Corpus Linguistics and Language Learning:

Bootstrapping linguistic knowledge and resources from text

Eric Steven Atwell

Published work submitted for the degree of Doctor of Philosophy

The University of Leeds
School of Computing

April 1st 2008

Acknowledgements

I must thank my wife Gillian and my mother-in-law Rose for coaxing me into completing my PhD despite years of procrastination… and I thank my Advisor, David Hogg, for sound advice and managing the examination arrangements. I could not have achieved the publications listed in Appendix A without the help and support of a long list of co-researchers as co-authors. These include research students I have supervised: Noorhan Abbas, Bayan Abu Shawar, Saleh Al-Osaimi, Latifa Al-Sulaiti, Simon Arnfield, Bogdan Babych, Claire Brierley, Gavin Churcher, George Demetriou, Xiao Yuan Duan, Debbie Elliott, John Elliott, Xunlei Rose Hu, John Hughes, Uwe Jost, Owen Nancarrow, Toshifuma Oba, Tim O'Donoghue, Rob Pocock, Andy Roberts, Majdi Sawalha, Michael Schillo, Clive Souter, Justin Washtell, Sean Wilcock, Xiaoda Zhang; and other researchers who have collaborated with me to write papers: Amal Alsaif, Junaid Arshad, Paulo Baldo, Roberto Bisiani, Paula Bonaventura, Adrian Brockett, David Cliff, Fabio Daneluzzi, Tony Denson, Nicos Drakos, Steven Elliott, Roger Garside, Paul Gent, Robin Haigh, Steve Hanlon, Tony Hartley, Dan Herron, Peter Howarth, Stig Johansson, Chien-Ming Lai, Gyuri Lajos, Geoffrey Leech, Yawei Liang, James Liu, Sam Lievesley, Paul McKevitt, Julia Medori, Wolfgang Menzel, Dan Modd, Rachel Morton, Lan Nim, Matthew Page, Noushin Rezapour-Ashegi, Geoffrey Sampson, Amanda Schiffrin, Jurgen Schmidt, Serge Sharoff, Richard Sutcliffe, Owen Taylor, Joerg Ueberla, Menno van Zaanen, Josiah Wang, Bill Whyte, Heinrich Wick. I am very grateful to you all.

Abstract

This submission for the award of the degree of PhD by published work must: "make a contribution to knowledge in a coherent and related subject area; demonstrate originality and independent critical ability; satisfy the examiners that it is of sufficient merit to qualify for the award of the degree of PhD." It includes a selection of my work as a Lecturer (and later, Senior Lecturer) at Leeds University, from 1984 to the present. The overall theme of my research has been bootstrapping linguistic knowledge and resources from text. A persistent strand of interest has been unsupervised and semi-supervised machine learning of linguistic knowledge from textual sources; the attraction of this approach is that I could start with English, but go on to apply analogous techniques to other languages, in particular Arabic. This theme covers a broad range of research over more than 20 years at Leeds University which I have divided into 8 sub-topics: A: Constituent-Likelihood statistical modelling of English grammar; B: Machine Learning of grammatical patterns from a corpus; C: Detecting grammatical errors in English text; D: Evaluation of English grammatical annotation models; E: Machine Learning of semantic language models; F: Applications in English language teaching; G: Arabic corpus linguistics; H: Applications in Computing teaching and research. The first section builds on on my early years as a lecturer at Leeds University, when my research was essentially a progression from my previous work at Lancaster University on the LOB Corpus Part-of-Speech Tagging project (which resulted in the Tagged LOB Corpus, a resource for Corpus Linguistics research still in use today); I investigated a range of ideas for extending and/or applying techniques related to Part-of-Speech tagging in Corpus Linguistics. The second section covers a range of co-authored papers representing grant-funded research projects in Corpus Linguistics; in this mode of research, I had to come up with the original ideas and guide the project, but much of the detailed implementation was down to research assistant staff. Another highly productive mode of research has been supervision of research students, leading to further jointly-authored research papers. I helped formulate the research plans, and guided and advised the students; as with research-grant projects, the detailed implementation of the research has been down to the research students. The third section includes a few of the most significant of these jointly-authored Corpus Linguistics research papers. A "standard" PhD generally includes a survey of the field to put the work in context; so as a fourth section, I include some survey papers aimed at introducing new developments in corpus linguistics to a wider audience.

Contents

Preface

I would like to start by clarifying for readers of this submission (if ever there are any other than my Examiners) what Leeds University Ordinances and Regulations say should be in a PhD by publications. The requirements for a "standard" PhD are that: "... the thesis must contain evidence of originality and independent critical ability and matter suitable for publication, and be of sufficient merit to qualify for the degree of Doctor of Philosophy." In my own experience of examining "standard" PhDs, I have generally tried to judge whether the thesis contains publishable material, as this yardstick is more tangible than an abstract judgement of "sufficient merit". This is because in practice the two coincide: journal editors would not publish a paper unless it showed "originality and independent critical ability", so surely the measure of a PhD is whether the research reported is publishable. The Research Degrees & Scholarships Office Circular 441 on 'Opportunity for Members of Staff - PhD by Published Work' states that: "... For the award of the degree of PhD by published work, the work must: make a contribution to knowledge in a coherent and related subject area; demonstrate originality and independent critical ability; satisfy the examiners that it is of sufficient merit to qualify for the award of the degree of PhD." The only difference seems to be that for the "standard" case, the examiners have to judge whether there is "matter suitable for publication", whereas a PhD by Published Work obviously meets this criterion. I am left with the first requirement, to "make a contribution to knowledge in a coherent and related subject area".

So, I was left with the challenge of selecting some of my publications to include in this submission; and of presenting them as a coherent and related "story". My first thought was to include ALL my publications, as this should be most impressive. Also it might save me having to write much, as the references alone would take up most of my word-count quota. Luckily my advisor David Hogg helped me realise that my Examiners would not be too pleased at the prospect of ploughing through over 150 publications. It would be better to identify and place

into context the principal contributions of my research career at Leeds University; with a selction of key papers to illustrate this. To help me decide what and how much to include, I sought out some previous PhDs by publication from Leeds University Library; but it turned out they just showed that there is no established precedent or pattern to follow. Past collections of publications ranged from a single book jointly-authored by the candidate[1], to a book plus 5 single-authored and 11 jointly-authored journal papers[2]. Perhaps Ordinances and Regulations (for PhDs, and also for RAE etc) have to be deliberately vague to allow for different norms of publication in disparate research fields. I was left to decide for myself, and so I chose a selection of papers in the four areas outlined in the Abstract: (1) extensions and/or applications of techniques related to Part-of-Speech tagging in Corpus Linguistics; (2) my grant-funded research projects in Corpus Linguistics; (3) joint work with research students I have supervised; and (4) surveys of the field.

Clearly the "weighting" of joint-authored papers in sections 2 and 3 towards my own PhD is less than that of my single-author papers from sections 1 and 4. Probably I could have submitted just the single-authored papers, as surely these are already enough to meet the Ordinances and Regulations for PhD by publications, as discussed above. However, I wanted to present a more rounded picture of my research work at Leeds University; an increasing part of this has been through collaboration and supervision. Indeed, it seems to me that this should be the main difference between a "standard" PhD and a PhD by a staff member: Leeds University academic staff should be actively involved in research grant projects and research student supervision, and I believe that their PhDs by Publications should reflect this.

---

[1] Hudson, Robert. 2006. Stock Market Investment: PhD by published work. University of Leeds.

[2] Green, Andrew. 2001. Published work submitted for the degree of Doctor of Philosophy. University of Leeds.

I'd like to say how interesting it has been to compile this submission. Or rather, I should say that writing the papers has generally been fun; however, the gathering together of the sources has been at times frustrating, as I had to hunt down or generate PDFs of papers from years ago … A full list of my publications can be found on my website (just google "Eric Atwell"). Like most web-sites, the online version is subject to change, but the References section serves as a snapshot at the time of submission of this PhD.

Eric Atwell, 2008.

Chapter 1:
Personal Research Beyond Part-of-Speech Tagging in Corpus Linguistics and Language Learning

The overall theme of my research has been bootstrapping resources from raw textual corpus data. A persistent strand of interest has been unsupervised and semi-supervised machine learning of linguistic knowledge from textual sources; the attraction of this approach is that I could start with English, but go on to apply analogous techniques to other languages, in particular Arabic. This theme covers a broad range of research over 20 years at Leeds University which I have divided into 8 sub-topics: A: Constituent-Likelihood statistical modelling of English grammar; B: Machine Learning of grammatical patterns from a corpus; C: Detecting grammatical errors in English text; D: Evaluation of English grammatical annotation models; E: Machine Learning of semantic language models; F: Applications in English language teaching; G: Arabic corpus linguistics; H: Applications in Computing teaching and research.

In my early years as a lecturer at Leeds University, my research was essentially a development from my previous work with Geoffrey Leech and Roger Garside at Lancaster University, on the LOB Corpus Part-of-Speech Tagging project [1,2,3,4,5,6]; the main tangible achievement of this project was the Tagged LOB Corpus, a resource for Corpus Linguistics research still in use today. After I moved to a Lectureship at Leeds University in 1984, I went on to investigate a range of ideas for extending and/or applying techniques related to Part-of-Speech tagging to research in Corpus Linguistics and Language Learning; my research involved using corpora for machine learning of language models, and/or for assisting student learning of language. Some of these ideas turned out to be theoretically interesting but not entirely successful in practice, and some of my papers actually ended up showing the problems with these approaches; but I suppose this is the nature of speculative research, you have to try new ideas to discover whether they really work.

The first paper in topic A: **Constituent-Likelihood statistical modelling of English grammar** reported ideas for future research on *Analysis of the LOB Corpus: progress and prospects* [7] (Atwell et al 1984). Much previous work on

parsing was based on hand-crafted sets of context-free grammar rules to be applied in a parsing algorithm searching top-down or bottom up for constituents of a context-free parse-tree. The paper introduced hypertags, which capture parse-tree structure as a sequence of bundles of opening and/or closing phrase brackets between PoS-tags. Parsing is modelled as predicting hypertags between PoS-tags, by machine-learning a hypertag-sequence model from a parsed training corpus., This was essentially an extension of the Constituent-Likelihood grammar model used in the LOB tagging program (CLAWS, Constituent-Likelihood Automatic Word-tagging System, an acronym I thought up in my previous job at Lancaster University). This *Constituent-Likelihood Grammar* model used in PoS-tagging, and by extension in hypertag-based parsing, was discussed in more detail in [18] (Atwell 1987). By modelling a parse-tree as a sequence of hypertags, parsing is simplified to prediction of hypertag-sequence. Two novel approaches to *Corpus-based statistical modelling of English grammar* were implemented and compared in [40] (Atwell 1993): Neural Network and Markov bi-gram models of hypertag sequence prediction were implemented and compared. In the end, neither model stood out as clearly better than the other. This model of grammar was taken up by other researchers developing grammatical "chunkers" or partial parsers for English, e.g. (Garside and Fanny Leech 1985), (Sampson 1986), (Garside et al 1987), (Lesk 1988), (Church 1988), (Souter 1990), (Oostdijk 1991), (Chen and Chen 1993), (Fang and Nelson 1994), (Lyon 1994), (Lyon and Dickerson 1995), (Karlsson 1995), (Arnfield 1996), (McMahon and Smith 1996), (Lyon and Frank 1997), (Qiao and Huang 1998), (Hong and Renje 1998), (Smith and McEnery 2000), (Pedler 2007). Similar models were also applied to segmentation in other languages, e.g Chinese (Chen et al 1996), (Huang et al 1997), (Feng 2001), Hungarian (Megyesi 1998), French (Thibeault 2004), Czech (Kralik and Hladka 2006); and multi-language or language-independent systems, eg (Paprotte and Schumacher 1993), (Rayson 2003), (Elliott 2007). Constituent-Likelihood Grammar research was also applied in research on spoken language annotation (Stenstrom and Svartvik 1994), (Arnfield 1996); lexical semantics (Demetriou 1993), (Harley and Glennon 1997), (Johnson et al 2001); language variation (Oostdijk 1988); and English language teaching (Chapelle 1988), (Sugiura 1990), (Saito et al 2002).

My research on Constituent Likelihood Grammar broadened into topic B: **Machine Learning of grammatical patterns from a corpus**. I investigated unsupervised machine learning of Part-of-Speech tags: *a parsing expert system which learns from corpus analysis* [13] (Atwell 1987). English corpus projects have used various different PoS-tag sets, as linguists disagree on definitions and boundaries of grammatical categories; so I thought it would be interesting to see if unsupervised Machine Learning could come up with an "objective" set of grammatical classes independently. Clustering algorithms had been applied to science data-sets, but the application of unsupervised machine learning to large-scale textual corpus data was novel. The paper reported on results of a pioneering experiment to automatically categorise word-types according to the lexical contexts they keep in a corpus, using a bespoke machine-learning clustering algorithm. Two word-types are clustered together if their tokens tend to appear with the same neighbour-words. The program started by compiling a context-profile for each word-type: a list of words it appears next to, and frequency of each pairing. Then starting with the most frequent word in the corpus ("the"), its context-profile is compared with context-profiles of all other words, to find the best match. If the best match is above a threshold score, the words are clustered, and their context-profiles are merged. The program went on to the next-most-frequent-word; and iterated until no more word-pairs were sufficiently similar, with a match-score above the threshold. Then, the threshold was lowered, and the process started again from the most frequent "word", which by this stage was not the word "the", but the cluster including "the". The results showed a clear cluster of prepositions, and other clusters for some other high-frequency function-words; but the method did not work well for lower-frequency words. So, I tried to extend this approach in *pattern recognition applied to the acquisition of a grammatical classification system from unrestricted English text* [15] (Atwell and Drakos 1987). Grammatical classes are indicated by inflectional patterns as well as by syntagmatic relations; so this paper investigated how to cluster or group words using word-ending or suffix information as well as lexical context in the machine learning algorithm, which included some lower-frequency words with common suffixes.

I experimented with *Transforming a Parsed Corpus into a Corpus Parser* by a different approach: machine-learning context-free grammar/parser rules from a parsed corpus [26] (Atwell 1988). The learning system extracted a set of context-free grammar rules from each parse-tree in the parsed corpus, and converting each context-free rule into a Prolog Definite Clause Grammar rule. In theory, the set of Prolog DCG rules constitutes not just a grammar but also a parser and generator: DCG grammars can be used both for parsing corpus sentences and generation of new sentences. The resulting Prolog rule-set was much larger than other AI knowledge bases of the time, and unfortunately proved much too large for the available Prolog interpreter.

A general drawback of simple clustering of word-types is that all tokens or instances of a word-type are assumed to belong to the same word-class, but in English many words can take two or more different PoS-tags depending on context. [115] (Atwell 2003) introduced an alternative approach to clustering similar words, using Prolog unification: *a new machine learning algorithm for neoposy: coining new parts of speech*. This paper proposed to group together (through Prolog unification) individual word tokens (rather than word-types) which appeared in equivalent contexts. However, it was not clear what should count as "equivalent context": if exactly the same word-types are needed then not many words are unified; but if equivalent unification-classes are allowed, then too many words are grouped together. [126] (Atwell 2004) compared and contrasted the two different approaches: *Clustering of word types and unification of word tokens into grammatical word-classes*. The speculative conclusion was that the ideal approach should be some hybrid of the two.

I advocated *Combinatory Hybrid Elementary Analysis of Text* [139] (Atwell and Roberts 2006) as a solution to another challenge for unsupervised machine learning: unsupervised segmentation of words into morphemes. The CHEAT approach involves Super-Sized unsupervised learning, as it combines not one, not two, not three but four different layers of unsupervised learning. The key idea is: acquire results from a number of other candidate systems; CHEAT will read in the

output files of each of the other systems, and then line-by-line select the "majority vote" analysis - the analysis which most systems have gone for. The first layer of unsupervised learning involved getting my class of students to develop their own entries for the contest independently: “unsupervised learning” is (coincidentally) a recognized term in Education research, referring to student learning with minimal explicit direction from teachers. The resulting unsupervised learning systems developed by students constituted the second layer. Our cheat.py program learnt from the students’ outputs, without knowing which was the correct answer; so this was the third layer of unsupervised learning. An anonymous reviewer of our draft paper pointed out that the CHEAT approached seemed similar to an approach already known in the Machine Learning literature: a committee of unsupervised learners; however, we had developed the CHEAT algorithm without use of training material such as this background literature, adding a fourth layer to the super-sized unsupervised learning model! We could argue that super-sized unsupervised learning is not only a valid approach to Machine Learning for Corpus Linguistics, but also a valid approach to Student Learning; which makes super-sized unsupervised learning good for teaching as well as research.

My research on Machine Learning of grammatical patterns from a corpus was followed up by other Corpus Linguistics researchers, including (Altenberg 1991), (Souter and O’Donoghue 1991), (O’Donoghue 1991), (Belmore 1991), (Lankhorst and Moddemeijer 1993), (Osborne 1994), (Wilms 1995), (Zavrel 1996), (Sutcliffe 1996), (Fang 1996), (Wermter et al 1996), (Clark 2001), (Bod 2003), (Kurimo et al 2006), (Sawalha and Atwell 2008).

As well as investigating new algorithms for tagging, parsing, and unsupervised learning of grammar and morphology, I have been interested in applications of these techniques. One such avenue of research is topic C: **Detecting grammatical errors in English text.** At the time of this research, Word Processing was a new technology; and at best error-detectors could flag “non-words” but not grammatical errors. Parsers were so “brittle” that they could not parse many sentences in ordinary English text, so a “grammar-checker” based on a parser would simply have rejected

many sentences as unparseable. My approach was to adapt PoS-tagging: *how to detect grammatical errors in a text without parsing it* [16] (Atwell 1987), by PoS-tagging the text and then flagging low-probability PoS-tag sub-sequences as suspected errors. The system was illustrated using an example text where most of the errors were pinpointed and flagged; however this early paper presented no larger-scale evidence that this would work more generally. [17] (Atwell and Elliott 1987) included further discussion of probabilistic corpus-based methods for *Dealing with ill-formed English text*, using the PoS-tagger to predict grammatical errors. Instead of testing on a small artificial text, this paper presented results on an Error Corpus compiled from "real" sources. Further research on spelling and grammar error-detection has continued to build on these pioneering findings, including: (Rapp 1995, 1996), (Mitton 1996), (Ingels 1996, 1997), (Mangu and Brill 1997), (Bolt and Yazdani 1998), (Gojenola and Oronoz 2000), (Min et al 2000), (Bigert 2004, 2005), (Carlberger et al 2004), (Foster 2005), (Hirst and Budanitsky 2005), (Rayson et al 2005), (Al-Sulaiti and Atwell 2006), (Andersen 2007), (de Ilarraza et al 2007), (Pedler 2007), (Ramshaw 2007), (Wagner et al 2007).

Another application of a linguistically-analysed corpus is topic D: **Evaluation of English grammatical annotation models**. *Comparative evaluation of grammatical annotation models* [65] (Atwell 1996) presents research on corpus-based evaluation of parsing results from a range of English parsing programs. One approach to comparative evaluation of rival parsing algorithms was to test them on a common text and compare their outputs; but in practice different parsers can operate with very different theories or models of English grammar, rendering their analyses incompatible and not directly comparable. In response to this problem, some researchers advocated ignoring the labels used in each parsing scheme, and falling back on a comparison of tree-structures or phrase-bracketing; but the grammatical classes and functions can be a major part of some parsers' outputs, so the "crossing-brackets" evaluation is unfair to these sophisticated parsers. The paper proposed a set of criteria for assessing the complexity or sophistication of the linguistic labelling scheme, as an orthogonal issue to bracketing accuracy. This work influenced subsequent research on evaluation and standards for grammatical annotation of

corpora, including (Sutcliffe et al 1996), (Carroll et al 1998, 1999), (Ide and Romary 2001), (Oepen et al 2004), (Kakkonen 2007).

My work on topic E: **Machine Learning of semantic language models** extended my research from grammar to semantics and pragmatics. I investigated a broad range of corpus annotations of linguistic knowledge which can be used in *Machine learning from corpus resources for speech and handwriting recognition* [66] (Atwell 1996); Part-of-Speech tags are just one of several layers of linguistic knowledge which can be encoded in a corpus and used to improve disambiguation of input speech, handwritten script, or multi-modal interaction, eg (Sutcliffe et al 1996), (Hess and Volk 1998), (Gorostiza et al 2006). Probably my most ambitious (and/or outlandish!) idea for applied corpus linguistics research was a proposal for *a corpus for interstellar communication* [96] (Atwell and Elliott 2001) in the event that the Search for Extra-Terrestrial Intelligence proves successful. To help our new ET friends learn English, we should build a richly-annotated corpus suitable for Machine Learning of semantics as well as grammar and lexis, combining a range of existing corpus linguistics resources. This paper was presented at the first International Conference on Corpus Linguistics, which was effectively Geoffrey Leech's 65th birthday party; Longman supplied a huge birthday cake. My paper proposed that an international expert committee should be set up to design the Corpus for Interstellar Communication, headed by our preeminent Corpus Linguist … but Geoffrey Leech declined the invitation! And as far as I know, no researchers have followed up this idea (yet).

Many corpus linguists are interested in another application area, topic F: **Applications in English language teaching**. A related strand of research on extracting linguistic knowledge, not from a corpus but a dictionary text-file, resulted in *a lexical database for English learners and users: the Oxford Advanced Learner's Dictionary* [31] (Atwell 1989). The OALD typesetting tape text-file was parsed to extract a lexical database usable by corpus linguistics researchers. For example, Appendix 2 of [31] showed an example search of the OALD database for taboo words: Abo, arse, ball, … 40 in total! I had access to the OALDCE file for personal

research through contact with the OALDCE Editor, Tony Cowie, at the time a colleague in the English Department at Leeds University; but I had no right to distribute or promote my extracted database. Other researchers around the same time also independently investigated the OALD "raw" typesetting file, i.e. the source that I started from, to extract lexical information for personal research, eg (Beckwith et al 1991), (Church et al 1991), (Zernik 1991), (Mitton 1992), (Karp et al 1992) (Strzalkowski and Vauthey 1992); but Mitton took the initiative by placing his OALD-derived database in the Oxford Text Archive for general reuse. Hence, most later re-use of OALD in NLP research was of Mitton's version; for example, (Chen and Xu 1995), (Yamashita and Matsumoto 2000), (Minnen et al 2001), (Brierley and Atwell 2008) cited not a journal or conference paper, but Mitton's "readme" file from the Oxford Text Archive.

I have also published a range of other papers extending and/or applying techniques relating to Part-of-Speech tagging, too many to cover in detail in this submission, including: [8, 9, 11, 12, 21, 23, 25, 27, 32, 33, 43, 47, 55, 57, 59, 73, 95, 122, 126, 128, 148, 149, 150, 156, 157, 158]; furthermore, most or all of the papers referred to in other sections could also be seen as extensions and/or applications of Part-of-Speech tagging.

Chapter 2:
Grant-Funded Research Projects
in Corpus Linguistics and Language Learning

As well as continuing with my individual research, I have managed a number of research projects supported by research grants. In this mode of research, I had to come up with original ideas and guide the project, but much of the detailed implementation was down to research assistant staff. This second section presents a range of co-authored papers arising from grant-funded research projects in Corpus Linguistics and Language Learning; the projects used corpora for machine learning of language models, and/or for assisting student learning of language. The projects continued on the research topics initiated in chapter 1; or extended my research into related topics.

Project APRIL [28] (Haigh et al 1988) developed topic A: **Constituent-Likelihood statistical modelling of English grammar.** We investigated another approach to Machine-Learning a parser from a parsed corpus. The parser started with an initial guess of the parse-tree for a given sentence, and then gradually tried to improve this first attempt by a series of minor changes, using patterns from the parsed corpus to evaluate proposed changes. Instead of hill-climbing or optimisation by only accepting improvements, the evolving tree sometimes accepted changes that seemed worse; this search strategy, Simulated Annealing, aimed to avoid getting stuck in local maxima in the search space. Unfortunately, the APRIL parser still made many mistakes, and we did not have conclusive evidence of its superiority over other parsers; but the approach was of interest to other researchers, e.g. (Sharkey 1989), (Souter 1990).

Further research on topic D: **Evaluation of English grammatical annotation models** included *AMALGAM: Automatic mapping among lexicogrammatical annotation models* [57] (Atwell et al 1994). We aimed to map between a range of rival English corpus Part-of-Speech tagging schemes, and analyse similarities and differences between the PoS-tag sets. *A comparative evaluation of modern English corpus grammatical annotation schemes* [87] (Atwell et al 2000) covered both Part-of-Speech tagging schemes and parsing schemes applied in a range of English corpus research projects. The AMALGAM project also collated definitions and

examples of each of the tag-sets, downloadable from the AMALGAM website, and provided an email Part-of-Speech tagging service. Users could submit their English text by email, stating which PoS-tagset(s) they wanted to apply to the text; and receive a PoS-tagged text by reply. This research was used by other working on comparing and combining tagsets and taggers, e.g. (Teufel 1995), (Sutcliffe et al 1996), (Zavrel and Daelemans 2000), (Dejean 2000), (Foth and Hagenstrom 2002), (Al-Sulaiti and Atwell 2006), (Pedler 2007), (Dickinson and Jochim 2008).

The ISLE project developed topic F: **Applications in English language teaching**. The project built tools and resources for Interactive Spoken Language Education. We developed a prototype system including English pronunciation exercises which used speech recognition to detect pronunciation errors and suggest corrections. Our paper in the Natural Language Engineering Journal special issue on best practice in spoken dialogue systems engineering introduced the concept of Meta-users in *User-guided system development in ISLE: Interactive Spoken Language Education* [90] (Atwell et al 2000). Meta-users are experts on the needs of users; in our case, English language teachers who guided systems designers on the needs of our end-users, English language learners. We also developed a novel corpus for the ISLE project, *the ISLE corpus: Italian and German spoken learner's English* [111] (Atwell et al 2003). This is a collection of recordings of English language learners' spoken English utterances, augmented with graphemic and phonetic transcripts, error-tagging and prosodic markup. The ISLE corpus is a unique resource for research on non-native English learner errors in pronunciation and prosody. Our research on applications in English language teaching was followed up by other researchers, e.g. (Oba and Atwell 2003), (Gabrielatos 2005), (Tepperman and Narayanan 2005a,b, 2008), (Al-Sulaiti and Atwell 2006).

Collaborating with other researchers in a joint project allowed me to take ideas originally developed for English and try them out on other languages, specifically topic G: **Arabic corpus linguistics**. The ABC project, Arabic By Computer, resulted in *an Arabic text database and glossary system for students* [29] (Brockett et al 1989). Computer and corpus resources were starting to be used for English language teaching, but these technologies were novel in Arabic language teaching; simply storing, processing and displaying Arabic script was a significant problem at the time. Few Arabic academics or departments had access to or interest in using computers; and other Arabic academics did not take up our system. So, I stopped working on Arabic, till later …

My first funded research project was on topic H: **Applications in Computing teaching and research.** Supported by Leeds University internal funding, we developed the *Leeds Unix Knowledge Expert: a domain-dependent Expert System generated with domain-independent tools* [19] (Cliff and Atwell 1987). We used the Unix online manual as a corpus, and LUKE was an English-language query front-end to this corpus. Students learning the Unix "language" could ask a simple question such as "how do I delete a file?"; LUKE used simple partial parsing and pattern-matching to find a matching Unix manual page presenting the appropriate Unix command as an answer. Another project which combined machine-learning of language models and student learning was our project on Customising a Copying-Identifier for Science Student Reports. A corpus of science student lab reports was used in corpus-based development and evaluation of plagiarism-detection systems applied to the specific task of *detecting student copying in a corpus of science laboratory reports* [114] (Atwell et al 2003). First-year science students have to learn the specialised language used in reporting on laboratory experiments: lab reports follow a quite rigid, standardised format, and can be expected to have significant overlap in content. However, too much overlap amounts to plagiarism. Our research resulted in corpus-based algorithms to detect this specialised form of plagiarism; and also, some pedagogical findings on the key individual contributions to expect of students: the Introduction, Methods, and Results sections of lab reports can justifiably overlap, but the Discussion and Conclusions should reflect individual insights. Our research has been followed up by others, eg (Gabrielatos 2005), (Thelwall 2005), (Abu Shawar and Atwell 2007).

I have also published a range of other papers reporting on results of grant-funded research projects, too many to cover in detail here, including: [22, 24, 30, 34, 35, 58, 59, 60, 62, 69, 71, 74, 80, 84, 89, 101, 102, 111, 114, 122, 146, 155, 159].

Chapter 3:
Working with Research Students
in Corpus Linguistics and Language Learning

Another highly productive mode of research has been supervision of research students, leading to jointly-authored research papers: I have supervised 18 research students to successful completion of their Theses, and hope to supervise many more in the future. I helped formulate the research plans (as in my experience few if any PhD applicants start out with a clear idea of what they want to do), and guided and advised the student. As with research-grant projects, the detailed implementation of the research was down to the research students. To spare the examiners (and any other readers) I mention only a few of these below, but this should not be taken as a criticism of the students who co-authored the papers left out. Of course, all their Theses have been significant contributions to research (or else they would not have been awarded their degrees!); I have limited this collection to a sample of jointly-authored papers which illustrate the theme of Corpus Linguistics and Language Learning: using corpora for machine learning of language models, and/or for assisting student learning of language.

I supervised several Phd students researching aspects of topic B: **Machine Learning of grammatical patterns from a corpus**. For example, [61] (Hughes and Atwell 1995) presented an approach to *the automated evaluation of inferred word classifications.* In supervised machine learning, the results can be compared to the training data, but it is not obvious how to evaluate the clusters found by unsupervised machine learning methods. This paper showed that unsupervised learning of word-classes from corpus data results in novel word-classifications; and these can be evaluated by comparing the classes against PoS-tag classes in the tagged LOB corpus. A range of different clustering algorithms and similarity metrics were compared, to find the combination which most closely resembles the LOB tag-set partition of words into classes. Other researchers who built on this approach to evaluation include (Sutcliffe et al 1996), (Zavrel 1996), (Thompson and Brew 1996), (McMahon and Smith 1996, 1998), (Dejean 2000).

I supervised several PhD students investigating the bootstrapping of resources from textual corpus data for topic E: **Machine Learning of semantic language models**. Corpus-based language modelling can be applied to semantics and pragmatics as well as syntax. [78] (Churcher et al 1997) presented *the semantic/pragmatic annotation of an air traffic control corpus for use in speech recognition.* A corpus of transcripts of air traffic control dialogues was annotated at several linguistic levels, to use in building dialogue models for speech recognition with the specialised sublanguage of air traffic control. [98] (Demetriou and Atwell 2001) described *a domain-independent semantic tagger for the study of meaning associations in English text,* derived from LDOCE, the Longman Dictionary of Contemporary English. In LDOCE, each word has a definition, written using a restricted terminology, the Longman Defining Vocabulary. This means that the set of words in each word-definition, stripped of stop-words, can serve as a set of semantic primitives representing the meaning of the word in computational analysis. The bundle of semantic primitives is in effect a semantic tag. The meaning association between any two words is measured by the overlap in semantic primitives: the number of Longman Defining Vocabulary words appearing in the definitions of both of the words in question. The paper describes a Prolog implementation and some applications. [103] (Duan and Atwell 2002) applied this LDOCE-derived lexical semantic model to measure *semantic association between web pages - a lexical knowledge based method.* Another use of corpora to derive language models involves *using corpora in machine-learning chatbot systems* [136] (Abu Shawar and Atwell 2005). The AIML chatbot architecture requires a set of input-reply patterns or templates, which allow the chatbot to find an appropriate response to any user input. This paper proposed that a set of input-reply patterns could be directly extracted from a dialogue corpus, and hence a chatbot can be retrained to chat in the language of any given dialogue corpus. This technique was illustrated by a range of chatbots trained in different genres, topics, and even different languages. Other corpus-based semantic researchers built on our approach to bootstrapping semantic knowledge and resources from text, e.g. (Piao et al 2004), (Rayson et al 2004), (Campbell-Laird 2004), (Geeb 2007), (Ravi and Kim 2007).

I have been able to renew my interest in topic G: **Arabic corpus linguistics**, through supervision of several research student projects which used Arabic corpora for machine learning of language models, and/or for assisting Arabic language teaching and learning. [123] (Atwell et al 2004) surveyed growing interest in Arabic Corpus Linguistics, and presented *a review of Arabic corpus analysis tools*. [141] (Al-Sulaiti and Atwell 2006) discussed the limitations of a range of existing Arabic

corpora, leading to *the design of a corpus of contemporary Arabic.* [140] (Roberts et al 2006) demonstrated the problems a number of existing concordancers have with analysis of Arabic corpora, and presented an alternative: *aConCorde: Towards an open-source, extendable concordancer for Arabic.* This research has started to have a wider impact, for example (Zribi et al 2007), (Smith et al 2008), (Abbes and Dichy 2008).

I have also published a range of other papers in collaboration with research students, too many to cover in detail, including: [41, 42, 44, 49, 50, 51, 52, 54, 56, 63, 64, 68, 72, 75, 76, 77, 82, 83, 85, 86, 88, 91, 92, 93, 100, 103, 104, 106, 107, 108, 110, 112, 113, 116, 117, 118, 119, 120, 121, 124, 127, 130, 131, 132, 133, 134, 135, 137, 138, 143, 145, 147, 148, 151, 153, 154, 155, 156, 157, 158]

Chapter 4:
Surveying Research in Corpus Linguistics and Language Learning

As a concluding section, I include survey papers aimed at introducing new developments in computing and corpus linguistics to a wider audience, particularly in English language teaching. Some of the "new developments" may look rather dated to the reader in 2008 (and will undoubtedly look even more dated to future readers!); this may help to make readers aware of the limited computational resources and technologies which formed the context of some earlier research papers in this collection.

Two survey papers focussed on topic F: **Applications in English language teaching**. The first of these survey papers [10] (Atwell 1986) was written for a British Council symposium on computers in English language teaching and research. The survey looked *beyond the micro, towards advanced software for research and teaching from computer science and artificial intelligence;* and aimed to introduce state-of-the-art issues from Computer Science and Artificial Intelligence to English language teachers and researchers. *The Language Machine* [81] (Atwell 1999) was a broader survey of language engineering and corpus linguistics techniques and applications, aimed at British Council staff and clients around the world, as a contribution to the "English 2000" initiative to promote the British English language industry and British English language teaching in the new millennium. Neither of these papers have had much impact in terms of citations by other researchers; however, the aim was to reach British Council "customers" and English language teaching practitioners, and this impact is less tangible or measurable.

Other survey papers dealt with topic H: **Applications in Computing teaching and research.** From 1990 to 1996, I took some time out on internal secondment from my Lectureship in a series of externally-funded research projects and initiatives: a postdoctoral Advanced Research Fellowship funded by the Science and

Engineering Research Council; a Senior Research Fellowship funded by the Defence Research Agency of the Ministry of Defence; National Coordinator of the Knowledge Based Systems Initiative, funded by the UK Universities Funding Council which became the Higher Education Funding Councils; and National Coordinator for Computer Analysis of Language And Speech projects within the New Technologies Initiative funded by the UK Higher Education Funding Councils. One of my aims in these secondments was to promote knowledge-based systems and language technologies to a broader academic audience. I include three survey papers from this period: [37] (Atwell 1993) introduced *the HEFCs' Knowledge Based Systems Initiative,* and a selection of the projects I coordinated; [38] (Atwell and Lajos 1993) looked at *Knowledge and constraint management: large scale applications*, including an early recognition of the emerging World Wide Web as a potentially useful research resource; and [39] (Atwell 1993) overviewed *linguistic constraints for large-vocabulary speech recognition*, introducing a range of corpus linguistics knowledge-sources at the levels of lexis, syntax and semantics. These Knowledge Based Systems Initiative papers (and, as far as I can tell, papers from other KBSI projects as well) had little impact in terms of citations by other researchers, which was disappointing both for me and for the KBSI sponsors. The lack of citations for the British Council initiatives is understandable as the target readership was not researchers; but the KBSI programme was supposed to promote KBS takeup by UK university academics. On reflection, it seems that funding "technology promotion" programmes may not be an effective use of Higher Education Funding Council funds.

I have also published a number of other papers which survey issues relating to corpus linguistics and language learning, too many to cover in detail in this submission, including: [14, 20, 36, 45, 46, 48, 53, 67, 70, 79, 94, 97, 99, 105, 109, 129, 142, 152, 154].

References (i)
Publications by Eric Atwell 1980-2008

The references are ordered most recent first, as on my web-site; for an up-to-date list of my publications, see http://www.comp.leeds.ac.uk/eric/

[159] Atwell, Eric; Abu Shawar, Bayan. *An AI-inspired intelligent agent/student architecture to combine language resources research and teaching* in: **Proceedings of LREC'08: Language Resources and Evaluation Conference**. 2008.

[158] Brierley, Claire; Atwell, Eric. *ProPOSEL: A prosody and POS English lexicon for language engineering* in: **Proceedings of LREC'08: Language Resources and Evaluation Conference**. 2008.

[157] Sawalha, Majdi; Atwell, Eric. *Comparative evaluation of Arabic language morphological analysers and stemmers* in: **Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics**. 2008.

[156] Brierley, Claire; Atwell, Eric. *ProPOSEL: a human-oriented prosody and PoS English lexicon for machine learning and NLP* in: **Proceedings of COLING 2008 CogALex Workshop on Cognitive Aspects of the Lexicon**. 2008.

[155] Atwell, Eric; Brierley, Claire. *Combining teaching and research in text-mining from social and cultural data* in: **Proceedings of 4th International Conference on e-Social Science, Workshop on Text Mining Applications in the Social Sciences**. 2008.

[154] Atwell, Eric; Abbas, Noorhan; Abu Shawar, Bayan; Alsaif, Amal; Al-Sulaiti, Latifa; Roberts, Andrew; Sawalha, Majdi. *Mapping Middle Eastern and North African diasporas: Arabic corpus linguistics research at the University of Leeds* in: **Proceedings of BRISMES'08 Conference.** 2008.

[153] Abu Shawar, Bayan; Atwell, Eric. *Different measurement metrics to evaluate a chatbot system* in: **Proceedings of the NAACL'07 Workshop: Bridging the Gap: Academic and Industrial Research in Dialog Technologies**, pp. 89-96 Association for Computational Linguistics. 2007.

[152] Abu Shawar, Bayan; Atwell, Eric. *Chatbots: Sind Sie wirklich nu"tzlich? (are they really useful?).* **LDV-Forum Journal for Computational Linguistics and Language Technology**, vol. 22, pp. 31-50. 2007.

[151] Abu Shawar, Bayan; Atwell, Eric. *Fostering language learner autonomy via adaptive conversation* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[150] Atwell, Eric; Roberts, Andy. *CHEAT: combinatory hybrid elementary analysis of text* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[149] Atwell, Eric. *A cross-language methodology for corpus Part-of-Speech tag-set development* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[148] Brierley, Claire; Atwell, Eric. *Corpus-based evaluation of prosodic phrase break prediction* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[147] Nancarrow, Owen; Atwell, Eric. *A comparative study of the tagging of adverbs in modern English corpora* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[146] Atwell, Eric; Arshad, Junaid; Lai, Chien-Ming; Nim, Lan; Rezapour Ashegi, Noushin; Wang, Josiah; Washtell, Justin. *Which English dominates the World Wide Web, British or American?* in: **Proceedings of Corpus Linguistics 2007**. 2007.

[145] Brierley, Claire; Atwell, Eric. *Prosodic phrase break prediction: problems in the evaluation of models against a gold standard*. **TAL Journal: Traitement Automatique des Langues**, vol. 48.1. 2007.

[144] Brierley, Claire; Atwell, Eric. *An approach for detecting prosodic phrase boundaries in spoken English*. **ACM Crossroads journal**, vol. 14.1. 2007.

[143] Brierley, Claire; Atwell, Eric. *Using Nltk_lite's Chunk Parser to Detect Prosodic Phrase Boundaries in the Aix-MARSEC Corpus of Spoken English* University of Leeds, School of Computing research report 2007.02. 2007.

[142] Atwell, E; Shawar, B A; Roberts, A; Al-Sulaiti, L. *Unsupervised learning of linguistic significance* in: Barber, S, Baxter, P D, Mardia, K V & Walls, R E (editors) **Interdisciplinary Statistics and Bioinformatics**, pp. 107-108 University of Leeds. 2006.

[141] Al-Sulaiti, Latifa; Atwell, Eric. *The design of a corpus of contemporary Arabic*. **International Journal of Corpus Linguistics**, vol. 11, pp. 135-171. 2006.

[140] Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. *aConCorde: Towards an open-source, extendable concordancer for Arabic*. **Corpora journal**, vol. 1, pp. 39-57. 2006.

[139] Atwell, Eric; Roberts, Andrew. *Combinatory hybrid elementary analysis of text* in: Kurimo, M, Creutz, M & Lagus, K (editors) **Proceedings of the PASCAL**

Challenge Workshop on Unsupervised Segmentation of Words into Morphemes. 2006.

[138] Abu Shawar, B; Atwell, E. *Modelling turn-taking in a corpus-trained chatbot* in: Fisseni, B, Schmitz, H-C, Schroder, B & Wagner, P (editors) **Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen**, pp. 549-561 Peter Lang Verlag. 2005.

[137] Abu Shawar, Bayan; Atwell, Eric. *A chatbot system as a tool to animate a corpus*. **ICAME Journal**, vol. 29, pp. 5-24. 2005.

[136] Abu Shawar, Bayan; Atwell, Eric. *Using corpora in machine-learning chatbot systems*. **International Journal of Corpus Linguistics**, vol. 10, pp. 489-516. 2005.

[135] Abu Shawar, Bayan; Atwell, Eric; Roberts, Andrew. *FAQchat as in Information Retrieval system* in: Vetulani, Z (editors) **Human Language Technologies as a Challenge for Computer Science and Linguistics: Proceedings of the 2nd Language and Technology Conference**, pp. 274-278. 2005.

[134] Abu Shawar, Bayan; Atwell, Eric. *Re-creating dialogues from corpora* in: **Proceedings of CL2005 Workshop on Using Corpora for Natural Language Generation**. 2005.

[133] Al-Sulaiti, Latifa; Roberts, Andrew; Atwell, Eric. *The use of corpora and concordance in the teaching of contemporary Arabic* in: **Proceedings of EuroCALL 2005**. 2005.

[132] Al-Sulaiti, Latifa; Atwell, Eric. *Extending the corpus of contemporary Arabic* in: **Proceedings of Corpus Linguistics 2005**. 2005.

[131] Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. *aConCorde: towards a proper concordance of Arabic* in: **Proceedings of Corpus Linguistics 2005**. 2005.

[130] Elliott, Debbie; Atwell, Eric; Hartley, Tony. *Using corpora to automatically detect untranslated and outrageous words in machine translation output* in: **Proceedings of Corpus Linguistics 2005**. 2005.

[129] Atwell, Eric. *Web chatbots: the next generation of speech systems?*. **European CEO journal**, vol. November-December, pp. 142-144. 2005.

[128] Atwell, Eric. *Sleeping with the enemy: infiltrating AI into the broader curriculum* in: Fasli, M (editors) **Proceedings of 1st UK Workshop on Artificial Intelligence in Education** Higher Education Academy. 2005.

[127] Elliott, Debbie; Atwell, Eric; Hartley, Anthony. *Compiling and using a shareable parallel corpus for MT evaluation* in: **Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora. Fourth International Conference on Language Resources and Evaluation (LREC)**, pp. 18-21. 2004.

[126] Atwell, Eric. *Clustering of word types and unification of word tokens into grammatical word-classes* in: Bel, B &Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 1**, pp. 27-32 ATALA. 2004.

[125] Atwell, Eric. *Machine learning approaches to analysis of corpora* in: Bel, B & Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2**, pp. 553-558 ATALA. 2004.

[124] Abu Shawar, Bayan; Atwell, Eric. *An Arabic chatbot giving answers from the Qur'an* in: Bel, B & Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2**, pp. 197-202 ATALA. 2004.

[123] Atwell, Eric; Al-Sulaiti, Latifa; Al-Osaimi, Saleh; Abu Shawar, Bayan. *A review of Arabic corpus analysis tools* in: Bel, B & Marlien, I (editors) **Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2**, pp. 229-234 ATALA. 2004.

[122] van Zaanen, Menno; Roberts, Andrew; Atwell, Eric. *A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation* in: **Proceedings of LREC'04 Workshop on The Amazing Utility of Parallel and Comparable Corpora**, pp. 58-61 European Language Resources Association. 2004.

[121] Abu Shawar, Bayan; Atwell, Eric. *A chatbot as a novel corpus visualization tool* in: **Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC04)**, vol. IV, pp. 2057-2060. 2004.

[120] Abu Shawar, Bayan; Atwell, Eric. *Evaluation of chatbot systems* in: **Proceedings of Eighth Maghrebian Conference on Software Engineering and Artificial Intelligence**. 2004.

[119] Abu Shawar, Bayan; Atwell, Eric. *Accessing an information system by chatting* in: Meziane, F & Metais,E (editors) **Natural Language Processing and Information Systems**, pp. 407-412 Springer. 2004.

[118] Al-Sulaiti, Latifa; Atwell, Eric. *Designing and developing a corpus of contemporary Arabic* in: **TALC 2004: Proceedings of the sixth Teaching And Language Corpora conference**, pp. 92-93. 2004.

[117] Elliott, Debbie; Hartley, Anthony; Atwell, Eric. *A fluency error categorization scheme to guide automated machine translation evaluation* in: Frederking, R E & Taylor, K B (editors) **Machine Translation: From Real Users to Research** , pp. 64-73 Springer. 2004.

[116] Abu Shawar, Bayan; Atwell, Eric. *Using dialogue corpora to train a chatbot* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 681-690 Lancaster University. 2003.

[115] Atwell, Eric. *A new machine learning algorithm for neoposy: coining new parts of speech* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 43-47 Lancaster University. 2003.

[114] Atwell, Eric; Gent, Paul; Medori, Julia; Souter, Clive. *Detecting student copying in a corpus of science laboratory reports* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 48-53 Lancaster University. 2003.

[113] Babych, Bogdan; Hartley, Anthony; Atwell, Eric. *Statistical modelling of MT output corpora for information extraction* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 62-70 Lancaster University. 2003.

[112] Elliott, Debbie; Hartley, Anthony; Atwell, Eric. *Rationale for a multilingual aligned corpus for machine translation evaluation* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 191-200 Lancaster University. 2003.

[111] Atwell, Eric; Howarth, Peter; Souter, Clive. *The ISLE corpus: Italian and German spoken learner's English*. **ICAME Journal**, vol. 27, pp. 5-18. 2003.

[110] Oba, Toshifumi; Atwell, Eric. *Using the HTK speech recogniser to anlayse prosody in a corpus of German spoken learner's English* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 591-598 Lancaster University. 2003.

[109] Hu, Xunlei Rose; Atwell, Eric. *A survey of machine learning approaches to analysis of large corpora* in: Simov, K & Osenova, P (editors) **Proceedings of SProLaC: Workshop on Shallow Processing of Large Corpora**, pp. 45-52 Lancaster University. 2003.

[108] Roberts, Andrew; Atwell, Eric. *The use of corpora for automatic evaluation of grammar inference systems* in: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) **Proceedings of CL2003: International Conference on Corpus Linguistics**, pp. 657-661 Lancaster University. 2003.

[107] Abu Shawar, Bayan; Atwell, Eric. *Machine learning from dialogue corpora to generate chatbots*. **Expert Update journal**, vol. 6, pp. 25-30. 2003.

[106] Abu Shawar, Bayan; Atwell, Eric. *Using the corpus of Spoken Afrikaans to generate an Afrikaans chatbot*. **Southern African Linguistics and Applied Language Studies journal**, vol. 21, pp. 283-294. 2003.

[105] Atwell, Eric; Abu Shawar, Bayan; Babych, Bogdan; Elliott, Debbie; Elliott, John; Gent, Paul; Hartley, Anthony; Hu, Xunlei Rose; Medori, Julia; Oba, Toshifumi; Roberts, Andy; Sharoff, Serge; Souter, Clive. *Corpus Linguistics, Machine Learning and Evaluation: Views from Leeds* University of Leeds, School of Computing research report 2003.02. 2003.

[104] Al-Sulaiti, Latifa; Atwell, Eric. *The Design of a Corpus of Contemporary Arabic (CCA)* University of Leeds, School of Computing research report 2003.11. 2003.

[103] Duan, Xiao Yuan; Atwell, Eric. *Semantic association between web pages - a lexical knowledge based method* in: **Proceedings of the 5th Annual CLUK Colloquium: Computational Linguistics in the United Kingdom, (CLUK 2002, University of Leeds, UK)**, pp. 66-76. University of Leeds. 2002.

[102] Medori, Julia; Atwell, Eric; Gent, Paul; Souter, Clive. *Customising a copying-identifier for biomedical science student reports: comparing simple and smart analyses* in: O'Neill, M, Sutcliffe, R, Ryan, C, Eaton, M, & Griffith, N (editors) **Artificial Intelligence and Cognitive Science, Proceedings of AICS02**, pp. 228-233 Springer. 2002.

[101] Lievesley, Sam; Atwell, Eric. *NAIL: artificial intelligence software for learning natural language* in: Adriaans, P, Fernau, H & van Zaanen, M (editors) **Grammatical Inference Algorithms and Applications, Proceedings of ICGI 2002**, pp. 306-309 Springer. 2002.

[100] Abu Shawar, Bayan; Atwell, Eric. *A Comparison Between Alice and Elizabeth Chatbot Systems* University of Leeds, School of Computing research report 2002.19. 2002.

[99] Roberts, Andrew; Atwell, Eric. *Unsupervised Grammar Inference Systems for Natural Language* University of Leeds, School of Computing research report 2002.20. 2002.

[98] Demetriou, G; Atwell, E. *A domain-independent semantic tagger for the study of meaning associations in English text* in: Harry Bunt, Ielka van der Sluis and Elias Thijsse (editors) **Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4)**, pp. 67-80. 2001.

[97] Liang, Y; Zhang, X; Fugere, B; Atwell, E. *The Internet and Web Design - Questions and Answers*., 401pp. 2001.

[96] Atwell, E; Elliott, J. *A corpus for interstellar communication* in: Rayson, P, Wilson, A, McEnery, T, Hardie, A & Khoja, S (editors) **Proceedings of CL2001: International Conference on Corpus Linguistics**, pp. 31-39. 2001.

[95] Atwell, E; Elliott, J. *Corpus linguistics and the design of a response message* in: **Proceedings of IAC'2001: the 52nd International Astronautical Congress**, pp. 9.2.08. 2001.

[94] Atwell, E. *Language engineering and electronic media* in: **Proceedings of the International European Year of Languages Seminar: Problems of Language and Identity in a Changing Europe**. 2001.

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

[92] Elliott, John; Atwell, Eric; Whyte, Bill. *A toolkit for visualisation of combinatorial constraint phenomena in linguistically interpreted corpora* in: **Proceedings of the 4th Annual CLUK Colloquium: Computational Linguistics in the United Kingdom**, pp. 90-95. 2001.

[91] Elliott, John; Atwell, Eric; Whyte, Bill. *First stage identification of syntactic elements in an extraterrestrial signal* in: **Proceedings of IAC'2001: the 52nd International Astronautical Congress**, pp. 9.2.07. 2001.

[90] Atwell, E; Howarth, P; Souter, C; Baldo, P; Bisiani, R; Bonaventura, P; Herron, D; Menzel, W; Morton, R; Wick, H. *User-guided system development in ISLE: Interactive Spoken Language Education*. **Natural Language Engineering journal**, vol. 6, pp. 229-241. 2000.

[89] Menzel, W; Atwell, E; Bonaventura, P; Herron, D; Howarth, P; Morton, R; Souter, C. *The ISLE Corpus of non-native spoken English* in: Gavrilidou, M, Carrayannis, G, Markantionadou, S, Piperidis, S & Stainhaouer, G (editors) **Proceedings of LREC2000: Language Resources and Evaluation Conference**, vol. 2, pp. 957-964 European Language Resources Association. 2000.

[88] Demetriou, G; Atwell, E; Souter, C. *Using lexical semantic knowledge from machine readable dictionaries for domain independent language modelling* in: Gavrilidou, M, Carrayannis, G, Markantionadou, S, Piperidis, S & Stainhaouer, G (editors) **Proceedings of LREC2000: Language Resources and Evaluation Conference**, vol. 2, pp. 777-782 European Language Resources Association. 2000.

[87] Atwell, E; Demetriou, G; Hughes, J; Schriffin, A; Souter, C; Wilcock, S. *A comparative evaluation of modern English corpus grammatical annotation schemes*. **ICAME Journal**, vol. 24, pp. 7-23. 2000.

[86] Elliott, John; Atwell, Eric. *Is anybody out there?: the detection of intelligent and generic language-like features*. **Journal of the British Interplanetary Society**, vol. 53, pp. 13-22. 2000.

[85] Elliott, John; Atwell, Eric; Whyte, Bill. *Language identification in unknown signals* in: **Proceedings of COLING'2000 - 18th International Conference on Computational Linguistics**, pp. 1021-1026 Morgan Kaufman. 2000.

[84] Atwell, Eric; Demetriou, G; Hughes, J; Souter, C; Wilcock, S. *Comparing linguistic interpretation schemes for English corpora* in: Thorsten Brants (editors) **Proceedings of COLING LINC-2000 Workshop on Linguistically Interpreted Corpora**, pp. 1-10. 2000.

[83] Elliott, John; Atwell, Eric; Whyte, Bill. *Increasing our ignorance of language: identifying language structure in an unknown signal* in: Daelemans, W (editors) **Proceedings of CoNLL-2000: International Conference on Computational Natural Language Learning**, pp. 25-30 Association for Computational Linguistics. 2000.

[82] Elliott, J; Atwell, E. *Language in signals: the detection of generic species-independent intelligent language features in symbolic and oral communications* in: **Proceedings of the 50th International Astronautical Congress**, pp. 9pp. International Astronautical Federation. 1999.

[81] Atwell, Eric. *The Language Machine*., 64pp The British Council. 1999.

[80]Herron, Daniel; Menzel, Wolfgang; Atwell, Eric; Bisiani, Roberto; Daneluzzi, Fabio; Morton, Rachel; Schmidt, Jurgen. *Automatic localization and*

diagnosis of pronunciation errors for second-language learners of English in: **Proceedings of EUROSPEECH'99**. 1999.

[79] Atwell, E S. *What can SALT offer the English teaching professional?*. **English Teaching Professional journal**, vol. 7, pp. 46-47. 1998.

[78] Churcher, G E; Atwell, E S; Souter, C. *The semantic/pragmatic annotation of an air traffic control corpus for use in speech recognition* in: Ljung, M (editors) **Corpus-based Studies in English: Papers from Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)**, pp. 353-374 Rodopi. 1997.

[77] Churcher, G E; Atwell, E S; Souter, C. *Generic template for the evaluation of dialogue management systems* in: **Proceedings of EUROSPEECH'97**, vol. 4, pp. 2247 - 225. 1997.

[76] Demetriou, G; Atwell, E S; Souter, C. *Large-scale lexical semantics for speech recognition support* in: **Proceedings of EUROSPEECH'97**, vol. 5, pp. 2755 - 27. 1997.

[75] Churcher, G E; Atwell, E S; Souter, C. *A generic template to evaluate integrated components in spoken dialogue systems* in: Hirschberg, J, Kamm, C and Walker, M (editors) **Proceedings of ACL/EACL'97 Workshop on Interactive Spoken Dialog Systems: bringing speech and NLP together in real applications**, pp. 9-16. 1997.

[74] Atwell, E S; Demetriou, G; Hughes, J S; Schriffin, A; Souter, C; Wilcock, S P. *Tagging email with 8 tagsets: lessons on evaluation* in: Gaizauskas, R (editors) **Evaluation in Speech and Language Technology: Proceedings of the SALT Club Workshop**, pp. 17-25. 1997.

[73] Atwell, E S; Sutcliffe, R. *Industrial parsing of software manuals: empirical qualitative comparison of parsers and parsing schemes* in: Gaizauskas, R (editors) **Evaluation in Speech and Language Technology: Proceedings of the SALT Club Workshop**, pp. 26-28. 1997.

[72] Churcher, G E; Atwell, E S; Souter, C. *A generic template to evaluate integrated components in spoken dialogue systems* in: Gaizauskas, R (editors) **Evaluation in Speech and Language Technology: Proceedings of the SALT Club Workshop**, pp. 51-58. 1997.

[71] Demetriou, G; Atwell, E S; Souter, C. *Measuring the uncertainty in grammatical classification* in: Gaizauskas, R (editors) **Evaluation in Speech and Language Technology: Proceedings of the SALT Club Workshop**, pp. 80-82. 1997.

[70] Atwell, E S. *The world of language: speech and language technology* in: Bowers, R (editors) **The World of Language: preliminary studies** The World of Language / British Council. 1997.

[69] Churcher, G E; Atwell, E S; Souter, C. *Using a commercial speech recogniser within the domain of Air Traffic Control* **WEB-SLS: WWW European Student Journal of Language and Speech** 97-03. 1997.

[68] Zhang, X; Liu, J; Atwell, E S. *A multilingual information retrieval tool hierarchy for a WWW virtual corpus* in: Spyropoulos, C (editors) **Proceedings of the 2nd Workshop on Multilinguality in the Software Industry : the AI Contribution(MULSAIC'97)** URL: http://www.iit.nrcps.ariadne-t.gr/skel/mulsaic97proc/Session2/zhang.ZIP. 1997.

[67] Churcher, Gavin E; Atwell, Eric S; Souter, Clive. *Dialogue Management Systems: a Survey and Overview* University of Leeds, School of Computing Research Report 1997.06. 1997.

[66] Atwell, E S. *Machine learning from corpus resources for speech and handwriting recognition* in: Thomas, J & Short, M (editors) **Using Corpora for Language Research : Studies in Honour of Geoffrey Leech**, pp. 151-166 Longman. 1996.

[65] Atwell, E S. *Comparative evaluation of grammatical annotation models* in: Sutcliffe, R, Koch, H & McElligott, A (editors) **Industrial Parsing of Software Manuals**, pp. 25-46 Rodopi. 1996.

[64] Churcher, G E; Atwell, E S; Souter, C. *Dialogues in air traffic control* in: **Proceedings of TWLT 11 - the 11th Twente Workshop on Language Technology**. 1996.

[63] Schillo, M; Atwell, E S; Souter, C; Denson, A. *Language modelling for the in-car intelligent personal assistant* in: Moghrabi, C (editors) **Proceedings of NLP+IA96 : International Conference on Natural Language Processing and Industrial Applications**. 1996.

[62] Churcher, Gavin E; Souter, Clive; Atwell, Eric S. *Using a Commercial Speech Recogniser Within the Domain of Air Traffic Control* University of Leeds, School of Computing Research Report 1996.04. 1996

[61] Hughes, J S; Atwell, E S. *The automated evaluation of inferred word classifications* in: Cohn, A G (editors) **ECAI-94 Proceedings of the 11th European Conference on Artificial Intelligence**, pp. 535-539 John Wiley & Sons. 1995.

[60] Hughes, J S; Souter, C; Atwell, E S. *Automatic extraction of tagset mappings from parallel annotated corpora* in: Tzoukermann E & Armstrong, S (editors) **From Text to Tags: Issues in Multilingual Language Analysis, Proceedings of ACL-SIGDAT Workshop**, pp. 10-17. 1995.

[59] Atwell, E S; Churcher, G; Souter, C. *Developing a corpus-based grammar model within a commercial continuous speech recognition package* in: Collingham, R (editors) **Proceedings of the Institute of Acoustics Workshop on Integrating Speech Recognition and Natural Language Processing Systems**, pp. 5-6. 1995.

[58] Atwell, Eric; Churcher, Gavin; Souter, Clive. *Developing a corpus-based grammar model within a continuous commercial speech recognition package* University of Leeds, School of Computing research report 1995.20. 1995.

[57] Atwell, E S; Hughes, J S; Souter, C. *AMALGAM: Automatic mapping among lexicogrammatical annotation models* in: Klavans, J (editors) **The Balancing Act: Combining Symbolic and Statistical Approaches to Language - Proceedings of the ACL Workshop**, pp. 21-28 Association for Computational Linguistics. 1994.

[56] Demetriou, G; Atwell, E S. *Machine-readable, non-compositional semantics for domain independent speech or text recognition* in: **Proceedings of the 2nd Hellenic-European Conference on Mathematics and Informatics**, pp. 103-104. 1994.

[55] Atwell, E S; Hughes, J S; Souter, C. *A unified multicorpus for training syntactic constraint models* in: Evett, L & Rose,T (editors) **Computational Linguistics for Speech and Handwriting Recognition AISB'94 Workshop**, pp. 111-118 University of Leeds/AISB. 1994.

[54] Demetriou, G; Atwell, E S. *A semantic network for large vocabulary speech recognition* in: Evett, L & Rose,T (editors) **Computational Linguistics for Speech and Handwriting Recognition AISB'94 Workshop**, pp. 21-28 University of Leeds/AISB. 1994.

[53] Demetriou, G; Atwell, E S. *Semantics in speech recognition and understanding : a survey* in: Evett, L & Rose,T (editors) **Computational Linguistics for Speech and Handwriting Recognition AISB'94 Workshop** University of Leeds/AISB. 1994.

[52] Hughes, J S; Atwell, E S. *A methodical approach to word class formation using automatic evaluation* in: Evett, L & Rose,T (editors) **Computational**

Linguistics for Speech and Handwriting Recognition AISB'94 Workshop, pp. 41-48 University of Leeds/AISB. 1994.

[51] Jost, U; Atwell, E S. *A hierarchical , mutual-information based probabilistic language model* in: Evett, L & Rose,T (editors) **Computational Linguistics for Speech and Handwriting Recognition AISB'94 Workshop** University of Leeds/AISB. 1994.

[50] Modd, D; Atwell, E S. *A word hypothesis lattice corpus - a benchmark for linguistic constraint models* in: Evett, L & Rose,T (editors) **Computational Linguistics for Speech and Handwriting Recognition AISB'94 Workshop**, pp. 191-197 University of Leeds/AISB. 1994.

[49] Jost, U; Atwell, E S. *Intrinsic error estimation for corpus-trained probabilistic language models* in: Cohn, A G (editors) **ECAI-94 Proceedings of the 11th European Conference on Artificial Intelligence**, pp. 550-554 John Wiley & Sons. 1994.

[48] Souter, C; Atwell, E S. *Using parsed corpora: a review of current practice* in: Oostdijk, N & de Haan, P (editors) **Corpus-based Research into Language**, pp. 143-158 Rodopi. 1994.

[47] Atwell, E S; McKevitt, P. *Pragmatic linguistic constraint models for large-vocabulary speech processing* in: McKevitt, P (editors) **Integrating Speech and Natural Language Processing : AAAI94 Workshop Proceedings**, pp. 58-64 AAAI Press. 1994.

[46] Atwell, E S (editor). **Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledged Based Systems Initiative**, 146pp University of Leeds. 1993.

[45] Souter, C, Atwell, E S (editors) **Corpus-Based Computational Linguistics**, 260pp Rodopi. 1993.

[44] Arnfield, S C; Atwell, E S. *A syntax based grammar of stress sequences* in: Lucas, S (editors) **Grammatical Inference: Theory, Applications and Alternatives**, pp. 71-77 IEE Colloquium Proceedings no. 1993/092. 1993.

[43] Atwell, E S; Arnfield, S C; Demetriou, G; Hanlon, S J; Hughes, J S; Jost, U; Pocock, R; Souter, C; Ueberla, J. *Multi-level disambiguation grammar inferred from English corpus, treebank and dictionary* in: Lucas, S (editors) **Grammatical Inference: Theory, Applications and Alternatives**, pp. 91-97 IEE Colloquium Proceedings no. 1993/092. 1993.

[42] Hughes, J S; Atwell, E S. *Automatically acquiring and evaluating a classification of words* in: Lucas, S (editors) **Grammatical Inference: Theory,**

Applications and Alternatives, pp. 81-88 IEE Colloquium Proceedings no. 1993/092. 1993.

[41] Jost, U; Atwell, E S. *Deriving a probabilistic grammar of semantic markers from unrestricted English text* in: Lucas, S (editors) **Grammatical Inference: Theory, Applications and Alternatives**, pp. 191-198 IEE Colloquium Proceedings no. 1993/092. 1993.

[40] Atwell, E S. *Corpus-based statistical modelling of English grammar* in: Souter, C, Atwell, E S (editors) **Corpus-Based Computational Linguistics: Proceedings of the ICAME International Conference**, pp. 195-214 Rodopi. 1993.

[39] Atwell, E S. *Linguistic constraints for large-vocabulary speech recognition* in: Atwell, E S (editors) **Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledged Based Systems Initiative**, pp. 26-32 University of Leeds. 1993.

[38] Atwell, E S; Lajos, G. *Knowledge and constraint management: large scale applications* in: Atwell, E S (editors) **Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledged Based Systems Initiative**, pp. 21-25 University of Leeds. 1993.

[37] Atwell, E S. *The HEFC's Knowledge Based Systems Initiative*. **Artificial Intelligence and Simulation of Behaviour Quarterly**, vol. 83/84, pp. 29-34. 1993.

[36] Atwell, E S. *Overview of grammar acquisition research* in: Thompson, H. (editors) **Workshop on Sublanguage Grammar and Lexicon Acquisition for Speech and Language : Proceedings**, pp. 65-70 Human Communication Research Centre, University of Edinburgh. 1992.

[35] Souter, C; Atwell, E S. *A richly annotated corpus for probabilistic parsing* in: Weir, C, Grishman, R (editors) **Proceedings of the AAAI Workshop on Statistically Based NLP Techniques**, pp. 28-38 American Association for Artificial Intelligence. 1992.

[34] Souter, D C; Atwell, E S. *A Richly Annotated Corpus for Probabilistic Parsing* University of Leeds, School of Computing research report 1992.13. 1992.

[33] Atwell, E S; O'Donoghue, T F; Souter, C. *Training Parsers with Parsed Corpora* University of Leeds, School of Computing research report 1991.20. 1991.

[32] Atwell, E S. *Measuring grammaticality of machine-readable text* in: Bahner W, Schildt J, Viehweger (editors) **Proceedings of the XIV International Congress of Linguists**, vol. III, pp. 2275-2277 Akademie-Verlag. 1990.

[31] Atwell, E S. *A lexical database for English learners and users: the Oxford Advanced Learner's Dictionary* in McCrank, L (editor), **Databases in the Humanities and Social Sciences 4: Proceedings of the International Conference**, pp21-34, New Jersey, Learned Information. 1989.

[30] Sampson, G; Haigh, R; Atwell, E S. *Natural language analysis by stochastic optimisation: a progress report on Project APRIL* in **JETAI: Journal of Experimental and Theoretical Artificial Intelligence**, Volume 1, pp271-287. 1989.

[29] Brockett, A; Atwell, E S; Taylor, O; Page, M. *An Arabic text database and glossary system for students* in **Proceedings of the Seminar on Bilingual Computing in Arabic and English**, pp154-162, University of Cambridge. 1989.

[28] Haigh, R; Sampson, G; Atwell, E S. *Project APRIL - a progress report* in **Proceedings of ACL, the 26th Conference of the Association for Computational Linguistics**, pp104-112, New Jersey, ACL. 1988.

[27] Atwell, E S. *Grammatical analysis of English by statistical pattern recognition* in Kittler, J (editor), **Pattern Recognition: Proceedings of the 4th International Conference**, pp626-635, Berlin, Springer-Verlag. 1988.

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

[25] Atwell, E, and Souter, C. *Experiments with a very large corpus-based grammar* in Choueka, Y (ed) **Fifteenth Annual Conference of the Association for Literary and Linguistics Computing (ALLC)**, Jerusalem. 1988.

[24] Cliff, D, and Atwell, E. *A plain English advice system for operating system commands* in Choueka, Y (ed) **Fifteenth Annual Conference of the Association for Literary and Linguistics Computing (ALLC)**, Jerusalem. 1988.

[23] Atwell, E, Sampson, G, and Haigh, R. *A corpus-based statistical parser* in Choueka, Y (ed) **Fifteenth Annual Conference of the Association for Literary and Linguistics Computing (ALLC)**, Jerusalem. 1988.

[22] Cliff, D; Atwell, E. *Leeds Unix Knowledge Expert: a domain-dependent Expert System generated with domain-independent tools.* Computer Studies Research Report 1988.3, University of Leeds. 1988.

[21] Atwell, E. *An expert system for the automatic discovery of particles* in Weydt, H (ed) **Internationaler Kongress uber Sprachpartikeln**, pp.34-35, West Berlin, Free University of Berlin. 1987.

[20] Atwell, E. *Current European research in Computational Linguistics* in **Artificial Intelligence and Simulation of Behaviour Quarterly** no.62 pp.22-23. 1987.

[19] Cliff, D; Atwell, E S. *Leeds Unix Knowledge Expert: a domain-dependent Expert System generated with domain-independent tools* in **BCS-SGES: Newsletter of the British Computer Society Specialist Group on Expert Systems** no.19 pp.49-51. 1987.

[18] Atwell, E S. *Constituent-likelihood grammar* in Garside, R, Sampson, G, Leech, G (editors) **The computational analysis of English: a corpus-based approach**, London, Longman. 1987.

[17] Atwell, E S; Elliot, S. *Dealing with ill-formed English text* in Garside, R, Sampson, G, Leech, G (editors) **The computational analysis of English: a corpus-based approach**, London, Longman. 1987.

[16] Atwell, E S. *How to detect grammatical errors in a text without parsing it* in Maegaard, B (editor), **Proceedings of EACL: the Third Conference of European Chapter of the Association for Computational Linguistics**, New Jersey, ACL. 1987.

[15] Atwell, E S; Drakos, N. *Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text* in Maegaard, B (editor), **Proceedings of EACL: the Third Conference of European Chapter of the Association for Computational Linguistics**, New Jersey, ACL. 1987.

[14] Atwell, E. *Natural Language Computing in the office* in McManus B and Salenieks P (eds) **Computers in tomorrow's society: selected papers from the 2nd BCS Young Professionals Group Conference** pp.34-40, British Computer Society. 1987.

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

[12] Atwell, E S. *Converting the Oxford Advanced Learner's Dictionary into a structured database*in Oakman, R, Pantonial, B (editors), **ICCH87: Eighth International Conference on Computers and the Humanities**, Columbia, South Carolina; Association for Computers and the Humanities. 1987.

[11] Johansson, S; Atwell, E S; Garside, R; Leech, G. **The Tagged LOB Corpus - User Manual**, 160pp, Bergen, Norwegian Computing Centre for the Humanities. 1986. (to accompany the Tagged LOB Corpus on ICAME CD-ROM)

[10] Atwell, E S. *Beyond the micro: advanced software for research and teaching from computer science and artificial intelligence* in Leech, G, Candlin, C (editors) **Computers in English language teaching and research: selected papers from the British Council Symposium,** pp167-183, London, Longman. 1986.

[9] Atwell, E S. *How to detect grammatical errors in a text without parsing it.* Computer Studies Research Report 216, University of Leeds. 1986.

[8] Atwell, E S. *Extracting a Natural Language grammar from raw text.* Computer Studies Research Report 208, University of Leeds. 1986.

[7] Atwell, E S; Leech, G; Garside, R. *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, Amsterdam, Rodopi. 1984.

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

[5] Atwell, E S. *Constituent-Likelihood Grammar* in **ICAME Journal of the International Computer Archive of Modern English** Vol.7. 1983.

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

[3] Johansson, S; Atwell, E S. *Seminar on the use of computers in English language research* in **ICAME Journal of the International Computer Archive of Modern English** Vol.7. 1983.

[2] Atwell, E S. **LOB Corpus Tagging Project: Manual Post-edit Handbook**, 45pp, Departments of Computer Studies and Linguistics, Lancaster University. 1982.

[1] Atwell, E S. **LOB Corpus Tagging Project: Manual Pre-edit Handbook**, Departments of Computer Studies and Linguistics, Lancaster University. 1981.

References (ii)
illustrating the historic context and impact of my research

Abbes, Ramzi and Joseph Dichy. AraConc, an Arabic Concordance Software Based on the DIINAR.1 Language. Proceedings of INFOS'2008 6th International Conference on Informatics and Systems, Cairo.

http://www.fci-cu.edu.eg/infos2008_old/infos/NLP_19_P127-134.pdf

Abu Shawar, Bayan, and Eric Atwell. 2007. Chatbots: Sind Sie wirklich nutzlich? (are they really useful?). LDV-Forum Journal for Computational Linguistics and Language Technology, vol. 22, pp. 31-50.

http://www.ldv-forum.org/2007_Heft1/Bayan_Abu-Shawar_and_Eric_Atwell.pdf

Al-Sulaiti, Latifa; Atwell, Eric. 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171.

http://www.comp.leeds.ac.uk/eric/alsulaiti06ijcl.pdf

Altenberg, Bengt. A bibliography of publications relating to English computer corpora. In Stig Johansson and Anna-Brita Stenstrom (eds), English Computer Corpora: Selected Papers and Research Guide. pp.127-148. Walter de Gruyter, 1991. http://books.google.co.uk/books?id=woopk294GpsC

Andersen, Øistein. 2007. Grammatical error detection using corpora and supervised learning. Proceedings of ESSLLI'2007 19th European Summer School in Logic, Language and Information, Dublin. pp.1-10.

https://www.cs.tcd.ie/esslli2007/content/CD_Contents/content/ss/ss.pdf#page=12

Arnfield, Simon. 1996. Word Class Driven Synthesis of Prosodic Annotations. Proceedings of ICSLP'96 Fourth International Conference on Spoken Language, vol.3 pp. 1978-1980. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=608024

Beckwith, Richard; Christiane Fellbaum, Derek Gross and George Miller. 1991. WordNet: a lexical database organized on psycholinguistic principles. in Uri Zernik (ed), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. pp. 211-232. Lawrence Erlbaum. http://books.google.co.uk/books?id=B4iIRuH1ylcC

Belmore, Nancy. 1991. Tagging Brown with the LOB Tagging Suite. ICAME Journal, volume 15, pp. 63-86.

http://icame.uib.no/archives/No_15_ICAME_Journal_index.pdf

Bigert, Johnny. 2004. Probabilistic Detection of Context-Sensitive Spelling Errors. Proceedings of LREC'2004 Language Resources and Evaluation Conference. Lisbon, Portugal. http://www.sdjt.si/bib/lrec04/pdf/122.pdf

Bigert, Johnny. 2005. Automatic and Unsupervised Methods in Natural Language Processing. PhD thesis, Numerisk Analys och Datalogi, Kungl Tekniska Högskolan, Stockholm, Sweden. http://www.nada.kth.se/~johnny/docs/thesis.pdf

Bod, Rens. 2003. Do all fragments count? Natural Language Engineering journal. Volume 9(4), pp. 307-323.

http://journals.cambridge.org/download.php?file=%2FNLE%2FNLE9_04%2FS1351324903003140a.pdf&code=c6f8f4e6f0f1637f91f4e05f6aebbcf2

Bolt P, Yazdani M. 1998. The Evolution of a Grammar-Checking Program: LINGER to ISCA. Computer Assisted Language Learning journal, 11(1), pp.55-112.

http://pdfserve.informaworld.com/Pdf/AddCoversheet?xml=/mnt/pdfserve/pdfserve/476835-731197585-725289251.xml

Campbell-Laird, K. Aviation English: a review of the language of International Civil Aviation. Proceedings of IPCC 2004 International Professional Communication Conference, pp.253-261.

http://ieeexplore.ieee.org/iel5/9464/30036/01375306.pdf?tp=&isnumber=&arnumber=1375306

Carlberger, J; R Domeij, V Kann, O Knutsson. 2004. The Development and Performance of a Grammar Checker for Swedish: A Language Engineering Perspective. Technical report, Kungl Tekniska Högskolan, Stockholm, Sweden.

http://www.csc.kth.se/tcs/projects/granska/rapporter/granskareport.pdf

Carroll, John, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: A survey and a new proposal. In Proceedings of LREC'98 First International Conference on Language Resources and Evaluation, pages 447–454, Granada, Spain. http://www.informatics.susx.ac.uk/research/groups/nlp/carroll/papers/lre98.pdf

Carroll, John, Guido Minnen and Ted Briscoe. 1999. Corpus Annotation for Parser Evaluation. In Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC), Bergen, Norway.

http://arxiv.org/PS_cache/cs/pdf/9907/9907013v1.pdf

Chapelle, Carol. 1988. Review of The computational analysis of English: a corpus-based approach. TESOL Quarterly, 22(4), pp. 668-669.

Chen, Kuang-hua and Hsin-hsi Chen. 1993. A Probabilistic Chunker. Proceedings of ROCLING VI, pp.99-117.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.5138&rep=rep1&type=pdf

Chen, Si-Qing and Luomai Xu. 1995. A full and efficient machine tractable dictionary for natural language processing: a revised version of the CUVOALD. Computers and the Humanities. Vol.28, pp.141-152.

http://www.springerlink.com/content/vm089034360w70xu/fulltext.pdf

Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu. 1996. SINICA CORPUS: Design Methodology for Balanced Corpora, Proceedings of PACLIC 11th Conference, pp.167-176.

http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/12025/1/PACLIC11-167-176.pdf

Church, Kenneth. 1988. A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of ANLP'88 Second Conference on Applied Natural Language Processing, pp.136-143. http://acl.ldc.upenn.edu/A/A88/A88-1019.pdf

Church, Kenneth; William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. in Uri Zernik (ed), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. pp. 115-164. Lawrence Erlbaum.

http://books.google.co.uk/books?id=B4iIRuH1ylcC

Clark, Alexander. 2001. Unsupervised language acquisition: theory and practice. D.Phil thesis, University of Sussex.

http://arxiv.org/ftp/cs/papers/0212/0212024.pdf

de Ilarraza, Arantza Dıaz; Koldo Gojenola, Maite Oronoz, Maialen Otaegi and Inaki Alegria. 2007. Syntactic Error Detection and Correction in Date Expressions using Finite-State Transducers. In FSMNLP'2007 Sixth International Workshop on Finite-State Methods and Natural Language Processing. Potsdam.

http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1184848110/publikoak/fsmnlp07-8.pdf

Déjean, Hervé. 2000. How to evaluate and compare tagsets: a proposal. Proceedings of LREC'2000 second international conference on language resources and evaluation. http://www.cnts.ua.ac.be/lcg/pdf/dejean.lrec2000.pdf

Demetriou, George. 1993. Lexical disambiguation using Constraint Handling in Prolog (CHIP). In Proceedings of EACL'93 sixth conference of the European chapter of the Association for Computational Linguistics. Utrecht, the Netherlands. Pages: 431 – 436 http://www.aclweb.org/anthology-new/E/E93/E93-1051.pdf

Dickinson, Markus. 2005. Error detection and correction in annotated corpora. Ph.D. thesis, Ohio State University.

http://www.ling.ohio-state.edu/~dickinso/papers/diss/dickinson05-alt.pdf.gz

Dickinson, Markus and Charles Jochim. 2008. A Simple Method for Tagset Comparison. Proceedings of LREC'2008 6th Language Resources and Evaluation Conference.

http://www.lrec-conf.org/proceedings/lrec2008/pdf/210_paper.pdf

Elliott, John. 2007. A post-detection decipherment matrix. Acta Astronautica. 61(7-8), pp. 712-715.

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V1N-4NHD91V-2&_user=65461&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000005458&_version=1&_urlVersion=0&_userid=65461&md5=139b255c48b8ce275a3e4b7323857dfb

Fang, Alex and Gerald Nelson. 1994. Tagging the Survey Corpus: a LOB to ICE Experiment using AUTASYS. Literary and Linguistic Computing. 9(3), pp.189-194. http://llc.oxfordjournals.org/cgi/reprint/9/3/189

Fang, Alex. 1996. The Survey parser: design and development. In Sidney Greenbaum (ed), Comparing English worldwide: the International Corpus of English, pp. 142-160, Oxford: OUP.

http://www.phon.ucl.ac.uk/home/alex/publish2/parsing.ps

Feng, Zhiwei. 2001. Hybrid Approaches for Automatic Segmentation and Annotation of a Chinese Text Corpus. International Journal of Corpus Linguistics, 6(2), pp. 35-42.

http://www.ingentaconnect.com/content/jbp/ijcl/2001/00000006/I00000s1/art00004

Foster, Jennifer. 2005. Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English. PhD thesis, Trinity College, University of Dublin.

http://www.cs.tcd.ie/research_groups/clg/Theses/jfoster.ps

Foth, Kilian and Jochen Hagenstrom. 2002. Tagging for robust parsers. In Proceedings of ROMAND2002 2nd Workshop on Robust Methods in Analysis of natural Language Data. Pp.21-32. Frascati, Italy.

http://citeseer.ist.psu.edu/cache/papers/cs2/135/http:zSzzSznatswww.informatik.uni-hamburg.dezSzpubzSzMainzSzNatsPublicationszSzromand2002.pdf/foth02tagging.pdf

Gabrielatos, Costas. 2005. Corpora and language teaching: Just a fling, or wedding bells? TESL-EJ, 8 (4). pp. 1-37. http://www.tesl-ej.org/ej32/a1.html

Garside, Roger and Fanny Leech. 1985. A Probabilistic Parser. Proceedings of EACL'85, Second Conference of the European Chapter of the Association for Computational Linguistics, pp.166-170.

http://acl.ldc.upenn.edu/E/E85/E85-1024.pdf

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson, (eds.). 1987. The computational analysis of English: a corpus-based approach. London: Longman.

http://books.google.com/books?id=peZZAAAAMAAJ&pgis=1

Geeb, Franziskus. 2007. Chatbots in der praktischen Fachlexikographie und Terminologie (Chatbots and specialised lexicography and terminology). LDV-Forum Journal for Computational Linguistics and Language Technology, vol. 22 (1), pp. 51-70. http://www.ldv-forum.org/2007_Heft1/Franziskus_Geeb.pdf

Gojenola, K. & Oronoz, M. 2000. Corpus-based syntactic error detection using syntactic patterns. In NAACL-ANLP'2000.

http://portal.acm.org/citation.cfm?id=974461&dl=GUIDE,

Gorman, Paul and Nigel Hardy. 1993. CLAWS, Ada and software components. In Clive Souter and Eric Atwell (eds), Corpus-based Computational Linguistics:

Papers Presented at the 12th Conference of the International Computer Archive of Modern English. pp. 163-180. Amsterdam: Rodopi.

http://books.google.com/books?id=67OSqA_3hykC&pg=PA163

Gorostiza, Javi, Ramon Barber, Alaa M. Khamis, Maria Malfaz Rakel Pacheco, Rafael Rivas, Ana Corrales, Elena Delgado and Miguel A. Salichs. 2006. Multimodal Human-Robot Interaction Framework for a Personal Robot. Proceedings of RO-MAN'06 15th IEEE International Symposium on Robot and Human Interactive Communication. Hatfield

http://roboticslab.uc3m.es/publications/2006%20RO-MAN%20Javi.pdf

Harley, Andrew and Dominic Glennon. 1997. Sense Tagging in Action: Combining Different Tests with Additive Weightings. In Marc Light (ed), Tagging Text with Lexical Semantics: Why, What and How? ACL SIGLEX Special Interest Group on the Lexicon workshop, Association for Computational Linguistics.

http://acl.ldc.upenn.edu/W/W97/W97-0212.pdf

Hess, Michael, and Martin Volk. 1998. Seminar in Computerlinguistik: Robustes Parsing. Technical report, University of Zurich.

http://www.ifi.unizh.ch/groups/CL/hess/classes/seminare/robustparsing/themen.ps

Hirst, Graeme, and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering journal, Vol.11 pp.87–111. http://ftp.cs.toronto.edu/pub/gh/Hirst+Budanitsky-2005ms.pdf

Huang, C.R., K.J. Chen, & Li-Li Chang, 1997. Segmentation Standard for Chinese Natural Language Processing. International Journal of Computational Linguistics and Chinese Language Processing Vol.2.2, pp. 74-62.

Ide, Nancy, and Romary, Laurent. 2001. A common framework for syntactic annotation. Proceedings ACL'2001, pp. 298–305. Toulouse, France.

http://acl.ldc.upenn.edu/P/P01/P01-1040.pdf

Ingels, Peter. 1997. A Robust Text Processing Technique Applied to Lexical Error Recovery. PhD thesis, Linkoping University, Sweden.

http://arxiv.org/PS_cache/cmp-lg/pdf/9702/9702003v1.pdf

Ingels, Peter. 1996. Connected Text Recognition Using Layered HMMs and Token Passing. Proceedings of NeMLaP-2 Conference on New Methods in Natural Language Processing. Bilkent University, Turkey.

http://arxiv.org/PS_cache/cmp-lg/pdf/9607/9607036v1.pdf

Johnson, Christopher, Charles Fillmore, Esther Wood, Josef Ruppenhofer, Margaret Urban, Miriam Petruck and Collin Baker. 2001. The FrameNet Project: Tools for Lexicon Building. Technical report, FrameNet Project.

http://ccl.pku.edu.cn/doubtfire/semantics/FrameNet/theory/FrameNet_book.pdf

Kakkonen, Tuomo. 2007. Framework and Resources for Natural Language Parser Evaluation. PhD Thesis, University of Joensuu.

http://books.google.co.uk/books?id=X4xog4y8aIkC

Karlsson, Fred, Atro Voutilainen, Juha Heikkila, and Arto Anttila, editors. 1995. Constraint Grammar : a language-independent system for parsing unrestricted text. Mouton de Gruyter, Berlin and New York.

http://books.google.com/books?id=70IvVPlH63cC

Karp, D, Schabes, Y, Zaidel, M, and Egedi, D. 1992. A Freely Available Wide Coverage Morphological Analyzer for English. Proceedings of COLING'92 15th International Conference on Computational Linguistics, Nantes, France.

http://acl.ldc.upenn.edu/C/C92/C92-3145.pdf

Králík, Jan and Barbora Hladká. 2006. Proměna Českého akademického korpusu (The transformation of the Czech Academic Corpus). Slovo a slovesnost journal, vol.3 pp. 179-194.

http://www.ceeol.com/aspx/issuedetails.aspx?issueid=52e0d972-9ee9-47c3-bdba-a467672acfad&articleId=f0483032-23b2-4600-a9a3-8cf21cdb6d64

Kurimo, Mikko, Mathias Creutz, Matti Varjokallio, Ebru Arsoy, Murat Saraclar. 2006. Unsupervised segmentation of words into morphemes – Morpho Challenge 2005, Application to automatic speech recognition. INTERSPEECH 2006 ICSLP Ninth International Conference on Spoken Language Processing. Pittsburgh, PA, USA.

http://www.isca-speech.org/archive/interspeech_2006/i06_1512.html

Lankhorst, M, and R. Moddemeijer. 1993. Automatic word categorization: An information-theoretic approach. in K. A. Schouhamer Immink and P. G. M. Bot (eds) Forteenth Symposium on Information Theory in the Benelux, pp. 62-69.

http://www.cs.rug.nl/~rudy/papers/documents/RM9201.ps.gz

Lesk, Michael. 1988. Review of The computational analysis of English: a corpus-based approach. Computational Linguistics. 14(4), pp. 90-91.

http://acl.ldc.upenn.edu/J/J88/J88-4007.pdf

Lyon, Caroline. 1994. The representation of natural language to enable neural networks to detect syntactic structures. PhD Thesis, University of Hertfordshire.

http://homepages.feis.herts.ac.uk/~comrcml/Lyon-thesis.ps

Lyon, Caroline and Bob Dickerson. 1995. A fast partial parse of natural language sentences using a connectionist method. Proceedings of EACL'95 7th Conference of the European Chapter of the Association of Computational Linguistics, pp.215-222. http://www.aclweb.org/anthology-new/E/E95/E95-1030.pdf

Lyon, Caroline and Ray Frank. 1997. Using Single Layer Networks for Discrete, Sequential Data: an Example from Natural Language Processing. Neural Computing Applications. 5(4), pp. 196-214.

http://www.springerlink.com/content/j05216820p672928/fulltext.pdf

Mangu, Lidia and Brill, Eric. 1997. Automatic rule acquisition for spelling correction. In Proc. 14th International Conference on Machine Learning. Morgan Kaufmann. http://research.microsoft.com/users/brill/Pubs/ICML97.ps

McMahon, John, and Francis Jack Smith. 1996. Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies. Computational Linguistics journal, vol. 22(2), pp. 217-247.

http://portal.acm.org/citation.cfm?id=230389

McMahon, John, and Francis Jack Smith. 1998. A Review of Statistical Language Processing Techniques. Artificial Intelligence Review, vol. 12, pp.347–391. http://www.springerlink.com/content/x3072u620n7v5651/fulltext.pdf

Megyesi, Beáta. 1998. Rule-Based Part of Speech Tagger for Hungarian. MSc Thesis, Stockholm University. http://stp.ling.uu.se/~bea/Duppsats.pdf

Min, Kyongho; William Wilson, and Yoo-Jin Moon. 2000. Typographical and orthographical spelling error correction. , Proceedings of LREC-2000 Second International Conference on Language Resources and Evaluation, pp. 1781-1785. Athens, Greece. http://www.cse.unsw.edu.au/~billw/reprints/LREC-2000.pdf

Minnen, Guido, John Carroll and Darren Pearce. 2001. Applied morphological processing of English. Natural Language Engineering journal. Vol.7, pp.207-223 Cambridge University Press.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=83353

Mitton, Roger. 1992. A Description of A Computer-Usable Dictionary File Based on The Oxford Advanced Learner's Dictionary of Current English. Technical report accompanying the computer-readable file, Oxford Text Archive

http://archive.alphanet.ch/local/old/alphanet/mvm/GLUE/tt/rsynth2.0/text710.doc

Mitton, Roger. 1996. English spelling and the computer. London: Longman. http://books.google.com/books?id=v_JZAAAAMAAJ&pgis=1

Oba, Toshifumi; Atwell, Eric. 2003. Using the HTK speech recogniser to anlayse prosody in a corpus of German spoken learner's English. In: Archer, D, Rayson, P, Wilson, A & McEnery, T (editors) Proceedings of CL2003: International Conference on Corpus Linguistics, pp. 591-598 Lancaster University.

http://ucrel.lancs.ac.uk/publications/CL2003/papers/oba.pdf

O'Donoghue, Tim. 1991. Taking a Parsed Corpus to the Cleaners: The EPOW Corpus. ICAME Journal, volume 15, pp.55-62.

http://icame.uib.no/archives/No_15_ICAME_Journal_index.pdf

Oepen, Stephan , Dan Flickinger, Kristina Toutanova and Christopher Manning. 2004. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. Research on Language & Computation. Vol2(4), pp. 575-596.

http://www.springerlink.com/content/t851781443373812/fulltext.pdf

Osborne, Miles. 1994. Learning unification-based natural language grammars. PhD Thesis, University of York.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.606&rep=rep1&type=pdf

Oostdijk, Nelleke. 1988. A Corpus Linguistic Approach to Linguistic Variation. Literary and Linguistic Computing journal. Volume 3(1), pages 12-25. http://llc.oxfordjournals.org/cgi/reprint/3/1/12

Oostdijk, Nelleke. 1991. Corpus Linguistics and the Automatic Analysis of English. Amsterdam: Rodopi.

http://books.google.com/books?id=ZdpZAAAAMAAJ&pgis=1

Owen, Marion. 1987. Evaluating automatic grammatical tagging of text. ICAME Journal 11, pp.18–26.

http://icame.uib.no/archives/No_11_ICAME_Journal_index.pdf

Paprotte, Wolf and Frank Schumacher, 1993. MULTILEX Final Report WP9: MLEXd. Technical ReportMWP8-MS Final Version, Westfalische Wilhelms Universitat Munster. http://santana.uni-muenster.de/Publications/wp9fin.ps

Pedler, Jennifer. 2007. Computer Correction of Real-word Spelling Errors in Dyslexic Text. PhD thesis, Birkbeck College, University of London.

http://www.dcs.bbk.ac.uk/research/recentphds/pedler.pdf

Piao, Scott S. L., Paul Rayson, Dawn Archer, Tony McEnery. 2004. Evaluating Lexical Resources for A Semantic Tagger. In proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004), 26-28 May 2004, Lisbon, Portugal, Volume II, pp. 499-502.

http://www.comp.lancs.ac.uk/~paul/publications/pram_lrec04.pdf

Qiao, Hong Liang and Renjie Huang. 1998. Design and implementation of the AGTS Probabilistic Tagger. ICAME Journal 22, pp.23–48.

http://icame.uib.no/ij22/hong.pdf

Ramshaw, Lance. 2007. Correcting real-word spelling errors using a model of the problem-solving context. Computational Intelligence, Vol. 10 Issue 2, Pages 185-211.
http://www3.interscience.wiley.com/journal/119971680/abstract?CRETRY=1&SRETRY=0

Rapp, Reinhard. 1995. Die Berechnung von Assoziationen: Ein korpuslinguistischer Ansatz. PhD thesis, Informationswissenschaft, University of Konstanz. http://www.fask.uni-mainz.de/user/rapp/papers/disshtml/main/main.html

Rapp, Reinhard. 1996. Die Berechnung von Assoziationen. Georg Olms Verlag http://books.google.co.uk/books?id=x03L9g-gBS4C

Ravi, Sujith and Jihie Kim. 2007. Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers. In Proceedings of AIED'2007 Artificial Intelligence in Education Conference.

http://isi.edu/~jihie/papers/ThreadAssessmt-AIED2007.pdf

Rayson, Paul. 2003. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD thesis, Lancaster University. http://eprints.comp.lancs.ac.uk/753/

Rayson, Paul, Archer, D., Piao, S., McEnery, T. 2004. The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

http://www.comp.lancs.ac.uk/~paul/publications/usas_lrec04ws.pdf

Rayson, Paul; Dawn Archer, and Nick Smith. 2005. VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historic corpora. In: Proceedings of the Corpus Linguistics 2005 Conference, Birmingham, UK. http://eprints.comp.lancs.ac.uk/1157/

Saito, Toshio, Junsaku Nakamura, Shunji Yamazaki. 2002. English Corpus Linguistics in Japan. Amsterdam: Rodopi.

http://www.ingentaconnect.com/content/rodopi/lang/2002/00000038/00000001

Sampson, Geoffrey. 1986. A Stochastic Approach to Parsing. Proceedings of COLING'86 International Conference on Computational Linguistics. Pp.151-155. http://acl.ldc.upenn.edu/C/C86/C86-1033.pdf

Sawalha, Majdi; Atwell, Eric. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers in: Proceedings of COLING 2008 22nd International Conference on Computational Linguistics.

http://aclweb.org/anthology-new/C/C08/C08-2027.pdf

Sharkey, Noel. 1989. Models of Cognition: A Review of Cognitive Science. Intellect Books. http://books.google.com/books?id=GkZajyzL_qMC

Smith, Nicholas and Tony McEnery. 2000. Can we improve Part-of-Speech tagging by inducing probabilistic Part-of-Speech annotated lexicons from large corpora? In Nicolas Nicolov and Ruslan Mitkov (eds), Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97. John Benjamins.

http://books.google.com/books?id=Ncj4672xHiMC

Smith, Nick, Sebastian Hoffmann, Paul Rayson. 2008. Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations. Literary and Linguistic Computing journal. Volume23(2), pp.163-180.

http://llc.oxfordjournals.org/cgi/content/abstract/23/2/163

Souter, Clive. 1990. Systemic-functional grammars and corpora. In Jan Aarts, Willem Meijs (eds), Theory and Practice in Corpus Linguistics. Amsterdam: Rodopi.

http://books.google.com/books?id=ZjRO8cY0xwEC

Souter, Clive, and O'Donoghue, Tim. 1991. Probabilist parsing in the COMMUNAL project. In Stig Johansson and Anna-Brita Stenstrom (eds), English Computer Corpora: Selected Papers and Research Guide. pp. 33-50. Walter de Gruyter. http://books.google.com/books?id=woopk294GpsC

Strenstrom, A.-B., & Svartvik, J. 1994. Imparsable speech: Repeats and other nonfluencies in spoken English. In N. Oostdijk & P. de Haan (Eds.), Corpus-based research into language. Amsterdam: Rodopi.

http://books.google.com/books?id=JquvB-uJJG4C

Strzalkowski, Tomek, and Barbara Vauthey. 1992. Information retrieval using robust natural language processing. Proceedings of ACL'92 30th conference of the Association for Computational Linguistics, pp.104-111.

http://portal.acm.org/citation.cfm?id=981981

Sugiura, Masatoshi. 1990. On the Lancaster-Oslo/Bergen Corpus. Journal of Language and Culture, 1, pp115-131.

http://oscar.gsid.nagoya-u.ac.jp/paper/sugiura1990LOB.pdf

Sutcliffe, Richard; Koch, Heinz-Detlev and McElligot, Annettte. 1996. Industrial parsing of software manuals. Amsterdam: Rodopi.

http://books.google.co.uk/books?id=ir0zI8aXmmIC

Tepperman and S. Narayanan. 2005. Automatic Syllable Stress Detection for Pronunciation Evaluation of Language Learners. Proc. ICASSP'05, pp. 937–940. Philadelphia. http://sail.usc.edu/~tepperma/ICASSP2005_tepperman.pdf

Tepperman and S. Narayanan. 2005. Hidden-articulator Markov models for pronunciation evaluation. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.174-179.

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1566471&isnumber=33227

Tepperman and S. Narayanan. 2008. Using Articulatory Representations to Detect Segmental Errors in Nonnative Pronunciation. IEEE Transactions on Audio, Speech, and Language Processing. Vol. 16(1), pp.8-22.

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4384605&isnumber=4407525

Teufel, Simone. 1995. A support tool for tagset mapping. Proceedings of EACL'95 SIGDAT workshop.

http://arxiv.org/PS_cache/cmp-lg/pdf/9506/9506005v2.pdf

Thelwall, Mike. 2005. Creating and using Web corpora. International Journal of Corpus Linguistics, Volume 10, Number 4, pp. 517-541.

http://www.ingentaconnect.com/content/jbp/ijcl/2005/00000010/00000004/art00005

Thibeault, Mélanie. 2004. La categorisation grammaticale automatique: adaption du categoriseur au francais et modification de l'approche. M.A. thesis, Faculte des Lettres, Universite Laval, Quebec.

http://archimede.bibl.ulaval.ca/archimede/files/e383ca7b-e7c5-443e-adb8-bcf296ea0014/22225.html

Thompson, Henry and Chris Brew. 1996. Automatic Evaluation of Computer Generated Text: Final Report on the TextEval Project. Human Communication Research Center, University of Edinburgh.

http://reference.kfupm.edu.sa/content/a/u/automatic_evaluation_of_computer_generat_147739.pdf

Wagner, Joachim, Jennifer Foster, and Josef van Genabith. 2007. A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In Proceedings of EMNLP-CoNLL-2007. http://acl.ldc.upenn.edu/D/D07/D07-1012.pdf

Wermter, Stefan, Ellen Riloff and Gabriele Scheler. 1996. Learning approaches for natural language processing. In Stefan Wermter, Ellen Riloff and Gabriele Scheler (eds), Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. pp1-16, Lecture Notes in Artificial Intelligence 1040, Springer. http://books.google.com/books?id=QleFEFRrj6UC

Wilms, Geert Jan. 1995. Automated Induction of a Lexical Sublanguage Grammar Using a Hybrid System of Corpus- and Knowledge-based Techniques. PhD Thesis, Computer Science department, Mississippi State University.

http://computerscience.uu.edu/faculty/jwilms/papers/dissert/dissertation.pdf

Yamashita, Tatsuo and Matsumoto, Yuji. 2000. Language Independent Morphological Analysis. 6th Applied Natural Language Processing Conference. Pp.232-238. http://acl.ldc.upenn.edu/A/A00/A00-1032.pdf

Zavrel, Jakub. Lexical space: learning and using continuous linguistic representations. PhD thesis, Utrecht University. 1996.

http://ilk.uvt.nl/downloads/pub/zavrel/cki-scriptie.tree.ps.gz

Zavrel Jakub and Walter Daelemans. 2000. Bootstrapping a tagged corpus through combination of existing heterogeneous taggers. Proceedings of LREC'2000 second international conference on language resources and evaluation, pp.17-20.

http://arxiv.org/PS_cache/cs/pdf/0007/0007018v1.pdf

Zernik, Uri. 1991. Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Lawrence Erlbaum. http://books.google.co.uk/books?id=B4iIRuH1ylcC

Zribi, Chiraz, Hanene Mejri and Mohamed Ahmed. 2007. Combining Methods for Detecting and Correcting Semantic Hidden Errors in Arabic Texts. In Computational Linguistics and Intelligent Text Processing: Proceedings of CICLing 2007, Mexico City, Mexico. Lecture Notes in Computer Science 4394, Springer.

http://www.springerlink.com/content/6532050l14k3625g/

Appendix:
Papers included in this PhD submission

1. [7] Analysis of the LOB Corpus: progress and prospects. 1984.
2. [10] Beyond the micro: advanced software for research and teaching from computer science and artificial intelligence. 1986.
3. [13] A parsing expert system which learns from corpus analysis 1987.
4. [15] Pattern recognition applied to the acquisition of a grammatical classification system from unrestricted English text. 1987.
5. [16] How to detect grammatical errors in a text without parsing it. 1987.
6. [17] Dealing with ill-formed English text. 1987.
7. [18] Constituent-likelihood grammar. 1987.
8. [19] Leeds Unix Knowledge Expert: a domain-dependent Expert System generated with domain-independent tools. 1987.
9. [26] Transforming a parsed corpus into a corpus parser. 1988.
10. [28] Project APRIL - a progress report. 1988.
11. [29] An Arabic text database and glossary system for students. 1989.
12. [31] A lexical database for English learners and users: the Oxford Advanced Learner's Dictionary. 1989.
13. [37] The HEFC's Knowledge Based Systems Initiative. 1993.
14. [38] Knowledge and constraint management: large scale applications. 1993.
15. [39] Linguistic constraints for large-vocabulary speech recognition. 1993.
16. [40] Corpus-based statistical modelling of English grammar. 1993.
17. [57] AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. 1994.
18. [61] The automated evaluation of inferred word classifications. 1995.
19. [65] Comparative evaluation of grammatical annotation models. 1996.
20. [66] Machine learning from corpus resources for speech and handwriting recognition. 1996.
21. [78] The semantic/pragmatic annotation of an air traffic control corpus for use in speech recognition. 1997.
22. [81] The Language Machine. 1999.
23. [87] A comparative evaluation of modern English corpus grammatical annotation schemes. 2000.
24. [90] User-guided system development in ISLE: Interactive Spoken Language Education. 2000.
25. [96] A corpus for interstellar communication. 2001.

26. [98] A domain-independent semantic tagger for the study of meaning associations in English text. 2001.
27. [103] Semantic association between web pages - a lexical knowledge based method. 2002.
28. [111] The ISLE corpus: Italian and German Spoken Learner's English. 2003.
29. [114] Detecting student copying in a corpus of science laboratory reports. 2003.
30. [115] A new machine learning algorithm for neoposy: coining new parts of speech. 2003.
31. [123] A review of Arabic corpus analysis tools. 2004.
32. [126] Clustering of word types and unification of word tokens into grammatical word-classes. 2004.
33. [136] Using corpora in machine-learning chatbot systems. 2005.
34. [139] Combinatory Hybrid Elementary Analysis of Text. 2006.
35. [140] aConCorde: Towards an open-source, extendable concordancer for Arabic. 2006.
36. [141] The design of a corpus of contemporary Arabic. 2006.