23. Development of tag sets for part-of-speech tagging

1. Introduction: Parts-of-speech and pos-tag sets

This article discusses tag sets used when pos-tagging a corpus, that is, enriching a corpus by adding a part-of-speech tag 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 in section 1, since English was the first language studied by corpus linguists. Pioneering corpus linguists thought that their English corpora could be more useful research resources if each word was annotated with a part-of-speech label or tag. Traditional English grammars generally provide eight basic parts-of-speech, derived from Latin grammar. However, most tag set developers wanted to capture finer grammatical distinctions, leading to larger tag sets. Pos-tagged English corpora have been used in a wide range of applications.

Section 2 examines criteria used in development of English corpus part-of-speech tag sets: 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.

To illustrate these issues, section 3 outlines a range of examples of tag set developments for different languages, and discusses how these criteria apply. First we consider tag sets for an online part-of-speech tagging service for **English**; then design of a tag set for another language from the same broad Indo-European language family, **Urdu**; then for a non-Indo-European language with a highly inflexional grammar, **Arabic**; then for a contrasting non-Indo-European language with isolating grammar, **Malay**.

Finally, we present some conclusions in section 4, and references in section 5.

1.1. General-purpose pos-tags for pioneering English corpora

English was the first language of Corpus Linguistics; the first journal of the new research field, the *ICAME Journal* of the International Computer Archive of Modern and Medieval English, reflected this initial focus in its title and contents. Later, Corpus Linguistics

extended to other languages. New journals have sprung up to cater for this wider range; for example, the first issue of *Corpora*, the latest Corpus Linguistics journal (founded nearly 30 years after *ICAME Journal*), included papers on Arabic and Spanish (as well as English).

The pioneering Corpus Linguists who collected 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; see also article 20) all thought that their corpora could be more useful research resources if the source text samples were enriched with linguistic analyses. 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-tag sets used in these English general-purpose corpora are illustrated in Table 23.1, 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/Stenström 1996); prepared speech for radio broadcasts, from SEC, the Spoken English Corpus (Taylor/Knowles 1988); and written text in software manuals, from IPSM, the Industrial Parsing of Software Manuals corpus (Sutcliffe/Koch/McElligott 1996). The example sentence in Table 23.1 is from the software manuals section.

The pos-tagging schemes illustrated in Table 23.1 include: Brown corpus (Greene/Rubin 1981), LOB: Lancaster-Oslo/Bergen corpus (Atwell 1982; Johansson et al. 1986), SEC: Spoken English Corpus (Taylor/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* (Hanks 1979), and the basic PARTS tag set used to tag the SCRIBE corpus (Atwell 1989).

1.2. Traditional parts-of-speech

School textbooks, in England at least, generally state that there are eight parts-of-speech in English, derived from traditional Latin grammatical categories: *noun, verb, adjective, preposition, pronoun, adverb, conjunction,* and *interjection*. These traditional English parts-of-speech are usually defined in terms of syntactic function (e. g. a noun can function as the head of a noun phrase, the subject or object of a verb), and morphological patterns of grammatical forms (e. g. a noun can have singular and plural forms, but an adjective cannot − in English). These distinctions are explained by showing typical examples. However, this overlooks problematic borderline cases; syntactic and morphological criteria can occasionally conflict. For example, I work in the School of Comput-

Tab. 23.1: Example sentence illustrating rival English pos-taggings (from the AMALGAM multi-tagged corpus)

	Collins English Dictionary	SCRIBE parts	Brown	LOB	Upenn	BNC-C5	BNC-C6	ICE	PoW	LLC
If	s.conjuntion	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)	.	.

ing at Leeds University; "computing" after the preposition "of" behaves syntactically as a noun, but morphologically is an inflected form of "compute", a verb. Idiosyncratic words which do not readily fit a category can also be problematic, for example "not". Some grammar descriptions try to cope with problems by extending the categories. For example, the *Collins English Dictionary* extends the traditional eight parts-of-speech by including *determiner* for "this", "that", "my", "his", "a", "some", "any" etc.; and introducing some sub-classifications, for example a distinction between *coordinating conjunctions* "and", "but", "or" etc., and *subordinating conjunctions* "where", "until", "before" etc.

1.3. Why not just use traditional parts-of-speech?

For most linguists developing part-of-speech tag sets and taggers, this is not enough: they may want to capture other grammatical distinctions, including morphological subcategories such as number for nouns and tense and person for verbs, and/or syntactic subcategories such as making a distinction between adjectives in attributive and predicative positions. This is why most of the pos-tag sets in Table 23.1 use far more than eight tags.

Before you develop a part-of-speech tag set, or decide to re-use an existing pos-tag set, you should be clear about why you want to pos-tag your corpus. For developers of general-purpose corpus resources, the aim may be to enrich the text with linguistic analyses to maximize the potential for corpus re-use in a wide range of applications. Since these applications are not known in advance, the level of enrichment required is also unknown, so it is tempting to add as much linguistic enrichment as feasible. Corpus linguists have tended to devise pos-tag sets with very fine-grained grammatical distinctions; these pos-tag sets reflect their expert interest in syntax and morphology, rather than specific predicted needs of end-users.

On the other hand, very fine-grained distinctions may cause problems for automatic tagging if some words in English can change grammatical tag depending on function and context. For example, if the tag set tries to distinguish between attributive adjectives and predicative adjectives, then most (but not all) English adjectives have more than one possible tag to choose from according to context, making the task of pos-tagging adjectives non-trivial. This has influenced some pos-tagger developers to favour pos-tag distinctions which avoid computational difficulties. Notwithstanding, other pos-tag designers have chosen to make linguistically-motivated distinctions despite computational problems this may bring; for example, the **Stuttgart-Tübingen Tag Set (STTS)** for German (Schiller/Teufel/Thielen 1995; Thielen et al. 1999; also see http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html) has exactly this distinction between attributive and predicative adjectives.

1.4. Corpus applications which use pos-tags

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 grammati-

cal analysis or tagging of English texts (e. g. Leech/Garside/Atwell 1983; Atwell 1983; Booth 1985; Owen 1987; Souter 1989a; O'Donoghue 1991; Belmore 1991; Kytö/Voutilainen 1995; Aarts 1996; Qiao/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/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/Mackie/Anderson 1989; Atwell 1993), or training statistical language processing models (Manning/Schütze 1999).

2. Criteria for tag set development: Differences between English corpus part-of-speech tag sets

Table 23.1 illustrates a range of alternative English corpus part-of-speech tag sets. The rival tag sets display differences (and similarities) along several dimensions. These dimensions are in effect choices to be made by developers of new pos-tag sets, for English or another language; in developing a new tag set, the designer must decide how to handle each dimension. Once a researcher has decided it would be useful to add part-of-speech tags to their corpus, they must decide on the tag set: decide on the set of grammatical tags or categories, and their definitions and boundaries. It may be attractive to simply adopt an existing tag set, but this still leaves the decision of which of several possible or rival tag sets to adopt, at least for English or other major European languages. If the language being studied is like a virgin, tagged for the very first time (cf. Madonna, 1984), then the researcher does not have the option to adopt an existing tag set; but they may still draw on parallels from other, more experienced languages.

The criteria to consider in deciding or developing the tag set include the underlying differences between the tag sets of Table 23.1. There are also a number of additional design criteria to take account of.

2.1. Mnemonic tag names

Generally the tag names are not arbitrary symbols, but chosen to help linguists remember the categories; for example several tag sets include CC for *Coordinating Conjunction*, and VB for *VerB*. However, sometimes a mnemonic value is not universally agreed. For example, Brown, LOB and UPenn contrast NN for *singular noun* and NNS for *plural*

noun, since -*s* is the standard suffix for plurals in English. However, the designers of the BNC tag sets decided that NNS might be mistakenly interpreted as *noun-singular*, so instead use NN1 for *singular noun* and NN2 for *plural noun*. The designers of the ICE tag set decided to use abbreviations rather than acronym-style mnemonics: for example, N(com,sing) for *singular common noun* and N(com,plu) for *plural common noun*.

2.2. Underlying linguistic theory

When a new tag set is developed by a linguist, they will inevitably be swayed by the linguistic theories they espouse. For example, the PoW corpus was collated and annotated by researchers interested in Systemic Functional Grammar, and the pos-tags reflect SFG analysis. Table 23.1 shows that words like "library" and "information" are tagged *noun* in most schemes, but H for *Head* (of a noun phrase) in PoW, showing the function of the noun in the syntactic structure. This makes it easier to parse the pos-tagged corpus with a SFG parser (e. g. O'Donoghue 1993; Souter 1996); but on the other hand a more traditional tagging in terms of nouns etc. would render a corpus more readily parseable by other parsers such as Principar (Lin 1994) or Sextant (Grefenstette 1996).

Another example of theoretical influence is in the ICE tagging scheme, developed later than others, at a time when grammar theories like Generalised Phrase Structure Grammar and Lexical Functional Grammar had promoted the notion that a category is composed of a bundle of features. Whereas earlier tag sets implicitly encoded some grammatical features (e. g. in LOB and Brown, a tag ending S was generally a plural) ICE tags explicitly show the bundle of features. This is more useful for feature-based parsers (e. g. Briscoe/Carroll 1993; Fang 2005).

However, most rival grammar theories like GPSG, LFG, GB etc. differ mainly in how they handle phrase structure, and more complicated structural issues such as the analysis of WH-questions. They generally had little to add to "traditional English grammar" on the issue of word categories, and there are no English corpus pos-tagging schemes which are closely tied to GPSG or LFG, for example.

Some researchers have been more interested in applications beyond linguistics. For example, the UPenn corpus was developed at least partly for researchers in Computer Science, Artificial Intelligence and Machine Learning. Machine Learning researchers using a part-of-speech tagged corpus for their ML experiments may not be concerned whether distinctions conform to a specific linguistic theory; but they will want a tagging which is readily Machine-Learnable.

Some corpus linguists may claim their part-of-speech tag sets are "theory-neutral"; but then why do so many rival part-of-speech tag sets abound? It is really not possible to have a theory-neutral annotation, every tagging scheme makes some theoretical assumptions.

2.3. Classification by form or function?

Traditionally, parts-of-speech are defined in terms of paradigmatic forms (for example, a word is a noun if it can be inflected to singular and plural forms), and syntagmatic functions (for example, a word is a noun if it can appear in specific sentence-slots such

as head of a noun phrase). Usually paradigmatic and syntagmatic criteria coincide, but there are some exceptional cases, and different English corpus tag sets may handle these borderline cases differently. For example, in English text, most words with suffix "-ing" are inflected forms of verbs, e. g. "dancing" is derived from the verb "dance", just as "computing" is derived from the verb "compute". So, it is tempting to always tag words with "-ing" suffix as verb-derivatives: VBG in several tag sets. However, "dancing" can also function as an adjective or a noun; and the LOB tag set designers decided to tag "-ing" words according to function rather than form, so "dancing" must be tagged as any one of VBG, NN or JJ depending on syntactic function in context.

2.4. Idiosyncratic words

English has a number of words with special, idiosyncratic behaviour; particles which do not fit into traditional parts-of-speech. Different tag sets may analyse these differently. For example, "a" is allowed a special *article* tag AT in the Brown and LOB tag sets, but is lumped in with *determiner* DT in the UPenn tag set. The word "to" is always a *preposition* in the *Collins English Dictionary* part-of-speech categories, even when preceding a verb infinitive (e. g. "to go"); whereas most other tag sets have a special tag for infinitival "to", different from the preposition tag for other uses of "to". Another example is the word "one", which has a range of grammatical roles in English. In the LOB tag set, "one" is simply tagged CD1 in all cases, but the ICE tag set has four separate tags for different functions, which a tagger has to try to separately identify.

2.5. Categorization problems

If a corpus linguist wants to design a detailed categorisation scheme, with many more than the eight basic categories, then it is not enough to provide a list of tags: each tag must be defined clearly and unambiguously, giving examples in a "case law" document. The definitions should include how to decide difficult, borderline cases, so that all examples in the corpus can be tagged consistently. For example, the Brown corpus manual specifies a general adverb tag **RB**, and a specialised tag for adverbs used as qualifiers, **QL**. However, it is not clear what limitations there are on the use of **QL**, leading to apparent internal inconsistency in the tagging of adverbs in Brown: a few adverbs appear tagged sometimes **RB** and sometimes **QL,** without any clear rationale. Another example is that most English tag sets have distinct tags for proper nouns and common nouns. It is easy to give prototypical examples of these two categories, but analysis of a corpus tends to throw up problem cases, so the tagging scheme guidelines must specify how to handle these grey areas. For example, product names like "Perrier Water", "International Journal of Corpus Linguistics" could be analysed as including common nouns, or alternatively the name could be tagged as a sequence of proper nouns.

In English, many words can belong to more than one grammatical category; for example, "water" can be used as a noun or a verb. Where a word can have different pos-tags in different contexts, tagging schemes should specify how to choose one tag as appropriate.

However, the UPenn tagging scheme incorporates special 'vertical slash' tags for (very rare) occasions when the part-of-speech is genuinely ambiguous. Consider the sentence: **The duchess was entertaining last night.** (This example is taken from Santorini 1990) Does *entertaining* mean that she was hosting an event, in which case the word would be a present participle verb, *VBG*, or does *entertaining* act adjectively (*JJ*) implying that the Duchess was good company? Either analysis is plausible, and even the surrounding context may not help in reaching a decision; so the Penn Treebank developers allow both tags to apply at the same time in this rare special case. In this case, *entertaining* is legitimately assigned the slash tag *JJ|VBG*. However, in the great majority of other uses of **entertaining**, the context can be used to disambiguate the word, so it should be tagged EITHER **JJ** or **VBG**. (Many cases of genuine ambiguity result in or perhaps reflect syntactic ambiguity, see articles 13 and 28.)

2.6. Tokenisation issues: What counts as a word?

Generally, English text is divided into words by spaces; punctuation and text-formatting can complicate this task of tokenisation, but not much (see article 24). In English, the main exceptions to this generalization are verb-contractions and genitives; and different pos-tag schemes deal with these exceptions differently. In the UPenn scheme, verb contractions and the Anglo-Saxon genitive of nouns are split into their component morphemes, and each morpheme is tagged separately; for example:

children's → children 's parents' → parents '
won't → wo n't
gonna → gon na
I'm → I 'm (from Treebank 1999)

In contrast, the London-Lund tagging scheme uses 'combined tags' for words such as *don't* (VD+0*AN) and *I've* (RA*VH+0). All combined tags have the same form: an asterisk separates the tags for the different tokens that make up the complete combined word. The Brown tagging scheme also uses 'combined tags' for words such as won't (MD*) and I'd (PPSS+HVD). Combined tags come in only these two forms: either negated words have an asterisk appended after their tag or the plus symbol separates the tags for the different tokens that make up the complete combined word.

2.7. Multi-word lexical items

A related problem area is the treatment of multi-word lexical items, also known as idiomatic phrases. For example, "as well" in Table 23.1 is equivalent to "also" or "too". The *Collins English Dictionary* does not specify how to tag this; and the Brown and UPenn tagging schemes insist on one tag per word, treating this as a sequence of adverb/qualifier + adverb. In contrast, the PoW tagging scheme simply supplies one tag (AL) for the phrase. Other tagging schemes include special tags for multi-word lexical items. The LOB tagging scheme introduced Ditto tags, applied to words whose role changes

from their normal syntax when applied in certain combinations. The first word of the combination is tagged as normal and all subsequent words are given the first word's tag plus the ditto symbol (''). For example, the combination "so as to" is tagged TO TO'' TO''. The BNC tag sets C5 and C6 have a more complicated equivalent of ditto-tags: the phrase is given a single tag, AV0 in BNC-C5 or RR in BNC-C6; then each word has a variant of this general tag, with a 2-digit suffix, showing the length of the phrase, and the position within the phrase of this word. So, RR21 means "adverb, 2-word phrase, first word"; and RR22 means "adverb, 2-word phrase, second word".

What counts as a "multi-word lexical item" is also variable. For example, the BNC tag set treats "for example" as a single adverb (RR21 RR22 or AV021 AV022), whereas other tag sets assume this is preposition + noun.

To summarize, there is not always a one-to-one mapping between token and pos-tag. Sometimes a token contains several pos-tags (at least on some level) and sometimes several tokens have a common pos-tag. For more on multi-word lexical patterns, see article 58.

2.8. Target users and/or application

This relates back to section 1. The most important criterion is to satisfy the customer; the final tag set should be evaluated in terms of fit for purpose, and/or customer satisfaction. For example, developers of the LOB corpus thought its main use could be in English language teaching and research, and developed a comparatively complex tag set to reflect fine distinctions of English grammar for learners and teachers. Developers of the UPenn tagged corpus saw more use in language engineering, as a training set for Machine Learning systems which would cope better with a smaller tag set.

Many specific uses of corpora do not need delicate, detailed tag sets. However, the corpus developer should bear in mind the potential for re-use: a small tag set aimed at an immediate customer/application may turn out to be too limited for wider re-use of the corpus in future research. This is one reason why most English corpus pioneers developed sophisticated pos-tag schemes.

2.9. Availability and/or adaptability of tagger software

It is convenient to be able to automate part-of-speech tagging of a corpus, so a part-of-speech tag set which comes with a part-of-speech tagger program has a clear advantage over a purely theoretical tag set. An additional criterion may be the accuracy level of the tagger: it is tempting to adopt a tag set because it can be computed highly accurately, such as the ENGCG tagging system. For a virgin, un-tagged language, it may be possible to adapt a tagger program developed for another language, and this is generally more straightforward if the tag set for the new language parallels the old language tag set.

For example, Brill's tagger (Brill, 1993, 1995) was originally developed for pos-tagging English texts, but has been adapted to several languages including French (Lecomte 1998) and Arabic (Freeman 2001). Similarly, the CLAWS tagger (Leech/Garside/Atwell 1983), originally developed to pos-tag the LOB corpus of British English, has been

adapted to other languages including Urdu (Hardie 2003) and Arabic (Khoja/Garside/Knowles 2001, Khoja 2003). In practice, a tagging scheme is inevitably influenced by the tractability of decisions made by the associated tagger program; for example, it is not always easy for a program to decide what function the word "one" has in an English text, so most English tagging schemes (except ICE) just have one tag for "one". It follows that the tag sets assigned to French, Urdu and Arabic texts by taggers originally built for tagging English may have been influenced by the English tag sets assigned by the original programs.

2.10. Adherence to standards

The EAGLES project (see Leech/Barnett/Kahrel 1996) embarked upon the task of setting standards for corpus annotation. The EAGLES guidelines propose a set of grammatical features to recognize in tag sets, including recommended and optional features. A language-neutral "intermediate tag set" is provided, using a numeric (non-mnemonic) coding for each feature; for example, *singular common noun* is N101000, *plural common noun* is N102000, *main verb infinitive* is V0002500100000. Each digit corresponds to a specific EAGLES-recognised grammatical feature; a zero shows this feature is not present (though it may be in other languages, e. g. *gender* in nouns). These intermediate tags are useful in allowing direct computational comparisons of tag sets across two or more languages, but clearly they are hard for humans to digest. Instead, corpus developers are free to use simpler mnemonics, as long as there is a simple mapping between intermediate tags and "human-friendly" tags.

Linguists devising a tag set for a virgin language may try to conform to agreed standards, for example tag sets partly conforming to the EAGLES guidelines have been applied beyond the original EU member-state languages, including to Urdu (Hardie 2003) and Arabic (Khoja/Garside/Knowles 2001). However, this may be an unwarranted imposition: for example, EAGLES guidelines come into conflict with some categories from traditional Arabic linguistics and grammar; and European part-of-speech categories may be quite inappropriate for Malay (Knowles/Don 2003).

Another type of standard is the de-facto widespread adoption of an existing tag set with significant credentials or backers. For example, the Brown Corpus was the first to be part-of-speech tagged, and it was the first major American corpus, which led to its widespread use in American computational linguistics research. Arguably other tag sets evolved from the Brown set, such as LOB and then ICE, have linguistic merit, but they have achieved less exposure in the American-dominated computational linguistics community.

For more on standards in Corpus Linguistics, see article 22.

2.11. Genre, register or type of language

To some linguists, the type of language may have a bearing on the grammatical categories to be employed; for example, they may think that to some extent spoken and written languages have different grammars. Spoken texts may include hesitations, repetitions,

false starts, incomplete or partly-inaudible phrases, and other disfluencies not found in written texts; and they may include more informal or non-standard vocabulary and grammar. Some tag sets were developed for specialised corpora, e. g. the PoW corpus of children's conversations, or the Brown and LOB Corpora of written, published English. However the tag sets may still apply to other types of language, for example the LOB tag set was readily applied to the SEC Spoken English Corpus. Other tag sets were developed for corpora which deliberately aimed for a diverse variety of language, e. g. the ICE and BNC corpora contain both written and spoken language, and the respective ICE and BNC tag sets cover both.

2.12. Degree of delicacy of the tag set

It may seem strange to leave this issue to last: arguably the most obvious difference between rival tag sets for English corpora, for instance, is the number of tags, indicating the level of fine-graininess of analysis. However, this decision is heavily influenced by other criteria, which in effect are also decisions about delicacy; and decisions about the target application, available tagger software, standards to be adopted, and genre of the language may leave little room for debate on the appropriate level of delicacy. For example, section 2.8. suggested that the main reason for the difference in number of tags, or degree of delicacy, between the LOB and UPenn tag sets was the target user-group foreseen by the tag set developers.

3. Case studies of tag set development

To illustrate these issues, we outline a range of examples of tag set development and discuss how these criteria apply. First we consider tag sets for an online part-of-speech tagging service for **English**; then design of a tag set for another language from the same broad Indo-European language family, **Urdu**; then for a non-Indo-European language with a highly inflexional grammar, **Arabic**; then for a contrasting non-Indo-European language with isolating grammar, **Malay**.

3.1. Tag sets for an online English corpus part-of-speech tagging service

The AMALGAM project set up a free-to-use part-of-speech tagging service for the English Corpus Linguistics community (Atwell et al. 2000). For this diverse audience, it was decided NOT to develop or adopt a single standard tag set, but to allow users to choose from a range of options from the pioneering English corpus pos-tag sets illustrated in Table 23.1.

The Amalgam project team also tried some experiments with devising a set of mapping rules from one tag set to another (Hughes/Atwell 1994; Atwell/Hughes/Souter 1994; Hughes/Souter/Atwell 1995). The main lesson learnt was that this is non-trivial: the differences between tag sets cover the range of dimensions listed in section 2, and it was

not feasible to draw up a simple set of mapping rules coping with all these dimensions of difference.

By offering a choice of the tag sets from Table 23.1, the Amalgam service managed to sit on the fence and leave users to make their own choices with regards to most of the tag set design criteria listed in section 2.

3.1.1. Mnemonic tag names

Users can choose from the mnemonic tag name schemes in Table 23.1, to suit their preferences and needs.

3.1.2. Underlying linguistic theory

Most of the tag sets in Table 23.1 are (claimed to be) "theory-neutral", although the PoW tag set does illustrate one modern "school" or theory of grammar, Systemic Functional grammar. Unfortunately, the Amalgam service could not offer tagging based on other modern theoretical approaches to grammar, such as GPSG or LFG, because no tagged training corpus was available to re-train the tagger.

3.1.3. Classification by form or function?

All of the alternatives illustrated in Table 23.1 were made available. These tag sets generally classify by function: words can have more than one pos-tag, varying according to syntactic context or function.

3.1.4. Idiosyncratic words

The tag sets in Table 23.1 include some variation in treatment of idiosyncratic words; for example, the different treatment of "one" in LOB and ICE mentioned in section 2.4.

3.1.5. Categorization problems

Table 23.1 illustrates one use of the Brown **QL** adverbial qualifier tag, on the word "as" when qualifying the adverb "well". However this is an idiosyncratic qualifier role, in that most other tagging schemes see "as well" as a single multi-word idiom: "as well" is not a typical example to illustrate or define the use of **QL**. Unfortunately, the tagged Brown corpus manual does not define **QL** much more clearly, beyond some more examples of its use.

Table 23.1 also illustrates different attitudes to avoiding inconsistency in distinguishing common and proper nouns, in the software name "the Dynix Gateways product". "Dynix" is clearly a proper name as it does not coincide with any common noun. "Gate-

ways" could also be a proper name; but later in the sentence, "gateways" (without word-initial capital) is unanimously voted as a common noun by every tagging scheme. The ~~LOB,~~ BNC~~,~~ and ICE schemes choose to categorise "Gateways" as a common noun consistently, regardless of word-initial-capital; whereas the Brown, UPenn, PoW and LLC schemes rule that the word-initial capitals consistently mark out nouns as proper, regardless of possible recurrence in the text in lower-case.

Most tag sets and tagging programs assume a word-type may have more than one pos-tag, but each specific word-token must have one pos-tag, decidable from the context (at least in principle). As explained in section 2.5., the UPenn tag set stands out by allowing "vertical slash" tag-pairs for genuinely ambiguous cases. However, the AMALGAM service was based on an automatic tagging program (the Brill tagger) which was unable to distinguish genuinely ambiguous words from the vast majority of words which can safely be given just one tag; so in practice, "vertical slash" tags are not used in processing texts to be tagged.

3.1.6. Tokenisation: Dividing text into words

The text to be tagged is first passed through a tokeniser which applies various formatting rules to divide the text into words. This can be turned off when mailing data to amalgam-tagger, by specifying 'notoken'. The different English corpus tagging schemes tokenise some special cases differently, as outlined in section 2.6. The AMALGAM tokeniser is in effect a compromise algorithm which tokenizes text the same way for all tag sets, to simplify alignment and direct comparisons of rival taggings of the same text, as illustrated by Table 23.1. Users who require a different tokenization are recommended to tokenise the text themselves, and then turn off tokenization (via 'notoken') when using the Amalgam sevice.

3.1.7. Multi-word lexical items

Table 23.1 illustrates some alternative treatments of the multi-word lexical item "as well". However, this is NOT the direct output of the AMALGAM tagger service: it has been proofread and hand-corrected to reflect the tagging described in the handbooks or other documentation defining each tag set. The Amalgam service (based on the Brill tagger) attempts to supply a pos-tag to every "token" passed from the tokeniser, and it does NOT include a special module for analysis of multi-word lexical items (on the basis that these account for a tiny proportion of the total words in a corpus, not worth the additional processing complexity which would be required to handle these properly). In general, this results in incorrect tagging of most multi-word lexical items.

3.1.8. Target users and/or application

The AMALGAM tagging service was aimed at casual users, who wanted to explore whether and how pos-tags might be useful in their research or teaching, without having to install and set up tagging software on their own computer. As part of such explorative,

speculative use, many users tried more than one available pos-tag set, to discover which would be most appropriate to their needs. This implies that an online pos-tagging service should indeed offer a variety of pos-tag sets to choose from, rather than just offering a single scheme.

3.1.9. Availability and/or adaptability of tagger software

The AMALGAM tagging service was powered by the Brill tagger, which could be re-trained to any pos-tag scheme embodied in an existing pos-tagged corpus. The Brill tagger can be freely downloaded from a website, and comes "pre-trained" on the tagged Brown corpus; it was fairly straightforward to re-train with other English pos-tagged corpora.

3.1.10. Adherence to standards

The AMALGAM project did not have access to a corpus tagged with EAGLES guidelines pos-tags, to re-train the Brill tagger; hence EAGLES-tagging is not available. However, the major English tagged corpora, particularly the Brown corpus, have become de-facto standards in Computational Linguistics research, so the Amalgam service does in effect allow users to work with "standard" tag sets. (Atwell et al. 2000, 18) notes: "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".

3.1.11. Genre, register or type of language

The tag sets offered by the AMALGAM service aim to cover a wide range of genres, including both spoken and written language. The only constraint noted in the help-file is: "Please note that the tagger is intended for English text − it will not work for languages other than English". This constraint was specified following user feedback that their French texts were not being pos-tagged correctly!

3.1.12 Degree of delicacy of the tag set

User feedback suggests that users select a tag set appropriate to their application on the basis of the above criteria, and NOT simply according to number of tags in the tag set.

3.2. Developing a tag set for another Indo-European language: Urdu

Urdu is outside the original European Union member state languages; however, it is an Indo-European language and hence one might expect that many of the EAGLES standard guidelines could be applied. (Hardie 2003, 2004) demonstrated this by developing a tag set for pos-tagging an Urdu corpus (~~Baker et al.~~ 2003).

3.2.1. Mnemonic tag names

The features recognized by the tag set were largely those of the EAGLES guidelines. However, instead of the numerical-code "intermediate tag set", (Hardie 2003, 2004) specifies mnemonic tags reminiscent of the BNC tag set; for example, II for *unmarked postposition* (in BNC-C6, *preposition)*; CC for *coordinating conjunction* (as in BNC-C6). In general, the first two letters show the broad word-class, then one character is used to denote each marked grammatical feature. Urdu has more complex inflexional morphology than English, so some tags are quite long; for example, JJF2N for *marked feminine plural nominative adjective*, NNMM1N for *common marked masculine singular nominative noun.*

3.2.2. Underlying linguistic theory

Urdu does not have a long-established tradition of grammatical description, so Hardie based his categories on a modern standard grammar textbook (Schmidt 1999). Other available sources tended to cover only specific aspects of Urdu grammar, for example features which learners need to learn first.

3.2.3. Classification by form or function?

The example tags given in section 3.2.1. are based on morphological form. The tag set also includes some distinct tags for different functions of certain words. However, the focus of Hardie (2003) is on the tag set rather than the tagging program or tagged corpus, and there are no examples which illustrate a form/function conflict.

3.2.4. Idiosyncratic words

Urdu has some idiosyncratic words which do not fit readily into the EAGLES categories; for example, there are special tags PA for the *honorific pronoun "āp"*, AL for *Arabic definite article "al"* found only with Arabic loan-words.

3.2.5. Categorization problems

The tags are defined by examples; Hardie (2003) does not discuss the issue of disambiguation of problem cases, although it does imply that some words may belong to more than one tag-class and hence a tagger will have to select a single tag in context.

3.2.6. Tokenisation: Dividing text into words

This was a challenge: "... many things described in the literature on Urdu grammar as suffixes are actually written as independent words ... For consistency, the (essentially arbitrary) decision was taken to treat every orthographic space as a word break even if

it occurs within a lexical word ... Word breaks are also introduced in some places where there is no orthographic space, e. g. where clitics precede/follow another word without a break." (Hardie 2003, 302). This led Hardie to add a tag LL for *non-grammatical lexical element*, to tag the initial token(s) of a multi-token lexical item. For example, Urdu for "telephone" is "teli fon" – the first token "~~tele~~" is tagged LL.

3.2.7. Multi-word lexical items

Hardie (2003, 302) notes that the above handling of *non-grammatical lexical elements* is "... only partially analogous to the problem of multi-word idioms in English and similar languages ... In these cases, there is also an analysable internal syntactic structure ... In the Urdu case, it would be very difficult to assign any internal structure to *teli fon* ...". However, Hardie does not discuss or illustrate how "true" equivalents of multi-word idioms would be treated in his tagging system.

3.2.8. Target users and/or application

There was no specific target user group or application, beyond the stated aim of developing a pos-tagger for one of the South Asian languages covered by the EMILLE project (Baker et al. 2003). Hence the tag set was generic and not constrained to a specific application, following the lead of pioneering English tagged corpora.

3.2.9. Availability and/or adaptability of tagger software

Hardie worked within the Lancaster UCREL tradition of Corpus Linguistics, and was undoubtedly influenced by the CLAWS heritage of taggers and tag sets for the LOB and BNC corpora. However, the tag set for Urdu was designed before a tagger, to ensure it was primarily based on linguistic principles, without compromising to suit computational feasibility or efficiency.

3.2.10. Adherence to standards

As already illustrated, the Urdu tag set fitted EAGLES guidelines, with some minor additions.

3.2.11. Genre, register or type of language

The tag set was developed for the Urdu component of the EMILLE corpus of South Asian languages (Baker et al. 2003), which includes both written texts (e. g. UK government advice leaflets) and spoken texts (e. g. transcripts of UK BBC Asian Network radio broadcasts). Hence, the tag set is not limited to one type of language.

3.2.12. Degree of delicacy of the tag set

Hardie does not explicitly state how many tags there are in his tag set; however, the list of tags in the Appendix shows that the complex morphology of Urdu generates a large number of tags to distinguish many possible combinations of grammatical features. For example, whereas English tag sets generally have a single *preposition* tag, the Urdu tag set has 10 different tags for *postpositions*, to capture possible feature combinations of *unmarked/marked/clitic, masculine/feminine, singular/plural, nominative/oblique.*

3.3. A tag set compatible with Arabic academic traditions

Arabic has been used and studied far longer than English. Classical Arabic was standardized around fourteen hundred years ago, when the Koran became in effect the definitive corpus of the language. Since then Muslim scholars have studied and documented the Arabic language and grammar, keeping it from straying too far from what they believed were the words of God, narrated to Mohammad by an angel, to be passed on verbatim to and by all believers. Modern Standard Arabic has added modern vocabulary, and avoids some of the more complicated grammatical forms, but is essentially the same language.

Western researchers have only recently shown much interest in Arabic, perhaps because of the very different script, morphology, lexis and grammar; and corpus linguists have only recently had open access to Arabic corpora (see Al-Sulaiti/Atwell 2006), and concordancers (Roberts/Al-Sulaiti/Atwell 2006). Corpus linguists have not attempted to apply EAGLES standards to Arabic, a non-Indo-European language. If they did, the tag set arrived at might well seem alien to Arabic linguists and grammarians. The Arabic tag set and part-of-speech Tagger developed by (Khoja/Garside/Knowles 2001; Khoja 2003) came from the Lancaster UCREL tradition of Corpus Linguistics, and like Hardie, Khoja was undoubtedly influenced by the CLAWS heritage of taggers and tag sets for the LOB and BNC corpora. However, Khoja's main influence was traditional Arabic grammatical theory, still used today in Modern Standard Arabic.

3.3.1. Mnemonic tag names

The tags cited in examples in (Khoja 2003) generally use one capital letter (sometimes with an extra lower-case letter) to show each grammatical feature, reminiscent of the LOB and BNC tag sets. For example, VPP12M is *verb perfect plural second-person masculine;* VISg2FI is *verb imperfect singular second-person feminine indicative*; NPrPDu3 is *noun pronoun personal dual third-person.* This illustrates the complex morphology of Arabic; but there are also some simpler tags, for example NP for *proper noun* (such as a name, by default singular). For the benefit of English-speaking corpus linguists, Khoja has used terminology from English grammar rather than Arabic tradition in naming categories and features; however, the tags do also have equivalents in Arabic script, for the benefit of Arabic linguists.

3.3.2. Underlying linguistic theory

Traditional Arabic grammarians recognize only three main parts-of-speech, which map roughly on to nouns, verbs, and particles. Hence, all pos-tags start with N, V or P. Other EAGLES traditional European categories are subclasses of one of these three, mainly on the basis that they share inflexional patterns; for example, pronouns and adjectives inflect like nouns so they are classed with nouns.

3.3.3. Classification by form or function?

As already stated, traditional Arabic grammar groups words according to their inflexional behaviour, which implies that word-class is dependent on form. A complication peculiar to Arabic is the writing system. Vowels are normally omitted in written Arabic, and left for the reader to infer; unfortunately, vowels can encode grammatical category or feature information. Typically a root or lexical item consists of three consonants, and the vowels between these consonants add grammatical information. For example, the three letters *ktb* can stand for the verb *kataba* meaning 'he wrote', or for the plural noun *kutub* 'books'. The result is that in a written text, many words are rendered ambiguous through lack of vowels, and a tagger has to work out classification of each word taking context or function into account. Human readers of Arabic text manage this disambiguation task, in effect subconsciously tagging the text to understand it.

One major exception to this is the Koran: to ensure it is pronounced (and parsed) correctly, vowels are traditionally included. This makes the Koran a potential "Gold Standard" corpus for Arabic tagging and NLP research.

3.3.4. Idiosyncratic words

Arabic grammatical tradition already recognizes a subclass of *Particle* translated as *Exceptions* by Khoja, to cover some idiosyncratic words which do not fit other patterns: "... These include the Arabic words that are equivalent to the word except and the prefixes non-, un-, and im-." (Khoja 2003, 52).

3.3.5. Categorization problems

Tags are explained by examples, but there is no detailed handbook of "case law" defining how to disambiguate problem cases. As explained above, absence of vowels in most printed text renders many of the words ambiguous. (Khoja 2003) reports results of some initial tagger program experiments, in which 18 %−23 % of words were unambiguous; but many of the remaining words were not in her lexicon or could not be handled by her stemmer, so these figures are only indicative.

The tagger progam uses the Viterbi algorithm (see article 24) to disambiguate words, and this assumes all words should only have one tag: there is no scope for accepting "truly ambiguous" cases, as in the UPenn tagging scheme.

3.3.6. Tokenisation: Dividing text into words

Particles are sometimes affixes; for example the definite article 'al' is well-known as a prefix in Arabic loan-words in other languages, e. g. algebra, Algarve. These are handled by a compound tag, reminiscent of the Brown tagging scheme. For morphologically complex words a combination of tags is used. For example, the word *walktab* 'and the book' is given the tag PC+NCSgMND, where PC indicates a particle that is a conjunction, and NCSgMND indicates a singular, masculine, nominative, definite noun.

3.3.7. Multi-word lexical items

The prototype tagger reported in (Khoja 2003) was based on a lexicon of under 10,000 word-types, extracted from a corpus of about 50,000 word-tokens. Multi-word lexical items were not considered separately.

3.3.8. Target users and/or application

The Arabic tag set was developed for general Corpus Linguistics research, and so aimed to be generic, analogous to the pioneer English corpus tag sets of Table 23.1.

3.3.9. Availability and/or adaptability of tagger software

The CLAWS system was available to Khoja and her Lancaster colleagues, but her Arabic tag set was not unduly influenced by this: the guiding principle was compatibility with Arabic grammar tradition.

3.3.10. Adherence to standards

The standards adhered to in this case were those of Arabic grammar tradition. However, the English translations of category and feature names were drawn from standard terminology found in the EAGLES guidelines.

3.3.11. Genre, register or type of language

The initial 50,000-word training corpus was extracted from the Saudi Al-Jazirah newspaper (date 03/03/1999); initial tagging experiments were done on other newspaper texts, and a social science paper. However, given that the tag set seems to be as general as analogous tag sets developed at Lancaster such as BNC, we can hope that the tag set will cover other genres, and spoken texts.

3.3.12. Degree of delicacy of the tag set

As with Hardie's tag set for Urdu, the complex inflexional morphology generates many possible combinations of grammatical features, leading to a large number of tags. Khoja states there are 131 tags, but (presumably) this does not include all possible combination-tags for morphologically complex words, such as the example cited above, PC+NCSgMND.

3.4. A tag set for Malay corpus linguistics

Western researchers have also tried to apply Indo-European grammatical concepts to Malay, another non-Indo-European language. In contrast to Arabic, there is no "home-grown" Malay tradition of grammar, other than the ex-colonial tradition of applying concepts from English. Knowles/Don (2003, 2005) are developing a tag set to use in tagging the Dewan Bahasa dan Pustaka (DBP) 80-million-word corpus of Malay texts; their experience suggests that the "English tradition" may be misplaced, and Malay may be better served by a drastically rethought notion of part-of-speech.

3.4.1. Mnemonic tag names

The example tags cited in (Knowles/Don 2003) use Malay tag-names; for example *kata sifat* 'adjective', *kata nama* 'noun', *kata kerja* 'verb'. They do this to distinguish form from function (see below), and also to dissuade users from assuming direct parallels with European or Arabic categories. Unfortunately, they do not present a complete list of tags, but say: "... some of our class labels look like traditional parts-of-speech, but the underlying definitions are entirely different" (ibid., 423).

3.4.2. Underlying linguistic theory

Knowles/Don (2003, 423) note that "... European parts-of-speech are the accepted point of departure for considering grammatical class in Malay (see Asmah, 1993; Sneddon, 1996)". An alternative is to apply the Arabic tradition of three major grammatical categories *noun*, *verb, particle* (e. g. Abdullah 1974). However, Knowles/Don (2003, 424) argue that a tag set for Malay must take account of 'syntactic drift': "... The class of many words in European languages is made unambiguous by their morphology ... However, in view of the lack of any inflectional morphology, Malay has a large number of simplex forms which belong to no clearly defined class, and appear to 'drift' from one class into another. For example, *masuk* is the normal word for 'enter', which makes it a kind of verb; but it is used in such a way on buildings and in carparks that it could also be taken to be a noun 'entrance' ...".

This 'drift' has been recognized by others, for example (Lewis 1947, xvii): "... Malay words change their function according to context. Be prepared for this, and do not attempt to force the language into a set mould. It will escape".

3.4.3. Classification by form or function?

Knowles and Don's radical response to 'syntactic drift' is to separate lexical class or form from syntactic function, and give each word in the lexicon only one class-tag. "... we use the term 'tag' to label a lexical class, and 'slot' to refer to a position in syntactic structure. We maintain the distinction consistently by giving lexical classes Malay labels, and syntactic slots and constructions English labels" (Knowles/Don 2003, 425).

3.4.4. Idiosyncratic words

Knowles and Don's examples include a couple of idiosyncratic words: *relative particle* 'yang' (e. g. 'pintu yang hijau': 'door that is green'); and *negative particle* 'tidak'. In general, Knowles and Don advocate that grammatically idiosyncratic words should NOT be given a special tag, but instead that the idiosyncratic grammatical behaviour should be captured by rules in the parser referring not to tags but to individual word forms: "... for example, the Malay expression corresponding to *o'clock* is a "verb" *pukul* normally meaning 'to hit, strike (e. g. a gong)'. To handle an expression such as *pukul tiga* 'three o'clock', the parser has to test for the specific word *pukul* before a numeral, and so the fact that *pukul* is tagged in the lexicon as a "verb" causes no problem at all" (Knowles/Don 2003, 425).

3.4.5. Categorization problems

As explained above, Knowles and Don's radical approach to categorization problems is to avoid tagging ambiguity by allowing only one tag for each word in the lexicon. They suggest this could even work for English: for example, instead of saying a word like *telephone* is ambiguous between noun and verb, they suggest a single tag or lexical category for words which can function as either nouns or verbs, distinct from lexical categories *noun (only)* and *verb (only)*. Interestingly, this has also been suggested by Machine Learning research on unsupervised learning of word-clusters, (e. g. Atwell 1987; Atwell/Drakos 1987; Hughes/Atwell 1994).

3.4.6. Tokenisation: Dividing text into words

Although Malay lacks inflexional morphology, it does have derivational morphology: for example, *besar* 'big', *membesarkan* 'enlarge', *kebesaran* 'size'; or *baca* 'read', *pembaca* 'reader', *bacaan* 'reading'. Sometimes this leads to problematic tokenisation: for example, "... the "verb" *berbaju* 'wear a shirt' is formed from the "noun" *baju* 'shirt', and this noun can still be followed by an "adjective" such as *merha* 'red' ... the structure is not strictly ((berbaju)(merah)) but (ber(baju merah))" (Knowles/Don 2003, 426).

3.4.7. Multi-word lexical items

The only examples cited by Knowles and Don are multi-word adverbs. There is no separate lexical class of *adverb (only)* in Malay. Instead, a *kata sifat* 'adjective' can also

function as a verbal modifier; and Malay also has idiomatic adverbial constructions in which *dengan* 'with' or *secara* 'manner' is followed by a *kata sifat*, e. g. *dengan betul* 'with correct, correctly', *secara betul* 'manner correct, correctly'.

3.4.8. Target users and/or application

The most obvious user of this tag set is Dewan Bahasa dan Pustaka (DBP), the government body responsible for coordinating the use of the Malay language in Malaysia and Brunei. The work should be of wider interest in corpus linguistics, particularly cross-language studies may suggest ways in which the radically different approach to tagging may apply to other languages. As this is a pioneering first attempt to develop a tag set for Malay, analogous to the pioneering English tag sets, Knowles and Don have focused on generic, theoretical issues rather than designing a tag set for a specific application.

3.4.9. Availability and/or adaptability of tagger software

There is no tagger for Malay; furthermore, most existing taggers focus on techniques for choosing the best tag for ambiguous words, which makes them inappropriate for Knowles and Don's model of tagging.

3.4.10. Adherence to standards

Of all the tag sets discussed in this article, this is the furthest from compliance to any standards!

3.4.11. Genre, register or type of language

The sample Knowles and Don have worked on contains only literary texts: four modern novels. However, they make no suggestion that their tag set is limited to literary texts.

3.4.12. Degree of delicacy of the tag set

Knowles/Don (2003, 423) state "... Our tagset currently contains 119 tags in 19 different major classes". However, they only illustrate a few major word-classes, and give no further indication of delicacy of distinctions in the tag set.

4. Conclusions

English corpus linguists have developed a variety of tag sets for part-of-speech tagging, reflecting a range of target applications for pos-tagged corpora, pos-tagging software, linguistic intuitions and theories about categories and degree of delicacy required, adher-

ence to standards, genre or type of language to be analysed, and other factors or constraints. In practice, for many applications of pos-tagged text, a small pos-tag set with few delicate disctinctions is sufficient, and using more complex or delicate tag sets makes little difference. English language researchers may be better off using a pos-tag set for which an accurate tagging software system is readily available; for example, the BNC tag set is widely used by English language researchers, not so much because of its intrinsic superiority over rival tag sets, but because an online pos-tagging service is freely available for tagging of researchers' own texts. This has the knock-on bonus effect of making research results involving part-of-speech information more directly comparable, as the same tag set is used in these results.

Corpus Linguistics has expanded beyond English, and there are now a range of tagging schemes and tagged corpora for at least a few other languages (e. g. German, French, Chinese). However, for corpus linguists studying more exotic languages, particularly non-European languages, there is not the same wealth of existing tag sets to choose from. If a pioneering researcher has developed a pos-tag set for such a language (e. g. Arabic, Urdu, Malay), it is tempting to adopt this as a de-facto "standard" rather than invest time and effort in developing a custom-made tag set. However, the researcher should still evaluate whether an existing tag set is "fit for purpose", as pioneers are not always perfect: the first tag set developed for a language may not suit all applications. If there is no existing tag set for a "virgin" language, the developer can still learn from the range of experiences of Corpus Linguists working with other languages. Hopefully this article can help the researcher in evaluating existing candidates, and if necessary, in developing or revising a tag set suited to the target use or application.

5. Literature

Aarts, Jan (1996), A Tribute to W. Nelson Francis and Henry Kučera: Grammatical Annotation. In: *ICAME Journal* 20, 104–107.

Aarts, Jan/van Halteren, Hans/Oostdijk, Nelleke (1996), The TOSCA Analysis System. In: Koster, C./Oltmans, E. (eds.), *Proceedings of the first AGFL Workshop.* Technical Report CSI-R9604, Computing Science Institute, University of Nijmegen, 181–191.

Abdullah, Hasan (1974), *The Morphology of Malay*. Kuala Lumpur: Dewan Behasa dan Pustaka.

Al-Sulaiti, Latifa/Atwell, Eric (2006), The Design of a Corpus of Contemporary Arabic. In: *International Journal of Corpus Linguistics* 11, 135–171.

Andersen, Gisle/Stenström, Anna-Brita (1996), COLT: A Progress Report. In: *ICAME Journal* 20, 133–136.

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

Asmah, Haji Omar (1993), *Nahu Melayu Mutakhir* (revised edition). Kuala Lumpur: Dewan Bahasa dan Pustaka.

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. In: *ICAME Journal* 7, 34–66.

Atwell, Eric (1987), A Parsing Expert System which Learns from Corpus Analysis. In: Meijs, Willem (ed.), *Corpus Linguistics and Beyond: Proceedings of the Seventh International Conference on English Language Research on Computerised Corpora*. Amsterdam: Rodopi, 227–235.

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 (1993), Corpus-based Statistical Modelling of English Grammar. In: Souter, Clive/Atwell, Eric (eds.), *Corpus-based Computational Linguistics*. Amsterdam: Rodopi, 195−215.

Atwell, Eric (1996), Comparative Evaluation of Grammatical Annotation Models. In: Sutcliffe/Koch/McElligott 1996, 25−46.

Atwell, Eric (2003), 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./Marlien, I. (eds.), *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*. ATALA, Volume 1, 27−32.

Atwell, Eric/Al-Sulaiti, Latifa/Al-Osaimi, Saleh/Abu Shawar, Bayan (2004), A Review of Arabic Corpus Analysis Tools. In: Bel, B./Marlien, I. (eds.), *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*. ATALA, Volume 2, 229−234.

Atwell, Eric/Demetriou, George/Hughes, John/Schriffin, Amanda/Souter, Clive/Wilcock, Sean (2000), A Comparative Evaluation of Modern English Corpus Grammatical Annotation Schemes. In: *ICAME Journal* 24, 7−23.

Atwell, Eric/Drakos, Nicos (1987), Pattern Recognition Applied to the Acquisition of a Grammatical Classification System from Unrestricted English Text. In: Maegaard, Bente (ed.), *Proceedings of EACL: Third Conference of the European Chapter of the Association for Computational Linguistics*. New Jersey: ACL, 56−62.

Atwell, Eric/Hughes, John/Souter, Clive (1994), AMALGAM: Automatic Mapping among Lexico-grammatical Annotation Models. In: Klavans, Judith/Resnik, Philip (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, NM, 11−20.

Baker, Paul/Hardie, Andrew/McEnery, Tony/Jayaram, Sri (2003), Constructing Corpora of South Asian Languages. In: Archer et al. 2003, 71−80.

Belmore, Nancy (1991), Tagging Brown with the LOB Tagging Suite. In: *ICAME Journal* 15, 63−86.

Benello, J./Mackie, A./Anderson, J. (1989), Syntactic Category Disambiguation with Neural Networks. In: *Computer Speech and Language* 3, 203−217.

Black, William/Neal, Philip (1996), Using ALICE to Analyse a Software Manual Corpus. In: Sutcliffe/Koch/McElligott 1996, 47−56.

Booth, Barbara (1985), Revising CLAWS. In: *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.

Brill, Eric (1995), Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. In: *Computational Linguistics* 21, 543−566.

Briscoe, Edward/Carroll, John (1993), Generalised Probabilistic LR Parsing of Natural Language (Corpora) with Unification-based Grammars. In: *Computational Linguistics* 19, 25−60.

Cure, The (1980), *A Forest*. Fiction Records.

Eeg-Olofsson, Mats (1991), *Word-class Tagging: Some Computational Tools*. PhD thesis. Department of Linguistics and Phonetics, University of Lund, Sweden.

Elliott, John/Atwell, Eric (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.

Fang, Alex (2005), Robust Practical Parsing of English with an Automatically Generated Grammar. PhD thesis, University College London.

Freeman, Andrew (2001), Brill's POS Tagger and a Morphology Parser for Arabic. In: *Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects*. Toulouse, France. Available at: http://www.elsnet.org/acl2001-arabic.html.

Garside, Roger (1996), The Robust Tagging of Unrestricted Text: The BNC Experience. In: Thomas, Jenny/Short, Mick (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman, 167−180.

Greenbaum, Sidney (1993), The Tagset for the International Corpus of English. In: Clive Souter/Atwell, Eric (eds.), *Corpus-based Computational Linguistics*. Amsterdam: Rodopi, 11−24.

Greene, Barbara/Rubin, Gerald (1981), *Automatic Grammatical Tagging of English*. Providence, RI: Department of Linguistics, Brown University.

Grefenstette, Gregory (1996), Using the SEXTANT Low-level Parser to Analyse a Software Manual Corpus. In: Sutcliffe/Koch/McElligott 1996, 139−158.

Hanks, Patrick (ed.) (1979), *Collins English Dictionary*. London and Glasgow: Collins.

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/Atwell, Eric (1994), The Automated Evaluation of Inferred Word Classifications. In: Cohn, Anthony (ed.), *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. Chichester: John Wiley, 535−539.

Hughes, John/Souter, Clive/Atwell, Eric (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, 10−17.

Johansson, Stig/Atwell, Eric/Garside, Roger/Leech, Geoffrey (1986), *The Tagged LOB Corpus: Users' Manual*. Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities.

Karlsson, Fred/Voutilainen, Atro/Heikkilä, Juha/Anttila, Arto (1995), *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text.* Berlin: Mouton de Gruyter.

Khoja, Shereen (2003), APT: An Automatic Arabic Part-of-speech Tagger. PhD thesis, Lancaster University.

Khoja, Shereen/Garside, Roger/Knowles, Gerry (2001), A Tagset for the Morphosyntactic Tagging of Arabic. In: Rayson, Paul/Wilson, Andrew/McEnery, Tony/Hardie, Andrew/Khoja, Shereen (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL Technical Paper 13, Lancaster University, 341.

Kilgarriff, Adam (2007), Message to *CORPORA@uib.no*, Discussion Forum on the History of Corpus Linguistics.

Knowles, Gerry/Don, Zuraidah Mohd (2003), Tagging a Corpus of Malay Texts, and Coping with "Syntactic Drift". In: Archer et al. 2003, 422−428.

Knowles, Gerry/Don, Zuraidah Mohd (2005), MALEX: Providing Linguistic Information and Tools for Automated Processing of Malay Texts. In: *Proceedings of O-COCOSDA-2005 International Conference on Speech Databases and Assessments*. Jakarta, Indonesia, 138−143.

Kytö, Merja/Voutilainen, Atro (1995), Applying the Constraint Grammar Parser of English to the Helsinki Corpus. In: *ICAME Journal* 19, 23−48.

Lecomte, Josette (1998), *Le categoriseur Brill14-JL5 / WinBrill-0.3*. Technical report, Institut National de la Langue Française.

Leech, Geoffrey/Barnett, Ros/Kahrel, Peter (1996), *EAGLES Final Report and Guidelines for the Syntactic Annotation of Corpora*. EAGLES Report EAG-TCWG-SASG/1.5. See also http://www.ilc.pi.cnr.it/EAGLES96/home.html.

Leech, Geoffrey/Garside, Roger/Atwell, Eric (1983), The Automatic Grammatical Tagging of the LOB Corpus. In: *ICAME Journal* 7, 13−33.

Lewis, M. (1947), *Teach Yourself Malay*. London: English Universities Press.

Lin, Dekang (1994), PRNCIPAR − an Efficient, Broad-coverage, Principle-based Parser. In: *Proceedings of COLING-94*. Kyoto, Japan, 482−488.

Lin, Dekang (1996), Using PRINCIPAR to Analyse a Software Manual Corpus. In: Sutcliffe/Koch/McElligott 1996, 103−118.

Lüdeling, Anke/Evert, Stefan (2003), Linguistic Experience and Productivity: Corpus Evidence for Fine-grained Distinctions. In: Archer et al. 2003, 475−483.

Madonna (1984), *Like a Virgin*. Sire Records.
Manning, Christopher/Schutze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
Marcus, Mitchell/Marcinkiewicz, Mary Ann/Santorini, Beatrice (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19, 313−330.
O'Donoghue, Timothy (1991), Taking a Parsed Corpus to the Cleaners: The EPOW Corpus. In: *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/Koch/McElligott 1996, 179−206.
Osborne, Miles (1996), Using the Robust Alvey Natural Language Toolkit to Analyse a Software Manual Corpus. In: Sutcliffe/Koch/McElligott 1996, 119−138.
Owen, Marion 1987. Evaluating Automatic Grammatical Tagging of Text. In: *ICAME Journal* 11, 18−26.
Qiao, Hong Liang/Huang, Renje (1998), Design and Implementation of AGTS Probabilistic Tagger. In: *ICAME Journal* 22, 23−48.
Roberts, Andrew/Al-Sulaiti, Latifa /Atwell, Eric (2006), aConCorde: Towards an Open-source, Extendable Concordancer for Arabic. In: *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.
Schiller, Anne/Teufel, Simone/Thielen, C. (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.
Schmidt, Ruth (1999), *Urdu: An Essential Grammar*. London: Routledge.
Sleator, Daniel/Temperley, Davy (1991), *Parsing English with a Link Grammar*. Technical Report CMU-CS-91−196, School of Computer Science, Carnegie Mellon University.
Sneddon, James (1996), *Indonesian: A Comprehensive Grammar*. London: Routledge.
Souter, Clive (1989a), The Communal Project: Extracting a Grammar from the Polytechnic of Wales Corpus. In: *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/Koch, Heinz-Detlev/McElligott, Annette (eds.) (1996), *Industrial Parsing of Software Manuals*. Amsterdam: Rodopi.
Sutcliffe, Richard/McElligott, Annette (1996), Using the Link Parser of Sleator and Temperley to Analyse a Software Manual Corpus. In: Sutcliffe/Koch/McElligott 1996, 89−102.
Taylor, Lolita/Knowles, Gerry (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.
Thielen, Christine/Schiller, Anne/Teufel, Simone/Stöckert, Christine (1999), *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Technical report, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart and Seminar für Sprachwissenschaft, University of Tübingen.
Treebank (1999), *The Penn Treebank Project website*. www.cis.upenn.edu/~treebank/home.html.

Eric Atwell, Leeds (UK)