Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 6*. Editor: Peter K. Austin

Documenting grammatical tone using Toolbox: an evaluation of Buseman's interlinearisation technique

STUART MCGILL

Documenting grammatical tone using Toolbox: an evaluation of Buseman's interlinearisation technique[1]

Stuart McGill

1. Introduction

For tone languages, particularly those found in Africa, it is often the case that important grammatical distinctions depend solely on the occurrence of different tone patterns on a particular noun or verb. The goal of this paper is to consider how the grammatical contribution of such tone patterns should be represented in the annotations within documentary corpora. After introducing certain tone phenomena from Cicipu, a Benue-Congo language spoken in northwest Nigeria (section 2), I will challenge the view that it is not appropriate to make explicit the effects of tone patterns in interlinear annotations (section 3), and review a technique devised by Alan Buseman of the Summer Institute of Linguistics for handling tone using the software program Toolbox[2] (section 4). Based on the evaluation of this technique as applied to a corpus of Cicipu texts (section 5)[3], I suggest some recommendations for the future development of interlinearisation software (section 6).

[1] Special thanks are due to Alan Buseman for posting details of his Toolbox technique in response to an initial query by Sophie Salffner and myself, and to Peter Austin, Oliver Bond, Mike Cahill, Tucker Childs, Lutz Marten and Alan Buseman again for helpful comments on an earlier version of this paper. I would also like to thank Heidi Rosendall for help with some technical details of Toolbox. An earlier version of this paper was presented at the Conference on Language Documentation and Linguistic Theory, held at the School of Oriental and African Studies in London 7th-8th December 2007.

[2] http://www.sil.org/computing/toolbox

[3] See http://www.cicipu.org/texts.html for an application of the methods discussed in this paper. This material was collected during two field trips to the Cicipu language area betwen 2006 and 2008, funded by Field Trip Grant FTG0102 from the Hans Rausing Endangered Languages Documentation Project and by a grant from the University of London Central Research Fund. Unless otherwise stated, all linguistic examples in this paper are from Cicipu.

Stuart McGill 2009. Documenting grammatical tone using Toolbox: an evaluation of Buseman's interlinearisation technique. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 6,* 236-250 London: SOAS.

2. Grammatical tone in Cicipu

Grammatical tone refers to a tonal change which signals not a lexical difference of the kind shown in (1), but a grammatical difference as in (2);

(1)	(a) *káayá*	‘room’	H H[4]
	(b) *káayà*	‘bean’	H L

(2)	(a) *ǹdúkwà*	‘I went’	L H L
	(b) *ńdùkwà*	‘I should go’	H L L

The tone patterns on the nouns in (1) make no isolatable contribution to the meaning of the word. Consequently the tones on the lexemes *káayá* and *káayà* must be memorised by learners along with the segmental phonemes. In contrast, the examples in (2) share a common component of meaning, namely a GO event involving the speaker as theme. The tone patterns do contribute a separable component of meaning, either REALIS in (2a) or IRREALIS in (2b). It will be argued in section 3 that the contribution of tone patterns such as these to the meaning of the utterance should be made explicit in interlinear annotations. The verbs themselves are inherently toneless, the tones with which they surface being determined entirely by mood and aspect.

Before turning to the more general discussion it will be helpful to consider some examples of grammatical tone in more detail:

(3) (a) *ù-dúkwà* L H L
3SG-go\RLS
‘he/she went’

(b) *ú-dùkwà* H L L
3SG-go\IRR
‘he/she should go’

[4] Cicipu has two contrasting levels of tone (written H and L here) plus a falling tone HL, which can be analysed as a sequence of H plus L. Other abbreviations used in this paper are 2 = second-person, 3 = third-person, AG = agreement, AOR = aorist, CAUS = causative, FUT = future, HAB = habitual, INJ = injunctive, IRR = irrealis, IMP = imperative, NEG = negative, RLS = realis, SG = singular.

(4) (a) *Ø-dúkwà* H L
2SG-go\RLS
‘you (s.) went’

(b) *dùkwá* L H
go\IMP
‘go!’

These two examples illustrate a three-way alternation in mood involving just tonal changes. In each case the segmental material remains the same in the (a) and (b) sentences, but the different tone patterns superimposed on the segments give rise to different values for the grammatical category of mood, and hence different meanings. In (3) the verb can be found in two different moods, depending on the tone pattern. In (3a) the verbal word *u-dukwa* (‘3SG-go’) takes an L H L pattern which indicates realis mood, while in (3b) *u-dukwa* has a H L L pattern with a high-tone prefix, indicating irrealis mood. Examples (4a) and (4b) illustrate a similar contrast, this time between realis and imperative, and involving the second person.

With vowel-initial roots such as *aya* (‘come’) the surface tones are slightly different, as shown in (5). The expression of realis mood results in a rising tone across the first syllable [waː], while irrealis mood leads to a falling tone. Nevertheless the underlying representations remain L H L and H L L, as long as contour tones are considered to be sequences of level tones (a standard assumption for African tone languages – see Clements 2000:153). This autonomous behaviour of the tone patterns is a strong argument for separating them out in interlinear annotations.

(5) (a) *wǎːyà* LH L
3SG:come\RLS
‘he/she came’

(b) *wâːyà* HL L
3SG:come\IRR
‘he/she should come’

While the examples above are all from Cicipu, grammatical tone is a common phenomenon in African languages, and so a further example is given from the Bantoid language Kakɔ. Again, the only formal difference between the examples in (6) is the tone pattern on the verb.

(6)	(a)	*à*	*tí*	*ɓēŋgwɛ̀* [M L]	*nyɛ́*	*nā*
		he	NEG	follow\FUT	him	NEG
		‘he won’t follow him’				
	(b)	*à*	*tí*	*ɓēŋgwɛ́* [M H]	*nyɛ́*	*nā*
		he	NEG	follow\AOR	him	NEG
		‘he does not follow him’ [stating a general principle]				
	(c)	*à*	*tí*	*ɓēŋgwɛ̄* [M M]	*nyɛ́*	*nā*
		he	NEG	follow\INJ	him	NEG
		‘he must not follow him’				

(Ernst 1996:3)

Similar systems can also be found in a number of Bantu languages, and are sometimes called “predictable” (Kisseberth and Odden 2003:61) – see also Crozier (1984:145) for the realis/irrealis distinction in Cishingini, a relative of Cicipu, and Bird (1999:15-16) for interesting data on the Bantoid language Etung.

3. Interlinear formats for grammatical tone

Most documentation projects will include a certain amount of interlinearised annotation. According to Schultze-Berndt (2006:239) “in an annotated corpus, it is also recommended practice to include interlinear glosses for all or at least part of the transcriptions”. Lehmann (2001) seems to argue for an interlinear annotation accompanying every text. Given that the aim of a documentary corpus is to “represent the language for those who do not have access to the language itself” (Lehmann 2001:88), while a recording and its accompanying free translation can give such a person access to the sound of the utterances and a grasp of their meanings, in order for the linguistic structure of the text to be accessible to the non-specialist linguist, further annotation is required. Lehmann therefore suggests that the minimum level of annotation should be to transcribe utterances with an morphophonemic representation, and to annotate this transcription with an interlinear morphemic gloss (IMG). Others recommend that only a small amount of the transcribed text should be interlinearised (e.g. Wittenberg 2003:124), in which case it is especially important that such interlinearised output is well-structured.

The question to be addressed here is how the kind of tone patterns illustrated in section 2 should be represented in such interlinear texts. An obvious starting point for the discussion is the Leipzig Glossing Rules (Bickel, Comrie, and Haspelmath 2004), a set of widely-accepted conventions for interlinear glossing. The relevant rule here is 4D (Bickel at al 2004:4) which states that "if a grammatical property in the object-language is signaled by a morphophonological change of the stem (ablaut, mutation, etc.), the backslash is used to separate the category label and the stem gloss". The application of this rule can be seen in examples (3-6) above. Lehmann (2004a:26) includes the same rule (R20), which:

> Distinguishes [stem-change processes] from other morphological processes, but not from each other. *Such a morpheme can hardly be signaled in the L1*[5] *representation* [my emphasis – S.M.].

As a result of this indeterminacy, in the annotations presented in (3-6) it is not apparent what it is in the text that contributes the verb mood. The process signalled by the backslash could refer to the first tone, the second tone, the tone pattern as a whole, or even a vowel change or consonant mutation.

The backslash notation treats a form such as *dúkwà* ('go\RLS') as formally unsegmentable. Nevertheless the two units *dukwa* and H L *can* be conceptually isolated and directly linked to separate components of meaning. Without making this link explicit, the contributions of the tone patterns in examples (3-6) can only be interpreted when viewed as members of a paradigm of word-forms, which an interlinear annotation does not provide. Tone-marked transcriptions are better than no tone-marking at all, and are of obvious benefit to linguists familiar with the language. However this practice by itself does not always meet Lehmann's criterion that the annotation should "represent the language for those who do not have access to the language itself" (Lehmann 2001:88).

Lieb and Drude (2000) have criticised traditional interlinear glossing in that it does not allow the degree of annotation necessary for what they consider best-practice language documentation. They provide an alternative annotation technique called Advanced Glossing (AG) which provides a large number of tiers for the representation of different kinds of linguistic data. Of relevance here is the fact that the segmental and suprasegmental parts of a

[5] L1 refers to the object language, and L2 the metalanguage.

word can be separated in the representation, and different glosses can be applied to each. Although AG allows the link between tone pattern and gloss to be made explicit, it could be argued that it does so at the expense of readability and conciseness, as well as being time-consuming to implement (Drude 2003, Schultze-Berndt 2006:251).

It is possible to imagine further complex annotation systems which would bring out the contribution of tone patterns, in particular using insights from autosegmental phonology (Goldsmith 1990), for example by splitting the transcription and gloss into two separate lines, one for the segmental tier and one for the tonal tier. However such systems would likely suffer from similar problems as AG regarding their readability and the time taken to produce them. Also the use of autosegmental formalism would introduce a theoretical sophistication into the annotations, something Lehmann (2001) argues against.

The methods discussed so far suffer either from a lack of explicitness or from undesirable complexity. The data format of Toolbox provides an opportunity for compromise, crucially because it uses one more tier than standard IMGS. It is worth making the differences between the two formats explicit. Lehmann (2004b) observes that standard IMGS consist of three lines, as in example (3-6). Example (3a) is repeated as (7) for convenience:

(7) *ù-dúkwà*
3SG-go\RLS
'he/she went'

The first line is a morphophonemic (or orthographic) transcription in L1, with morph breaks indicated by hyphens. The second line is a morphemic representation, where the L1 morphemes are given mnemonic names in the L2 metalanguage. The third line is the L2 free translation. By contrast, Toolbox interlinear glosses typically have (at least) four lines:

(8)	\tx[6]	*ùdúkwà*		
	\mb	*ù-*	*dúkwà*	
	\ge	3SG-	go\RLS	
	\ft	'He went'		

The first, third, and fourth lines correspond to the three lines in a traditional IMG. The second line `\mb` shares properties with both the `\tx` and `\ge` lines. Like the transcription it is written in L1, but like the gloss it is morphemic rather than morphophonemic. This extra line provides an opportunity to link tone patterns and glosses without disturbing the integrity of either the L1 morphophonemic transcription or the L2 morphemic gloss, as in (9):

(9)	\tx	*ùdúkwà*		
	\mb	*u-*	*dukwa-*	L H L
	\ge	3SG-	go	RLS
	\ft	'He went'		

The resulting annotation is concise, intuitive, and makes clear the link between the L H L tone pattern and realis mood. This solution relies on the fact that although the Toolbox `\mb` line is populated with the *names* of morphemes, the fact that these names are written in L1 serves as a pointer to the form of the corresponding morph represented in the `\tx` line. Ideally the tone patterns represented in the `\mb` line should remain in a one-to-one relationship with the gloss. For example, the realis forms of monosyllabic Cicipu verbs have a falling (HL) tone on the verb root, and so the tone on the verbal word (including the subject prefix) is L HL. Similarly for trisyllabic verb stems the pattern is L H L L. These tone patterns should both be represented as L H L in the `\mb` line to avoid mixing morphemic and morphophonemic representations on the same tier.

In certain cases the discrepancies between the 'citation' form of the tone pattern and its realisation in a particular example may be too great for the `\mb` line to serve as a reliable mnemonic, in which case a representation in the format of (9) may end up obscuring rather than elucidating the link between

[6] This paper uses standard Toolbox field markers: \a = alternate form, \ft = free translation, \ge = gloss (English), \lx = lexeme, \mb = morpheme break, \tx = text, \u = underlying form.

tone and meaning. We can however note that certain properties of tonal systems (e.g. the no-crossing constraint) will limit the extent of this problem.

Before turning to how such annotations can be produced using Toolbox, it should be admitted that the format in (9) is not compatible with all morphological theories, since it treats the realis tone pattern as a 'suprasegmental morpheme'. Such analyses were common under the Item-and-Arrangement morphological model favoured by American structuralists (Hockett 1954, Matthews 1974:79), but are less popular today. Nevertheless it is not always easy to see how the insights of more modern theories can be concisely represented in interlinear form. As Lieb and Drude (2000) point out, the interlinear format itself is inherently biased towards the Item-and-Arrangement model.

4. Buseman's method

The interlinear format set out in (9) (minus field markers) does not presuppose the involvement of Toolbox. It could equally well be used in the production of interlinear texts by other means, and it seems a sensible presentation format to use in descriptive works where the contribution of grammatical tone is at issue. Nevertheless this particular format has been considered here because it is the one generated using a technique for interlinearising grammatical tone devised by Alan Buseman[7]. Buseman's technique allows annotations such as the one shown in (9) to be generated without sacrificing Toolbox's semi-automatic interlinearisation, a facility that makes the program indispensable to many linguists. The method is briefly summarised here, and then evaluated in section 5.

The main difficulty in parsing tone through Toolbox is that although tones can vary independently of the segmental material to which they are attached, Toolbox by default understands them as an intrinsic part of the string of characters[8]: for example *dùkwá* is composed of d + u + ̀ + k + w + a + ́. The challenge is to extract from such a string a morpheme break line of *dukwa* ('go') + L H (IMP). The crucial first step of Buseman's method is to add tone marks to the lexicon as both suffixes and infixes. So for example L will be represented as:

[7] http://www.sil.org/computIng/toolbox/extras.htm (posted in January 2007).

[8] Using combining diacritics rather than pre-composed characters is essential for Buseman's method to work.

(10) \lx -L
\a -̀-
\a -̀
\ge tone.mark[9]

This technique relies on the fact that when Toolbox detects an infix, it extracts it out to the end of the word (by default), and so if we add lexical entries such as (10) for both low and high tones then *dùkwá* will be parsed as:

(11)	\tx	*dùkwá*		
	\mb	*dukwa*	-L	-H
	\ge	go	-tone.mark	-tone.mark
	\ft	‘go!’		

The final step is to register the combination of L + H as a named tone pattern, in this case the imperative, by adding a further lexical entry as in (12). The resulting interlinear parse shown in (13) is in precisely the same format as (9).

(12) \lx -L -H
\ge IMP

(13)	\tx	*dùkwá*	
	\mb	*dukwa*	-L -H
	\ge	go	IMP
	\ft	‘go!’	

5. Evaluation

This section provides an evaluation of the technique just described, based on its application to over six hours of transcribed and tone-marked spoken Cicipu texts. Overall the technique has been successful, and I continue to use it for

[9] Care must be taken when typing the `\a` forms in Toolbox, since the program superimposes the accents on top of the hyphens. There must be no spaces in these fields.

interlinearisation. There were however certain difficulties in applying the technique, as well as more general issues. These will now be described in turn.

5.1 One-infix-per-word restriction

For performance reasons Toolbox, by default, contains a restriction that allows only one infix per word to be parsed. Cicipu verbs ending with a digraph in the transcription therefore failed to parse, for example items ending in diphthongs, or long or nasal vowels (written with an *n* following the vowel) e.g. *ù-tínàa* ('he swore') and *ù-wónsòn* ('it barked'). As a workaround, the last accent in such words can be moved to the end of the word, as in *ù-tínaà*. Fortunately Cicipu roots only have CV structure, but for languages with CVC syllables it may look ungainly to place tone marks over coda consonants.

Trisyllabic roots initially failed to parse for the same reason as forms ending with a digraph. This is because the first two tones fall in the middle of the string of characters, but only one of them can be handled as an infix. This time the problem cannot be solved by shifting accents, but requires alternate forms (\a) to be added to the lexical entries of all trisyllabic roots, with the middle vowel marked for tone (e.g. *jungònu* 'shut'). The form *júngònù* will then be parsed as *jungonu*-RLS, with the first and third tones being glossed as the realis tone pattern H L, the second tone being already included in the alternate form. Adding the extra forms was less arduous than it might have been otherwise since few Cicipu roots have more than two syllables, and there are only two contrastive tones to consider. For languages with more complex word-structure or tone systems, many more entries may be required.

It was mentioned above that the 'one-infix-per-word' restriction is the default behaviour of Toolbox. In fact, in 2008 a new version of the software was produced which optionally allows multiple infixes[10]. I have found this very helpful, although even on a well-specified computer it can slow down parsing considerably, depending on the morphological complexity and the number of homonymous affixes/tone patterns in the language[11]. Consequently it may not be possible to use this option for some Toolbox projects.

[10] In the dialog used to modify the interlinear parse process there is a check box called 'Allow multiple infixes'. It is unchecked by default.

[11] Word formulas (§5.2) do not assist with parsing performance in Toolbox, since the candidate still has to be parsed before it can be rejected.

Languages with ‘real’ infixes are especially problematic given the one-infix-per-word restriction, and they cannot be easily handled, even in the new version of the program. Cicipu has both iterative <*il*> and causative <*is*> verbal infixes, and verb-forms containing these cannot be parsed without manual intervention. Buseman has proposed a solution which involves applying a simple batch program (CCT consistent changes table) prior to interlinearisation, in order to extract the tone marks from the text and place them at the end of each word. So for example *sùkùlìsú* (‘cause something to move’) becomes *sukulisu*ˋ ˋ ˋ ˊ, which would then be parsed as *sukulu*-*is*-IMP (‘move-CAUS-IMP’) without the need to treat any of the tone marks as infixes.

The problem with this method is that the batch program has no way of referencing parts of speech. Therefore the process is applied indiscriminately to all words of all lexical categories, which is unlikely to be appropriate – unless, of course, only grammatical tone is marked in the transcription and not lexical tone. This might be the case for an orthographic transcription (see Bird 1999 for orthographies of Kakɔ and Etung which mark only grammatical tone), in which case this technique will be helpful. However it will not work with the morphophonemic transcriptions which are generally recommended for language documentation.

5.2 Ambiguity

In the initial stages of using this method, the existence of non-unique tone patterns gave rise to problems with ambiguity. However this was overcome by subcategorising the ‘part of speech’ field for each tonal pattern in the lexicon, and then fine-tuning the use of Toolbox word formulas accordingly. In fact, this was a useful exercise in itself in terms of understanding of the grammar of Cicipu, quite apart from the benefits for interlinear parsing.

5.3 Content of the `\mb` tier

It was mentioned in section 3 above that a desideratum for IMGS in the format of (9) is that the `\mb` tier is morphemic rather than morphophonemic. However the technique being described here results in morphophonemic representations of tone patterns. So, for irrealis Cicipu verbs, say, the various values in the `\mb` tier will be HL, H L, H L L, H L L L, depending on the number of syllables in the word, whereas ideally we would like them all to resolve to H L. Unfortunately there does not seem to be an easy way around this. Toolbox does offer the `\u` and `\a` fields as a means of handling segmental allomorphy, but this resolution comes into effect too soon in the parsing process to be of any use here. In particular, it happens before sequences of individual tones are recognised as meaningful patterns as a result of the existence of lexical entries such as (12). Resolving the allomorphy by hand is not straightforward[12].

5.4 Complexity

At this stage of the Cicipu documentation project the interlinearisation set-up is stable and transparent to the user. Nonetheless the application of this technique amounts to a significant increase in the complexity of the interlinearisation process. This is not necessarily a problem for long-term language documentation – as Alan Buseman (pers. comm.) has pointed out, the final clarity of the interlinearised text is more important than how it was arrived at, since the hypothetical linguist of five hundred years time will not be using Toolbox[13]. It should nevertheless be made clear that any extra `\a` fields (see 5.1) in the lexicon were put there for technical rather than linguistic reasons. One way to do this would be to use a separate marker altogether for such alternate forms (e.g. `\at`), and to add this to the 'Markers to find' list for parsing.

6. Software recommendations

Some of the problems mentioned in section 5 above can be overcome or mitigated by adjustments within Toolbox, but others are more serious and would require a change to the program, or the development of new software. In this section I offer some general recommendations that can be made to the designers of future interlinearisation software.

First is a plea that developers consider the autosegmental model of phonology from the outset of program design. In particular, the option to specify a subset of characters (e.g. ˊˉˋ) to be treated independently of the remainder is highly desirable, so that tone marks could participate as a *separate* input to the parsing process (as was done in SIL's TonePars program, see Black 1997). Thus word-forms such as *ùdúkwà* could contribute two separate inputs to the parser, namely *udukwa* together with ˋ ˊˋ, as well as the unanalysed input *ùdúkwà*. This would allow the automatic parsing of verb

[12] If the aim is simply to create XML for archiving or use in another program, then the problem can be solved by either XSLT or Regular Expressions operating on the XML export from Toolbox. However there is currently no way to import the Toolbox XML export back into Toolbox.

[13] It may be more of a problem in the short-term if the solution proposed here has to be migrated to another program. FieldWorks Language Explorer is (for good reasons) designed to be less flexible than Toolbox and does not support the technique outlined here (Heidi Rosendall pers. comm.).

forms in Cicipu and similar phenomena involving purely grammatical tone. Handling interaction between lexical and grammatical tone would also be possible, because once the distinction has been made between the segmental and suprasegmental representations of word-forms and lexemes, the ability to specify rules which handle the interaction of lexical and grammatical tone reduces to the familiar Toolbox technique of setting up underlying/alternate forms. Naturally these rules could be constrained according to the principles of autosegmental phonology, although this should not be the default behaviour of the program.

The second recommendation concerns the content of the interlinearised text. As was noted in 5.3, the technique described in this paper populates the L1 'intermediate' tier (\mb) with an unfortunate mix of two different kinds of representations: the usual morphemes, plus the tone patterns, which are actually more like suprasegmental 'morphs'. Future interlinearisation programs should make it possible to create a unitary morphemic tier able to contain both straightforward lexemes and morphemic tone patterns[14].

Thirdly, while arguing for a certain amount of tonal sophistication in the software, I would also recommend that input methods should not make undue demands on the user, so that it is feasible to apply the technique to all texts to be interlinearised rather than just a subset. For example, a linear tone-marked transcription of the kind shown in the \tx fields in this paper should be an acceptable input. It should also be possible for the user to set up the rules required in a simple format without making it necessary for the user to deal with autosegmental formalism. This paper and the recommendations made here are not concerned with tonal analysis, for which Toolbox is not an appropriate tool, but with the representation of tonal patterns in the annotations included in documentary corpora. With that in mind, it should be acceptable to sacrifice a certain amount of theoretical elegance in order to avoid complexity in the annotation technique

[14] In Toolbox one way to avoid this mix, but retain the explicit link between tone pattern and meaning, would be to allow the option for the intermediate tier to be wholly *morphophonemic* rather than morphemic and hence populated with allomorphs rather than morphemes. Even in the case of allomorphs which are quite distinct from the citation form of the morpheme, the link to the morpheme would still be recoverable from the mnemonic properties of the metalanguage gloss, and of course this is exactly how standard (i.e. non-Toolbox) IMG functions. Although the outputs from the Toolbox parsing and glossing processes are configurable, certain technical difficulties mean that the program is currently unable to produce such a tier.

7. Summary

This paper has considered a number of options for the interlinearisation of utterances involving grammatical tone. The guiding assumption has been the view that the annotation accompanying a documentary corpus should "represent the language for those who do not have access to the language itself" (Lehmann 2001:88), and it was argued that strict adherence to the conventions in Lehmann (2004a) and Bickel, Comrie, and Haspelmath (2004) can obscure the linguistic structure of constructions involving grammatical tone. There is a tension between the readability of annotations and their feasibility of creation with respect to timescales on the one hand, and the degree of linguistic structure represented therein on the other. The technique devised by Alan Buseman is suggested as a compromise between more complex techniques such as Advanced Glossing which make detailed linguistic annotation possible at the expense of visual compactness and speed of processing, and simpler techniques which fail to capture the link between tone patterns and meanings at all. Although as pointed out in section 5 above there are deficiencies with the method, it is nevertheless proving valuable for the on-going documentation of Cicipu. As well as the recommendations made to software developers in the previous section, I hope that this paper will benefit other linguists working on tone languages by helping them to judge the applicability of Buseman's method to their own documentary corpora.

References

Bickel, Balthasar, Bernard Comrie & Martin Haspelmath. 2004. *The Leipzig glossing rules. Conventions for interlinear morpheme by morpheme glosses*. Leipzig: Max-Planck-Institut für Evolutionäre Anthropologie.

Bird, Steven. 1999. Strategies for representing tone in African writing systems. *Written Language and Literacy* 2, 1-44.

Black, H. Andrew. 1997. TonePars: A computational tool for exploring autosegmental tonology. *SIL Electronic Working Papers 1997-007*. http://www.sil.org/silewp/1997/007/SILEWP1997-007.html.

Clements, G. N. 2000. Phonology. In Bernd Heine & Derek Nurse (eds.), *African languages: an introduction*, 123-60. Cambridge: Cambridge University Press.

Crozier, David. 1984. *A study in the discourse grammar of Cishiningi*. PhD dissertation, University of Ibadan.

Drude, Sebastian. 2003. Digitizing and Annotating Texts and Field Recordings in the Awetí Project. Paper presented at Third EMELD conference on "Digitizing and Annotating Texts and Field Recordings", Lansing, Michigan. http://www.emeld.org/workshop/2003/paper-drude.html.

Ernst, Urs. 1996. Tone Orthography in Kakɔ (Kakɔ East). Unpublished ms. http://www.sil.org/Africa/Cameroun/bydomain/linguistics/orthography /kako_ernst1996_2113_p.pdf.

Goldsmith, John. 1990. *Autosegmental and metrical phonology*. Oxford: Blackwell.

Hockett, Charles. 1954. Two models of grammatical description. *Word* 10, 210-34.

Kisseberth, Charles and David Odden. 2003. Tone. In Derek Nurse and Gérard Phillipson (eds.) *The Bantu Languages*, 59-70. London: Routledge.

Lehmann, Christian. 2001. Language documentation: a program. In Walter Bisang (ed.) *Aspects of typology and universals*, 83-97. Berlin: Akademie Verlag.

Lehmann, Christian. 2004a. Interlinear morphemic glossing. http://www.uni-erfurt.de/sprachwissenschaft/personal/lehmann/CL_Publ/IMG.PDF.

Lehmann, Christian. 2004b. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan & Stavros Skopeteas (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2. Halbband,* 1834-57. Berlin: Mouton de Gruyter.

Lieb, Hans-Heinrich & Sebastian Drude. 2000. Advanced glossing: A language documentation format . http://www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf.

Matthews, P. 1974. *Morphology*. Cambridge: Cambridge University Press.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelman and Ulrike Mosel (eds.), *Essentials of language documentation*, 213-52. Berlin: Mouton de Gruyter.

Wittenberg, Peter. 2003. The DOBES model of language documentation. In Peter K. Austin (ed.), *Language Documentation and Description,* Vol 1, 122-139. London: SOAS.