elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 7*. Editor: Peter K. Austin

Language documentation and linguistic theory

PETER SELLS

Cite this article: Peter Sells (2010). Language documentation and linguistic theory. In Peter K. Austin (ed.) *Language Documentation and Description, vol 7*. London: SOAS. pp. 209-237

Link to this article: http://www.elpublishing.org/PID/086

This electronic version first published: July 2014

Language documentation and linguistic theory

Peter Sells

1. Documentation and theory

This chapter is about the relationship between language documentation and linguistic theory.

1.1 The starting point

There was an article in the Chronicle of Higher Education (1st June 2009) called 'Languages on Life Support', with the subtitle 'Linguists debate their role in saving the world's endangered tongues', which quoted Michael Krauss who has been a very active field linguist for many years working in Alaska. It is all very well to generate ideas about how languages work (in other words, linguistic theory), Krauss and his fellow critics say, but those ideas will be next to useless without primary material to test them against. But the loss of languages affects more than just linguists. The world of our languages is a 'very fragile membrane that humanity depends on, that we evolved in, that makes us human', Krauss is quoted as saying. When languages disappear, cultures do, too – ways of thinking and describing, and of adapting to the globe's varied environments.

Recently, advocates of preserving dying languages can point to some signs of hope: students at the 3L Summer School, for example. Master's and doctoral programs emphasising documentary linguistics have grown in number and enrolments in several countries.

The article also quotes someone unexpected – Noam Chomsky from MIT. Chomsky, who spawned the theoretical turn in the field of linguistics, says that the loss of a language 'is much more of a tragedy for linguists whose interests are mostly theoretical, like me [Chomsky], than for descriptive linguists who focus on specific languages, since it means the permanent loss of the most relevant data for general theoretical work'. In that sense, a descriptive linguist working in, say, Africa, is far less affected by the death of a language in New Guinea than a theoretical linguist. This is view is open to debate. I am not going to enter into this discussion, but of course, for theoretical linguists all language data is terribly relevant.

Peter Sells 2010. Language documentation and linguistic theory. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 7,* 209-237. London: SOAS.

However, Chomsky says that his sympathy for endangered language communities does not mean that MIT, or any other department, should award PhDs for descriptive work alone. In linguistics, he says, 'just as in every other field, you can't do descriptive work without a theoretical understanding'.

Let us start making some sense of all of this, and for that we can turn to Nick Evans, another very well-known field linguist, from the Australian National University. In this same article he says that to compile a grammar is to live and breathe theory. The process of immersion, extraction, analysis, and summation of a language is, he argues, 'the most demanding intellectual task a linguist can engage in'.

1.2 Documentation and theory

In this chapter I focus on that part of language documentation which is directed towards grammatical description of a language ('compiling a grammar'). There are three ways in which language documentation in this sense and linguistic theory might interact:

> 1. grammatical description presupposes **some** theory – even basic description involves organisation of the data into categories and parts, and underlyingly this must suppose some linguistic theory. Linguistic theory can help provide a context within which to present a grammatical description. We will look at some examples below;
>
> 2. theory needs (more) data – linguistic theory can and surely must be informed by more and better data (this is Chomsky's point): this requires that the properties of the data that are fed into theory development are theoretically accurate. (It is of course a fully valid result to discover that some prior construct of linguistic theory was a mistake.) So theory needs data and theory needs more data;
>
> 3. a theoretical outlook can be useful, in the field and 'at home' – in terms of presentation of data to linguists, it is important to know what is theoretically interesting or relevant, what might be unusual, etc. Sensitivity to linguistic theory might invite language documenters to look for various phenomena in the language they are studying, without presupposing their necessary existence. This, of course, could involve elicitation (which is one of the research methods used in documentation, see Luepke's chapter in this volume), and for compiling complete grammatical sketches is probably necessary. So a theoretical outlook can be useful, in the field (in guiding work on the language) and also 'at home' (on return to the home institution).

1.3 Linguistic theory?

There are at least two possible senses of the term 'linguistic theory'; one that I mean and one that I do not mean. Let me begin with the one I do not mean:

(1) a particular representational system designed to account for properties of, and generalisations within, a specific set of data

In other words, a particular linguistic theory might work for that set of data but it might not work for anything else. Instead, what I think linguistic theory should give us, as we go out about language documentation, is:

(2) a flexible representational system used to account for various sets of data across different languages without giving primacy to any one set of data

I have been doing linguistic theory for several years and over those years I have come to feel that what linguistic theory tells us about language is that it has the following properties:

- **structure**
- **hierarchies –** this means that in certain situations when two possible outcomes could occur in the language data, one is privileged over the other
- **relationships** – e.g. agreement between the subject and the verb, or agreement between a pronoun and its antecedent
- **systematicities** – there is a system, there is a grammar
- **shared inheritances** – one might think of these as constructions, but most linguists have a sense that when they look at the grammar of any individual language there are certain kinds of commonalities – it has a certain character which it displays – and these are through its constructions.

In formal linguistic theory, we can have a notion of construction that would actually give some theoretical substance to the final intuition. But none of the above are necessarily properties from English or French or whatever the dominant language of theoretical development happens to have been. I myself over the years have been drawn to linguistic theories that allow themselves the kinds of properties that I described in definition (2) – a flexible representational system. Some good sources for this topic are Foley & Van Valin (1984), Comrie (1981), Payne (1997), Van Valin & LaPolla (1997) and Kroeger (2004) (see also Bond's chapter on typology in this volume).

2. Constructs in linguistic theory

Let us now concentrate on some things in linguistic theory that will actually be useful as we go out and start thinking about writing grammars. I will begin with a notion from semantics – the concept of thematic roles.

2.1 Thematic roles

It turns out to be very useful to classify arguments and predicates according to the role each participant has in an event. For example, a particular person might be an Agent in an event, or it might be an emotional state oriented towards a Goal. Over almost forty years of work, linguists have realised that there is a hierarchy to these thematic roles, as in:

(3) Agent > Experiencer > Instrument > Goal > Source > Theme/Patient > Location

What does it mean to classify arguments and predicates according to roles like these? Consider the following:

(4) (a) *give* < Agent, Goal, Theme >
(b) *remember* < Experiencer, Theme >
(c) *build* < Agent, Patient >

We might say that the verb *give* in (4a) is a 3-place relation between an Agent, a Goal and a Theme. A verb like *remember* in (4b) does not really involve an Agent in the same way that *give* does, and we might say that *remember* is a 2-place relation between an Experiencer and a Theme. Meanwhile, we might take *build* in (4c) as a 2-place relation between an Agent and a Patient. A Patient is something that is affected by the action: if you build something it comes into existence as a result, so in that sense it is seriously affected.

When we look at phenomena in different languages, we find that often there is sensitivity to these properties. One classic example, which goes back to Fillmore (1968), which had the misleading title *The case for Case* (it should have been *The role of Roles*), relates to how arguments of predicates can be expressed as Subjects in English (Fillmore 1968: 33):

(5) If there is an A [=Agent], it becomes the Subject; otherwise,
if there is an I [=Instrument], it becomes the Subject;
otherwise, the Subject is the O [=Objective, i.e., Theme/Patient].

The following examples show this:

(6) (a) The door opened.
(b) Dana opened the door.
(c) The chisel opened the door.
(d) Dana opened the door with a chisel.
(e) *The door opened by Dana.
(f) *The chisel opened the door by Dana.

Example (6b) has an Agent and a Theme. In (6c) the chisel is used to prise the door open, and is different from the Subject in (6b). If all three entities are included, we get (6d) – here the Agent is expressed as the Subject, the Theme is the Object, and the Instrument is in a Prepositional Phrase. However, not all possible combinations are available. So, for example, we cannot say something like (6e) in the sense of (6b). Neither can we say sentence (6f) in the sense of (6d). Fillmore proposed that, if we take part of the thematic hierarchy of (3) shown in (7a), then in (6c) where the Instrument is expressed as the Subject, an Agent cannot be included as it would be higher up the hierarchy. If Agent is chosen as the Subject, everything else is available, as in (6d). This means that the highest ranking semantic role is always expressed as the subject.

(7) (a) Agent > Instrument > Theme/Patient
(b) The argument of a verb bearing the highest-ranked semantic role is its Subject.

Exactly what the thematic hierarchy is or whether it is the same in every language is a matter of theoretical debate and theoretical elaboration, however this example from English shows the kind of thing that we would use thematic roles for. The labels for each role are intended to be helpful in figuring out what they are. Here are some examples of what all of these different roles might be:

- *Agent*: deliberately performs the action (e.g. **Bill** ate his soup quietly).
- *Experiencer*: receives sensory or emotional input (e.g. The smell of lilies delighted **Jennifer**).
- *Instrument*: used by an Agent to carry out the action (e.g. Jamie cut the ribbon **with a pair of scissors**).
- *Goal*: what the action is directed towards (e.g. The caravan continued on **toward the distant oasis**).
- *Source*: where the action originates (e.g. The rocket was launched **from Central Command**).

- *Theme*: undergoes the action but does not change its state; or is in a location (e.g. The baby kissed **the rabbit**).
- *Patient*: undergoes the action and has its state changed (e.g. The falling rocks crushed **the car**).
- *Location*: where the action occurs (e.g. Johnny and Linda played carelessly **in the park**).

Thematic roles and their hierarchy are implicated in the analysis of various phenomena in many languages – for example, the linking to grammatical functions (see below), or the binding of reflexive pronouns in some languages, or conditions on PP scrambling in Tongan, an Oceanic language (Otsuka 2005). PP scrambling refers to taking a Prepositional Phrase and moving it around from its expected position within the clause. Otsuka argued that the conditions on this scrambling are sensitive to the thematic hierarchy. The hierarchy can thus have very specific ramifications.

2.2 Grammatical functions

Grammatical functions are concepts such as SUBJect, OBJect, Second OBJect and OBLique. There are many subtypes of OBLique. Gramatical functions also fall into a hierarchy:

(8) SUBJECT > OBJECT > SECOND OBJECT > OBLIQUE

We are already familiar with at least the top end of the hierarchy. For example, for any sentence in a language, if the verb has any dependents it will almost certainly have a Subject, and if it has an Object it will almost certainly have a Subject as well. In Western European languages, in particular, Subjects and Objects are expressed by Noun Phrases (NPs) and Obliques are expressed by Prepositional Phrases (PPs). But this is only a rough correspondence.

What are Grammatical Functions for? Usually the grammatical information for dependents is defined with respect to their relationship to a predicate. For example, thematic roles associated with a predicate (e.g. Agent, Goal, Theme) are expressed through their grammatical functions of arguments which carry the roles. So, in the examples that we saw in (6), an Agent might be the Subject, or an Instrument might be the Subject – the roles remain the same but correspond to different grammatical functions, depending on particular uses of the predicate.

Consider this example from Spanish:[1]

(9)	*le*	*di*	*un regalo*	*a mi madre*
	her.DAT	gave.1SG	a present	to my mother
	'I gave a present to my mother.'			

The verb here *di* ('give') has a first person Subject (expressed inflectionally in the verb), there is an Object *un regalo* ('a present'), and the sentence has an Indirect Object which is actually expressed in two different parts: *le* ('to her') and *a mi madre* ('to my mother') – the two elements agree.

We can certainly say that the Subject of the sentence is the first person singular and the Indirect Object ('to my mother') has a dative in it. However, we cannot necessarily see that from the structure. If I asked 'does this verb have an Indirect Object?' the answer would be 'yes'. If I said 'show me where it is', you would have to say 'it's here and here'. And where is the Subject? It is somewhere! But you know what the functions are. You know what the Subject, Object and Indirect Object are, independent of the specific expression.

Doing this exercise involves using the concept of grammatical functions. The idea is that many of the grammatical properties that we want to describe hold (of a given verb or predicate) independently of any specific expression of the verb's dependents. That is, some grammatical properties hold independently of any specific thematic roles of the verb's dependents. For example, we are not talking about Agents having some property, we are talking about Subjects.

Linguistic theory has been concerned for a long time with the way that thematic roles relate to grammatical functions – this is often known as 'linking'. How do we link thematic roles to grammatical functions? A verb like *build* in (4c) will be expressed in a clause with a Subject and an Object if it is used transitively. Similarly, *remember* in (4b) will occur with a Subject and an Object if it is used transitively.

---

[1] The abbreviations used in this chapter are 1 = 1st person, 2 = 2nd person, 3 = 3rd person, ABS = absolutive, AV = Agent Voice, DAT = dative, DYN = dynamic, ERG = ergative, GEN = genitive, I = irrealis, LNK = link, LOC = locative, MUT = mutation, NEG = negative, NOM = nominative, NR = nominaliser, OBJ = object, OBJ2 = second object, OBL = oblique, PASS = passive, PL = plural, PN = proper name marker, POSS = possessive, PRED = predicate, R = realis, SG = singular, ST = stative, SUBJ = subject, TOP = topic, TV = Theme Voice.

If we find ways in which *build* and *remember* work identically in some language, we might think it has something to do with grammatical functions. If we find ways in which they work differently, that might be something to do with the thematic roles. In the following English examples, the identical forms on the associated dependent noun phrases show that the two clauses have the same grammatical functions but different thematic roles:

(10) (a) *He builds them*
(b) *He remembers them*

We can express this in terms of linking as follows:

(11) (a) *build* < Agent, Patient >

SUBJ, OBJ

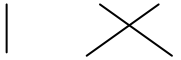
(b) *remember* < Experiencer, Theme >

SUBJ, OBJ

The verb *give* is a little bit more complicated – in English, for example, *give* can have multiple expressions where the same thematic roles are expressed with different grammatical functions:[2]

(12) *He gives the book to her*

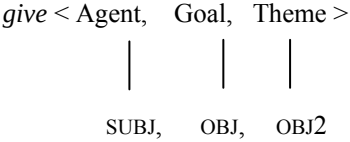
give < Agent, Goal, Theme >



SUBJ, OBJ, OBL

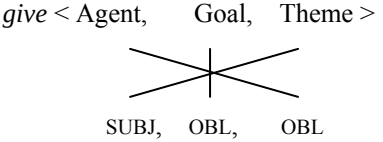
[2] The diagram in (12) looks a bit odd because of the crossing lines, but I have done that deliberately to maintain the thematic hierarchy and the function hierarchy.

(13) *He gives her the book*



Another way that languages have of providing more expressions for a given verb's thematic content is to have something like a passive voice. If we draw the linking for a passive verb, we have this:

(14) *The book was given to her by him*



The Theme is the Subject, the Goal is an Oblique, and the Agent is also an Oblique. Note that we can have several types of Oblique, each with its own thematic role: a Goal Oblique *to her*; and an Agent Oblique *by him*.

Why do we want grammatical functions? Grammatical functions and their hierarchy are implicated in the analysis of various phenomena in many languages, e.g. the choice of 'controller' in so-called 'control constructions' (Bresnan 1982), or the target of relativisation (Keenan & Comrie 1977). Surface coding properties such as agreement and case-marking often refer to grammatical functions. So in German (and in English for pronouns only), for instance, the Subject is typically Nominative, where we see the Nominative Case form, regardless of the thematic role that Subject has.

2.3 Accusative and ergative systems

Case-marking, mentioned in the previous section, shows a famous differences between languages in terms of Accusative and Ergative types. This is a very simple way of getting started with language description, but we need to go beyond it and use slightly more elaborated theoretical constructs.

The concept of Ergativity uses certain primitive terms to describe the arguments of transitive and intransitive verbs:

(15) (a) The single argument of an intransitive verb is called S for Subject

(b) The two arguments of a transitive verb are called A and P for Agent and Patient

(c) Any phenomenon which groups together S and A (i.e. the Subject of an intransitive and the Agent of a transitive) to the exclusion of P is known as a **Nominative** (or Nominative-Accusative) system. (This is what we find in familiar Western European languages.) In terms of case marking, P (alone) typically receives Accusative case

(d) Any phenomenon which groups together S and P to the exclusion of A is known as an **Ergative** (or Absolutive-Ergative) system. In terms of case marking, A (alone) typically receives Ergative case.

Notice that actually this terminology has a grammatical function name in (15a) and thematic role functions in (15b). The names do not really matter, but what underlies them does, and that is what the theory ultimately will tell us about.

2.4 Examples of ergative case marking

Yup'ik Eskimo is an Ergative language. The single argument of an intransitive verb is marked with a suffix *-aq*, which is the Absolutive (see 16a). The same case marking is used on the Object of the transitive verb while the Subject of the transitive verb has a special Ergative marking (see Payne 1997):

(16) (a) *Doris-**aq*** *ayallruuq*.
Doris-ABS travelled

(b) *Tom-**am*** *Doris-**aq*** *cingallrua*.
Tom-ERG Doris-ABS greeted

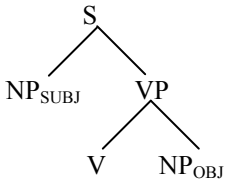
Now, if we translate these into a Nominative-Accusative language, then of course the Subject *Doris* in (16a) would be Nominative, the Subject *Tom* in (16b) would also be Nominative, and the Object of (16b) *Doris* would be Accusative – a totally different system.

This is primarily about one-argument verbs and two-argument verbs, and it is a good place to start, but we need to go beyond that, moving beyond thematic roles and grammatical functions.

2.5 Surface constituent structure

Linguistic theory has a further construct: phrase structure. Phrase structure representations combine syntactic categories of Verbs, Noun Phrases, Adpositional Phrases (Prepositional and Postpositional Phrases), Adverbs and so on into Sentences. In some languages, e.g. many Western European languages, Verbs and Noun Phrases can combine together into a Verb Phrase, to the exclusion of the Subject. The following is a phrase structure representation for English:

(17)



(18) (a) *The bird build**s** a nest.*
(b) *The bird**s** build∅ a nest.*

Here we have two Noun Phrases and a Verb Phrase. The higher NP is the Subject; the lower NP is the Object. The forms of the higher subject noun and the verb in (18) illustrate Subject-Verb agreement: a singular subject requires *–s* on the verb in the present tense (18a) while a plural subject requires no *–s* on the verb (18b). The form of the Verb co-varies with properties of the higher NP. What are the linguistic properties of that NP? In terms of phrase structure, it is the NP outside of the VP, rather than the one inside. In terms of grammatical functions, it is the Subject, rather than the Object. In terms of thematic relations, it is the Agent, rather than the Patient. Which of those is controlling Agreement? It could be thematic role, grammatical function or phrase structure. English does not help us answer this question; we need to find languages where the properties come apart more.

2.6 Verb classes (aspectual classes)

In the discussion so far we have been talking about 'verbs' in general, however we can actually get very fine-grained differences between verbs. An interesting example comes with *buy* and *sell*. Obviously, conceptually *buying* and *selling* are related, and we often find in many languages that one is derived from the other, whether in some transparent manner or not.

Which one is basic: *buying* or *selling*? It has been claimed at least for English that *buying* is basic, and the reason is because of the patterns in the following examples:

(19) (a) *I recently bought a concert ticket from my friend*.
(b) *I recently bought a coffee from the machine*.

(20) (a) *My friend recently sold me a concert ticket*.
(b) ??*The machine recently sold me a coffee*.

The strangeness of (20b) suggests that *sell* is more restricted than *buy*. The same might not be true in other languages – it will be necessary to test.

Cross-linguistically however we do find much bigger classes of verbs, namely Aspectual Classes, comprising State, Achievement, Activity and Accomplishment (Vendler 1967, Dowty 1979). A Stative verb in English would be *know*. An instance of an achievement verb is *melt*, in its intransitive sense – e.g. 'It is so hot today that I am melting'. An Activity verb would be *run* – e.g. 'He is running'. An Accomplishment is an event that has an endpoint and an example would be *build* in 'He built a house' (once you have built something, you stop building).

Why would this kind of thing be interesting? The answer comes when we look at a couple of examples from Japanese. Japanese has a verb-ending *-te iru* which is very much like the progressive 'is doing' in English. It is a diagnostic for these aspectual classes. Consider the following examples from Japanese:

(21) (a) *gakusei-tati-ga* *kyositu-de* *sawaide iru*.
student-PL-NOM classroom-in make-noise-TEIRU
'The students are making noise in the classroom.'

(b) *sono* *syozyo-wa* *ippo ippo* *nobotte iru*.
that girl-TOP step by step climb-TEIRU
'The girl is climbing step by step.'

(c) *sono* *syozyo-wa* *ki-no* *teppen-ni* *nobotte iru*.
that girl-TOP tree-GEN top-on climb-TEIRU
'The girl is on top of the tree.'

The verb in Japanese is clause-final and *-te iru* (or sometimes *-de iru*) is attached to the verb in (21a) and (21b) to express a progressive meaning – just like 'is doing' in English. In fact the *iru* component is the verb 'to be' in Japanese (used with animate subjects). Notice in (21c) the verb has a resultative meaning, literally '...has climbed'.

The 'be doing' form is traditionally used as a test for aspectual classes of verbs. Thus, in English *know* is a Stative verb because it is incompatible with

the 'be doing' form: we cannot say *'I am knowing'. Similarly, we say 'I understand', but not *'I am understanding'. In Japanese, on the other hand, Statives are not incompatible with the *-te iru* form: the way to say 'I know' is *sitte iru*, with the progressive form, and 'I understand' is *wakatte iru*, again with the *-te iru* form (literally 'I am understanding').

In Japanese, the semantic interpretation of the *–te iru* form is diagnostic of the difference between Achievement and Activity verbs. Thus, for an Achievement verb like *kuru* ('to come') *kite iru* does not mean 'is coming', but rather 'has come' (cf. *noboru* 'climb' in (21c)). On the other hand, for an Activity verb like *aruku* ('to walk') the *aruite iru* form means 'is walking'.

Similarly, in Japanese *kekkonsuru* ('to get married') is an Accomplishment verb because its *-te iru* form, *kekkonsite iru* does not mean 'to be getting married' but rather 'to be married' – it has happened. Use of the *-te iru* form with an Accomplishment verb (that has an end point) gives a resultative meaning, while use with an Activity verb gives a progressive meaning. Note that *–te iru* is not compatible with true Stative verbs in Japanese, and with Accomplishments it gives both progressive and resultative interpretations. This gives us the following distribution:

(22)

	progressive	resultative
Stative	NO	NO
Achievement	NO	YES
Accomplishment	YES	YES
Activity	YES	NO

(Note that in Japanese *siru* (glossed as 'know') is actually an Achievement verb that means something like 'to come to know' (to get knowledge). It is not a Stative verb, and so when used in the *-te iru* form it has a resultative sense. The same is true of *wakaru* 'to come to understand'.)

There is still a big theoretical question: 'how can this form *-te iru* sometimes mean progressive and sometimes mean resultative?' We know when it means one or the other, but not why. That is for linguistic theory to keep investigating.

2.7 Complex verb constructions

It is quite common cross-linguistically to find that sentences do not have just one verb in them, but they often have several. Understanding the relation between the verbs is crucial to understanding even some quite basic grammatical structures in some languages.

Consider the following examples from a Papuan language called Barai, an endangered language whose population 10 years ago was only around 2,000 (the data comes from Foley & Olsen 1985). Each sentence involves two verbs namely *fi* ('sit') and *isoe* ('write'):

(23) (a) Control verb ('sit down to V')

Fu ***fi*** *fase* ***isoe***.
3SG sit letter write
'He sat down to write a letter.'

(b) V-V complex predicate

Fu *fase* ***fi*** ***isoe***.
3SG letter sit write
'He sat writing a letter.'

In example (23a) we see that Barai is a so-called OV language (i.e. the Verb follows its Object): *fase isoe* (literally 'letter write'). I have called (23a) a 'Control verb construction' – this is a term from linguistic theory suggested by Foley & Olsen's analysis since it means 'to sit down to do something'. I have called (23b) on the other hand a 'V-V complex predicate' where the Agent is doing two things at the same time, rather than doing one in order to do the other.

We can see from the meaning here that these are probably different constructions. But we can also look at syntactic differences between them to check that they really are different constructions – this is what Foley & Olsen do.

(24) (a) *Fu* ***isema*** *fi* *fase* *isoe*.
3SG wrongly sit letter write
'He sat wrongly and wrote a letter.'

(b) *Fu* *fi* *fase* ***isema*** *isoe*.
3SG sit letter wrongly write
'He sat down and wrote a letter wrongly.'

(c) *Fu* *fase* ***isema*** *fi* *isoe*.
3SG letter wrongly sit write
'He wrongly sat writing a letter.'

(d) **Fu* *fase* *fi* ***isema*** *isoe*.
3SG lettersit wrongly write

If we take an adverb like *isema* ('wrongly') and put it before the first verb *fi* ('sit') in the Control structure, it means 'He wrongly sat and wrote a letter', as in (24a). You can put *isema* before the second verb *isoe* ('write') and then it means 'The sitting was alright but it was the writing the letter that was wrong', as in (24b).

If we look at the Complex verb structure, we can put 'wrongly' before the first verb, but we cannot put it between the first and second verbs – see (24c, d). This shows that in (24c) *fi isoe* is some kind of unit which does not allow itself to be interrupted by the adverb.

Here are some further examples, involving the negative *naebe*:

(25) (a) *Fu* ***naebe*** *fi* *fase* *isoe*.
3SG NEG sit letter write
'He did not sit down but did write a letter.'

(b) *Fu* *fi* *fase* ***naebe*** *isoe*.
3SG sit letter NEG write
'He sat down but did not write a letter.'

(c) *Fu* *fase* ***naebe*** *fi* *isoe*.
3SG letter NEG sit write
'He did not sit and write a letter.'

(d) **Fu* *fase* *fi* ***naebe*** *isoe*.
3SG letter sit NEG write

Notice that naebe cannot appear between the two verbs in the second type. In terms of a description of the constructions, the Control structure might be indicated: [sit [letter write]]. The Complex Verb one is: [letter [sit write]]. We cannot put an adverb between the two verbs in the second case. Is that because they are really a single word? Is it because of something in the syntactic structure? Is it to do with the way the semantics works? It is not clear from these examples, and we would have to explore Barai further to be able to answer the questions. There is an enormous theoretical literature on the ways in which complex verbal constructions can be built, from true compounding to clausal coordination. The Barai examples must fall somewhere within that space, theoretically. And if they do not, the theory needs to change so that they do. It would be really interesting to take the analysis of Barai further.

3. Austronesian

This section is about Austronesian languages. The first is Nias (also called Nias Selatan) spoken on an island near Sumatra. The population is around 600,000, and all the data is taken from a dissertation by Brown (2001) on this variety.

3.1 Nias Selatan

Japanese and Barai are languages where the verb is canonically at the end of the sentence, but in Austronesian languages the verb will typically be initial in the sentence. Thus, in example (26a), we see the verb *manavuli* ('return') followed by a particle *sui* ('again'), then the Subject *nama Gumi* ('father Gumi'), and then the Oblique *ba Maenamölö* ('to Maenamölö'). The word for 'father' is *ama*, but in certain circumstances it takes a different form, which is known as a mutation. The mutations are marked in the glosses as MUT.

(26) (a) *manavuli* *sui* [***n**-ama Gumi*] *ba* *Maenamölö*.
return again MUT-father Gumi LOC Maenamölö
'Father Gumi came back again to Maenamölö.'

(b) *I-a* [***m**-bavi*] [*ama Gumi*].
3SG.R-eat MUT-pig father Gumi
'Father Gumi eats pig.'

We see that *ama* is mutated to *nama* when it is the Subject of an intransitive verb. In terms of the Ergative account given above, the S argument of an intransitive verb mutates. However, with a transitive verb, it is the Patient that gets the mutation, not the Agent, as we can see from example (26b). We can summarise Nias case marking as:

(27) The S argument of an IV mutates.
The P argument of a TV mutates; but not the A.

This looks like an Ergative system, however we need to explore further. In Nias, as in many languages, when you ask a question the answer can just be a simple Noun Phrase that gives the information wanted. There are at least two ways of answering the question 'What did they steal?' in Nias: you can say 'his money' or 'they stole his money'.

(28) (a) *Haija* *ni-tagö?*
what PASS-steal
'What did they steal?'

(b) *Kefe-nia /* *La-tagö* ***g****efe-nia.*
money-his / 3SG.R-steal MUT.money-his
'His money / They stole his money.''

Notice that in the second answer the Object of the transitive verb is in the mutated form, but in the first answer, it is not. In the first answer there is no Verb Phrase present, or at least, this is not the Object of a verb in the relevant sense. So whatever structure the second answer has that causes the mutation to happen, the first answer does not have that structure.

Is this really an ergative system? It *could* be, but there are some things about it that follow the Nominative system, namely verb agreement. In the following examples, we have the mutation (given in bold), (29a) is an intransitive verb, (29b) is a transitive verb whose Subject is not formally expressed but is just part of the verb (an inflection):

(29) (a) *Ya-ma-nana* ***n****-ono-nia* *ba* *va-a-lio*.
3SG.I-DYN-crawl MUT-child-3SG.POSS LOC MUT.NR-ST-quick
'Her child will be crawling soon.'

(b) *Ya-mbu'a* ***g****-ömö-nia*.
3SG.I-repay MUT-debt-3SG.POSS
'He might repay his debt.'

Verb agreement always goes with the Subject of the sentence: in the case of an intransitive verb this is the Noun Phrase that mutates, but in the case of a transitive verb the verb it is the non-mutating NP. From a theoretical perspective, case marking follows an Ergative system, but agreement in the Irrealis mood follows the Nominative system (the verb agrees with S and A).

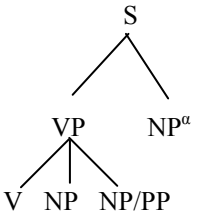
This is not the whole story however. When we look at different kinds of predicates we find something interesting. Here is an Experiencer predicate:

(30) *A-ta'u* ***m****-ba'e* ***n****-ono* *matua*.
ST-fear MUT-monkey MUT-child male
'The monkey is afraid of the boy.'

Here we have a transitive predicate but both Noun Phrases are mutated. Crysmann (2009) proposes an analysis to account for this, namely that there is a Verb Phrase constituent in the syntactic structure in Nias, and this VP hosts all of the arguments of the verb except for the Agent of a Transitive Verb – this is why you get the appearance of ergativity. Only NP arguments inside the VP mutate, so the mutation is not actually signalling anything about the

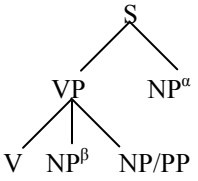
thematic or grammatical relations, but is actually about the chunking of the structure. An Experiencer predicate takes arguments with semantic roles Experiencer and Theme, and because there is no Agent both arguments are inside the VP and hence mutate. In other words, mutation (case marking) reflects the syntactic structure, which we can analyse as follows:

(30)



The NP on the right labelled NP^{α} is the Agent of a transitive verb, and does not mutate. Everything else is inside the VP: the single argument of an intransitive verb, two arguments of an Experiencer predicate. Anything inside that domain mutates. Agreement is with the Subject of the verb (the highest argument on the thematic hierarchy).

(31)



In English we can we can identify NP and VP, and there are some things that go inside the VP. If Crysmann is right about Nias, there are more things that go inside the VP in Nias than in English. In other Austronesian languages the VP is smaller, possibly having nothing more than NP^{β}.

What should Linguistic Theory give us in a case like this? It should give us the idea that there are things like NPs and VPs, but I do not think that it should give us the idea that there are VPs just like English VPs. We want a *flexible* representation system that allows you to look for patterns without prejudging what it is we are going to find.

Nias Selatan looks like it is an Ergative language, but it is not really – actually the mutation Case Marking has to do with phrase structure, and there is also some connection to thematic roles. Nias does something special with Agents of transitive verbs and has some interesting phrase structure (e.g. quite a 'big' VP). Why does it do this? Do other languages nearby do this? Do other languages far away do this? We do not know; we should go and find out.

3.2 Austronesian voice

This section is about other Austronesian languages; in particular it will involve the notion of voice. I mentioned Passive above, and we often think from an Indo-European tradition of Active voice and Passive voice. In the Passive, the Theme or Patient argument of a transitive verb is the one that is the Subject. Many Austronesian languages have a different system known as a Symmetric voice system.

Typically, across languages for a transitive verb that has an Agent and a Theme, the Agent is the Subject and the Theme is the Object (this is the default mapping mentioned above). What many Austronesian languages do is they allow that linking to be 'flipped', so the Agent is the Object and the Theme is the Subject. Such a system is therefore Symmetrical in that sense.

Passive actually involves 'demotion' of the Agent so that it is typically coded like an oblique, as in 'The house was built *by* the carpenter'. In the Austronesian system, there is no demotion: there is simply Subject and Object and the alignment of thematic roles with grammatical functions is flipped.

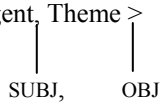
Consider Toba Batak, spoken in Sumatra, Indonesia. The first linking called Agent Voice (Agent=Subject, Theme=Object) is marked with *mang-* on the verb, and the second called Theme Voice (Theme=Subject, Agent=Object) is marked with *di-* on the verb, as in the following schema:

(32) (a) The student ***mang***-read the book.
(b) The book ***di***-read the student.

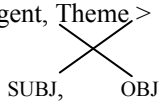
Examples (32a) and (32b) have the same propositional meaning but have different information structure properties, and different effects on specificity. Note that (32b) is not a passive since the agent is expressed as a direct argument of the verb and is as obligatory as the Agent in (32a).

This is a Symmetric voice system where we find the following two alternatives for linking:

(33) *Agent Voice* < Agent, Theme >



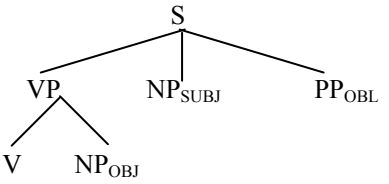
(34) *Theme Voice* < Agent, Theme >



This is important for linguistic theory because for many years there were theories of linking that said the only possible analysis is (33) and not (34); documentation of Austronesian languages shows that both are possible.

Toba Batak has a small VP which can only have an Object NP inside it. Subjects and Obliques are outside the vP, in structures like the following:

(35)



mang-V	Thm	Agt
di-V	Agt	Thm

There is evidence for this structure from various phenomena in Toba Batak which are explored in the following section.

3.3 Toba Batak binding

Recall that we have proposed the following linking rules for Toba Batak:

(36) Linking
- (a) Agent Voice: the thematically highest argument is the SUBJ.
- (b) Theme Voice: a thematically non-highest argument (Patient or Theme) is the SUBJ.

When we look at 'binding of reflexives', that is what NP a reflexive pronoun can corefer with, we find the following principles hold:

(37) Antecedents of a Reflexive Pronoun ('Binding')
(a) Core Grammatical Functions (SUBJ/OBJ) antecede OBLs (GF hierarchy).
(b) For arguments of equal GF rank, the Thematic Hierarchy governs binding.

Consider the following examples (the subject in each is underlined):

(38) (a) *Mang-ida* *diri-na*$_i$ *si John*$_j$.
AV.see self-his$_i$ PN John
'John saw himself.'

(b) **Mang-ida* *si John*$_i$ *diri-na*$_i$.
AV.see PN John self-his$_i$
'Himself saw John.'

(c) **Di-ida* *diri-na*$_i$ *si John*$_i$.
TV.see self-his$_i$ PN John
'Himself saw John.'

(d) *Di-ida* *si John*$_i$ *diri-na*$_i$.
TV.see PN John self-his$_i$
'John saw himself.'

In (38a) in the *mang-* Agent Voice form of the verb (indicating the Agent is the Subject and the Theme is the Object) the antecedent of the pronoun **follows** it in the sentence. Switching the order of antecedent and reflexive results in the ungrammatical (38b). In *di-* Theme Voice the opposite is true: the antecedent (Object) must **precede** the reflexive (as in (32d) and if it follows the sentence in ungrammatical (32c). The generalization is that neither linear order nor grammatical functions determine the binding, rather Agents always bind Themes (even in cases like (32d) where the Agent is Object and the Theme is Subject.

When the reflexive is an Oblique, either Core argument (SUBJ or OBJ) can be the antecedent in either voice, regardless of surface configuration. In the following examples *ho* always binds *dirim* regardless of whether it is subject (39a, 39d) or object (39b, 39c):

(39) (a) *Manghatahon* *si* *Torus* *ho*$_i$ *tu* *dirim*$_i$.
AV.talk.about PN Torus 2SG to self
'You talk about Torus to yourself.'

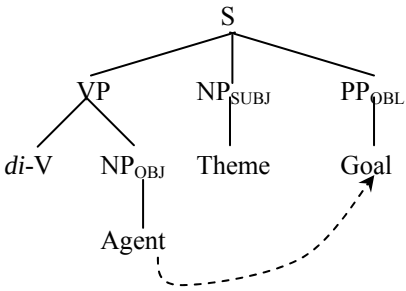
(b) *Manghatahon* *ho*$_i$ *si Torus* *tu* *dirim*$_i$.
AV.talk.about 2SG PN Torus to self
'Torus talks about you to yourself.'

(c) *Dihatahon* *ho*$_i$ *si* *Torus* *tu* *dirim*$_i$.
TV.talk.about 2SG PN Torus to self
'You talk about Torus to yourself.'

(d) *Dihatahon* *si* *Torus* *ho*$_i$ *tu* *dirim*$_i$.
TV.talk.about PN Torus 2SG to self
'Torus talks about you to yourself.'

The binding in (39c) is as shown by the dotted line in (40), with the verb as in (41):

(40)



(41) Verb in *di*-form (39C):

SUBJ, OBJ, OBL

PRED	'*talk-about* < AGENT, GOAL, THEME >'
SUBJ	['Torus']
OBJ	['you$_i$']
OBL	['yourself$_i$']

We can see in (41) that the Agent is OBJ, due to the *di*-marker for voice. In terms of the phrase structure tree (40), this Agent phrase is somewhat embedded; but in terms of the thematic hierarchy, Agent is always highest, and it is this property that allows it to bind the antecedent of the reflexive within the oblique PP.

3.4 Batak fronting

Binding in Toba Batak follows the thematic hierarchy and has nothing to do with the overt syntactic structure. However, there are some things that do, namely various constituent positionings within the clause. Recall that for Batak (see (35)) we have proposed a VP followed by the Subject which is then followed by other things such as adverbs like *nantoari* 'yesterday'.

Adverbs can be placed in front of the verb at the beginning of the clause (affecting emphasis, but not changing propositional meaning), as in (42a). The adverb can never be placed between the verb and its Object, as in the ungrammatical (42b). Placing the adverb between the VP and the Subject is fine, as in (42c), as is sentence-final position, as in (42d).

(42) (a) ***Nantoari*** [*mangida si Ria*] <u>*si Torus*</u>.
yesterday [AV.saw PN Ria] PN Torus
'Torus saw Ria yesterday.'

(b) *[*Mangida* ***nantoari*** *si Ria*] <u>*si Torus*</u>.
(*Adverb in VP)

(c) [*Mangida si Ria*] ***nantoari*** <u>*si Torus*</u>.

(d) [*Mangida si Ria*] <u>*si Torus*</u> ***nantoari***.

In the following examples we see exactly the same pattern, but with the *di-* form of the verb rather than the *mang-* form:

(43) (a) ***Nantoari*** [*diida si Torus*] <u>*si Ria*</u>.
yesterday [TV.saw PN Torus] PN Ria
'Ria was seen by Torus yesterday.'

(b) *[*Diida* ***nantoari*** *si Torus*] <u>*si Ria*</u>. (*Adverb in VP)

(c) [*Diida si Torus*] ***nantoari*** <u>*si Ria*</u>.

(d) [*Diida si Torus*] <u>*si Ria*</u> ***nantoari***.

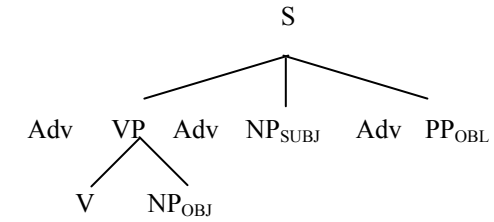
There is no necessary fixed order between adverb and subject, even when both are post-VP. These examples are exactly the same as those in (41) in terms of what they show about the distribution of the different phrases. This is important because it demonstrates that the overt syntax of the two voices is the same. They **are** symmetrical in the sense that the language treats both of them identically.

Again, this is different from Active and Passive constructions: in most languages, the syntax of Passive sentences is a little different from the syntax

of Active sentences. But in these Symmetric voice languages the syntax is the same.

What does this tell us about the structure? Considering the positions for Adv in (43), these are the positions where the adverb **can** go – (42a, c, d). It **cannot** go inside the VP, between the verb and the Object – (42b). The simplest explanation for that would be that the adverb is structurally at the same level, just under the S, as the NP and the VP do – see example (43). It is excluded from the VP, and as an immediate daughter of S it can permute with the VP, NP and PP.

(44)



mang-V	Theme	Agent
di-V	Agent	Theme

These examples show that adverbs, even though intuitively modifying the verb are not inside the VP constituent. The fact that example (42b) is ungrammatical shows that adverbs are **never** inside the VP.

Schachter (1984) gives the following coordination data, with the structure of [[V O] & [V O] S], as the structure in (43) would predict:

(45) (a) [*Mangantuk si John*] *jala* [*disipak si Bob*] *si Fred*.
[AV.hit PN John] and [TV.kick PN Bob] PN Fred
'Fred hit John and was kicked by Bob.'

(b) [*Diantuk si John*] *jala* [*manipak si Bob*] *si Fred*.
[TV.hit PN John] and [AV.kick PN Bob] PN Fred
'Fred was hit by John and kicked Bob.'

There is another very similar construction, termed 'preposing'. Normally, the predicate is first in the clause, as in (46a), but sometimes other things can precede it. Example (46b) shows a fronted Subject. But, keeping the form in

(45a), if we try to front the Object (in bold), we get an ungrammatical example (45c).

(46) (a) *Mamboan ulos angka sisolhot*.
AV.bring cloth PL relative
'The relatives bring cloth.'

(b) *Angka sisolhot mamboan ulos*.
(c) ***Ulos** mamboan angka sisolhot*.

Fronting the underlined NP (the Subject) is good; but fronting the Object is bad. Notice that we **can** front an adverb in the presence of a Subject, however an Object cannot be fronted in the presence of a Subject.

That suggests that grammatical functions are important. In fact, the voice system indicates, for a given argument of the predicate, that it could be the Subject or it could be the Object. In order to front the noun meaning 'cloth' the other voice must be used:

(47) (a) *Diboan angka sisolhot ulos i*.
TV.bring PL relative cloth the
'The relatives brought the cloth.'

(b *Ulos i diboan angka sisolhot*.
(c) ***Angka sisolhot** diboan ulos i*.

These phenomena are found across Western Austronesia. This is exactly the same as examples (46), except that we are using the other voice – so the Theme is the Subject and the Agent is the Object. It still means 'The relatives brought the cloth', but now 'the cloth' is the Subject, and can be fronted.

Binding is sensitive to the thematic hierarchy. Adverb placement is sensitive to surface phrase structure. These examples clearly show that whatever fronting is about it is not to do with the thematic hierarchy, because sometimes you can front an Agent and sometimes you cannot, sometimes you can front a Theme and sometimes you cannot – it is dependent, rather, on the grammatical functions and their position in the phrase structure.

Relativisation data shows a 'Subject-only' property, and the Object cannot be relativised (the c-examples in (48)-(49)).

(48) (a) *Manjaha buku guru i.*
AV-read book teacher the
'The teacher is reading a book.'

(b) *Guru na manjaha buku i*
teacher LNK AV-read book the
'the teacher who is reading a book'

(c) **Buku na manjaha guru i*
(Object (Theme) cannot relativise)

(49) (a) *Dijaha guru buku i.*
TV.read teacher book the
'A teacher read the book.'

(b) *Buku na dijaha guru*
book LNK TV.read teacher
'the book which a teacher read'

(c) **guru na dijaha buku i*
(Object (Agent) cannot relativise)

In Toba Batak the preposing operation cannot apply to Objects. From a theoretical point of view, there could be two reasons:

1. fronting does not apply to Objects;
2. fronting does not apply to things inside the VP (and the VP has the Object inside it).

Question formation also appears to have something to do with phrase structure and not with grammatical functions, even though the idea that these constructions make reference to grammatical funtions is very very strong in the Austronesian literature. In example (50) we go back to the verb *give*, with three arguments – an Agent, a Goal and a Theme.

(50) (a) ***Ise*** *mang-alean missel i tu soridadu?*
who AV-give missile the to soldier

(b) ****Aha*** *mang-alean jeneral i tu soridadu?*
what AV-give general the to soldier

(c) ***Tu ise*** *mang-alean missel i jeneral i?*
to who AV-give missile the general the

To form a question in this language, the question word must be sentence-initial. Fronting the Subject as in (50a) to mean 'Who gave the missile to the soldier?' is fine. However, fronting the Object in (49b) to mean 'What did the general give to the soldier?' is bad. In this example 'the general' is the Subject (it is underlined). Finally, fronting an Oblique, as in (50c) is good. The question word is thus behaving just like an adverb.

If preposing (above) and question fronting are the same kind of construction, it is not a matter that Subjects do it and nothing else does. The right generalisation is that Objects do **not**, and we have to find out in the language why that is true.

Earlier I mentioned that languages have structure, hierarchies, systematicities and shared inheritances. I have just made reference to a shared inheritance in the form: 'if preposing and question fronting are the same kind of construction...'. I do not know whether it is or not – I have to look at more examples to try to find out. But it might well be, and then you would find these kind of regularities coming over and over again.

Note also that the post-VP PP and the subject can be freely ordered with respect to each other:

(51) (a) *Mangalean missel i jeneral i tu soridadu.*
AV.give missile the general the to soldier
'The general gave the missile to the soldier.'

(b) *Mangalean missel i tu soridadu jeneral i.*
AV.give missile the to soldier general the
'The general gave the missile to the soldier.'

This suggests that phrases external to the VP may easily reorder (with each other).

4. Conclusion

In conclusion, in this chapter I have addressed three relationships between language documentation and linguistic theory:

- grammatical description presupposes **some** theory – and getting it right means getting some theoretical distinctions right (as I have tried to illustrate with Nias and Batak, for example);
- theory needs (more) data – there is so much we do not know;
- a theoretical outlook can be useful, in the field and 'at home' – it can help generate hypotheses about what kind of data to test, what kind of things to look for, etc.

And finally, I hope I have succeeded in conveying to you that a theoretical outlook can be fun!

References

Bresnan, Joan. 1982. Control and complementation. *Linguistic Inquiry* 13(3), 343-434.

Brown, Lea. 2001. *A grammar of Nias Selatan*. Ph.D. thesis, University of Sydney.

Comrie, Bernard. 1981. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

Crysmann, Berthold. 2009. Deriving superficial ergativity in Nias. In Stefan Müller (ed.), *The proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar (HPSG09)*, 68-88. Stanford: CSLI Publications. http://csli-publications.stanford.edu/HPSG/2009/ (accessed 2010-01-28).

Dowty, David R. 1979. *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Dordrecht: Reidel.

Fillmore, Charles. 1968. The case for case. In Emmon Bach & Robert T. Harms (eds.), *Universals in linguistic theory*, 1-88. New York: Holt, Rinehart and Winston.

Foley, William & Mike Olsen. 1985. Clausehood and verb serialization. In Johanna Nichols & Anthony C. Woodbury (eds.), *Grammar inside and outside the clause: Some approaches to theory from the field*, 17-60. Cambridge: Cambridge University Press.

Foley, William A. & Robert D. Van Valin, Jr. 1984. *Functional syntax and universal grammar* (Cambridge Studies in Linguistics 38). Cambridge: Cambridge University Press.

Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1), 63-99.

Kroeger, Paul R. 2004. *Analysing syntax: A lexical-functional approach*. Cambridge: Cambridge University Press.

Otsuka, Yuko. 2005. Syntax and/or pragmatics: PP-scrambling in Tongan and the thematic hierarchy. In Jeffrey Heinz & Dimitrios Ntelitheos (eds.), *Proceedings of the twelfth Annual Conference of the Austronesian Formal Linguistics Association (AFLA)* (UCLA Working Papers in Linguistics 12), 343-357. Los Angeles, CA: University of California, Los Angeles.

Payne, Thomas. 1997. *Describing morphosyntax: A guide for field linguists*. Cambridge: Cambridge University Press.

Schachter, Paul. 1984. Semantic-role-based syntax in Toba Batak. In Paul Schachter (ed.), *Studies in the structure of Toba Batak* (UCLA Occasional Papers in Linguistics 5), 122-149. Los Angeles, CA: University of California, Los Angeles.

Van Valin, Robert D., Jr. & Randy LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

Vendler, Zeno. 1967. *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.

Discussion Questions

1. Are Agents always Subjects?

2. If a verb or predicate agrees with something (say in terms of person, number), what does it typically agree with?

3. What kinds of verb classes do languages have? How do we look for them? How can we see the difference? (Things to keep in mind: intransitive, transitive, ditransitive, different aspectual classes, different morphosyntax for stative predicates, apparently transitive verbs that take oblique second arguments, e.g. ‘hit at/on (something)’, ‘meet with (someone)’, ‘wait for (someone)’.)