

Refining Parts-of-speech in the Lexicon

Charles E. Grimes

*Australian National University
Pattimura University & Summer Institute of Linguistics*

1. Introduction

There is a story about a baseball umpire who was asked, "How do you know which ones are balls and which ones are strikes?" He replied thoughtfully, "Well, some balls are clearly strikes, some balls are balls, and some are nothing until I call them." In this paper I address similar problems in the categorization of parts-of-speech in the lexicon.¹

What I want to focus on here are some frequently observed discrepancies between principles of linguistics and the practice of indicating parts-of-speech in the lexicon.² The discussion is aimed particularly at linguists and lexicographers in the early stages of compiling a dictionary.³ While some of the specific issues addressed in this paper have not, to my knowledge, been addressed elsewhere, the general principles of determining parts-of-speech are not new, and are addressed to one degree or another in standard works on grammar or lexicography (Nida 1949, Zgusta 1971, Bartholomew & Schoenhals 1983, Givón 1984, Schachter 1985, Wierzbicka 1988). The issues highlighted in this paper are relevant on a broad scale, but particularly so for Austronesian languages of Asia and the Pacific.

I take it as a given that dictionaries are meant to be used and should therefore be user-centric (user-friendly) rather than compiler-centric. I also recognize that dictionaries can be aimed for different groups of users, such as international academic audiences or local audiences. The way in which information is packaged in a dictionary must be adapted to the specific audience, but such adaptation must not compromise an accurate representation of the language.⁴ Regardless of which group of users is in focus, the information in a lexical entry on parts-of-speech (also referred to as 'word class' or 'form class') should enable the uninitiated user of a dictionary to understand — and hopefully to use — the lexeme in its appropriate syntactic contexts in conjunction with the description of the grammar. *Parts-of-speech* in a lexicon is simply a tag that identifies the lexeme as a member of a *category* that shares a cluster of properties in its morphosyntactic network with other members of the same category. The parts-of-speech tag is a link between meaning and grammar.

¹ I am grateful to Adrian Clynes, Barbara Dix Grimes, and Darrell Tryon for their helpful comments on an earlier draft of this paper. The views expressed in this paper reflect my own current thinking and are not necessarily shared by them.

² The basis for my comments comes from compiling lexicons of the Buru and Tetun languages in eastern Indonesia, as well as from consulting on a wide variety of dictionary projects in Asia and the Pacific. The Buru lexicon currently has around 5,000 entries. The Tetun lexicon is still in beginning stages (a little over 300 entries), but highlights a different set of complications than those encountered in Buru.

³ Major dictionary projects of well-described languages occasionally also fall prey to the problems described here.

⁴ It is often not economically feasible to produce more than one version of a dictionary for different audiences. If only one dictionary can be produced, it is my feeling that it is better to add information for a scientific audience to a dictionary packaged for local users than the other way around. This is particularly true with languages that have any significant degree of derivational morphology.

The minimal information necessary to enable the dictionary user to make effective use of any part-of-speech category varies from language to language. The trouble is, however, very often such information is not in a dictionary, or it is misleading, or it is insufficient to be useful. We may get an approximation of the meaning, but the information on how the lexeme behaves, or how to use the lexeme is inadequate.

One could well ask whether we need parts-of-speech information in a dictionary at all, especially since such information seems to be of little interest or relevance to proficient speakers. We must recognise, however, that a dictionary is very often the first and most frequent resource a person (including linguists) consults when learning or studying a new language. It is for these 'outsiders' that information on parts-of-speech categories is most useful.

I also recognise as a secondary consideration the utility of the principle of transferability from one terminological system to another. For example, the cluster of properties for what is labelled as *noun* should overlap significantly with the cluster of properties that are generally labelled as *noun* cross-linguistically. For a group of local users, category labels should be adapted to the labels used for the national language, where the associations with those categories are not in conflict. This facilitates a transfer to and from dictionaries in other languages, such as the national language or an international language.

For bilingual dictionaries, the introduction must clarify categories to reflect the source language or the target language. This information is often missing.

2. Common principles behind determining parts-of-speech

Using traditional parts-of-speech categories, and using the terms commonly accepted in the nation or region in which the language is spoken is certainly a place to start, but is something that must not be simply assumed or blindly accepted. In determining or refining parts-of-speech categories there is a fairly broad acceptance of basic principles. Pinning linguistic labels on bits and pieces of a language is justifiable only where the structures of the language itself indicate contrastive patterns. A fundamental principle underlying all analysis is determining whether two things are considered the *same* or *different* within the scope under scrutiny. An operating assumption is that it is preferable in a dictionary to associate similar forms that share a common thread of meaning. Parts-of-speech categories for a language are generally determined by comparing and contrasting the following criteria:

- a) **Form:** In some cases the structural form of an entire form class distinguishes it from other form classes. In Buru, prefixes can be distinguished from proclitics on the basis of form — prefixes always take the shape eC-, while proclitics can take any V and are of the shape CV (Grimes 1991:60). Also in Buru, certain classes of functors may be monosyllabic, but classes using content words (e.g. nouns and verbs) are never monosyllabic.
- b) **Function:** When we talk about an entire form class, or the behavior of a single lexeme in the syntax we usually refer to its 'function', or its range of functions — what it does, or how it (and things like it) behaves in different contexts. For example, in many languages that have prepositions, the function of the class of prepositions is to relate non-core arguments to the verb and to identify the semantic role that argument is playing. The function of prepositions contrasts with the function of verbs.⁵ Schachter

⁵ Except, of course, where there are propositional verbs or serial verbs where a verb functions as a preposition might in another language.

(1985:4) observes the preference “that the assignment of parts-of-speech classes is based on properties that are grammatical rather than semantic.” Thus, defining nominals as the head of grammatical arguments in a clause is preferable to defining them as words that name persons, places, or things.⁶

- c) **Distribution:** The distributional behavior of a lexeme or a form class must also be taken into account. This includes the syntactic slot(s) it fills, as well as combinatory possibilities with affixes and with other form classes. In compiling a dictionary the well-attested phenomenon of *complementary distribution* is often overlooked in determining parts-of-speech categories relevant to a given language.

I refer to the combination of the above criteria as the ‘morphosyntactic network’ of a lexeme or a form class. In many languages assigning parts-of-speech to a lexeme is quite straightforward for the bulk of the lexicon — a noun is clearly a noun, a verb is a verb, and a preposition is a preposition. This paper focuses on situations where categorisation is not so straightforward.

3. Common areas of discrepancy between principle and practice

Problems with assigning parts-of-speech in the lexicon often occur when there are discrepancies between principles and actual practice in lexicography. Such discrepancies are often encountered in the following areas (which can be seen as a variety of ‘temptations’ in assigning parts-of-speech that lexicographers experience to one degree or another, and to which some succumb):

- 1) Lexicographers may assign parts-of-speech on the basis of the gloss in the national (or international) language, rather than on the syntactic behavior of the form class in the language itself. In Buru, for example, we might be tempted to call *saa* an ‘article’ because it most commonly translates into English with the indefinite article ‘a’. However, in exploring the whole morphosyntactic network it becomes clear that *saa* is a member of a closed class of what I call ‘deictics’ that share a variety of formal, functional, and distributional properties (Grimes 1991:167ff.).
- 2) Lexicographers tend to remain committed to the parts-of-speech labels that they first assigned to a lexeme in the early analysis of a language (with associated assumptions about the behavior of that part-of-speech), even after those labels are shown to be inappropriate. Ideas about parts-of-speech categories need to be refined and updated in the development of a lexicon to reflect developments in the understanding of the grammar.
- 3) Lexicographers generally assume ‘word class’ or ‘part-of-speech’ is inherent to the lexicon, and that every lexeme belongs fundamentally to a single part-of-speech category. Most lexicographers (and linguists) are not aware of operating on this assumption, but freely acknowledge it when it is brought to their attention. However, the empirical and theoretical basis for the assumption is problematic, and I discuss later the possibility that parts-of-speech for some parts of the lexicon may need to be defined syntactically, rather than lexically. After all, the whole notion of parts-of-speech is with reference to the syntax of a language.

⁶ This characterisation of a nominal is not tight enough for languages such as Tagalog, which can also use verbs as clausal arguments (see Schachter 1985:9).

- 4) When a lexeme can function in two or more classes (e.g. both nominally and verbally, or as a preposition and a conjunction), lexicographers tend to assume that it must be primarily one class, and only secondarily the other, assigning primacy on the basis of external (etic, rather than internal, emic) criteria. This is the 'flaw of the excluded middle'.
- 5) There is a tendency to assume certain word classes, such as 'adjective', are universal to all languages, and must therefore be in the language whose lexicon they are compiling.
- 6) Lexicographers often fail to distinguish verbal subcategories that are relevant to the language, often assuming the only relevant division for verbs in all languages is limited to 'transitive' or 'intransitive'.⁷ As described later in this paper, the fundamental division for some types of languages is more complex than a simple binary distinction.
- 6) Lexicographers often tag multiple pronominal sets with terminology that is not appropriate to the type of language, such as using case terms (e.g. nominative-accusative or ergative-absolutive) for split-S languages or for pragmatically driven systems such as switch-reference systems.⁸ In such languages labelling something as an 'ergative pronoun' or a 'nominative pronoun' reflects an inappropriate typology for the language.

4. Specific areas to watch out for

In the following sections I address various problem areas and suggest some ways in which parts-of-speech categories can more accurately reflect the language.

4.1 Views about the basis for assigning parts-of-speech

The traditional (and perhaps necessary) nature of a dictionary is as an artificial catalogue of the lexicon, presenting a serial list of lexemes isolated from natural speech and organised around principles of retrievability of information. That, together with ideas about what comprises a 'lexical entry' encourages linguists and lexicographers to slip into incorrect Aristotelian thinking that:

This lexeme cannot be both A and not A at the same time.

In other words, the thinking goes, this lexeme cannot be, for example, both a noun and a verb; therefore it must be primarily one and only secondarily the other (for example, through a zero-derivation),⁹ or they must be two different lexemes. However, the problem arises out of the artificial nature of the dictionary in trying to assign parts-of-speech to lexemes in isolation. It is not the case in normal speech that a lexeme is functioning as both a noun and a verb *at*

⁷ At a recent lecture I heard a world-renowned linguist reiterating the notion that "all languages divide verbs into two types: transitive and intransitive." This simplification encourages linguists and lexicographers to be blinded to what distinctions languages actually *are* making where the fundamental divisions are more complex, such as in split-S languages, and blinded to notions such as 'ambitransitive' and 'intra-directive'.

⁸ This fallacy was reinforced at another recent lecture by a well-known linguist with the statement "75% of the world's languages are nominative-accusative and 25% are ergative-absolutive." This characterisation blinds newcomers to well-documented language types such as split-S, active/non-active (stative-active), and Philippine-type languages which are numerically significant in the world's languages.

⁹ This view often surfaces at linguistic seminars in lively debate over whether lexeme X is primarily category A or category B - and the implications for syntactic arguments that follow from that. The linear nature of a dictionary forces sense A to precede sense B, and it is part of the conventional culture of dictionary users to *assume* that the sense presented first is more basic - and for other reasons this makes good lexicographic sense.

the same time. Where a lexeme is functioning in more than one category, it is either in different utterances, or in different syntactic slots within the same utterance. I explore below two areas in which the conflict commonly arises.

4.1.1 Are they adpositions or conjunctions?

A problem often occurs in assigning parts-of-speech to certain types of functors which operate in a variety of syntactic slots. For example, in English:

- | | | |
|-----|--------------------------------|--|
| (1) | He went to the store. | [relates verb to a non-core argument] |
| | He went to take a bath. | [relates verb to object complement purpose clause] |

These are commonly handled in English dictionaries as separate lexemes (homonyms), yet they share the 'meaning' of energy directed toward a goal — one locative and the other purpose (see also Wierzbicka 1988). They are relating different types of syntactic units, and with the similarity in meaning they could be analysed as the same lexeme with different functions in complementary distribution.

It is, in many cases, quite misleading to characterise functors of this sort merely as (or primarily as) a discourse particle, a clause-level conjunction, or an adposition (i.e. preposition or postposition). Many can function across a range of syntactic levels, linking constructions of varying scopes. The following contexts are from the Buru language (Grimes 1991:398).

- | | | |
|-----|---|--|
| (2) | PARAGRAPH ₁ . Petu PARAGRAPH ₂ | Linking paragraphs in a discourse |
| | SENTENCE ₁ . Petu SENTENCE ₂ | Linking sentences in a paragraph |
| | CLAUSE ₁ , petu CLAUSE ₂ | Linking clauses in a sentence (paratactic) |
| | Subject Verb petu CLAUSE ₂ | Subordinating a result clause (hypotactic) |
| (3) | Tu dii , DISCOURSE | 'At that time,...' Introduces (cataphorically) the time setting in a discourse |
| | SENTENCE ₁ . Tu SENTENCE ₂ | Linking sentences in a paragraph |
| | CLAUSE ₁ , tu CLAUSE ₂ | Linking clauses in a sentence (paratactic) |
| | [N tu N]subject, Predicate | Coordinating nouns in an NP |
| | S - V - (O) - tu NP | Preposition |

While some of these types of lexemes function exclusively as adpositions, some as conjunctions, and some as discourse particles, in a dictionary it is misleading to assign one of these classes to lexemes that can relate units of varying scopes. For this latter more flexible type, I prefer the broad term *relator*. This issue of scope should be addressed in the grammatical introduction to a dictionary, but rarely is.

4.1.2 Are they nouns or verbs?

In many languages a portion of the lexicon is inherently and unambiguously nominal, while another portion is unambiguously verbal. But in many languages there is also a portion of the lexicon that may be either, according to its distribution and function in an utterance, such as the following examples from English.¹⁰

¹⁰ The flexibility of this ambivalent portion of the lexicon may also vary between dialects of the same language. For example, Australian English can verbalise many words that are not able to be verbalised in American English. *She is flattening* (= *She is renting a flat/apartment*).

- | | | |
|-----|---|-----------------------|
| (4) | She is going to <i>sail</i> around the world.
He mended the <i>sail</i>. | [verbal]
[nominal] |
| (5) | He went to <i>photocopy</i> the manuscript.
He took the <i>photocopy</i> of the document away. | [verbal]
[nominal] |
| (6) | It looked like it was going to <i>rain</i>.
The <i>rain</i> got her wet. | [verbal]
[nominal] |
| (7) | He would like to <i>shower</i> under the tree.
The <i>shower</i> is no longer working. | [verbal]
[nominal] |
| (8) | They found it hard to <i>laugh</i>.
They had a <i>laugh</i>. | [verbal]
[nominal] |

Some lexicographers are tempted to argue etymologically for the primacy of membership in one form class over another, but unless there are clear synchronic derivational processes, the arguments may be much more difficult to substantiate and tend to appeal to elusive processes such as 'zero-derivation', which have no surface marking and which assume the primacy of one part-of-speech over another. Where 'zero-derivation' is warranted, there must be surface evidence somewhere in the morphosyntactic networks of the forms in question. Otherwise the claim of 'zero-derivation' is simply linguistic hocus-pocus.

Like English and other languages, Malay also has a number of lexemes whose function is distinguished only by its distribution within an utterance in an informal register,¹¹ such as:

- (9) **Orang-nya jalan di jalan situ.**
 person-ANAPH walk LOC path DISTAL.LOC
 'The person went along that path.'

Such lexemes of ambivalent category membership are handled in different dictionaries variously as 1) different lexemes (homonyms), 2) the same lexeme in different distributions, 3) a compromise where they are viewed as separate but related lexemes (i.e. partial homonymy), or 4) by avoiding addressing the parts-of-speech issue altogether. Any four-year-old speaker of Malay knows the two are related, and not just because they sound the same.

There is extensive discussion in the literature on Austronesian languages of the Philippines and Taiwan (Formosan languages) as to whether the verbal construction should be interpreted as primarily nominal or verbal (see, for example, the discussion in Starosta, Pawley, and Reid 1982, and Ross, forthcoming). The following derivations from the Paiwan root *kan* 'eat' are from Ferrell (1982:17,106), adapted from Ross (forthcoming).¹²

¹¹ Formal Malay would require derivational affixes such as *ber-jalan* for the verbal predicate use. One could argue on the basis of formal Malay that there is simply affix ellipsis for informal Malay. But this leaves at least two difficulties. First, what is the status of the unmarked base to which the verbal affixes (e.g. *ber-*, *meN--kan*) attach in the first place? Secondly, how can we argue for the ellision of affixes that simply are not used in these contexts in informal Malay?

¹² The symbol [ə] here represents schwa for typographical convenience.

(10)	Paiwan	Verbal construction	Nominal interpretation
	k < èm > an	Actor pivot (neutral)	'eater' or 'someone who eats'
	kan-èn	Undergoer pivot (neutral)	'food' or 'something to be eaten'
	k < in > an	Undergoer pivot (perfective)	'consumed food' or 'something eaten'
	kan-an	Locative pivot (neutral)	'place where one eats'
	si-kan	Instrumental pivot (neutral)	'eating utensil' or 'something to eat with'

The point is, the interpretation of these constructions as nominal or verbal depends largely on their distribution in the syntax.

There continues to be debate on the distinction between nouns and verbs in some Oceanic languages, such as Samoan (Mosel 1991, Vonen 1991), Tokelau (Vonen 1992), and Tongan (Taumoefolau 1991). One possibility that has entered the debate¹³ is to define nouns and verbs syntactically, rather than lexically, where the distinction is at issue. Unfortunately lexicographers do not tend to think in syntactically-oriented terms because of the lexically-oriented nature of dictionaries. But as Wierzbicka (MS:9) observes about definitions, "words don't have any meaning in isolation, but only in sentences." And Halliday (1961:261) notes, "a class is always defined with reference to the structure of the unit next above, and structure with reference to classes of the unit next below." In other words, word class — like meaning — is with reference to context.

4.1.3 Handling as homonymy versus polysemy

When a single form can function in more than one category without any explicit derivation, the lexicographer must decide whether to handle them as *homonymy* (same form but different meaning, therefore separate lexemes), or as *polysemy* (same form with range of related meanings, therefore subentries of the same lexeme).¹⁴

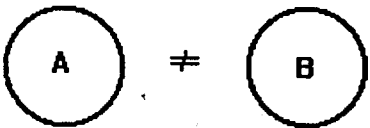
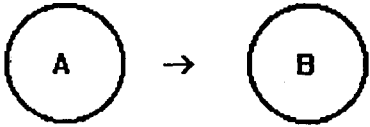
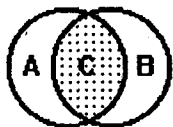
In one sense it is a moot point whether we should view the problem of lexemes like *sail* (*n*) and *sail* (*v*) as a zero derivation or as part of the lexicon whose form class membership is syntactically defined, if both views result in them being handled the same way in the dictionary — as subentries of a single entry.

However, if there is a distinction in the lexicon between, for example, category A — that part of the lexicon that is inherently nominal and must take verbal derivations to function verbally, category B — that part of the lexicon that is inherently verbal and must take nominal derivations to function nominally, and category C — that part of the lexicon that can function in either capacity with either no derivation or with either derivation, then we must indicate this latter portion of the lexicon as a different category.¹⁵

¹³ Not without objections.

¹⁴ Zgusta (1971:80-89) recognises a vague intermediate status which he calls "partial homonymy" and acknowledges some of the complexities of the issue.

¹⁵ I'm sure there are a variety of solutions for different types of languages and regions of the world. One possibility, as I suggested earlier, is to use a broader term, such as *relator*, where the membership is more flexible than strictly *preposition* or *conjunction*. Another possibility is to distinguish something like *Headword n* (= inherently nominal), from *Headword As n* (= flexible membership syntactically defined).

Categories				
1		meaning not related [distribution irrelevant]	two entries	
2		meaning inherently related [A derived as B]	{subentries of one entry}*	
3		meaning related; [a portion of the lexicon can function in either of two categories]	subentries of one entry	

*NOTE: For certain types of audiences (e.g. local audiences), and certain types of derivational morphology (such as prefixes), it may be preferable to handle derived forms as separate entries with cross references.

Figure 1: Homonymy vs. polysemy

4.1.4 Handling 'precategorials'

Many Austronesian languages have a number of lexical roots (content words) that are bound roots which never occur in an utterance without derivational morphology. For some bound roots, there is no internal evidence to say that one derived usage is more basic than another. Thus, one cannot, except by etic speculation, declare the root to be primarily 'nominal', 'verbal', or whatever. Such bound roots, with reference to their form class membership are sometimes called *precategorials*.¹⁶ For example in Buru:

- (11) **tea-** 'involving the planting of a post?'
- tea-k** 'to jam s.t. postlike into the ground for use'
- tea-n** 1) 'a (house)post'
- 2) 'point of reference for kin group origins (place of original post?)'
- ep-tea** 1) 'to live, stay, dwell (figurative from planting housepost?)'
- 2) 'to sit down (extended sense from 1?)'
- (12) **mae-** 'involving a rigid object graspable in one hand'
- mae-t** 'a fighting staff also used as a walking stick'
- mae-n** 'shaft (e.g. of spear); handle (e.g. of sword)'
- mae-k** 'to make a handle (e.g. of spear or sword)'

¹⁶ Adelaar (1985:223) defines precategorials for his study of Proto-Malayic as "roots that do not occur in isolation, that is, roots which only occur in derivations and in compounds." For Buru I expand the definition to include reduplication. E.g. *pani-n* 'wing', *p-e-pani* 'HAVE wing (s.t. of which wing is the most salient feature)', but never *[*pani-*] by itself. Some languages have a number of inherently reduplicated roots which never occur in the unreduplicated form. These roots could also be considered precategorials.

- (13) **bidu-** '(involving a cast-net)'
bidu-k 'to cast a cast-net'
bidu-t 'a cast-net'

In the last example above, one could argue either that 1) the nominal form uses /-t/ to derive the instrument that is characteristically used to perform the action of the verb, or 2) the verbal form uses /-k/ to derive a verb that is characteristically done using the noun. Both are legitimate explanations in the derivational paradigms of the language.

For an academic audience, precategorials can be handled as bound root lexemes (e.g. *mae-*, *tea-*, *bidu-*) with the surface derivations as subentries. But for a local audience this option is often not possible, since these roots do not constitute a minimal possible utterance. For such an audience, one can work with the community to choose one derivation as the citation form with the other forms as minor senses, or else list each surface derivation as a separate lexeme.¹⁷

4.2 Verbal subclasses

For some languages more information is required than simply tagging verbal lexemes as *vi* (verb-intransitive) or *vt* (verb-transitive).

4.2.1 Split-S languages

One type of language that requires a greater number of basic distinctions are split-S languages (Dixon 1979). To describe a split-S language it is helpful to first clarify some terms. *Subject* and *object* are grammatical roles that must be defined for each language (see Schachter 1976, 1977). I use the terms *Actor* and *Undergoer* in the sense of the semantic macroroles described by Foley & Van Valin (1984) which encompass a variety of case roles, with overlapping hierarchies of accessibility. The prototypical Actor is *agent*; the prototypical Undergoer is *patient*.

- (14) Subject as Actor (Foley & Van Valin 1984:30)
- | | |
|--|------------------|
| Colin killed the taipan. | [agent] |
| The rock shattered the mirror. | [instrument] |
| The lawyer received a telegram. | [recipient/goal] |
| The dog sensed the earthquake. | [experiencer] |
| The sun emits radiation. | [source] |
- (15) Object as Undergoer (Foley & Van Valin 1984:30)
- | | |
|---|------------------|
| Phil threw the ball to the umpire. | [them e/figure] |
| The avalanche crushed the cottage. | [patient] |
| The arrow hit the target. | [locative/goal] |
| The mugger robbed Fred of \$50.00. | [source] |
| The announcer presented Mary with the award. | [recipient/goal] |

The semantic macroroles (Actor and Undergoer) operate independently from the grammatical roles (subject and object), as can be seen in such constructions as the passive.

- (16) Subject as Undergoer (Foley & Van Valin 1984:31)
- The taipan was killed by Colin.**

¹⁷ The Buru community, at this stage, prefers each of these types of derived forms as a separate entry, with no precategorials having an entry by themselves.

The mirror was shattered by the rock.
The ball was thrown to the umpire by Phil.

Some languages give clues about the role structure interpretation of verbal arguments by the types of pronominal systems that collocate with the verbs or by the case marking systems on core arguments. Thus, when a language uses one set of pronouns (or a certain case marking for NPs) to encode both the subject of a transitive verb and the subject of an intransitive verb, but a different set (or case) to encode the object of a transitive verb, the language is characterised as *nominative-accusative*. A different language that encodes both the subject of intransitive verbs and the object of transitive verbs with one set of pronouns (or case marking), but the subject of transitive verbs with a different set of pronouns is characterised as *ergative-absolutive*. A third pattern may be found in languages in which the semantic Actor is encoded one way on both transitive and intransitive verbs and the semantic Undergoer is encoded a different way on both transitive and intransitive verbs. This pattern is called *split-S* (Dixon 1979).¹⁸ A very different pattern driven by pragmatic factors in discourse, rather than by semantics is a *switch-reference system*. In these systems one uses one set of pronouns if the referent is the same as in the preceding clause [SS = same subject/referent], and a different set if the referent is not a continuation from the one in the preceding clause [DS = different subject/referent]. In the figure below, circles represent a contrast between types of case marking in the pronominal systems or full NPs of a language.

	Nominative-accusative	Ergative-absolutive
intransitive		
transitive		
	Split-S	Switch-reference
intransitive		
transitive		

Figure 2: Some typologies for interpreting core arguments [A = Actor; U = Undergoer]

These typologies can be manifested separately in different subsystems in a language. For example, a language which reflects an ergative-absolutive orientation in its pronominal systems, but a nominative-accusative system in full NPs is called a *split-ergative* system. The relevant subsystems include:

- (17) a) NP marking
 b) NP order (in relation to verb)
 c) Pronominal marking
 d) Pronominal order (in relation to verb)
 e) Verbal morphosyntax

¹⁸ S in Dixon's system is the single argument of intransitive verbs. In a split-S system Actor and Undergoer are encoded differently on intransitive verbs - hence the name split-S. Dixon (1979) does not use Actor and Undergoer as primitives, but rather S, A, and O.

Following Dixon's (1979) system, **A** is the subject of a transitive verb; **O** is the object of a transitive verb; **S** is the subject of an intransitive verb. These are *grammatical* categories and are represented in CAPS. Subscript **A** for A(ctor) and **U** for U(ndergoer) are *semantic* macroroles. Active verbs are DO or CAUSE type verbs (e.g. *do, make, go, hit, kill, break, return*); Non-active verbs are BE or BECOME type verbs (e.g. *dark, ripe, white, sick, hungry, big, small, die, good, bad*). If the different typologies are illustrated using the two English pronoun sets, they would operate something like the following:

- (18) Nominative-accusative system (S patterns with A)
- | | | | | | | |
|----------------------------|---------------------------|---|---|---|---|--|
| <i>he</i> hit <i>him</i> . | [Active transitive; | A | V | O |] | |
| <i>he</i> ran. | [Active intransitive; | S | V | |] | |
| <i>he</i> is sick. | [Non-active intransitive; | S | V | |] | |
- (19) Ergative-absolutive system (S patterns with O)
- | | | | | | | |
|----------------------------|---------------------------|---|---|---|---|----|
| <i>he</i> hit <i>him</i> . | [Active transitive; | A | V | O |] | |
| <i>him</i> ran. | [Active intransitive; | | V | S |] | |
| <i>him</i> is sick. | [Non-active intransitive; | | V | S |] | or |
| <i>he</i> is sick. | [| S | V | |] | |
| <i>he</i> ran. | [| S | V | |] | |
- (20) Split-S system (S_A patterns with A; S_U patterns with O)¹⁹
- | | | | | | | |
|----------------------------|----------------------|-------|---|-------|---|----|
| <i>he</i> hit <i>him</i> . | [Active transitive | A | V | O |] | |
| <i>he</i> ran. | [Active intransitive | S_A | V | |] | |
| <i>him</i> is sick. | [Non-active; | | V | S_U |] | or |
| <i>he</i> is sick. | [| S_U | V | |] | |
- (21) Switch-reference (SS = same subject; DS = different subject)
- | | |
|---|-----------------------|
| <i>he_{SS}/him_{DS}</i> hit <i>him_{DS}</i> . | [Active transitive] |
| <i>he_{SS}/him_{DS}</i> ran. | [Active intransitive] |
| <i>he_{SS}/him_{DS}</i> is sick. | [Non-active] |

Split-S systems are fairly widespread within the Austronesian world; for example in Acch, North Sumatra (Durie 1985), and in many languages in eastern Indonesia (e.g. Buru, Grimes 1991). Selaru, a language of southern Tanimbar (Coward 1990) represents a classic split-S orientation, and is representative of many Austronesian languages in eastern Indonesia.

Verbal semantics:	Active / non-active ²⁰
Active transitive:	(A) A-V O [actor-indexed]
Active intransitive:	(SA) A-V [actor-indexed]
Non-active:	(SU) V-U [undergoer-indexed]
Nominal predicate:	Predicate-Subject [parallels Undergoer patterns]

Figure 3: Profile of Selaru

¹⁹ Relational grammarians call the S_A type verbs unergative and the S_U type verbs unaccusative. While there is nothing wrong with the terms for linguistic purposes, I do not recommend using these labels as parts-of-speech categories in a published dictionary as they severely limit the audience of effective users.

²⁰ The Selaru data and primary analysis are from Coward (1990). Some of the terminology and split-S framework reflect my adaptation of Coward's material. I avoid the label 'stative-active' that is more widespread in the general literature for these types of languages, because the 'non-active' verbs are typically ambiguous in their internal aspectual interpretation as imperfective (process) or perfective (state). The label 'stative' is thus highly misleading (see discussion in Grimes 1991:93-108).

	FREE PRONOUN	UNDERG. ENCLIT.	ACTOR PREFIXES ²¹	GENITIVE ENCLITICS	GENERAL POSSESSION	EDIBLE POSSESSION
1s	yaw	yaw	kU-	-kU	wasi-kw	hina-kw
2s	oa	-o	mU-	-mU	wasi-mw	hina-mw
3s	ia	-i	I-	-na	wasi	hina
INAN	θ	θ	kI-	θ	wasi	hina
1pe	aramy	aramy	aramI-	ara...-mi	ara wasi-my	ara hina-my
1pi	ity	ity	ta-	ity...-t	ity wai-t	ity hina-t
2p	ea	-e	mI-	-my	wasi-my	hina-my
3p	sira	sir	ra-	(sir)...-t	(sir) wai-t	(sir) hina-t

Figure 4: Pronominal systems in Selaru

Selaru Actor prefixes are phonologically conditioned by the root or stem to which they attach.

Stem begins w/		1s	2s	3s	3sINAN	1pi	1pe	2p	3p
SET 1	-V	k-	m-	y-	ky-	t-	aramy-	my-	r-
SET 2	-CC	ku-	mu-	i-	ki-	ta-	arami-	mi-	ra-
SET 3	-C	kC ^w	mc ^w	C ^y	kC ^y	t-	aramy-	mC ^y	r-
Abstract form		kU-	mU-	I-	kI-	ta-	aramI-	mI-	ra-

Figure 5: Selaru Actor prefixes

(22) Active transitive

Naman-ke *y- oban* **asw-re.** [AVO; actor-indexed verb]
 child-ART 3sA-hit dog-PL
 'The child hit the dogs.'

(23) **Sawa-na** *y- al* **turi -ke** **ti** **naman-desy-kre.**
 wife-3sGEN 3sA-transfer machete-ART DAT child-DIST-ART.PL
 'His wife gave the machete to those children.'

(24) Active intransitive

Ama -ku *i- ris.* [SV; actor-indexed verb]
 father-1sGEN 3sA-bathe
 'My father is bathing.'

(25) **kb^wa.** **mb^wa.** **b^ya.**
 kU- ba mU- ba I- ba
 1sA-go 2sA-go 3sA-go
 'I'm going.' 'You're going.' 'She's going.'

²¹ See following chart for phonological conditioning of these prefixes.

- (26) **Aro-ke** **kbya** **de.**
 aro-ke kI-ba de
 boat-ART INAN-go PERF
 'The boat is already gone.'
- (27) Non-active
- Hahkye** **lan i.** [VS; undergoer-indexed verb]
 hahy-ke lan i
 pig -ART big 3sU
 'The pig is big.'
- (28) **Bat-batak-ke** **lanθ .**
 RED-chest-ART big INAN
 'The chest is big.'
- (29) **Atiat i.** **Bob atiat i.**
 bad 3sU (name)bad 3sU
 'He's bad.' 'Bob is bad.'
- (30) Nominal predicate [parallels Undergoer-oriented constructions]
- Guru** **yaw.**
 teacher[Skt] 1s
 'I am a teacher.'

Other languages in the region such as Dobel (Aru Islands - J. Hughes, MS) are very similar to Sclaru in their basic orientation. Buru is split-S in its verbal semantics, but shows an incipient switch-reference system in its pronominal typology (Grimes 1991). All split-S languages must minimally distinguish *three* types of verbs in the lexicon, not just two, but dictionaries and word lists published over the last century for split-S languages in eastern Indonesia have failed to do so. For Buru I abbreviate the three types as *vt* (active transitive), *vi* (active intransitive), and *vn* (non-active verbs).

An additional verb type shows up in many Austronesian languages in eastern Indonesia and out into the Pacific. Active intransitive verbs tend to be verbs of motion or posture, such as *go*, *return*, *stand*, *sit*, in which the person doing the action is also the one undergoing the action (their location or position is changed). For example, in *I go*, I am volitionally doing something, which results in my location being changed. There is only *one* semantic referent, but some languages mark some (or all) active intransitive verbs of this sort as *morphologically transitive*. In some literature on Oceanic languages these are referred to as *intradirective*, or *reflexive* verbs (see Pawley 1973), and in other areas of the world as *quasi-reflexive* verbs.

- (31) South Nuaulu (South-central Seram - R. Bolton 1990)

Ia **pina** **ona-te** **i- sipu -i,** **i -eu-i** **ria** **manahane.** [S_A V_{-II}]
 3s female big-NOM 3s-descend-3sU 3s-go-3sU inland outside
 'The old woman got down and went outside.'

(32) Buru (archaic)

Kae oli -m beka. [S V-U]
 2sA return-2sU first
'You should go home now.'

If these types of verbs contrast in a language with other active intransitive verbs that cannot be morphologically transitive, they must be indicated as a separate category in the lexicon.

4.2.2 Handling morphologically defined subclasses

It is still not always sufficient to identify a part-of-speech as, for example, *vt*, *vi*, or *vn*. Sometimes there are morphologically motivated subclasses within each category. For example in Buru, with a non-active verb (*vn*) we must also know whether it is an **em-** verb, an **eb-** verb, or a **-t** verb to know how it behaves in its morphosyntactic network.²² Thus *vn-t* is minimal part-of-speech information for a non-active **-t** verb. Similarly Buru active transitive verbs must distinguish whether they are **-k** verbs or **-h** verbs to know how they indicate pronominalised singular objects. Thus, *vt-h* is minimal information for a transitive **-h** verb in the lexicon. (See Grimes 1991:93ff.).

4.2.3 Pragmatically motivated variants

There are some languages which have a clear morphologically motivated distinction between transitive and intransitive verbs or usages, such as Fijian (Dixon 1988). There are other languages which have a group of verbs that are clearly and exclusively transitive, another group that are clearly intransitive, and a portion that may function in either capacity with no morphological distinction. For example, the Buru data parallel the English:

(33) Da ba kaa mangkau.
 3s DUR eat cassava
'He is eating cassava.'

(34) Da ba kaa-h.
 3s DUR eat-it
'He is eating it.'

(35) Da ba kaa.
 3s DUR eat
'He is eating.'

The above pattern of reducing the referential prominence of an argument through pronominalisation and omission is a common strategy in discourse for languages that allow it (see Givón 1990). It is pragmatically motivated. The referential prominence of the object in example (35) above is completely reduced through omission. Constructions like example (35) occur commonly as predicate-focus constructions (action-prominence) where the referential identity of the object is unimportant or irrelevant, as in the situation: *Q: What is he doing now? A: He is eating [so don't bother him].* But there is no morphological difference between the

²² Numerically indicated subclasses (e.g. Class I, Class II, Class III, etc.) seem to be very frustrating to everybody except the linguist who assigned those labels. An alternative such as the actual affixes that distinguish the subclasses in defined contexts broadens the audience of potential users (e.g. **em-** verbs, **eb-** verbs, etc.). This kind of morphological subclass is conventionalised in Spanish dictionaries in the citation form of verbs as **-ar**, **-ir**, or **-er** verbs.

transitive use and the intransitive use of *kaa* 'eat', other than the presence or absence of the object.

Some linguists, perhaps motivated by the view that parts-of-speech is always and exclusively inherent to the lexicon, ignore syntactic and pragmatic issues, preferring to say there are two lexemes *eat*₁ *vt*, and *eat*₂ *vi*. I find this approach of (partial) homonymy highly unsatisfying and prefer to say verbs like *eat* are included in that portion of the lexicon that is ambitransitive according to pragmatic issues. The portion of *ambitransitive* verbs in English is fairly restricted, but in Buru most verbs that can take a syntactic object without morphological derivation are ambitransitive.

In languages where a portion of the vocabulary is ambitransitive in contrast with a portion that is obligatorily transitive, this contrast must be noted in the dictionary as a separate category.

4.3 Adjectives (versus nouns or verbs)

Some languages clearly have 'adjective' as a distinct part-of-speech, expressing such things as dimension, physical property, colour, human propensity, age, value and speed. In some languages attributive modifiers in a noun phrase [NP] pattern closely with nouns, in other languages with verbs, and in others as a mixture (see Dixon 1982, Schachter 1985, Wierzbicka 1986 and forthcoming). Buru, like many Austronesian languages, has no canonical (underived) class of 'adjective' — all attributive modifiers in an NP are derived from verb roots (both active and non-active).²³ For a few Austronesian languages in eastern Indonesia there does seem to be a closed class of a handful of underived adjectives (often in the form of inherently reduplicated roots), with the bulk of attributive modifiers in NPs being derived from verbs. With the exception of these true 'adjectives', there is often no morphological distinction between predicative and attributive uses of verbs.

(36) Buru

Huma di em- kele. [predicative]
house DIST STAT-tall
'That house is high.'

Da puna huma em- kele. [attributive]
3s do [house STAT-tall]_{NP}
'He made a pile house.' [Lit. 'a tall house']

Where there is a morphological distinction between predicative and attributive uses, it is clear that the attributive (i.e. adjectival) use is derived from the predicative use, not the other way around.

(37) **Da ba haa hede.** [predicative]
3s DUR big still
'He is still growing.'

Da puna huma haa-t. [attributive]
3s make [house big-NOM]_{NP}
'He is making a big house.'

²³ Nouns can also modify other nouns in Buru NPs, but behave quite differently from verb-derived modifiers in their morphosyntactic networks (Grimes 1991:178ff.).

(38) **Kau di beha.** [predicative]
 wood DIST heavy
 'That wood is heavy.'

Da wada kau beha-t. [attributive]
 3s shoulder.carry [wood heavy-NOM]_{NP}
 'He is carrying heavy wood (on his shoulder).'

(39) **Feten boti mohede.** [predicative]
 millet white not.yet
 'The millet isn't yet ripe.'

Da ego labu -n boti -t. [attributive]
 3s get [shirt-GEN white-NOM]_{NP}
 'She took the white shirt.'

To label what translates into an English adjective as 'adjective' for these languages fails to recognise the behavior of the lexemes as verbs in the greater morphosyntactic networks of the language.

4.4 Related issues

Finally, there are a few issues related to handling parts-of-speech which can be handled in separate fields in a computerised lexicon.

4.4.1 Checking paradigms

Some languages have obligatory indexing on the verb for one or more core arguments. For many Austronesian languages in the string of islands east of Bali, consonant-initial verb roots are not inflected for person and number of the subject, but vowel-initial roots are. In some languages, however, the paradigms are not complete for all possible combinations (also noted generally by Zgusta 1971:122). For example in Tetun (central and east Timor - Therik and Grimes, MS) some verbs take the complete paradigm and others are only partial — the citation form of these verbs is the *h*-form. The completeness of the paradigm can also vary across dialects.

	Complete Paradigms				Incomplete Paradigms			
PERSON	eat	make	bring	kill	look	remain	wait	pass by
1s	kaa	kalo	kodi	ko'o	karee	kela	—	—
2s	maa	malo	modi	mo'o	maree	mela	mein	mosi
3s	naa	nalo	nodi	no'o	naree	nela	nein	nosi
1pe	haa	halo	hodi	ho'o	haree	hela	hein	hosi
1pi	haa	halo	hodi	ho'o	haree	hela	hein	hosi
2p	haa	halo	hodi	ho'o	haree	hela	hein	hosi
3p	raa	ralo	rodi	ro'o	—	—	—	—

Figure 6: Completeness of verb paradigms in Tetun

Where there is inconsistency in the completeness of paradigms it is not economical to indicate the complete paradigm for every verb, but only for those that deviate from a norm.

4.4.2 Indicating semantic domain

For some parts of a language, indicating a semantic class may provide more grammatical information than simply indicating the part-of-speech.²⁴ For example, in the Buru lexicon certain generalisations become available by knowing a verb is a cutting verb:

- | | | |
|------|--------------------------------|--|
| (40) | \le hete | [\le = lexeme] |
| | \ps vt-k | [\ps = part-of-speech] |
| | \ge cut into sections for use; | [\ge = gloss (English) ²⁵] |
| | \cl Vcut | [\cl = semantic class] |

This information tells us (Grimes 1991) that this entry shares with other cutting verbs the basic structure:

Subject:Actor:agent — DO:cut — (Object:Undergoer:patient) -- (tu instrument)

What distinguishes one cutting verb from another tends to be differences in manner, typical instrument, typical object, and occasionally typical agent. A carrying verb looks something like the following:

- | | |
|------|---------------------------------------|
| (41) | \le leba |
| | \ps vt-h |
| | \ge carry on the shoulder with a pole |
| | \cl Vcarry |

It shares with other carrying verbs the following general structure:

Subject:Actor:agent - DO:carry - (Object:Undergoer:figure)²⁶ - (tu instrument)
— (fi di locative source) — (gam di locative goal)

Similarly, identifying the semantic class of certain types of nouns tells us (again from the grammar description) how this lexeme should behave in certain constructions, (such as in this example, in the vocative).

- | | |
|------|------------|
| (42) | \le ama |
| | \ps n |
| | \ge father |
| | \cl Nkin |

5. Summary

Indicating parts-of-speech in the lexicon has been traditionally useful. The task of doing so is often straightforward and uncomplicated, but as this paper has shown there are many potential pitfalls. The lexicographer must continually refine notions about parts-of-speech categories in a language and update the lexicon as understanding of the grammar increases. Parts-of-speech categories should be adequately defined to fit the language and to make the dictionary a useful and productive tool.

²⁴ In computerised databases this also facilitates extracting, cross-referencing, and studying groups of related words.

²⁵ This is distinguished from \gi (Indonesian) and \ga (Ambonese Malay) in the multilingual database.

²⁶ 'Figure' is the object whose location is in question. Foley and Van Valin (1984) use the term 'theme'. With carrying verbs only one oblique argument is normally expressed - the one most salient to the discourse.

REFERENCES

- Adelaar, A.K. 1985. Proto-Malayic: the reconstruction of its phonology and parts of its lexicon and morphology. PhD dissertation. Rijksuniversiteit te Leiden. (Published 1992 in *Pacific Linguistics*.)
- Bartholomew, Doris A. and Louise C. Schoenhals. 1983. *Bilingual dictionaries for indigenous languages*. Mexico D.F.: Summer Institute of Linguistics.
- Bolton, Rosemary. 1990. A preliminary description of Nuaulu phonology and grammar. MA thesis. University of Texas at Arlington.
- Coward, David F. 1990. An introduction to the grammar of Selaru. MA thesis. University of Texas at Arlington.
- Dixon, R.M.W. 1979. 'Ergativity'. *Language*, 55:59-138.
- . 1982. *Where have all the adjectives gone? and other essays in semantics and syntax*. Amsterdam: Mouton.
- . 1988. *A grammar of Boumaa Fijian*. University of Chicago Press.
- Durie, Mark. 1985. *A grammar of Achehnese: on the basis of a dialect of north Aceh*. Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde 112. Cinnaminson, N.J.: Foris Publications.
- Ferrell, Raleigh. 1982. *Paiwan Dictionary*. Pacific Linguistics, C-73.
- Foley, William A. and Robert D. Van Valin Jr. 1984. *Functional syntax and universal grammar*. Cambridge Studies in Linguistics 38. Cambridge: Cambridge University Press.
- Givón, Talmy. 1984. *Syntax: a functional- typological introduction*, Vol. 1. Amsterdam: John Benjamins.
- . 1990. *Syntax: a functional-typological introduction*, Vol. 2. Amsterdam: John Benjamins.
- Grimes, Charles E. 1991. The Buru language of eastern Indonesia. PhD dissertation. Australian National University.
- Halliday, M.A.K. 1961. Categories of the theory of grammar. *Word*, 17:241-292.
- Hughes, Jock. MS [1991]. *Dobel, a language of the Aru Islands*. Ambon: Pattimura University and Summer Institute of Linguistics.
- Mosel, Ulrike. 1991. Markedness theory and the distinction of major word classes in Samoan. Seminar presented at the Australian National University.
- Nida, Eugene. 1949. *Morphology*. Ann Arbor: University of Michigan Press.

- Pawley, Andrew K. 1973. Some problems in Proto- Oceanic grammar. *Oceanic Linguistics*, 12/1-2:103-188.
- Ross, Malcolm D. (forthcoming). Reconstructing Proto Austronesian verbal morphology: evidence from Taiwan. Paper to be presented at International Symposium on Austronesian Studies relating to Taiwan (December 1992).
- Schachter, Paul. 1976. The subject in Philippine languages: topic, actor, actor-topic or none of the above? In *Subject and Topic*. Pp. 491-518. Edited by Charles Li. New York.
- . 1977. Reference-related and role-related properties of subjects. In *Syntax and semantics 8: grammatical relations*. Pp. 279-306. Edited by Cole and Sadock. New York.
- . 1985. Part-of-speech systems. In *Language typology and syntactic description I: clause structure*. Edited by Timmothy Shopen. Cambridge: Cambridge University Press. Pp. 3-61.
- Starosta, Stanley, Andrew K. Pawley, and Lawrence A. Reid. 1982. The evolution of focus in Austronesian. *Pacific Linguistics* C-75:145-170.
- Taumoefolau, Melenaita. 1991. Verbal senses of concrete nouns in Tongan. Paper presented at the Sixth International Conference on Austronesian Linguistics, Honolulu, Hawaii.
- Therik, Tom, and Charles E. Grimes. MS [1992]. Baria Ulu: a Tetun text. Canberra: Australian National University.
- Vonen, Arnfinn M. 1991. Hunting for nouns and verbs in Samoan. Seminar presented at the Australian National University, 22 November 1991.
- . 1992. Nominalisations in Tokelau. Seminar presented at the Australian National University, 15 May 1992.
- Wierzbicka, Anna. 1986. What's in a noun? (Or: How do nouns differ in meaning from adjectives?) *Studies in Language*, 10(2):353-389.
- . 1988. The semantics of grammar. Amsterdam: John Benjamins. *Studies in Language*, Companion Series 18.
- . (forthcoming). Adjectives vs. verbs: the iconicity of part-of-speech membership. In *Proceedings of a symposium on iconicity*. Edited by M. Landsberg. Zagreb.
- . MS [1992]. Back to definitions: cognition, semantics, and lexicography. Typescript, Australian National University.
- Zgusta, Ladislav. 1971. *Manual of lexicography*. The Hague: Mouton.