language game. *J. Theoretical Biology*, 200:147–162, 1999.

C E Shannon. Prediction and Entropy of Printed English (1951). In N J A Sloane and Aaron D Wyner, editors, *Shannon: Collected Papers*. IEEE Press, 1993.

D Sleator, D Temperly, and J Lafferty. *Link Grammar*. Carnegie Mellon University, http://www.link.cs.cmu.edu/link/, 2005. Visited 17 Jan 2005.

P Taylor and A Black. Assigning phrase breaks from part-of-speech sequences. In *Proceedings of Eurospeech'97, Rhodes*, pages 995–998, 1998.

S Warren. *Phonological Acquisition and Ambient Language: a Corpus Based, Cross-Linguistic Exploration*. PhD thesis, University of Hertfordshire, UK, 2001.

Appendix A

The tagset of the British National Corpus is mapped onto our tagset. Each of the BNC tags is mapped onto an integer, as shown below, so that functionally similar tags are grouped together.

Tag	Code for our mapping
AJ0	1
	Adjective (general or positive) (e.g. *good, old, beautiful*)
AJC	1
	Comparative adjective (e.g. *better, older*)
AJS	1
	Superlative adjective (e.g. *best, oldest*)
AT0	2
	Article (e.g. *the, a, an, no*)
AV0	3
	General adverb: an adverb not subclassified as AVP or AVQ (see below) (e.g. *often, well, longer* (adv.), *furthest*).
AVP	3
	Adverb particle (e.g. *up, off, out*)
AVQ	3
	Wh-adverb (e.g. *when, where, how, why, wherever*)
CJC	4
	Coordinating conjunction (e.g. *and, or, but*)
CJS	4
	Subordinating conjunction (e.g. *although, when*)
CJT	4
	The subordinating conjunction *that*
CRD	2
	Cardinal number (e.g. *one, 3, fifty-five, 3609*)

DPS   5
Possessive determiner-pronoun (e.g. *your, their, his*)

DT0   2
General determiner-pronoun: i.e. a determiner-pronoun which is not a DTQ or an AT0.

DTQ   2
Wh-determiner-pronoun (e.g. *which, what, whose, whichever*)

EX0   6
Existential *there*, i.e. *there* occurring in the *there is ...* or *there are ...* construction

ITJ   7
Interjection or other isolate (e.g. *oh, yes, mhm, wow*)

NN0   8
Common noun, neutral for number (e.g. *aircraft, data, committee*)

NN1   9
Singular common noun (e.g. *pencil, goose, time, revelation*)

NN2   10
Plural common noun (e.g. *pencils, geese, times, revelations*)

NP0   11
Proper noun (e.g. *London, Michael, Mars, IBM*)

ORD   1
Ordinal numeral (e.g. *first, sixth, 77th, last*) .

PNI   12
Indefinite pronoun (e.g. *none, everything, one* [as pronoun], *nobody*)

PNP   13
Personal pronoun (e.g. *I, you, them, ours*)

PNQ   14
Wh-pronoun (e.g. *who, whoever, whom*)

PNX   15
Reflexive pronoun (e.g. *myself, yourself, itself, ourselves*)

POS   16
The possessive or genitive marker *'s* or *'*

PRF   17
The preposition *of*

PRP   18
Preposition (except for *of*) (e.g. *about, at, in, on, on behalf of, with*)

PUL   0
Punctuation: left bracket - i.e. *(* or *[*

PUN   0
Punctuation: general separating mark - i.e. *. , ! , : ; - * or *?*

PUQ   0
Punctuation: quotation mark - i.e. *'* or *"*

PUR   0
Punctuation: right bracket - i.e. *)* or *]*

TO0   19
Infinitive marker *to*

UNC   7
Unclassified items which are not appropriately considered as items of the English lexicon.

VBB   20
The present tense forms of the verb BE, except for *is, 's*: i.e. *am, are, 'm, 're* and *be* [subjunctive or imperative]

VBD   20
The past tense forms of the verb BE: *was* and *were*

VBG   21
The *-ing* form of the verb BE: *being*

VBI   22
The infinitive form of the verb BE: *be*

VBN   23
The past participle form of the verb BE: *been*

VBZ   24
The *-s* form of the verb BE: *is, 's*

VDB   20
The finite base form of the verb DO: *do*

VDD   20
The past tense form of the verb DO: *did*

VDG   21
The *-ing* form of the verb DO: *doing*

VDI   22
The infinitive form of the verb DO: *do*

VDN   23
The past participle form of the verb DO: *done*

VDZ   24
The *-s* form of the verb DO: *does, 's*

VHB   20
The finite base form of the verb HAVE: *have, 've*

VHD   20
The past tense form of the verb HAVE: *had, 'd*

VHG   21
The *-ing* form of the verb HAVE: *having*

VHI   22
The infinitive form of the verb HAVE: *have*

VHN   23
The past participle form of the verb HAVE: *had*

VHZ   24
The *-s* form of the verb HAVE: *has, 's*

VM0   25
Modal auxiliary verb (e.g. *will, would, can, could, 'll, 'd*)

VVB   26
The finite base form of lexical verbs (e.g. *forget, send, live, return*) [Including the imperative and present subjunctive]

VVD   26

> The past tense form of lexical verbs (e.g. *forgot, sent, lived, returned*)

VVG   27

> The *-ing* form of lexical verbs (e.g. *forgetting, sending, living, returning*)

VVI   28

> The infinitive form of lexical verbs (e.g. *forget, send, live, return*)

VVN   29

> The past participle form of lexical verbs (e.g. *forgotten, sent, lived, returned*)

VVZ   30

> The *-s* form of lexical verbs (e.g. *forgets, sends, lives, returns*)

XX0   31

> The negative particle *not* or *n't*

ZZ0   7

> Alphabetical symbols (e.g. *A, a, B, b, c, d*)

Appendix B

The derivation of the formula for calculating conditional entropy

This is derived from Shannon's work on the entropy of symbol sequences. 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 $p(b_i, j) \div p(b_i)$

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

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 that used, for instance, by Bell et al. (1990).