Syntax and semantics for sentence processing in English and Māori

Ian Bayard, Alistair Knott and John Moorfield
Dept of Computer Science, School of Māori Studies
University of Otago*

September 30, 2002

1 Introduction

This paper provides an account of the syntactic and semantic representations used in Te Kaitito, a bilingual English/Māori natural language processing system. With the exception of some very simple systems (e.g. Leslie, 1998), the grammar given in Te Kaitito is the first implemented computational grammar for Māori, indeed (we believe) for any Polynesian language. It is therefore of interest to linguists studying these languages. For linguists not specifically working with Māori or Polynesian languages, it might also be interesting as an indication of the extensibility of HPSG to Polynesian languages.

Section 2 outlines the criteria we are using to evaluate the role of syntactic and semantic representations in our system. Sections 3 and 4 describe the syntactic and semantic frameworks we are using, and Section 6 details some of the grammatical structures currently covered by the system.

2 Criteria for the model of syntax and semantics

Our development of a model of English and Māori syntax and semantics was directed by a number of requirements. The first requirement was that the same declarative syntactic and semantic resources should be used for generation and interpretation of sentences, and that the syntactic resources for the two languages should be incorporated into a single grammar (see Knott *et al.*, 2002 for the motivation behind this requirement.) The second requirement was that sentences which are paraphrases or translations of one another should receive the same semantic representation. This requirement is probably too strong; the machine translation literature is full of cases where the 'literal' semantic representation of a sentence differs from that of its translation. However, it makes sense for us to begin by looking at simple cases, and the modifications necessary for more complex cases are well accommodated within our chosen semantic framework (see Copestake *et al.*, 1995). The final requirement was that our models of syntax and semantics should be theoretically well-motivated, and draw as much as possible on existing research in both English and Māori.

3 Syntactic Background

3.1 LKB

Our grammar development environment is the LKB (Linguistic Knowledege Building) system (Copestake *et al.*, 2000). This system supports bidirectional sentence generation and interpretation for unification-based grammar formalisms. Our grammars are fairly closely based on one such formalism—HPSG (Head-driven Phrase Structure Gram-

*This work was supported by University of Otago Research Grant MFHB10, and by the New Zealand Foundation for Research in Science and Technology (FRST) grant UOOX02.

mar; Pollard and Sag, 1994).

3.2 HPSG

The details of the HPSG formalism are mostly not important for our presentation of the English and Māori grammars. The best way of thinking about our syntactic model is simply as a method of associating particular English or Māori syntactic constructions with particular hierarchical constituent structures, and many different grammar formalisms would be able to do this. Having said that, there are a few principles from HPSG which relate directly to the kind of trees which the model uses, which will be described below.

X-bar theory Firstly, the model conforms to **X-bar theory**, in which key lexical items in a sentence determine or **project** the syntactic contexts in which they appear. These projected contexts have a standard hierarchical structure: the lexical item (termed the **head**) first projects its **complements** (obligatory syntactic elements), then zero or more **adjuncts**, and finally a **specifier**, as shown in Figure 1 (i). Figure 1 (ii) shows a standard HPSG analysis of an English sentence (note that the subject is the specifier of the verb, and thus subsumed within the verb phrase).

Gap features Analyses within the tradition of generative grammar trade heavily on the idea of 'movement of constituents' between syntactic positions. In HPSG, movement is modelled declaratively, by the use of **gap features**. When a sentence requires that a word be extracted from its usual place, such as the word *is* in a yes-no question (*is the dog barking*, for instance) we use a rule to allow the formation of a phrase which is missing one of its usually-required components, and mark this phrase with a feature which can be passed up the tree in the GAP feature of the head of the phrase. This GAP feature is passed up, like any other feature, to the higher levels of the parse tree. Higher up the parse tree a second rule is invoked which references and removes this GAP feature, and allows the addition of the missing item.

Functional projections Finally, it is in the spirit of HPSG to avoid as much as possible the postulation of phrase projections with phonologically empty heads. We have adhered to this principle, mainly to keep the complexity of our grammars to a minimum.

3.3 Multilinguality

We work with a combined grammar for English and Māori, using a feature `language` with alternative values for lexemes and grammatical rules in the two languages. Sentences containing a mixture of words from different languages are ruled out by feature agreement constraints. The parser can therefore accept sentences in either English or Māori, and the generator produces sets of paraphrases in both English and Māori.

4 Semantic background

4.1 Minimal Recursion Semantics

The representation LKB uses for sentence semantics is minimal recursion semantics or **MRS** (Copestake *et al.*, 2001). Again, the details of this formalism are not very important: the main point is that the semantic representation of a sentence is built compositionally from the semantics of its component lexical items, in a way which is guided by the sentence's constituent structure. There are a few relevant features of LKB's semantic framework, however, which will be discussed below.

Words with null semantics In both English and Māori there are a small number of words that make no semantic contribution to the semantics of a sentence in which they appear. LKB allows for such words, which are designated as words with **null semantics**.

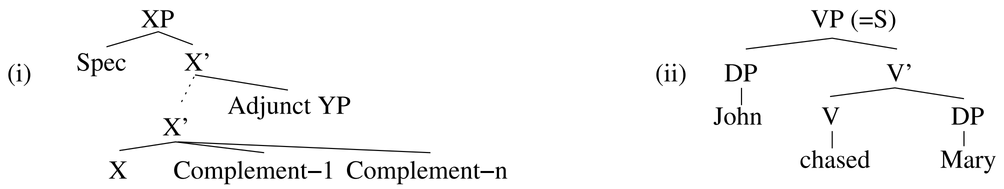

Figure 1: Diagram of X-bar phrase structure, and an example for English

Such words are sometimes needed in a sentence to satisfy a syntactic rule; for instance, the English copula is commonly analysed as having no semantic contribution, but is required because in English all sentences must have a verb. Particles which are syntactically optional can also be given null semantics; the Māori particle *ai* is an example of this (see Section 6 for details).

Non-lexicalist elements of the grammar In certain other cases semantics are required which are not added by the words but rather are added by a rule. Answer sentences provide a good example of this type of rule. *It was the cat* is semantically identical with its shortened form *the cat*. Obviously the words in the second example cannot contain all the neccesary semantics to make the two sentences semantically identical. Instead, this additional semantics is contributed by the syntactic rule which allows a simple Determiner Phrase (DP) to function as a full sentence. Again, this departure from purely lexicalist semantics is supported by LKB.

4.2 Semantic granularity

As is frequently the case for any pair of languages, there are semantic concepts encoded in the English lexicon and morphosyntax which are not encoded in that of Māori, and vice versa. For instance, the English system of auxiliary verbs permits a finer-grained encoding of tense and aspectual information than is possible in Māori. Conversely, the Māori pronoun system is much richer than that of English, distinguishing between inclusive and exclusive *we*, and between dual and plural pronouns. The LKB system allows semantic information to be expressed within a typed feature hierarchy, which provides a good way of encoding different levels of semantic granularity. Our policy to deal with such discrepancies is firstly to assume that the semantic representation which serves as input to the generator is as detailed as possible, and secondly to make it possible for a generated sentence to express this input at a coarser level of detail if the constraints of the language require this. LKB does not currently handle this kind of climbing of the semantic hierachy, but it is a facility we would like to add at some point.

5 Key Syntactic Differences Between English and Māori

In this section, we review the main syntactic differences between English and Māori which we have focussed on. (For a comprehensive introduction to Māori syntax, see Bauer, 1997.)

VSO order versus SVO order The VSO order of Māori sentences makes them structurally quite different from their English counterparts. The first decision that we had to make was how to handle complements and specifiers of verb phrases in light of the VSO order.

Borsley (1995) suggests two alternative HPSG analyses for VSO languages, one used for examples with Syrian Arabic and the other

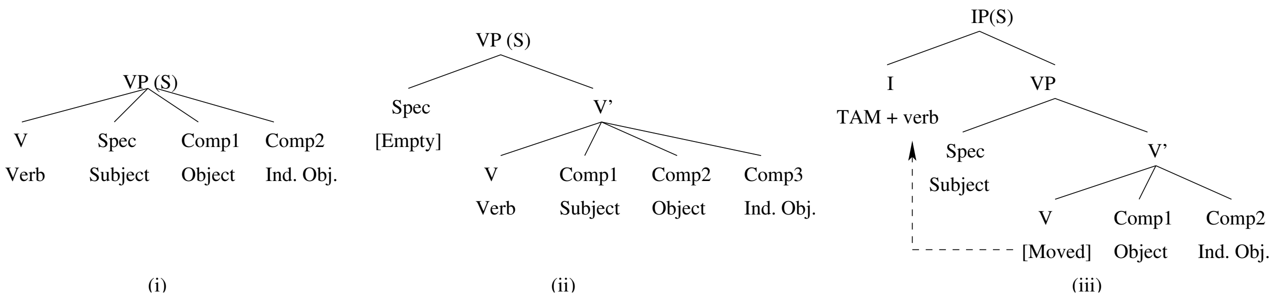

Figure 2: Three possible analyses of VSO order for Māori

from Welsh. Borsley's suggestion for Syrian is to have the subject still kept in the specifier category and to have a rule which discharges the specifier and the complements at the same level and places the specifier after the verb (see Figure 2 (i)). Borsley calls this rule the head-subject-complement rule. The alternative Borsley presents for Welsh is to have the subject stored as the first item in the list of complements to the verb. Thus the subject and object still get discharged at the same level but the specifier of the verb phrase remains empty and is not used (see Figure 2 (ii)). A third alternative, within a GB framework, is that the parse tree for VSO Māori is much the same as SVO English except that the head of the verb phrase (the verb) is moved out from its position to adjoin to the head of an Inflectional Phrase at the next level up (see Figure 2 (iii)). This analysis is due to Pearce (1997).

We decided to adopt Borsley's treatment of VSO order for Welsh (Figure 2 (ii)) for Māori. This option has the dual advantages of keeping the same HPSG order of features as for English (specifier - head - complements) and also avoids using movement (or in our case, gapping).

Tense and Aspect Markers A second key feature of Māori which contrasts significantly with English is the use of a separate word to give the tense and aspect of a sentence. Tense and aspect markers (TAMs) occur at the start of most sentences. They convey the information that would, in English, be carried by a combination of verb inflections and auxiliary verbs. While for English the notion of an IP goes against our principles of avoiding empty functional projections and movement, in Māori the IP analysis does not violate these principles at all. We can place the TAM at the head of an inflectional phrase, introducing the VP as its complement (see Figure 3 (i)). In summary, our Māori grammar contrasts with our English grammar as regards the general structure of a sentence, which has a verbal projection as its highest constituent (see Figure 3 (ii)). In our model, therefore, we do not adhere to the principle that sentences in different languages share the same underlying syntactic structure.

Non-verbal Sentences Another characteristic of Māori is the existence of sentences which do not contain a main verb. These sentences correspond (in all cases we have encountered) to English sentences involving the copula. Like verbal sentences, the non-verbal sentences begin with a particle carrying tense information. These particles have a function and syntactic position closely related to the verbal TAMs, and are included in the same category in Te Kaitito's grammar. We have so far covered three types of nonverbal sentence types. See Section 6 for specific examples of our coverage of non-verbal sentences.

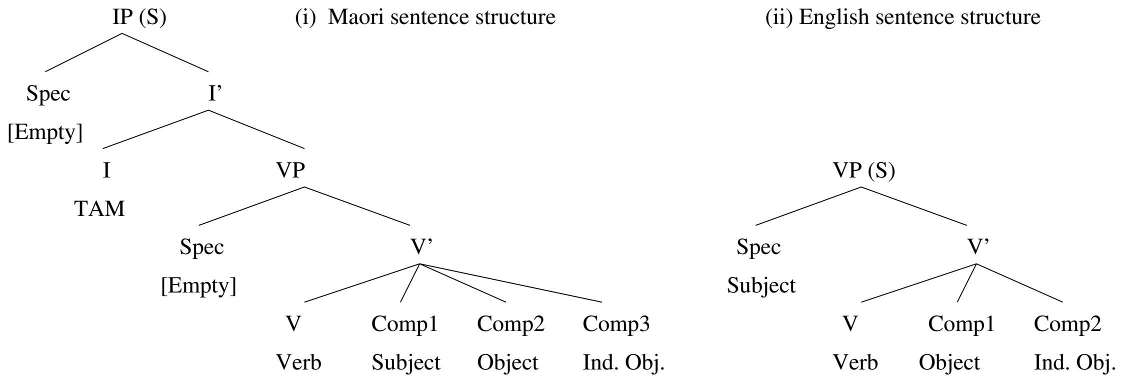

Figure 3: Māori sentence structure compared with English

The passive The passive is rather different in Māori than in English. The use of the passive mood is much more common in Māori and English active sentences are frequently translated to passive form in Māori. In some form of deference to this, Te Kaitito at the moment makes no semantic differentiation between active and passive in either language. Thus when a sentence is paraphrased or generated the passive and active equivalents are generated in both languages. Eventually we plan to accommodate the frequent translation of Māori passive to English active by using a semantic feature to distinguish between theme and rheme, and allowing Māori sentences to be unspecified for this feature.

6 Some constructions covered in the Māori grammar

A range of Māori grammatical constructions which the grammar covers are discussed below. (In each case, our grammar also covers the corresponding English constructions.)

Verbal sentences Intransitive, transitive and ditransitive verbal sentences are analysed using exactly the structure given in Figure 3 (i) above. Examples of each type of sentence are given below.

(1)

The man was eating			
I te	kai	te	tangata
TAM	eat	the	man.

(2)

The man is chasing the cat						
Kei te	whai	te	tangata	i	te	ngeru
TAM	chase	the	man	OBJ	the	cat.

(3)

The man gave the dog the ice cream									
I	hoatu	te	tangata	i	te	aihikirimi	ki	te	kurī.
TAM	give	the	man	OBJ	the	ice cream	IND OBJ	the	dog.

Both the subject and object(s) in Māori sentences are treated as complements. The system treats the Māori case markers of *i* (object) and *ki* (indirect object) as modifiers to the determiner phrase.

Predicative Sentences These sentences duplicate a number of the functions of the copula in English. They can be used to attribute a property to an object, either an adjectival-type property ('redness', 'bigness' etc) or a nominal property (e.g. being a dog) to an entity. An example of this type of sentence is shown in Example 4.

(4)

The house is big			
He	nui	te	whare
TAM	big	the	house

The parse trees for sentences like this one are also closely based on Figure 3 (i), except they contain an 'attribute phrase' (headed by an adjective or a DP functioning as a predicate) instead of a verb phrase headed by a verb.

Locative Sentences There are two forms of Māori sentences which correspond to the English form *X is in/at/on Y*. The form used in Māori for *at* is different from the form used for other words of location. Examples of each form are given below.

(5)

The monkey is at the house				
Kei	te	whare	te	makimaki
TAM	the	house	the	monkey

(6)

The girl is in the house						
Kei	roto	te	kotiro	i	te	whare
TAM	in	the	girl	OBJ	the	house

The parse trees for these sentences are given in Figures 4 and 5. These two sentences look almost identical in English but have somewhat different structures in Māori. In Māori the two sentences both use the same range of TAMs, but the canonical order of the arguments is reversed in Example 5 and the object (*whare*) lacks the case marker (*i*).

Given our assumption that a Māori sentence and its English translation have the same semantic representation, these sentences pose an interesting problem: how is the semantics of the English word *at* expressed in the Māori sentence in Example 5? We want to assume the TAM *kei* is the same for both types of locative sentence. The 'at' semantics of the second sentence, therefore, has to be added not by any lexical item, nor by any morphology (morphology of any kind is very rare in Māori), but by the syntactic rule which allows the TAM to introduce a pair of DPs. Our grammar is, as a result of tactics like this, not 'purely' lexicalist.

Yes/No Questions Māori yes/no questions are only distinguished from assertions by intonation or punctuation (or domain and context knowledge). To sidestep the need for inference in distinguishing between yes/no questions and assertions, our grammar requires yes/no questions to be terminated with a question mark (the one piece of punctuation in the grammar so far). An example of a yes/no question is given below.

(7)

Is the dog barking?				
Kei te	auau	te	kurī	qmark
TAM	bark	the	dog	qmark

Wh-questions We have Māori equivalents for the English *who*, *which*, and *what*. In Māori questioned NPs are sometimes fronted but need not always be. *Tēhea* is the Māori determiner used for *which*. It operates very similarly to the English determiner *which* but does not need to be fronted when in the object position. Whether it can be left un-fronted when questioning a subject is an issue we still need to resolve. At the moment our system does not support fronting in Māori questions so the un-fronted version of all questions have been left as acceptable. This will most likely have to be changed. An example is shown below.

(8)

Which dog ate the ice cream						
i	kai	tēhea	kurī	i	te	aihikirimi
TAM	eat	which	dog	OBJ	the	ice cream

Wai is the Māori equivalent of *who*. Like *tēhea* it need not always be fronted (especially if it is the object being questioned). In a similar manner to the English *who*, *wai* operates as a determiner phrase by itself; see Example 9.

(9)

Who was chasing the dog						
I te	whai	a	wai	i	te	kurī
TAM	chase	{sc det	who	OBJ	the	dog

Aha is used in the same way as *wai* but with the ordinary noun determiner *te* in front. It is the Māori equivalent of *what* and is used for non-person questioning. It can also be used with *he* but this use is not yet supported by Te Kaitito. Example 10 demonstrates a sentence using *aha*.

(10)

What is the man eating?						
Kei te	kai	te	tangata	i	te	aha
TAM	eat	the	man	OBJ	the	what

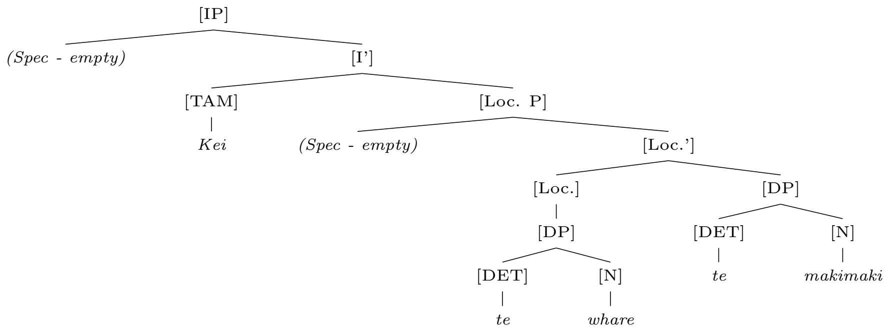

Figure 4: Parse of *Kei te whare te makimaki.*

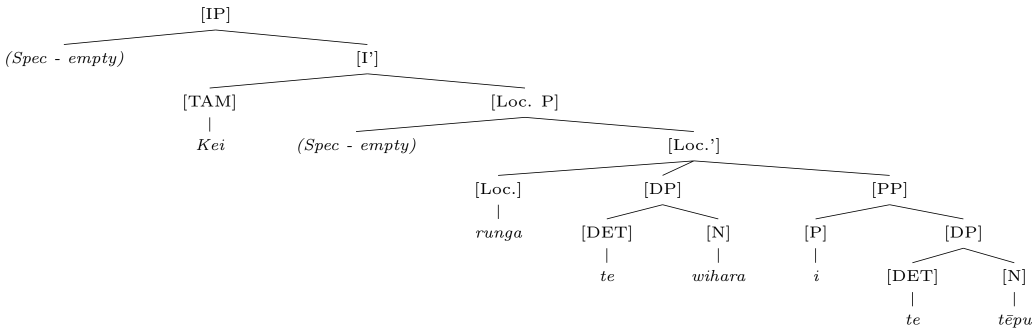

Figure 5: Parse of *Kei runga te wihara i te tēpu.*

Actor Emphatic answers These sentences correspond to English clefted sentences. They come in two slightly different forms, one using *ko* (Example 11) and one using *nā* (Example 12).

(11)

It is the cat which chased the mouse							
Ko	te	ngeru	i	whai	i	te	kiore
TAM	the	cat	TAM	chase	OBJ	the	mouse

(12)

It was the cat which chased the mouse						
Nā	te	ngeru	i	whai	te	kiore
TAM	the	cat	TAM	chase	the	mouse

Parse trees for these sentences are shown in Figures 6 and 7. The most noticeable difference between the two is the lack of an accusative case marker in the *nā...* sentences. We believe that this is an indicator of a deeper structural difference between the two forms; the verbs in this type of sentence are treated as passive even though they appear active in form, following Bauer (1997: ch 33). In addition to the full forms of the actor emphatic answers Te Kaitito also allows the shortened answers of the form *nā te ngeru/ko te ngeru.* Examples of the two forms of actor emphatic so far implemented as well as a parse tree are given below. Note that we have arbitrarily assigned the sentences using *ko* to the present tense, at least until we implement a mechanism to allow the sentence generator to select a higher level of semantic granularity if necessary (as mentioned in Section 4.2).

Relative Clauses Māori relative clauses are handled in a similar manner to English ones; they are treated as modifiers on the noun. However, relative clauses are structured very differently depending on which argument is extracted. At the moment we only handle relative clauses with a gap in subject position, because these are fairly similar to English relative clauses. To translate an English sentence with an object-gap rela-

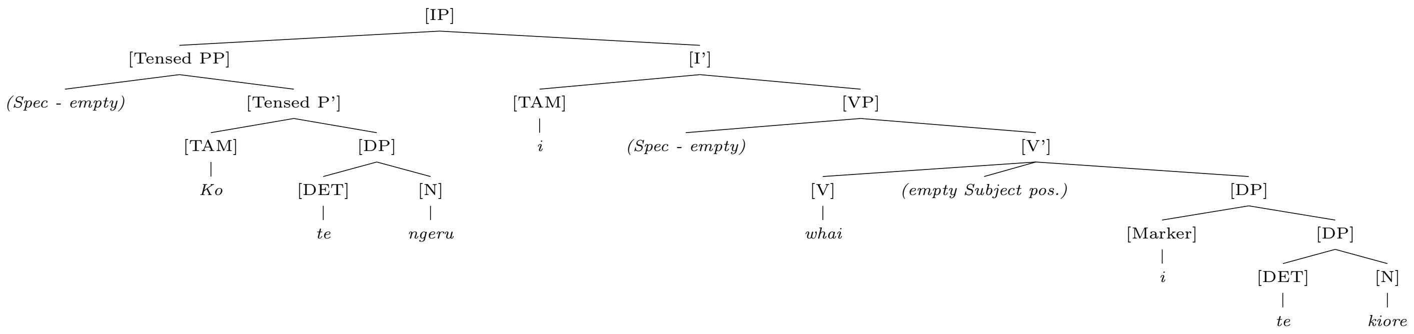

Figure 6: Parse of *Ko te ngeru i whai i te kiore*

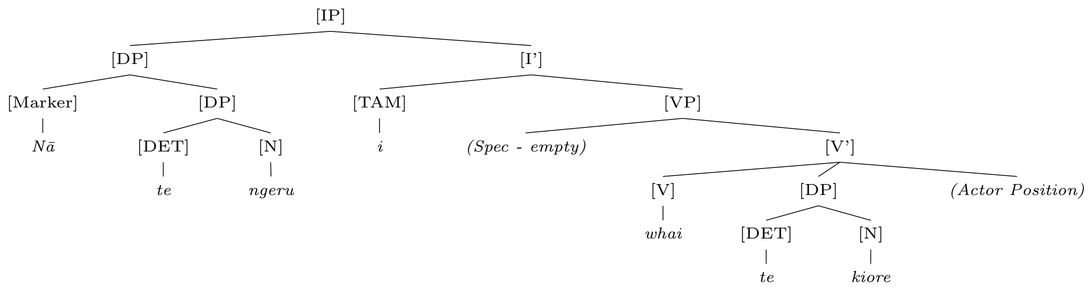

Figure 7: Parse of *Nā te ngeru i whai te kiore*

tive clause, the system is therefore currently obliged to passivise the relative clause. An example is given below.

(13)

The dog which sandy chased barked								
i	auau	te	kurī	i	whāia	e	a	sandy
TAM	bark	the	dog	TAM	bark	PASS	DET	sandy

7 Summary

This paper has presented an overview of the syntactic and semantic treatment of Māori currently implemented in Te Kaitito. Although we are still at an early stage in development of a useful grammar, HPSG is so far extending fairly well to the Māori constructions we have investigated.

References

Bauer, W. (1997). *The Reed reference grammar of Māori*. Reed books, Auckland, NZ.

Borsley, R. (1995). On some similarities and differences between Welsh and Syrian Arabic. *Linguistics*, **33**.

Copestake, A. (2000). The (new) LKB system. CSLI, Stanford University.

Copestake, A., Flickinger, D., Malouf, R., Riehemann, S., and Sag, I. (1995). Translation using Minimal Recursion Semantics. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium.

Copestake, A., Lascarides, A., and Flickinger, D. (2001). An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France.

Knott, A., Bayard, I., de Jager, S., and Wright, N. (2002). An architecture for bilingual and bidirectional nlp. Paper submitted to ANLP.

Leslie, N. (1998). Towards a formal grammar for māori. In *Proceedings of the NZ Asia conference*, Palmerston North, NZ.

Pearce, E. (1997). Negation and indefinites in Māori. *Current Issues in linguistic Theory*, **155**.

Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.