In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 25–30, Lisbon, Portugal, 2000.

# Increasing our Ignorance of Language: Identifying Language Structure in an Unknown 'Signal'

**John Elliott** and **Eric Atwell** and **Bill Whyte**
Centre for Computer Analysis of Language and Speech, School of Computer Studies
University of Leeds, Leeds, Yorkshire, LS2 9JT England
{jre, eric, billw}@scs.leeds.ac.uk

## 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. Outside our own immediate NLP sphere, generic communication techniques are of particular interest in the astronautical community, where two sessions are dedicated to SETI at their annual International conference with topics ranging from detecting ET technology to the ethics and logistics of message construction (Elliott and Atwell, 1999; Ollongren, 2000; Vakoch, 2000).

## 1 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? We need a 'language detector': or, to put it more accurately, something that separates language from non-language. But what is special about the language signal that separates it from non-language? Is it, indeed, separable?

The problem goal is to separate language from non-language without dialogue, and learn something about the structure of language in the passing. The language may not be human (animals, aliens, computers...), the perceptual space can be unknown, and we cannot assume human language structure but must begin somewhere. We need to approach the language signal from a naive viewpoint, in effect, increasing our ignorance and assuming as little as possible.

Given this standpoint, an informal description of 'language' might include that it:

- has structure at several interrelated levels
- is not random
- has grammar
- has letters/characters, words, phrases and sentences
- has parts of speech
- is recursive
- has a theme with variations
- is aperiodic but evolving
- is generative
- has transformation rules
- is designed for communication
- has Zipfian type-token distributions at several levels

Language as a 'signal'

- has some signalling elements (a 'script')
- has a hierarchy of signalling elements? ('Words', 'phrases' etc.)
- is serial?
- is correlated across a distance of several signalling elements applying at various levels in the hierarchy
- is usually not truly periodic
- is quasi-stationary?
- is non-ergodic?

We assume that a language-like signal will be encoded symbolically, i.e. with some kind of character-stream. Our language-detection algorithm for symbolic input uses a number of statistical clues such as entropy, "chunking" to find character bit-length and boundaries, and matching against a Zipfian type-token distribution for "letters" and "words".

## 2 Identifying Structure and the 'Character Set'

The initial task, given an incoming bit-stream, is to identify if a language-like structure exists and if detected what are the unique patterns/symbols, which constitute its 'character set'. A visualisation of the alternative possible byte-lengths is gleaned by plotting the entropy calculated for a range of possible byte-lengths (fig 1).

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 and Atwell, 2000; Elliott and Atwell, 1999). 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. Given the probability of a bit is Pi; the message entropy of a string of length N will be given by the first order measure:

$$E = SUM[P_i ln P_i]; i = 1, N$$

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 1, we see what happens when we applied this to samples of 8-bit ASCII text. 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.**

The next task, still below the stages normally tackled by NLL researchers, is to chunk the incoming character-stream into words. Looking at a range of (admittedly human language) text, if the text includes a space-like word-separator character, this will be the most frequent character. So, a plausible hypothesis would be that the most frequent character is a word-separator[1]; then plot type-token frequency distributions for words, and for word-lengths. If the distributions are Zipfian, and there are no significant 'outliers' (very large gaps between 'spaces' signifying very long words) then we have evidence corroborating our space hypothesis; this also corroborates our byte-length hypothesis, since the two are interdependent.

## 3 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 which character 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 2).

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 (1949), which predicts this behaviour on the grounds of minimum ef-

[1] Work is currently progressing on techniques for unsupervised word separation without spaces.

fort in a communication act. Conversely, results obtained similar to the 'flatter' distributions above, when using the most frequent character, is likely to indicate the absence of word separators in the signal.

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 3). Using statistical measures of significance, it was found that most groups fell well within 5- only two individual languages were near exceeding these limits - of the proposed Human language word-length profile (Elliott et al., 2000).

**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. Natural phenomena such as molecular distribution in yeast DNA possess characteristics of power laws (Jenson, 1998). Nevertheless, it is worth noting, that such non-language possessors of power law characteristics generally display distribution ranges far greater than language with long repeats far from each other (Baldi and Brunak, 1998); characteristics detectable at this level or at least higher order entropic evaluation.

## 4 Identifying 'Phrase-like' chunks

Having detected a signal which satisfies criteria indicating language-like structures at a physical level (Elliott and Atwell, 2000; Elliott and Atwell, 1999), second stage analysis is required to begin the process of identifying internal grammatical components, which constitute the basic building blocks of the symbol system. With the use of embedded clauses and phrases, humans are able to represent an expression or description, however complex, as a single component of another description. This allows us to build up complex structures far beyond our otherwise restrictive cognitive capabilities (Minsky, 1984). Without committing ourselves to a formal phrase structure approach, (in the Chomskian sense) or even to a less formal 'chunking' of language (Sparkle Project, 2000), it is this universal hierarchical structure, evident in all human languages and believed necessary for any advanced communicator, that constitutes the next phase in our signal analysis (Elliott and Atwell, 2000). It is from these 'discovered' basic syntactic units that analysis of behavioural trends and inter-relationships amongst terminals and non-terminals alike can begin to unlock the encoded internal grammatical structure and indicate candidate parts of speech. To do this, we make use of a particular feature common to many known languages, the 'function' words, which occur in corpora with approximately the same statistics. These tend to act as boundaries to fairly self-contained semantic/syntactic 'chunks.' They can be identified in corpora by their usually high frequency of occurrence and cross-corpora invariance, as opposed to 'content' words which are usually less frequent and much more context dependent.

Now suppose the function words arrived in a text independent of the other words, then they would have a Poisson distribution, with some long tails (distance between successive function words.) But this is NOT what happens. Instead, there is empirical evidence that function word separation is constrained to within short limits, with very few more than nine words apart (see fig 4). We conjecture that this is highly suggestive of chunking.

## 5 Clustering into syntactico-semantic classes

Unlike traditional natural language processing, a solution cannot be assisted using vast amounts of training data with well-documented 'legal' syntax and semantic interpretation or known statistical behaviour of speech categories. Therefore, at this stage we are endeavouring to extract the syntactic elements without a 'Rossetta' stone and by making as few assumptions as possible. Given this, a generic system is required to facilitate the analysis of behavioural trends amongst selected pairs of terminals and non-terminals alike, regardless of the target language.

Therefore, an intermediate research goal is to apply Natural Language Learning techniques to the identification of "higher-level" lexical and grammatical patterns and structure in a linguistic signal. We have begun the development of tools to visualise the correlation profiles between pairs of words or parts of speech, as a precursor to deducing general principles for 'typing' and clustering into syntactico-semantic lexical classes. Linguists have long known that collocation and combinational patterns are characteristic features of natural languages, which set them apart (Sinclair, 1991). Speech and language technology researchers have used word-bigram and n-gram models in speech recognition, and variants of PoS-bigram models for Part-of-Speech tagging. In general, these models focus on immediate neighbouring words, but pairs of words may have bonds despite separation by intervening words; this is more relevant in semantic analysis, eg Wilson and Rayson (1993), Demetriou (1997). We sought to investigate possible bonding between type tokens (i.e., pairs of words or between parts of speech tags) at a range of separations, by mapping the *correlation profile* between a pair of words or tags. This can be computed for given word-pair type (w1,w2) by recording each word-pair token (w1,w2,d) in a corpus, where d is the distance or number of intervening words. The distribution of these word-pair tokens can be visualised by plotting d (distance between w1 and w2) against frequency (how many (w1,w2,d) tokens found at this distance). Distance can be negative, meaning that w2 occurred *before* w1 and for any size window (i.e., 2 to n). In other words, we postulate that it might be possible to deduce part-of-speech membership and, indeed, identify a set of part-of-speech classes, using the joint probability of words themselves. But is this possible? One test would be to take an already tagged corpus and see if the parts-of-speech did indeed fall into separable clusters.

Using a five thousand-word extract from the LOB corpus (Johansson et al., 1986) to test this tool, a number of parts-of-speech pairings were analysed for their cohesive profiles. The arbitrary figure of five thousand was chosen, as it both represents a sample large enough to reflect trends seen in samples much larger (without loosing any valuable data) and a sample size, which we see as at least plausible when analysing ancient or extra-terrestrial languages where data is at a premium.

Figure 5 shows the results for the relationship between a pair of content and function words, so identified by looking at their cross-corpus statistics. It can be seen that the function word has a 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.

Examining language in such a manner also lends itself to summarising ('compressing') the behaviour to its more notable features when forming profiles. Figure 6 depicts a 3D representation of results obtained from profiling VB-tags with six other major syntactic categories; figure 7 shows the main syntactic behavioural features found for the co-occurrence of some of the major syntactic classes ranging over the chosen window of ten words.

Such a tool may also be useful in other areas, such a lexico-grammatical analysis or tagging of corpora. Data-oriented approaches to corpus annotation use statistical n-grams and/or constraint-based models; n-grams or constraints with wider windows can improve error-rates, by examining the topology of the annotation-combination space. Such information could be used to guide development of Constraint Grammars. The English Constraint Grammar described in (1995) includes constraint rules up to 4 words either side of the current word (see Table 16, p352); the peaks and troughs in the visualisation tool might be used to find candidate patterns for such long-distance constraints.

Our research topic NLL4SETI (Natural Language Learning for the Search for Extra-Terrestrial Intelligence) is distinctive in that - it is potentially a VERY useful application of unsupervised NLL; - it starts from more basic assumptions than most NLL research: we do not assume tokenisation into characters and words, and have no tagged/parsed training corpus; - it focuses on utilising statistical distributional universals of language which are computable and diagnostic; - this focus has led us to develop distributional visualisation tools to explore type/token combination distributions; - the goal is NOT learning algorithms which analyse/annotate human language in a way which human experts would approve of (eg phrase-chunking corresponding to a human linguist's parsing of English text); but algorithms which recognise language-like structuring in a potentially much wider range of digital data sets.

## 6 Summary and future developments

To summarise, our achievements to date include - a method for splitting a binary digit-stream into characters, by using entropy to diagnose byte-length; - a method for tokenising unknown character-streams into words of language; - an approach to chunking words into phrase-like sub-sequences, by assuming high-frequency function words act as phrase-delimiters; - a visualisation tool for exploring word-combination patterns, where word-pairs need not be immediate neighbours but characteristically combine despite several intervening words.

So far, our approaches have involved working with languages with which we are most familiar and, to a certain extent, making use of linguistic 'knowns' such as pre-tagged corpora. It is early days yet and we make no apology for this initial approach. However, we feel that by deliberately reducing our dependence on prior knowledge ('increasing our ignorance of language') and by treating language as a 'signal', we might be contributing a novel approach to natural language processing which might ultimately lead to a better, more fundamental understanding of what distinguishes language from the rest of the signal universe.

## References


P. Baldi and S. Brunak. 1998. *Bioinformatics - The Machine Learning Approach.* MIT press, Cambridge Massachusetts.

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

George Demetriou. 1997. *PhD thesis.* School of Computer Studies, University of Leeds.

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

John Elliott and Eric Atwell. 2000. Is there anybody out there?: The detection of intelligent and generic language-like features. *Journal of the British Interplanetary Society*, 53:1/2:13–22.

John Elliott, Eric Atwell, and Bill Whyte. 2000. Language identification in unknown signals. In *Proceedings of COLING'2000 International Conference on Computational Linguistics.* Saarbruecken.

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

Stig Johansson, Eric Atwell, Roger Garside, and Geoffrey Leech. 1986. *The Tagged LOB corpus: users' manual.* Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities. Available from http://www.hit.uib.no/icame/lobman/lob-cont.html.

Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text.* Berlin: Mouton de Gruyter.

Geoffrey Leech, Roger Garside, and Eric Atwell. 1983. The automatic grammatical tagging of the lob corpus. *ICAME Journal*, 7:13–33.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

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

Alexander Ollongren. 2000. Large-size message construction for eti. In *Proceedings of the 50th International Astronautical Congress.* paper IAA-99-IAA.9.1.09, International Astronautical Federation, Paris.

Sparkle Project. 2000. http://www.ilc.pi.cnr.it/sparkle/wp1-prefinal/node25.html.

John Sinclair. 1991. *Corpus, concordance, collocation describing English language.* Oxford University Press.

Doug Vakoch. 2000. Communicating scientifically formulated spiritual principles in interstellar messages. In *Proceedings of the 50th International Astronautical Congress.* paper IAA-99-IAA.9.1.10, International Astronautical Federation, Paris.

Andrew Wilson and Paul Rayson. 1993. The automatic content analysis of spoken discourse. In C. Souter and E. Atwell, editors, *Corpus based computational linguistics.* Rodopi, Amsterdam.

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