*Nordic Journal of African Studies 13(1): 95–102 (2004)*

# HANDLING TONE IN TWO-LEVEL MORPHOLOGY: THE CASE OF HA*

LOTTA HARJULA
*University of Helsinki, Finland*

## ABSTRACT

The problem of handling tone in morphological parsing has not yet been widely addressed. The morphological parsers are typically developed for languages without contrastive tonal systems, and the application of such parsers need to be adapted to handle tonal phenomena. The major problem, especially with African tone languages, is the fact that the tonal features often function on a level separate from the segmental level and would thus require a level of description of their own. The purpose of this paper, however, is to show that the Two-Level approach is sufficient to handle at least the pitch-accent type system of tone in Ha.

*Keywords: morphological parsing, tone, Two-Level Morphology*

## 1. THE HA LANGUAGE AND THE PROBLEM OF THE PARSING TONE

Ha (also Kiha or Igiha) is a Bantu language spoken by nearly a million people in Western Tanzania. It is one of the 132 languages spoken in Tanzania as a whole (Grimes & Grimes 2000). 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.

The Ethnologue (Grimes & Grimes 2000) lists the number of Ha speakers as 800 000, but the current number is difficult to estimate. The population of the Kigoma region has more than doubled between the censuses of 1988 and 2002 (from 860,000 to 1,680,000), and part of the refugees from Burundi have merged into the Ha community. There are still ca. 400,000 refugees in the region (compared with the 500,000 in the countryas a whole; figures ReliefWeb).

The Ha language is closely related to the Rundi of Burundi and Nyarwanda of Rwanda. Guthrie (1971) has classified Ha as a group D60 Bantu language (D66), together with Rundi, Nyarwanda, Vinza, Hangaza and Shubi. In the

classification of Eastern Bantu languages made by Nurse and Philippson (in Hinnebusch et al. 1981: 10-11), Ha, along with the other five languages in Guthrie's D60, is grouped together as the West Highlands subgroup of the Interlacustrine group.

Ha is a typical Bantu language with 18 noun classes and complex verbal morphology. It is highly agglutinative. 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.

In my PhD dissertation (Harjula 2004) I have analysed the tonal system of Ha as a pitch-accent system, and according to this analysis there are two types of accents, i.e. lexical and grammatical accents. On the lexical level this means that all the verbal roots fall into two tonal classes, one accented and the other accent-less. For the other word classes the place of the accent also has to be defined in the lexicon. The lexical accents are realised as H tones on the mora with which they are lexically connected, unless they are moved or removed because of the grammatical accents. The phonetic changes are not taken into account here.

The lexical accents, together with the grammatical accents or the absence of the accent, define the grammatical forms of the verbs. The verb forms are called lexical tone-keeping forms if the tonal contrast of the verbal roots is preserved, and lexical tone-neutralising forms if the possible accent of the root is deleted. In both the lexical tone-keeping and -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. A complete account of the verbal system can be found in Harjula (2004).

In addition to the grammatical accents of verbs, there are some 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.

Thus the morphological parser should be able to handle several different tonal phenomena: 1) the lexical tones, which are deleted in some grammatical environments; 2) the grammatical tones and their different places of realisation; 3) the floating tones of the index forms; and 4) the vowel lengthening caused by the accent.

## 2. THE PROPOSED APPROACH

The first attempt to handle the tone in Ha was made by marking the lexical accents in the lexicon and by writing rules that would produce the correct input for the moving and floating grammatical tones. However, this approach soon turned out to be far too risky in the sense that it became almost impossible to control the effects of the overlapping rules. In addition, writing separate rules for

all the possible tone shifts or defining the environment uniquely enough turned out to be too messy.

The solution chosen in this experiment was to include as much tone information as possible in the lexicon and thus keep the rules simple. In addition to the lexical accents, the places for the possible grammatical accents also were defined in the lexicon.

The application was designed with two separate sets of rules, one of which can be used with tone-marked texts (herafter called Rules 1), while the other gives the underlying tones in the output for texts in which the tones are not marked (Rules 2). The same lexicon can be used for both types of analysis.

## 3. PARSING OF NOMINAL TONES[1]

The lexical accent of the noun stems is marked in the lexicon with H\, which follows the vowel with which it is connected. The noun class prefixes do not carry accents of their own. In Rules 1 the default surface value of H\ is H\, and thus the correct analysis is produced only if the accent is marked in the right place in the text analysed (1). The default value for H\ in Rules 2 is 0, and thus the place of the underlying accent is marked in the output only (2).

(1) "<urukwaH\avu>"
"#uru+kwaH\avu" CP11 11/14:'hare'
"<urukwaaH\vu>"
***

(2) "<urukwaavu>"
"#uru+kwaH\avu" CP11 11/14:'hare'

When the noun stem is monosyllabic the possible accent of the stem is realised on the second syllable of the noun class prefix. With these stems the accent (H\) is written at the beginning of the stem in the lexicon.

(3) "<umuH\bu>" (Rules 1)
"#umu+H\bu#" CP3 3/4:'mosquito'
"<umubu>" (Rules 2)
"#umu+H\bu#" CP3 3/4:'mosquito'

---

1 Abbreviations:

| | |
|---|---|
| AI | associative index |
| Con | connexive |
| CP | noun class prefix |
| DP | determiner prefix |

Thus, the analysis of accents on noun stems is rather straightforward. However, the tonal analysis of the nouns becomes more of a challenge when the nouns are preceded by the connexive *-a* or the index forms (associative index *na*, comparative index *nka* and presentative index *nga*). If the following word has an augment, the vowel of these elements is deleted and the accent they carry is realised on the augment (4). If the following word does not have an augment, the vowel is not deleted and the accent is realised on the first syllable of the following word (5). When the following noun has an accent on the first mora of the stem and the class prefix is monosyllabic, the augment vowel is lengthened (6).

(4) umugabo ‘man’ yúmugabo ‘of the man’
ingoga ‘hurry’ níngoga ‘with hurry’

(5) weéne ‘it’ wawéene ‘of it’

(6) inyáabu ‘cat’ níinyáabu ‘with the cat’

The target mora of the potentially shifting accent is marked in the lexicon with T\ after the vowel with which it may be connected, and the floating accent of the connexive and the index forms is marked in the lexicon with A\. The default value for T\ and A\ is 0 in both Rules 1 and Rules 2. The realisation of the accent is controlled by a rule which connects the underlying T\ with the surface H\ when A\ is present.

(7) "<umugabo>"
"#umu+gabo#" CP1 1/2:'man'
"<yuH\mugabo>" (Rules 1)
"#yi+aA\+uT\mu+gabo#" DP9 Con CP1 1/2:'man'
"<yumugabo>" (Rules 2)
"#yi+aA\+uT\mu+gabo#" DP9 Con CP1 1/2:'man'

All the augment vowels that can be lengthened with the connexive or the index forms are marked with V\ in the lexicon. In addition, as with the tone shift, there is a rule that connects V\ with a surface vowel when the other conditions prevail, i.e. there is an accent on the first syllable of the stem, a monosyllabic noun class prefix with an augment, and an index form or the connexive.

(8) "<niH\inyaH\abu>" (Rules 1)
"#naA\+iT\V\n+nyaH\abu#" AI CP9 9/10:'cat'
"<niinyaabu>" (Rules 2)
"#naA\+iT\V\n+nyaH\abu#" AI CP9 9/10:'cat'

## 4. PARSING OF VERBAL TONES[2]

Since almost any syllable of a verbal form may have a grammatical accent associated with it, it is insufficient, in the tonal parsing, to mark only the moras with which an accent may be connected, but the possible realisation slots of the accents have to be defined for each verbal form. In this proposed solution for parsing the grammatical accents the fact that the grammatical accents are connected with the segmental markers is exploited. Since in most instances the segmental markers of the verbal forms precede the stem, it is possible that the sublexicons for the tense markers lead to different stem sublexicons, according to the way in which the grammatical accents are realised. This means, for example, that in one sublexicon the possible place of the accent is the first mora of the stem, while another tense leads to a sublexicon in which the place of the accent is the second mora of the macrostem.

It is recognised that this solution is not very economical, since all of the verbal stems have to be listed in separate sublexicons several times, but the extension of the lexicon resulting from the multiple sublexicons does not challenge the performance of the automata. The analysis of the verbal forms is the same, regardless of the presence or absence of tone-marking in the input.[3]

## 5. LEXICAL TONE-KEEPING FORMS

In the lexical tone-keeping forms the lexical accent of the verb stem is realised on the first mora of the macrostem[4], as, for example, in the infinitive (*ku-*) and the present disjunct (*-ra-*) (9). In some of these tenses there is also an accent on the tense marker, as, for example, in the persistive (*-cháa-*) (10).

| | | | | |
|---|---|---|---|---|
| (9) | -bóna | ‘see’ | -seka | ‘laugh’ |
| | kubóna | ‘to see’ | kuseka | ‘to laugh’ |
| | arabóna | ‘he sees’ | araseka | ‘he laughs’ |
| | aramúbona | ‘he sees him’ | aramuseka | ‘he laughs at him’ |
| (10) | acháabóna | ‘he still sees’ | acháaseka | ‘he still laughs’ |

---

2 Abbreviations:

| | |
|---|---|
| FV | final vowel |
| OP | object prefix |
| SP | subject prefix |
| VH | accented stem |
| VL | accent-less stem |

3 With some of the parses multiple analyses are produced, but the lexemes are labelled in a manner that permits disambiguation. This does not, however, fall within the scope of this paper.

4 The macrostem consists of the stem, derivational extensions and possible object prefixes.

With these tenses the tense marker sublexicon leads to a sublexicon where the lexical accents are marked with H\ after the vowel with which they are lexically connected. When no object prefix is present, the parse outputs the stems with the possible lexical accent (11). However, when an object prefix precedes the stem, the possible lexical accent is realised on the object prefix. This shift is realised by marking the object prefixes in the sublexicon for lexical tone-keeping forms with L\ and by rules (12). These rules associate the underlying L\ with the surface H\ and the underlying H\ of the stem with 0 when there is an object prefix.

(11) "<araboH\na>"
"#a+ra+boH\n+a#" SP1 PresDisj VH:'see' FV
"<araseka>"
"#a+ra+sek+a#" SP1 PresDisj VL:'laugh' FV

(12) "<aramuH\bona>"
"#a+ra+muL\+boH\n+a#" SP1 PresDisj OP1 VH:'see' FV
"<aramuseka>"
"#a+ra+muL\+sek+a#" SP1 PresDisj OP1 VL:'laugh' FV

## 6. TONE-NEUTRALISING FORMS

In the present conjunct (-) and the recent past conjunct (*-a-*), for example, there are no tense accents and the lexical accent of the stem is neutralised. For these tenses, the stem sublexicon is constructed in such a manner that there will be no accents or places of accents marked for the lexemes, but the information of the lexical accent is stored only in the labels of the lexemes. The surface forms are the same for both types of analysis (tone-marked and toneless) (13), and the presence of an object prefix does not change the tonal realisation (14).

(13) "<abona>"
"#a+bon+a#" SP1 Pres VH:'see' FV
"<aseka>"
"#a+sek+a#" SP1 Pres VL:'laugh' FV

(14) "<amubona>"
"#a+mu+bon+a#" SP1 Pres OP1 VH:'see' FV
"<amuseka>"
"#a+mu+sek+a#" SP1 Pres OP1 VL:'laugh' FV

In some tenses there is an accent of the tense either on the tense marker or on the subject prefixes. In the case of the tenses where the tense accent is realised on the segmental tense marker, for example the remote past conjunct (*-á-*), the sublexicon of the tense leads to a sublexicon where there are no accents marked on the stem lexemes (15). When the tense accent is on the subject prefixes, as, for example, in the participial present (´-), the subject prefixes are in a separate sublexicon and marked with P\. The default surface realisation of P\ is H\ in Rules1, while in Rules 2 the default realisation is 0 (16).

(15) "<baaH\bona>"

"#ba+aH\+bon+a#" SP2 RemPast VH:'see' FV

"<baaH\seka>"

"#ba+aH\+sek+a#" SP2 RemPast VL:'laugh' FV

(16) "<baH\bona>"

"#baP\+bon+a#" SP2 PresPart VH:'see' FV

"<baH\seka>"

"#baP\+sek+a#" SP2 PresPart VL:'laugh' FV

In the tenses where the grammatical accent is assigned to the first mora of either the first or the second syllable of the macrostem, the tense sublexicons lead to the appropriate stem sublexicons, where the place of the accent is marked with R\ for the first or the second syllable. These are, for example, the remote past disjunct (*-ára-*) (17) and the negated present (*nti-*) (18), respectively. The default surface value for R\ is H\ in Rules 1, and 0 in Rules 2.

(17) "<baaH\raboH\na>"

"#ba+aH\ra+boR\n+a#" SP2 RemPastDisj VH:'see' FV

"<baaH\raseH\ka>"

"#ba+aH\ra+seR\k+a#" SP2 RemPastDisj VL:'laugh' FV

(18) "<ntibabonaH\>"

"#nti+ba+bon+aR\#" Neg SP2 PresNeg VH:'see' FV

"<ntibasekaH\>"

"#nti+ba+sek+aR\#" Neg SP2 PresNeg VL:'laugh' FV

In the tenses where the grammatical accent is realised on the first syllable of the macrostem, the accent is associated with the vowel of the possible object prefix. The tense sublexicon leads to the accented object prefix sublexicon and from there to the accentless stem sublexicon (19). In the tenses where the grammatical tone is realised on the second syllable of the macrostem, the accent is associated with the first root vowel, when the object prefix is present (20).

(19) "<baaH\ramuH\bona>"
"#ba+aH\ra+muL\+bon+a#" SP2 RemPastDisj OP1 VH:'see' FV
"<baaH\ramuH\seka>"
"#ba+aH\ra+muL\+sek+a#" SP2 RemPastDisj OP1 VL:'laugh' FV

(20) "<ntibamuboH\na>"
"#nti+ba+mu+boR\n+a#" Neg SP2 PresNeg OP1 VH:'see' FV
"<ntibamuseH\ka>"
"#nti+ba+mu+seR\k+a#" Neg SP2 PresNeg OP1 VL:'want' FV

## CONCLUSION

As shown above, the Two-Level approach can be adapted to handle both lexical and grammatical tone, and even simultaneously so. Some of the solutions, such as listing the verbal stems several times in different sublexicons, do not at first glance seem neat, but they do not exceed the performance of the parses nor cause too much extra work when the lexicon is compiled.

## REFERENCES

Grimes, Barbara F. and Joseph E. Grimes (eds.). 2000. *Ethnologue*. Dallas, Texas: sil Inc.

Harjula, Lotta. 2004 (forthcoming). The Ha Language of Tanzania: Grammar, Texts, and Vocabulary. Köln: Rüdiger Köppe Verlag.

Hinnebusch, Thomas J., Derek Nurse and Martin Mould. 1981. *Studies in the Classification of Eastern Bantu Languages.* Hamburg: Helmut Buske Verlag.

Kreigler, René. 1999. *An approach to tone in Two-Level Morphology.* Unpublished manuscript.

Koskenniemi, Kimmo. 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.