LANGUAGE IDENTIFICATION IN UNKNOWN SIGNALS

Contact author: John Elliott, jre@scs.leeds.ac.uk

Co-authors: Eric Atwell, eric@scs.leeds.ac.uk
Bill Whyte, billw@scs.leeds.ac.uk

Organisation: Centre for Computer Analysis of Language and Speech,
School of Computer Studies, University of Leeds, Leeds, Yorkshire, LS2 9JT England

Abstract

This paper describes algorithms and software developed to characterise and detect generic intelligent language-like features in an input signal, using Natural Language Learning techniques: looking for characteristic statistical "language-signatures" in test corpora. As a first step towards such species-independent language-detection, we present a suite of programs to analyse digital representations of a range of data, and use the results to extrapolate whether or not there are language-like structures which distinguish this data from other sources, such as music, images, and white noise. We assume that generic species-independent communication can be detected by concentrating on localised patterns and rhythms, identifying segments at the level of characters, words and phrases, without necessarily having to "understand" the content.

We assume that a language-like signal will be encoded symbolically, i.e. some kind of character-stream. Our language-detection algorithm for symbolic input uses a number of statistical clues: data compression ratio, "chunking" to find character bit-length and boundaries, and matching against a Zipfian type-token distribution for "letters" and "words". We do not claim extensive (let alone exhaustive) empirical evidence that our language-detection clues are "correct"; the only real test will come when the Search for Extra-Terrestrial Intelligence finds true alien signals. If and when true SETI signals are found, the first step to interpretation is to identify the language-like features, using techniques like the above. Our current research goal is to apply Natural Language Learning techniques to the identification of "higher-level" grammatical and semantic structure in a linguistic signal.

Introduction

A useful thought experiment is to imagine eavesdropping on a signal from outer space. How can you decide that it is a message between intelligent life forms, without dialogue with the source? What is special about the language signal that separates it from non-language? What special ’zone’ in the signal universe does language occupy? Is it, indeed, separable from other semi-structured sources, such as DNA and music (fig 1).

Solving this problem might not only be useful in the event of detecting such signals from space, but also, by deliberately ignoring preconceptions based on human texts, may provide us with some better understanding of what language really is.

The Signal Universe

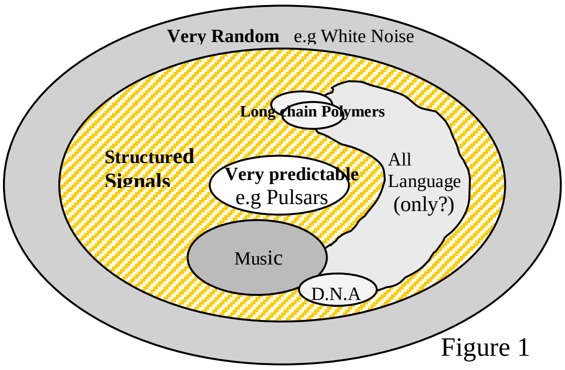


Figure 1

However, we need to start somewhere, and our initial investigations - which this paper summarises - make some basic assumptions (which we would hope to relax in later research). Namely, that identifiable script will be a serial string, possessing a hierarchy of elements broadly equivalent to ‘characters,’ ‘words’, and 'spaces', and possess something akin to human grammar.

Identifying the 'Character Set'

In 'real' decoding of unknown scripts it is accepted that identifying the correct set of discrete symbols is no mean feat (Chadwick 1967). To make life simple for ourselves we assume a digital signal with a fixed number of bits per character. Very different techniques are required to deal with audio or analogue equivalent waveforms (Elliott & Atwell 99, 00). We have reason to believe that the following method can be modified to relax this constraint, but this needs to be tested further.

The task then reduces to trying to identify the number of bits per character.

Suppose the probability of a bit is P_I. Then the message entropy of a string of length N will be given by:

$$\mathbf{E = SUM\ [P_I \ln P_i];\ i = 1,N}$$

If the signal contains merely a set of random digits, the expected value of this function will rise monotonically as N increases. However, if the string contains a set of symbols of fixed length representing a character set used for communication, it is likely to show some decrease in entropy when analysed in blocks of this length, because the signal is 'less random' when thus blocked. Of course, we need to analyse blocks that begin and end at character boundaries. We simply carry out the measurements in sliding windows along the data. In figure 2 below, we see what happens when we apply this to samples of 8-bit ASCII text:

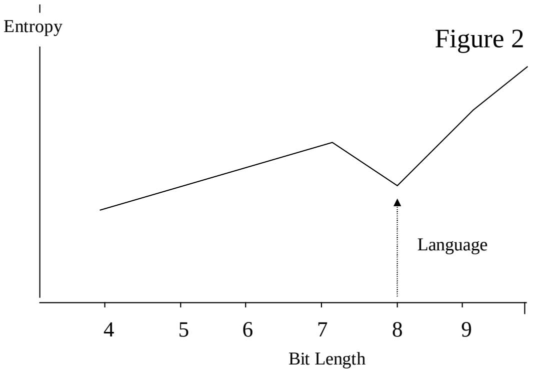


Figure 2

Entropy profile as an indicator of character bit-length

We notice a clear drop, as predicted, for a bit length of 8.Modest progress though it may be, it is not unreasonable to assume that the ***first piece of evidence for the presence of language-like structure, would be the identification of a low-entropy, character set within the signal.***

Identifying 'Words'

Again, work by crytopaleologists suggests that, once the character set has been found, the separation into word-like units, is not trivial and again we cheat, slightly: we assume that the language possesses something akin to a 'space' character. Taking our entropy measurement described above as a way of separating characters, we now try to identify the one, which represents 'space'. It is not unreasonable to believe that, in a word-based language, it is likely to be one of the most frequently used characters.

Using a number of texts in a variety of languages, we first identified the top three most used characters. For each of these we hypothesised in turn that it represented 'space'. This then allowed us to segment the signal into words-like units ('words' for simplicity). We could then compute the frequency distribution of words as a function of word length, for each of the three candidate 'space' characters (fig 3).

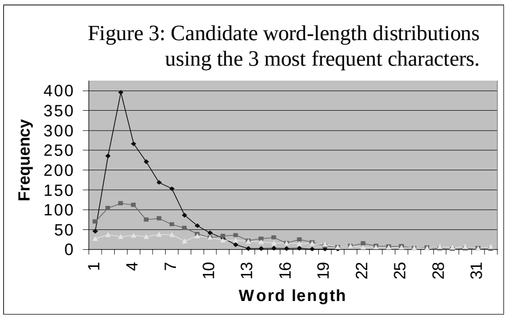


Figure 3: Candidate word-length distributions using the 3 most frequent characters.

It can be seen that one 'separator' candidate (unsurprisingly, in fact, the most frequent character of all) results in a very varied distribution of word lengths. This is an interesting distribution, which, on the right hand side of the peak, approximately follows the well-known 'law' according to Zipf (Zipf, 1949), which predicts this behaviour on the grounds of minimum effort in a communication act.

To ascertain whether the word-length frequency distribution holds for language in general, multiple samples from 20 different languages from Indo-European, Bantu, Semitic, Finno-Ugrian and Malayo-Polynesian groups were analysed (fig 4).

Word length distributions in multiple samples from Indo-European, Semitic, Finno-Ugrian, and Malayo-Polynesian language groups

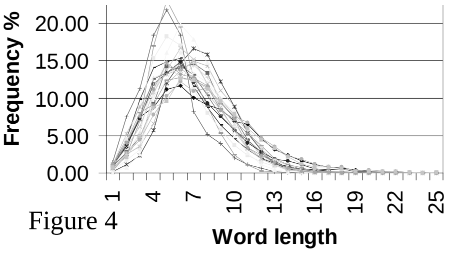
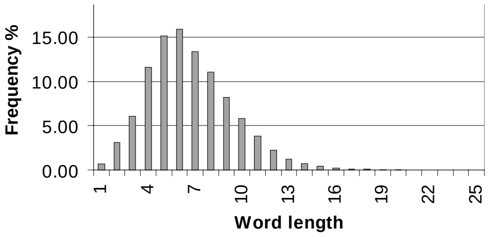


Figure 4

Using statistical measures of significance, it was found that most groups fell well within 5% limits – only two individual languages were near exceeding these limits – of the proposed Human language word-length profile shown in fig 5.

Figure 5: Human language word-length frequency distribution profile



Zipf's law is a strong indication of language-like behaviour. It can be used to segment the signal provided a 'space' character exists.

However, we should not assume Zipf to be an infallible language detector. Other natural phenomena such as molecular distribution in yeast DNA possess characteristics of power laws. Analyses of protein length distributions also display Poisson distributions where the number of proteins is plotted against the lengths of amino acids (Jenson 1998).

Identifying 'Phrases'

Although alien brains may be more or less powerful than ours (Norris 1999), it is reasonable to assume that all intelligent problem solvers are subject to the same ultimate constraints of computational power and storage and their symbol systems will reflect this.

Thus, language must use small sets of rules to generate a vast world of implications and consequences. Perhaps its most important single device is the use of embedded clauses and phrases (Minsky 1984), with which to represent an expression or description, however complex, as a single component of another description.

In serial languages, this appears to be achieved by clustering words into 'chunks' (phrases, sentences) of information, which are more-or-less consistent and self-contained elements of thought. Furthermore, in human language at least, these 'chunks' tend to consist of *content* terms, which describe what the chunk is 'about' and *functional* terms, which attribute references and context by which the content terms convey their information unambiguously. 'King' is usually a content term; 'of' and 'the' are functional. We use 'term' rather than word, because many languages make far less use of full words for functional operations than does English: in Latin the transformation 'rex' ('king') to 'regis' (of the king) is one such example.

Functional terms in a language tend to be short, probably attributable to the principle of least effort, as they are used frequently.

A further distinguishing characteristic of functional and content terms is that different texts will often vary in their content but tend to share a common linguistic structure and therefore make similar use of functional terms. That is, the probability distribution of content terms will vary from text to text, but the distribution of function terms will not.

Using English text, which had been enciphered using a simple substitution cipher (to avoid cheating), we identified

across a variety of texts, the most common words, with least inter-text variation. These we call ‘candidate function words’.
Now, suppose these words occurred at random in the signal: we would expect to see the spacing between them to be merely a function of their individual probabilities of occurrence. Analysing this statistically (as a Poisson distribution) or simply simulate it practically, we find that there are a non-insignificant number of cases wherein there are very large gaps (of the order of several tens of words) between successive occurrences. Compare this with the results from our analysis (fig 6).

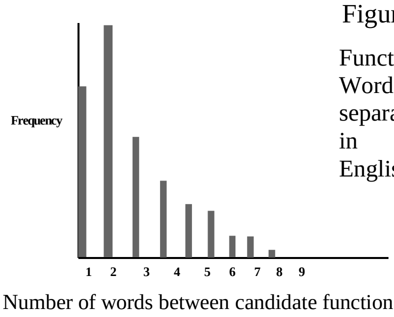


Figure 6
Function Word separation in English.

Initial findings show that the frequency distribution of these lengths of text – our candidate phrases – follow a Zipfian distribution curve and rarely exceed lengths of more than eight.

We might conclude from this, that our brains tend to ‘chunk’ linguistic information into phrase-like structures of the order of seven or so word units long.

Interestingly enough, this fits in well with human cognition theory (Ally & Bacon 1991), which states that our short-term mental capacity operates well only up to 7 (+ or - 2) pieces of information, but any causal connection between this and our results must be considered highly speculative at this stage!

Directions for Future Research

We are familiar with parts of speech (commonly, ‘nouns’, verbs’ etc) in language. Identification of patterns indicative of these would be further evidence of language-like characteristics and, by allowing us to group together the numerous word tokens in any language into smaller, more manageable collections would facilitate statistical analysis. Some attempts have been made in the past to use n-gram probabilities in order to define word classes or ‘parts of speech’ (Charniak 1993).
In our own work we have begun the development of tools that measure the correlation profile between pairs of words, as a precursor to deducing general principles for ‘typing’ and clustering into syntactico-semantic classes.

Correlation profile for word pair

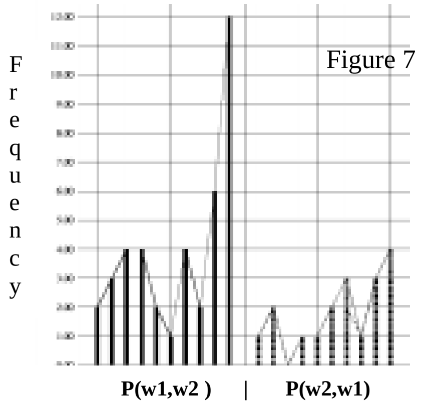


Figure 7

The figure 7 above shows the results for the relationship between a pair of unknown (because of the substitution cipher approach) content and functional words, so identified by looking at their cross-corpus statistics as described above. It can be seen that the functional word has a very high probability of preceding the content word but has no instance of directly following it. At least metaphorically, the graph can be considered to show the ‘binding force’ between the two words varying with their separation. We are looking at how this metaphor might be used in order to describe language as a molecular structure, whose ‘inter-molecular forces’ can be related to part-of-speech interaction and the development of potential semantic categories for the unknown language.
So far we have mainly been working with English, but we have begun to look at

languages which represent their functional relationships by internal changes to words or by the addition of prefixes or suffixes. Although the process for separating into functional and content terms is more complex, we believe the fundamental results should be consistent. This will be one test of the theories presented above.

In general, we realise that testing our language detection algorithms will be a significant issue. We do not have examples that we know to be definitely from non-human, but intelligent origins, and we need to look extensively at signals of non-intelligent origin which may mimic some of the language characteristics described above. This will form a significant part of our future work and we welcome discussion and suggestions.

Conclusion

Language in its written format has proved to be a rich source for a variety of statistical analyses - some more conclusive than others - which when combined, give a comprehensive algorithm for identifying the presence of language-like systems. Analysis stages include compression, entropy profile, type-token distribution, word-length Zipfian analysis, finding a frequency distribution signature by successive chunking, stemming, cohesion analysis, phrase-length frequency distribution and pattern comparison across samples.

REFERENCES

Ally & Bacon, Cognitive Psychology, (third edition), Solso, Massachusetts, USA, 1991.

Baldi, P., & Brunak, S., Bioinformatics – The Machine Learning Approach, MIT press, Cambridge Massachusetts, 1998.

Chadwick, J. The Decipherment of Linear B, Cambridge University Press, 1967.

Charniak E., Statistical language learning Bradford/MIT Press, Cambridge. 1993.

Elliott, J & Atwell, E, Language in signals: the detection of generic species-independent intelligent language features in symbolic and oral communications, Proceedings of the 50th International Astronautical Congress, paper IAA-99-IAA.9.1.08, International Astronautical Federation, Paris, 1999.

Elliott, J & Atwell, E., Is anybody out there?: the detection of intelligent and generic language-like features, Journal of the British Interplanetary Society, Vol53 No 1 & 2.

Elliott, J, Decoding the Martian Chronicles, MSc project report, School of Computer Studies, University of Leeds 1999.

Hughes J & Atwell E., The automated evaluation of inferred word classifications in Proceedings of the European Conference on Artificial Intelligence (ECAI'94), pp550-554, John Wiley, Chichester. 1994.

Jenson, H. Self Organised Criticality, Cambridge University Press, 1998.

Minsky, M., Why Intelligent Aliens will be Intelligible, Cambridge University Press, 1984.

Norris, R,. How old is ET?, Proceedings of 50th International Astronautical Congress, paper IAA-99-IAA.9.1.04, International Astronautical Federation, Paris. 1999.

Zipf, G. K., Human Behaviour and The Principle of Least Effort, Addison Wesley Press, New York, 1949 (1965 reprint)