A Richly Annotated Corpus for Probabilistic Parsing

Clive Souter and Eric Atwell

Centre for Computer Analysis of Language and Speech
Division of Artificial Intelligence
School of Computer Studies
University of Leeds
Leeds LS2 9JT
United Kingdom

Tel: +44 532 335460 and 335761
Email: cs@ai.leeds.ac.uk and eric@ai.leeds.ac.uk

Abstract

This paper describes the use of a small but syntactically rich parsed corpus of English in probabilistic parsing. Software has been developed to extract probabilistic systemic-functional grammars (SFGs) from the Polytechnic of Wales Corpus in several formalisms, which could equally well be applied to other parsed corpora. To complement the large probabilistic grammar, we discuss progress in the provision of lexical resources, which range from corpus wordlists to a large lexical database supplemented with word frequencies and SFG categories. The lexicon and grammar resources may be used in a variety of probabilistic parsing programs, one of which is presented in some detail: The Realistic Annealing Parser. Compared to traditional rule-based methods, such parsers usually implement complex algorithms, and are relatively slow, but are more robust in providing analyses to unrestricted and even semi-grammatical English.

1. Introduction

1.1 Aim

The aim of this paper is to present resources and techniques for statistical natural language parsing that have been developed at Leeds over the last 5-6 years, focusing on the exploitation of the richly annotated Polytechnic of Wales (POW) Corpus to produce large scale probabilistic grammars for the robust parsing of unrestricted English.

1.2 Background: Competence versus performance

The dominant paradigm in natural language processing over the last three decades has been the use of competence grammars. Relatively small sets of rules have been created intuitively by researchers whose primary interest was to demonstrate the possibility of integrating a lexicon and grammar into an efficient parsing program. A fairly recent example of the competence approach for English is the UK Government-sponsored Alvey Natural Language Toolkit, which contains a GPSG-like grammar which expands to an object grammar of over 1,000 phrase-structure rules (Grover et al 1989). The toolkit also includes an efficient chart parser, a morphological analyser, and a 35,000 word lexicon.

However, over the last five years or so, interest has grown in the development of robust NLP systems with much larger lexical and grammatical coverage, and with the ability to provide a best fit analysis for a sentence when, despite its extended coverage, the grammar does not describe the required structure. Some researchers have attempted to provide ad hoc rule-based solutions to handle such lexical and grammatical shortfalls (see, for example, ACL 1983, Charniak 1983, Cliff and Atwell 1987, Kwasny and Sondheimer 1981, Heidorn et al 1982, Weischedel and Black 1980). Others, including some at Leeds, have investigated the use of probabilistic parsing techniques in the quest for more reliable systems (e.g. Atwell and Elliott 1987, Atwell 1988, 1990). One of the central principles behind the adoption of probabilities (unless this is done simply to order the solutions produced by a conventional parser) is the forfeiture of a strict grammatical/ungrammatical distinction for a sliding scale of likelihood. Structures which occur very frequently are those which one might call grammatical, and those which occur infrequently or do not occur at all, are either genuinely rare, semi-grammatical or ungrammatical. The source of grammatical frequency information is a parsed corpus: A machine-readable collection of utterances which have been analysed by hand according to some grammatical description and formalism. Corpora may be collected for both spoken and written text, transcribed (in the case of spoken texts), and then (laboriously) grammatically annotated. Such a grammar may be called a performance grammar.

1.3 Background at Leeds

Research on corpus-based parsing at Leeds has revolved round more than one project and corpus. In 1986 the RSRE Speech Research Unit (now part of the UK Defence Research Agency) funded a three-year project called APRIL (Annealing Parser for Realistic Input Language) (see Haigh et al 1988, Sampson et al 1989). In this project, Haigh, Sampson and Atwell developed a stochastic parser based on the Leeds-Lancaster Treebank (Sampson 1987a), a 45,000-word subset of the Lancaster-Oslo/Bergen (LOB) Corpus (Johannson et al 1986) annotated with parse trees. The grammar was extracted from the corpus in the form of a probabilistic RTN, and used to evaluate the solution trees found by a simulated annealing search for some new sentence. At the end of the 'annealing run', provided the annealing schedule has been properly tuned, a single best-fit parse tree is found. (To simplify matters, in the early stages of the APRIL project, an ordered list of word tags was used as input, rather than a list of words). The parsing technique is quite complex compared to rule-based alternatives, and is much slower, but has the benefit of always producing a solution tree.

In 1987, the RSRE Speech Research Unit, ICL, and Longman sponsored the COMMUNAL project (COnvivial Man-Machine Understanding through NAtural Language) at University of Wales College at Cardiff (UWCC) and the University of Leeds. The COMMUNAL project aimed to develop a natural language interface to knowledge-based systems. The Cardiff team, led by Robin Fawcett, were responsible for knowledge representation, sentence generation and the development of a large systemic functional grammar. At Leeds, Atwell and Souter concentrated on developing a parser for the grammar used in the generator. A number of parsing techniques were investigated. One of these involved modifying the simulated annealing technique to work with words, rather than tags, as input, parsing from left to right, and constraining the annealing search space with judicious use of probability density functions. In the COMMUNAL project, two corpus sources were used for the grammatical frequency data: Firstly, the Polytechnic of Wales (POW) Corpus, 65,000 words of hand-analysed children's spoken English (Fawcett and Perkins 1980, Souter 1989). Secondly, the potentially infinitely large Ark Corpus, which consists of randomly generated tree structures for English, produced by the COMMUNAL NL generator, GENESYS (Wright 1988, Fawcett and Tucker 1989). The Realistic Annealing Parser (RAP) produces a solution tree more quickly than APRIL, by focusing the annealing on areas of the tree which have low probabilities (Atwell et al 1988, Souter 1990, Souter and O'Donoghue 1991).

Other ongoing research at Leeds related to corpus-based probabilistic parsing is surveyed in (Atwell 1992), including projects sponsored by the Defence Research Agency and British Telecom, and half a dozen PhD student projects.

2. Parsed Corpora of English

A corpus is a body of texts of one or more languages which have been collected in some principled way, perhaps to attempt to be generally representative of the language, or perhaps for some more restricted purpose, such as the study of a particular linguistic genre, of a geographical or historical variety, or even of child language acquisition. Corpora may be raw (contain no additional annotation to the original text), tagged (parts of speech, also called word-tags, are added to each word in the corpus) or parsed (full syntactic analyses are given for each utterance in the corpus). A range of probabilistic grammatical models may be induced or extracted from these three types of corpus. Such models may generally be termed constituent likelihood grammars (Atwell 1983, 1988). This paper will restrict itself to work on fully parsed corpora.

Because of the phenomenal effort involved in hand analysing raw, or even tagged, text, parsed corpora tend to be small and few (between only 50,000 to 150,000 words). This is not very large, compared to raw corpora, but still represents several person-years of work. Included in this bracket are the Leeds-Lancaster Treebank, the Polytechnic of Wales Corpus, the Gothenburg Corpus (Ellegard 1978), the related Susanne Corpus (Sampson 1992), and the Nijmegen Corpus (Keulen 1986). Each of these corpora contain relatively detailed grammatical analyses. Two much larger parsed corpora which have been hand analysed in less detail are the IBM-Lancaster Associated Press Corpus (1 million words; not generally available) and the ACL/DCI Penn Treebank (which is intended to consist of several million words, part of which has recently been released on CD-ROM). The Ark corpus of NL generator output contains 100,000 sentences, or approximately 0.75 million words. For a more detailed survey of parsed corpora, see (Sampson 1992). Each of these corpora contains analyses according to different grammars, and uses different notations for representing the tree structure. Any one of the corpora may be available in a variety of formats, such as verticalized (one word per line), 80 characters per line, or one tree per line (and hence, often very long lines). Figure 1 contains example trees from a few of these corpora.

Nijmegen Corpus (numerical LDB form):

```
0800131 AT 9102 THIS 2103 MOMENT, 3101 WE 5301 'VE F201 BEEN F801 JOINED A801
0800131 BY 9102 MILLIONS 7803 OF 9104 PEOPLE 3103 ACROSS 9104
0800201 EUROPE, 3501 THIS 2104 ER 1104 COVERAGE 3103 BEING F903 TAKEN A803
0800201 BY 9104 QUITE 2805 A 2505 NUMBER 3105 OF 9106 EUROPEAN 4107
0800202 COUNTRIES 3202 AND 6102 ALSO 8103 BEING F903 TAKEN A803 IN 9104 THE 2105
0800202 UNITED 9906 STATES. 3600 [[ 9400
```

Leeds-Lancaster Treebank:

```
A01  68 001
[S[Nns[NPT[ Mr ]NPT][NP[ James ]NP][NP[ Callaghan ]NP][,[ , ],][Ns[NN$[ labour' s ]NN$][JJ[ colonial ]JJ][NN[
spokesman ]NN]Ns][,[ , ],]Nns][V[VBD[ said ]VBD]V] [Fn[Nns[NPT[ Sir ]NPT][NP[ Roy ]NP]Nns][V[HVD[ had
]HVD]V][Ns[ATI[ no ]ATI][NN[ right ]NN][Ti[Vi[TO[ to ]TO][VB[ delay ]VB]Vi][Ns[NN[ progress ]NN]Ns][P[IN[
in ]IN][Np[ATI[ the ]ATI][NNS[ talks ]NNS]Np]P][P[IN[ by ]IN][Tg[Vg[VBG[ refusing ] VBG]Vg][Ti[Vi[TO[ to
]TO][VB[ sit ]VB]Vi][P[IN[ round ]IN][Ns[ATI[ the ]ATI][NN[ conference ]NN][NN[ table
]NN]Ns]P]Ti]Tg]P]Ti]Ns]Fn][.[ . ].]S]
```

IBM-Lancaster Associated Press Corpus (Spoken English Corpus Treebank):

```
SK01   3 v
[Nr Every_AT1 three_MC months_NNT2 Nr] ,_, [ here_RL [P on_II [N Radio_NN1 4_MC N]P]] ,_, [N I_PPIS1 N][V
present_VV0 [N a_AT1 programme_NN1 [Fn called_VVN [N Workforce_NP1 N]Fn]N]V] ._.
```

Polytechnic of Wales (POW) Corpus:

```
189
Z 1 CL FR RIGHT 1 CL 2 C PGP 3 P IN-THE-MIDDLE-OF 3 CV 4 NGP 5 DD THE 5 H TOWN 4 NGP 6 & OR 6 DD
THE 6 MOTH NGP H COUNCIL 6 H ESTATE 2 S NGP HP WE 2 OM 'LL 2 M PUT 2 C NGP DD THAT 2 C QQGP
AX THERE 1 CL 7 & AND 7 S NGP H WE 7 OM 'LL 7 M PUT 7 C NGP 8 DQ SOME 8 H TREES 7 C QQGP AX
THERE
```

Ark Corpus:

```
[Z [Cl [M close] [C2/Af [ngp [ds [qqgp [dds the] [a worst]]] [h boat+s]]] [e !]]]
[Z [Cl [S/Af [ngp [dq one] [vq of] [h [genclr [g mine]]]]] [O/Xf is] [G going_to] [Xpf have] [Xp been] [M cook+ed] [e
.]]]
[Z [Cl [O/Xp isn't] [S/Af [ngp [h what]]] [M unlock+ed] [e ?]]]
[Z [Cl [S it] [O/Xpd is] [M snow+ing] [e .]]]
```

Figure 1: Examples of trees from different parsed corpora.

The most obvious distinction is the contrast between the use of brackets and numbers to represent tree structure. Numerical trees allow the representation of non-context-free relations such as discontinuities within a constituent. In the bracketed notation this problem is normally avoided by coding the two parts of a discontinuity as separate constituents. A less obvious distinction, but perhaps more important, is the divergence between the use of a coarse- and fine-grained grammatical description. Contrast, for example, the Associated Press (AP) Corpus and the Polytechnic of Wales Corpus. The AP Corpus contains a skeletal parse using a fairly basic set of formal grammatical labels, while the POW Corpus contains a detailed set of formal and functional labels. Furthermore, there may not be a straightforward mapping between different grammatical formalisms, because they may assign different structures to the same unambiguous sentence.

2.1 The Polytechnic of Wales Corpus: Its origins and format

In the rest of the paper, the POW corpus will be used as an example, although each of the corpora shown in Figure 1. have been or are being used for probabilistic work at Leeds. Many of the techniques we have applied to the POW corpus could equally well be applied to other corpora, but some work is is necessarily corpus-specific, in that it relates to the particular grammar contained in POW, namely Systemic Functional Grammar (SFG).

The corpus was originally collected for a child language development project to study the use of various syntactico-semantic English constructs between the ages of six and twelve. A sample of approximately 120 children in this age range from the Pontypridd area of South Wales was selected, and divided into four cohorts of 30, each within three months of the ages 6, 8, 10, and 12. These cohorts were subdivided by sex (B,G) and socio-economic class (A,B,C,D). The latter was achieved using details of the 'highest' occupation of either of the parents of the child and the educational level of the parent or parents.

The children were selected in order to minimise any Welsh or other second language influence. The above subdivision resulted in small homogeneous cells of three children. Recordings were made of a play session with a Lego brick building task for each cell, and of an individual interview with the same adult for each child, in which the child's favourite games or TV programmes were discussed.

2.1.1 Transcription and syntactic analysis

The first 10 minutes of each play session commencing at a point where normal peer group interaction began (the microphone was ignored) were transcribed by 15 trained transcribers. Likewise for the interviews. Intonation contours were added by a phonetician, and the resulting transcripts published in four volumes. (Fawcett and Perkins 1980).

For the syntactic analysis ten trained analysts were employed to manually parse the transcribed texts, using Fawcett's systemic-functional grammar. Despite thorough checking, some inconsistencies remain in the text owing to several people working on different parts of the corpus, and no mechanism being available to ensure the well-formedness of such detailed parse trees. An edited version of the corpus (EPOW), with many of these inconsistencies corrected, has been created (O'Donoghue 1990).

The resulting machine-readable fully parsed corpus consists of approximately 65,000 words in 11,396 (sometimes very long) lines, each containing a parse tree. The corpus of parse trees fills 1.1 Mb. and contains 184 files, each with a reference header which identifies the age, sex and social class of the child, and whether the text is from a play session or an interview. The corpus is also available in wrap-round form with a maximum line length of 80 characters, where one parse tree may take up several lines. The four-volume transcripts can be supplied by the British Library Inter-Library Loans System, and the machine readable versions of both POW and EPOW are distributed by ICAME [1] and the Oxford Text Archive [2].

2.1.2 Systemic-Functional Grammar

The grammatical theory on which the manual parsing is based is Robin Fawcett's development of a Hallidayan Systemic-Functional Grammar, described in (Fawcett 1981). Functional elements of structure, such as subject (S), complement (C), modifier (MO), qualifier (Q) and adjunct (A) are filled by formal categories called units, (cf phrases in TG or GPSG) such as nominal group (ngp), prepositional group (pgp) and quantity-quality group (qqgp), or clusters such as genitive cluster (gc). The top-level symbol is Z (sigma) and is invariably filled by one or more clauses (cl). Some areas have a very elaborate description, eg: adjuncts, modifiers, determiners, auxiliaries, while others are relatively simple, eg: main-verb (M), and head (H).

2.1.3 Notation

The tree notation employs numbers rather than the more traditional bracketed form to define mother-daughter relationships, in order to capture discontinuous units. The number directly preceding a group of symbols refers to their mother. The mother is itself found immediately preceding the first occurrence of that number in the tree. In the example section of a corpus file given in Figure 1., the tree consists of a sentence (Z) containing three daughter clauses (CL), as each clause is preceded by the number one.

In the POW corpus when the correct analysis for a structure is uncertain, the one given is followed by a question mark. Likewise for cases where unclear recordings have made word identification difficult. Apart from the numerical structure, the grammatical categories and the words themselves, the only other symbols which may occur in the trees are three types of bracketing:

i) square [NV...], [UN...], [RP...] for non-verbal, repetition, etc.
ii) angle <...> for ellipsis in rapid speech.
iii) round (...) for ellipsis of items recoverable from previous text.

3. Extracting a lexicon and grammar from the POW corpus

A probabilistic context-free phrase-structure grammar can be straightforwardly extracted from the corpus, by taking the corpus one tree at a time, rewriting all the mother-daughter relationships as phrase structure rules, and deleting all the duplicates after the whole corpus has been processed. A count is kept on how many times each rule occurred. Over 4,500 unique rules have been extracted from POW, and 2820 from EPOW. The 20 most frequent rules are given in Figure 2.

8882	S --> NGP
8792	NGP --> HP
8251	Z --> CL
6698	C --> NGP
4443	QQGP --> AX
2491	PGP --> P CV
2487	CV --> NGP
2283	NGP --> DD H
2272	CL --> F
1910	Z --> CL CL
1738	NGP --> DQ H
1526	NGP --> H
1496	C --> PGP
1272	C --> QQGP
1234	CM --> QQGP
1221	NGP --> DD
1215	NGP --> HN
1182	C --> CL
1011	MO --> QQGP
1004	CL --> C

Figure 2. The 20 most frequent rules in the POW corpus

2679	I	HP
2250	THE	DD
1901	A	DQ
1550	AND	&
1525	IT	HP
1298	YOU	HP
1173	’S	OM
1117	WE	HP
1020	THAT	DD
897	YEAH	F
691	GOT	M
610	THEY	HP
585	NO	F
554	IN	P
523	TO	I
482	PUT	M
417	HE	HP
411	DON’T	ON
401	ONE	HP
400	OF	VO

Figure 3. The 20 most frequent word-wordtag pairs in the POW corpus

Similarly, each lexical item and its grammatical tag can be extracted, with a count kept for any duplicates. The extracted wordlist can be used as a prototype probabilistic lexicon for parsing. The POW wordlist contains 4,421 unique words, and EPOW 4,618. This growth in the lexicon is the result of identifying ill-formatted structures in the POW corpus in which a word appears in the place of a word tag, and hence (wrongly) contributes to the grammar instead of the lexicon. This normally occurs when the syntactic analyst has omitted the word tag for a word. The 20 most frequent words are given in Figure 3. For once, the word ‘the’ comes second in the list, as this is a spoken corpus. The list is disambiguated, in that it is the frequency of a word paired with a specific tag which is being counted, not the global frequency of a word irrespective of its tag.

Of course, context-free phrase structure rules are not the only representation which could be used for the grammatical relationships which are in evidence in the corpus. The formalism in which the grammar is extracted depends very much on its intended use in parsing. Indeed, the rules given above already represent two alternatives: a context-free grammar (ignoring the frequencies) and a probabilistic context-free grammar (including the frequencies).

We are grateful to Tim O’Donoghue [3] for providing two other examples which we have used in different probabilistic parsers: Finite state automata (or probabilistic recursive transition networks) which were used in the Realistic Annealing Parser (RAP), and a vertical strip grammar for use in a vertical strip parser (O’Donoghue 1991). The RTN fragment (Figure 4) contains 4 columns. The first is the mother in the tree, in this example A (= Adjunct). The second and third are possible ordered daughters of the mother (including the start symbol (#) and the end symbol ($)). The fourth column contains the frequency of the combination of the pair of daughters for a particular mother.

A	#	CL	250
A	#	NGP	156
A	#	PGP	970
A	#	QQGP	869
A	#	TEXT	1
A	CL	$	250
A	CL	CL	6
A	NGP	$	156
A	PGP	$	970
A	PGP	PGP	2
A	QQGP	$	869
A	QQGP	QQGP	4
A	TEXT	$	1

Figure 4. A fragment of probabilistic RTN from EPOW

The vertical strip grammar (Figure 5) contains a list of all observed paths to the root symbol (Z) from all the possible leaves (word-tags) in the trees in the corpus, along with their frequencies (contained in another file). The fragment shows some of the vertical paths from the tag B (Binder).

```
B CL A CL A CL Z
B CL A CL AL CL C CL Z
B CL A CL AL CL Z
B CL A CL C CL Z
B CL A CL Z
B CL A CL Z TEXT C CL Z
B CL AF CL Z
B CL AL CL AL CL C CL AL CL Z
B CL AL CL AL CL Z
B CL AL CL C CL AL CL Z
B CL AL CL C CL Z
B CL AL CL CREPL CL Z
B CL AL CL CV PGP C CL Z
B CL AL CL Z
B CL AL CL Z TEXT C CL Z
B CL AM CL Z
B CL AML CL Z
```

Figure 5. A fragment of a vertical strip grammar (EPOW)

3.1 The non-circularity of extracting a grammar model from a parsed corpus

A natural question at this stage is "why extract a grammar from a fully parsed corpus, because a grammar must have existed to parse the corpus by hand in the first place?" Such a venture is not circular, since, in this case, the grammar had not yet been formalised computationally for parsing. SFG is heavily focused on semantic choices (called systems) made in NL generation, rather than the surface syntactic representations needed for parsing. So the grammar used for the hand parsing of the corpus was informal, and naturally had to be supplemented with analyses to handle the phenomena common to language performance (ellipsis, repetition etc). As we have hopefully shown, extracting a model from a parsed corpus also allows a degree of flexibility in the choice of formalism. Most importantly, though, it is also possible to extract frequency data for each observation, for use in probabilistic parsing.

3.2 Frequency distribution of words and grammar 'rules'

The frequency distribution of the words in the corpus matches what is commonly known as Zipf's law. That is, very few words occur very frequently, and very many words occur infrequently, or only once in the corpus. The frequency distribution of phrase-structure rules closely matches that of words, and has led at least one researcher to conclude that both the grammar and the lexicon are open-ended (Sampson 1987b), which would provide strong support for the use of probabilistic techniques for robust NL parsing. If the size of the corpus is increased, new singleton rules are likely to be discovered, which makes the idea of a watertight grammar ridiculous.

Conflicting evidence has been provided by Taylor et al (1989), who argue that it is the nature of the grammatical formalism (simple phrase-structure rules) which prompts this conclusion. If a more complex formalism is adopted, such as that used in GPSG, including categories as sets of features, ID-LP format, metarules and unification, then Taylor et al demonstrate that a finite set of rules can be used to describe the same data which led Sampson to believe that grammar is essentially open-ended.

However interesting this dispute might be, one cannot avoid the fact that, if one's aim is to build a robust parser, a very large grammar is needed. The Alvey NL tools grammar represents one of the largest competence grammars in the GPSG formalism, but the authors quite sensibly do not claim it to be exhaustive (Grover et al 1989; 42-43). To our knowledge, no corpus exists which has been fully annotated using a GPSG-like formalism, so it is a necessity to resort to the grammars that have been used to parse corpora to obtain realistic frequency data.

The coverage of the extracted lexicons and grammars obviously varies depending on the original corpus. In the POW corpus, the size and nature of the extracted wordlist is a more obvious cause for concern. Even with much of the 'noise' removed by automatic editing using a spelling checker, and by manual inspection, there are obvious gaps. The corpus wordlist can be used for developing prototype parsers, but to support a robust probabilistic parser, a large-scale probabilistic lexicon is required.

4. Transforming a lexical database into a probabilistic lexicon for a corpus-based grammar

The original POW wordlist is woefully inadequate in terms of size (4,421 unique words), and also contains a small number of errors in the spelling and syntactic labelling of words (see Figure 6 for a fragment of the unedited POW wordlist). Although the latter can be edited out semi-automatically (O'Donoghue 1990), we can only really address the lexical coverage problem by incorporating some large machine-readable lexicon into the parsing program.

Our aims in providing an improved lexical facility for our parsing programs are as follows: The lexicon should be large-scale, that is, contain several tens of thousands of words, and these should be supplemented with corpus-based frequency counts. The lexicon should also employ systemic functional grammar for its syntax entries, in order to be compatible with the grammar extracted from the POW corpus.

```
 1 ABROAD AX
 1 ABROARD CM
 2 ACCIDENT H
 1 ACCOUNTANT H
 1 ACHING M
 5 ACROSS AX
 1 ACROSS CM
14 ACROSS P
 3 ACTING M
 4 ACTION-MAN HN
 1 ACTUALLY AL
 7 ADD M
 1 ADD? M
 1 ADDED M
 1 ADDING M
 1 ADJUST M
 3 ADN &
 1 ADN-THEN &
 1 ADRIAN HN
 1 ADVENTRUE H
```

Figure 6. A fragment of the unedited POW wordlist

The source of lexical information we decided to tap into was the CELEX database. This is a lexical database for Dutch, English and German, developed at the University of Nijmegen, Holland. The English section of the CELEX database comprises the intersection of the word stems in LDOCE (Procter 1978, ASCOT version, see Akkerman et al 1988) and OALD (Hornby 1974), expanded to include all the morphological variants; a total of 80,429 wordforms. Moreover, the wordform entries include frequency counts from the COBUILD (Birmingham) 18 million word corpus of British English (Sinclair 1987), normalised to frequencies per million words. These are sadly not disambiguated for wordforms with more than one syntactic reading. A word such as "cut", which might be labelled as an adjective, noun and verb, would only have one gross frequency figure (177) in the database. There are separate entries for "cuts", "cutting" and other morphological variants. CELEX offers a fairly traditional set of syntax categories, augmented with some secondary stem and morphological information derived mainly from the LDOCE source dictionary. The format of the lexicon we exported to Leeds is shown in Figure 7.

4.1 Manipulating the CELEX lexicon

Various transformations were performed on the lexicon to make it more suitable for use in parsing using systemic functional grammar. Using an AWK program and some UNIX tools, we reformatted the lexicon along the lines of the bracketed LDOCE dictionary which is perhaps the nearest to an established norm. A substantial number of the verb entries were removed, which would be duplicates for our purposes, reducing the lexicon to only 59,322 wordforms. Columns of the lexicon were reordered to bring the frequency information into the same column for all entries, irrespective of syntactic category.

```
abaci abacus N Y N N N N N N N N Y 0 N irr
aback aback ADV Y N N N N N 3 N
abacus abacus N Y N N N N N N N N N 0 N
abacuses abacus N Y N N N N N N N N Y 0 N +es
abandon abandon N N Y N N N N N N N N 16 Y
abandon abandon V Y N N N N N N N N N N N  16 Y
abandoned abandon V Y N N N N N Y N N Y N N +ed 36 Y
abandoned abandoned A Y N N N N 36 Y N N
```

KEY: wordform stem category ...stem info... frequency ambiguity ...morphol. info..

Figure 7. A sample of the condensed CELEX lexicon.

The most difficult change to perform was the transforming of the CELEX syntax codes into SFG. Some mappings were achieved automatically because they were one to one, eg: prepositions, main verbs and head nouns. Others resorted to stem information, to distinguish subordinating and co-ordinating conjunctions, and various pronouns, for example. Whereas elsewhere, the grammars diverge so substantially (or CELEX did not contain the relevant information) that manual intervention is required (determiners, auxiliaries, temperers, proper nouns). The development of the lexicon is not yet complete, with the 300 or so manual additions requiring frequency data. A mechanism for handling proper nouns, compounds and idioms has yet to be included, and the coverage of the lexicon has yet to be tested against a corpus. A portion of the resulting lexicon is illustrated in Figure 8.

When the development and testing work is complete, the lexicon will be compatible with the grammar extracted from the POW corpus and be able to be integrated into the parsing programs we have been experimenting with.

5. Exploiting the probabilistic lexicon and grammar in parsing

Perhaps the most obvious way to introduce probability into the parsing process is to adapt a traditional chart parser to include the probability of each edge as it is built into the chart. The search strategy can then take into account the likelihood of combinations of edges, and find the most likely parse tree first (if the sentence is ambiguous). A parser may thereby be constrained to produce only one optimal tree, or be allowed to continue its search for all possible trees, in order of decreasing likelihood. For an example of this approach

```
((abaci)(abacus)(H)(0)(N)(Y)(N)(N)(N)(N)(N)(N)(N)(N)(Y)(irr))

((aback)(aback)(AX)(3)(N)(Y)(N)(N)(N)(N)(N))

((abacus)(abacus)(H)(0)(N)(Y)(N)(N)(N)(N)(N)(N)(N)(N)(N)())

((abacuses)(abacus)(H)(0)(N)(Y)(N)(N)(N)(N)(N)(N)(N)(N)(Y)(+es))

((abandon)(abandon)(H)(16)(Y)(N)(Y)(N)(N)(N)(N)(N)(N)(N)(N)())
((abandon)(abandon)(M)(16)(Y)(Y)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)(N)())

((abandoned)(abandon)(M)(36)(Y)(Y)(N)(N)(N)(N)(N)(Y)(N)(N)(Y)(N)(N)(+ed))
((abandoned)(abandoned)(AX)(36)(Y)(Y)(N)(N)(N)(N)(N)(N))
KEY: ((wordform)(stem)(category)(frequency)(ambiguity)(...stem info...)(...morphol. info...))
```

Figure 8. A sample of the CELEX lexicon reformatted to SFG syntax.

see (Briscoe and Carroll 1991). In the projects described here, we have abandoned the use of rule-based grammars and parsers altogether, on the grounds that they will still fail occasionally due to lack of coverage, even with a very large grammar. The main alternative we have experimented with, as discussed briefly in section 1, is the use of simulated annealing as a search strategy.

5.1 The APRIL Parser

In the earlier APRIL project, the aim was to use simulated annealing to find a best-fit parse tree for any sentence. The approach adopted by the APRIL project involves parsing a sentence by the following steps:

[1] Initially, generate a "first-guess" parse-tree at random, with any structure (as long as it is a well-formed phrase marker), and any legal symbols at the nodes.

[2] Then, make a series of random localised changes to the tree, by randomly deleting nodes or inserting new (randomly-labelled) nodes. At each stage, the likelihood of the resulting tree is measured using a constituent-likelihood function; if the change would cause the tree likelihood to fall below a threshold, then it is not allowed (but alterations which increase the likelihood, or decrease it only by a small amount, are accepted)

Initially, the likelihood threshold below which tree-plausibility may not fall is very low, so almost all changes are accepted; however, as the program run proceeds, the threshold is gradually raised, so that fewer and fewer 'worsening' changes are allowed; and eventually, successive modifications should converge on an optimal parse tree, where any proposed change will worsen the likelihood. This is a very simplified statement of this approach to probabilistic parsing; for a fuller description, see (Haigh et al 1988, Sampson et al 1989, Atwell et al 1988).

5.2 The Realistic Annealing Parser

The parse trees built by APRIL were not Systemic Functional Grammar analyses, but rather followed a surface-structure scheme and set of labelling conventions designed by Geoffrey Leech and Geoffrey Sampson specifically for the syntactic analysis of LOB Corpus texts. The parse tree conventions deliberately avoided specific schools of linguistic theory such as Generalised Phrase Structure Grammar, Lexical Functional Grammar (and SFG), aiming to be a 'theory-neutral greatest common denominator' and avoid theory-specific issues such as functional and semantic labelling. The labels on nodes are atomic, indivisible symbols, whereas most other linguistic theories (including SFG) assume nodes have compound labels including both category and functional information. To adapt APRIL to produce parse trees in other formalisms, such as for the SFG used in the POW corpus and the COMMUNAL project, the primitive moves in the annealing parser would have to be modified to allow for compound node labels.

Several other aspects of the APRIL system were modified during the COMMUNAL project to produce the Realistic Annealing Parser. APRIL attempts to parse a whole sentence as a unit, whereas many researchers believe it is more psychologically plausible to parse from left to right, building up the parse tree incrementally as words are 'consumed'. When proposing a move or tree-modification, APRIL chooses a node from the current tree at random, whereas it would seem more efficient to somehow keep track of how 'good' nodes are and concentrate changes on 'bad' nodes. The above two factors can be combined by including a left-to-right bias in the 'badness' measure, biasing the node-chooser against changing left-most (older) nodes in the tree. APRIL (at least the early prototype) assumes the input to be parsed is not a sequence of words but of word-tags, the output of the CLAWS system; whereas the COMMUNAL parser would deal with 'raw word' input.

The Realistic Annealing Parser was developed by Tim O'Donoghue [3], and employed the wordlist and probabilistic RTN extracted from either the POW or the Ark corpora, but the same method could equally have been used for a number of other parsed corpora. The way the RAP deals with all the above problems is explained in detail in COMMUNAL Report No. 17 (Atwell et al 88), so here we will just list briefly the main innovations:

5.2.1 Modified primitive moves

The set of primitive moves has been modified to ensure 'well-formed' Systemic Functional trees are built, with alternating functional and category labels. The set of primitive moves was also expanded to allow for unfinished constituents only partially built as the parse proceeds from left to right through the sentence.

5.2.2 Pruning the search space by left-to-right incremental parsing

The RAP uses several annealing runs rather than just one to parse a complete sentence: annealing is used to find the best partial parse up to word N, then word N+1 is added as a new leaf node, and annealing restarted to find a new, larger partial parse tree. This should effectively prune the search space, since (Marcus 1980) and others have shown that humans parse efficiently from left to right most of the time. However, it is not the case that the parse tree up to word N must be kept unchanged when merging in word N+1; in certain cases (eg 'garden path sentences') the new word is incompatible with the existing partial parse, which must be radically altered before a new optimal partial parse is arrived at.

5.2.3 Optimising the choice of moves by Probability Density Functions (PDFs)

The rate of convergence on an optimal solution can be improved by modifying the random elements inherent in the Simulated Annealing algorithm. When proposing a move or tree-change, instead of choosing a node purely at random, all nodes in the tree are evaluated to see how 'bad' they are. This 'badness' depends upon how a node fits in with its sisters and what kind of mother it is to any daughters it has. These 'badness' values are then used as a PDF in choosing a node, so that a 'bad' node is more likely to be modified. Superimposed on this 'badness' measure is a weight corresponding to the node's position from left to right in the tree, so that relatively new right-most nodes are more likely to be altered. This left to right bias tends to inhibit backtracking except when new words added to the right of the partial parse tree are completely incompatible with the parse so far.

5.2.4 Recursive First-order Markov model

To evaluate a partial tree at any point during the parse, the partial tree is mapped onto the probabilistic recursive transition network (RTN) which has been extracted from the corpus. The probability of the path through the RTN is calculated by multiplying together all the arc probabilities en route. Each network in the RTN is used to evaluate a mother and its immediate daughters. The shape of each network is a straightforward first-order Markov model, with the further constraint that each category is represented in only one node. There is no need to 'hand-craft' the shape of each transition network, since this constraint applies to all recursive networks. This simplifies statistics extraction from the parsed corpus, and we have found empirically that it appears to be sufficient to evaluate partial trees. This empirical finding is theoretically significant, as it would seem to be at odds with established ideas founded by Chomsky (1957). However, although Chomsky maintained that a simple Markov model was insufficient to describe Natural Language syntax, he did not consider Recursive Markov models, and did not contemplate their use in conjunction with a stochastic optimisation algorithm such as Simulated Annealing.

6. Conclusions

The Polytechnic of Wales Corpus is one of a few parsed corpora already in existence which offer a rich source of grammatical information for use in probabilistic parsing. We have developed tools for the extraction of constituent likelihood grammars in a number of formalisms, which could, with relatively minor adjustment, be used on other parsed corpora. The provision of compatible large-scale probabilistic lexical resources is achieved by semi-automatically manipulating a lexical database or machine-tractable dictionary to conform with the corpus-based grammar, and supplementing corpus word-frequencies. Constituent likelihood grammars are particularly useful for the syntactic description of unrestricted natural language, including syntactically 'noisy' or ill-formed text.

The Realistic Annealing Parser is one of a line of probabilistic grammatical analysis systems built along similar general principles. Other projects are ongoing which utilise corpus-based probabilistic grammars, supplemented with suitably large probabilistic lexicons. They include the use of neural networks, vertical strip parsers, and hybrid parsers which combine the efficiency of chart parsers with the robustness of probabilistic alternatives. Whilst we would not yet claim that these have progressed beyond the work bench and onto the production line for general use, they offer promising competition to the established norms of rule-based parsers. We believe that corpus-based probabilistic parsing is a new area of Computational Linguistics which will thrive as parsed corpora become more widely available.

7. Notes

[1] ICAME (International Computer Archive of Modern English), Norwegian Computing Centre for the Humanities, P.O. Box 53, Universitetet, N-5027 Bergen, Norway.

[2] The Oxford Text Archive (OTA), Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, UK.

[3] Tim O'Donoghue was a SERC-funded PhD student in the School of Computer Studies, who contributed to the Leeds' research on the COMMUNAL project.

8. References

ACL 1983 "Computational Linguistics: Special issue on dealing with ill-formed text." Journal of Computational Linguistics volume 9 (3-4).

Akkerman, Eric, Hetty Voogt-van Zutphen and Willem Meijs. 1988 "A computerized lexicon for word-level tagging". ASCOT Report No 2, Rodopi Press, Amsterdam.

Atwell, Eric Steven, 1983 "Constituent-likelihood grammar". ICAME Journal no. 7 pp. 34-66.

Atwell, Eric Steven, 1988 "Grammatical analysis of English by statistical pattern recognition" in Josef Kittler (ed), Pattern Recognition: Proceedings of the 4th International Conference, Cambridge, pp 626-635, Berlin, Springer-Verlag.

Atwell, Eric Steven, 1990 "Measuring Grammaticality of machine-readable text" in Werner Bahner, Joachim Schildt and Dieter Viehweger (eds.), Proceedings of the XIV International Congress of Linguists, Volume 3, pp 2275-2277, Berlin.

Atwell, Eric Steven, 1992 "Overview of Grammar Acquisition Research" in Henry Thompson (ed), "Workshop on sublanguage grammar and lexicon acquisition for speech and language: proceedings", pp. 65-70, Human Communication Research Centre, Edinburgh University.

Atwell, Eric Steven and Stephen Elliot, 1987, "Dealing with ill-formed English text" in Garside et al 1987, pp. 120-138.

Atwell, Eric Steven, D. Clive Souter and Tim F. O'Donoghue 1988 "Prototype Parser 1" COMMUNAL report 17, CCALAS, Leeds University.

Briscoe, Ted and John Carroll, 1991 "Generalised Probabilistic LR Parsing of Natural Language (Corpora) with Unification-based Grammars" University of Cambridge Computer Laboratory Technical Report No. 224, June 1991.

Charniak, Eugene, 1983 "A parser with something for everyone" in Margaret King (ed.) Parsing Natural Language, Academic Press, London.

Chomsky, Noam 1957 "Syntactic Structures" Mouton, The Hague

Ellegard A. 1978. "The Syntactic Structure of English Texts". Gothenburg Studies in English 43. Gothenburg: Acta Universitatis Gothoburgenis.

Fawcett, Robin P. 1981 "Some Proposals for Systemic Syntax", Journal of the Midlands Association for Linguistic Studies (MALS) 1.2, 2.1, 2.2 (1974-76). Re-issued with light amendments, 1981, Department of Behavioural and Communication Studies, Polytechnic of Wales.

Fawcett, Robin P. and Michael R. Perkins. 1980. "Child Language Transcripts 6-12" Department of Business and Communication Studies, Polytechnic of Wales.

Fawcett, Robin P. and Gordon Tucker. 1989. "Prototype Generators 1 and 2". COMMUNAL Report No. 10, Computational Linguistics Unit, University of Wales College of Cardiff.

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

Grover, C., E.J. Briscoe, J. Carroll and B. Boguraev. 1989. "The Alvey natural language tools grammar". University of Cambridge Computer Laboratory Technical Report No. 162, April 1989.

Haigh, Robin, Geoffrey Sampson, and Eric Steven Atwell. 1988. "Project APRIL - a progress report" Proceedings of the 26th ACL Conference Buffalo, USA.

Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow. 1982. "The EPISTLE text-critiquing system" in IBM Systems Journal 21(3): 305-326.

Hornby, A.S. (ed.) 1974 "Oxford Advanced Learner's Dictionary of Contemporary English". Oxford, OUP.

Johannson, Stig, Eric Atwell, Roger Garside, and Geoffrey Leech. 1986. "The Tagged LOB Corpus". Norwegian Computing Centre for the Humanities, University of Bergen, Norway.

Keulen, Francoise. 1986. "The Dutch Computer Corpus Pilot Project". In Jan Aarts and Willem Meijs (eds.) Corpus Linguistics II, pp 127-161, Amsterdam: Rodopi Press.

Kwasny, S. and Norman Sondheimer. 1981. "Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems" in American Journal of Computational Linguistics 7(2): 99-108.

Marcus, Mitchell. 1980. "A theory of syntactic recognition for natural language". MIT Press, Cambridge, Massachusetts.

O'Donoghue, Tim F. 1990. "Taking a parsed corpus to the cleaners: the EPOW corpus". ICAME Journal no. 15. pp 55-62.

O'Donoghue, Tim F. 1991. "The Vertical Strip Parser: A lazy approach to parsing". Research Report 91.15, School of Computer Studies, University of Leeds.

Procter, Paul. 1978. "Longman Dictionary of Contemporary English" London, Longman.

Sampson, Geoffrey. 1987a. "The Grammatical Database and Parsing Scheme". In Garside et al 1987, pp 82-96.

Sampson, Geoffrey. 1987b. "Evidence against the "grammatical"/"ungrammatical" distinction". In Willem Meijs (ed.) Corpus Linguistics and Beyond. Amsterdam: Rodopi.

Sampson, Geoffrey. 1992. "Analysed corpora of English: a consumer guide". In Martha Pennington and Vance Stevens (eds.) Computers in Applied Linguistics. Multilingual Matters.

Sampson, Geoffrey R., Robin Haigh and Eric S. Atwell. 1989 "Natural language analysis by stochastic optimization: a progress report on Project APRIL", Journal of Experimental and Theoretical Artificial Intelligence 1: 271-287.

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

Souter, Clive. 1989. "A Short Handbook to the Polytechnic of Wales Corpus". ICAME, Norwegian Computing Centre for the Humanities, Bergen University, Norway.

Souter, Clive. 1990. "Systemic Functional Grammars and Corpora". In Jan Aarts and Willem Meijs (eds.) Theory and Practice in Corpus Linguistics, pp 179-211. Amsterdam: Rodopi Press.

Souter, Clive and Tim O'Donoghue. 1991. "Probabilistic Parsing in the COMMUNAL Project". In Stig Johannson and Anna-Brita Stenström (eds.) English Computer Corpora: Selected Papers and Research Guide, pp 33-48. Berlin: Mouton de Gruyter.

Taylor, Lita, Claire Grover and Ted Briscoe. 1989. "The syntactic regularity of English noun phrases". Proceedings of the 4th European ACL Conference, Manchester: 256-263.

Weischedel, Ralph and John Black. 1980. "Responding intelligently to unparsable inputs". In American Journal of Computational Linguistics 6(2) 97-109.

Wright, Joan. 1988. "The development of tools for writing and testing systemic functional grammars". COMMUNAL Report No. 3, Computational Linguistics Unit, University of Wales College of Cardiff.