Structural and semantic dependencies in word class

William A. Foley

Tuesday 31st January, 2017

William A. Foley

Is there a dependency between the type of phrase structure that a language has and its inventory of lexical classes? This chapter will argue that there may well be, although not one that is strictly determinative. The claim is that a certain phrase structure pattern, i.e. left-headed and with overt functional categorial heads like determiners and tense-aspect-mood markers, is correlated with an attenuated distinction between nouns and verbs. A striking fact about the languages of the world is a widespread asymmetry between nouns and verbs. It is a salient but remarkable observation that languages often have many more monomorphemic nouns than they do verbs: in Yimas, a language of New Guinea, for instance, there are over 3000 noun roots, but only around 100 verb roots, a skewing commonly found in other languages of the region (**Pawley1994**). Even in languages with large inventories of both classes of words, such as English, there is a marked differential in behavior. Basic nouns in English typically have fewer meanings and usages than verbs. The *Webster's New World College Dictionary* (**Webster2009**), for example, lists seven meanings for the noun *chair*, but no less than seventy for the verb *take*. Furthermore, while the noun *chair* can be used in extended meanings such as *chair a meeting*, there is a clear semantic link between such uses and its basic noun meaning, while with verbs this is commonly not the case; what does *take* contribute when we contrast *to nap* with *to take a nap*?

In most language families around the world the predisposition to distinguish nouns and verbs is strong, and the distinction remains diachronically robust. But in a few, it is a family wide fact that the distinction is not so clear, and very many or most lexemes are flexible, i.e. can be freely used without clear derivational morphology either as a noun or verb. This does not mean a distinction between noun and verbs cannot be recognized; in some languages it may, and perhaps in others, it shouldn't, but that is not my concern here. I am strictly concerned with the fact and prevalence of such flexibility and the attenuation of a sharp distinction between them. This will be looked at briefly in this chapter in the Austronesian and Salish language families, for which the status of the noun-verb distinction have long been controversial, and mainly concentrating on the former. What is it about the grammatical organization of Austronesian and Salish languages that leads to a recurring predilection to attenuate the noun-verb contrast? And as this attenuation is relatively rare crosslinguistically, what happens to this structural trait when languages bearing it come into contact with languages with the much more common property of a sharp noun-verb contrast and a different type of phrase structure? This question will be briefly looked at in areas of heavy Austronesian-Papuan language contact in the New Guinea

region. Papuan languages across diverse language families exhibit sharp noun-verb distinctions, even sharper than classical languages like Latin, the source of our descriptive grammatical tradition. But it does seem that in Austronesian-Papuan contact situations, the selective pressures of areal features do outweigh inheritance.

Our earliest, more sophisticated grammatical treatment of word classes goes back to the first century BC grammar of Greek by the Alexandrian grammarian Dionysius Thrax, building on the work of Aristotle and the Stoic philosophers before him. He defined the categories of noun and verb and their distinction in the following terms:

- *Ónoma* (noun): a part of speech inflected for case, signifying a person or thing

- *Rhêma* (verb): a part of speech without case inflection, but inflected for tense, person and number, signifying an activity or process performed or undergone

Thrax's definitions are notable for two reasons, and both of these have influenced descriptive grammatical traditions ever since. Note that neither relies on a single criterion, both invoke two, one semantic and the other morphosyntactic. In the hothouse multicultural and multilingual atmosphere of Hellenistic Alexandria, Thrax would have been well aware of the wide differences in grammatical organization across languages, so he knew that a straightforward definition of word classes in semantic terms would not do, as items with very similar meanings could behave very differently in different languages and hence belong to different word classes. Yet he didn't abandon semantic criteria entirely, as he was also aware of the semantic commonality of core members of each word class and the use of this as a heuristic in a first pass at identifying members of a given word class. Still, the semantic criterion on its own wouldn't do, not only because of crosslinguistic differences, but also because the match between the typical meaning of a word class and the meanings of its individual members even in a language like Greek wasn't perfect; there were simply too many exceptions to what would be expected. So he dragged morphosyntactic behavior into use for delineating word class differences, for example, case for nouns and tense for verbs.

In his two pronged approach, Thrax was greatly aided by the grammatical structure of the classical languages; his description was based on Greek. In these languages, the distinction between noun and verb is over-determined; it is virtually impossible to miss it. Consider Table 1 a map of lexical organization in terms of word class membership in Latin.

Table 1: Lexical Organization in Latin

(i)	phonology	-*a* first declension	-*ē* second conjugation
		↑	↑
(ii)	inflection	DECLENSIONS	CONJUGATIONS
		↑	↑
(iii)	syntax	N + CASE	V + TENSE
		↑	↑
(iv)	semantics	ARGUMENT (thing) +	PREDICATE (event)

The reason, for instance, that the noun-verb distinction was so salient to the Ancient Greek and Latin grammarians is the sharp differentiation in morphological behavior between them in these two languages. Not only do Ancient Greek and Latin have distinct grammatical categories for nouns and verbs due to their syntactic properties (level iii), e.g. case for nouns and tense for verbs, but in addition different noun and verb lexemes belong to distinct inflectional patterns (level ii), declensions for nouns and conjugations for verbs, and these in turn correlate to clear phonological contrasts in their forms (level i) (nouns belong to five phonologically contrastive declensions and verbs to four conjugations). There is overkill in the distinctiveness of these two classes in these languages; grammarians could not fail to notice it. Ancient Greek does have word types that blend the morphosyntactic properties of nouns and verbs, such as participles, gerunds and infinitives, but these are clearly derived secondary forms and do not eclipse the very salient noun-verb distinction in these languages.

The classical languages with their robust distinction of word classes have provided a largely taken for granted model for thinking about lexical distinctions ever since. Classical languages have provided us with categories of nouns, verbs and adjectives, and linguists have mainly approached language descriptions with these categories in mind (although adjectives have been more controversial) by trying to find analogs of these classical categories in the languages under description, in spite of often very different syntactic properties and inflectional categories. It is almost as if, as **Riemer2010** points out, that knowing baseball and its terminology well, like first base, shortstop, home run etc., we use these familiar categories to describe all ball games: football, volleyball, tennis, basketball, etc. The real question is how much communality there is across languages that permits us to believe that we are talking about the same or even similar categories. In some languages, rather than pervasive difference as attested in Latin and An-

cient Greek, what we find is pervasive similarity in the grammatical behavior of lexemes which are prototypically divided into these two word classes, noun and verbs, those which denote objects and those which denote events respectively. Nouns function as arguments, and verbs as predicates. St'át'imcets, a Salish language of British Columbia, is one such language, and as such, is typical of its language family and indeed the languages of its region (**DemirdacheEtAl1994**):

    use as a verb/predicate qwatsáts-kacw
leave-2.ɴᴏᴍ
event 'you leave/left' smúlhats-kacw
woman-2.ɴᴏᴍ
object 'you are a woman'
    use as noun/argument: qwatsáts-∅ ti smúlhats-a
leave-3. D woman-D
object 'the woman left' smúlhats-∅ ti qwatsáts-a
woman-3ꜱɢ.ᴀʙꜱ D leave-D
event 'the leaver (one who left) is a woman'

When used as verbs, roots occur clause initially and are cliticized by a set of subject and object marking pronominals, here *–(lh) kacw* for 2SG.NOM (St'át'imcets is morphologically split ergative, so first and second person pronouns are inflected on a nominative-accusative basis, while third person exhibits an ergative absolutive pattern; third person absolutive is realized by zero, but the ergative form is *–ás*). When used as nouns, the same lexemes occur in a DP headed by the determiner *ti -a*; arguments are typically realized as DPs in Salish languages. Even more striking is that lexemes of both semantic types can co-occur with such prototypical markers of verbs (going right back to Thrax's definition more than two thousand years ago) as tense clitics like *tu7* PAST and *kelh* FUT:

    event
qwatsáts-∅ tu7 kw-s Gertie
leave-3. ᴘᴀꜱᴛ D-ɴᴏᴍ ᴘɴ
'Gertie left'
    qwatsáts-∅ kelh kw-s Gertie
leave-3. D-ɴᴏᴍ ᴘɴ
'Gertie will leave'
    object
plísmen tu7 kw-s Bill
policeman ᴘᴀꜱᴛ D-ɴᴏᴍ ᴘɴ
'Bill was a policeman'
    plísmen kelh kw-s Bill

policeman D-ɴᴏᴍ ᴘɴ
'Bill will be a policeman'

I am not claiming that no noun-verb distinction can be found in S'át'imcets and other Salish languages. That depends on wider empirical findings and how one weighs conflicting evidence, and I do not regard this question as settled yet (see the discussion in **Beck2002 Beck2002 DavisEtAl1999 DavisEtAl1999 DemirdacheEtAl1994 DemirdacheEtAl1994 JelinekEtAl1994 JelinekEtAl1994 Kinkade1983 Kinkade1983 Kuipers1968 Kuipers1968 vanEijkEtAl1986 vanEijkEtAl1986**). What I am simply doing here is exemplifying the pattern of flexibility in the language, and further pointing out that in the survey reported below, over 90% of all its roots exhibited flexibility.

Pretty much the same pattern is found in many Austronesian languages and is also widespread across this vast family, although the rate is variable, as will be reported below. I illustrate here with data from Tagalog, a language with a very high rate of flexibility:

use as a verb/predicate um-alis ang lalake
ᴀᴠ.ᴘᴇʀꜰ-leave D man
event 'the man left' titser ang lalake
teacher D man
object 'the man is a teacher'
use as a noun/argument: lalake ang um-alis
man D ᴀᴠ.ᴘᴇʀꜰ-leave
event 'the leaver (one who left) is a man'
lalake ang titser
man D teacher
object 'the teacher is a man'

In Tagalog, both event denoting words like *umalis* 'left' and object denoting words like *titser* 'teacher' function freely as arguments, the usual function of nouns and the reason for their common grammatical categories like case, by being the complements of DPs headed by a set of case marking determiners; the one illustrated in (5) and (6) is *ang,* the nominative determiner. But they both are also equally good predicates, the function associated with verbs; predicates in Tagalog are indicated by their normal initial position in the clause. Predicates are commonly specified for a number of aspectual, voice and other categories by a rich set of affixes. Crucially these are not restricted to only event denoting roots; most object denoting roots also can co-occur with them: *abogado* 'lawyer', *magabogado* 'study to become a lawyer, engage a lawyer'; *tao* 'person', *ma-tao* 'populated'; *manok* 'chicken', *magmanok* 'raise chickens'; *ipis* 'cockroach', *ipis-in*

'be infested with cockroaches'. These cannot be claimed as verbalizing suffixes because they occur also on underived verbs like *mag-linis* 'to clean', *linis-in* 'be cleaned', *ma-nood* 'to watch'.

St'át'imcets and Tagalog share a number of structural traits, and these in fact facilitate the high rate of flexibility in these languages. There may be other structural patterns that some languages may have hit upon to facilitate high rates of flexibility (e.g. Mundari **EvansEtAl2005mundari**), but that found in these two languages is crosslinguistically the most common, even accounting for languages like English. In languages like Latin the functions of arguments and predicates, prototypical uses of nouns and verbs, are built into the word forms themselves, into the inflections they take. But in languages like St'át'imcets, Tagalog and indeed English, this is not the case; rather these functions are indicated syntactically, not morphologically, and commonly phrasally, that is, there are phrasal functional categories like case and determiners to mark argument function and nouns and other functional categories like aspect or tense or agreement or just fixed syntactic constituent structure or perhaps a combination of these to mark predicate function and verbs. Predicate function is indicated by clause initial position and by the possibility of tense or aspect inflection in St'át'imcets and Tagalog, and also by subject agreement clitics in the former. Argument function is marked by being the complement of a determiner head in a DP in both languages (the theoretical model in which these phrase structures are cast is Lexical Functional Grammar; **Bresnan2001** the phrase structure may look different in other frameworks and even more so in dependency based frameworks, but the basic point here about heads being functional categories would still hold):

Predicates   Arguments

Figure 1: Favored Structures for Flexible Languages.

TAM indicates tense-aspect-mood inflection and IP, the projection of these inflections, and X any flexible lexeme. The phrase structure of a basic clause in both languages is identical and is shown in figure 2.

But there is an interesting contrast as well between St'át'imcets and Tagalog: the direction of derivation, in other words, which of the two meanings, object denoting or event denoting corresponds to the unmarked form. In St'át'imcets it is event denoting, while in Tagalog, it is object denoting. Consider the data in Table 2.

[IP, for tree=parent anchor=south, child anchor=north [I′ [I] [S [PRED] [(DP*)
]]]]

: :   fnum@figure

Austronesian/Salish phrase structure

XX	
St'át'imcets	Tagalog
7úqwa 'drink'	
s-úqwa 'a beverage'	*inom* 'drinking'
	um-imom 'drink'
núkw7-am 'help'	
s-núk'wa7 'friend'	*tulong* 'help'
	tulung-an 'help someone'
cwil'-em 'seek'	
s-cwil'-em 'something sought'	*bigay* 'gift'
	mag-bigay 'give'
náq'w 'steal'	
s-náq'w 'something stolen'	*nakaw* 'something stolen'
	mag-nakaw 'steal'

The prefix *s-* in St'át'imcets marks words which are object denoting, but in no sense can it be claimed as a derivational affix that regularly outputs nouns from basic verbs, because probably the majority of object denoting words in the language, derived or not, occur with it: *skuza* 'child', *sqáycw* 'man', *smúlhats* 'woman', *spzúza7* 'bird', *sqaxwá7* 'dog'. The point of the above examples is that in St'át'imcets, the root form has a verb-like event denoting meaning and the noun-like object denoting meaning is derived, but in Tagalog it is the opposite. This is most obviously brought out in the final examples of 'steal', 'something stolen'.

But even in languages with very high rates of flexibility, it is, as we shall see, not universal, and for the classical languages like Latin, the source of our contrasting categories of noun and verb, it is not the case that there is no flexibility (although there are certainly cases of languages with zero flexibility; this is common among the Papuan languages of mainland New Guinea). In Latin, about 10% of the lexemes of basic vocabulary in a survey I carried out with Johanna Nichols, me concentrating on Pacific languages, she on Eurasian and North American

languages, turn out to be flexible, close to the mean crosslinguistically that we established for this feature, as in Table 3.

Table 3: Flexible categorization in Latin

(i)	stem form	*-or* third declension *calor* 'warmth' ↑		*-ē* second conjugation *calēre* 'be warm' ↑
(ii)	inflection	DECLENSION ↑		CONJUGATION ↑
(iii)	syntax	N + CASE ↑		V + TENSE ↑
(iv)	semantics	Subject (*ónoma*) (thing/object)	+	PREDICATE (*rhêma*) (event)

To measure rates of flexibility across languages, we drew up a list of nearly 200 basic vocabulary items and then carefully pored over grammars and dictionaries of languages to determine whether each word base was flexible or not. The list of words we used covered a wide range of semantic categories:

- Properties: *heat, cold, length, width, dry, red, black, big, good*

- Experiential states: *fear, anger, shame, hunger, happy, sad*

- Bodily activities: *cry, sweat, sneeze, laugh, sleep, pee, poo*

- Posture: *sit, sit down, stand, stand up, lie, lie down*

- Activities: *run, walk, swim, fly, shout , sing*

- Actions on objects: *eat, bite, tear, hit, cut, open, break, throw*

- Transfer: *give, buy, say, tell*

- Perception: *see, look at, hear, listen to, know, forget*

- Contact: *pour, spill, load, empty, fill*

- Weather: *rain, thunder, lightning*

- Body parts: *ear, eye, hand, tongue, tooth, bone, elbow, hair, blood*

- Environment: *sun, moon, water, fire, sand, earth*

- Kin: *mother, father, child, sibling, spouse, name*

- Natural kinds: *dog, snake, fish, bird, pig, mouse, louse, ant, tree, leaf*

- Artifacts: *axe, spear, arrow, knife, house, broom, needle, string, clothing*

Flexibility was calculated as follows. A root was counted as inflexible either if 1. it had no derivational processes that shifted it from being object denoting or event denoting, or 2. any derivational affix which had such a shifting function was restricted to that use only and was never used on underived forms which had the same function as the derived form. Consider the following two entries from the corpus for Tagalog:

- 'snake': *ahas* 'snake'

- 'gone': *wala* 'not be, gone, extinct' *mawala* 'to be lost, to vanish, disappear' *mawalan, iwala* 'to lose something' *ikawala* 'to lose, cause one to lose something' *magwala* 'lose something from carelessness' *makawala* 'to miss, let slip by'/ 'to escape' *magpakawala, pakawala* 'to unbind, loosen, let free' *kawalan* 'want, deficiency' *pagkawala* 'disappearance'

ahas 'snake' is not flexible because it has no possible derived forms at all, never mind those which are event denoting. Note that this would not hold for English: *the road snaked its way around the mountain;* for English *snake* would count as flexible. *wala* 'gone, disappear' is also not flexible, because it is an event denoting predicate and all its derived forms, bar the last two, are also event denoting, predicating expressions. The only exceptions are the forms with the 'nominalizing' prefix *pagka-*, *pagkawala* 'disappearance' and the circumfix *ka-...-an, ka-wala-(a)n* 'want, deficiency', but these also fail to qualify the root for flexibility. The prefix *pagka-* has the sole role of deriving event nominalizations and never occurs on underived object denoting words; **pagka-ahas* 'snaking' is impossible (the circumfix *ka-...-an* is more complex, the details of which I cannot go into here; it turns out that it occurs with both object and event denoting roots, see **SchachterEtAl1972** on its functions). Now consider the following example of a flexible root in Tagalog:

- 'give': *bigay* 'gift' *magbigay/ibigay/bigyan* 'give someone something' *magbigayan* 'to compromise' *mamigay/ipamigay* 'to distribute, give out' *mapagbigayan* 'to accommodate someone in providing something'

The root *bigay* with no affixation at all means 'gift'. When it takes one of the voice affixes, it then takes on the meaning 'give' or some other closely semantically related event type. But these voice affixes cannot be claimed to be deriving a verb from a noun root (see **Foley2008 Foley2008** on this point), because voice is a necessary affixation for *any* event denoting predicating word, not just those seemingly derived from object denoting ones, as the voice affixes on all the event denoting forms for *wala* 'gone' above demonstrate. Hence *bigay* counts as a flexible lexeme, used either as 'gift' or 'give' (in the latter meaning requiring, as all event denoting predicating words do, voice affixation).

I surveyed fourteen Austronesian languages; I report on the data from seven of them in Figure 3:

Tolai, Mbula and Gedaged are all New Guinea region Austronesian languages of the Oceanic subgroup, and all of them have lower rates of flexibility than their sisters further afield. But even among these three, there is a significant difference in rates of flexibility; it is much lower in Mbula and Gedaged, approaching nil, than in Tolai. There are good sociolinguistic reasons for this. The contact with Papuan languages with their norm of zero flexibility has been much more intense for these two than it has been for Tolai. So intense indeed for Gedaged that its overall typology closely resembles its Papuan neighbors, with verb final clausal word order, postpositions and clause chaining constructions. These data strongly support the claim that flexibility is selected *against* in normal contact situations (those resulting in the genesis of pidgin and ultimately creole languages may be different). The mechanism by which language contact would lead to an increase or decrease in flexibility is not at this point entirely clear; more detailed research is needed. Is it piecemeal lexeme by lexeme as they are borrowed and either adapted or adapted to, importing a flexibility pattern for particular lexical items and then extending that to other items at a later stage through lexical diffusion? Or is it the case that speakers of the importing language abstract a general principle of flexibility or lack thereof from the source language and apply that to different lexical roots in their own language?

The contrastive situation between Tolai and Tongan is also remarkable. Tolai and Tongan like the other two New Guinea languages belong to the same Oceanic subgroup of Austronesian languages, and on archeological grounds, we know that the homeland of the proto-language of this subgroup was somewhere in the Bismarck Archipelago, the region where Tolai is spoken today. The ancestral language of Tongan like that of Tolai, not to mention Mbula and Gedaged, was spoken there. Note that the rate of flexibility of Tongan is the same as that of

Indonesian much further to the west and generally closer to that of the languages spoken in the western region of the Austronesian family. The languages of the west belong to a number of different high order subgroups and typically have high rates of flexibility, so on standard assumptions of historical linguistics, we would regard the high rate in Tongan as a retention from its ancestral language. So the question arises why do we find such high retention in Tongan, but not in Tolai? Tolai is in the New Guinea region, but its flexibility rate is much higher than Mbula and Gedaged, and its overall typology is that of a relatively conservative Oceanic language. Its speakers are originally from New Ireland, an island where today almost exclusively Austronesian languages are spoken. That may be the case, but there has been very significant contact with Papuan languages in its history. The genetic data tell the story. The speakers of Austronesian languages originally migrated out of Southeast Asia, so there are certain genetic markers that are closely linked with them. Speakers of Papuan languages, on the other hand, have been in situ in New Guinea for a very long time, at least forty thousand years, so they too are correlated with certain genetic markers. If we compare the Y-chromosomal DNA, which is inherited in the male line, from the father, and mitochondrial DND, which passes only through the female line, from the mother, for both Tolai and Tongan, we find a very interesting contrast (Table 4, **KayserEtAl2006**).

	Tolai		Tongan	
	Asian	Papuan	Asian	Papuan
Y-chromosome DNA	5.3	94.7	41.4	55.2
mitochondrial DNA	29.4	70.6	92.3	7.7

Tolai speakers have been swamped by Papuan genes, an indication of heavy contact through interbreeding. The percentage of Papuan Y-chromosomal DNA in Tolai is particularly high, and this is a signal of a favored cultural pattern of Papuan men marrying into Tolai communities (Tolai society like that of Proto-Oceanic is matrilineal), though many Papuan speaking women also contributed to the Tolai gene pool. For Tongan the percentages of Y-chromosomal DNA is more equally balanced, indicating that the ancestors of Tongan speakers did interbreed with speakers of Papuan languages as they migrated through the New Guinea region on their way to Polynesia, but to a much lesser extent. This is to be expected, as their presence in the New Guinea region could not have lasted more

than a few hundred years on current archeological evidence, while the ancestors of today's Tolai speakers have been there for three thousand years. But really remarkable are the percentages for mitochondrial DNA among Tongan speakers; it is almost exclusively of Asian origin. Part of this could be due to founder effects of small populations arriving in Polynesia, but not all. What it does tell us is that very few Papuan women entered the gene pool in the ancestral community. Papuan men commonly interbred with Austronesian women, but the reverse was very uncommon (again the matrilineal basis of early Oceanic society would have had a lot to do with this). This explains the preservation of the high flexibility rate in Tongan from Proto-Oceanic. There is much common wisdom in the term 'mother tongue'. Children were learning their language mainly from their mothers and other female relatives, and as these were Austronesian speakers and very rarely Papuan speakers, there was much less opportunity for the Papuan pattern to diffuse into ancestral Tongan.

Flexibility rates vary across the Austronesian languages surveyed. And even in languages with very high rates, such as those of the Philippines and Formosa, it is never the case that flexibility is universal; some lexemes strongly resist flexibility. But this is not random. It is tied to specific semantic categories. Consider Figure 4.

The central 'zero' line below each semantic category heading represents the baseline for each language, that is, the mean rate of flexibility across all categories for that language (Mbula and Gedaged are omitted from this figure because their mean is so low that nothing meaningful can be said about the distribution across categories). Under each semantic category heading, values are given for the degree to which words in that category depart from the language's baseline flexibility value. Note that certain semantic categories are mostly above the baseline, kin terms and particularly artifacts, so that they have higher flexibility rates, while others, natural kinds and especially caused actions are always below the baseline, with lower flexibility rates. This gives us empirical evidence for something we could call 'natural ontology'. For certain semantic categories, humans are strongly cognitively predisposed to classify the words labeling them as denoting objects or events and thereby further predisposed to only provide them with a grammatical categorization consonant with the expression of an object or an event (**GentnerEtAl2001**'s **GentnerEtAl2001** cognitive dominance). In languages with a sharp noun-verb distinction this feeds directly into that grammatical and lexical distinction. But in languages not so organized, the question is more complex. What criterion do we have for saying we have a grammatical and lexical category of noun, if all clear members are restricted to denoting natural kinds? This is just erasing difference, largely due to a theoretical preference for

Figure 4: Flexibility Rates across Semantic Types

assimilating languages to a shared base structure. I question the desirability of this move. We need to be more careful about the differences between languages before jumping to conclusions about similarities, largely on theoretical preference. If someone were to describe the difference between Latin and languages like St'át'imcets and Tagalog as one 'of degree, not of kind', I would ask then empirically what would count as a difference in kind if not the data reported here? Or are our theories so poorly framed that we cannot recognize a difference in kind when we see one? Or even worse, that differences in kind simply don't exist by virtue of theoretical fiat?