Language Documentation (LD) has generated a wealth of textual resources for minority languages over the past two decades. These resources are generally created in the form of digital-born texts, which sometimes raises the issue of which technology is best suited for reaching a compromise between proper orthography of previously unwritten languages on the one hand, and computability of existing text corpora on the other hand. Textual resources stemming from LD should be designed in such a way that they "represent the language for those who do not have access to the language itself" (Lehmann 2001:88). In the case of Bantu languages, most of which are tone languages, this entails representation of tone sequences as anchored in the melody pattern of speech production. The bundling of tone and letters into pre-composed or ad hoc binary or ternary graphical units in mainstream orthographic practices as well as in the creation of text resources for endangered languages, actually adds an additional layer of digital information whose interpretation by the computer is not necessarily anticipated by most LD practitioners ; e.g. the unique pre-composed character **<é>** will be interpreted as not a single character, but as a combination of two separate elements of grapheme and diacritic.

Within both the scholarly community and speech groups of minority languages, there is currently much emphasis on developing typesetting technologies such as virtual keyboards and extension of Unicode character sets, which aim to facilitate the creation and visualization of text resources which use indigenous or IPA-based writing symbols. In IPA-based writing systems, tones are represented supra-segmentally; they stand on an upper graphical tier above the segmental tier as in (1), an example from the Basaa language spoken in the center and coastal regions of Cameroon.

(1) mà-lép má ń-sóbì
CL6-water PERF-pour
"water is poured"

The above example is made up of a sequence of six Tone Bearing Units (TBU) which are host to an equal number of tone realizations at the surface level from a purely structuralist perspective, namely L-H-H-H-H-L. It should be noted, however, that surface realizations of tones as perceived and interpreted by human cognition may differ from their deeper structure, as will be discussed below.

Tone representation as supra-segmentals or diacritics is a convenient graphical method with regard to human cognitive ability to process visual information, but much less convenient from a computational perspective. This is so because :

1) Graphical clustering of tone markers (acute or grave accent in (1)) with letters (a, e, n, o, i) is interpreted as a single unit of writing by humans, but as two separate digital objects by the computer, as Latin characters and accents bear distinct digital codes[1] ;

2) Deciphering of pitch associated with a given tone, whether High, Low, rising or falling by humans does not require intrinsic knowledge of lexical and grammatical information encoded by the tone marker. In (1) for example, High tone marking on <n> in <ńsóbì> signals perfective aspect; however, this grammatical knowledge is not a prerequisite for proper decoding and pronunciation of the nasal prefix associated with the stem <sobi>. For the computer, however, being able to retrieve the grammatical information encoded through the nasal prefix and the High tone associated with it requires explicit and non ambiguous one-to-one mapping of meaning and form.

[1] See the most current Unicode character map at: http://www.unicode.org/Public/UCD/latest/charts/CodeCharts.pdf

3) Tones in African languages are not merely supra-segmentals, but indeed auto-segmentals (Goldsmith, 1990); they form independent units from the segments which bear them. This can be seen in (2), where the tone melody on the verb root <sɔb> "pour", is modified to form a contour rising tone (LH).

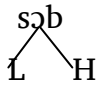
(2)

sɔ̌b	mà-lép
pour.IMP	CL6-water

"pour water !"

The contour tone on the verb root <sɔb>, is the result of the combination of a lexical low tone with a grammatical high tone which signals imperative mood. An auto-segmental representation of the tone melody on the verb root <sɔb>, yields two distinctive association lines linking both Low and High to an identical TBU, as in (3).

(3)



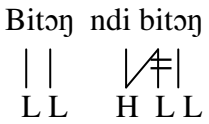
Other tone variation phenomena are found in Bantu languages, which need explicit semantic markup in corpora building for accurate computability of text data; this is true in cases of tone spreading (4) and downstep (6), among others. Example (4) is drawn from the Bati language belonging to the Mbam group in the central region of Cameroon, whereas (6) is another example from Basaa.

(4)

bì-tɔ̀ŋ ndí bí-tɔ̀ŋ
cl8-village and cl8-village
"village after village"

Tone spreading is manifested through propagation of the High tone on the coordinating item <ndi> over the TBU of the noun prefix in the word < bi-tɔŋ> in (4). As a result, the lexical Low tone of the noun prefix is replaced with the spreading High tone. An auto-segmental representation of this process further shows delinking of the lexical Low tone as in (5).

(5)



Tone spreading as illustrated in (4) through (5) shows that a single tone can stand on more than one TBU. Therefore, what may surface as two or more consecutive High or Low tones on adjacent TBUs could indeed stand as a single tone which holds on more than one TBU. Most situations involving adjacent Highs or Lows in Bantu languages are better accounted for by the OCP (Obligatory Contour Principle), which states that adjacent identical tones are disallowed (Leben, 1973).

Graphical marking of tones in writing as in (4) might be appropriate for on-line mapping of pitch variation in speech melody, but could prove inefficient for semantically representing complex tone patterns. This is further illustrated in tone variation as a result of downstep. Downstep is an

intonational phenomenon whereby one or more High tones are lowered after a preceding High tone, as in (6), where the downwards arrow signals downstep.

(6) á ꜜɲɔ́
1.SM.COND drink
"Should he/she drink"

The subject marker (SM) in (6) is toneless underlyingly; however, this morpheme is assigned a High tone which marks conditional mood. The conditional marker (H) on the SM surfaces with a higher pitch than the High tone on the verb <ɲɔ>, which results in the latter being downstepped.

Most software used in the compilation of text corpora by language documentarians are tools which have been designed for general-purpose usage. They are not always meant to address the structural peculiarities of every specific language to be documented, particularly as languages which are of interest to current LD investment are also those about whose structural properties little is known. Two of the most widely used tools in for LD are Toolbox and FLEx. These pieces of software allow for automatic glossing of grammatical features, as grammatical information is stored incrementally in the backend database. In addition to automatically parsing grammatical elements at the structural level, Toolbox is able to parse lexical and grammatical information encoded through tone marking thanks to a specific data input method designed by Buseman, as cited in McGill (2009). Buseman's method has been chiefly motivated by the need to overcome the software's understanding of tones as intrinsic parts of the string of characters. Overall, Buseman's method consists in separating tone marking and the corresponding segmental string of characters triggering TBUs, and then glossing them each on their own, as in (7)[2], which I have borrowed from McGill (2009: 244).

(7) \tx dùkwá
\mb dukwa -L -H
\ge go IMP
\ft 'go'

Enhancement and enrichment of Buseman's method has been suggested by McGill (2009) in view of further development of Toolbox as well as development of new software. Among other suggestions, McGill strongly advocates the implementation of SIL's TonePars program (Black 1997) into future software development for corpora creation for endangered languages. "The TonePars program […] allow for modeling an auto-segmental approach to tone"[3].

While further refinement of Toolbox (or FLEx) to include auto-segmental modeling of tone along with multiple-tier representations of character strings, tone and semantics might offer a workable and interesting solution for automatic glossing of text resources for tone languages in Africa, and despite the ability of Toolbox (and FLEx) to generate XML files for multiple purposes and multiple platforms, Toolbox-made corpora do not yet conform to current standards of corpus building in the digital humanities, for at least three reasons:

1) Text representation and processing of tone languages with Toolbox is usually biased towards an exclusively scholarly stance, over other possible language usage models, where tones might not be represented at all. As a matter of evidence, it should be noted that non-marking of tones in

[2] Data in this example come from Cicipu, a Niger-Congo language spoken in Northwest Nigeria. Backslash is standard markup signaling the Toolbox input field; '\tx' is the transcription field; '\mb' is the morpheme-by-morpheme field; '\ge' is the English gloss field; and '\ft' is the free translation field. Each field stands on a separate tier.
[3] Weblink: http://www-01.sil.org/silewp/1997/007/silewp1997-007.html, accessed on 14 October 2016.

orthography in tone language communities is the norm rather than the exception. An edifying illustration is the case of Google Translate, where none of the five African tone languages which are available for online translation (namely Hausa, Igbo, Shona, Xhosa and Yoruba) uses tone marking in their writing system. Plain Latin scripts are preferred over the IPA-based writing systems mostly advocated by linguists.

2) Assuming TonePars algorithms are incorporated into Toolbox, the issue of character encoding is left unresolved. TonePars interprets the content value associated with a given tone, whether High or Low, on the basis of its graphical shape, namely acute or grave. However, there may exist multiple character input possibilities for representing the same toned-segment graphically. For example <é> (the character <e> bearing and acute accent), may have as possible typesetting inputs:

a) The unique pre-composed character **<é>** (*Unicode C1 Controls and Latin-1 Supplement*, code point 00E9);

b) **<e>** (*Unicode C0 Controls and Basic Latin*, codepoint 0065) and acute accent **< ´ >** (*Unicode C1 Controls and Latin-1 Supplement*, codepoint 00B4);

c) **<e>** (*Unicode C0 Controls and Basic Latin*, codepoint 0065) and acute accent **< ´ >** (*Unicode Spacing Modifier Letters*, codepoint 02CA);

d **<e>** (*Unicode C0 Controls and Basic Latin*, codepoint 0065) and acute accent **< ´ >** (*Unicode Combining Diacritical Marks*, codepoint 0301);

etc.

In other words, the existence of multiple graphical representations of the same prosodic reality creates the possibility of arbitrary representation of tones in a program such as Toolbox. This is definitely not good practice with regards to current standards in data sharing and dissemination as advocated by such data consortia as the Text Encoding Initiative (TEI), the Data Research Alliance (RDA), Component MetaData Infrastructure (CMDI), etc.

3) As a consequence of the above two issues, text corpora created with tools such as Toolbox and FLEx cannot lend themselves appropriately to open data sharing and re-use, in a world where sustainability of data infrastructures rely heavily on interoperability.

While the TEI (Text Encoding Initiative) provides a comprehensive framework for the development of text encoding standards and schemes for virtually any living language, specific issues in representing textual information in African languages have not yet been accounted for. This situation hampers the development of reference infrastructures for sharing and re-use of language data in Africa, such as multilingual language corpora using TEI standards.

This presentation aims to raise awareness of the increasing marginalization of African languages in the global textual data infrastructure, and to suggest possible solutions to reverse this negative trend. Some of the suggestions for inversing the trend in textual resources production are as follows:

- Language development should no longer only entail book literacy development and language corpora for linguistic research, but also digital anchoring of textual resources and development of digital textual infrastructures to meet the requirements of digital data dissemination, sharing, and re-use;
- Adoption of mainstream writing systems for yet-to-be-developed languages, such as Latin scripts, devoid of superscripts and diacritics, for use in the web and in digital devices (phone texting, etc.), and complementing them in corpora building with TEI-based markups for tones and grammatical information where necessary; this could allow for both automatic retrieval and

processing of linguistic information on the computational interface of text corpora, while making the user interface more flexible and more accessible as with Google Translate.

- Development of text encoding schemes for semantic markup of grammatical and prosodic features such as tones, aspiration, labialization, vowel lengthening, etc. in compliance with the TEI framework.

References

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.

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

Leben, William. 1973. *Suprasegmental phonology*. Ph.D. thesis, MIT, Cambridge, MA.

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

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. London: SOAS. pp. 236 – 250.

TEI Guidelines. Weblink: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.