

Comparing linguistic interpretation schemes for English corpora

Eric ATWELL¹, George DEMETRIOU², John HUGHES³,
Amanda SCHIFFRIN¹, Clive SOUTER¹, Sean WILCOCK¹

1: School of Computer Studies, University of Leeds, LEEDS LS2 9JT, England;

2: Department of Computer Science, University of Sheffield, SHEFFIELD S1 4DP, England;

3: BT Laboratories, Adastral Park, Martlesham Heath, England

EMAIL: amalgam-tagger@scs.leeds.ac.uk

WWW: <http://www.scs.leeds.ac.uk/amalgam/>

Abstract

Project AMALGAM explored a range of Part-of-Speech tagsets and phrase structure parsing schemes used in modern English corpus-based research. The PoS-tagging schemes and parsing schemes include some which have been used for hand annotation of corpora or manual post-editing of automatic taggers or parsers; and others which are unedited output of a parsing program. Project deliverables include:

- a detailed description of each PoS-tagging scheme, and multi-tagged corpus;
- a “Corpus-neutral” tokenization scheme;
- a family of PoS-taggers, for 8 PoS-tagsets;
- a method for “PoS-tagset conversion”,
- a sample of texts parsed according to a range of parsing schemes: a MultiTreebank;
- an Internet service allowing researchers worldwide free access to the above resources, including a simple email-based method for PoS-tagging any English text with any or all PoS-tagset(s).

We conclude that the range of tagging and parsing schemes in use is too varied to allow agreement on a standard; and that parser-evaluation based on ‘bracket-matching’ is unfair to more sophisticated parsers.

1. Introduction

The International Computer Archive of Modern and medieval English, ICAME, is an international research network focussing on English Corpus Linguistics, including the collation and linguistic annotation of English language corpora, and applications of these linguistically interpreted corpora. ICAME

publishes an annual ICAME Journal (now in its 24th volume) and holds an annual ICAME conference (ICAME’2000, the 19th ICAME conference, was held in Sydney, Australia). Many English Corpus Linguistics projects reported in ICAME Journal and elsewhere involve grammatical analysis or tagging of English texts (eg Leech et al 1983, Atwell 1983, Booth 1985, Owen 1987, Souter 1989a, Benello et al 1989, O’Donoghue 1991, Belmore 1991, Kyto and Voutilainen 1995, Aarts 1996, Qiao and Huang 1998). Each new project reviewed existing tagging schemes, and chose which to adopt and/or adapt.

The project AMALGAM (Automatic Mapping Among Lexico-Grammatical Annotation Models) has explored a range of Part-of-Speech tagsets and parsing schemes used in ICAME corpus-based research. The PoS-tagging schemes include: Brown (Greene and Rubin 1981), LOB (Atwell 1982, Johansson et al 1986), parts (man 1986), SEC (Taylor and Knowles 1988), POW (Souter 1989b), UPenn (Santorini 1990), LLC (Eeg-Olofsson 1991), ICE (Greenbaum 1993), and BNC (Garside 1996). The parsing schemes include some which have been used for hand annotation of corpora or manual post-editing of automatic parsers; and others which are unedited output of a parsing program.

2. Defining the PoS-tagging schemes

ICAME researchers have used a range of different PoS-tag annotation schemes or models. Table 1 shows how an example sentence from the IPSM Corpus (Sutcliffe et al 1996), ‘Select the text you want to protect’, is tagged according

to several alternative tagging schemes and vertically aligned.

Table 1. An example sentence tagged according to eight rival PoS-tagging schemes

	Brown	ICE	LLC	LOB	PARTS	POW	SEC	UPenn
select	VB	V(montr,imp)	VA+0	VB	adj	M	VB	VB
the	AT	ART(def)	TA	ATI	art	DD	ATI	DT
text	NN	N(com,sing)	NC	NN	noun	H	NN	NN
you	PPSS	PRON(pers)	RC	PP2	pron	HP	PP2	PRP
want	VB	V(montr,pres)	VA+0	VB	verb	M	VB	VBP
to	TO	PRTCL(to)	PD	TO	verb	I	TO	TO
protect	VB	V(montr,infin)	VA+0	VB	verb	M	VB	VB
.	.	PUNC(per)

As Corpus Linguists, we preferred to see the tagged corpus as definitive of the meanings and uses of tags in a tagset. We have compiled a detailed description of each PoS-tagging scheme, at a comparable level of detail for each Corpus annotation scheme: a list of PoS-tags with descriptions and example uses from the source Corpus.

We have also compiled a multi-tagged corpus, a set of sample texts PoS-tagged in parallel with each PoS-tagset, and proofread by experts. We selected material from three quite different genres of English (see Table2): informal speech of London teenagers, from COLT, the Corpus of London Teenager English (Andersen and Stenstrom 1996); prepared speech for radio broadcasts, from SEC, the Spoken English Corpus (Taylor and Knowles 1988); and written

text in software manuals, from IPSM, the Industrial Parsing of Software Manuals corpus (Sutcliffe et al 1996).

3. A neutral tokenization scheme

An analysis of the different lexical tokenization rules used in the source Corpora has led us to a “Corpus-neutral” tokenization scheme, and consequent adjustments to the PoS-tagsets in our study to accept modified tokenization. The performance of the tagger could be improved by incorporating bespoke tokenisers for each scheme, but we have compromised by using only one for all schemes, to simplify comparisons. This results in errors of the kind exemplified in Table 3, using examples from the POW scheme.

Table 2. Text sources for the multi-tagged corpus.

	Sentences	Words	Average Sentence Length
London teenager speech (COLT)	60	407	6.8
Radio broadcasts (SEC)	60	2016	33.6
Software manuals (IPSM)	60	1016	16.9
Total:	180	3439	19.1

Table 3. Examples where the standardised tokenizer clashes with a specific tagging scheme (POW)

	Tokeniser/ Tagger Output	Correct analysis in POW corpus
<i>Negatives</i>	are/OM n't/OXN	aren't/OMN
<i>Enclitics</i>	where's/H	where/AXWH 's/OM
<i>Possessives</i>	God's/HN	God/HN 's/G
<i>Expressions</i>	for/P example/H have/M to/I	for-example/A have-to/X

(similarly for set-up, as-well-as, so-that, next-to, Edit/Copy, Drag & Drop, Options... etc.

4. The multi-tagger: a family of PoS-taggers

We trained a publicly-available machine learning system, the Brill tagger (Brill, 1993), to re-tag according to all of the schemes we are working with. As the Brill tagger was the sole automatic annotator for the project we achieved greater consistency. The Brill system is first given a tagged corpus as a training set, from which it extracts a lexicon and two sets of non-stochastic rules: *contextual*, indicating which tag should be chosen in the context of other tags or words, and *lexical*, used to guess the tag for words which are not found in the lexicon. Table 4 shows the model size gleaned from each

training set, and accuracy of the re-trained Brill tagger on 10,000 words from the source Corpus. The most common errors (as a percentage of all errors for that scheme), are listed in Table 5.

A more realistic evaluation of tagger accuracy across a range of text types was derived in building the multi-tagged corpus, after the outputs of the multi-tagger were proof-read and post-edited by experts in each scheme. Table 6 shows the accuracy of each tagger for the multi-tagged corpus. All the tagging schemes performed significantly worse on this test material than they did on their training material, which indicates how non-generic they are.

Table 4. Model size and accuracy of the re-trained Brill multi-tagger

Tagger	Lexicon	Context Rules	Lexical Rules	Accuracy %
Brown	53113	215	141	97.43
ICE	8305	339	128	90.59
LLC	4772	253	139	93.99
LOB	50382	220	94	95.55
Unix Parts	2842	36	93	95.8
POW	3828	170	109	93.44
SEC	8226	206	141	96.16
Upenn	93701	284	148	97.2

Table 5. The most common PoS-tagging errors.

<i>Brown</i>	VBN/VBD 14.6%	JJ/NN 4.9%	NN/VB 4.2%			
<i>ICE</i>	V(cop,pres,encl)/V(intr,pres,encl) 4.1%	ADJ/N(prop,sing) 3.1%	PUNC(oquo)/PUNC(cquo) 2.6%			
<i>LLC</i>	PA/AC 4.1%	PA/AP 2.7%	RD/CD 2.7%			
<i>LOB</i>	IN/CS 5.8%	TO/IN 4.1%	VBN/VBD 4%			
<i>POW</i>	AX/P 4.3%	OX/OM 2.9%	P/AX 2.5%			
<i>SEC</i>	TO/IN 6.3%		JJ/RB 5.6%		JJ/VB 4.8%	

Table 6. Accuracy found after manual proof-reading of multi-tagged corpus

TAGSET	TOTAL	IPSM60	COLT60	SEC60
Brown	94.3	94.3	87.7	95.6
Upenn	93.1	91.6	88.7	94.6
ICE	89.6	87.0	85.3	91.8
Parts (Unix)	86.7	89.9	82.3	86.0
LLC	86.6	86.9	84.3	87.0
POW	86.4	87.6	87.7	85.4

5. Mapping between tagging schemes

To re-tag the old parts of speech of a corpus with a new scheme of another, we apply our tagger to just the words of the corpus. This might appear to be ‘cheating’; but earlier experiments with devising a set of mapping rules from one tagset to another (Hughes and Atwell 1994, Atwell et al 1994, Hughes et al 1995) concluded that one-to-many and many-to-many mappings predominated over simple one-to-one (and many-to-one) mappings, resulting in more errors than the apparently naïve approach of ignoring the source tags.

6. Comparing tagging schemes

The descriptions of each tagset and multitagged corpus on our website enable corpus-based comparisons between the tagsets. However, quantitative measures are not straightforward. As a simple metric, consider the number of tags in the tagset: this is generally not as simple as it first seems. Most tagsets use tags which are actually a combination of features; this is clearest in ICE (eg **N(com,sing)** for singular common noun), but is also implicit in other tagsets (eg LOB **NN** is also singular common noun, in contrast with **NNS** plural common noun, and **NP** singular proper noun). Our website lists all the tags occurring in the multitagged corpus, but this does not include rare but possible feature-combinations which happen not to occur in the corpus (eg ICE has a tag for plural numbers (as in *three fifths*) which is not used in our corpus). Also, Brown and UPenn tagsets have some tags which are two ‘basic’ tags combined. In Brown, these tags are for enclitic or fused wordforms (eg *I’d* **PPSS+HVD**, *whaddya* **WDT+DO+PPS**); in UPenn, these tags are for words whose analysis is ambiguous or indeterminate (eg *entertaining* **JJ|VBG** = adjective|verb-ing-form).

A general observation is that tagsets developed later in time were designed to be ‘improvements’ on earlier tagsets; for example, LOB and UPenn tagsets designers took Brown as a starting-point. So an informal ranking based on age (as given by definitive references) is: Brown (Greene and Rubin

1981), parts (man 1986), LOB (Atwell 1982, Johansson et al 1986), SEC (Taylor and Knowles 1988), POW (Souter 1989b), UPenn (Santorini 1990), LLC (Eeg-Olofsson 1991), ICE (Greenbaum 1993). The ICE tagset is the only one to incorporate explicit features or subcategories, making it more readily digestible by non-expert users: informal feedback from users of our multi-tagger suggests that linguists (and others) find it easier to use tags like **N(com,sing)** than **NN**, since the division into major category and features in brackets is more intuitive. Another class of users of tagged texts are Machine Learning researchers, who want tagged text to train a learning algorithm, but want a small tagset to reduce the problem space; another advantage of the ICE tagset is that it is easy to reduce the tagset to major categories only by ignoring the bracketed features.

7. A MultiTreebank

The differences between English corpus annotation schemes are much greater between parsing schemes for full syntactic structure annotation than they are at word class level. The following are parses of the sentence ‘*Select the text you want to protect.*’ according to the parsing schemes of several English parsed corpora or treebanks:

==> *ENGCG-BankOfEnglish* <==

```
"select" <*> V IMP VFIN @+FMAINV
"the" DET CENTRAL ART SG/PL @DN>
"text" N NOM SG @OBJ
"you" PRON PERS NOM SG2/PL2 @SUBJ
"want" V PRES -SG3 VFIN @+FMAINV
"to" INFMARK> @INFMARK>
"protect" V INF @-FMAINV
"<$.>"
```

==> *L-ATR* <==

```
[start [sprpd1 [sprime1 [sc3 [v2
Select_VVIOPTATIVE [nbarq14 [nbar4 [d1
the_AT d1] [n1a text_NN1MESSAGE n1a] nbar4]
[sd1 [nbar6 you_PPY nbar6] [vbar1 [v33
want_VVOMENTAL-ACT [vibar1 to_TO [v1
protect_VVIHELP v1] vibar1] v33] vbar1] sd1]
nbarq14] v2] sc3] sprime1] _ . sprpd1] start]
```

==> **LOB-TREEBANK** <==

```
[S[V[VB[ select ]VB]V][Ns[ATI[ the ]ATI][NN[
text ]NN][Fr[N[PP2[ you ]PP2]N][V[VB[ want
]VB]V][Ti[Vi[TO[ to ]TO][VB[ protect
]VB]Vi]Ti]Fr]Ns][. . ]S]
```

==> **POW** <==

```
Z CL 1 M SELECT 1 C NGP 2 DD THE 2 H
TEXT 2 Q CL 3 S NGP HP YOU 3 M WANT 3 C
CL 4 I TO 4 M PROTECT 1 ? .
```

==> **SEC** <==

```
[V Select_VV0 [N the_AT text_NN1 [Fr[N
you_PPY N][V want_VV0 [Ti to_TO
protect_VV0 Ti]V]Fr]N]V] ....
```

==> **SUSANNE** <==

```
VV0t Select select [O[S*[V.V]
AT the the [Ns:o101.
NN1n text text .
PPY you you[Fr[Ny:s103.Ny:s103]
VV0v want want [V.V]
YG - - [Ti:o[s103.s103]
TO to to [Vi.
VV0t protect protect .Vi]
YG — -
[o101.o101]Ti:o]Fr]Ns:o101]S*]
YF +. - O]
```

There are two main approaches to format : one word per line, with parsing annotations (ENGCG, SUSANNE), aimed at human proofreaders, to make it easier to scan parses and correct errors; and tree-structure captured via lisp-like bracketting (L-ATR, LOB-TREEBANK, SEC, POW), assuming the textfile is processed by a tree-viewing program for human end-user consumption. The POW format uses a numerical code capable of capturing crossing branches, but in principle encodes the phrase structure.

There is even greater diversity in the parsing schemes (and formats) used in alternative NLP parsing *programs*. The example sentence was actually selected from a test-set used at the Industrial Parsing of Software Manuals workshop (Sutcliffe et al 1996); it is one of the shortest test sentences, which one might presume to be one of the most grammatically straightforward and uncontroversial. The following are outputs of several rival NLP

parsing programs, given the example sentence to parse:

==> **alice** <==

```
Fragment No. 1
>From 0 To 5
(SENT (SENT-MOD (UNK-CAT "Select") (NP
(DET "the") (NOUN "text"))))
(SENT (VP-ACT (NP "you") (V-TR "want"))) (NP
NULL-PHON)))
Fragment No. 2
>From 5 To 7
(SENT-MOD (UNK-CAT "to") (NP "protect"))
```

==> **despar** <==

```
VB select 1 --> 8 -
DT the 2 --> 3 [
NN text 3 --> 1 + OBJ
PP you 4 --> 5 " SUB
VBP want 5 --> 3 ]
TO to 6 --> 7 -
VB protect 7 --> 5 -
.. 8 --> 0 -
```

==> **principar_constituency** <==

```
(S
(VP (Vbar (V (V_NP
(V_NP Select)
(NP
(Det the)
(Nbar
(N text)
(CP
Op[1]
(Cbar (IP
(NP (Nbar (N you)))
(Ibar (VP (Vbar (V (V_CP
(V_CP want)
(CP (Cbar (IP
PRO
(Ibar
(Aux to)
(VP (Vbar (V (V_NP
(V_NP protect)
t[1]))))))))))))))))))))
.)
```

==> **principar_dependency** <==

```
(
(Select ~ V_NP *)
(the ~ Det < text spec)
(text ~ N > Select comp1)
```

```
(you ~ N < want subj)
(want ~ V_CP > text rel)
(to ~ I > want comp1)
(protect ~ V_NP > to pred)
(.)
)
```

==> *ranlp* <==

```
(VP/NP select
(N2+/DET1a the
(N2-
(N1/INFMOD
(N1/RELMOD1 (N1/N text)
(S/THATLESSREL (S1a (N2+/PRO you) (VP/NP
want (TRACE1 E))))))
(VP/TO to (VP/NP protect (TRACE1 E))))))
```

==> *sextant* <==

```
134 -----
Select the text you want to protect .
134 VP 101 Select select INF 0 0
134 NP 2 the the DET 1 1 2 (text) DET
134 NP* 2 text text NOUN 2 1 0 (select) DOBJ
134 NP* 3 you you PRON 3 0
134 VP 102 want want INF 4 0
134 VP 102 to to TO 5 0
134 VP 102 protect protect INF 6 1 3 (you) SUBJ
134 -- 0 . . . 7 0
```

This sentence is part of our multi-parsed corpus or MultiTreebank (Atwell 1996). The parsing schemes exemplified in our MultiTreebank include some which have been used for hand annotation of corpora or manual post-editing of automatic parsers: EPOW (O'Donoghue 1991), ICE (Greenbaum 1992), POW (Souter 1989a,b), SEC (Taylor and Knowles 1988), and UPenn (Marcus et al 1993). Linguist experts in each of these corpus annotation schemes kindly provided us with their parsings of the 60 IPSM sentences. Others are unedited output of parsing

programs: Alice (Black and Neal 1996), Carroll/Briscoe Shallow Parser (Briscoe and Carroll 1993), DESPAR (Ting and Shiuan 1996), ENGCG (Karlsson et al 1995, Voutilainen and Jarvinen 1996), Grammatik (WordPerfect 1998), Link (Sleator and Temperley 1991, Sutcliffe and McElligott 1996), PRINCIPAR (Lin 1994, 1996), RANLP (Osborne 1996), SEXTANT (Grefenstette 1996), and TOSCA (Aarts et al 1996, Oostdijk 1996). Language Engineering researchers working with these systems kindly provided us with their parsings of the 60 IPSM sentences.

The MultiTreebank illustrates the diversity of parsing schemes available for modern English language corpus annotation. The (EAGLES 1996) guidelines recognise layers of syntactic annotation, which form a hierarchy of importance. None of the parsing schemes included here contains all the layers (*a-h*, in Table 7 below). Different parsers annotate with different subsets of the hierarchy.

7. Website and e-mail tagging service

The multi-tagged corpus, multiTreebank, tagging scheme definitions and other documentation are available on our website. Email your English text to *amalgam-tagger@scs.leeds.ac.uk*, and it will be automatically processed by the multi-tagger, and then the output is mailed back to you. Users can select any or all of the eight schemes (Brown, ICE, LLC,LOB, Parts, POW, SEC, UPenn). The tagged text is returned one email reply message per scheme. A verbose mode can also be selected, which gives the long name for each tag as well as its short form in the output file.

Table 7. Evaluation of MultiTreebank parse schemes in terms of EAGLES layers of syntactic annotation :

- (a) Bracketing of segments
- (b) Labelling of segments
- (c) Showing dependency relations
- (d) Indicating functional labels
- (e) Marking sub-classification of syntactic segments
- (f) Deep or ‘logical’ information
- (g) Information about the rank of a syntactic unit
- (h) Special syntactic characteristics of spoken language

Parse Scheme	EAGLES layer								Score
	a	b	c	d	e	f	g	h	
ALICE	yes	yes	no	no	no	no	no	no	2
CARROLL	yes	yes	no	no	no	no	no	no	2
DESPAR	no	no	yes	no	no	no	no	no	1
ENGCG	no	no	yes	yes	yes	no	no	no	3
EPOW	yes	yes	no	yes	no	no	no	yes	4
GRAMMATIK	yes	yes	no	yes	no	no	no	no	3
ICE	yes	yes	no	yes	yes	no	no	yes	5
LINK	no	no	yes	yes	no	no	no	no	2
POW	yes	yes	no	yes	no	yes	no	yes	5
PRINCIPAR	yes	yes	yes	no	no	yes	yes	no	5
RANLT	yes	yes	no	no	no	yes	yes	no	4
SEC	yes	yes	no	no	yes	no	no	yes	4
SEXTANT	yes	yes	yes	yes	no	no	no	no	4
TOSCA	yes	yes	no	yes	yes	yes	no	yes	6
UPENN	yes	yes	no	no	no	No	no	no	2

The service has been running since December 1996, and usage is logged on our website; up to December 1999, it processed 19,839 email messages containing over 628 megabytes of text. The most popular schemes are LOB, UPenn, Brown, ICE, and SEC (in that order), with relatively little demand for parts, LLC, and POW; this reflects the popularity of the source corpora in the Corpus Linguistics community. Apart from obvious uses in linguistic analysis, English language teaching and learning, and teaching Natural Language Processing and Artificial Intelligence university students, some unforeseen applications have been found, e.g. in using the tags to aid data compression of English text (Teahan 1998); and as a guide in the search for extra-terrestrial intelligence (Elliott and Atwell 2000, Elliott et al 2000).

8. Conclusions

NLP researchers have not agreed a standard lexico-grammatical annotation model for English, so the AMALGAM project has investigated a range of alternative schemes. We have trained a ‘machine learning’ tagger with several lexico-grammatical annotation models, to enable it to annotate according to several rival modern English language corpus Part-of-Speech tagging schemes. Our main achievements are:

Software: *PoS-taggers* trained to annotate text according to several rival lexico-grammatical annotation models, accessible over the Internet via email.

Data-sets: a *multi-tagged corpus* and *multi-treebank*, a corpus of English text where each sentence is annotated according to several rival

lexico-grammatical annotation models. We have also collected together definitions of eight major English corpus word-tagging schemes. All are available over the Internet via WWW.

We conclude that there is still work to be done on agreeing a truly generic PoS-tagging scheme; and that it is not possible, to map between all parsing schemes. Unlike the tagging schemes, it does not make sense to make an application-independent comparative evaluation. No single standard can be applied to all parsing projects. Even the presumed lowest common denominator, bracketing, is rejected by some corpus linguists and dependency grammarians. The guiding factor in what is included in a parsing scheme appears to be the author's theoretical persuasion or the application they have in mind.

Acknowledgements

We are grateful to the UK Engineering and Physical Sciences Research Council (EPSRC) for funding this research project. We are also indebted to the numerous researchers worldwide who helped us by providing advice, data, and documentation, and by proofreading the multi-tagged Corpus and MultiTreebank.

This COLING Workshop paper is an abridged version of a full paper published in *ICAME Journal*, (Atwell et al 2000); we are grateful for the Journal's permission to present our findings to this complementary Workshop audience. To get the full *ICAME Journal* paper, see <http://www.hd.uib.no/icame/journal.html>

References

Aarts, Jan. 1996. A tribute to W. Nelson Francis and Henry Kucera: grammatical annotation. *ICAME Journal* 20:104-107.

Aarts, Jan, Hans van Halteren and Nelleke Oostdijk. 1996. The TOSCA analysis system. In C. Koster and E Oltmans (eds). *Proceedings of the first AGFL workshop*. 181-191. Technical Report CSI-R9604, Computing Science Institute, University of Nijmegen.

Andersen, Gisle, and Anna-Brita Stenstrom. 1996. COLT: a progress report. *ICAME Journal* 20:133-136.

Atwell, Eric. 1982. *LOB Corpus tagging project: post-edit handbook*. Department of Linguistics and

Modern English Language, University of Lancaster.

Atwell, Eric. 1983. Constituent Likelihood grammar. *ICAME Journal* 7:34-66.

Atwell, Eric. 1996. Comparative evaluation of grammatical annotation models. In (Sutcliffe et al 1997), 25-46.

Atwell, Eric, John Hughes and Clive Souter. 1994. AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. In Judith Klavans and Philip Resnik (eds.), *The balancing act - combining symbolic and statistical approaches to language. Proceedings of the workshop in conjunction with the 32nd annual meeting of the Association for Computational Linguistics*. New Mexico State University, Las Cruces, New Mexico, USA.

Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter, and Sean Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24, to appear.

Belmore, Nancy. 1991. Tagging Brown with the LOB tagging suite. *ICAME Journal* 15:63-86.

Benello, J., A. Mackie and J. Anderson. 1989. Syntactic category disambiguation with neural networks. *Computer Speech and Language* 3:203-217.

Black, William, and Philip Neal. 1996. Using ALICE to analyse a software manual corpus. In (Sutcliffe et al 1996), 47-56.

Booth, Barbara. 1985. Revising CLAWS. *ICAME Journal* 9:29-35.

Brill, Eric. 1993. *A Corpus-based approach to language learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Briscoe, Edward and John Carroll. 1993. Generalised probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics* 19:25-60.

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

Eeg-Olofsson, Mats. 1991. *Word-class tagging: Some computational tools*. PhD thesis. Department of Linguistics and Phonetics, University of Lund, Sweden.

- Elliott, John, and Eric Atwell. 2000. Is there anybody out there?: the detection of intelligent and generic language-like features. In *Journal of the British Interplanetary Society*, 53:1/2, 13-22.
- Elliott, John, Eric Atwell, and Bill Whyte. 2000. Language identification in unknown signals. Proc COLING'2000.
- Garside, Roger. 1996. The robust tagging of unrestricted text: the BNC experience. In Jenny Thomas and Mick Short (eds) *Using corpora for language research: studies in the honour of Geoffrey Leech*, 167-180. London: Longman.
- Greene, Barbara and Gerald Rubin. 1981. *Automatic grammatical tagging of English*. Providence, R.I.: Department of Linguistics, Brown University.
- Greenbaum, Sidney. 1993. The tagset for the International Corpus of English. In Clive Souter and Eric Atwell (eds) *Corpus-based Computational Linguistics*. 11-24. Amsterdam: Rodopi.
- Grefenstette, Gregory. 1996. Using the SEXTANT low-level parser to analyse a software manual corpus. In (Sutcliffe et al 1996), 139-158.
- Hughes, John and Eric Atwell. 1994. The automated evaluation of inferred word classifications. In Anthony Cohn (ed.), *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. 535-539. Chichester, John Wiley.
- Hughes, John, Clive Souter and Eric Atwell. 1995. Automatic extraction of tagset mappings from parallel-annotated corpora. In *From texts to tags: issues in multilingual language analysis. Proceedings of SIGDAT workshop in conjunction with the 7th Conference of the European Chapter of the Association for Computational Linguistics*. University College Dublin, Ireland.
- Johansson, Stig, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB corpus: users' manual*. Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities. Available from <http://www.hit.uib.no/icame/lobman/lob-cont.html>
- Karlsson, Fred, Aro Voutilainen, Juha Heikkila, and Arto Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyter.
- Kyto, Merja and Aro Voutilainen. 1995. Applying the Constraint Grammar parser of English to the Helsinki corpus. *ICAME Journal* 19:23-48.
- Leech, Geoffrey, Roger Garside and Eric Atwell. 1983. The automatic grammatical tagging of the LOB corpus. *ICAME Journal* 7:13-33.
- Lin, Dekang. 1994. PRNCIPAR – an efficient, broad-coverage, principle-based parser. *Proceedings of COLING-94, Kyoto*. 482-488.
- Lin, Dekang. 1996. Using PRINCIPAR to analyse a software manual corpus. In (Sutcliffe et al 1996), 103-118.
- man 1986. *parts*. The on-line Unix manual.
- Marcus, Mitch, M Marcinkiewicz, and Barbara Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19:313-330.
- O'Donoghue, Tim. 1991. Taking a parsed corpus to the cleaners: the EPOW corpus. *ICAME Journal* 15:55-62
- Oostdijk, Nelleke. 1996. Using the TOSCA analysis system to analyse a software manual corpus. In (Sutcliffe et al 1996), 179-206.
- Osborne, Miles. 1996. Using the Robust Alvey Natural Language Toolkit to analyse a software manual corpus. In (Sutcliffe et al 1996), 119-138.
- Owen, M. 1987. Evaluating automatic grammatical tagging of text. *ICAME Journal* 11:18-26.
- Qiao, Hong Liang and Renje Huang. 1998. Design and implementation of AGTS probabilistic tagger. *ICAME Journal* 22: 23-48.
- Santorini, Barbara. 1990. *Part-of-speech tagging guidelines for the Penn Treebank project*. Technical report MS-CIS-90-47. University of Pennsylvania: Department of Computer and Information Science.
- Sleator, D. and Temperley, D. 1991. *Parsing English with a Link grammar*. Technical Report CMU-CS-91-196. School of Computer Science, Carnegie Mellon University.
- Souter, Clive. 1989a. The COMMUNAL project: extracting a grammar from the Polytechnic of Wales corpus. *ICAME Journal* 13:20-27.
- Souter, Clive. 1989b *A short handbook to the Polytechnic of Wales Corpus*. Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities. Available from <http://kht.hit.uib.no/icame/manuals/pow.html>
- Sutcliffe, Richard, Heinz-Detlev Koch and Annette McElligott (eds.). 1996. *Industrial parsing of software manuals*. Amsterdam: Rodopi.
- Sutcliffe, Richard, and Annette McElligott. 1996. Using the Link parser of Sleator and Temperley to analyse a software manual corpus. In (Sutcliffe et al 1996), 89-102.
- Taylor, Lolita and Gerry Knowles. 1988. *Manual of information to accompany the SEC corpus: The machine readable corpus of spoken English*.

- University of Lancaster: Unit for Computer Research on the English Language. Available from <http://kht.hit.uib.no/icame/manuals/sec/INDEX.HTM>
- Teahan, Bill. 1998. *Modelling English text*. PhD Thesis, Department of Computer Science, University of Waikato, New Zealand.
- Ting, Christopher, and Peh Li Shiuan. 1996. Using a dependency structure parser without any grammar formalism to analyse a software manual corpus. In (Sutcliffe et al 1996), 159-178.
- Voutilainen, Aro, and Timo Jarvinen. 1996. Using the English Constraint Grammar Parser to analyse a software manual corpus. In (Sutcliffe et al 1996), 57-88.