

**THE FILTRATION OF INTER-GALACTIC OBJETS TROUVÉS
AND THE IDENTIFICATION OF THE LINGUA EX MACHINA HIERARCHY**

John Elliott, jre@comp.leeds.ac.uk

Eric Atwell, eric@comp.leeds.ac.uk, Bill Whyte, billw@comp.leeds.ac.uk

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

The ultimate aim of any system designed to analyse a signal for structured and intelligent-like behaviour, is not merely to provide a Boolean filter but to have the ability to deduce what category the signal fits – language, image, music noise etc – and then to decipher its contents. Up until now, my research has concentrated on the lower level universals of language structure and how they can be detected^{1,2}. However, in this paper, I begin to look into developing techniques, which use only surface structure information deduced from the signal sample itself, for learning what underlies the higher level elements of the language (lingua ex machina) hierarchy, where syntax meets semantics. The aim therefore is to move towards developing techniques, which will unlock a signal's meaning using its internal cohesion and constraints. In doing so, I look at how the most disparate of human language orthographic forms mirror each other's underlying structure, despite their encoding strategies and how classes of words give their secrets away by the functional 'friends' they keep.

To address such linguistic problems, which reach down to and beyond the very core fundamental concepts of what we currently understand as language, I have had to develop new techniques to provide ways of learning and understanding syntactic behaviour from surface structure alone. Although each word or symbol is arbitrarily paired with its meaning, empirical evidence has shown that the semantic class or part-of-speech to which it belongs is constrained by its behaviour and according to underlying interactive rules. By using functional elements/words detected from lower level analysis and visualisation techniques for bonding, I have begun to develop algorithms, which group a communication's word classes, using the notion of polyposicⁱ and uniposicⁱⁱ behaviour within minimal constrained pairings. Initial trials with English confirm such behaviour and show how these semantic classes can be detected and words attributed their membership, without prior knowledge, using unsupervised positional rule-based criteria and statistical variance.

Finally, I present techniques developed for non-language structured signals:

- How the statistical 'fingerprint' of a binary image, such as the recent challenge set by Dr Dutil³ and the Arecibo transmission,⁴ can be detected when encoded as a binary bit-stream by the analysis of its type-token distribution.
- Further analysis of the structure of music and how it differs from language.

References

1. Elliott J, Atwell E & Whyte, W. 2001. First Stage Identification of Syntactic Elements in an Extraterrestrial Signal in: Proceedings of IAC'2001: the 52nd International Astronautical Congress, paper IAA-01-IAA.9.2.07, Toulouse, France.
2. Elliott, John & Atwell, Eric, 1999. 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.
3. Dutil, Y, 2001. University of Catalunya, Barcelone, Spain & ABB Bomem Inc,
4. Arecibo Observatory. <http://www.naic.edu/>.

ⁱ Homonyms with multiple part-of-speech membership

ⁱⁱ Homonyms with membership restricted to only one part-of-speech