Proceedings Paper:
Atwell, ES and McKevitt, P (1994) Pragmatic linguistic constraint models for large-vocabulary speech processing. In: Proceedings of the Twelfth National Conference on Artificial Intelligence. Integrating Speech and Natural Language Processing: AAAI94 Workshop, 31 July 1994, Seattle, Washington USA. AAAI Press , 58 - 64.

Pragmatic linguistic constraint models for large-vocabulary speech processing

Eric Atwell
CCALAS & Artificial Intelligence Division
School of Computer Studies
University of Leeds
GB- LS2 9JT, Leeds
England, EU.
E-mail: eric@scs.leeds.ac.uk

Paul Mc Kevitt
Department of Computer Science
Regent Court, 211 Portobello Street
University of Sheffield
GB- S1 4DP, Sheffield
England, EU.
E-mail: p.mckevitt@dcs.shef.ac.uk

Abstract

Current systems for speech recognition suffer from uncertainty: rather than delivering a uniquely-identified word, each input segment is associated with a set of recognition-candidates or word-hypotheses. Thus an input sequence of sounds or images leads to, not an unambiguous sequence of words, but a *lattice* of word-hypotheses. To choose the best candidate from each word-hypothesis set (i.e. to find the best route through the lattice), linguistic context needs to be taken into account, at several levels: lexis and morphology, parts-of-speech, phrase structure, semantics and pragmatics.

We believe that an intuitively simple, *naive* model will suffice at each level; the sophistication required for full Natural Language Understanding (NLU) (e.g. Alvey Natural Language Toolkit (ANLT)) is inappropriate for real-time language recognition. We describe here models of each linguistic level which are simple but robust and computationally straightforward (hence 'pragmatic' in the everyday sense) and which have clear theoretical shortcomings in the eyes of linguistic purists, but which nevertheless *do the job*.

1 Background

Output from an English recognition system whether it is speech, handwriting or optical character) is typically a sequence of candidate sets, referred to as a recognition lattice. For example, on 'hearing' the sentence "Stephen left school last year", an English speech recognition system may produce the following lattice of candidates in order of decreasing similarity to the input speech signal:

```
stephen --- stiffen
left --- lift --- loft
school --- scowl --- scull
lest --- last --- least
yearn --- your --- year
```

This is in fact an oversimplification, as several variations are possible on this pattern using lattices. Lattices can have varying numbers of candidates. For example 'DragonDictate[1] 30K' may produce only one candidate when it is sure of the spoken word, or many more (up to 10) when there are many close matches to the acoustic signal. The ordering, in terms of decreasing similarity to the acoustic signal, is based on a confidence value attatched to each candidate. These are not displayed to the user, but may be accessible internally to be used by a language model. In some speech recognisers, lattices are first built at a sub-word level, with candidate-sets of phonemes, syllables, triphones etc. To apply syntactic and semantic linguistic constraints, sub-word lattices must first be converted into word lattices via lexical analysis. With a continuous speech recogniser, word boundaries are not pre-ordained, and alternative candidates may overlap or have unassigned gaps between them, significantly complicating the application of linguistic constraint models. However, for our initial experiments, we assume the simplified model where word-boundaries are known as is the case with a discrete-word recogniser, and candidates are words rather than sub-word units.

To disambiguate lattices, a standard technique is to use a language model to constrain the possible choices, so that the chosen sequence of words is the most linguistically plausible. Most language models for lattice disambiguation provide only a limited coverage of the linguistic knowledge available, restricted to word or wordtag n-grams [Jelinek 90]. Analysis of recognition lattices involves traversing a much larger search space than when analysing sentences, and the necessity of real-time computability acts as a constraint on language model complexity. Because of this, sophisticated language analysis systems have not been successful in disambiguating recognition lattices. [Keenan 92] found that the ANLT parser [Phillips and Thompson 87] was too powerful for such a task, requiring long computation times to discover a very large number of ambiguous analyses of even simple sentences. There is a clear need for a language model incorporating a broader range of linguistic knowledge than word and wordtag n-grams, while remaining computationally feasible.

N-grams or Markov Models are a conceptually simple mathematical means for representing an observable, real-

[1] DragonDictate is developed by DragonSystems Inc., USA.

world sequence of events or symbols. They are equivalent to a non-deterministic finite automata, where the transition from the current symbol to the next is determined by probability, based upon a small fixed-size *window* of previous symbols. Markov theory is computationally efficient and provides simple but very general and powerful models for applications throughout science. In NLP, common applications of Markov theory are in speech-processing where symbols are *acoustic chunks* such as phonemes and in grammatical tagging where symbols are part-of-speech wordtags. However, Markov models might not be readily applicable to higher levels of linguistic analysis (e.g. semantics/pragmatics) involving links between units beyond a small fixed-size window. With respect to language modelling *collocations* are a variation on Markov models or n-grams. An n-gram model records *all* n-length symbol-sequences in a given training set. For example, a word bigram model records all pairs of words in the training set and their frequencies of occurrence, while a word trigram model records all word-triples. A word collocation model records combinations which occur together significantly more frequently than predicted by their probabilities in isolation using some application-specific measure of significance. As only significant combinations and their frequencies are recorded, a much larger *window* can be used than for a strict n-gram model of equivalent size.

One attraction of n-gram and collocational models is that they are *not* compositional. [Gazdar and Mellish 89] state that "...one principle attributed to the philosopher Frege stands out in just about every approach that has been made...known as the *principle of compositionality*, the meaning of a sentence can be expressed in terms of the meanings of the phrases within it." (p. 280). Most NLP researchers, like Gazdar and Mellish, see sentence-understanding as the natural goal of NLP. However, for speech and handwriting recognition, as distinct from understanding, non-compositional models of syntax and semantics are not merely adequate, but more efficient and effective.

N-gram and collocational models also have the advantage of being automatically extractable from appropriate training data [Atwell 87a,b, 88a,b, 92, Souter and Atwell 92, Atwell et al 93]. Each model will be automatically extracted using a variety of large-scale linguistic resources such as tagged corpora, treebanks and machine readable dictionaries. N-gram and collocational models can be learnt even for minority languages without rich NLP resources or expertise, e.g. Slovene [Gros et al 94], Basque [Agirre et al 94]. This is in contrast to many other NLP systems, where linguistic knowledge is supplied from expert introspection. We recognise that each individual model is theoretically and practically inadequate as a model of linguistic knowledge, but believe that taken as a combination they will provide a holistic model of constraints sufficient for the application we propose. The optimal analysis is not required to be fully correct at all levels; its purpose is to indicate the correct words. Furthermore, an integrated collocational model allows a wider variety of knowledge types to be combined straightforwardly, in a clean, simple holistic model; this contrasts with many of the complex architectures developed to integrate disparate knowledge sources since the ARPA SUR projects [Lea 80].

2 Pragmatic linguistic constraints

We briefly outline a selection of 'sub-optimal, linguistically naive' yet robust models of a range of types of linguistic knowledge: word-sense overlap, semantic tags, Markovian parse-trees, wordtag n-grams and word collocations. These are not necessarily models of pragmatics, but pragmatic models in the everyday sense of this word. For example, the following definition of pragmatics is given in Collins English Dictionary:

```
PRAGMATIC: advocating behaviour that is
dictated more by practical consequences
than by theory or dogma
```

Each linguistic knowledge type is represented by a form of minimalist sub-context-free collocational model, rather than a truly compositional model. The linguistic knowledge sources to be integrated are:

2.1 LDOCE semantic primitives

All word sense-definitions in the Longman Dictionary Of Contemporary English (LDOCE) are written in terms of the *Longman Defining Vocabulary*. This is a closed set of approximately 2000 words which effectively constitute semantic primitives. [Demetriou 93], [Demetriou and Atwell 94], [Guthrie 93], [Guthrie et al. 91], [Rose and Evett 92], have shown that the LDOCE text defining a word can be used as the basis for semantic constraints by maximising semantic overlap between words in a recognition lattice. In analysing the earlier illustrative lattice, Demetriou's algorithm looks up the LDOCE definition of each candidate, to calculate a semantic overlap score for every possible path through the lattice (every possible sequence of candidates, e.g. *stiffen lift scowl last year*). For example, the LDOCE definition of *last* includes:
"...(in time) one or ones before the one mentioned or now..."
and the LDOCE definition of *year* includes:
"...a measure of time equal to about 365 days..."
These definitions both contain the word *time*, indicating a semantic overlap favouring coocurrence of these two candidates; so the score of all sequences including *last year* is incremented. This procedure is applied to all candidate-pairs in a sequence to evaluate each possible sequence, and the highest-scoring candidate sequence should be the most semantically consistent.

2.2 Semantic tagging

LDOCE also has a set of semantic field markers which provide a hierarchical taxonomic semantics at a higher-level of abstraction than the sense-definitions. Words have associated a small number of semantic field markers and [Jost and Atwell 93] has shown that these can be used as semantic tags in a Markovian disambiguation algorithm. An alternative semantic tag set has been produced at Lancaster University [Wilson and Rayson 93] and we hope to investigate its applicability.

2.3 Non-compositional phrase structure

A Markovian collocation model parser derived from the Spoken English Corpus (SEC) Treebank has been developed at Leeds [Atwell 83, 87, 93, 94, Pocock and Atwell 93], for the M.O.D. funded Speech-Oriented Probabilistic Parsing (SOPP) project. The model used is a variant of standard Markov theory, in that both the training set and desired output are required to be an alternating sequence of wordtags and labelled bracket combinations. The parser implementation uses this adapted model for a "bracket-insertion" procedure, augmented with a collocational "tree-closing" procedure to ensure parse trees are correctly balanced. With experiments in parsing lattices, using equivalent sized training sets, [Pocock and Atwell 93, Atwell 94] found that the Markov Model based parser is much faster and more robust than a probabilistic chart parser developed as part of the SOPP project. Its optimal parsetree is unlikely to be structurally correct, but it dominates the correct word-sequence which is adequate for lattice disambiguation.

2.4 Wordtag n-grams

These are widely used in handwriting, speech and optical character recognition (e.g. [Jelinek 90], [Keenan 92]). They have also been successfully used for the automatic part-of-speech tagging of corpora [Atwell 83, Owen 87]. [Leech et al. 83, Atwell et al. 84] describe the CLAWS system for tagging the LOB Corpus [Johansson et al. 86]. CLAWS was the first NLP system to go beyond a Markov model to wider collocations. The "augmented first-order model" [Atwell 83] added only *significant* trigrams to the core bigram model, avoiding the size and computational problems of a full trigram model. Other variants of wordtag n-gram models, useful for specific tasks, are discussed in [Jost and Atwell 94a,b, Hughes and Atwell 93, 94a,b, Arnfield and Atwell 93].

2.5 Word-collocational preferences

Word collocations are recognised within English Language Teaching (ELT) and applied linguistics as indicators of the naturalness of native speakers [Howarth 93]. Lexicographers have long known that word-collocations are an alternative source of lexical semantic patterns or constraints (see [Sinclair 87]). More recently, speech and handwriting researchers [Rose and Evitt 92] have used word-collocations as a readily trainable surrogate for traditional NLP semantics in disambiguation of handwriting lattices.

3 Intention modelling

The above models apply to unconstrained large-vocabulary "data-capture" tasks such as automated dictation. Many NLP researchers are more concerned with NL interfaces and dialogue systems; for these, 'low-level' syntactic and semantic constraints are still important, but we also need to model dialogue structure. There has been little work in AI on looking at the empirical side of modelling intentions such as goals, plans and beliefs in dialogue or text (see [Mc Kevitt 92], [Mc Kevitt et al. 92a]). Most of the work has been looking at how formal and intuitive models of intentions can be constructed (see [Allen 93], [Cohen et al. 82], [Grosz and Sidner 86], and [Litman and Allen 84]). However, there has been a history of looking at adjacent pairs of utterances in dialogue, called *adjacency pairs*, and analysing these empirically (see [Heritage 86, 88]) and little or none of this work has been carried across from Sociology into AI.

[Mc Kevitt 91] has conducted Wizard-of-Oz experiments to collect data on the types of questions people ask about computer operating systems. Questions were categorised into a number of basic intention types, such as requsts for information, confirmation, elaboration and so on. One experiment showed that there was a significant difference in the frequencies of intention types under χ^2 and t-test's between two groups of subjects, one experienced and the other unexperienced. The subjects were asking questions about the UNIX[2] operating system. Unexperienced subjects asked many more requests for explanation, guidance, elaboration and confirmation that experienced subjects. Graphs of the frequencies of pairs of subject questions or intentions showed that unexperienced subjects had a tendancy to move from standard requests for information *to* explanations, elaborations, etc. and also tended to repeat those intention types. Even more interesting was the fact that particular subjects had high frequencies of particular intention pairs in their dialogues. Details of the experiments are given in [Mc Kevitt and Ogden 89a, 89b].

A theory of intention analysis (see [Mc Kevitt 91]) has been proposed as a model, in part, of the coherence of natural-language dialogue. A central principle of the theory is that coherence of natural-language dialogue can be modelled by analysing sequences of intention. The theory has been incorporated within a computational model in the form of a computer program called the Operating System CONsultant (OSCON) (see [Guthrie et al. 89], [Mc Kevitt 86, 91], [Mc Kevitt and Wilks 87], and [Mc Kevitt et al. 92b, 92c, 92d]). OSCON, which is written in Quintus Prolog, understands, and answers in English, English queries about computer operating systems.

The computational model has the ability to analyse sequences of intention. The *analysis of intention* has at least two properties: (1) that it is possible to recognise intention, and (2) that it is possible to represent intention. The syntax, semantics and pragmatics of natural-language utterances can be used for intention recognition. Intention sequences in natural-language dialogue can be represented by what we call *intention graphs*. Intention graphs represent frequencies of occurrence of intention pairs in a given natural-language dialogue. An ordering of intentions based on *satisfaction* exists, and when used in conjunction with intention sequences, indicates the *local*[3] and *global* degree of expertise of a speaker in a dialogue.

The architecture of the OSCON system consists of six basic modules and two extension modules. There

[2] UNIX is a trademark of AT&T Bell Laboratories.

[3] By *local* expertise we wish to stress the fact that sometimes experts can act as novices on areas of a domain which they do not know well.

are at least two arguments for modularising any system: (1) it is much easier to update the system at any point, and (2) it is easier to map the system over to another domain. The six basic modules in OSCON are as follows: (1) ParseCon: natural-language syntactic grammar parser which detects query-type, (2) MeanCon: a natural-language semantic grammar (see [Brown et al. 75], and [Burton 76]) which determines query meaning, (3) KnowCon: a knowledge representation, containing information on natural-language verbs, for understanding, (4) DataCon: a knowledge representation for containing information about operating system commands, (5) SolveCon: a solver for resolving query representations against knowledge base representations, and (6) GenCon: a natural-language generator for generating answers in English. These six modules are satisfactory if user queries are treated independently, or in a context-free manner. However, the following two extension modules are necessary for dialogue-modelling and user-modelling: (1) DialCon: a dialogue modelling component which uses an intention matrix to track intention sequences in a dialogue, and (2) UCon: a user-modeller which computes levels of user-satisfaction from the intention matrix and provides information for both context-sensitive and user-sensitive natural-language generation. A diagram of OSCON's architecture is shown in Figure 1.

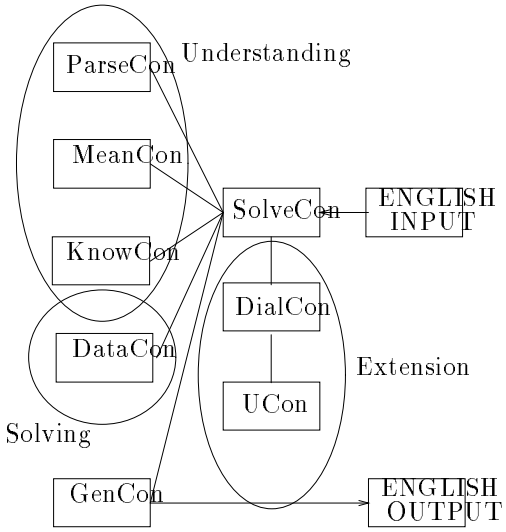

Figure 1: Architecture of the Operating System CONsultant (OSCON) system

Hence, by integrating the processing of higher level information such as intention sequences in dialogue with lower level information such as semantic primitives, semantic tagging and wordtag n-grams we hope that holistic models of integrated speech and language processing can emerge.

4 Conclusion and future work

As previously alluded, an advantage of an integrated collocational model is that it allows a wide variety of knowledge types to be combined straightforwardly, in a clean, simple holistic architecture. This approach is particularly appropriate to parallel architectures, and to Constraint Logic Programming (CLP). At present, only one component [Demetriou 1993] utilises ICL's CLP software development environment DecisionPower/CHIP (Constraint Handling In Prolog); others are coded in Pop11 and Quintus Prolog, so we clearly need to translate submodels to a common implementation language. The different linguistic constraint models will be integrated in a parallel lattice disambiguation model. Dynamic lattice-traversal modules for each level, with separate windows on the same section of the lattice, will map a search space for CLP optimisation. Each word-hypothesis will be annotated with a set of probabilities, one with respect to each level, and these probabilites are then combined into an overall cost function used by inbuilt DecisionPower/CHIP optimisation procedures.

As a resource for evaluating the success of the implemented lattice disambiguation system, we propose to collect together recognition lattices, along with the correct sequence of words for each lattice, from the NLP/Pattern Recognition research community. As the lattices will be gathered from many different sources, we have proposed a standard format, to which all lattices will be converted [Modd and Atwell 94]. We will consult with the research community via SALT (UK) and ELSNET (European) networks for language and speech, both in the gathering of lattice data and in formulation of formatting standards; these must also conform to Text Encoding Initiative (TEI) Guidelines. The Lattice Corpus will be the first of its kind. To be reasonably representative a large sample is required. Initially we aim for a number of recognition lattices equivalent to 50,000 words, which is comparable in size to current richly-annotated parsed corpora such as the Spoken English Corpus (SEC) [Atwell et al 94a,b]. The Corpus will become a standard test resource, and we will distribute it through text archivers and file servers, including the International Computer Archive of Modern English (ICAME) at Bergen University, and the Oxford Text Archive at Oxford University.

When we have a prototype integrated linguistic constraints system and Lattice Corpus, we propose to empirically test and evaluate the system against the corpus, extensively assessing and comparing different weightings and combinations of the component knowledge sources.

5 References

Agirre, E., X. Arregi, X. Artola, A. Diaz de Ilarraza, K. Sarasola (1994) *Conceptual distance and automatic spelling correction.* In [Evett and Rose 94]: 101-108.

Allen, James F. (1983) Recognising intentions from natural language utterances. In *Computational Models of Discourse*, M. Brady and R.C. Berwick (Eds.), 107-166. Cambridge, MA: MIT Press.

Arnfield, Simon, and Eric Atwell (1993) *A syntax based grammar of stress sequences.* In [Lucas 93]: 71-78.

Atwell, Eric Steven (1983) Constituent Likelihood Grammar. In *International Computer Archive of Modern English (ICAME) Journal*, 7: 34-67.

Atwell, Eric, Geoffrey Leech and Roger Garside (1984) Analysis of the LOB Corpus: progress and prospects. In *Corpus Linguistics: Proceedings of the ICAME 4th International Conference*, Jan Aarts and Willem Meijs (Eds.), 40-52. Amsterdam: Rodopi.

Atwell, Eric (1987a) A parsing expert system which learns from corpus analysis. In *In Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference*, Willem Meijs (Ed.), 227-235. Amsterdam: Rodopi.

Atwell, Eric (1987b) Constituent Likelihood Grammar. In *The Computational Analysis of English: a corpus-based approach*, Roger Garside, Geoffrey Leech and Geoffrey Sampson (Eds.), 57-65. London and New York: Longman.

Atwell, Eric (1988a) Grammatical analysis of English by statistical pattern recognition. In *Pattern Recognition: Proceedings of the 4th International Conference*, Josef Kittler (Ed.), 626-635. Berlin: Springer-Verlag.

Atwell, Eric (1988b) Transforming a Parsed Corpus into a Corpus Parser. In *Corpus Linguistics, hard and soft: Proceedings of the ICAME 8th International Conference*, Merja Kyto, Ossi Ihalainen and Matti Risanen (Eds.), 61-70. Amsterdam: Rodopi.

Atwell, Eric (1992) Overview of Grammar Acquisition Research. In *Workshop on sublanguage grammar and lexicon acquisition for speech and language: proceedings*, Henry Thompson (Ed.), 65-70. HCRC, Edinburgh University.

Atwell, Eric (1993) *Corpus-Based Statistical Modelling of English Grammar.* In [Souter and Atwell 93]: 195-215.

Atwell, Eric, Simon Arnfield, George Demetriou, Steve Hanlon, John Hughes, Uwe Jost, Rob Pocock, Clive Souter and Joerg Ueberla (1993) *Multi-level Disambiguation Grammar Inferred from English Corpus, Treebank and Dictionary.* In [Lucas 93]: 91-98.

Atwell, Eric (1994) *Speech-Oriented Probabilistic Parser Project: Final Report to MoD.* CCALAS, School of Computer Studies, Leeds University, England.

Atwell, Eric,John Hughes, and Clive Souter (1994a) *A Unified MultiCorpus for training syntactic constraint models.* In [Evett and Rose 94]: 111-118.

Atwell, Eric,John Hughes, and Clive Souter (1994b) AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Judith Klavans and Philip Resnik (Eds.). ACL'94 Workshop Proceedings, New Mexico State University.

Brown, John Seely, Richard R. Burton and Alan G. Bell (1975) SOPHIE: a step towards creating a reactive learning environment. In *International Journal of Man-Machine Studies*, 7, 675-696.

Burton, R. (1976) *Semantic grammar: An engineering technique for constructing natural-language understanding systems.* BBN Report No. 3453, Bolt, Beranek and Newman, Cambridge, MA, USA.

Cohen, Phil R., C. Raymond Perrault and James F. Allen (1982) Beyond question answering. In *Strategies for natural language processing*, Wendy G. Lehnert and Martin H. Ringle (Eds.), 245-274. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Demetriou, George (1993) Lexical Disambiguation Using CHIP (Constraint Handling In Prolog). In *Proceedings of 1993 European Conference of the Association for Computational Linguistics*, Utrecht: ACL, 431-436.

Demetriou, George and Eric Atwell (1994) *A semantic network for large vocabulary speech recognition.* In [Evett and Rose]: 21-28.

Evett, Lindsay, and Tony Rose (Eds.) (1994) *Computational Linguistics for Speech and Handwriting Recognition: AISB'94 Workshop.* Leeds University: AISB.

Gazdar, Gerald, and Christopher Mellish (1989) *Natural Language Processing in Pop-11 / Prolog / Lisp.* Wokingham, England: Addison Wesley.

Gros, Jerneja, France Mihelic and Nikola Pavesic (1994) *Sentence Hypothesization in a speech recognition and understanding system for the Slovene spoken language.* In [Evett and Rose]: 81-86.

Grosz, B.J. and C.L. Sidner (1986) Attention, intentions and the structure of discourse. In *Computational Linguistics*, Vol. 12, 3, 175-204.

Guthrie, Louise (1993) *A Note on Lexical Disambiguation.* In [Souter and Atwell 93]: 217-239.

Guthrie, Louise, Paul Mc Kevitt and Yorick Wilks (1989) OSCON: An operating system consultant. In *Proceedings of the Fourth Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-89), Subtitled, 'Augmenting Human Intellect By Computer'*, 103-113, Registry Hotel, Denver, Colorado, USA, June.

Guthrie, Joe, Louise Guthrie, Yorick Wilks, and H. Aidinejad (1991) Subject-Dependent Co-Occurrence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 146-152, Berkeley, California. Also Memoranda in Computer and Cognitive Science, MCCS-91-206, CRL, NMSU.

Heritage, John (1986) *Garfinkel and ethnomethodology.* Cambridge: Polity Press.

Heritage, John (1988) Explanations as accounts: a conversation analytic perspective. In *Analysing everyday explanation: a casebook of methods*, Charles Antaki (Ed.), 127-144. London: Sage Publications.

Howarth, Peter (1993) *Automatic analysis of conventional language in written English.* Technical Report, English Language Unit, Leeds University, England.

Hughes, John, and Eric Atwell (1993) *Automatically acquiring and evaluating a classification of words.* In [Lucas 93]: 81-88.

Hughes, John, and Eric Atwell (1994a) *A Methodical Approach to Word Class Formation Using Automatic Evaluation.* In [Evett and Rose 94]: 41-48.

Hughes, John, and Eric Atwell (1994b) The Automated Evaluation of Inferred Word Classifications. In *ECAI'94: 11th European Conference on Artificial Intelligence*, Anthony Cohn (Ed.). John Wiley and Sons.

Jelinek, Fred (1990) Self-organized language modeling for speech recognition. In *Readings in Speech Recognition*, Alex Waibel and Kai-Fu Lee (Eds.), 450-506. Los Altos, CA: Morgan Kaufmann.

Johansson, Stig, Eric Steven Atwell, Roger Garside and Geoffrey Leech (1986) *The Tagged LOB Corpus.* Norwegian Computing Centre for the Humanities, Bergen University.

Jost, Uwe, and Eric Atwell (1993) *Deriving a probabilistic grammar of semantic markers from unrestricted English text.* In [Lucas 93]: 191-198.

Jost, Uwe, and Eric Atwell (1994a) *A hierarchical, mutual information based probabilistic language model.* In [Lucas 93]: 51-58.

Jost, Uwe, and Eric Atwell (1994b) Intrinsic error estimation for corpus-trained probabilistic language models. In *ECAI'94: 11th European Conference on Artificial Intelligence*, Anthony Cohn (Ed.). John Wiley and Sons.

Keenan, Francis (1992) *Large vocabulary syntactic analysis for text recognition.* Unpublished Ph.D. Thesis, Department of Computing, Nottingham Trent University, England.

Lea, Wayne (1980) Speech recognition: past, present, and future. In *Trends in speech recognition*, Wayne Lea (Ed.), 39-98. Prentice-Hall.

Leech, Geoffrey, Roger Garside, and Eric Atwell (1983) The automatic grammatical tagging of the LOB corpus. In *International Computer Archive of Modern English (ICAME) Journal*, 7: 13-33.

Litman, Diane and James Allen (1984) A plan recognition model for clarification subdialogues. In *Proceedings of the COLING-84*, 302-311.

Simon Lucas (Ed.) (1993) *Grammatical Inference - theory, applications and alternatives.* Colloquium Proceedings no.1993/092, London: Institution of Electrical Engineers.

Mc Kevitt, Paul (1986) *Building embedded representations of queries about UNIX.* Memorandum in Computer and Cognitive Science, MCCS-86-72, Computing Research Laboratory, Dept. 3CRL, Box 30001, New Mexico State University, Las Cruces, NM 88003-0001, USA.

Mc Kevitt, Paul (1991) *Analysing coherence of intention in natural-language dialogue.* Ph.D. Thesis, Department of Computer Science, University of Exeter, Exeter, England, EU, September.

Mc Kevitt, Paul (Editorial) (1992) Natural language processing. In *Artificial Intelligence Review*, Vol. 6, No. 4, 327-332, Dec..

Mc Kevitt, Paul and William C. Ogden (1989a) *Wizard-of-Oz dialogues in the computer operating systems domain.* Memorandum in Computer and Cognitive Science, MCCS-89-167, Computing Research Laboratory, Dept. 3CRL, Box 30001, New Mexico State University, Las Cruces, NM 88003-0001, USA.

Mc Kevitt, Paul and William C. Ogden (1989b) *OSWIZ II: Wizard-of-Oz dialogues in the computer operating systems domain.* Memorandum in Computer and Cognitive Science, MCCS-90-181, Computing Research Laboratory, Dept. 3CRL, Box 30001, New Mexico State University, Las Cruces, NM 88003-0001, USA.

Mc Kevitt, Paul, Derek Partridge and Yorick Wilks (1992a) *A survey of approaches to natural language discourse processing.* Technical Report 235, Department of Computer Science, University of Exeter, GB-EX4 4PT, Exeter, England, EU.

Mc Kevitt, Paul, Derek Partridge and Yorick Wilks (1992b) *Analysing coherence of intention in natural language dialogue.* Technical Report 227, Department of Computer Science, University of Exeter, GB-EX4 4PT, Exeter, England, EU.

Mc Kevitt, Paul, Derek Partridge and Yorick Wilks (1992c) *Why machines should analyse intention in natural language dialogue.* Technical Report 233, Department of Computer Science, University of Exeter, GB- EX4 4PT, Exeter, England, EU.

Mc Kevitt, Paul, Derek Partridge and Yorick Wilks (1992d) *Experimenting with intention in natural language dialogue.* Technical Report 234, Department of Computer Science, University of Exeter, GB- EX4 4PT, Exeter, England, EU.

Modd, Dan, and Eric Atwell (1994) *A word hypothesis lattice corpus - a benchmark for linguistic constraint models.* In [Evett and Rose 94]:191-7.

Owen, Marion (1987) Evaluating automatic grammatical tagging of text. In *ICAME Journal 11*, 18-26.

Phillips, J.D., and Henry Thompson (1987) *A parsing tool for the Natural Language theme.* Software Paper no.5, Department of Artificial Intelligence, Edinburgh University, Scotland, EU.

Pocock, Robert and Eric Atwell (1993) *Treebank-trained probabilistic parsing of lattices.* Technical Report 93.30, School of Computer Studies, Leeds University, Leeds, England, EU.

Rose, Tony, and Lindsay Evitt (1992) A large vocabulary semantic analyser for handwriting recognition. In *AISB Quarterly*, 80: 34-39.

Sinclair, John (Ed.) (1987) *Looking Up: an account of the COBUILD project in lexical computing.* London: Collins.

Souter, Clive and Eric Atwell (1992) A Richly Annotated Corpus for Probabilistic Parsing. In *Proceedings of AAAI workshop on Statistically-Based NLP Techniques*, San Jose, California, July 12-17.

Souter, Clive, and Eric Atwell (Eds.) (1993) *Corpus-Based Computational Linguistics.* Amsterdam: Rodopi.

Wilson, Andrew, and Paul Rayson (1993) *Automatic content analysis of spoken discourse.* In [Souter and Atwell 93]: 215-227.