Constituent-Based Morphological Parsing: A New Approach to the Problem of Word-Recognition.

Richard Sproat
Linguistics Department
AT&T Bell Laboratories
600 Mountain Ave
Murray Hill, NJ 07974.

*Barbara Brunson**
AT&T Bell Laboratories
and
Department of Linguistics
University of Toronto
Toronto, Ontario, Canada M5S 1A1.

Abstract

We present a model of morphological processing which directly encodes prosodic constituency, a notion which is clearly crucial in many widespread morphological processes. The model has been implemented for the Australian language Warlpiri and has been successfully interfaced with a syntactic parser for that language (Brunson, 1986). We contrast our approach with approaches to morphological parsing in the KIMMO framework.

1. Introduction

The "Two-Level" Model of morphological processing developed by Kimmo Koskenniemi (1983), henceforth KIMMO, has spawned much subsequent research in the same framework (Karttunen, 1983; inter alia). Important design features of this model include a set of morpheme lexicons and a set of parallel finite state transducers which implement phonological rules mapping surface strings to lexical representations. Not only are phonological rules finite state, but the control structure of the model is itself finite state.

Two criticisms of this model can be put forth. First, KIMMO is not guaranteed to be computationally efficient (Barton, 1986). Second, there are many interesting morphological phenomena that KIMMO cannot cover without significantly redesigning the model. In this paper we will address the second point. We will present a model of word-structure recognition which, unlike the KIMMO model, makes heavy use of prosodic constituent structure. Not only is reference to prosodic constituency necessary to provide a principled way of dealing with certain morphological processes, but such an approach to phonological processing is crucial for any interface of current parsing systems with speech recognition systems (Church, 1983). The model has been implemented for the Australian language Warlpiri. We will describe how the parser works, and how it handles morphological phenomena that would, at best, require inelegant mechanisms within the KIMMO model. We will also show how we can handle morphological phenomena that are not exemplified in Warlpiri but which are of a similar ilk.

2. Two Facts about Morphology

We will now consider two issues in morphology, namely prosody and the non-isomorphism of syntactic and phonological structure. We maintain that these are are central to the task of a morphological analyzer and, hence, have incorporated them into our model.

2.1 The Relevance of Prosody to Morphology

It has become increasingly evident from research within Generative Linguistics that morphology cannot be limited to the concatenation and subsequent modification of strings of segments, but must recognize prosodic constituents devoid of segmental content (McCarthy, 1979; Levin, 1985). Work on reduplication[1] by Marantz (1982) and by Levin (1985) has argued convincingly that reduplication involves the prefixation or suffixation of a prosodic constituent which is empty of segmental information but which receives segmental specification by copying the segmental melody from the base. Furthermore, it has been suggested that infixation[2] must be viewed as prefixation or suffixation of an affix to a prescribed prosodic subconstituent of a word rather than to the whole word.

All of this work argues that prosody is a crucial component of morphology. It is necessary, therefore, that morphological processing systems should have a mechanism for dealing with prosody in a general way. KIMMO does not provide such a mechanism. Instead, it assumes that the problem of morphological recognition is one of matching some input *string* to a set of lexical *strings*. Prosodic considerations do not even enter the picture. The KIMMO model probably could be extended in various ways to cover such phenomena, but such extensions would constitute a significant change in the theory. Reduplication would require a particularly significant revision since it both involves reference to prosodic structure as well as a copy mechanism which is not finite state in any interesting sense. Note that although reduplication is strictly speaking bounded by the maximal size of some well-defined prosodic unit, and hence is *effectively* finite state, finite state recognition for reduplication would require the anticipation — i.e., precompilation — of all possible reduplicative-affix/stem sequences. Reduplication in natural language involves recognition of the language *ww*, a language which is well known not to be regular. As we shall see, reduplication is handled in our model by directly encoding prosody, and allowing for a bounded matching mechanism.

2.2 The Non-Isomorphism of Morphophonology and Morphosyntax

Another fundamental property of morphology is the fact that the structure required for the phonology is not necessarily isomorphic to the structure required for the morphosyntax. This point has been argued extensively in work such as Marantz (1984) and Sproat (1985). For example, in Warlpiri a number of clitics which are suffixes as far as the phonology is concerned (i.e., they undergo Vowel Harmony[3] with the word to which they attach) are separate words from the point of view of the syntax. For instance, the auxiliary in Warlpiri tensed clauses generally occurs as the second syntactic constituent of the sentence; phonologically, however, it is part of the first constituent. This phenomenon is by no means limited to scattered examples in a few languages, but apparently represents a very important generalization about the interaction of phonology and syntax in the morphology — they operate over different, though related structures. We propose to capture this observation by making the syntactic module of the parser largely independent of the phonological module, as we shall outline below.

3. A Description of the Warlpiri Parsing System

The main reason for choosing Warlpiri for our test domain is that Warlpiri provides a sufficient number of interesting morphological and phonological phenomena — such as Vowel Harmony and reduplication — without having an overabundance of phonological rules (unlike Finnish which has roughly 20 rules in the KIMMO description). It is thus possible to build a system which has a reasonable coverage of the morphological and phonological processes evident in the language. At the same time, in order to cover the Warlpiri data the system must be designed to handle morphological processes whose description crucially depends upon prosodic constituency.

The task of the morphophonological parser is to find out where the word boundaries are and then where the morphemes are. It receives as input a stream of segments and a parallel stream of suprasegmental stress information.

The input streams may represent a single word or they may represent a sequence of words; in any case, no word or morpheme boundaries are provided in the input. The parser checks to see if a morpheme sequence can correspond to the input stream by verifying that the appropriate phonological rules apply in the appropriate domains. It then passes a 'flattened representation' of the morphological structure, consisting merely of the morphemes in their linear order with word boundaries, off to the syntactic parser.

The syntactic parser for Warlpiri which we have been using is due to Brunson (1986). This parser was designed to take as input a sequence of morphemes rather than a sequence of fully formed words as most syntactic parsers do. Such a parser embodies our belief that the the task of building a *syntactic* representation for words should be handled by the syntactic parser and not by a separate morphosyntactic parser. In this way clitics can readily be identified in their syntactic roles independent of their phonological constituency.

Let us now turn to a concrete example from Warlpiri and show how we parse the morphemes and pass on the 'flattened representation' to the syntactic parser.

4. Parsing the Morphophonology

We will take as an example for discussion the word /pangupangurnu/, which means 'dug repeatedly' and which is composed of the morphemes *Reduplication+pangi+rnu*, (*pangi* = 'dig', *rnu* = 'past') (Nash, 1980), where *Reduplication* is the verbal reduplication morpheme. Of interest in this example are regressive Vowel Harmony[4], and, of course, reduplication. The input consists of the stream of segments and a stream of stresses[5]:

```
p a n g u p a n g u r n u
  1         2
```

There is a question of course as to whether one could reliably derive stress information from connected speech input. Preliminary studies of Warlpiri intonation suggest that main word stress at least is extractable from acoustic input (see Figure 1). We presume, however, that other phonetic facts may also help determine the prosody; see Church (1983) for a method for determining English prosodic constituents from observable allophonic variation.

The first task is to find the prosodic constituents, i.e. to find where the syllables are, where the feet[6] are, and where the prosodic words are. The particular parsing algorithm we adopt is that of Church (1983), which is not left-to-right, but nothing hinges on this decision; indeed, as we point out below, we will ultimately want a left-to-right parsing algorithm so that the phonological and syntactic parsing can be interleaved. The prosody of Warlpiri is simple in that syllable types are limited and phonological words are reliably left-stressed. In the particular example, the parser will tell us that the syllables are /pa/, /ngu/, /pa/, /ngu/ and /rnu/ (the sequences ng and rn represent single segments), that the feet are /pangu/ and /pangurnu/ and that there is a single prosodic word, namely /pangupangurnu/.

Having done the prosody, we proceed to look up the morphemes which might plausibly comprise the word. Warlpiri quite generally requires that morphemes be syllabifiable strings. The only exceptions to this are suffixes which consist of the sequence [sonorant][stop][vowel], for example the imperfective auxiliary base *lpa*. We can therefore find all possible morphological decompositions for a word by checking all [sonorant][stop][vowel] sequences and all well-formed syllable sequences and seeing if the strings spanning them correspond to known morphemes.

Lexical lookup is complicated due to the fact that the surface string can differ from the underlying representation of the morpheme in several ways. This can come about by the application of phonological rules. We implement lexical access in such cases by hashing on underspecified feature representations. In Warlpiri the only complication of this sort involves rounding of high vowels: for example, lexical /i/ may surface as /i/ or /u/ depending upon the harmony context. In the verb root *pangi* will therefore match the input sequences /pangi/ and /pangu/.

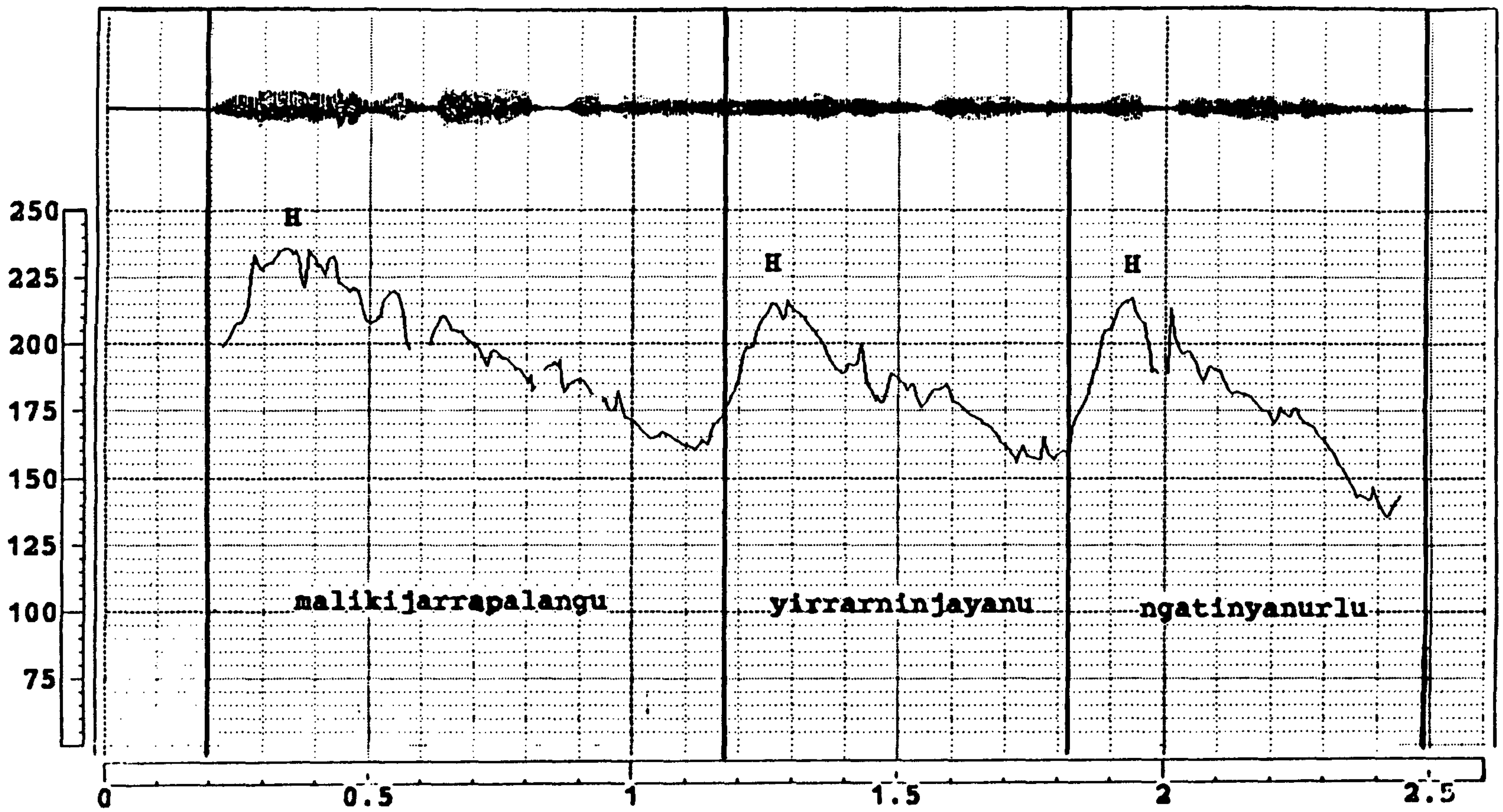


Figure 1

This pitch track shows word stress reliably positioned as a high target at the end of the first syllable of words. This sentence is from a tape provided by Mary Laughren of a Warlpiri speaker reading children's stories.

maliki-jarra-palangu	yirra-rninja-ya-nu	ngati-nyanu-rlu
dog-dual-aux(dual obj)	put-inf-go-past	mother-self's-erg

'Their mother was going along putting the two dogs.'

Another way in which the surface representation of a morpheme may differ from its underlying representation is if it does not contain any segmental information, but merely information about prosodic shape. This type of morphology manifests itself in Warlpiri as reduplication. Briefly, the verbal reduplicative prefix is listed as a bimoraic foot: i.e., a foot of the form CV(C)(C)V. Whenever we see such a constituent, we posit the existence of verbal reduplication subject to immediate verification if it matches the phonological material to its right. For Warlpiri, "matches" is "string equivalent to". For other languages, a more sophisticated notion of matching would be necessary. This would be necessary when phonological rules apply to only one part of the reduplicated pair. In /pangupangurnu/, the first sequence /pangu/ is a bimoraic foot, and furthermore it matches appropriately with the sequence to its right. Therefore we can here posit the existence of a verbal reduplicative affix.

Having found the possible morphemes, we have a lattice of morphemes spanning the input. In the example case, we have a lattice with a unique path comprising *Verbal-Reduplication*, *pangi*, *rnu*. We now wish to check that, from a phonological point of view alone, the affixes can be combined in the order given. That is, the affix path must be well-formed according to a morphophonological grammar for Warlpiri. We can state the morphophonological grammar simply as follows (where VHD stands for 'Vowel Harmony Domain'):

Word → (Prefix) VHD

VHD → [Root Suffix*] ∩ Vowel-Harmony

The first rule indicates that a word consists of an optional prefix followed by a Vowel-Harmony-Domain; the second claims that a Vowel-Harmony-Domain is a string analyzable as a root followed by some number of suffixes taken together with the Vowel Harmony process. We check the application of phonological rules, such as Vowel Harmony, by checking to see that the sequence of surface segments can be paired with the sequence of lexical segments in the underlying morphemes and that the surface string is well-formed according to the statement of the rules. This we do by a mechanism formally equivalent to the finite state transducer mechanism of the KIMMO model. In particular, we implement phonological rules as rejection sets (Koskenniemi, 1983), which are stated as regular expressions over the set of possible lexical/surface segment correspondences. However, in our model, phonological rules are defined for particular domains of application rather than continuously applying as in the KIMMO parser for Finnish. For example, Warlpiri Vowel Harmony is defined to apply over the sequence consisting of a root followed by its suffixes, but not over prefixes.[7]

Having established the identity of the morphemes of the word, and having further established that each potential morphological analysis is well-formed from a phonological point of view — i.e. the morphemes are in the right order and the relevant phonological rules have applied correctly over the appropriate domains — we then pass the morphological analysis off to the syntactic parser. More specifically, we pass off what we call a "flattened representation" which encodes only the information as to what order the morphemes occur in and where the word boundaries are. Arguably the syntactic parser does need to know where the phonological words and phrases are, but the fine details of the phonological structure are not needed. The potential non-isomorphism between phonological and syntactic structure is derived from the narrow bandwidth of the channel between the phonological and syntactic components of the parser. This non-isomorphism is illustrated when a morpheme which is phonologically an affix is syntactically a separate word — this is the case with cliticization.

Also exemplary of the division of duty between the morphophonological parser and the syntactic parser is the dual status of subcategorization in Warlpiri. For example, the ergative case suffix has two forms — /rlu/ and /ngku/. Both are subcategorized to occur with nominals, a fact that is crucial in the projection and selection of syntactic constituency. The choice between /rlu/ and /ngku/, on the other hand, is conditioned by subcategorization with respect to the prosodic structure of the stem — /ngku/ being restricted to bimoraic stems. This subcategorization is only an issue for the morphophonological parser, and is never even visible to the syntactic parser.

In Figure 2 we give an illustration of the behavior of the morphological and syntactic parsers on a more complicated example: *Ngarrka-ngku-ka marlu marna-kurra luwa-rnu ngarni-nja-kurra* (man-ergative-aux kangaroo grass-obj shoot-past eat-infinitive-obj) 'The man is shooting the kangaroo while it is eating grass.' This example illustrates a number of instances of phonological and syntactic mismatch.

5. Extensions and Improvements to the Current Work

The model proposed here, although designed and implemented for Warlpiri, is intended to be a general approach to morphological parsing. A number of extensions can easily be made and a number of design improvements are necessary.

First, reduplication, as we have noted, is only one of the kinds of morphology which are best defined in terms of prosodic constituents. The morphology of Arabic verbs (McCarthy, 1979) is another example of this, as is infixation. While Warlpiri does not exhibit these morphological processes, there would be no problem extending the parser to cover languages which do, since it is already designed to handle prosodically defined morphology.

Another problem which comes up in the current implementation is that the ordering of syntactic parsing after morphological parsing fails to identify syntactically ill-formed words as early as possible. To give a simple example from English, the string *analyz-iti-able* is arguably well-formed as far as the phonology is concerned, but is ill-formed syntactically since *-ity* attaches to adjectives, not to verbs, and *-able* attaches to adjectives, not to words ending in *-ity*, which are themselves invariably nouns. The current parsing system would discover that such a word was well-formed phonologically, only to realize that the word was in fact ill-formed when the syntax was reached. Needless to say, the solution is to interleave the phonological and syntactic analyses. Sequences like *analyz-iti-able* would then be detected early as ill-formed .

6. Summary

To summarize, we have built a morphological parsing system for Warlpiri which directly encodes prosodic notions and which also encodes the kind of non-isomorphy between phonological and syntactic representations exhibited in natural languages. We have argued that it is necessary for any general theory of morphological processing to encode these notions. We view the parsing system as a partial but general theory of morphological processing, and the work we have done on Warlpiri as a particular instantiation of this general model.

Acknowledgments

We would like to thank Mary Laughren and Ken Hale for their advice on Warlpiri.

Notes

* This work was partially supported by the *Social Sciences and Humanities Research Council of Canada.*

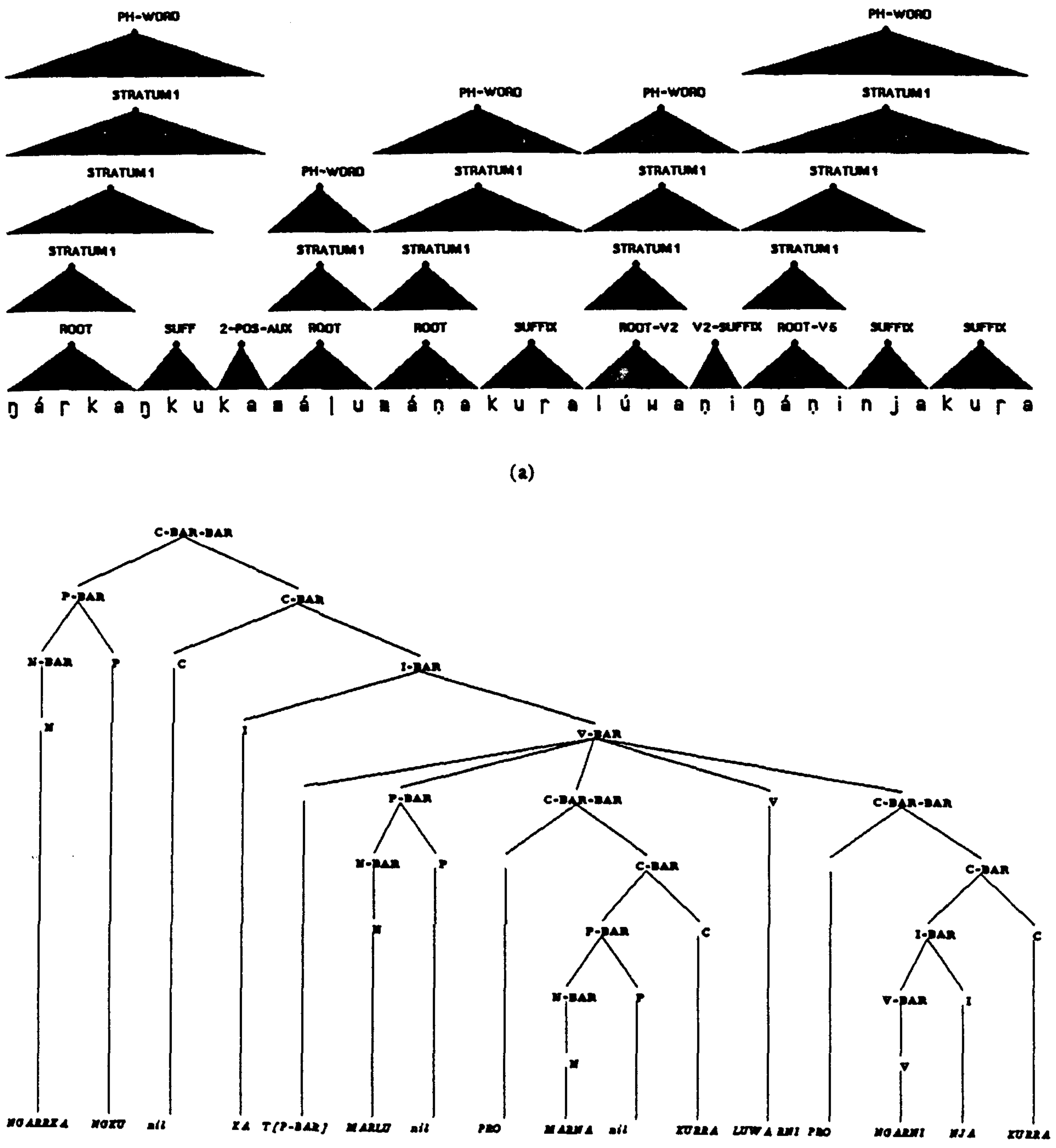
[1] Reduplication is a word formation process involving the repetition of a word or a part of a word. As an example, in Warlpiri there is a process of nominal reduplication to form the plural: *kurdu* 'child' — *kurdukurdu* 'children'.

[2] Infixation, like prefixation and suffixation, involves the attachment of an affix to a word; but, unlike these other two processes, an infixed affix occurs within the word rather than at the edge of the word.

[3] Vowel Harmony is a phonological process in which the vowels within a certain domain (usually a word) must agree in some set of features.

[4] The /i/ of the verb stem is changed due to the following /u/ of the past tense morpheme. This contrasts with /pangipangirni/ 'dig

Figure 2



(a)

(b)

Figure 2a is the phonological representation for the sentence:

ngarrka-ngku-ka marlu marna-kurra luwa-rnu ngarni-nja-kurra
'The man is shooting the kangaroo while it is eating grass.'

Figure 2b is the syntactic representation for that sentence. Note that the bracketing into phonological words is not isomorphic with the syntactic bracketing.

repeatedly, where the nonpast morpheme, *rni*, does not trigger such a stem change.

[5] Vowels bearing primary stress are aligned with 1, those bearing secondary stress are aligned with 2.

[6] A foot is a level of metrical structure intermediate between the syllable and the word.

[7] These domains correspond to the strata of Lexical Phonology (Kiparsky, 1982; Mohanan, 1982; inter alia).

References

Barton, E. (1986). "Computational Complexity in Two-Level Morphology." *Proceedings of the 24th Conference of the Association for Computational Linguistics*, 53-59, Columbia University, New York.

Brunson, B. (1986). *A Processing Model for Warlpiri Syntax and Implications for Linguistic Theory.* M.A. Thesis, University of Toronto, forthcoming as a TR of the Computer Science Department, University of Toronto.

Church, K. (1983). *Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints.* Ph.D. Thesis, MIT, published by IULC.

Karttunen, L. (1983). "KIMMO: A Two-Level Morphological Analyzer." *Texas Linguistic Forum*, 22, 165-186.

Kiparsky, P. (1982). "Lexical Phonology and Morphology." in *Linguistics in the Morning Calm*, Linguistic Society of Korea. Seoul: Hanshin.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* Ph.D. Thesis, University of Helsinki.

Levin, J. (1985). *A Metrical Theory of Syllabicity.* Ph.D. Thesis, MIT.

Marantz, A. (1982). "Re Reduplication." *Linguistic Inquiry.* 13(3): 435-482.

__________. (1984). *On the Nature of Grammatical Relations.* Cambridge, MA: MIT Press.

McCarthy, J. (1979). *Formal Problems in Semitic Phonology and Morphology.* Ph.D. Thesis, MIT, published by IULC.

Mohanan, K.P. (1982). *Lexical Phonology.* Ph.D. Thesis, MIT, published by IULC.

Nash, D. (1980). *Topics in Warlpiri Grammar.* Ph.D. Thesis, MIT.

Sproat, R. (1985). *On Deriving the Lexicon.* Ph.D. Thesis, MIT.