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

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
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 Zaanne, 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

Acknowledgements............................................................................................................ ii
Abstract.......................................................................................................................... iii
Contents ......................................................................................................................... iv
Preface.......................................................................................................................... v

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

Chapter 2: Grant-Funded Research Projects in Corpus Linguistics and
Language Learning.................................................................................................... 9

Chapter 3: Working with Research Students in Corpus Linguistics and
Language Learning................................................................................................... 12

Chapter 4: Surveying Research in Corpus Linguistics and Language
Learning....................................................................................................................... 15

References (i) Publications by Eric Atwell 1980-2008 ............................................. 17

References (ii) illustrating the historic context and impact of my research.......... 33

Appendix: Papers included in this PhD submission ................................................. 50
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 selection 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.


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.

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 unparsable. 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, e.g. (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].
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/


[154] Atwell, Eric; Abbas, Noorhan; Abu Shawar, Bayan; Alsaiif, 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.


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


[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.


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


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.


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


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.


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


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


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


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.comp.leeds.ac.uk/eric/alsulaiti06ijcl.pdf

http://books.google.co.uk/books?id=woopk294GpsC

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


http://www.informatics.susx.ac.uk/research/groups/nlp/carroll/papers/lre98.pdf


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

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


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


http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V1N-4NHD91V-2&_user=65461&_rdoc=1&_fmt=&_origin=search&_sort=d&view=c&_acct=C00000458&_version=1&_userid=65461&md5=139b255c48b8ce275a3e4b7323857dfb


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

http://www.ingentaconnect.com/content/jbp/ijcl/2001/00000006/00000s1/art00004


Ingels, Peter. 1997. A Robust Text Processing Technique Applied to Lexical

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.

Johnson, Christopher, Charles Fillmore, Esther Wood, Josef Ruppenhofer,
Margaret Urban, Miriam Petrucci and Collin Baker. 2001. The FrameNet Project:

Kakkonen, Tuomo. 2007. Framework and Resources for Natural Language
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
http://books.google.com/books?id=70IvVPfH63cC

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
corpusu (The transformation of the Czech Academic Corpus). Slovo a slovesnost
http://www.ceeol.com/aspx/issuedetails.aspx?issueid=52e0d972-9ee9-47c3-bdba-
a467672acfad&articleId=f0483032-23b2-4600-a9a3-8cf21cdb6d64
http://www.isca-speech.org/archive/interspeech_2006/i06_1512.html


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

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


http://llc.oxfordjournals.org/cgi/reprint/3/1/12

http://books.google.com/books?id=ZdpZAAAAMAAJ&pgis=1
http://icame.uib.no/archives/No_11_ICAME_Journal_index.pdf


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

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

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


http://eprints.comp.lancs.ac.uk/753/


http://www.ingentaconnect.com/content/rodopi/lang/2002/00000138/00000001

http://acl.ldc.upenn.edu/C/C86/C86-1033.pdf
http://aclweb.org/anthology-new/C/C08/C08-2027.pdf


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

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

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

http://books.google.com/books?id=woopk294GpsC

http://books.google.com/books?id=JquvB-uJJG4C
http://portal.acm.org/citation.cfm?id=981981


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


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

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


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


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


Appendix:
Papers included in this PhD submission


32. [126] Clustering of word types and unification of word tokens into grammatical word-classes. 2004.


34. [139] Combinatory Hybrid Elementary Analysis of Text. 2006.
