Proceedings Paper:
Atwell, ES and Elliott, JR (2001) Corpus linguistics and the design of a response message. In: Proceedings of IAC'2001: the 52nd International Astronautical Congress. IAC'2001: the 52nd International Astronautical Congress, 01-05 Oct 2001, Toulouse, France. International Astronautical Federation.

Paper Number IAA-01-IAA.9.2.08

CORPUS LINGUISTICS AND THE DESIGN OF A RESPONSE MESSAGE

E. Atwell and J. Elliott

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

52nd International Astronautical Congress
1-5 Oct 2001/Toulouse, France

CORPUS LINGUISTICS AND THE DESIGN OF A RESPONSE MESSAGE

Paper Number IAA-99-IAA.9.1.08

Eric Atwell, eric@comp.leeds.ac.uk and John Elliott, jre@comp.leeds.ac.uk
Centre for Computer Analysis of Language And Speech,
School of Computing, University of Leeds, Leeds, Yorkshire, LS2 9JT England

ABSTRACT

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, and algorithms for analysis of these signals to identify intelligent language-like characteristics. 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. 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. 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-layered sequencing, structural and functional patterns; some corpora have been annotated with several levels or layers of linguistic knowledge. Tagged and parsed corpora can be used by corpus linguists as a testbed to guide their development of grammars; and they can be used to train Natural Language Learning or data-mining models of complex sequence data. 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, such as 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, such as LDOCE, the Longman Dictionary of Contemporary English, which uses 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; 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.

INTRODUCTION

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.

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).

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 should 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 (Ollongren 1999, Freudenthal 1960).

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).

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);

Part-of-Speech word-tagging
mapping word-sequences onto wordclass-tag sequences, e.g. (Leech et al 1983, Atwell 1983, Eeg-Olofsson 1991, Brill 1993, Atwell et al 1984, 2000a);

Parsing: Sentence-structure analysis
mapping word- and/or word-class sequences onto parses, e.g. (Sampson et al 1989, Atwell 1987, 1988, 1993, Black et al 1993, Bod 1993, Briscoe 1994, Jelinek et al 1992, Joshi and Srinivas 1994, Magerman 1994, O'Donoghue 1993, Schabes, Roth and Osborne 1993, Sekine and Grishman 1995)

Lexical semantics or word-sense tagging
mapping word-sequences onto semantic tags or meaning-analyses, e.g. (Demetriou 1993, Demetriou and Atwell 1994, 2001, Bod et al 1996, Kuhn and de Mori 1994,

Weischedel et al 1993, Wilson and Rayson 1993, Wilson and Leech 1993, Jost and Atwell 1993)

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)

Speech Recognition
mapping a speech signal onto a phonetic and graphemic transcription word-sequence, e.g. (Demetriou and Atwell 1994, Giachin 1995, Jelinek 1991, Kneser and Ney 1995, Yamron 1994, Young and Bloothooft 1997).

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.

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 should include experts in lexis, grammar and semantics of English and other natural languages, English language learning and teaching, and language corpus design, implementation, annotation, standardisation, and analysis.

Expert Advisory Panels or Steering Groups are common practice in computational linguistic research projects, to advise on research ideas and techniques and monitor progress; examples of large projects which have benefited from Advisory Panels include BNC (British National Corpus), MATE (Multi-level Annotation Tools Engineering), and EAGLES (Expert Advisory Group on

Language Engineering Standards). An Advisory Panel of corpus linguists could design and implement an extended Multi-annotated Interstellar Corpus of English. This Interstellar Corpus Advisory Panel should bring corpus linguists into contact with "mainstream" SETI researchers.

ENRICHING THE RESPONSE MESSAGE WITH LINGUISTIC ANNOTATIONS

The following are ideas for the Advisory Panel to consider:

Enrich with Corpora

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;

Enrich with learners' dictionaries

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;

Enrich with multi-level linguistic annotation

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 of format, (iv) PoS-tagging, (v) phrase structure parses, (vi) dependency structure analyses, (vii) coreference markup, (viii) dialogue act markup, (ix) semantic analyses.

Enrich with multilingual translations

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.

The resultant enriched message corpus would not only be more readily understood by alien contacts, but it would also be a rich research resource for computational linguists here on Earth.

ACKNOWLEDGEMENT

This paper is a revision of an earlier paper presented to a Corpus Linguistics research audience: Atwell E and Elliott J. 2001, *A corpus for interstellar communication*, in Rayson P, Wilson A, McEnery T, Hardie A, and Khoja S (editors) **Proceedings of CL2001: International Conference on Corpus Linguistics**, University Centre for computer corpus Research on Language (UCREL) Technical Paper 13, Lancaster University, Lancaster, UK.

REFERENCES

Atwell E 1983 Constituent Likelihood Grammar. *ICAME Journal* 7: 34-65.

Atwell E 1987 A parsing expert system which learns from corpus analysis. In Meijs W (ed) *Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference on English Language Research on Computerised Corpora*, Amsterdam, Rodopi, pp227-235.

Atwell E 1988 Transforming a Parsed Corpus into a Corpus Parser. In Kyto M, Ihalainen O, Risanen M (eds) *Corpus Linguistics, Hard and Soft: Proceedings of the ICAME 8th International Conference on English Language Research on Computerised Corpora*. Amsterdam, Rodopi, pp61-70.

Atwell E 1993 Corpus-based statistical modelling of English grammar. In Souter C, Atwell E (eds), *Corpus-based computational linguistics: proceedings of the 12th conference of the International Computer Archive of Modern English.* Amsterdam, Rodopi, pp195-214.

Atwell E 1993 Linguistic Constraints for Large-Vocabulary Speech Recognition. In Atwell E (ed), *Knowledge at Work in*

Universities: Proceedings of the second annual conference of the Higher Education unding Council's Knowledge Based Systems Initiative. Leeds University Press, pp26-32.

Atwell E 1996 Machine Learning from corpus resources for speech and handwriting recognition. In Thomas J, Short M (eds), *Using corpora for language research: studies in the honour of Geoffrey Leech*, Harlow, Longman, pp151-166.

Atwell E 1996 Comparative Evaluation of Grammatical Annotation Models. In Sutcliffe R, Koch H, McElligott A (eds), *Industrial Parsing of Software Manuals*. Amsterdam, Rodopi.

Atwell E, Leech G, Garside R 1984 Analysis of the LOB Corpus: progress and prospects. In Aarts J, Meijs W (eds) *Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research*, Amsterdam, Rodopi, pp40-52.

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, Wilcock S 2000 A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24: 7-23.

Atwell E, Howarth P, Souter C, Baldo P, Bisiani R, Bonaventura P, Herron D, Menzel W, Morton R, Wick H 2000 User-Guided System Development in Interactive Spoken Language Education. *Natural Language Engineering*, 6(3): 188-202.

Atwell E, Drakos N 1987 Pattern recognition applied to the acquisition of a grammatical classification system from unrestricted English text. In Maegaard B (ed), *Proceedings of the Third Conference of European Chapter of the Association for Computational Linguistics*. New Jersey, Association for Computational Linguistics, pp56-63.

Atwell E, Hughes J, Souter C 1995 Automatic extraction of tagset mappings from Parallel-Annotated Corpora. In Tzoukermann E, Armstrong S (eds) *From text to tags - issues in multilingual language analysis: Proceedings of EACL-SIGDAT*. Dublin, Association for Computational Linguistics, pp10-17.

Berger A, Brown P, Cocke J, Pietra S, Pietra V, Gillett J, Lafferty J, Mercer R, Printz H, Ures L 1994 The Candide system for Machine Translation. In *Proceedings of the ARPA workshop on Human Language Technology*. San Mateo, Morgan Kaufmann, pp152-157.

Black E, Garside R, Leech G (eds) 1993 *Statistically-driven computer grammars of English: the IBM/Lancaster approach*. Amsterdam, Rodopi.

Bod R 1993 Using an annotated corpus as a stochastic grammar. In *Proceedings of the 6th EACL.* Utrecht, Association for Computational Linguistics.

Bod R, Bonnema R, Scha R 1996 A data-oriented approach to semantic interpretation. *In Proceedings of the ECAI'96 workshop on corpus-oriented semantic analysis.* Budapest, ECAI.

Brill E 1993 *A Corpus-based approach to language learning.* PhD thesis, University of Pennsylvania.

Briscoe E 1994 Prospects for practical parsing of unrestricted text: robust statistical parsing techniques. In Oostdijk N, de Haan P (eds) *Corpus-based research into language*. Amsterdam, Rodopi.

Brown P, Cocke J, Pietra S, Pietra V, Jelinek F, Lafferty J, Mercer R, Roossin P 1990 A statistical approach to Machine Translation. *Computational Linguistics* 16(2): 79-85.

Burke-Ward R 2000 Possible existence of extra-terrestrial technology in the solar system. . *Journal of the British Interplanetary Society* 53(1&2): 1-12.

Chomsky N 1957 *Syntactic Structures*. The Hague, Mouton.

Couper H, Henbest N 1998 *Is anybody out there?* London, Dorling Kindersley.

Day P (ed) 1998 *The search for extraterrestrial life.* Oxford, Oxford University Press and the Royal Institution.

Demetriou G 1993 Lexical disambiguation using CHIP. In *Proceedings of the 6th EACL*. Utrecht, Association for Computational Linguistics, pp431-436.

Demetriou G, Atwell E 1994 Machine-learnable, non-compositional semantics for domain independent speech or text recognition. In *Proceedings of the 2nd Hellenic-European Conference on Mathematics and Informatics (HERMIS)*, Athens.

Eeg-Olofsson M 1991 *Word-class tagging: Some computational tools*. PhD thesis, University of Lund.

Elliott R, Lakhdar A, Aggoun J, Moore R 1995 *Hidden Markov models: estimation and control*. London, Springer-Verlag.

Elliott J, Atwell E 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*. Amsterdam, paper IAA-99-IAA.9.1.08.

Elliott J, Atwell E 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.

Elliott J, Atwell E, Whyte B 2000 Language identification in unknown signals. In *Proceedings of COLING'2000 18th International Conference on Computational Linguistics*, Saarbrucken, Association for Computational Linguistics (ACL) and San Francisco, Morgan Kaufmann Publishers, pp1021-1026.

Elliott J, Atwell E, Whyte B 2000 Increasing our ignorance of language: identifying language structure in an unknown signal. In Daelemans W (ed) *Proceedings of CoNLL-2000: International Conference on Computational Natural Language Learning*. Lisbon, Association for Computational Linguistics.

Elliott J, Atwell E, Whyte B 2001 A toolkit for visualisation of combinational constraint phenomena in linguistically interpreted corpora. In *Proceedings of CLUK'4: Computational Linguistics in the United Kingdom*. Sheffield.

Finch S 1993 *Finding structure in language.* PhD thesis, Edinburgh University.

Freudenthal H 1960 *LINCOS, Design of a language for cosmic intercourse*. North Holland.

Gale W, Church K, 1993 A program for aligning sentences in bilingual corpora. Computational Linguistics 19(1): 75-102.

Giachin E 1995 Phrase bigrams for continuous speech recognition. In *Proceedings of ICASSP'95*. Detroit.

Herron D, Menzel W, Atwell E, Bisiani R, Daneluzzi F, Morton R, Schmidt J 1999 Automatic localization and diagnosis of pronunciation errors for second language learners of English. In *Proceedings of EUROSPEECH'99: 6th European Conference on Speech Communication and Technology*. Budapest.

Huang X 1990 *Hidden Markov models for speech recognition*. Edinburgh, Edinburgh University Press.

Hughes J 1993 *Automatically acquiring a classification of words* PhD thesis, University of Leeds.

Hughes J, Atwell E 1994 The automated evaluation of inferred word classifications. In Cohn A (ed) *Proceedings of ECAI'94: 11th European Conference on Artificial Intelligence*. Chichester, John Wiley, pp535-539.

Jelinek F 1991 Self-organised language modeling for speech recognition. In Waibel A, Lee K (eds), *Readings in speech* *recognition*. San Mateo, Morgan Kaufmann, pp450-506.

Jelinek F, Lafferty J, Mercer R 1992 Basic methods of probabilistic context-free grammars. In Laface P, de Mori R (eds) *Speech recognition and understanding*. Berlin, Springer-Verlag, pp347-360.

Joshi A, Srinivas B 1994 Disambiguation of Super Parts of Speech (or Supertags): almost parsing. In *Proceedings of COLING'94*. Kyoto.

Jost U, Atwell E 1993 Deriving a probabilistic grammar of semantic markers from unrestricted English text. In Lucas S (ed) *Grammatical Inference: theory, applications, and alternatives, IEE Colloquium Proceedings 1993/092*. London, Institute of Electrical Engineers, pp91-97.

Jurafsky D, Martin J 2000 *Speech and Language Processing*. Prentice-Hall

Karlsson F, Voutilainen A, Heikkila J, Anttila A (eds) 1995 *Constraint Grammar*. Berlin, Mouton de Gruyter.

Kneser R, Ney H 1995 Improved backing-off for n-gram language modelling. In *Proceedings of IEEE ICASP'95.* Detroit, pp49-52.

Kuhn R, de Mori R 1994 Recent results in automatic learning rules for semantic interpretation. In *Proceedings of the International Conference on Spoken Language Processing*. Yokohama, pp75-78.

Leech G, Garside R, Atwell E 1983 The automatic grammatical tagging of the LOB corpus. *ICAME Journal* 7: 13-33.

Leech G, Barnett R, Kahrel P 1996 *EAGLES Final Report and guidelines for the syntactic annotation of corpora*, EAGLES Report EAG-TCWG-SASG/1.5. http://www.ilc.pi.cnr.it/EAGLES96/home.html

MacDonald I 1997 *Hidden Markov and other models for discrete-valued time series*. London, Chapman and Hall.

Magerman D 1994 *Natural Language Parsing as statistical pattern recognition*. PhD thesis, Stanford University.

Manning C, Schutze H 1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press.

McDonough T 1987 *The Search for Extra-Terrestrial Intelligence*. John Wiley and Sons.

Menzel W, Atwell E, Bonaventura P, Herron D, Howarth P, Morton R, Souter C 2000 The ISLE Corpus of non-native spoken English. in Gavrilidou M, Carayannis G, Markantionatou S, Piperidis S, Stainhaouer G (eds) *Proceedings of LREC2000: Second International Conference on Language Resources and Evaluation*. Athens, European Language Resources Association (ELRA), vol.2 pp.957-964.

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

O'Donoghue T 1993 *Reversing the process of generation in systemic grammar*. PhD thesis, University of Leeds.

Ollongren A 1999 Large-size message construction for ETI. In *Proceedings of the 50th International Astronautical Congress*. Amsterdam, paper IAA-99-IAA.9.1.06.

Sagan C (ed) 1978 *Murmers of Earth: the Voyager interstellar record*. New York, Random House.

Sagan C 1988 *Contact*. London, Legend.

Sampson G, Haigh R, Atwell E 1989 Natural language analysis by stochastic optimisation. *Journal of Experimental and Theoretical Artificial Intelligence* 1: 271-287.

Schabes Y, Roth M, Osborne R 1993 Parsing of the Wall Street Journal with the inside-outside algorithm. In In *Proceedings of the 6th EACL*. Utrecht, Association for Computational Linguistics, pp341-46.

Sekine S, Grishman R 1995 A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of IWPT International Workshop on Parsing Technologies.* Prague University.

Sivier D 2000 SETI and the historian: methodological problems in an interdisciplinary approach. . *Journal of the British Interplanetary Society* 53(1&2): 23-26.

Souter C, Churcher G, Hayes J, Hughes J, Johnson S 1994 Natural language identification using corpus-based models. *HERMES Journal of Linguistics.* 13: 183-204.

Souter C, Atwell E 1994 Using parsed corpora: a review of current practice. In Oostdijk N, de Haan P (eds) *Corpus-based research into language*. Amsterdam, Rodopi, pp143-158.

Taylor L, Knowles G 1988 *Manual of information to accompany the SEC corpus: The machine readable corpus of spoken English*. University of Lancaster: Unit for Computer Research on the English Language. Available from http://kht.hit.uib.no/icame/manuals/sec/INDEX.HTM

Teahan B 1998. *Modelling English text.* PhD Thesis, University of Waikato, New Zealand.

Voutilainen A, Jarvinen T 1996 Using the English Constraint Grammar Parser to analyse a software manual corpus. In Sutcliffe R, Richard, Koch H, McElligott A (eds.) *Industrial parsing of software manuals*. Amsterdam, Rodopi, pp57-88.

Turnbull M, Smith L, Tarter J 1999 Project Phoenix: Starlist2000. In *Proceedings of the 50th International Astronautical Congress*. Amsterdam, paper IAA-99-IAA.9.1.02.

Vakoch D 1998a Pictorial messages to extraterrestrials, *SETIQuest* 4(1): 8-10 (Part 1), 4(2): 15-17 (Part 2).

Vakoch D 1998b Constructing messages to extraterrestrials: an exosemiotic perspective. *Acta Astronautica* 42(10-12): 697-704.

Vakoch D 1998c The dialogic model: representing human diversity in messages to extraterrestrials. *Acta Astronautica* 42(10-12): 705-710.

Vakoch D 1999 Communicating scientifically formulated spiritual principles in interstellar messages. In *Proceedings of the 50th International Astronautical Congress*. Amsterdam, paper IAA-99-IAA.9.1.10.

Weischedel R, Meteer M, Schwarz R, Ramshaw L, Palmucci J 1993 Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics* 19(2): 359-382.

Wilson A, Rayson P 1993 The automatic content analysis of spoken discourse. In Souter C, Atwell E (eds), *Corpus based computational linguistics*. Amsterdam, Rodopi, pp215-226.

Wilson A, Leech G 1993 Automatic content analysis and the stylistic analysis of prose literature. *Revue Informatique et Statistique dans les Sciences Humaines* 29: 219-234.

Young S, Bloothooft G (eds) 1997 *Corpus-based Methods in Language And Speech Processing*. Dordrecht/Boston, Kluwer Academic Publishers.

Zipf G 1935 *The psycho-biology of language*. Boston, Houghton-Mifflin

Zipf G 1949 *Human Behaviour and The Principle of Least Effort*. New York, Addison Wesley.