Autosegmental representations in an HPSG of Hausa

Berthold Crysmann
Universität Bonn
Poppelsdorfer Allee 47, D–53115 Bonn
crysmann@ifk.uni-bonn.de

Abstract

In this paper I shall present a treatment of lexical and grammatical tone and vowel length in Hausa, as implemented in an emerging bidirectional HPSG of the language based on the Lingo Grammar Matrix (Bender et al., 2002). I shall argue in particular that a systematic treatment of suprasegmental phonology is indispensible in an implemented grammar of the language, both for theoretical and practical reasons. I shall propose an LKB representation that is strongly inspired by linguistic and computational work on Autosegmental Phonology. Finally, I shall show that the specific implementation presented here is flexible enough to accommodate different levels of suprasegmental information in the input.

1 Introduction

Hausa is a tone language spoken by over 30 million speakers in Northern Nigeria and bordering areas of Niger. Genetically, the language belongs to the Chadic sub-branch of the Afroasiatic family.

In this language, both tone and length are lexically and grammatically distinctive: Hausa distinguishes two vowel lengths, as well as two underlying tones, H(igh) and L(ow). At the surface level, we can observe two level tones, as well as one contour tone (fall). Wolff (1993) cites the following minimal pairs for tone:

(1) a. *fàrī* — 'look (n)'
b. *farȉ̄* — 'dry season'
c. *farī* — 'white/whiteness'

Rising tone only results from the interaction of grammatical and intonational tone (Sharon Inkelas and Cobler, 1987; Inkelas and Leben, 1990).

In addition to its function of differentiating lexical items, tone is also grammatically distinctive: the paradigms of subjunctive and preterite (=relative completive) TAM markers partially overlap in terms of their segments (*kà* '2sg.subj, *yà* '3sg.m.subj', *tà* '3sg.f.subj' vs. *ka* '2sg.rel.compl, *ya* '3sg.m.rel.compl', *ta* '3sg.f.rel.compl'). Further, the bound possessive linker and the previous reference (=specificity) marker are systematically distinguished by tonal means alone.

(2) a. rȉ̄ga-r Audù
gown.f-of.f Audù.m
'Audu's gown'
b. rȉ̄gâ-r
gown.f-spec.f
'the (aforementioned) gown

(3) a. birni-n Kanȍ
town.m-of.m Kano
'Kano town'
b. birnî-n
town.m-spec.m
'the (aforementioned) town'

Similarly, vowel length is also distinctive on both lexical and grammatical levels: Newman (2000) cites the following pair (inter alia): *fāsà̄* 'postpone' vs. *fasà̄* 'smash'. Examples of grammatical length distinctions can again be found in the areas of TAM marking: in relative clauses and focus constructions, completive aspect is expressed by means of the relative completive set (or preterite), using short vowel *na* '1.sg.rel.compl, *ka* '2.sg.rel.compl', *ya* '3.sg.m.rel.compl' and *ta* '3.sg.f.rel.compl', inter alia, which contrasts with the long vowel absolute completive *nā*, *kā*, *yā*, and *tā* used elsewhere (see Jaggar (2006) for discussion of the use of the preterite in narratives). Furthermore, Hausa uses verb-final vowel length to signal presence of a following in-situ direct object (Hayes, 1990; Crysmann, 2005).

Despite the fact that the sophisticated models of suprasegmental phonology developed more than a quarter of a century ago within Autosegmental

Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks, ACL-IJCNLP 2009, pages 28–36,
Suntec, Singapore, 6 August 2009.

Theory (Goldsmith, 1976; Leben, 1973) have already been rigorously formalised in the nineties in the context of feature-structure-based computational phonology (Bird, 1995; Scobbie, 1991; Bird and Klein, 1994; Walther, 1999), the representation of tone and length has received little or no attention in the area of grammar engineering. This may be partly due to the fact that the languages for which substantial grammars have been developed are not tone languages. Existing grammar implementations of tone languages like Chinese (Fang and King, 2007) do not appear to make use of autosegmental models either, possibly because the assignment of tone in an isolating language is not as intimately connected to inflectional and derivational processes, as it is in a morphologically rich language like Hausa.

In this paper, I shall argue that the issue of suprasegmental phonology is an integral part of any implemented grammar of Hausa, not only from the point of view of linguistic adequacy, but also under grammar-engineering and application-oriented perspectives. I shall propose a treatment of tone and length in an LKB-grammar of Hausa that systematically builds on separate representations of segments, tone and length and discuss how various salient aspects of Hausa syntax and morphology can be addressed using a representation inspired by Autosegmental Theory. Furthermore, I shall address how different levels of suprasegmental information encoded in the different writing systems employed in the language can be robustly integrated into a single grammar, and explore its application potential.

2 Suprasegmental information in Hausa writing systems

2.1 Latin script

2.1.1 Standard orthography (Boko)

Modern Hausa is standardly written using (a modified version of) the Latin script, called *bōkòo*. In addition to the standard 26 letters of the Latin alphabet, Boko uses hooked letters, the apostrophe, as well as digraphs to represent glottalised consonants (ɓ, ɗ, ƙ, ts [s'], 'y [ʔj], ' [ʔ]). Yet, neither tone nor length are represented in the standard orthography.

2.1.2 Tone & length in scientific and educational literature

In contrast to the standard orthography, tone and length are typically fully represented in the academic literature on Hausa. Besides reference grammars and other scientific publications on the language, this includes lexica, some of which exist in machine-readable form (e.g., the on-line version of Bargery (1934) at `http://bargeryhausa.gotdns.com/`).

Length in scientific publications is typically marked using one of the following strategies: diacritical marking of long (macron or post-fixed colon; Newman (2000; Jaggar (2001)) or short vowels (ogonek; Newman and Ma Newman (1977)), and segmental gemination of vowels (long) (Wolff, 1993). Regardless of whether the strategy is diacritic or segmental, there is a strong tendency to have short vowels unmarked, representing the length information on long vowels only.

Tone, by contrast, is exclusively marked by means of diacritics: again, two systems are typically used, one marking low tone with a grave accent leaving high tone unmarked, the other marking high tone with an acute accent, leaving low tone unmarked. Besides that, fully toned representations can also occasionally be found (using acute and grave accents). Falling tone, which phonologically corresponds to a H-L contour associated with a single heavy syllable, is standardly marked with a circumflex accent. Rising tone, by contrast, which only ever plays a role in intonational phonology, as mentioned in section 1, is typically not represented.[1]

Apart from the scientific literature, full representation of suprasegmental information is also provided in most of the Hausa language teaching literature, e.g. Cowan and Schuh (1976; Jungraithmayr et al. (2004). Conventions tend to follow those found in the scientific literature, given that Hausa language teaching often forms an integral part of African linguistics curricula.

The marking strategy assumed in this paper follows the one found in Newman (2000) and Jaggar (2001), using diacritics for low and falling tones, taking high tone as the default. Long vowels are marked by a macron.

2.2 Arabic script (Ajami)

Besides the now standard Latin orthography, Hausa has been written traditionally using a slightly modified version of the Arabic script called *àjàmi*. Today, Ajami is still used occasionally, mainly in the context of religious texts.

Just like Boko, Ajami does not represent tone. Owing to the Semitic origin of the script, however, length distinctions are indeed captured: while short vowels are solely marked by diacritics, if at all, long vowels are represented using a combination of letters and diacritics: long front vowels (/i:/ and /e:/) using the letter *ya* (ي), otherwise used for the palatal glide /j/, long back vowels using the letter *wau* (و), also used for the labio-velar glide /w/, and

[1] Lexical L-H sequences associated with a single syllable undergo tonological simplification rules (Leben, 1971; Newman, 1995).

long /a:/ being represented by alif (ا).[2] Vowel quality (/i:/ vs. /e:/ and /o:/ vs. /u:/) is differentiated by means of diacritics.

Thus, depending on the writing system, different levels of suprasegmental information need to be processed, ranging from full representation in scientific and educational texts, over partial representation (Ajami), to complete absence of any tone or length marking (Boko). This means that the grammar should be able to extract what information is available, and robustly deal with both specified and underspecified input. This is even more important, if we want to include applications, where input in parsing is an underspecified representation, but output in generation requires full specification of suprasegmentals, e.g., in TTS or CALL scenarios.

3 Morphology and suprasegmental phonology

Hausa morphological processes, like derivation and inflection, display close interaction between segmental and suprasegmental marking. Affixation in Hausa is predominantly suffixal, although prefixes and circumfixes are also attested. On the segmental level, affixes can be divided into fully specified suffixes, and reduplicative suffixes. Although partial and full reduplication of entire CV-sequences can also be observed, probably the most common reduplicative pattern involves reduplication and gemination of root consonants, with vowel melodies prespecified.

Tonally, affixes fall into one of three categories: affixes lexically unspecified for tone (only prefixes), tone-integrating affixes (suffixes only) and non-integrating affixes[3]. While non-integrating affixes only specify their own lexical tone, possibly affecting the segmental and suprasegmental realisation of a preceding syllable, tone-integrating suffixes holistically assign a tonal melody to the entire word they attach to.

In contrast to tone, which is often assigned to the entire morphological word, alternations in length do not tend to affect the entire base, but rather only syllables at morpheme boundaries.

3.1 Tone-integrating suffixes

Hausa plurals represent the prototypical case of tone-integrating affixation. The language has an extremely rich set of morphological patterns for plural formation: Newman (2000) identifies 15 classes, many of which have between 2 and 6 subclasses. Quite a few Hausa nouns form the plural according to more than one pattern. Among these 15 plural classes, three are particularly productive, most notably classes 1-3. All these three classes are tone integrating, as are almost all plural formation patterns. Thus, regardless of the tonal specification in the singular, plural formation assigns a regular tone melody to the entire word:

(4) -ōXī (H) (Class I)

a. gulà (HL) — gulōlī ‘drum stick’

b. tāgà (HL) — tāgōgī ‘window’

c. gyàlè (LL) — gyalōlī ‘shawl’

d. tàmbayà (LHL) — tambayōyī ‘question’

e. kamfànī (HLH) — kamfanōnī ‘company’

f. kwàmìtî (LLHL) — kwamitōcī ‘committee’

(5) -ai (LH) (Class II)

a. àlhajì (LHL) — àlhàzai ‘Hadji’

b. ɗālìbī (HLH) — ɗàlìbai ‘pupil’

c. sankacè (HHL) — sànkàtai ‘reaped corn laid down in a row’

d. àlmùbazzàrī (LLHLH) — àlmùbàzzàrai ‘spendthrift’

Class I plural formation involves affixation of a partially reduplicative suffix *-ōXī* replacing the base-final vowel, if there is one. Tone in class I plurals is all H, regardless of whether the base is HL, LH, LL, HLH, or LHL. Length specifications, by contrast are carried over from the base, except of course for the base-final vowel. The quality of the affix-internal consonant is determined by reduplication of the base-final consonant, possibly undergoing regular palatalisation.

Class II plurals are formed by means of the fully specified suffix *-ai*, with an associated integrating LH. Tone assignment in Hausa is right to left: thus, L automatically spreads to the left. Again, the tonal shape of the base gets entirely overridden by the LH plural pattern. Non-final length specifications, however, are identical between the singular and the plural.

3.2 Toneless prefixes

As we have seen above, tonal association in Hausa proceeds from right to left. As a result, suffixes carry a lexical specification for tone. Amongst

[2] Ajami letter names are the Hausa equivalent of original Arabic names. For a more complete description of Ajami, see Newman (2000, pp. 729–740).

[3] Among the non-integrating affixes, there is a subclass bearing polar tone, i.e., the surface tone is opposite to that of the neighbouring syllable.

Hausa prefixes, however, one must distinguish between those prefixes carrying a (non-integrating) lexical tone specification themselves, and those prefixes which are inherently unspecified for tone but have their surface tone determined by means of automatic spreading. An example of a prefix of the latter type is provided by the reduplicative prefixes C_1VC_1- and C_1VC_2 found with pluractional verbs. These prefixes consists of an initial consonant that copies the first consonant of the base, followed by a short vowel copying the first vowel of the base (possibly undergoing centralisation). The prefix-final consonant either forms a geminate with the following base-initial consonant, or else copies the second consonant of the base.

(6) C_1VC_1-

a. darnàcē (HLH) — daddarnàcē (HHLH) 'press down/oppress (gr 1)'

b. karàntā (HLH) — kakkaràntā (HHLH) 'read (gr 1)'

c. dà̰gurà̰ (LHL) — dàddà̰gurà̰ (LLHL) 'gnaw at (gr 2)'

d. gyà̰ru (LH) — gyàggyà̰ru (LLH) 'be well repaired (gr 7)'

With trisyllabic bases, it is evident that the tone assumed by the prefix is just a copy of the initial tone of the base.

The tonal pattern assigned to Hausa verbs are determined by paradigm membership, the so-called grade (Parsons, 1960), together with the number of syllables. Tone melodies range from monotonal, over bitonal, to maximally tritonal patterns. Thus, tone-assignment to quadrisyllabic verbs, as derived by pluractional prefixes, is an effect of automatic spreading.

Pluractional affixation to bisyllabic verbs constitutes a slightly more complicated case: Since some paradigms assign different tone melodies to bisyllabic and trisyllabic verbs, prefixation to bisyllabic bases triggers a change in tonal pattern. Note, however, that the tonal pattern assigned to the derived trisyllabic pluractional verb is just the one expected for trisyllabic underived verbs of the same paradigm (cf. underived grade 1 *karàntā* and grade 2 *dà̰gurà̰* above to the pluractional grade 1 and grade 2 verbs below).

(7) a. tākà̰ (HL) — tattà̰kā (HLH) 'step on (gr 1)'

b. jḛ̀fā (LH) — jàjjḛ̄fà̰ (LHL) 'throw at (gr 2)'[4]

[4]Owing to the inherent shortness of the reduplicated vowel, long /e:/ and /o:/ undergo regular reduction to [a] in the reduplicant.

Thus, instead of the affix carrying lexical tone, tone is rather assigned holistically to the entire derived word (Newman, 2000).

3.3 Non-integrating affixes

The third class of affixes we shall discuss are lexically specified for tone again (if vocalic). Yet, in contrast to tone-integrating suffixes, they do not override the entire tonal specification of the base. Examples of tonally non-integrating *suffixes* are manifold. They include nominal and verbal suffixes like the bound accusative (polar) and genitive pronouns, the genitive linker (*-n/-r*), the inherently low-tone specificity marker (*-ǹ/-ǹ*), and the regular gerundive suffix -*`wā*, among many others. What is common to all these suffixes is that they only affect the segmental and suprasegmental specification of the immediately preceding base-final syllable.

Regular gerunds of verbs in grades 1, 4, 5, 6 and 7 are formed by affixation of a floating tone-initial suffix -*`wā*. When attached to a verb ending in a long high syllable, the base final high tone and the floating low tone combine into a falling contour tone. If the base ends in a high short syllable, as in grade 7, or if the base-final vowel is already low, no tonal change to the base can be observed.

(8) a. karàntā — karàntầwā 'read (gr1)'

b. sayar — sayârwā 'sell (gr5)'

c. kāwō — kāwồwā 'come (gr6)'

d. kāmà̰ — kāmà̰wā 'catch (gr1)'

e. gyà̰ru — gyà̰ruwā 'be repaired (gr7)'

Note that apart from tonal change of high long to falling, the base undergoes no segmental or length change.

Consonantal suffixes, like the genitive linker and the specificity marker, by contrast, necessarily integrate into the coda of the preceding syllable. Since Hausa does not allow long vowels in closed syllables, base-final long vowels and diphthongs are shortened. The specificity marker is identical to the genitive linker, as far as truncation of long vowels and diphthongs is concerned. It differs from the genitive linker, in that it is inherently specified as low, giving rise to a falling tone with high-final bases. With low-final bases, no tonal change can be observed.

(9) a. ƙwai — ƙwa-n-tà '(her) egg'

b. rī̀gā — rī̀ga-r-tà '(her) gown'

c. mōtà̰ — mōtà-r-tà '(her) car'

(10) a. ƙwai — ƙwâ-n 'the (aforementioned) egg'

b. rȉgā — rȉgâ-r ‘the (aforementioned) gown’

c. mōtȁ — mōtà-r ‘(her) car’

Note that in contrast to tone-integrating suffixes, segmental and suprasegmental changes are strictly local, affecting material in adjacent syllables only.

Besides non-integrating suffixes there are some very rare prefixes that can be regarded as inherently specified for tone. One such prefix is low tone bà- that features in singular ethnonyms, like, e.g. *bàhaushȅ* ‘Hausa person’. Typically, the prefix bà- is accompanied by a final tone-integrating HL suffix *-ȅ* (masc) or HLH *-ā/-ìyā* (fem), but not always. With regular ethnonyms, the initial tone of the suffix (H) spreads to the left, up to but excluding the low tone prefix. The plural of such ethnonyms is formed without a prefix. Instead, a tone-integrating H or LH suffix *-āwā* is used. Vowel length of the base is retained throughout:

(11) Fàransà ‘France’ — Bàfaranshȅ (m), Bàfaranshìyā (f) , Faransāwā (pl) ‘French’

(12) Jāmùs ‘Germany’ — Bàjāmushȅ (m), Bàjāmushìyā (f) , Jāmusāwā (pl) ‘French’

Besides the regular pattern, there are a few ethnonyms that use a non-integrating *-ī* e.g. *Bàgȍbirī* from *Gȍbir*, thus preserving the tonal pattern of the place name base. According to Newman (2000), however, many Hausa speakers prefer to use the regular tone-integrating suffix *-ȅ* instead. Thus, entirely non-integrating formation of ethnonyms has ceased to be a part of productive Hausa morphology.

Moreover, even the productivity of tonally specified *bà-* seems to be diminished: while the plural is still productive, new ethnonyms tend to be formed using alternate periphrastic constructions *ɗan/mùtumìn* ‘son/man of’ (Newman, 2000).

(13) a. Pàlàsɗīnù ‘Palestine’ — ɗan/mùtumìn Pàlàsɗīnù (m) — Palasɗīnāwā (pl) ‘Palestinian’

b. Bosniyà ‘Bosnia’ — ɗan/mùtumìn Bosniyà (m) — Bosniyāwā (pl) ‘Bosnian’

To summarise, I shall take integrating and non-integrating suffixation as the standard case in Hausa, together with toneless prefixation. As we shall see in the description of our implementation in the following section, the treatment of isolated cases of tonally specified prefixes will be treated as a non-productive sub-regularity.

4 Representing autosegmental phonology in the LKB

4.1 Orthographemics in the LKB

The LKB (Copestake, 2002) has built-in support for orthographemic alternations, providing support for inflectional and derivational morphology. Technically, the orthographemic component of the LKB adopts a string-unification approach. Below is an example of the spelling part of regular -ōXī plural formation, together with the definitions of letter sets and wild-cards used. Patterns on the right preempt patterns further to the left.

(14)
```
%(wild-card (?v aeiou))
%(letter-set (!c bcdfghjklmnpqrstvwxyzɓɗƙ))
noun_pl1_vow_ir :=
   %suffix (!c?v !co!ci) (t?v toci)
   (s?v soshi) (w?v woyi) (ts?v tsotsi)
   noun-plural-infl-rule &
   ...
```

In the above rule, the letter set `!c` is string unified with the corresponding consonantal letter in the input. Note that in contrast to wild cards (e.g. `?v`), multiple occurrences of letter set identifiers within the same pattern are bound to the same consonant, providing a convenient solution to gemination and partial reduplication.

Orthographemic rules are unary (lexical) rules consisting of a feature structure description and an associated spelling change. The orthographemic part is applied to surface tokens in order to derive potential stem forms. The parser’s chart is then initialised with lexical entries that have a corresponding stem form. The orthographemic rules that have been applied in order to derive the stem are recorded on an agenda such that the feature structure part can be applied to the lexical entries thus retrieved.

Recall from section 2 that Hausa standard orthography does not represent tone or length. Thus, suprasegmentally unmarked strings define the common denominator for retrieving entries from the lexicon. But even if the input is marked diacritically for suprasegmentals, tone-integrating morphology can lead to drastic tonal changes, which are superficially encoded as segmental alternations (since á $\neq$ à). Moreover, we hope to have shown above that tone and segmental phonology should best be treated separately. Consequently, orthographic representations unmarked for tone constitute the common denominator for all orthographic input representations.

In a first preprocessing step, tone and length specifications on input tokens are extracted by means of a regular expression preprocessing engine built into the LKB (Waldron et al., 2006).

Instead of simply removing this potentially valuable information, the preprocessor rules convert the (diacritical) marking of tone and length into an inverse suffixal representation, separated from the segmental string by _. Overtly marked high will be represented as _H, overtly marked low as _L, and lack of tonal marking is recorded as _*. Similarly, length information, if present, will be recorded by means of a colon next to the corresponding tone. E.g., input ɗàlìbai 'pupils' will be converted into ɗalibai_*_L_L:, whereas tonally unspecified ɗalibai will become ɗalibai_*_*_*. Input partially specified for length (ɗaalibai), as, e.g., in Ajami, will receive a representation as ɗalibai_*_*_*:.

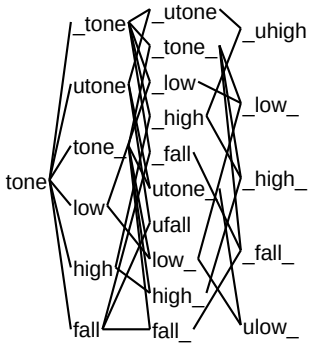
Once we have separated suprasegmental information from the orthography proper and stored it in the form of suffixal annotations, we can use LKB's standard orthographemic machinery to convert the suffixal annotation into feature structure constraints.[5]

4.2 Phonological representation

As we have seen above, there are several strategies of tone and length marking in Hausa. While overtly marked tone and length is both unambiguous in itself and directly enables us to infer what marking strategy is used, the interpretation of vowels unmarked for tone or length depends entirely on the context: if a low-marking strategy is employed, unmarked segments (=_*) can be interpreted as high. However, if no marking of tone occurs at all in the input, unmarked segments should be compatible with any tone. The very same goes for length. In order to enable the grammar to flexibly infer the meaning of these underspecified annotations, we introduce the following type hierarchy of tonal marking. The only assumption made here is that the marking strategy being adopted is used consistently across the entire input sentence.[6]

Lexical and grammatical tones will be one of *high*, *low*, or *fall*.[7] In addition to these three linguistic tones, the type hierarchy features tonal types that correspond to tonal annotations found in the input: *utone* is the type associated with tonally unmarked syllables, *tone_* is the type associated with a high-marking strategy, *_tone* corresponds to low-marking, and *_tone_* to full tonal marking (overt high and low).

(15)



Depending on which annotations are present in the input, the meaning of underspecified annotations can be determined on the basis of type inference. The orthographemic rules that consume tonal annotations do exactly two things: first, they record the tone specification just found as the first member of the TONE list of the daughter, successfully building up a list of surface tones from right to left.

(16)
```
_HH_ir :=
  %suffix (* _H:)
  diacritic-irule &
  [SUPRA [TONE [LIST #tones,
                LAST #tl],
          LEN  [LIST #lens,
                LAST #ll]],
   DTR [SUPRA [TONE [LIST
             high-marked-list &
             <high . #tones>,
                      LAST #last],
                LEN  [LIST
             long-marked-list &
             <long . #lens>,
                      LAST #ll]]]].

_*_ir :=
  %suffix (* \*)
  diacritic-irule &
  [SUPRA [TONE [LIST #tones,
                LAST #tl],
          LEN  [LIST #lens,
                LAST #ll]],
   DTR [SUPRA [TONE [LIST
             <utone . #tones>,
                      LAST #last],
                LEN  [LIST
             <ulength . #lens>,
                      LAST #ll]]]].
```

If the annotation is that of an overtly unmarked tone, the underspecified type *utone* is inserted, otherwise *high* or *low*, as appropriate. H or L tone rules simultaneously constrain the entire tone list according to the marking strategy, using list constraints.

(17)
```
high-marked-list :=
  tone-marked-list.
high-marked-null :=
  high-marked-list &
  tone-marked-null.
```

[5] In the near future, we plan to supplant this two-step solution with a direct conversion of using diacritical information into feature structure annotations, using the advanced token-mapping developed by Adolphs et al. (2008). At present, however, this token-mapping has only been integrated into the Pet run-time system (Callmeier, 2000), but not yet into the LKB.

[6] In principle, even this assumption can be relaxed, at the peril of having reduced cross-sentence disambiguation.

[7] I do not decompose falling tone into HL sequences, thereby simplifying the alignment between tone specifications, length specifications and segments.

```
high-marked-cons :=
  high-marked-list &
  tone-marked-cons &
  [FIRST tone_,
   REST high-marked-list].
```

Presence of a single overtly marked high tone will constrain every element of the tone list to be a subtype of *high_*. According to the hierarchy of tonal types given above, the greatest lower bound of *utone* and *high_* however, is *low_*, denoting (unmarked) low tone under a high-marking strategy. Thus, whatever tonal marking is found, unmarked tones are coerced to represent the opposite tones. The way the type hierarchy is set up, 4 different marking strategies are possible: completely unspecified tone, high-tone marking, low-tone marking and fully explicit high- and low-tone marking.

With the constraints we have just seen, we only get disambiguation of unmarked tone (and length) within the same word. In order to disambiguate across the entire sentence, we use difference lists of these tone and length lists to propagate the marking regime to preceding and following words. In essence, we use two difference lists _LTONE and _RTONE to propagate from left to right and vice versa.[8] Lexically, every word inserts its own tone list as the singleton member of each difference list. The general phrasal types from which all grammar rules inherit now concatenate the _LTONE and _RTONE values of their daughters left to right and right to left, respectively.

The tone marking rules given above are then further constrained according to the types of _LTONE and _RTONE. Using list-of-list type constraints as given below, every word marked for tone will constrain the marking regime found to its left and to its right.

(18)
```
hm-llist := tm-llist.
hm-clist := tm-clist &
    hm-llist &
    [FIRST high-marked-list,
     REST hm-llist].
hm-nlist := hm-llist & tm-nlist.
```

The treatment of length marking, as we have hinted at already, is entirely analogous to that of tone, imposing the corresponding constraints on a list of vowel length specifications.

With these constraints in place, we get the following disambiguation results (note that the verb *zō* is lexically specified as long):

(19)
a. Fully unspecified: `Ya zo` (3 readings: *yā zō, ya zō, yà zō*)
b. Length specified: `Ya zoo` (2 readings: *ya zō, yà zō*)
c. Length specified: `Yaa zoo` (1 reading: yā zō)
d. Tone/length specified: `Ya kaawoo shì` (1 reading: *ya kāwō shì*)
e. Fully specified: `Yá zóó` (1 reading: *ya zō*)
f. Inconsistent: `Yaa zo` (0 readings)

As witnessed above, presence of length marking coerces vowels not marked as long into the short vowel reading. Similarly, presence of a single low tone marking enforces a high tone reading of overtly unmarked tones.

In generation, the grammar only uses fully specified tone marking, i.e., application of rules such as `_*_ir` is blocked. As a result, we always get a surface representation with full tone and length information. Post-generation Lisp functions are used to convert the suffixal notation into the appropriate diacritic format.

4.3 Morphology

The main motivation for having tone and length represented on separate lists is two-fold: first, as witnessed by Ajami, writing systems may overtly mark one distinction but not the other. Second, and more importantly, we have seen in section 3, that morphological processes tend to leave length intact, even if the entire word is holistically marked with a completely new tonal melody, unrelated to that of the base. Having two separate lists, we can replace the tonal structure in the course of morphological derivation but still have the rhythmic structure shared between base and derived form by means of reentrancies.

Here we investigate in more detail the role these representations play in morphological derivation.

In the previous section, we provided a general representation of segmental and suprasegmental information, the latter being encoded by means of two lists and showed how preprocessor rules and orthographemic rules are used to extract this information from the input and associate it with parts of the feature structure, such that it can be matched against morphological and lexical constraints on length and tone.

Since both tone and length are lexically distinctive, every lexical item specifies the contents of its SUPRA|TONE and SUPRA|LEN lists. The order of the elements on these two lists is right to left, facilitating a treatment of tone spreading by means of list types. At the same time, this encoding provides convenient access to the right-most length and tone

[8] Since only overtly marked items can disambiguate tonally unmarked ones, and the position of these disambiguating items in the string is not known a priori, we need two lists of lists, one for disambiguation of preceding material (_LTONE), the other for following material _RTONE.

specification. Since Hausa is predominantly suffixal, non-holistic morphophonological changes to tone and length specifications exclusively target the right-most syllable of the base.

As we have observed above, tonal changes can be far more global than segmental and length alternations. Thus, we will use the LEN list to synchronise the segmental and suprasegmental representations. Consequently, length specifications will always be a closed list. Tone, by contrast, may involve spreading, i.e. the exact number of individual H of an all H tone melody is determined by the number of available tone bearing units, which corresponds to vowel length specifications in our grammar. Since the number of tone bearing units is already fixed by the length of LEN, and because the tone marking rules operate synchronously on TONE and LEN, we are free to underspecify the tonal representation as to the exact length of the melody. Therefore, we can provide a straightforward account of right-to-left association and left-ward tone spreading in terms of open tone list types.

(20)
```
h*-list := list.
h*-cons := h*-list &
           cons & [FIRST high,
   REST h*-list].

h*-null := h*-list & null.

h*-l-list := list.
h*-l-cons := h*-l-list &
             cons & [FIRST low,
                     REST h*-list].
```

As we shall see shortly, these list types provide a highly general way to constrain holistic tonal assignment, independently of the segmental make-up of the base.

In order to illustrate the interplay between segmental and suprasegmental constraints in morphological derivation, I provide a treatment of the two major types of morphological rules: tone-integrating and non-integrating.[9]

(21)
```
noun_pl1_vow_ir :=
%suffix (!c?v !co!ci) ...
noun-plural-infl-rule  \&
[SUPRA
   [TONE [LIST h*-list],
    LEN [LIST < long, long . #ll>,
         LAST #llast] ],
 DTR [SYNSEM.LKEYS.--MCLASS n-pl-1,
      SUPRA.LEN [LIST  < [] . #ll>,
                 LAST #llast]]].
```

Tone integrating affixes In our discussion of the Class I plural inflection rule above, we have only specified the segmental changes. As detailed in the version below, holistic assignment of tone is achieved by means of a list type constraint on the TONE of the mother, paired with the absence of any tonal restrictions regarding the morphological daughter (the base). The length marking of the two inherently long suffix vowels is captured by means of the addition of two *long* specification at the front of LEN. Affixation of *-ōXī* replaces the base final vowel. Accordingly, the associated initial length specification of the daughter is skipped and the remaining list is passed on to the length specification of the mother.

Non-integrating affixes In feminine singular specificity marking, both non-integrating tone and length changes can be observed. As depicted below, high-final bases undergo a tone change to fall. The remainder of the TONE list is structure-shared between mother and daughter, carrying over any list constraints that might be imposed there.

(22)
```
f-sg-noun_def_high_ir :=
 %suffix (!v !vr) (!vi !vr) ...
 noun-def-f-sg-irule &
 [SUPRA [TONE [LIST <fall . #tl >,
               LAST #tlast],
         LEN [LIST <short . #ll>,
              LAST #llast]],
  DTR [SUPRA
       [TONE [LIST <high . #tl>,
              LAST #tlast],
        LEN [LIST <[] . #ll>,
             LAST #llast] ]]].
```

Likewise, final shortening, which is triggered by the affixation of a syllable-final consonant, is captured by an analogous constraint on LEN.

5 Conclusion

In this paper, we have proposed a treatment of tone and length in Hausa in terms of distinct representations of segments, tone and length. We have shown that this separation is not only needed to accommodate different orthographic representations in the input, but that it also paves the way for a more general account of Hausa morphology, most notably holistic assignment of tonal melodies combined with tone spreading. At present, the grammar is not only capable of extracting different levels of suprasegmental annotations contained in the input, but can also resolving tone and length ambiguities on the basis of grammatical constraints: e.g., the ambiguity between genitive linker and previous reference marker, or the ambiguity between subjunctive, preterite, and absolute completive in relative and focus constructions. In the future, we intend to equip the grammar with parse selection models, to further enhance disambiguation. Given the bidirectionality of the grammar and its flexible support for tone and length, we plan to use it in the context of TTS and CALL applications in the near future.

[9] Toneless prefixation with automatic spreading constitutes just a special sub-case of tone-integrating rules.

References

Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger, and Bernd Kiefer. 2008. Some fine points of hybrid natural language parsing. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008), May, Marrakesh*.

G. P. Bargery. 1934. *A Hausa–English Dictionary and English–Hausa Vocabulary*. Oxford University Press, London.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammar. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14.

Steven Bird and Ewan Klein. 1994. Phonological analysis in typed feature systems. *Computational Linguistics*, 20(3):455–491.

Steven Bird. 1995. *Computational Phonology. A Constraint-based Approach*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Ulrich Callmeier. 2000. PET — a platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering*, 6(1):99–108.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

J. Ronayne Cowan and Russell Schuh. 1976. *Spoken Hausa*. Spoken Language Services, Ithaca.

Berthold Crysmann. 2005. An inflectional approach to Hausa final vowel shortening. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 2004*, pages 73–112. Kluwer.

Ji Fang and Tracy Holloway King. 2007. An LFG Chinese grammar for machine use. In Tracy Holloway King and Emily Bender, editors, *Proceedings of the GEAF 2007 Workshop*, CSLI Studies in Computational Linguistics ONLIN. CSLI Publications.

John A. Goldsmith. 1976. *Autosegmental Phonology*. Ph.D. thesis, MIT.

Bruce Hayes. 1990. Precompiled phrasal phonology. In Sharon Inkelas and Draga Zec, editors, *The Phonology-Syntax Connection*, pages 85–108. University of Chicago Press.

Sharon Inkelas and William R. Leben. 1990. Where phonology and phonetics intersect: The case of Hausa intonation. In Mary E. Beckman and John Kingston, editors, *Between the Grammar and the Physics of Speech*, Papers in Laboratory Phonology, pages 17–34. Cambridge University Press, New York.

Philip Jaggar. 2001. *Hausa*. John Benjamins, Amsterdam.

Philip Jaggar. 2006. The Hausa perfective tense-aspect used in wh-/focus constructions and historical narratives: A unified account. In Larry Hyman and Paul Newman, editors, *West African Linguistics: Descriptive, Comparative, and Historical Studies in Honor of Russell G. Schuh*, Studies in African Linguistics, pages 100–133.

Herrmann Jungraithmayr, Wilhelm J. G. Möhlig, and Anne Storch. 2004. *Lehrbuch der Hausa-Sprache*. Rüdiger Köppe Verlag, Köln.

William R. Leben. 1971. The morphophonemics of tone in Hausa. In C.-W. Kim and Herbert Stahlke, editors, *Papers in African Linguistics*, pages 201–218. Linguistic Research, Edmonton.

William Leben. 1973. *Suprasegmental Phonology*. Ph.D. thesis, MIT.

Paul Newman and Roxana Ma Newman. 1977. *Modern Hausa–English Dictionary*. University Press, Ibadan and Zaria, Nigeria.

Paul Newman. 1995. Hausa tonology: Complexities in an 'easy' tone language. In John Goldsmith, editor, *The Handbook of Phonological Theory*, pages 762–781. Blackwell, Oxford.

Paul Newman. 2000. *The Hausa Language. An Encyclopedic Reference Grammar*. Yale University Press, New Haven, CT.

F. W. Parsons. 1960. The verbal system in Hausa. *Afrika und Übersee*, 44:1–36.

Jim Scobbie. 1991. *Attribute-Value Phonology*. Ph.D. thesis, University of Edinburgh.

William R. Leben Sharon Inkelas and Mark Cobler. 1987. The phonology of intonation in Hausa. In *Proceedings of the North-Eastern Linguistic Society 17*, pages 327–341.

Ben Waldron, Ann Copestake, Ulrich Schäfer, and Bernd Kiefer. 2006. Preprocessing and tokenisation standards in DELPH-IN tools. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 2263–2268, Genova, May.

Markus Walther. 1999. *Deklarative Prosodische Morphologie*, volume 399 of *Linguistische Arbeiten*. Niemeyer, Tübingen.

Ekkehard Wolff. 1993. *Referenzgrammatik des Hausa*. LIT, Münster.