A Cross-Language Methodology for Corpus Part-of-Speech Tag-Set Development[1]

Eric Atwell[2]

This paper examines criteria used in development of Corpus Part-of-Speech tag sets used when PoS-tagging a corpus, that is, enriching a corpus by adding a part-of-speech category label to each word. This requires a tag-set, a list of grammatical category labels; a tagging scheme, practical definitions of each tag or label, showing words and contexts where each tag applies; and a tagger, a program for assigning a tag to each word in the corpus, implementing the tag-set and tagging-scheme in a tag-assignment algorithm.

We start by reviewing tag-sets developed for English corpora, since English was the first language studied by corpus linguists. Traditional English grammars generally provide 8 basic parts of speech, derived from Latin grammar. However, most tag-set developers wanted to capture finer grammatical distinctions, leading to larger tag-sets. Figure 1 illustrates a range of rival English PoS-tag-sets applied to a short example sentence; even with this simple sentence, it is easy to see some significant similarities and differences between these rival tag-sets for English.

The pioneering Corpus Linguists who collected the first large-scale English language corpora all thought that their corpora could be more useful research resources if the source text samples were enriched with linguistic analyses. These pioneering English corpus linguistics projects included projects to collect the Brown corpus, the Lancaster-Oslo-Bergen corpus (LOB), the Spoken English Corpus (SEC), the Polytechnic of Wales corpus (PoW), the University of Pennsylvania Corpus (UPenn), the London-Lund Corpus (LLC), the International Corpus of English (ICE), the British National Corpus (BNC), the Spoken Corpus Recordings In British English (SCRIBE), etc.; for references see below. In nearly every case (except PoW), the first level of linguistic enrichment was to add a Part-of-Speech tag to every word in the text, labeling its grammatical category.

The different PoS-tagsets used in these English general-purpose corpora are illustrated in Figure 1, derived from the AMALGAM multi-tagged corpus (Atwell et al. 2000). This corpus is PoS-tagged according to a range of rival English corpus tagging schemes, and also parsed according to a range of rival parsing schemes, so each sentence has not just one parse-tree, but "a forest" (Cure 1980). The AMALGAM multi-tagged corpus contains text from three quite different genres of English: 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

[1] This paper is an abridged summary version of an article on "Development of tag-sets for part-of-speech tagging" to appear in Anke Lüdeling and Merja Kytö (editors) Corpus Linguistics: An International Handbook, Mouton de Gruyter.

[2] University of Leeds
e-mail: eric@comp.leeds.ac.uk

	Collins English Dictionary	SCRIBE parts	Brown	LOB	UPenn	BNC-C5	BNC-C6	ICE	PoW	LLC
If	s.conjunction	subcj	CS	CS	IN	CJS	CS	CONJUNC(subord)	B	CC
your	determiner	pos	PP$	PP$	PRP$	DPS	APPGE	PRON(poss)	DD	TB
library	noun	noun	NN	NN	NN	NN1	NN1	N(com,sing)	H	NC
is	verb	be	BEZ	BEZ	VBZ	VBZ	VBZ	V(cop,pres)	OM	VB+3
on	preposition	prep	IN	IN	IN	PRP	II	PREP(ge)	P	PA
a	determiner	art	AT	AT	DT	AT0	AT1	ART(indef)	DQ	TF
network	noun	noun	NN	NN	NN	NN1	NN1	N(com,sing)	H	NC
and	c.conjunction	conj	CC	CC	CC	CJC	CC	CONJUNC(coord)	&	CA
has	verb	verb	HVZ	HVZ	VBZ	VHZ	VHZ	V(montr,pres)	M	VH+3
the	determiner	art	AT	ATI	DT	AT0	AT	ART(def)	DD	TA
Dynix	noun	noun	NP	NP	NNP	NP0	NP1	N(com,sing)	HN	NP
Gateways	noun	noun	NPS	NNS	NNPS	NN2	NN2	N(com,sing)	HN	NP
product	noun	noun	NN	NN	NN	NN1	NN1	N(com,sing)	H	NC
,	(unspecified)	,	,	,	,	PUN	YCOM	PUNC(com)	,	,
patrons	noun	noun	NNS	NNS	NNS	NN2	NN2	N(com,plu)	H	NC+2
and	c.conjunction	conj	CC	CC	CC	CJC	CC	CONJUNC(coord)	&	CA
staff	noun	noun	NN	NN	NNS	NN0	NN	N(com,plu)	H	NC
at	preposition	prep	IN	IN	IN	PRP	II	PREP(ge)	P	PA
your	determiner	pos	PP$	PP$	PRP$	DPS	APPGE	PRON(poss)	DD	TB
library	noun	noun	NN	NN	NN	NN1	NN1	N(com,sing)	H	NC
can	verb	aux	MD	MD	MD	VM0	VM	AUX(modal,pres)	OM	VM+8
use	verb	verb	VB	VB	VB	VVI	VVI	V(montr,infin)	M	VA+0
gateways	noun	noun	NNS	NNS	NNS	NN2	NN2	N(com,plu)	H	NC+2
to	preposition	verb	TO	TO	TO	TO0	TO	PRTCL(to)	I	PD
access	verb	verb	VB	VB	VB	VVI	VVI	V(montr,infin)	M	VA+0
information	noun	noun	NN	NN	NN	NN1	NN1	N(com,sing)	H	NC
on	preposition	prep	IN	IN	IN	PRP	II	PREP(ge)	P	PA
other	determiner	adj	AP	AP	JJ	AJ0	JJ	NUM(ord)	MOC	JS
systems	noun	noun	NNS	NNS	NNS	NN2	NN2	N(com,plu)	H	NC+2
as	(unspecified)	prep	QL	RB	RB	AV021	RR21	ADV(add)	AL	AC
well	(unspecified)	adv	RB	RB”	RB	AV022	RR22	ADV(add)		AC
.	(unspecified)	.	.	.	.	PUN	YSTP	PUNC(per)	.	.

Figure 1: Example sentence illustrating rival English PoS-taggings (from the AMALGAM multi-tagged corpus)

(Sutcliffe et al. 1996). The example sentence in Figure 1 is from the software manuals section. The PoS-tagging schemes illustrated in Figure 1 include: Brown corpus (Greene and Rubin 1981), LOB: Lancaster-Oslo/Bergen corpus (Atwell 1982, Johansson et al. 1986), SEC: Spoken English Corpus (Taylor and Knowles 1988), PoW: Polytechnic of Wales corpus (Souter 1989b), UPenn: University of Pennsylvania corpus (Santorini 1990), LLC: London-Lund Corpus (Eeg-Olofsson 1991), ICE: International Corpus of English (Greenbaum 1993), and BNC: British National Corpus (Garside 1996). For comparison, also included are the simpler "traditional" part-of-speech categories used in the Collins English Dictionary, and the basic PARTS tag-set used to tag the SCRIBE corpus (Atwell 1989).

As already mentioned, in deciding on the range and number of PoS-tags, it makes sense to take into account the potential uses of the PoS-tagged corpus. Many English Corpus Linguistics projects reported in ICAME Journal and elsewhere have involved grammatical analysis or tagging of English texts (eg Leech et al. 1983, Atwell 1983, Booth 1985, Owen 1987, Souter 1989a, O'Donoghue 1991, Belmore 1991, Kytö and Voutilainen 1995, Aarts 1996, Qiao and Huang 1998). Apart from obvious uses in linguistic analysis, some unforeseen applications have been found. As Kilgarriff (2007) put it, "... two external influences need mentioning: (i) lexicography - different agenda but responsible for lots of the actual corpus-building work and innovation, at least in UK; BNC was lexicography-led; (ii) NLP / computational linguistics, which has come into the field like a schoolyard bully, forcing everything that's not computational into submission, collusion or the margins.” Further applications include using the tags to aid data compression of English text (Teahan 1998); and as a possible guide in the search for extra-terrestrial intelligence (Elliott and Atwell 2000). Specific uses and results make use of part-of-speech tag information. For example, searching and concordancing can be made more efficient through use of part-of-speech tags to separate different grammatical forms of a word. An indelicate annotation is sufficient for many NLP applications, e.g. grammatical error detection in Word Processing (Atwell 1983), training Neural Networks for grammatical analysis of text (Benello et al. 1989, Atwell 1993), or training statistical language processing models (Manning and Schütze 1999).

EAGLES guidelines for PoS-tagging (Leech et al 1996) aimed to extend PoS-tagging standards beyond the pioneering English corpora to corpus linguistics research in other languages. The EAGLES guidelines focus on enumerating the categories and sub-categories which apply across a range of European Union languages. However, developers of a tag-set for a corpus must also take into account a range of other issues, including: mnemonic tag names; underlying linguistic theory; classification by form or function; analysis of idiosyncratic words; categorization problems; tokenisation issues: defining what counts as a word; multi-word lexical items; target user and/or application; availability and/or adaptability of tagger software; adherence to standards; variations in genre, register, or type of language; and degree of delicacy of the tag-set.

In our presentation, we will examine a range of examples of tag set developments for different languages, to illustrate how these criteria apply. We consider standard tag-sets for an online Part-of-Speech tagging service for **English** (Atwell et al 2000); design of a tag-set for a closely related language, **German** (Schiller et al 1995); a tag-set for a language from a far-off branch of the broad Indo-European language family, **Urdu** (Hardie 2004); a tag-set for a non-Indo-European language with a highly inflexional grammar, **Arabic** (Khoja 2003); and a Part-of-Speech tag-set for a contrasting non-Indo-European language with isolating grammar, **Malay** (Knowles and Mod 2003). These criteria constitute a design checklist for Part-of-Speech tag-set developments for new corpora and languages.

A survey of previous practice is potentially more useful if it ends with some recommendations for the future. Corpus Linguistics and Natural Language Processing researchers are increasingly working with very large corpora; whereas pioneering Brown and LOB corpus projects took several years to collate and PoS-tag one million words of text, the current “web-as-corpus” approach is allowing corpus linguists to collate corpora of one hundred million words in weeks or even days. When PoS-tagging a very large web-as-corpus, it is not practical to consider manual analysis or even manual post-editing and correction of tagging-program output; we have to rely on a highly-accurate PoS-tagger program. So, it is even more important to decide at the outset on a part-of-speech tag-set which can minimize error-rate while maintaining linguistic integrity; and also to use a PoS-tagger program which can use all the tricks of the trade to apply this tag-set with minimal errors. We conclude by recommending a combination of strategies to improve accuracy of future PoS-tagging: we advocate the development of an Open-source Knowledge-rich Hybrid Adaptive Adaptable Multilingual Architecture for Web-As-Corpus PoS-Tagging.

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*. Technical Report CSI-R9604, Computing Science Institute, University of Nijmegen.

Al-Sulaiti, Latifa, and Eric Atwell. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11: 135–71.

Andersen, Gisle, and Anna-Brita Stenstrom. 1996. COLT: a progress report. *ICAME Journal* 20: 133–36.

Archer, Dawn, Paul Rayson, Andrew Wilson, and Tony McEnery (editors) 2003. *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster University: UCREL Technical Paper 16.

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. 1989. *Grammatical analysis of SCRIBE: Spoken Corpus Recordings In British English*. SERC Advanced Research Fellowship proposal, Science and Engineering Research Council.

Atwell, Eric. 1987. A parsing expert system which learns from corpus analysis. In Meijs, Willem, (editor), *Corpus Linguistics and Beyond: Proceedings of the ICAME 7th International Conference on English Language Research on Computerised Corpora*, pp 227–35, Amsterdam, Rodopi.

Atwell, Eric. 1993. Corpus-based statistical modelling of English grammar. In Clive Souter and Eric Atwell (eds) *Corpus-based Computational Linguistics*. Amsterdam: Rodopi.

Atwell, Eric. 1996. Comparative Evaluation of Grammatical Annotation Models. In (Sutcliffe et al. 1997).

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 Schriffin, Clive Souter, and Sean Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal*, 24: 7–23.

Atwell, Eric. A word-token-based machine learning algorithm for neoposy: coining new parts of speech in: Archer et al 2003, 43–47.

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

Atwell, Eric, Latifa Al-Sulaiti, Saleh Al-Osaimi, Bayan Abu Shawar. 2004. A review of Arabic corpus analysis tools. In: Bel, B and Marlien, I (editors). Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2, pp. 229–34 ATALA.

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)

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.

Corpora journal. 2006-present. http://www.eup.ed.ac.uk/journals/content.aspx?pageId=1&journalId=12505

Cure, The. 1980. *A forest*. Fiction Records.

EAGLES. 1996. WWW site for European Advisory Group on Language Engineering Standards, http://www.ilc.pi.cnr.it/EAGLES96/home.html Specifically: Leech, Geoffrey, Ros 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.

Fang, Alex. 2005. Robust practical parsing of English with an automatically generated grammar. PhD thesis, University College London.

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*, pp. 167–80. London: Longman.

Greene, B. and G. 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*. pp. 11–24. Amsterdam: Rodopi.

Grefenstette, Gregory. 1996. Using the SEXTANT low-level parser to analyse a software manual corpus. In (Sutcliffe et al 1996).

Hardie, Andrew. 2003. Developing a tagset for automated Part-of-Speech Tagging in Urdu. In: Archer et al 2003, 298–307.

Hardie, Andrew. 2004. The computational analysis of morphosyntactic categories in Urdu. PhD thesis, University of Lancaster.

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)*. pp. 535–39. 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.

ICAME Journal of the International Computer Archive of Modern and medieval English. 1978-present. http://icame.uib.no/journal.html

International Journal of Corpus Linguistics. 1996-present. http://www.benjamins.nl/cgi-bin/t_seriesview.cgi?series=ijcl

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.

Karlsson, F., A Voutilainen, J Heikkila, and A Anttila. 1995. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.* Berlin: Mouton de Gruyter.

Khoja, Shereen, Roger Garside and Gerry Knowles. 2001. A tagset for the morphosyntactic tagging of Arabic. In: Rayson, Paul; Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds.), Proceedings of the Corpus Linguistics 2001 Conference, p. 341, UCREL Technical Paper 13, Lancaster University

Khoja, Shereen. 2003. APT: an Automatic Arabic Part-of-speech Tagger. Ph.D. thesis, Lancaster University.

Kilgarriff, Adam. 2007. Message to CORPORA@uib.no discussion forum on the history of corpus linguistics

Knowles, Gerry, and Zuraidah Mohd Don. 2003. Tagging a corpus of Malay texts, and coping with "syntactic drift". In: Archer et al 2003, 422–28.

Kytö, Merja and Atro 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.* pp. 482–88.

Lin, Dekang. 1996. Using PRINCIPAR to analyse a software manual corpus. In (Sutcliffe et al 1996).

Lüdeling, Anke, and Stefan Evert (2003), Linguistic experience and productivity: corpus evidence for fine-grained distinctions. In: Archer et al 2003, 475–83.

Madonna. 1984. *Like a virgin*. Sire Records.

man 1986. *parts.* The on-line Unix manual.

Manning, Christopher and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Marcus, Mitch, M Marcinkiewicz, and Barbara Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19: 313–30.

O'Donoghue, Timothy. 1991. Taking a parsed corpus to the cleaners: the EPOW corpus. *ICAME Journal* 15: 55–62

O'Donoghue, Timothy. 1993. Reversing the process of generation in Systemic Grammar. PhD thesis, University of Leeds.

Oostdijk, Nelleke. 1996. Using the TOSCA analysis system to analyse a software manual corpus. In (Sutcliffe et al 1996).

Osborne, Miles. 1996. Using the Robust Alvey Natural Language Toolkit to analye a software manual corpus. In (Sutcliffe et al 1996).

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.

Roberts, Andrew, Latifa Al-Sulaiti and Eric Atwell. 2006. aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora journal*, 1: 39–57

Santorini, Beatrice. 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.

Schiller, Anne, Simone Teufel, and C. Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.
http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

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.

Souter, Clive. 1996. A corpus-trained parser for systemic-functional syntax. PhD thesis, University of Leeds.

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).

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.

Teahan, Bill. 1998. *Modelling English Text.* PhD Thesis, Department of Computer Science, University of Waikato, New Zealand.

Treebank 1999. The Penn Treebank Project www.cis.upenn.edu/~treebank/home.html