IAA-02-IAA.9.2

CORPUS LINGUISTICS AND THE DESIGN OF A RESPONSE MESSAGE

Atwell, E. eric@comp.leeds.ac.uk
Elliott, J. jre@comp.leeds.ac.uk
University of Leeds, School of Computing, Leeds LS2 9JT, England
Phone: +(44) 113-2335430; Fax: +(44) 113-2335468

Most research related to SETI, the Search for Extra-Terrestrial Intelligence, 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, another issue for research and debate is the nature of our response, should a signal arrive and be detected. The design of potentially the most significant communicative act in history should not be decided solely by astrophysicists; the Corpus Linguistics research community has a contribution to make to what is essentially a Corpus design and implementation project. (Vakoch 1998) advocated that the message constructed to transmit to extraterrestrials should include a broad, representative collection of perspectives rather than a single viewpoint or genre; this should strike a chord with Corpus Linguists for whom a central principle is that a corpus must be “balanced” to be representative (Meyer 2001).

One idea favoured by SETI researchers is 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 (Ollongren 1999, Freudenthal 1960). The idea of “enriching” the message corpus with annotations at several levels should also strike a chord with Corpus Linguists who have long known that Natural language exhibits highly complex multi-layering sequencing, structural and functional patterns, as difficult to model as sequences and structures found in more traditional physical and biological sciences. Some corpora have been annotated with several levels or layers of linguistic knowledge, for example the SEC corpus (Taylor and Knowles 1988), the ISLE corpus (Menzel et al. 2000). Tagged and parsed corpus can be used by corpus linguists as a testbed to guide their development of grammars (e.g. Souter and Atwell 1994); and they can be used to train Natural Language Learning or data-mining models of complex sequence data (e.g. Brill 1993, Hughes 1993, Atwell 1996). Corpus linguists have a range of standards and tools for design and annotation of representative corpus resources, and experience of which annotation types are more amenable to Natural Language Learning algorithms. An Advisory panel of corpus linguists could help design and implement an extended Multi-annotated Interstellar Corpus of English, incorporating ideas from Corpus Linguistics such as:

- 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 (Johannson et al. 1986, Leech 1993);

- 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, a small set of “semantic primitives” to define all other words;

- 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) SML markup, (iv) PoS-tagging, (v) phrase structure parses, (vi) dependency structure analyses, (vii) co-reference markup, (viii) dialogue act markup, (ix) semantic analyses, (x) proper noun markup.

- Add translations of the English text into other human languages: Humanity should not be represented by English alone, and multilingual annotations may actually be useful in natural language learning algorithms.

This calls for a large-scale corpus annotation project, requiring an Interstellar Corpus Advisory Panel, analogous to the BNC or MATE advisory panels, to include experts in English grammar and semantics, English language learning, computational Natural language Learning algorithms, and corpus design, implementation, annotation, standardisation, and analysis.