University of Leeds

SCHOOL OF COMPUTING

RESEARCH REPORT SERIES

Report 2003.02

Corpus Linguistics, Machine Learning and Evaluation: Views from Leeds

by

Eric Atwell, Bayan Abu Shawar, Bogdan Babych[§], Debbie Elliott, John Elliott, Paul Gent[¶], Anthony Hartley[§], Xunlei Rose Hu[‡], Julia Medori, Toshifumi Oba, Andy Roberts, Serge Scharoff[§] & Clive Souter[†]

April 2003

[§]Centre for Translation Studies, School of Modern Languages and Cultures

[¶]School of Biomedical Sciences

[‡]School of Computing and Centre for Translation Studies, School of Modern Languages and Cultures

[†]Centre for Joint Honours in Science

Contents

1. Introduction

This collection of short papers is a bird's eye view of current research in Corpus Linguistics, Machine Learning and Evaluation at Leeds University. The papers are extended abstracts submitted to CL2003, the International Conference on Corpus Linguistics, Lancaster University, March 2003; the full papers appear in the CL2003 Proceedings (Archer et al 2003), except for (Hu and Atwell 2003), which is in the Proceedings of the pre-conference workshop on Shallow Processing of Large Corpora, SProLaC (Simov and Osenova 2003). The references from the full papers have been collated into an extensive bibliography of related research.

As background, we have added an Appendix listing abstracts of Corpus Linguistics theses completed 1993-2002, to give an overview of the past ten years of language research in the School of Computing.

We compiled this collection of short papers to illustrate both diversity and interrelatedness of current research at Leeds University. We welcome additions to our research group: research students, visiting scholars, and collaborative research projects. If this collection of short papers interests or even inspires you, please get in touch!

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

2. Using Dialogue Corpora to Train a Chatbot (Bayan Abu Shawar, Eric Atwell)

A chatbot is a conversational program that interacts with users using natural language. ALICE (ALICE 2002), (Abu Shawar and Atwell 2002) is a chatbot system that implements various human dialogues, using AIML (Artificial Intelligent Markup Language), a version of XML format to represent the patterns and templates underlying these dialogues. Elizabeth (Abu Shawar and Atwell 2002), (Millican 2002) is an alternative Chatbot system developed at Leeds University by Peter Millican; our comparative evaluation of the two Chatbots (Abu Shawar and Atwell 2002) concluded that ALICE is more suitable for Corpus-based "retraining" because AIML is closer to the markup formats used in annotated corpora. We are investigating a Corpus-based approach to generalizing these AIML files to cover different dialogue domains and registers in different languages. Since natural dialogue between a person and a computer should resemble a dialogue between humans as much as possible, we have learned from dialogue corpora to generate AIML files, using a program to read dialogue transcripts from the Dialogue Diversity Corpus (DDC) (Mann 2002) and convert these to dialogue patterns and templates in AIML format. The DDC is not a single Corpus, but a collection of links to different dialogue corpuses in different fields. A lot of problems arise when dealing with the diversity in these corpuses. For example:

1- MICASE Corpus (MICASE 2002) (Academic Speech): more than two speakers, long turns/monologues, irregular turn taking (overlapping).
2- CIRCLE Corpus (Ringenberg 2002) (Physics and Algebra tutoring): different formats to distinguish speakers, extra-linguistic annotations.
3- CSPA (Athel 2002) (Corpus of Spoken Professional American English): long turns/ monologues.
4- TRAINS dialogue Corpus (CISD 2002): extra -linguistic annotations.
5- ICE (Nelson 2002) (ICE Singapore English): unconstrained conversations, and great variation in turn length.
6- H. Mischler Book (Mischler 1985) (Medical Interviews): scanned text -image not converted to text format, extra-linguistic annotations.

In order to solve these problems to obtain a good dialogue model, a filtering process must be applied first to these transcripts to remove all unnecessary tags and other markup. Because of the variation in formats and annotations for these transcripts even in the same corpus, each transcript has to be filtered and processed differently, which contradicts the generalization objective.

When re-engineering ALICE to a new domain or conversation style, the patterns and templates learnt from a training corpus are only a raw prototype: the chatbot and AIML files must be tested and revised in user trials. One of the main design considerations is how to plan the dialogue. A good dialogue design would mean less time testing and re-implementing AIML files. In order to train a chatbot system with minimal need to post-edit the learnt AIML, dialogue corpuses should have the following characteristics: two speakers, structured format, short, obvious turns without overlapping, and without any unnecessary notes, expressions or other symbols that are not used when writing a text. Even such "idealised" transcripts may still lead to a chatbot which does not seem entirely "natural": although we aim to mimic the natural conversation between humans, the chatbot is constrained to chatting via typing, and the way we write is different from the way we speak. To date our machine-learnt models have not included linguistic analysis markup, such as grammatical, semantic or dialogue-act annotations (Atwell 1996, Atwell et al 2000), as ALICE/AIML makes no use of such linguistic knowledge in generating conversation responses. However, the Elizabeth chatbot (Abu Shawar and Atwell 2002), (Millican 2002) does allow for more sophisticated conversation modelling including such linguistic features. Future research will include investigating how to incorporate corpus-derived linguistic annotation into Elizabeth-style chatbot pattern files.

Our main conclusion relating to Corpus Linguistics is that the Dialogue Diversity Corpus (DDC) illustrates huge diversity in dialogues, not just in the subject area and speaker background/register but also in mark-up and annotation practices. We urge the dialogue corpus research community to agree standards for transcription and markup format: this would help us, and others too.

3. A Word-Token-Based Machine Learning Algorithm for Neoposy: coining new Parts of Speech (Eric Atwell)

According to Collins English Dictionary, "neology" is: *a newly coined word, or a phrase or familiar word used in a new sense; or the practice of using or introducing neologies*. We propose "neoposy" as a neology meaning "a newly coined classification of words into Parts of Speech; or the practice of introducing or using neoposies".

Unsupervised Natural Language Learning (UNLL), the use of machine learning algorithms to extract linguistic patterns from raw, un-annotated text, is a growing research subfield (e.g. see Proceedings of annual conferences of CoNLL). A first stage in UNLL is the partitioning or grouping of words into word-classes. A range of approaches to clustering words into classes have been investigated (eg Atwell 1983, Atwell and Drakos 1987, Hughes and Atwell 1994, Finch and Chater 1992, Elliott 2002, Roberts 2002). In general these researchers have tried to cluster word-types whose representative tokens in a Corpus appeared in similar contexts, but varied what counts as "context" (eg all immediate neighbour words; neighbouring function-words; wider contextual templates), and varied the similarity metric and clustering algorithm.This approach ultimately stems from linguists' attempts to define the concept of word-class in term of syntactic interchangeability; the Collins English Dictionary explains "part of speech" as: *a class of words sharing important syntactic or semantic features; a group of words in a language that may occur in similar positions or fulfil similar functions in a sentence.* For example, the previous sentence includes the word-sequences *a class of* and *a group of*; this suggests *class* and *group* belong to the same word-class as they occur in similar contexts.

Clustering algorithms are not specific to UNLL: a range of generic clustering algorithms for Machine Learning can be found in the literature (eg Witten and Frank 2000). A common flaw, from a linguist's perspective, is that these clustering algorithms assume all tokens of a given word belong to one cluster: a word-type can belong to one and only one word-class. This results in neoposy which passes a linguist's "looks good to me" evaluation (Hughes and Atwell 1994, Jurafsky and Martin 2000) for some small word-clusters corresponding to closed-class function-word categories (articles, prepositions, personal pronouns), but which cannot cope adequately with words which linguists and lexicographers perceive as syntactically ambiguous. This is particularly problematic for isolating languages, that is, languages where words are generally not inflected for grammatical function and may serve more than one grammatical function; for example, in English many nouns can be used as verbs, and vice versa.

The root of the problem is the general assumption that the word-type is the atomic unit to be clustered, using the set of word-token contexts for a word-type as the feature-vector to use in measuring similarity between word-types, applying standard statistical clustering techniques. For example, (Atwell 1983) assumes that a word-type can be characterised by its set of word-types and contexts in a corpus, where the context is just the immediately preceding word: two word-types are merged into a joint word-class if the corresponding word-tokens in the training corpus show that similar sets of words tend to precede them. Subsequent researchers have tried varying clustering parameters such as the context window, the order of merging, and the similarity metric; but this does not allow a word to belong to more than one class.

One answer may be to try clustering word **tokens** rather than word types. In the earlier example, we can say that the specific word-tokens *class* and *group* in the given sentence share similar contexts and hence share word-class, BUT we need not generalise this to all other occurrences of *class* or *group* in a larger corpus, only to occurrences which share similar context. To illustrate, a simple Prolog implementation of this approach, which assumes "relevant context" is just the preceding word, produces the following:

```
?- neoposy([the,cat,sat,on,the,mat],Tagged).
Tagged = [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat, T2]]
```

We see that the two tokens *the* have distinct tags T1 and T5 since they have different contexts; but the token *mat* is assigned the same tag as token *cat* because they have the same context (preceding word-type). This also illustrates an interesting contrast with word-type clustering: word-type clustering works best with high-frequency words for which there are plenty of example tokens; whereas word-token clustering, if it can be achieved, offers a way to assign low-frequency words and even hapax legomena to word-classes, as long as they appear in a context which can be recognised as characteristic of a known word-class. In effect we are clustering or grouping together word-contexts rather than the words themselves.

The best solution appears to be a hybrid of type- and token-clustering; but our initial investigation has shown that this has heavy computational cost, as a very large constraint-satisfaction problem.

4. Detecting Student Copying in a Corpus of Science Laboratory Reports: Simple and Smart Approaches (Eric Atwell, Paul Gent, Julia Medori, Clive Souter)

This case study is an evaluation of generic, general-purpose plagiarism detection systems applied to a specific domain and task: detecting intra-class student copying in a corpus of Biomedical Science laboratory reports. This involved:

1) *Detailed requirements analysis:*
We examined example student reports in the specific genre, to identify measurable characteristic features (e.g. use of identical diagrams and graphs may be indicative of copying but generic checkers like (Turnitin 2002) assume text-only essays). We also interviewed Biomedical Science teaching staff to elicit/confirm significant diagnostic features which can identify copying, and identify the level of copying considered unacceptable (as some overlap in lab reports is expected).

2) *Survey of what is available and how well this matches the requirements specification*
Unlike other surveys of generic plagiarism detection systems, we aimed to evaluate systems against the detailed specific requirements identified above. A number of candidate systems are available, for example: (CheatChecker 2002) - an n-gram based tool; (SIM 2002) and (YAP 2002) based on longest common sub-sequence approach; detecting 'unusual' similarities between a closed set of essays (or programs) is called collusion and is the basis for (Copycatch 2002); (Copyfind 2002) is an example of a system aimed at scientific rather than humanities documents; (Turnitin 2002) is probably the most widely-known system, but the website gives scant details of the underlying methods; (Clough 2002) proposes use of more sophisticated techniques from Natural Language Processing. We decided to develop two 'simple' copying-detection solutions to assess; and compared these against 'smarter' commercial-strength systems.

3) *Testing and evaluation on a test corpus of Biomedical Science student reports*
Evaluation using a test corpus is a well-established methodology in Corpus Linguistics research, and is appropriate to this task: in our case, the aim is to test candidate systems and compare detection (and 'red-herring') rates.
The initial test corpus used in our study was composed of 103 Biomedical Science first-year student reports. At the outset we did not know whether this included real cases of copying, or if so how many; none had been detected by teaching staff. We found 2 pairs of reports were almost identical, and called these Albert1, Albert2 and Barbara1, Barbara2 (all filenames are randomly generated and are not the real names of the students). As there turned out to be few examples of `real' plagiarism in our initial corpus, we extended it by adding 94 reports written by second year students, covering 3 more challenging, less constrained tasks; during subsequent tests, we discovered 4 pairs which appeared to involve copying: Geraldine2, Fred2; Jane2, Naomi2; Colin2, David2; and Jane3, Naomi3. There appeared to be no other clear cases of copying. To ensure our corpus included samples which we knew were definitely copied, we also added an example of first-year plagiarism that had been detected previously by Biomedical Science staff, which we called Scanfile1, Scanfile2; we were given the hardcopy of two reports to scan using an Optical Character Recognition tool. To check the sensitivity of each plagiarism detection system tested, we created 19 new TestFiles: each composed of one file from which 5%(then 10%,..., 95%) was removed and replaced by a portion of another file. We also created 2 extra TestFiles: TestFileAB made up of half SourceA and half SourceB; and TestFileABC, made up of TestFileAB with added text from SourceC from a different genre (en email message). This gave a final corpus of 220 student laboratory reports including artificial TestFiles.

The first year assignment described the experiment and gave instructions on how to use a computer simulation of the laboratory apparatus to generate results; then the students had to produce tables and graphs and answer a few questions to analyse the results. It seemed likely that any plagiarism we would find would be in the text of the answers. Unlike the first year student reports, the second year students did not have to answer specific questions; so the text should be more `free' and show more originality. However, they still follow the same generic structure of all science laboratory reports: Introduction, Methods, Results, Discussion. A characteristic of this genre is that a lot of the vocabulary from the specific science domain will be shared between all the reports. The answers were quite short (generally about 25 words, occasionally up to 100 words or more), so there is not much scope for originality from each student. As most of the plagiarism checkers available at the moment appear to be developed to detect plagiarism either in humanities essays (high level of originality) or in programming (with limited vocabulary), there is no existing copying checker for the specific problem of scientific reports: low level of originality but not a limited vocabulary.

We used our Corpus to test contrasting approaches to detection. We developed two "simple" programs: Zipping, based on a standard file-compression tool, and Bigrams, a basic comparison of frequent bigrams; and we tested "smart" methods using commercial-strength plagiarism-checking systems Turnitin, Copycatch, and Copyfind. We found little difference in detection rates between "simple" and smart" approaches, and in practice the main performance discriminator was ease of use rather than accuracy. A growing array of tools have appeared for detection of plagiarism and copying; some authorities such JISC in Britain advocate standardised adoption of a generic centralised web-based system. We conclude that our task is an exception meriting a specialised corpus-based study and tailored solution; and that such language analysis tools can be a catalyst for reform in assessment practices.

5. Statistical modeling of MT output corpora for Information Extraction (Bogdan Babych, Anthony Hartley, Eric Atwell)

The purpose of our project is to investigate the usability and performance of Information Extraction (IE) systems such as GATE (Cunningham et al 1996) for processing English output of state-of-the-art commercial Machine Translation (MT) systems. These systems do not yet achieve fully automatic high quality MT, but their mistakes are often stylistic, so IE systems can extract the factual information, which is still conveyed correctly. A part of our project is the issue of automatically acquiring IE templates, using MT output. Automatic acquisition of templates is an important topic in present-day IE research, aiming to make IE technology more adaptive (Wilks and Catizone 1999). There have been suggestions to use lexical statistical models of a corpus and a text for IE to automatically acquire templates: statistically significant words (i.e., words in text that have considerably higher frequencies than expected from their frequencies in the corpus) could be found in the text; templates could be built around sentences where these words are used (Collier 1998).

However, it can be shown that the output of traditional knowledge-based MT systems produces significantly different statistical models from the models for built on “natural” English texts (either original texts, or human translations of texts, done by native speakers). This is due to the fact that translation equivalents in target texts are triggered primarily by source language structures, but not by statistical laws of the target language, which could be the case for statistical MT systems. The translation equivalents in knowledge-based MT output may have substantially different distribution in source and target corpora. As a result, many words that would not be statistically significant in “natural” English texts become significant in MT output. Mistakes in word sense disambiguation, done by MT systems, also deteriorate the value of the statistical model for accurate automatic acquisition of IE templates.

Our paper reports the results of comparison of statistical models built on “natural English” and those built on “MT English” corpora (using the output of different commercial MT systems). We examine the lists of statistically significant words obtained in both cases for several parallel texts, and we evaluate the usability of these words for automatic acquisition of IE templates. We concentrate on contrasting the baseline performance of this method for “natural” English texts and “MT” English texts.

Finally, we propose directions for future work within this project, aimed at improving automatic template acquisition techniques and developing adaptive multilingual IE systems:

- combining knowledge-based template acquisition approaches (Hartley 2002) with statistical methods;
- investigating stochastic models of statistical and example-based MT systems;
- comparing the quality of summaries produced by MT and by summary generation modules of IE systems, after automatically acquired templates are filled.

6. Rationale for a Multilingual Aligned Corpus for Machine Translation Evaluation (Debbie Elliott, Anthony Hartley, Eric Atwell)

The human evaluation of the quality of machine translation (MT) output is costly and time-consuming. Unlike evaluation of PoS-taggers, parsers or speech recognisers, (e.g. Atwell et al 2000) it is not simply a matter of comparing output to some "gold standard" human translation, since translation is subject to stylistic and other variation; instead, evaluation relies on somewhat subjective quality assessments by experts. As a result, some researchers have begun to explore automated methods for evaluating the quality of machine-translated texts. Papineni et al. (2001) rely on a very small corpus that includes human reference translations. Other research (eg. Rajman and Hartley 2002, White and Forner 2001, Reeder et al. 2001, Vanni and Miller 2002) has made use of the much larger DARPA (Defense Advanced Research Projects Agency) corpus, along with results from the largest DARPA human MT evaluation, carried out in 1994. Researchers have used the DARPA corpus and evaluation results to validate (or not, as the case may be) experimental automated evaluation methods, by seeking correlations between DARPA scores and those from new methods. It is clear, however, that this corpus, although still a reliable resource, has its limitations. It comprises only newspaper articles, representing only a small part of MT use; the 300 source texts are in only three languages (French, Spanish and Japanese) and all target texts (human and machine translations) are in American English.

Since expectations of MT systems have become more realistic, a greater number of uses have been found for imperfect raw MT output. Consequently, a variety of text types, genres and subject matter are now machine-translated for different text-handling tasks, including filtering, gisting, categorising, information gathering and post-editing. It is crucial, therefore, to represent this variety of texts, ranging from emails to academic papers, in a corpus for the purpose of MT evaluation. In order to fulfil this requirement, the first step towards compiling this corpus has involved a survey of MT users to obtain information on the kinds of texts, genres and topics regularly translated using MT systems. Results from our survey provide valuable guidance for corpus design.

The number of language pairs that MT systems are now able to deal with is constantly increasing. Our new corpus must, therefore, comprise source texts in several languages, to include not only French, Italian, German and Spanish, but other, typologically different languages, such as Arabic, Chinese, Greek, Japanese and Russian. Furthermore, as much research to date has focussed on the evaluation of MT systems translating into English, it is clear that our corpus should include translations into additional languages to allow for more extensive research. We want to see how well existing MT evaluation methods and metrics transfer to other language-pairs; and we aim to develop new Machine Learnt metrics which generalise across language-pairs. As well as machine translations, expert human translations of each source text will form an essential part of the corpus, to be used for comparison when evaluating machine-translated text segments, and for evaluations in which MT output and human translations are scored against each other. We must also explore the possibility of establishing a reference corpus of untranslated texts in the same subject domains, for evaluating the extent to which translated texts differ from original texts in the same language.

Constraints in terms of time and cost mean that informed decisions must be made with respect to corpus size. An initial MT quality evaluation, specifically designed to compare the validity of scores from differing numbers of texts, is providing guidelines to the number of words required per source language in order for evaluation results to be reliable. This paper describes the results from this experiment, information obtained from the MT user survey and our detailed rationale for corpus design in response to these findings. Our intention is to make the corpus publicly available, along with the results from our detailed human MT evaluations using these texts, to serve the needs of the MT community.

7. The Human Language Chorus Corpus HULCC (John Elliott, Debbie Elliott)

Many aspects of linguistic research, whatever their aims and objectives, are reliant on cross-language analysis for their results. In particular, any research into generic attributes, universals, or inter-language comparisons, requires samples of languages in a readily accessible format, which are 'clean' and of adequate size for statistical analysis. As computer-based corpus linguistics is still a relatively recent discipline, currently available corpora still suffer from a lack of breadth and conformity. Probably due in part to restrictions dictated by funding, many of the machine-readable resources publicly available are for English or one of the major Indo-European languages and, although this is often frustrating for researchers, it is understandable. An equally problematic aspect of inter-corpus analysis is the lack of agreement between annotation schemes: their format, constituent parts-of-speech, granularity and classification, even within a single language such as English.

The aim of HuLCC is to provide a corpus of sufficient size to expedite such inter-language analysis by incorporating languages from all the major language families, and in so doing, also incorporating all types of morphology and word order. Parts-of-speech classification and granularity will be consistent across all languages within the corpus and will conform more closely to the main parts-of-speech originally conceived by Dionysius Thrax than to the fine-grained systems used by the BNC (British National Corpus) and LOB (Lancaster-Oslo/Bergen) corpora. This will then enable cross-language analysis without the need for cross-mappings between differing annotation systems, or for writing/adapting software each time a different language or corpus is analysed. It is also our intention to encode all text using Unicode to accommodate all script types with a single format, whether they traditionally use standard ASCII, extended ASCII or 16 b its.

An added feature will be the inclusion of a common text element, which will be translated across all languages to provide both useful translation data and a precise comparable thread for detailed linguistic analysis. Initially, it is planned to provide at least 20,000 words for each chosen language, as this amount of text exceeds the point where randomly generated text attains 100% bigram and trigram coverage (Elliott, 2002). This significantly contrasts statistically with the much lower percentages attained by natural languages and provides a statistical rationale for what is often a hotly debated point.

Finally, as all constituent language samples within HuLCC conform to the same format and mark-up, a single set of tools will accompany what will be a freely available corpus for the academic world, to facilitate basic analytical needs. This paper outlines the rationales and design criteria that will underlie the development and implementation of this corpus.

8. A survey of Machine Learning approaches to analysis of large corpora (Xunlei Rose Hu, Eric Atwell)

Linguistic study of small corpora can be achieved via a combination of software analysis and human 'manual' annotation. However, hand-crafted analysis or even post-editing is impractical with large corpora, particularly in languages where linguistic expertise is scarce: linguistic annotation must depend on analysis models Machine Learnt froma small 'bootstrap' training corpus. This paper surveys a range of approaches to corpus-based Machine Learning of linguistic annotations, categorised on several dimensions:

- Machine Learning model: neural network, n-gram, stochastic model, constraint-rules, genetic search, hybrid or combined;
- Level of supervision: unsupervised, semi-supervised, supervised;
- Linguistic level of analysis: tokenisation, language identification, PoS-tagging, dependency analysis, phrase structure parsing, semantics, pragmatics, discourse analysis;
- Training text genre: newspapers, dialogue, web-documents;
- Training text language: English, French, German, Arabic, Chinese, others
- Application area: language teaching, information extraction, machine translation, dialogue systems;
- Implementation language: Java, Perl, C, C++, others.

These dimensions provide an ontology or framework for further development. We present examples of systems which fit some of the 'cells' in this multi-dimensional space; but some potential combinations are not represented, for example, we know of no supervised constraint-rules system for semantic tagging of Arabic dialogue transcripts. Our aim is to develop a generic framework to integrate and comparatively evaluate Machine Learning corpus analysis systems. This can also provide a framework for the development of new components, via adaptation of solutions from 'neighbouring cells' in the solution space. Since this approach implies no human expert intervention to validate and/or enrich the output via postediting, there is no longer a need to produce and store annotation files. Instead we store only the raw corpus text, and the layer(s) of linguistic annotation can be generated (and regenerated) dynamically on demand.

9. Using the HTK speech recogniser to analyse prosody in a corpus of German spoken learners' English (Toshifumi Oba, Eric Atwell)

The ISLE project collected a corpus of audio recordings of Italian and German Spoken Learners' English, in which subjects read aloud samples of English text and dialogue selected from typical second language learning exercises (Menzel et al 2000), (Atwell et al 2000, 2003). The audio files were time-aligned to graphemic and phonetic transcriptions, and speaker errors were annotated at the word and the phone level, to highlight pronunciation errors such as phone realisation problems and misplaced word stress assignments.

This paper describes reuse of the ISLE corpus in experiments using speech recognition to investigate intonation in the German speakers. There were three main stages to the research: prosodic annotation of the English text in the corpus, following a model devised for speech synthesis; native speakers' assessments of the intonation abilities of the 23 German speakers; and speech recognition experiments using the HTK Hidden Markov Model ToolKit (Young et al 2001).

Prosodic annotation was done following the set of instructions or informal algorithm in (Knowles 1996), to predict 'model' intonation patterns for written English text, to be passed to a speech synthesiser. We hoped to use these predicted intonations as a 'native speaker target', against which to compare learners' actual intonation patterns, so we investigated automated methods to extract intonation features from the learners' speech-files. Unfortunately, we were unable to automatically predict markup equivalent to the synthesizer cues, so could not directly compare the learners against this model.

Instead, we turned to expert human evaluation: a linguistics researcher and an English language teacher subjectively assessed the intonation of the recorded utterances from German learners of English, by listening to the recorded utterances, comparing against the 'model' marked-up script, and counting perceived intonation errors. Using these judgments, the learners were divided into two groups: 'poor' intonation group and 'good' intonation group. Speakers with exceptionally poor pronunciation were excluded in this grouping by referring to data in the corpus so that results of the following recognition experiments are independent from pronunciation ability.

Finally, the HTK Hidden Markov Model ToolKit was used to train monophone and triphone Hidden Markov Models for 'poor' intonation group and 'good' intonation group separately. In each case, the training set excluded test speakers; this was achieved via cross-validation, repeating the experiment 4 times, taking out a different test-subset each time, and averaging the results. Every trained model was tested with test speakers from both 'poor' and 'good' intonation groups. Results reveal that recognition accuracy becomes higher when models are trained with a group of same intonation ability as test speakers. Cross-merging experiments confirm that these results are consistent.

10. The Use of Corpora for Automatic Evaluation of Grammar Inference Systems (Andrew Roberts, Eric Atwell)

In the past few years, the Natural Language Learning (NLL) community have produced many systems targeted at the complex task of Grammar Inference: automatically inferring or learning grammatical descriptions of a language from a corpus of language examples. A low-level Grammar Inference task is to group or cluster words into tentative categories according to their distributional behaviour, e.g. (Atwell 1983), (Atwell and Drakos 1987), (Hughes and Atwell 1994), (Elliott 2002), (Roberts 2002). A more ambitious aim of Grammar Inference is to propose grammatical phrase structure; a number of mature solutions have emerged, for example, GraSp (Henrichsen 2002), CLL (Watkinson and Manandhar 2002), ABL (van Zaanen 2001), EMILE (Adriaans 1992). All systems mentioned aim to be unsupervised, and thus can only rely on raw (unlabelled) text for their learning process. Results from these systems are quite promising - a variety of linguistic phenomena are induced.

Despite the advances in this field, the issue of thorough evaluation has been poorly addressed. Evaluation within Natural Language Processing tasks in general is problematic, not least because there is an obvious single correct output to measure against. For example, for PoS-tagging and parsing, different linguists advocate different tagsets and parsing schemes, making it difficult to compare accuracy metrics (Atwell 1996), (Atwell et al 2000). Ambiguity poses a similar threat in Grammar Inference: the basic problem is given a training corpus, there is no single correct grammar that represents it. In the majority of systems, a 'looks good to me' approach has been used: success is illustrated by presenting some grammar classifications or structures proposed by the system which appeal to the linguists' intuition. This is not to suggest that such an method is necessarily inappropriate. Verification of a system carried out independently by one or more expert linguists would provide a more reliable measure of performance. However, it is clear that it would be time consuming and would rely of the subjectivity of the experts, and may even be prone to unknown external factors that can affect humans. Evaluation of groups of systems would only be consistent as long as the same people were evaluating them, which is clearly infeasible in the long term.

This paper investigates the feasibility of harnessing corpora that can be used to develop a standard set of procedures (which are reliable and accurate) for evaluating Grammar Inference systems in the future. (Watkinson and Manandhar 2002) have already made some ground on a method to evaluate Natural Language Learning. An existing annotated corpus, in this instance, the Penn Treebank, has been translated into a format that can be easily compared with the output of a Grammar Inference system. For computational reasons, Categorial Grammars (CG) are often used in Grammar Inference; therefore, it is necessary to either develop corpora with CG labels and/or to devise a translation system between CG and existing grammar mark-up of corpora. The next step is to create multi-annotated corpora - that is a body of text that has been parsed by a variety of systems, as exemplified in (Atwell et al 2000). The resulting labels will all be different, and thus, evaluation can take place by comparing the output of a GI system to more than one evaluation corpora, and an overall score can be calculated. Alternative approaches are also discussed.

There is a need to develop methods for consistent evaluation of grammar induction systems. Without a reliable way of determining the performance of such systems, it will become increasing difficult to assess how competently they are doing the task they are supposed to do. Nor will it be trivial to compare two or more systems, which would be very valuable in deciding which techniques and algorithms work best for Natural Language Learning.

11. Methods and tools for development of the Russian Reference Corpus (Serge Sharoff)

From the viewpoint of corpus linguistics, Russian is one of few major world languages lacking a comprehensive corpus of modern language use, even though the need for constructing such a corpus is growing in the corpus linguistics community both in Russia and in the rest of the world.

1. The history of development of Russian corpora

The best known attempt to develop a comprehensive Russian corpus has been made in Uppsala. The Uppsala Corpus consists of 1 mln words of fiction and non-fiction texts, so it is too small and restricted in the genre coverage for modern standards. It also lacks morphosyntactic annotations. Another attempt has been made in the Soviet Union in the mid 1980s under the heading of the Machine Fund of Russian, though it did not produce the expected outcome. There are also multiple ad hoc collections of Russian texts, but they are not balanced and representative.

2. The objective

The objective of the project is to develop the Russian equivalent of the BNC, namely a corpus of 100 mln words with proportional coverage of various functional registers, POS annotation and lemmatisation (the latter is required for Russian, which has dozens of word forms for a lemma). The annotation scheme (based on TEI) also allows to mark noun phrases and prepositional phrases, because they are important for the resolution of ambiguity.

3. Problems and solutions

First, there are problems in obtaining source texts. Some sources are readily available: fiction and news texts are widely accessible via the Internet and can be legally available for the corpus. Other types of the discourse, like business or private correspondence, are hard to obtain and make available in a corpus because of legal obstacles. Yet other types of sources, like samples of spontaneous speech, are rare for technical reasons. The decision is to increase the amount of ephemera whenever possible, because news and fiction texts will take the rest of the share. Personal and business letters are subjected to an anonymization procedure with respect to names of persons and companies. Another set of problems with sources concerns the choice of diachronic sampling, because the turbulent history of Russia in the 20th century radically affected the language. For instance, according to the frequency list (Zasorina, 1977) that was compiled on the basis of texts from 1930-1960, such words as sovetskij (Soviet) and tovarishch (comrade) belonged to the first hundred of Russian words on a par with function words, but this is no longer valid in modern texts. The decision on the chronological limits of the study is different for different functional registers, for instance, fiction texts are taken from 1970, scientific texts from 1980, while news texts from 1997. Second, there are problems with resolving the ambiguity of word forms. Many word forms correspond to several lemmas and POS classes, for instance, the pole is an instance of three different nouns pol (floor), pole (field) and pola (lap). Since they have distinct morphological properties (the case, number and gender), the ambiguity can be resolved on the basis of simple syntactic analysis, like the agreement in noun phrases or between the subject and the predicate in a sentence. Yet other frequent types of ambiguity can be resolved only on the basis of semantic and pragmatic constraints: Xranite svoi denjgi v banke (keep your money in a bank/in a jar). Such cases of genuine ambiguity are kept in the corpus using multiple <ana> tags. Third, there are problems with the query language for accessing the corpus. Typically corpus query languages (e.g. SARA or CQP) assume the fixed order of tokens, while in Russian the order of participants in a clause is not fixed, but depends on thematic development conditions. Special operators are introduced for expressing such conditions.

4. The current state of the project

Currently tools and techniques for working with the reference corpus are tested using a corpus of 40 mln words. Its subcorpus of about 1 million words of fiction texts (The Russian Standard) has automatically assigned and manually inspected POS annotations (it is available from http://corpora.yandex.ru). It can be also used for correcting POS taggers used for processing the corpus.

Bibliography

Abu Shawar B, Atwell E 2002 *A comparison between ALICE and Elizabeth chatbot systems*. School of Computing research report 2002.19, University of Leeds.

Abu Shawar B, Atwell E 2003 *Using dialogue corpora to train a chatbot* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.681-690.

Adriaans P W 1992 *Language learning from a categorial perspective*. PhD thesis, Unversiteit van Amsterdam.

Aitchison J. 1996 *The Seeds of Speech: Language Origin and Evolutio*n. Cambridge University Press,

Ajdukiewicz K 1935 Die syntaktische Konnexiät. *Studia Philosophica* 1:1-27.

ALICE 2002 *A.L.I.C.E AI Foundation* , http://www.alicebot.org/ or http://alicebot.franz.com/

Andryuschenko, V.M. 1989. *Konzepziya i arhitectura Mashinnogo fonda russkogo jazyka* (The concept and design of the Computer Fund of Russian Language), Moscow: Nauka, 1989

Athel 2002 Corpus of Spoken Professional American-English: description, http://www.athel.com/corpdes.html

Atkins S, Clear J H and Ostler N. 1992 Corpus Design Criteria in *Literary and Linguistic Computin*g, Vol.7(1), pp. 1-16.

Atwell E 1983 Constituent-Likelihood Grammar. In ICAME Journal Vol.7

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”, pp61-70, Amsterdam, Rodopi

Atwell E 1993. Corpus-based statistical modelling of English grammar. In S Souter and E Atwell (eds), Corpus-based computational linguistics: Proc 12th ICAME, pp195-214, Amsterdam, Rodopi

Atwell E 1996 *Comparative Evaluation of Grammatical Annotation Models* in Sutcliffe R, Koch H-D, and McElligott A (editors), Industrial Parsing of Technical Manuals, pages 25-46, Rodopi, Amsterdam.

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 2003 *A word-token-based machine learning algorithm for neoposy: coining new parts of speech* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.43-47.

Atwell E 2003. Combining Corpus Linguistics resources and machine Learning in a Language Discovery Toolkit. Internal sabbatical research proposal, School of Computing, University of Leeds

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S. 2000 A comparative evaluation of modern English corpus grammatical annotation scheme*s. ICAME Journa*l, volume 24, pages 7-23, International Computer Archive of Modern and medieval English, Bergen.

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 EACL’87: the Third Conference of European Chapter of the Association for Computational Linguistics”, Copenhagen, ACL

Atwell E, Gent P, Medori J, Souter C 2003 *Detecting student copying in a corpus of science laboratory reports* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.48-53.

Atwell E, Gent P, Souter C, Medori J, 2002. Final Report of the C&IT in the Curriculum project: Customising a copying-identifier for Biomedical Science student reports. School of Computing, University of Leeds. http://www.comp.leeds.ac.uk/eric/citFinalReport.doc

Atwell E, Herron, D., Howarth, P., Morton, R. and Wick, H. 1999 *Pronunciation Training: requirements and solutions*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D1.4. Cambridge: Entropic.

Atwell E, Howarth, P., and Souter, C. 2003 The ISLE Corpus: Italian and German Spoken Learners' English In: *International Computer Archive of Modern and Medieval English (ICAME) Journal*, vol. 27.

Atwell E, Howarth, P., Souter, C., Baldo, P., Bisiani, R., Pezzotta, D., Bonaventura, P., Menzel, W., Herron, D., Morton, R., and Schmidt, J. 2000 User-Guided System Development in Interactive Spoken Language Education. In: *Natural Language Engineering journal, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering,* vol.6 (3-4), pp.229-241.

Atwell E, Lajos G 1993 Knowledge and Constraint Management: Large Scale Applications. In Atwell E (ed), Knowledge at Work in Universities: Proc 2nd HEFCs-KBSI pp21-25, Leeds, Leeds University Press

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

Babych B, Hartley A, Atwell E 2003 *Statistical modelling of MT output corpora for information extraction* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.62-70.
BEEP dictionary 1996 http://www-svr.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

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. *Proceedings of the ARPA workshop on Human Language Technology*. San Mateo, Morgan Kaufmann. pp. 152-157.

Berkling K, Zissman M, Vonwiller J, and Cleirigh C 1998 Improving Accent Identification through Knowledge of English Syllable Structure In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, vol.2, 30 November-4 December 1998, Sydney. pp.89-92.

BNC Index 2003. http://www.comp.lancs.ac.uk/ucrel/bncindex/

Bod, R, 1993. *Using an annotated corpus as a stochastic grammar.* In Proceedings of EACL, the sixth conference of the European chapter of the Association for Computational Linguistics, pp.37-44.

BOKR 2003. Boljshoj Korpus Russkogo yazyka (the Russian Reference Corpus, a description of the Project http://bokrcorpora.narod.ru

Brill E 1993 Automatic grammar induction and parsing free text: a transformation-based approach. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics (ACL)*, Columbus, Ohio, USA, pp 259-265.

Brill, E. 1995. *Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging*. Computational Linguistics, volume 21(4), pages 543-566.

Brown, G. 1977 *Listening to Spoken English*. London: Longman.

CFRL 2003. The Computer Fund of Russian Language, http://irlras-cfrl.rema.ru/

CheatChecker. 2002. http://www.cse.ucsc.edu/~elm/Software/CheatChecker/

CISD 2002 TRAINS Dialogue Corpus, http://www.cs.rochester.edu/research/cisd/resources/trains.html

Cloeren, Jan 1993 Towards a cross-linguistic tagset. In *Proceedings of the ACL Workshop on Very Large Corpora*, Ohio State University, Columbus.

Clough P 2002. http://www.dcs.shef.ac.uk/~clough

Coder 2003. A markup and classification tool: http://www.wagsoft.com/Coder/

Cohen, P, 1995. *Empirical methods for Artificial Intelligence.* MIT Press, Cambridge MA.

Collier R 1998 Automatic template creation for information extraction. PhD thesis. UK.

CopyCatch. 2002. http://www.copycatch.freeserve.co.uk

CopyFind. 2002. http://plagiarism.phys.virginia.edu/software.html

Cruttenden, A. 1997 *Intonation*, 2nd edition. Cambridge: Cambridge University Press.

Cunningham H, WilksY, Gaizauskas R 1996 GATE -- a General Architecture for Text Engineering. *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*, Copenhagen

Déjean H 2000 ALLiS: a symbolic learning system for natural language learning. In Cardie C, Daelemans W, Nédellec C, Tjong Kim Sang E (eds), *Proceedings of the fourth conference on computational natural language learning and of the second learning language in logic workshop*. Lisbon, Portugal, pp 95-98.

Demetriou G and Atwell E. 2001. *A domain-independent semantic tagger for the study of meaning associations in English text*. In Bunt H, van der Sluis I,and Thijsse E (editors), Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4) pp.67-80. Tilburg, Netherlands.

Dialing 2003. The Russian morphological analyser: http://www.aot.ru/

EAGLES 1996 WWW site for European Advisory Group on Language Engineering Standards, http://www.ilc.pi.cnr.it/EAGLES96/home.html Specifically: Leech G, Barnett R and Kahrel P, *EAGLES Final Report and guidelines for the syntactic annotation of corpora*, EAGLES Report EAG-TCWG-SASG/1.5.

Elliott D 2002 *Machine Translation Evaluation: Past, Present and Future.* MA dissertation, University of Leeds.

Elliott J 2002 Detecting Languageness:.In proceedings of 6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI2002), Orlando, Florida, USA: volume XI, pp 323-328.

Elliott J 2002 The Filtration of Inter-Galactic Objets Trouvés and the Identification of the Lingua ex Machina Hierarchy. In: Proceedings of *World Space Congress*: *The 53rd International Astronautical Congress*, pp. 9.2.10. Houston, USA.

Elliott J, Atwell E 2001 Visualisation of long distance grammatical collocation patterns in language. In: IV2001: *Proceedings of 5th International Conference on Information Visualisation*, pp.297-302.

Elliott J, Atwell E, Whyte B 2000 Language identification in unknown signals. In: *Proceeding of COLING'2000, 18th International Conference on Computational Linguistics*, pp 1021-1026, Association for Computational Linguistics (ACL) and Morgan Kaufmann Publishers, San Francisco.

Elliott D, Hartley A, Atwell E 2003 *Rationale for a multilingual aligned corpus for machine translation evaluation* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.191-200.

Elliott J, Elliott D 2003 *The Human Language Chorus Corpus (HULCC)* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.201-210.

Eskenazi, M. 1996 Detection of foreign speakers' pronunciation errors for second language training In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-1996),Philadelphia*, vol.3, pp.1465-1468.

Eskenazi, M. 1999 Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype In: *Language & Technology (LLT) Journal*, vol.2 (2), pp.62-76.

Erjavec T, Ide N, Tufis D 1998 Development and assessment of common lexical specifications for six central and eastern european languages. *Proceedings of LREC'9*8.

Evermann, G. 2002 *HTK History* [Online] http://htk.eng.cam.ac.uk/history.shtml

Finch S, Chater N 1992 Bootstrapping syntactic categories using statistical methods. In Daelemans W, Powers D (eds), *Backgrounds and experiments in machine learning and natural language:Proceedings first SHOE workshop*. Institute for Language Technology and AI, Tilburg University, pp 230-235.

Fox, A. 1984 *German Intonation*. Oxford: Clarendon Press.

Friederici A, Hickok G, Swinney D (eds) 2001 Brain Imaging and sentence processing. In *Journal of Psycholinguistic Research*, Volume 30. New York: Kluwer Academic/Plenum Publishers.

Grabe, E. 1998 *Comparative Intonational Phonology: English and German*, PhD thesis. Max-Planck-Institute for Psycholinguistic and University of Nijmegen.

Hansen, J.H.L. and Arslan, L.M. 1995 Foreign Accent Classification Using Source Generator Based Prosodic Features In: *Proceedings of the 1995 International Conference on Acoustic, Speech, and Signal Processing (ICASSP-1995)*, vol.1, 9-12 May 1995, Detroit. pp.836-839.

Henrichsen P J 2002 GraSp: Grammar learning from unlabelled speech corpora. In Roth D, van den Bosch A (eds), *Proceedings of CoNLL-2002*. Taipei, Taiwan, pp 22-28.

HTK 2000 [Online] http://htk.eng.cam.ac.uk/index.shtml

Hu X R, Atwell E 2003 *A survey of machine learning approaches to analysis of large corpora* In: Simov K, Osenova P (eds.) **Proceedings of SProLaC: Workshop on Shallow Processing of Large Corpora, held in conjunction with the Corpus Linguistics 2003 conference**, UCREL technical paper number 17. UCREL, Lancaster University, pp.45-52.

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*, John Wiley, pp 535-540.

Hunt, J. 1996 *The Ascent of Everest*. Stuttgart: Ernst Klett Verlag, English Readers Series.

Hutchins J, Hartmann W 2002. *IAMT Compendium of Translation Software* 1.5. http://www.eamt.org/compendium.html

Ide, N., Romary, L. (2002). Standards for language resources. In *Proc. of Language Resources and Evaluation Conference (LREC02)*. May, 2002, Las Palmas, Spain. Pp.59-65.

Isahara H 1995 JEIDA's Test-Sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View. In *Proceedings of Machine Translation Summit V*, Luxembourg.

ISLE 1999 *Interactive Spoken Language Education Non-Native Speech Data* 1999 [CD-ROM]. Cambridge: Entropic.

Jurafsky D, Martin J 2000 *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall, New Jersey

Jurafsky, D., Wooters, C., Tajchman, G., Segel, J., Stolcke, A., Folser, E., and Morgan, N. 1994 The Berkeley Restaurant Project In: *Proceedings of the 3nd International Conference on Spoken Language Processing (ICSLP-1994)*, September 1994, Yokohama. pp.2139-2142.

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

Knowles G 1996 From text structure to prosodic structure. In: Knowles G, Wichman A, Alderson P (editors). *Working with Speech*. Harlow: Addison Wesley Longman Limited. pp. 146-167.

Leech G, Garside R, Atwell E, 1983 *Recent developments in the use of computer corpora in English language research*, in Transactions of the Philological Society, pp.23-40.

Leech G, Garside R, Atwell E, 1983. *The Automatic Grammatical Tagging of the LOB Corpus* ICAME Journal of the International Computer Archive of Modern English Vol.7

Leech, G. 1997. A brief users' guide to the grammatical tagging of the British National Corpus, UCREL, Lancaster University. http://www.hcu.ox.ac.uk/BNC/what/gramtag.html

Lesher G, Moulton B, Higginbotham D 1999 Effects of ngram order and training text size on word prediction. *Proceedings of the RESNA '99 Annual Conferenc*e, 52-54, Arlington, VA: RESNA Press.

Lönngren, Lennart (ed.), 1993. *Chastotnyj slovar' sovremennogo russkogo jazyka.* (A Frequency Dictionary of Modern Russian. With a Summary in English.) Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32. 188 pp. Uppsala.

Losee R 1996 Learning syntactic rules and tags with genetic algorithm for information retrieval and filtering: An empirical basis for grammatical rules. *Information Processing and Management 32(2):185-197*.
Mann W 2002 *Dialog Diversity Corpus* http://www-rcf.usc.edu/~billmann/diversity/DDivers-site.htm

Martin, J.R. 1987. The meaning of features in systemic linguistics. In M.A.K. Halliday, R.P. Fawcett (eds.) *New Developments in Systemic Linguistics.* Vol. 1. London: Pinter Publishers. Pp.14-40.

Mast, M, Kompe, R, Harbeck, S, Keissling, A, Niemann, H, Noth, E, Schukat-Talamazzini, E, and Warnke, V, 1996. *Dialog act classification with the help of prosody*. In Proceedings of ICLSP-96, Philadelphia, volume 3, pp1732-1735.

Medori J, Atwell E, Gent P, Souter C 2002 *Customising a copying-identifier for biomedical science student reports: comparing simple and smart analyses*. In: O'Neill M, Sutcliffe R, Ryan C, Eaton M, and Griffith N (eds) **Artificial Intelligence and Cognitive Science, Proceedings of AICS02**, pp. 228-233 Springer-Verlag.

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 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, vol.2, pp.957-964.

MICASE 2002 *MICASE online homepage*, http://www.hti.umich.edu/m/micase/

Miller K, Vanni M 2001 Scaling the ISLE Taxonomy: Development of Metrics for the Multi-Dimensional Characterisation of Machine Translation Quality. In *Proceedings of Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Millican P 2002 *Elizabeth's home page* http://www.etext.leeds.ac.uk/elizabeth

Mishler E 1985 *The discourse of medicine: dialectics of medical interviews*, New Jersey, Ablex http://www-rcf.usc.edu/~billmann/diversity/Tr.5.1a.gif

Morton, R. 1999 *Recognition of Learner Speech*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D3.3. Cambridge: Entropic.

Moshkow Library 2003. http://lib.ru/

Nagao M, Tsujii J, Nakamura J 1985 The Japanese government project for machine translation. *Computational Linguistics* 11: 91-109.

Nelson G 2002 *International Corpus of English: the Singapore Corpus user manual*, http://www-rcf.usc.edu/~billmann/diversity/ICE-SIN_Manual.PDF

O'Connor J, Arnold G 1970 *Intonation of Colloquial English*, 7th edition. London: Longman.

Oba T Atwell E 2003 *Using the HTK speech recogniser to analyse prosody in a corpus of German spoken learner's English* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.591-598.

Oepen S, Netter K, Klein J 1997 TSNLP – Test Suites for Natural Language Processing. *Linguistic Databases. CSLI Lecture Notes*, CSLI Stanford.

Papineni K, Roukos S, Ward T, Zhu W 2001 BLEU: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report* RC22176. Yorktown Heights, NY:IBM.

Papineni K, Roukos S, Ward T, Zhu W-J 2001 Bleu: a method for automatic evaluation of machine translation. IBM research report RC22176 (W0109-022) September 17, 2001

Piao S 2000 Sentence and Word Alignment between Chinese and English. PhD Thesis, Lancaster University.

Piao S 2000 Chinese Corpus adapted from CEPC Corpus, Sheffield University.
Pierce J (Chair) 1966 Language and Machines: computers in Translation and Linguistics. *Report by the Automatic Language Processing Advisory Committee (ALPAC)*. Publication 1416. National Academy of Sciences National Research Council.

Rajman M, Hartley A 2001 Automatically predicting MT systems rankings compatible with Fluency, Adequacy and Informativeness scores. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Rajman M, Hartley A 2002. Automatic ranking of MT systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp 1247-1253.

Rayson Paul, 2002 "Matrix: A statistical method and software tool for linguistic analysis through corpus comparison", PhD thesis, Department of Computing, Lancaster University

Reeder F, Miller K, Doyon J, White J 2001 The Naming of Things and the Confusion of Tongues. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Reithinger, N, and Klesen, M, 1997. *Dialogue act classification using language models.* In Proceedings of EUROSPEECH-97, volume 4, pp.2235-2238.

Reithinger, N, Engel, R, Kipp, M, and Klesen, M 1996. *Predicting dialogue acts for a speech-to-speech translation system.* In Proceedings ICLSP-96, Philadelphia, volume 2, pp.654-657.

Ringenberg M 2002 CIRCLE's tutorial archive http://www.pitt.edu/~circle/Archive.htm

Roberts A 2002 *Automatic acquisition of word classification using distributional analysis of content words with respect to function words.* Technical report, School of Computing, University of Leeds.

Roberts A, Atwell E 2003 *The use of corpora for automatic evaluation of grammar inference systems* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.657-661.

Rock F 2001. Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics* 6(1).

Rodman, R. D. 1999 Computer Speech Technology. Norwood: Artech House Inc.

RS 2003. The Russian Standard (online access to a subcorpus), http://corpora.yandex.ru/

Samuel, K, Carberry, S, and Vijay-Shanker, K, 1998. *Dialogue act tagging with transformation-based learning.* In Proceedings of COLING/ACL-98, Montreal, volume 2, pp.1150-1156.

Sharoff S 2003 *Methods and tools for development of the Russian Reference Corpus* In: Archer D, Rayson P, Wilson A and McEnery T (eds.) **Proceedings of CL2003: International Conference on Corpus Linguistics**, UCREL technical paper number 16. UCREL, Lancaster University, pp.680-681.

Shiwen Y 1993 Automatic evaluation of output quality for machine translation systems. *Machine Translation vol.*8 pp.117-126.

SIM. 2002. http://www.few.vu.nl/~dick/sim.html

Sinclair J 1991 *Corpus Concordance Collocation. Describing English Language*. Oxford: Oxford University Press.

Sinclair J 1996 Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

Souter, C., Howarth, P., and Atwell, E. 1999 *Speech Data Collection and Annotation*, Project Report, Interactive Spoken Language Education Project LE4-8353, Deliverable D3.1. Cambridge: Entropic.

Sperberg-McQueen C, Burnard L (eds.) 2001 *Guidelines for Electronic Text Encoding and Interchange.* http://www.hcu.oxac.uk/TEI/P4X/index.html

Stemmer G, Nöth E, Niemann H 2001 Acoustic Modeling of Foreign Words in a German Speech Recognition System In: Dalsgaard P, Lindberg B, and Benner H, (eds) *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, vol.4, Aalborg. pp.2745-2748.

Stevenson M, Wilks Y 2001 The integration of knowledge sources in word sence disambiguation. *Computational Linguistics* 27(3):321-349.

Stolcke A, Shriberg E, Bates R, Coccaro N, Jurafsky D, Martin R, Meteer M, Ries K, Taylor P, Van Ess-Dykema C 1998. *Dialog Act Modeling for Conversational Speech*, in Chu-Carroll J, Green N, (eds) Applying machine learning to discourse processing: AAAI Spring Symposium, pp.98-105.

Sutcliffe R, Koch H, McElligott A (eds) 1996 *Industrial parsing of software manuals*. Rodopi, Amsterdam

Taylor P, King S, Isard S, Wright H 1998 Intonation and Dialog Context as Constraints for Speech Recognition. In: *Language and Speech*, vol.41 (3-4), pp.493-512.

Taylor P, King S, Isard S, Wright H, Kowtko J 1997 Using Intonation to Constrain Language Models in Speech Recognition. In: *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-1997)*, Rhodes, vol.5, pp.2763-2766.

Teixeira C, Trancoso I, Sarralheiro A 1996 Accent Identification. In: *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-1996)*, vol.3, 3-6, pp.577-580.

Tench P 1996 *The Intonation Systems of English*. London: Cassell.

Thambiratnam D 2001 *[HTK-Users] Problem with HERest*
http://htk.eng.cam.ac.uk/pipermail/htk-users/2001-August/001145.html

Turnitin. 2002. http://www.turnitin.com/

UC 2003. the Uppsala Corpus, available from the University of Tübingen,
http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html

Uebler U, Schüßler M, Niemann H 1998 Bilingual and Dialectal Adaptation and Retraining In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP-1998)*, vol.15, Sydney. pp.1815-1818.

van Zaanen M 2001 *Bootstrapping structure into language: alignment-based learning*. PhD thesis, School of Computing, University of Leeds.

Vanni M, Miller K 2001 Scaling the ISLE Framework: Validating Tests of Machine Translation Quality for Multi-Dimensional Measurement. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

Vanni M, Miller K 2002 Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain.

Verbitskaya L A, Kazanskij N N, Kassevich V B, (in press). Nekotorye problemy sozdanija natsional'nogo korpusa russkogo jazyka. In *NTI, Series 2.* (in Russian)

Vervoort M R 2000 *Games, walks and grammars*. PhD thesis, Unversiteit van Amsterdam.
Watkinson S, Manandhar S 2001 A psychologically plausible and computationally effective approach to learning syntax. In *CoNLL '01: workshop on computational natural language learning*, ACL/EACL.

Watkinson S, Manandhar S 2001 Translating treebank annotation for evaluation. In *Proceedings of the workshop on evaluation methodologies for language and dialogue systems*, ACL/EACL

Werner S, Keller E 1994 Prosodic Aspects of Speech. In Keller E (ed) *Fundamentals of Speech Synthesis and Speech Recognition*. Chichester: John Wiley & Sons Ltd. pp.23-40.

White J (Forthcoming) How to evaluate Machine Translation. In Somers H (ed), *Machine translation: a handbook for translators*. Amsterdam, Benjamins.

White J 1997 MT Evaluation: Old, New and Recycled Methods. *Tutorial, Machine Translation Summit VI,* San Diego.

White J 2000 Toward an Automated, Task-Based MT Evaluation Strategy. In *Proceedings of the Workshop on the Evaluation of Machine Translation, Third International Conference on Language Resources and Evaluation*, Athens.

White J, Forner M 2001 Predicting MT fidelity from noun-compound handling. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, Machine Translation Summit VIII,* Santiago de Compostela, Spain.

White J, O'Connell T, O'Mara F 1994 The ARPA MT evaluation methodologies: evolution, lessons and future approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*. Columbia, pp. 193-205.

White, L. 2000 *English speech timing: a domain and locus approach*, PhD thesis. University of Edinburgh.

Wilks Y 1994 Developments in MT research in the US. *Aslib Proceedings*, vol.46(4) pp.111-116.

Wilks Y, Catizone R 1999 Can we make information extraction more adaptive? In M. Pazienza (ed.) *Proceedings of the SCIE99 Workshop,* Springer-Verlag, Berlin. Rome.

Witt, S. and Young, S. 1997 Language Learning Based on Non-Native Speech Recognition In: *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-1997)*, Rhodes, vol.2, pp.633-636.

Woodland, P. 2000 *HTK History* [Online] http://htk.eng.cam.ac.uk/history.shtml

Woszczyna, M and Waibel, A, 1994. *Inferring linguistic structure in spoken language.* In ICSLP-94, Yokohama, pp.1363-1366.

Yan, Q. and Vaseghi, S. 2002 A Comparative Analysis of UK and US English Accents in Recognition and Synthesis In: *Proceedings of the 2002 International Conference on Acoustic, Speech, and Signal Processing (ICASSP-2002)*, Florida, vol.1, pp.413-416.

Yandex 2003. The search engine: http://www.yandex.ru/

YAP. 2002. http://www.cs.su.oz.au/~michaelw/YAP.html

Young S, Everman G, Kershaw D, Moore G, Odell J, Ollason D, Valtchev V, Woodland P. 2001 *The HTK Book 3.1*. Cambridge: Entropic.

Zasorina, L.N. (ed.) 1977. *Chastotnyj slovar' russkogo jazyka*. Moscow: Russkij Jazyk.

Zipf G K 1949 *Human Behaviour and The Principle of Least Effor*t, Addison Wesley Press, New York.

Appendix: abstracts of Corpus Linguistics theses completed 1993-2002: The past ten years of language research in the School of Computing at Leeds University

A1. Reversing the process of Generation in Systemic Grammar (Timothy O'Donoghue, 1993)

Systemic Grammar is a linguistic formalism that has mainly been exploited in text generation. This thesis explores the process of sentence analysis within systemic grammar, specifically concentrating on the task of syntactic analysis. Syntactic analysis presents a particular challenge since systemic grammars encode syntactic potential in a very distributed manner and, in the formalism utilized in this thesis, a not wholly declarative manner.

In a systemic grammar, the syntactic potential is not explicitly defined as it is in structural formalisms; rather it is implicitly defined via the complex interaction of various grammatical features which trigger certain syntactic relation. The work presented here demonstrates a method for making explicit that syntactic potential; namely stochastic generation and corpus - based extraction.

A large computational systemic grammar is used to stochastically generate a set of syntax trees which is an explicit albeit not very compact representation of the syntactic potential of the grammar. From this corpus, a number of different syntactic models can be extracted and these can in turn can be used as the basis for a number of different parsing schemes. Two schemes are explored: the Dynamic Annealing Parser, a Simulated Annealing-based parser which uses a recursive first order markov model of syntax to evaluate its analyses; and the Vertical Strip Parser, a parser based on a novel syntactic model, a Vertical Strip Grammar. A Vertical Strip Grammar is a type 3 model of syntax approximately a more powerful phrase structure model which allows the task of parsing to be reduced to that of word-tagging. Having found a syntactic analysis, this is then used to drive a further analysis process which recovers (some of) the grammatical features that licensed the syntactic structure.

A2. Prosody and Syntax in Corpus Based Analysis of Spoken English (Simon Arnfield, 1994)

This thesis attempts to show that it can be productive to analyse English prosody in terms of syntax. Although differing prosodies are possible for a fixed syntax, it is demonstrated that an utterances syntax can be used to generate an underlying "baseline" prosody regardless of the actual words, semantics or context. In order to analyse this a British English spoken corpus is needed which has both syntactic and prosodic information. Such a corpus (the Spoken English Corpus (SEC) now known as the Machine Readable Spoken English Corpus (MARSEC)) is used to calculate a number of statistical measures relating the prosodic (specifically the tonic stress mark annotations) and the syntactic (specifically the part of speech tags) information.

This thesis explores the mapping between this information. Models are devised around this information which implement the mappings and select from the search space of possible annotations those with the highest scores. The mapping is applied in the models for prediction of stress and prosodic annotations in new (part of speech tagged) text.

The models are used to demonstrate that there is a clear relationship between parts of speech and the prosodic annotations in the Spoken English Corpus. The models may be exploited to generate stress and prosodic annotations for text-to-speech applications in order to increase the intelligibility and naturalness of the synthesised speech.

A3. Generating Cohesive Texts from Simulations used in Computer Aided Instruction (Alec Grierson, 1994)

There is a need in a wide range of today's computer systems for automatic generation of cohesive text which describes the contents of a knowledge-base. In cases when the knowledge is static it is quite often sufficient to prestore such texts, however when the knowledge-base is dynamic it is not often possible to predict all possible states which the system can attain. This thesis argues that, in particular, the use of simulations in Computer Assisted Instruction is an area which would benefit from automatically generated text, both in response to users' queries and in presenting students with debriefings at the end of simulation episodes. In this application domain, not only is the state of the system undergoing constant flux, but each individual user varies in the type and amount of debriefing they require. In addition, since such systems are intended to be didactic, it is argued that educational considerations should be given to the structuring of the text, as well as linguistically oriented goals. Furthermore, it claims that such knowledge-bases should be constructed with the aim of driving an accurate, pedagogic simulation, not with the express goal of text generation.

An analysis of the text generation software available reveals a dearth of cohesive multiple sentence construction tools yet several single sentence systems, which are assessed. The requirement of the representation for additional qualitative information with which to furnish texts describing the quantitatively defined simulation is discussed.

This research presents a system for automatic generation of such texts. It is grounded on the use of schemata of rhetorical predicates which are filled in accordance with the user's needs, then linearized in such a manner as to ensure that the corresponding generated text is cohesive. The system uses a state-of-the-art multi- paradigm language for specifying explicit, declarative domain models.

Deficiencies of the generated texts are identified and steps which may be taken in order to rectify them are suggested.

A4. A Computational Theory of Contextual Knowledge in Machine Reading (Stephen J Hanlon, 1994)

Machine recognition of off-line handwriting can be achieved by either recognising words as individual symbols (word level recognition) or by segmenting a word into parts, usually letters, and classifying those parts (letter level recognition). Whichever method is used, current handwriting recognition systems cannot overcome the inherent ambiguity in writing without recourse to contextual information.

This thesis presents a set of experiments that use Hidden Markov Models of language to resolve ambiguity in the classification process. It goes on to describe an algorithm designed to recognise a document written by a single-author and to improve recognition by adapting to the writing style and learning new words. Learning and adaptation is achieved by reading the document over several iterations. The algorithm is designed to incorporate contextual processing, adaptation to modify the shape of known words and learning of new words within a constrained dictionary. Adaptation occurs when a word that has previously been trained in the classifier is recognised at either the word or letter level and the word image is used to modify the classifier. Learning occurs when a new word that has not been in the training set is recognised at the letter level and is subsequently added to the classifier. Words and letters are recognised using a nearest neighbour classifier and used features based on the two-dimensional Fourier transform. By incorporating a measure of confidence based on the distribution of training points around an exemplar, adaptation and learning is constrained to only occur when a word is confidently classified.

The algorithm was implemented and tested with a dictionary of 1000 words. Results show that adaptation of the letter classifier improved recognition on average by 3.9% with only 1.6% at the whole word level. Two experiments were carried out to evaluate the learning in the system. It was found that learning accounted for little improvement in the classification results and also that learning new words was prone to misclassifications being propagated.

A5. Automatically Acquiring a Classification of Words (John Hughes, 1994)

This Thesis is an investigation into how a meaningful classification of words can be automatically extracted from ordinary unaltered English text on the assumption that similar words of similar role usually appear in similar contexts.

Experiments are conducted, initially on a moderate sample of two hundred words, to find the patterns that provide the most useful contextual information about the role of each word and to discover how best to convert that information into a structured classification.

The metrics determining the dissimilarity measure between words and the clustering methods which convert these measures into a classification are scored explicitly with an evaluation tool. This tool demonstrates clearly and methodically which techniques are more appropriate for the lexical acquisition task.

The highest scored automatic classification techniques from the initial experiments are used to conduct a large-scale clustering of 2000 words. This clustering is itself evaluated to an accuracy of over 88%.

A6. Probabilistic Language Modelling for Speech Recognition (Uwe Jost, 1994)

Language models are used in speech and handwriting recognition to capture regularities in languages and in this way to provide information about the possibility or likelihood of certain language constructs.

In this thesis, a range of alternative language modelling approaches are reviewed; and two new probabilistic language models are presented. One models relationships between semantic classes of words, using the main subject field codes that are attached to each sense of each word in the machine-readable version of the Longman Dictionary of Contemporary English. The other model is a very simple probabilistic context-free grammar that can be derived fully-automatically from unrestricted English text, using a mutual-information based measure to recursively cluster units of words.
Both models are implemented and first results are presented and discussed.

A7. A Computational Treatment of Conversational Contexts in the Theory of Speech Acts (Amanda Schiffrin, 1995)

A large amount of work has been carried out in the field of pragmatics known as *Speech Act Theory* since its introduction by the philosopher J. L. Austin (1962). The importance of the theory has been widely recognised by linguists and philosophers alike; for in the ability to classify each sentence/utterance into its corresponding act, a large step will have been made towards understanding the meaning of each utterance in a conversation. The preliminary emphasis of this research was to try and implement a computational model to identify so called speech act *infelicities* (illegal discourse moves) in conversation. An example of an infelicitous speech act would be when some speaker agrees with one of the other participants in the conversation that some proposition is so, when in the context of the previous conversation, no other participant has asserted this proposition. The speech act is deemed inappropriate because one of the **preconditions** of the successful act of agreeing is that there must be some proposition in the conversational context with which to agree. Similar preconditions can be specified for a range of speech acts. The utterances expressed by participants are seen as **commitments** to their propositional content, and are stored on **commitment slates**. After the felicitous performance of a speech act, the commitment slates are updated according to the **effects** of the specified speech act. Thus, the commitment slates represent the set of propositional contents to which each participant in a conversation is currently committed. The idea behind this approach was to produce a tool for research to test and refince the analysis of speech acts and attempt to discover what it is that defines each speech act within a conversational context, for the purposes of recognising incoherent conversation. Early on, it was observed that such a theory could be turned on its head so that, assuming that the conversation is coherent, a search could be performed for a speech act that matched the conditions of the current state of the speaker's commitment slate (context), and each implicit speech act might be uniquely identified. If this could be shown to be true, this would fit in very well with current existing theories (such as Bach and Harnish 1979 and Reichman 1985). The principle aim of this thesis has been to demonstrate computationally how far this idea may be taken, and also illuminate its limitations and further developments.

A8. Inferencing Methods using Systemic Grammar with Minimal Semantic Representation (Nicholas Silver, 1995)

This thesis investigates and defends a "surface" approach to knowledge representation. The contrast between this and the standard "deep" approach is described. Arguments are presented from philosophy, psychology and computer science as to why surface reasoning is important. A strong theme running throughout is that a text-only computer, the kind usually used in natural language processing (NLP) applications, lives in a "world of text" because it has no other input or experience from the world. Therefore the deep approach has little relevance to such a computer, as it might a human. By living in a world of text, any semantics such a computer might use should be a semantics of its grammar rather than of the world. Several existing systems which take the surface approach are reviewed.
The grammar used for the demonstration of the surface approach is systemic functional grammar (SFG). This is described in a computationally-oriented manner, with particular reference to the two major implementations of SFG, Communal and Penman. The description is balanced between both implementations, which is an approach not taken previously. It is also suggested that SFG naturally presents its user with a semantics of grammar rather than of the world.
A formal model of SFG is developed and a method of inferencing is introduced using situation theory. It is explained how well the SFG formalism is adapted to the surface approach. A small example grammar is developed and syllogistic rules are added. Finally a computer implementation of this surface reasoning is demonstrated. The implementation uses the COMMUNAL grammar, which has not had any ad hoc (non-linguistic) additions for the purpose. The system uses a knowledge base of parsed natural language sentences. It is able to answer both polar and wh- questions, and performs logical (syllogistic) and grammatical reasoning.
The conclusion is that within certain specified bounds a good natural language grammar can be used for knowledge representation and reasoning, and that grammar does not need any non-linguistic additions for the task.

A9. The Formal Description of Aerobic Dance Exercise: a Corpus-based Computational Linguistics Approach (Adam Bull, 1996)

Over the past three decades the fitness activity known as Aerobics has evolved into one of the ,most popular and widely used forms of exercise in the world. It is now taught by specifically trained fitness professionals with high levels of knowledge in the fields of human kinesiology, physiology, anatomy and exercise theory. In the past, little attempt has been made to formally record and analyse the choreography itself, or to involve the use of computer technology.

There is a clear need for development of a formalism for representing Aerobic choreography, and this is investigated in the research presented. The development of such formal model for the language of Aerobic Dance Exercise using Computational Linguistics techniques is described, and some of its potential uses explored.

In particular the Bodytronix Project, as it became known, involved the collection of a corpus, or representative set, of Aerobics workout routines from qualified and practising fitness leaders; the conversion of the corpus into a standardised knowledge-representation language, or tree bank, held on computer; the extraction of a lexicon of Aerobics moves, annotated with formalised semantic and syntactic descriptions; and the extraction of a formalised grammar, or language model, of an Aerobics workout based on the theory of Generalised Phrase-Structure Grammars (GPSG's).

A10. Exploring the Use of Pattern Recognition, Pattern Identification and Machine Learning to Support Active Collaboration Between the User and the Computer (Daniel N Crow, 1996)

This thesis examines the use of pattern recognition techniques to support the active collaboration between the user and the computer. It focuses on the way that humans interact with computer systems, and how the tasks, specifically the habitual tasks, that users performs using computer can be recognised and automated. Several themes run through this thesis: examining the nature and utility of forms of knowledge, the types and styles of interaction users perform and the question of whether interfaces can support group users or must be adapted to the needs of individuals.

A review of part of the literature of Human-Computer Interaction, concentrating on Intelligent and Adaptive User Interfaces, and the field of Machine Learning describes the related works that inform, support and challenge the work presented here. This is followed by discussion of the historical, theoretical and conceptual issues that underpin this thesis. Both these chapters are informed by the themes identified in full in the introduction.

The thesis also describes a prototype system, known as DB_Habits, which implements many of the ideas and theories of interaction and adaptivity developed in chapter three. A detailed description of the system, its architecture and algorithms, is given, and a set of experimental results obtained using DB_Habits is presented, along with analysis of their implications for the theories presented herein. A final concluding chapter summaries the work, drawing together the various threads of the work, and examining how the basic themes are answered by the experimental results reported. This chapter also discusses possible future directions that might come out of the DB_Habits work.

A11. Working While Driving: Corpus based language modelling of a natural English Voice-User Interface to the in-car Personal Assistant (Michael Schillo, 1996)

This Thesis presents a pilot study of a potential industrial application of Natural Language Processing. An industrial sponsor is providing access to hardware and a practical application: the in-car "Intelligent Personal Assistant" system for travelling executives. The in-car application involves a Psion hand-held computer ("Personal Organiser"), including diary, databases of addresses, appointments, etc. and a portable PC-based commercial Speech Recognition system.

Our long-term aim is: to add a computer system that can access the information in the Psion personal organiser and the connected cellular telephone via a Voice-User Interface (VUI), i.e. via a speech recognition system, so that it is possible to access any desired information via natural audio language input; furthermore, to extend this computer system beyond an interface, to include some of the "intelligence" and organisational functionality expected of a human personal assistant. The Thesis focuses specifically on the modelling of an English-like sublanguage for the in-car personal assistant and pinpointing constraints on the feasibility of the project.

Ideally we would like to allow an unconstrained range of possible utterances; but a fully-comprehensive language model that would cover all possible spoken English would force the Speech Recogniser's performance to be very low. So a central issue is the design of a Voice-User Interface (VUI) sublanguage, a subset of English which will allow users of a speech interface to issue commands in a reasonably 'natural' way while achieving acceptable speech recognition accuracy. The sublanguage should be tailored to cover the underlying functionality of the software system being interfaced to, but less constrained than a menu-keyword system. This would simplify the syntax and increase the speech recogniser's accuracy and overall performance.

Our approach involved a survey of potential users and the Psion itself to elicit the desired functionality. Then an experiment was designed to elicit samples of natural English as it would be used in such a setting covering the whole functionality as found earlier. The experiment was partially conducted in an office setting and a driving simulator to guarantee both good sound quality and natural environment for the recording and provide more background for the progress of the project. The samples taken in this experiment form our corpus. Furthermore, we analysed the corpus of the spoken interactions to point out general properties (the complexity of the language) and some specific problems with regards to the needs of the application. Last but not least, the corpus was used in evaluating a very simple language model with a state-of-the-art speech recogniser, to estimate an upper limit on the accuracy of recognition. We conclude that even an apparently simple command-and-control application can correspond to a complex natural English sublanguage, which casts doubt on the feasibility of a natural English Voice-User Interface.

A12. A Corpus-Trained Parser for Systemic-Functional Syntax (David Clive Souter, 1996)

This thesis presents a language engineering approach to the development of a tool for the parsing of relatively unrestricted English text, as found in spoken natural language corpora.

Parsing unrestricted English requires large-scale lexical and grammatical resources, and an algorithm for combining the two to assign syntactic structures to utterances of the language. The grammatical theory adopted for this purpose is systemic functional grammar (SFG), despite the fact that it is traditionally used for natural language generation. The parser will use a probabilistic systemic functional syntax (Fawcett 1981, Souter 1990), which was originally employed to hand-parse the Polytechnic of Wales corpus (Fawcett and Perkins 1980, Souter 1989), a 65,000 word transcribed corpus of children's spoken English. Although SFG contains mechanisms for representing semantic as well as syntactic choice in NL generation, the work presented here focuses on the parallel task of obtaining syntactic structures for sentences, and not on retrieving a full semantic interpretation.

The syntactic language model can be extracted automatically from the Polytechnic of Wales corpus in a number of formalisms, including 2,800 simple context-free rules (Souter and Atwell 1992). This constitutes a very large formal syntax language, but still contains gaps in its coverage. Some of these are accounted for by a mechanism for expanding the potential for co-ordination and subordination beyond that observed in the corpus. However, at the same time the set of syntax rules can be reduced in size by allowing optionality in the rules. Alongside the context-free rules (which capture the largely horizontal relationships between the mother and daughter constituents in a tree), a vertical trigram model is extracted from the corpus, controlling the vertical relationships between possible grandmothers, mothers and daughters in the parse tree, which represent the alternating layers of elements of structure and syntactic units in SFG. Together, these two models constitute a quasi-context-sensitive syntax.

A probabilistic lexicon also extracted from the POW corpus proved inadequate for unrestricted English, so two alternative part-of-speech tagging approaches were investigated. Firstly, the CELEX lexical database was used to provide a large-scale word tagging facility. To make the lexical database compatible with the corpus-based grammar, a hand-crafted mapping was applied to the lexicon's theory neutral grammatical description. This transformed the lexical tags into systemic functional grammar labels, providing a harmonised probabilistic lexicon and grammar. Using the CELEX lexicon, the parser has to do the work of lexical disambiguation. This overhead can be removed with the second approach: The Brill tagger trained on the POW corpus can be used to assign unambiguous labels (with over 92% success rate) to the words to be parsed. While tagging errors do compromise the success rate of the parser, these are outweighed by the search time saved by introducing only one tag per word.

A probabilistic chart parsing program which integrated the reduced context-free syntax, the vertical trigram model, with either the SFG lexicon or the POW trained Brill tagger was implemented and tested on a sample of the corpus. Without the vertical trigram model and using CELEX lexical look-up, results were extremely poor, with combinatorial explosion in the syntax preventing any analyses being found for sentences longer than five words within a practical time span. The seemingly unlimited potential for vertical recursion in a context-free rule model of systemic functional syntax is a severe problem for a standard chart parser. However, with addition of the Brill tagger and vertical trigram model, the performance is markedly improved. The parser achieves a reasonably creditable success rate of 76%, if the criteria for success are liberally set at at least one legitimate SF syntax tree in the first six produced for the given test data. While the resulting parser is not suitable for real-time applications, it demonstrates the potential for the use of corpus-derived probabilistic syntactic data in parsing relatively unrestricted natural language, including utterances with ellipted elements, unfinished constituents, and constituents without a syntactic head. With very large syntax models of this kind, the problem of multiple solutions is common, and the modified chart parser presented here is able to produce correct or nearly correct parses in the first few it finds.

Apart from the implementation of a parser for systemic functional syntax, the re-usable method by which the lexical look-up, syntactic and parsing resources were obtained is a significant contribution to the field of computational linguistics.

A13. MIRTH Chinese/English Search Engine: A Multilingual Information Retrieval Tool Hierarchy For World Wide Web 'Virtual Corpus' and Training Resource in Computing and Lingustics & Literature (Xiaoda Zhang, 1996)

The thesis addresses the design of a natural language understanding toolkit which works with virtual corporation on the World Wide Web. It is motivated by the desire to create a multilingual search engine to retrieve in both Chinese and English. Using a large distributed hypertext information retrieval system such as the World Wide Web, users find resources by following hypertext links, manually. As the size of the system increases, this becomes impractical. MIRTH is a tool that attempts to help people to find information more efficiently by automatically exploring the Web according to users' instructions. In MIRTH, we first create an index file which contains key information about different Web pages. MIRTH indexes both document titles and document contents. Users can key in queries of search terms directly via a Web browser. Then, MIRTH starts a search program that explores the pre-computed index files in real time and yields search results accordingly.

On the Web, information is stored in many languages other than English, so many multilingual sites exist. Under these circumstances, to design a multilingual information retrieval tool becomes a demanding task for information retrieval from bilingual corpora (English and Chinese) on the Web. This thesis starts to address some problems of the World Wide Web relating to information retrieval. Then, it introduces existing information retrieval tools on the Web. The need to create a multilingual search engine is discussed. Next, a general hierarchy of MIRTH is illustrated. Furthermore, techniques to set up MIRTH search engine are explored. These include building up KBS and SALT virtual corporation, and index files, search engine, and constraints of query syntax. In addition, how to create a multilingual search engine for Chinese information retrieval is dealt with.

In brief, this thesis creates a general model of multilingual information retrieval for Web searching. It copes with both English and Chinese information retrieval.

A14. Lexical Semantic Information Processing for Large Vocabulary Human-Computer Speech Communication (George C Demetriou, 1997)

Speech has always been the most natural communication modality for humans so that unrestricted speech recognition by computers, if attainable, would be the ultimate solution to the human-machine interfacing problem. In spite of past and recent achievements in the area of acoustic modelling, it is widely recognised that the correct automatic transcription of speech is very difficult, if not impossible, without the use or proper analysis, modelling and application of linguistic knowledge in the form of syntactic, semantic or other contextual constraints.

This thesis presents a study on the use of semantic knowledge for large vocabulary (theoretically unrestricted) domain-independent speech recognition. The type of semantic knowledge considered is knowledge about the meanings of the words and a methods has been devised to compute the conceptual association between two words directly from the textual representations of their meanings.

A machine readable dictionary (the Longman Dictionary of Contemporary English) was used to provide the semantic information about the words and experiments were conducted to assess the amount of linguistic constraint implied by the model. The findings suggest that the model is capable of capturing phenomena of semantic associativity or connectivity between words and reducing the lexical ambiguity in natural language.

Experiments with simulated speech recognition hypotheses revealed that the model can be used to reduce the error rates by a considerable factor. Two algorithms for parsing word recognition lattices were developed and the results indicate that, although the algorithms are not admissible, they are quite efficient in restricting the search space and improve the recognition performance.

The important contribution of this research is that it provides a method of modelling limited or incomplete lexical semantic knowledge in a way that it can be efficiently used for large-scale "noisy channel" applications.

A15. A Framework for Dynamic Structuring of Information (Ceila Ralha, 1997)

Systematic classification generally precedes human understanding. Machine understanding will also depend on organizing concepts in order to allow the systematization and compression of large amounts of information. Several problems may frustrate such an attempt at organization. Firstly, information may be distributed among many different files, possibly on different computer systems, perhaps networked over the Internet. A second problem is the time-consuming and labour-intensive process of extracting knowledge from ill-structured expertise sources.

These problems have received attention from within different fields of the Artificial Intelligence community (for example, knowledge acquisition research has tried to develop methods to identify important concepts in ill-structured information elicited from experts). However, this thesis reports on a multi-disciplinary research project to integrate hypertext and hypermedia technology with knowledge acquisition. The thesis proposes a new approach which involves a multi-purpose framework to dynamically structure information in a distributed hypermedia system such as the world-wide web. The new approach co-ordinates aspects of automatic linking of nodes in hyperspace with intelligent mapping of the domain material by the application of qualitative spatial reasoning. Also cognitive aspects of human memory recovery and association, and an adequate visual interface to display large maps of supporting material are involved.

A16. Dialogue Management in Speech Recognition Applications (Gavin E Churcher, 1998)

This thesis combines Linguistics, engineering and empirical paradigms, in the design and evaluation of spoken language dialogue management systems. It begins with surveys of speech recognition systems and products, and dialogue management systems and techniques. In a case study, a commercial speech recognition system is applied to Air Traffic Control (ATC) dialogues; the speech recognition results reported in this case study evidence the need to use discourse and contextual information, in the form of a Dialogue Management System (DMS). A generic model for a DMS is presented, extrapolated from existing models and systems; to establish the generic qualities of the model, a DMS is designed in a further case study, the In-car Automated Personal Assistant. The DMS is implemented and evaluated for an in-car spoken system application. The generic model is also proposed as an evaluation tool for comparing and assessing rival DMSs. There are two aspects to DMS evaluation: quantitative evaluation of the speech recognition performance based on the DMS's ability to predict the language model, and qualitative evaluation of the "naturalness" of the dialogue itself; the generic model is a useful theoretical tool for qualitative comparisons, an alternative to costly and subjective user trials.

A17. Lexical Semantic Association Between Web Documents (Xiao Yuan Duan, 2002)

The rich variety of knowledge available on the World Wide Web makes it an attractive target for data mining and also language processing. In this project, a linguistic method, which was developed to measure the lexical semantic association between two words, is adapted to the task of measuring the semantic similarity between web pages. This lexical knowledge based method could also be used for meaning trend representation or theme representation. Themes are connotative meanings separated or contained in the textual units, such as text, or web pages, and are difficult to represent quantitatively and properly. This work also tries to propose a scientific description of themes that are viewed from the point of lexical semantics. Once textual units have been tagged with Lexical Semantic Tags contained in the Lexical Knowledge Base, themes within the units can be generated. The semantic representation of each web page is a bag of words after tagging which is believed contains certain themes. A program implements the measurement of lexical semantic similarity between two web pages. Various experiments have been undertaken to test the impact of text distance, noisy words and text length. To assess the precision of the methodology, we compare the result of the system with existing commercial information retrieval system and human judgments. A theme space was created to support the evaluation.

A18. Bootstrapping Structure into Language: Alignment-Based Learning (Menno Matthias Van Zaanen, 2002)

This thesis introduces a new unsupervised learning framework, called Alignment-Based Learning, which is based on the alignment of sentences and Harris's (1951) notion of substitutability . Instances of the framework can be applied to an untagged, unstructured corpus of natural language sentences, resulting in a labelled, bracketed version of that corpus. Firstly, the framework aligns all sentences in the corpus in pairs, resulting in a partition of the sentences consisting of parts of the sentences that are equal in both sentences and parts that are unequal. Unequal parts of sentences can be seen as being substitutable for each other, since substituting one unequal part for the other results in another valid sentence. The unequal parts of the sentences are thus considered to be possible (possibly overlapping) constituents, called hypotheses.

Secondly , the selection learning phase considers all hypotheses found by the alignment learning phase and selects the best of these. The hypotheses are selected based on the order in which they were found, or based on a probabilistic function. The framework can be extended with a grammar extraction phase. This extended framework is called parseABL. Instead of returning a structured version of the unstructured input corpus, like the ABL system, this system also returns a stochastic context-free or tree substitution grammar.

Different instances of the framework have been tested on the English ATIS corpus, the Dutch OVIS corpus and the Wall Street Journal corpus. One of the interesting results, apart from the encouraging numerical results, is that all instances can (and do) learn recursive structures.