This is a repository copy of *A corpus for interstellar communication*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/82306/

---

**Proceedings Paper:**

---

**Reuse**
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher’s website.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
A corpus for interstellar communication

Eric Atwell eric@comp.leeds.ac.uk http://www.comp.leeds.ac.uk/eric/
John Elliott jre@comp.leeds.ac.uk http://www.comp.leeds.ac.uk/jre/

Centre for Computer Analysis of Language And Speech,
School of Computing,
University of Leeds,

1. Introduction: SETI, the Search for Extra-Terrestrial Intelligence

Many researchers in Astronomy and Astronautics believe the Search for Extra-Terrestrial Intelligence is a serious academic enterprise, worthy of scholarly research and publication (e.g. Burke-Ward 2000, Couper and Henbest 1998, Day 1998, McDonough 1987, Sivier 2000, Norris 1999), and large-scale research sponsorship attracted by the SETI Institute in California. Most of this research community is focussed on techniques for detection of possible incoming signals from extra-terrestrial intelligent sources (e.g. Turnbull et al 1999), and algorithms for analysis of these signals to identify intelligent language-like characteristics (e.g. Elliott and Atwell 1999, 2000).

However, recently debate has turned to the nature of our response, should a signal arrive and be detected. For example, the 50th International Astronautical Congress devoted a full afternoon session to the question of whether and how we should respond to an initial message identified to be of extra-terrestrial origin. Interestingly, we (the authors of this paper) were the only corpus linguists present at this session: the Congress seemed to assume that the design of potentially the most significant communicative act in history should be decided by astrophysicists. We believe that others should be aware of and contribute to what is effectively a corpus design project; and that the Corpus Linguistics research community has a particularly significant contribution to make.

2. Past ideas on how to signal our existence to extra-terrestrials

Speculations about how to signal our existence to extraterrestrials began at least a century ago. Early ideas focussed on pictorial messages, transmitted visually by drawing over very large expanses of the Earth’s surface. “For example, the Pythagorean theorem could be illustrated visually during the daytime by clearing vast expanses of forest in Siberia to show the areas surrounding a right-angled triangle. Or during the night, canals dug into the Sahara desert in the shape of a circle could be filled with kerosene; when lit, the flames would provide a pictorial signal of our existence.” (Vakoch 1998a).

More recently, the Pioneer and Voyager spacecraft, sent to explore planets in our solar system but then left to drift out into interstellar space, carried messages to any extraterrestrials who might intercept them in their travels beyond the solar system. On the Pioneer plaque, an outline of the Pioneer spacecraft is seen behind figures of two humans. At the bottom of the plaque, the same spacecraft is shown in a smaller scale as it passes through the solar system on its journey from Earth. A diagram of fifteen converging lines shows the Earth’s location in time and space in relation to prominent pulsars. (Sagan et al 1972, Vakoch 1998a). The Voyager spacecraft each bear similar diagrams, and in addition a record (with player and encoded instructions on how to play) illustrating basics of human knowledge of mathematics and physics, and a wide variety of pictures of our world. (Sagan 1978, Vakoch 1998a).

There have also been attempts to deliberately transmit messages from the Earth’s surface. Most notably, in 1974 astronomers at the Arecibo radio-telescope in Puerto Rico sent a signal of 1,679 radio-wave pulses to M13, a star-cluster 25,000 light-years away. 1679 is the product of two prime numbers, 23 and 73; arranging the pulses into a rectangle of 23 columns by 73 rows creates a pictogram showing a radio-dish, a human, and some basic scientific information. (Couper and Henbest 1998, Vakoch 1998a).
3. Current SETI ideas on message construction

The Arecibo experiment was a deliberate attempt at message transmission. Humanity has been transmitting radio signals on a much larger scale for decades, since radio transmissions intended for terrestrial reception are also beamed into outer space; thus an extra-terrestrial first encounter with human culture may well be through accidental reception of television and radio broadcasts, as foreseen in the novel and subsequent film Contact (Sagan 1988). Reception of such “unintended” messages may prompt Extra-Terrestrials to initiate first contact; but many in the SETI research community (e.g. Vakoch 1999) feel it is important to plan a more deliberately designed, well-thought-out response message. (Vakoch 1998b) argues for “… the need for more intensive investigations of the linguistic aspects of SETI before a message is received”. (Vakoch 1998c, p705) also identifies several benefits of beginning work on construction of a reply message immediately, even before an incoming extraterrestrial message has been received and recognised:

“(1) concretely understanding the challenge of creating an adequate reply; (2) helping decode messages from extraterrestrials; (3) creating interstellar compositions as a new form of art; (4) having a reply ready in case we receive a message; (5) providing a sense of concrete accomplishment; (6) preparing for an active search strategy; and (7) gaining public support for SETI.”

In 1974 a signal of 1,679 bits was considered potentially significant and challenging to technology of the time, e.g. it took three minutes to transmit; a quarter of a century later, we are used to processing messages of megabytes, gigabytes, or bigger in terrestrial communication networks such as the Internet. It is clear that we could look beyond a single pictogram or collection of diagrams, to design a much larger Corpus of data to represent humanity. (Vakoch 1998c) advocates that the message constructed to transmit to extraterrestrials should include a broad, representative collection of perspectives rather than a single viewpoint or genre; this would strike a chord with Corpus Linguists for whom a central principle is that a corpus must be “balanced” to be representative.

The consensus at the 50th International Astronautical Congress seemed to be to transmit an encyclopaedia summarising human knowledge, such as the Encyclopaedia Britannica, to give ET communicators an overview and “training set” key to analysis of subsequent messages. Furthermore, this should be sent in several versions in parallel: the text; page-images, to include illustrations left out of the text-file; and perhaps some sort of abstract linguistic representation of the text, using a functional or logic language (Öllongren 1999, Freudenthal 1960).

4. Enriching the message corpus with multi-level linguistic annotations

The idea of “enriching” the message corpus with annotations at several levels should also strike a chord with Corpus Linguists. Natural language exhibits highly complex multi-layered sequencing, structural and functional patterns, as difficult to model as sequences and structures found in more traditional physical and biological sciences. Corpus Linguists have long known this, on the basis of evidence such as the following:

- Language datastreams exhibit structural patterns at several interdependent linguistics levels, including: phonetic and graphemic transcription, prosodic markup, part-of-speech wordclasses, collocations, phraseological and collegational patterns, semantic word-sense classification, syntax or grammatical phrase structure, functional dependency structure, semantic predicate structure, pragmatic references, discourse or dialogue structure, communication act or speech act patterns.

- Even within one such linguistic level, structural analysis is complex, with further interdependent sublevels. For example, the European Expert Advisory Group on Language Engineering Standards (EAGLES) report on parsing annotations (Leech et al 1996) recognises at least 7 separate yet interdependent sublayers of grammatical analysis which a full parser should aim to recognise; yet none of the state-of-the-art parsers evaluated in (Atwell 1996, Atwell et al 2000a) were capable of providing all 7 layers of analysis in their output. Different parsers analysed different subsets of these sublayers of grammatical information, making cross-parser comparisons and performance evaluations difficult if not meaningless.
Furthermore, linguistic analysis at one level may depend on or require other levels of linguistic information; for example, (Demetriou and Atwell 2001) demonstrated that lexical-semantic word-tagging subsumes or combines several knowledge sources including thesaurus class, semantic field, collocation preferences, and dictionary definition.

Some corpora have been annotated with several layers or levels of linguistic knowledge in parallel; for example, the SEC corpus (Taylor and Knowles 1988) has speech recordings, transcriptions, prosody markup, PoS-tags, parse-trees; the ISLE corpus (Menzel et al 2000, Herron et al 1999, Atwell et al 2000b) has language-learner speech recordings, transcriptions, corrections, prosody, expert evaluations. Other annotations can be added automatically by software, e.g. semantic tags (Demetriou and Atwell 2001), ENGCG Constraint Grammar dependency structures (Karlsson et al 1995, Voutilainen et al 1996).

5. Natural language learning

In the 1980s, most NLP researchers used their ‘expert intuitions’ to guide development of large-scale grammars; a language model was essentially an ‘expert system’ encoding the knowledge of a human linguistics expert. This kind of knowledge model was harder to ‘scale up’ to cover more and more language data, and it relied on existing expert knowledge. More recently, this has given way to the use of corpora or large text samples, some of which are annotated or ‘tagged’ with expert analyses. Tagged and parsed corpora can be used by linguists as a testbed to guide their development of grammars (see, for example Souter and Atwell 1994); and they can be used to train Natural Language Learning or data-mining models of complex sequence data. Several initiatives are under way to collect language datasets for language modelling research, for example, ICAME, the International Computer Archive of Modern and medieval English (based in Bergen); ELRA, the European Language Resources Association (based in Paris); LDC, the Linguistic Data Consortium (based at the University of Pennsylvania).

A growing number of NLP researchers are looking into ways to utilise these new training-set resources: the Association for Computational Linguistics has established a Special Interest Group in Natural Language Learning (machine-learning of language sequence-patterns from corpus data) which holds annual conferences, e.g. CoNLL’2000. Given appropriate annotated Corpus data, many NLP problems can be generalised to “mappings” between linguistic levels of analysis, for example:

- **Word-class identification** (mapping words into syntactic/semantic sets or classes), e.g. (Atwell and Drakos 1987, Hughes 1993, Finch 1993, Hughes and Atwell 1994, Teahan 1998)


- **Machine Translation** (mapping a source-language word sequence onto a target-language word-sequence), e.g. (Brown et al 1990, Berger et al 1994, Gale and Church 1993)

Researchers have tried casting these NLP mapping subtasks in terms of Natural Language Learning models, such as Hidden Markov Models (HMMs), Stochastic Context Free Grammar (SCFG) parsers, Data-Oriented Parsing (DOP) models. The complex patterns found in language data call for sophisticated stochastic modelling. For example, Hidden Markov Models have become widely used in Language Engineering applications because they are well-understood and computationally tractable (e.g. Young and Bloothooft 1997, Manning and Schutze 1999, Jurafsky and Martin 2000, Huang 1990, MacDonald 1997, Elliott et al 1995, Woodward 1997). Although (Chomsky 1957) famously demonstrated that a finite-state model is a theoretically inadequate approximation for certain aspects of language modelling, Language Engineers have come to realise that HMMs can be adapted to work most of the time, and that the theoretically problematic cases alluded to by Chomsky are infrequent enough in “real” applications to be ignored in practice. Language Engineering researchers have been searching for higher-level models which effectively extend Hidden Markov Models in limited ways without extending the computational cost prohibitively, for example higher-order Markov models, limited stochastic context-free grammars, hybrid statistical/knowledge-based models. Linguists have found ‘Universal’ features which appear to be common to and characteristic of all human languages, (e.g. Zipf 1935, 1949); but few of these have been stated in terms of or related to stochastic models.

We know how to extract low-level linguistic patterns from raw text using unsupervised learning algorithms (e.g. Atwell and Drakos 1987, Hughes 1993, Finch 1993, Hughes and Atwell 1994, Elliott and Atwell 1999, 2000, Elliott et al 2000a,b, 2001, Manning and Schutze 1999, Jurafsky and Martin 2000); a “Rosetta Stone” key to English, annotated with rich linguistic analyses, should help ET communicators map between symbols and meanings using supervised as well as unsupervised learning algorithms.

6. A corpus linguistics SETI advisory panel

Astronomers have not sought to consult Corpus Linguists on the design of this Corpus for Interstellar Communication; but we can and should make an informed contribution. The parallel corpus and multi-annotated corpus are not new concepts to Corpus Linguistics. We have a range of standards and tools for design and annotation of representative corpus resources. Furthermore, we know which analysis schemes are more amenable to supervised learning algorithms; for example, the BNC tagging scheme and the ICE-GB parsing scheme have been demonstrated to be machine-learnable in a tagger and parser respectively.

An Advisory Panel of corpus linguists could design and implement an extended Multi-annotated Interstellar Corpus of English. The following are ideas for the Advisory Panel to consider:

- augment the Encyclopaedia Britannica with a collection of samples representing the diversity of language in real use. Candidates include the LOB and/or BNC corpus;
- as an additional “key”, transmit a dictionary aimed at language learners which has also been a rich source for NLP learning (e.g. Demetriou and Atwell 2001); a good candidate would be LDOCE, the Longman Dictionary of Contemporary English, which uses the Longman Defining Vocabulary;
- supply our ET communicators with several levels of linguistic annotation, to give them a richer training set for their natural language learning attempts. We suggest that initial (i) raw text and (ii) page-images should be augmented with some or all of (iii) XML markup, (iv) PoS-tagging, (v) phrase structure parses, (vi) dependency structure analyses, (vii) coreference markup, (viii) dialogue act markup, (ix) semantic analyses.
- Add translations of the English text into other human languages; although the International Astronautical Congress seemed to assume Humanity should be represented by English, multilingual annotations may actually be useful in natural language learning algorithms.

This calls for a large-scale corpus annotation project, which may not seem immediately justifiable to computational linguistics research funding bodies such as the UK Engineering and Physical Sciences Research Council (EPSRC). However, the International Astronautical Congress also discussed plans to
proactively make interstellar contact using existing astronomical technology, by firing a satellite-based laser cannon at a range of nearby (in astronomical terms) potentially suitable targets. If this succeeds and we receive a message back, the need for our Interstellar Corpus Advisory Panel becomes more urgent.


7. References


Chomsky N 1957 Syntactic structures. The Hague, Mouton.


Jurafsky D, Martin J 2000 *Speech and language processing*. Prentice-Hall


