Nordic Journal of African Studies 14(4): 452–463 (2005)

Morphological Parsing of Tone
An Experiment with Two-Level Morphology on the Ha language

LOTTA HARJULA
University of Helsinki

Morphological parsers are typically developed for languages without contrastive tonal systems. Ha, a typical Bantu language of Western Tanzania, proposes a challenge to these parses with both lexical and grammatical pitch-accent that would, in order to describe the tonal phenomena, seem to require an approach with a separate level for the tones. However, since the Two-Level Morphology (Koskenniemi 1983) has proven successful with another Bantu language, Swahili (Hurskainen 1999), it is worth testing its possibilities with the tonally more challenging Bantu languages. The purpose of this paper is to show that morphological parsing of a fairly complex pitch-accent system is indeed possible with the Two-Level approach, but the solutions do not always describe the actual tonal system of the language.

Keywords: tone, Bantu languages, parsing

1. THE HA LANGUAGE

Ha[1] (Kiha being the Swahili form of the name and Igiha the Ha form of the name) is an Interlacustrine Bantu language spoken in Western Tanzania. It is not possible to state the exact number of Ha speakers since the language affiliation is not listed in the official census, but the estimation of a million speakers is probably not far from the truth. The speakers of the Ha language form the majority of the population in the three districts east and north from Lake Tanganyika: Kigoma, Kasulu, and Kibondo. Nowadays there are numerous Ha speakers also living in the other parts of Tanzania, as well as in Uganda, Kenya, Rwanda, Burundi, the Democratic Republic of Congo, and Zambia.

With its closest relatives, the Rundi language of Burundi and Nyarwanda language of Rwanda Ha forms a large language continuum, with possibly as many as 12 million speakers. These languages with some smaller languages are classified by Guthrie (1971) as group D60 of the Bantu language.

The orthography of the Ha language has not been standardised yet. An orthography workshop was held in the Ha area (in Kasulu) in June 2004, and the outcome of the workshop was an orthography proposal that is currently been tested with the speakers of the language. The proposed orthography more or less follows the linguistic choices (Harjula 2004), except that tone is not marked.

---

[1] The data cited in this paper was collected by myself over a period of several months in 1997, 2000, and 2003 in the Kigoma region.

Ha is a typical Bantu language with a noun class system and fairly complex verbal morphology, both inflectionally and derivationally. The tonal system of Ha has been analysed as a pitch-accent system (Harjula 2004), and according to this analysis there are two types of accents, i.e. lexical and grammatical accents. On the lexical level all the verbal roots fall into two tonal or accentual classes, one accented and the other accent-less. For the other word classes also the place of the accent has to be defined in the lexicon.

The lexical accent or the absence of the accent, together with the grammatical accents, define the grammatical forms of the verbs. The verb forms are called lexical tone-keeping forms if the tonal contrast of the verbal root is preserved and lexical tone-neutralising forms if the possible accent of the root is deleted. In both the lexical tone-keeping and tone-neutralising forms there may also be grammatical accents that are realised on the prefixes, or on the first or the second syllable of the stem. Both lexical and grammatical accents are realised as high tones (H) on certain tone-bearing units. An account of the verbal system can be found in Harjula (2004).

In addition to the grammatical accents of verbs, there are some grammatical elements, called index forms that have floating accents. The accents of these forms are realised as H either on the augment or on the first vowel of the stem of the following word. When the possible lexical accent of the noun stem falls on the syllable following the accent of the index, the vowel of the augment is lengthened.

2. Challenges for the Parsing of Tone

Although tone is not marked in the proposed official orthography of the Ha languages, tonal information produced in parsing is needed for disambiguation of lexical items and grammatical forms and further for example for producing tone marked texts for linguistic purposes. And as a language with both lexical and grammatical tone Ha serves as a good test case for morphological parsing of tone.

The morphological parser for Ha should be able to handle several different tonal phenomena: 1) the lexical accents, which are deleted or moved in some grammatical forms; 2) the grammatical accents and all their possible places of realisation; 3) the floating accents of the index forms; and 4) the vowel lengthening caused by the accent and compensatory lengthening.

Since the grammatical accents (see below) are usually not lexically associated with any single vowel but a morpheme and their realisation is determined by the complete segmental verb form, it is convenient and illustrative to describe the tonal level as a level of its own, as in Autosegmental Phonology (Goldsmith 1976). An autosegmental account of the Ha tonal system is found in Harjula (2004). However, a different approach is tested here.

2.1 Lexical Accents

The lexical tones are naturally marked in the lexicon. With the verbal roots it is enough to mark the whether the root has an accent or whether it has not (1), but with noun stems also the place of the accent has to be indicated in the lexicon (2).

(1)	-kund*[2]	'love'	a verb root with lexical accent
	-bon*	'see'	a verb root with lexical accent
	-gomb	'want'	a verb root without lexical accent
	-sek	'laugh'	a verb root without lexical accent
(2)	ukuguru[3]	'leg'	a noun without lexical accent
	ukubóko	'arm'	a noun with lexical accent on the first vowel of the stem
	umupfumú[4]	'doctor'	a noun with lexical accent on the second vowel of the stem

In some verbal forms the lexical accent is realised on the first vowel of the stem, but in some grammatical constructions the lexical accent seems to be moved. For example, the lexical accent of a verb stem is realised on the first vowel of the stem in the infinitive (*ku-*) and the consecutive form (*-ka-*) when there is no object prefix, but when an object prefix is present the accent is realised on the object prefix. Thus the accent is realised on the first mora of the macrostem, i.e. the stem with the possible object prefix.

(3)	kukúnda	'to love'	H realised on the first vowel of the stem
	tukakúnda[5]	'and we love'	H realised on the first vowel of the stem
	tukamúkunda	'and we love him/her'	H realised on the object prefix

Nominal lexical accents may also move. This takes place when the noun stem is monosyllabic and the noun class prefix disyllabic. The lexical accent of the noun stem is realised on the second syllable of the class prefix.

(4) umu-[6] bú → umúbu 'mosquito'

2.2 Grammatical Accents of Verbs

In majority of the verbal forms the possible lexical accent is deleted altogether, e.g. in the future *-roo-* (5) and in the present (6), no tense marking, and either the form does not have an accent at all or there is an accent/accents of the verbal

---

2 * indicates an accent that is associated with the morpheme, not with any certain vowel, before the application of the rules

3 uku- is the class prefix for noun class 15

4 umu- is the class prefix for noun class 1

5 tu- is the subject marker for the first person plural 'we'

6 umu- is the class prefix of class 3

form, e.g. remote past *-ára-* and an accent on the first mora of the macrostem (7). These so called grammatical accents come with or without a segmental marker of the verb form.

(5)	turookunda	'we will love'	no grammatical accent, lexical accent deleted
	turoogomba	'we will want'	
(6)	tukunda	'we love'	no grammatical accent, lexical accent deleted
	tugomba	'we want'	
(7)	twaárakúnda[7]	'we loved'	two grammatical accents, lexical accent deleted
	twaáragómba	'we wanted'	

Some verbal forms are differentiated only by the grammatical accent. For example, in the present tense (see 6 above) the lexical accent is deleted and there is neither grammatical accent nor segmental marking of the tense. When the same form is put into relative, the only difference is a grammatical accent on the second mora of the macrostem (8). In the participial the grammatical accent falls on the subject marker (9). When the subject marker is only a vowel, as for example in class 1, the accent of the participial is realised on the following vowel, i.e. on the first vowel of the stem, the object prefix, or a tense marker (10).

(8)	tukundá	'we who love'
	tumukúnda	'we who love him/her'
	tukundána	'we who love each other'
	tugombá	'we who want'
	tumugómba	'we who want him/her'
(9)	túkunda	'if we love'
	túmukunda	'if we love him/her'
	túgomba	'if we want'
	túmugomba	'if we want him/her'
(10)	akúnda	'if he/she wants'
	amúkunda	'if he/she wants him/her'

Adjacent accents on sequential syllables are not allowed in Ha, except in certain verb forms. The rule is called Meeussen's Rule (Clements & Goldsmith 1984: 245). Meeussen's Rule deletes the right-most accent of the two adjacent accents. In the remote past, for example, the accent of the tense prefix *-ára-* (see 7 above) is moved to the second syllable of the tense prefix when the subject prefix is only a vowel (11). The second accent, due to fall on the syllable following the already accented syllable, is deleted.

(11) yarákunda 'he loved' (a+ára+kund+a)

[7] When the subject prefix *tu-* is followed by a vowel the initial consonant is labialised and the following vowel lengthened by compensatory lengthening. This rule is not dealt with in this paper.

2.3 Floating Nominal Accents and Vowel Lengthening

The analysis of nouns provides a challenge for morphological parsing when nouns are preceded by the index forms (associative index *na**, comparative index *nka** and presentative index *nga**) or the connexive *-a**[8]. If the following word has an initial vowel the vowel of the connexive or index form is deleted and the accent carried by the connexive or index form is realised on the initial vowel (12). If the following word does not have an initial vowel, the vowel is not deleted and the accent is realised on the first syllable of the following word (13). Also, if these words have a lexical accent on the second vowel it is deleted (Meeussen's Rule). When the following noun has an accent on the first mora of the stem and the class prefix is monosyllabic, the initial vowel is lengthened and the lexical accent is preserved (14).

(12)	umugabo 'man'	númugabo 'with the man'	(na*+umu+gabo)
	umugabo 'man'	yúmugabo 'of the man'	(i+a*+umu+gabo)
	ingoga 'hurry'	níngoga 'with hurry'	(na*+in+goga)
(13)	weéne 'it'	wawéene 'of it'	(u+a*+weéne)
(14)	inyáabu 'cat'	níinyáabu 'with the cat'	(na*+i+nyáabu)

3. Designing the Parser for Tone in Ha

Since non-linguistic texts in Ha would not be tone-marked but tone is crucial for the linguistic analysis of the Ha language two sets of rules were designed in this test: the first set of rules handles text with tonal marking (Rules 1) and the other set of rules that handles text not marked for tone (Rules 2, not discussed in this paper). The lexical tones are marked in the lexicon. The output provides morphological analysis both of the segmental elements as well as of the tones. Syntactic disambiguation is needed with the forms that are not distinguishable only by morphology, e.g. some of the tenses. When the input text is not marked for tone, the analysis produces more ambiguity which, however, is mostly solvable by syntactic disambiguation.

3.1 Parsing Lexical Accents

The lexical accents of nouns and other word classes (except verbs) are marked in the lexicon with character H after the vowel on which they are realised.

(15) <ukuguru>

#uku+guru# CP15 15/6:'leg'

[8] * indicates a floating accent

<ukubóko>

#uku+boHko# CP15 15/6:'arm'

<umupfumú>

#umu+pfumuH# CP1 1/2:'doctor'

The lexical accent of the verbs is not necessarily realised on a vowel of the root itself. Therefore, the lexical accent of a verb is marked on the lexeme with an R (as in 'root'), and it is realised as 0 by default. Since writing a rule where a surface character does not have an equivalent in the lexical level is not desirable (e.g. 0:H) the first vowels of the root and the object prefix (the possible vowels on which the lexical accent may be realised) are lexically marked with a character Q marking a possible place of an accent. For the tenses where the lexical accent is realised on the first vowel of the macrostem, the rule states that the left-most Q is realised as H, and both R and the appropriate tense prefix on the lexical level are the context required.

(16) <tukakúnda>

#tu+ka+RkuQnd+a# SP1pp Cons VH:'love' FV

<tukamúkunda>

#tu+ka+muQ+RkuQnd+a# SP1pp Cons OP1 VH:'love' FV

With the monosyllabic noun stems the lexical accent can be written at the beginning of the stem instead of the normal position after the vowel it is realised on. This may not reflect the underlying connection of the accent, but it will produce a correct out put by the parser.

(17) <umúbu>

#umu+Hbu# CP3 3/4:'mosquito'

3.2 PARSING GRAMMATICAL ACCENTS

With the tenses where the lexical accent is not realised and there are no grammatical accents, e.g. the future no tonal rules are needed, since the lexical accent of the verb is realised as zero by default. The surface forms are identical with both rules.

(18) <turookunda>

#tu+roo+RkuQnd+a# SP:1pp Fut VH:'love' FV

<turoomukunda>

#tu+roo+muQ+RkuQnd+a# SP:1pp Fut OP1 VH:'love' FV

In the majority of the verbal forms the possible lexical accent is deleted and there is an accent/accents of the verbal form (e.g. remote past with *-ára-* and an accent on the first mora of the macrostem). These so called grammatical accents come with or without a segmental marker of the verb form. The grammatical accents that are realised on a certain vowel of the tense prefix are marked in the lexicon and realised as H by default (e.g. *-ára-* of the remote past), and the grammatical accents realised on the macrostem are assigned to the appropriate vowel by a rule. The accent position marker (Q) may be used to insert the grammatical accent on the first vowel of the macrostem.

(19) <twaárakúnda>

#tu+aHra+RkuQnd+a# SP:1pp RemPastFoc VH:'love' FV

<twaáramúkunda>

#tu+aHra+muQ+RkuQnd+a# SP:1pp RemPastFoc OP1 VH:'love' FV

<twaáragómba>

#tu+aHra+goQmb+a# SP:1pp RemPastFoc VL:'want' FV

<twaáramúgomba>

#tu+aHra+muQ+goQmb+a# SP:1pp RemPastFoc OP1 VL:'want' FV

In the present tense (see 6 above) the lexical accent is deleted and there is neither grammatical accent nor segmental marking of the tense, thus no rules are needed (20). However, when present tense form occurs as a relative clause, it has a grammatical accent on the second mora of the macrostem, i.e. on the first or second vowel of the stem, depending on the presence or absence of the object prefix. When the accents falls on the first vowel of the stem (i.e. when there is an object prefix), the accent position marker Q may again be used. But for the forms without an object prefix another position marker is needed (M), realised as zero by default. With monomoraic macrostems the accent falls on the final vowel (*-a*) or the perfective marker (*-ye*), but with longer stems the accented vowel would be the second vowel from the left. The macrostems of the verbal forms with an accent on the second vowel of the macrostem are prefixed with symbol $, which is then used as a condition for the accent to be realised on the second vowel (21).

(20) <tukunda>

#tu+RkuQnd+aN# SP:1pp Pres VH:'love' FV

<tumukunda>

#tu+muQ+RkuQnd+a# SP:1pp Pres OP1 VH:'love' FV

<tugomba>

#tu+goQmb+a# SP:1pp Pres VL:'want' FV

<tumugomba>

#tu+muQ+goQmb+a# SP:1pp Pres OP1 VL:'want' FV

(21) <tukundá>

#tu+$RkuQnd+aM# SP:1pp PresRel VH:'love' FV

<tumukúnda>

#tu+$mu+RkuQnd+aM# SP:1pp PresRel OP1 VH:'love' FV

<tukundána>

#tu+$RkuQnd+anM+a# SP:1pp PresRel VH:'love' Rec FV

<tugombá>

#tu+$goQmb+aM# SP:1pp PresRel VL:'want' FV

<tumugómba>

#tu+$mu+goQmb+aM# SP:1pp PresRel OP1 VL:'want' FV

In the present participial (as well as in most of the other participial forms, as well) the grammatical accent falls on the subject prefix. Since there is a limited and reasonable number of subject prefixes, the prefixes for the participial forms are listed separately in the lexicon, with a lexical accent (P) that is realised as H by default.

(22) <túkunda>

#tuP+RkuQnd+a# SP:1pp PresPart VH:'love' FV

<túmukunda>

#tuP+muQ+RkuQnd+a# SP:1pp PresPart OP1 VH:'love' FV

<túgomba>

#tuP+goQmb+a# SP:1pp PresPart VL:'want' FV

<túmugomba>

#tuP+muQ+goQmb+a# SP:1pp PresPart OP1 VL:'want' FV

However, when the subject prefix is only a vowel or a nasal, as for example in class 1 (*a-*), the accent of the participial is realised on the following vowel. With these forms, there is a rule that turns P into zero (P:0) and another rule that matches a surface accent with the initial accent position marker of the macrostem (Q:H). The possible contexts of the rule are the appropriate subject prefixes (*a-*, *i-*, *u-* and *n-*).

(23) <akúnda>

#aP+RkuQnd+a# SP1 PresPart VH:'love' FV

<amúkunda>

#aP+muQ+RkuQnd+a# SP1 PresPart OP1 VH:'love' FV

When the participial accent removed from the subject prefix falls on a tense marker instead of the macrostem, e.g. in the persistive participial (*-ki-*), the accent of the subject prefixes (P) matches zero on the surface level (by a rule) and, in the lexicon, these certain subject prefixes lead to another lexicon where the tense marker lexically carries an accent. This solution does not reflect the fact that the accent is lexically not associated with the tense marker, but with it some possibly overlapping rules are avoided.

(24) <akíkunda>

#aP+kiH+RkuQnd+a# SP1 PersPart VH:'love' FV

<akígomba>

#aP+kiH+goQmb+a# SP1 PersPart VL:'want' FV

In the remote past with a short subject prefix (see 11 above) the replacement of the accent of the tense marker *-ára-* is dealt with a rule (H:0), where the context is the appropriate subject prefixes and the tense marker. Since the accent is realised on another vowel a position marker (B) for this accent is written in the lexicon, and another rule with the same context fills the position with the accent (B:H). The rule that makes the floating grammatical accent to be realised on the macrostem with the longer subject prefixes (see 19 above) does not apply here. Again, the solution does not describe the actual steps how the correct tonal realisation is formed (e.g. Meeussen's Rule is not represented), but it produces the correct output in parsing.

(25) <yarákunda>

#a+aHraB+RkuQnd+a# SP1 RemPastFoc VH:'love' FV

<yarámukunda>

#a+aHraB+muQ+RkuQnd+a# SP1 RemPastFoc OP1 VH:'love' FV

3.3 Parsing Floating Nominal Accents

When a noun is prefixed with an element with a floating accent, the vowel of an index form or the connexive is deleted, and the accent is realised on the initial vowel of the noun class prefix. Lexically the accent is marked on the index forms and the connexive with A, surfaced as zero by default, since the accent is never

realised on the vowel of an index form or the connexive. The place of realisation of this floating accent is marked on the noun class prefixes with T, another type of accent position marker. In the rule T:H the accent (A) serves as the context, together with the segmental part of the morpheme.

(26) <múmugabo>

#naA+uTmu+gabo# AI CP1 1/2:'man'

<yúmugabo>

#i+aA+uTmu+gabo# DP9 Con CP1 1/2:'man'

<níngoga>

#naA+iTn+goga# AI CP9 9/10:'hurry'

The words without initial vowel (mainly pronouns) need to have the accent position marker on the first vowel of the stem. In addition to the rule T:H, another rule that deletes the lexical accent is needed, if the two accents should fall on sequential vowels (H:0).

(27) <wawéene>

#u+aA+weTeHne# DP3 Con Pro1:'it'

With the noun class prefixes with which the accent of the index form or the connexive may cause vowel lengthening (in order to avoid the application of Meeussen's Rule) the position of the possible lengthening is marked in the lexicon. This marker is realised as a surface vowel if there is a floating accent (A) and a lexical accent on the first vowel of the noun stem. The only vowel that occurs in monosyllabic noun class prefixes is *i*. The floating accent is realised on the left-most vowel of the noun class prefix.

(28) <níinyáabu>

#naA+iTVn+nyaHabu# AI CP9 9/10:'cat'

4. CONCLUSIONS

This experiment shows that it is indeed possible to morphologically parse a language with both lexical and grammatical accents with Two-Level Morphology. The basic idea is to mark the possible vowels on which the grammatical or replaced lexical accents may fall in the lexicon, and write rules that allow the positions to be realised as surface accents when the appropriate tense marker or other segmental element is present. When there is no segmental but only tonal marking of a tense, the macrostem is lexically prefixed with a symbol that can be used as a contex in the rules.

However, the parsing formalism presented here does not always describe the tonal phenomena that are found in the Ha language. With some type of accents the lexixal accents can be mapped directly to the surface realisations, but with others the interaction of the accents causes changes on the segmental level, or the morphophonemic changes of the segmental level affect the realisations of the accents. Thus, for proper description of the language, a formalism which would allow the tones or accents to be mapped with the segmental level only after certain rules have applied in the two levels separately, is required.

REFERENCES

Clements, G.N. and Goldsmith, J. 1984.
Autosegmental Studies in Bantu Tone. Dortrecht: Foris.

Goldsmith, J.A. 1976.
Autosegmental Phonology. Bloomington, Ind.: IULC.

Guthrie, M. 1971.
Comparative Bantu. Vol. I-IV. Farnborough: Gregg International.

Harjula, L. 2004.
The Ha Language of Tanzania: Grammar, Texts, Vocabulary. Cologne: Rüdiger Köppe Verlag.

Hurskainen, A. 1999.
SALAMA: Swahili Language Manager. Nordic Journal of African Studies 8(2): 139-157.

Koskenniemi, K. 1983.
Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications of the Department of General Linguistics, No. 11, University of Helsinki.

ABBREVIATIONS

AI	associative index
Con	connexive
Cons	consecutive
CP4	noun class prefix of class 4
DP4	determiner prefix of class 4
Foc	verbal focus
Fut	future
FV	final vowel
OP4	object prefix of class 4
Part	participial
Pres	present

Pro7	noun class 7 substitutive pronoun
Pp	person plural
Rec	reciprocal extension
Rel	relative
RemPast	remote past
SP4	subject prefix of class 4
VH	verb root with lexical accent
VL	verb root without lexical accent