Case Study from Africa: Use of Latin Script for African Languages

Meikal Mumin, M.A. (meikal.de/profile)
Institute for African Studies and Egyptology, University of Cologne
&
Department of Asian, African and Mediterranean Studies,
"L'Orientale" University of Naples

The use of Roman/Latin Script in Africa

The use of RLS in Africa

- Today, Roman/Latin script (RLS) is most widely used writing system in Africa
  - It is estimated that around 2000 languages are spoken in Africa
  - However, only around 500 out of these 2000 languages are written in established orthographies (Bendor-Samuel 1996: 689)
- Orthography development is an ongoing process and new orthographies continue to be developed for further languages
  - Other scripts used in Africa include the Arabic script, the Ethiopic script, Tifinagh, N‘ko, the Vai syllabary, or the Mandombe script

Representation of African Languages in RLS

- By design, RLS is an alphabet. As such it represents consonants and vowels equally
- African languages make use of different consonants and vowels than European languages
- Since the mid-19th century, the RLS was developed systematically to represent African languages
  - Over time, written representations of speech sounds became harmonized with most orthographies used today re-employing at least some conventions of the International Phonetic Alphabet (IPA)
- RLS had to be significantly extended and modified to represent African languages

The development of orthographies for African languages

- While first attempts to write African languages were undertaken since the 16th century, orthography development became more common only by the late 18th century (Pasch 2008)
  - Orthography development was conducted first by Christian missionaries and later on mostly by Western linguists
- All of these were trained in classical languages of Europe (e.g. Latin, Greek or less frequently Hebrew, Sanskrit, or Arabic) and speakers of Western European languages
  - All languages (including those) only have limited repertoires of consonants, vowels, and other relevant features
  - By consequence, also RLS was designed and extended first to express features common to such languages

From individual representations to harmonized transcriptions

- By the time, Christian missionaries and European linguists realized the variety of speech sounds in Africa they started to harmonize written representation of African languages
  - Most harmonizations were intended for scientific transcription instead of practical orthographies, but were re-employed nonetheless
  - Early examples include the Standard Alphabet by Lepsius, or the Africa Alphabet from 1928
  - Even though such Pan-African alphabets are mostly out of use today, such aspirations for unified representations of African languages didn‘t stop as for example in the case of the African reference alphabet developed at a UNESCO-organized conference in 1978 (revised 1982)

Representing languages individually or collectively

- Most important languages of Africa now have individual (or several) orthographies
  - Particularly in countries with a large number of languages, such as Nigeria (over 527) or Cameroon (over 284), or for languages spoken across national borders such as Fula (spoken across some 20 countries of Western Africa and Central Africa) National or regional orthographies exist next to individual orthographies
- Examples include the
  - Berber Latin alphabet (since 19th century)
  - Guinean languages alphabet (until 1989)
  - General Alphabet of Cameroon Languages (since 1979)
  - Pan-Nigerian alphabet (since 1981)
  - l'Alphabet National du Tchad for Chadian languages (since 2009)

Competing orthographies

- Frequently, individual languages were written in several orthographies, often in different scripts at the same time
- Occasionally, there are also different orthographies based on the same script used by the same languages today
  - This is the case particularly with languages spoken across national borders
  - Examples include Tuareg/Tamasheq (Berber), is written differently today, depending on wether it is used in Niger, Mali, or Burkina Faso, and the consonant /ʒ/ for example is written as <ž> in Niger and Burkina Faso but as <ʒ> in Mali

Areal/Regional strategies of orthographic representation

- Local orthographies have frequently reemployed commonly known conventions
  - Such conventions were often based on colonial languages, such as French, English, or Portugese, or other commonly used languages, such as Amharic
  - For example where French was a colonial language or lingua franca, <ou> was used to represent either /o/, /u/ or /w/, e.g. <Oumari> or <Ouagadougou>
  - Where English was a colonial language, the same speech sounds are represented as simple <o>, <u>, or <w>, e.g. <Omari> or <Watamu> (a town in Kenya)
- While this was much more common with vowels, the same can also be observed with consonants
  - Word-initial /k/ was represented as <k> in English colonies such as <Kenya>, but as <c> as in <Cameroon> as in French

Representing orthographies in Internationalized Domain Names

Encoding orthographies in IDNs: IDNA 2008

- Internationalized domain names for applications (IDNA 2008/IETF RFC5895) is the underlying standard for internationalized domain names
  - IDNA 2008 is based on Unicode 6.3
- IDNA is effectively a subset of Unicode since it restricts the use of Unicode code points
  - Code points are classified as Protocol Valid (PVALID), Contextual Rule Required (CONTEXTJ/CONTEXTO), Disallowed (DISALLOWED), or Unassigned (UNASSIGNED)
  - Only the former two can be used in IDNs, which leaves 97,973 code points which are classified as PVALID or CONTEXTJ/CONTEXTO
- IDNA permits only letters, digits, and hyphens (LDH) for use in IDNs
  - By consequence, any non-LDH code point used in orthographies is excluded from use in IDNs

IDNs in the root zone: From IDNA 2008 to MSR-2

- For the root zone, Integration Panel (IP) has defined the Maximal Starting Repertoire Version 2 (MSR-2), which is a subset of IDNA 2008, containing 33,490 code points from 28 scripts
- Only code points contained in MSR-2 can be included in label generation rules (LGRs) for the root zone
- Label generation rules are defined for entire script using communities, such as all languages using Latin/Arabic/Ethiopic script
  - Arabic and Ethiopic script using communities have already submitted LGRs, which probably leaves only Latin among scripts relevant to Africa, before the addition of other relevant script in any future revisions of MSR

Restrictions on IDNs of the root zone from MSR-2

- However, MSR-2 has restricted this to letters only
  - „Letter Principle in the [Procedure], which states that only code points exclusively used in writing words are to be included in root zone labels. There are the occasional orthographies that use digits and punctuation as part of words. Where these code points do not occur exclusively inside words, they are prohibited by the Letter Principle.“ (MSR-2: 18)
- By consequence, any digits or hyphens used in orthographies cannot be included in Label Generation Rules for the root zone
  - In Africa, no cases come to mind for RLS but in other areas of the world digits for example are used to write tones

Representing phonological features of African languages in RLS

Representing Consonants in RLS

- Numerous types of consonants used in African languages occur only rarely or not at all in European languages making use of RLS
  - Certain types of consonants such as ejectives occur only rarely in European languages (e.g. Georgian which traditionally makes use of Georgian script)
  - Other types of consonants such as clicks, implosives, labio-dentals, or pre-nasalized consonants do not occur in European languages
  - Few types of consonants are even-near exclusive to African languages such as labial flaps or labio-velar stops
- Apart from consonants, numerous suprasegmental features such as distinctive aspiration of stops are frequently encountered in African languages

Representing Ejectives and Implosives in African Orthographies

- Ejectives are commonly represented by the IPA strategy of an combining apostrophe, e.g. <ts’>, or <k’>, as for example in Sandawe (Isolate, Tanzania)
  - In some orthographies, such as Xhosa (Nguni, South Africa), such consonants are not orthographically distinguished, /p’/ <p>, /t’/ <t>, /k’/ <k>
  - In other orthographies, such as Hadza (Isolate, Tanzania) double letters are used, e.g. <tt> /t’/, <kk> /k’/, <qq> /q’/
- Also implosives are often represented using the IPA-strategy of hooked letters as in Fula (Senegambian, used in numerous countries), e.g. <ɓ> /ɓ/ or <ɗ> /ɗ/
  - There are however irregularities as in Hausa (Chadic, Nigeria), where <ɓ> represents the implosive /ɓ/, but <ƙ> represents the ejective /kʻ/
  - In other languages, implosives are represented by sequences of two letters (digraphs), as in Igbo (Volta-Niger, Nigeria), which represents /ɓ͡gb/ as <gb>

Combining marks in African orthographies

- Most consonantal distinctions are expressed by combining diacritical marks, including dots <◌̣>, curves <◌̃>, or accents <◌̌>, or superscripted <ʷ> letters
  - The Pan-Nigerian alphabet designed to represent hundreds of languages for example uses combining marks below (in both upper and lower case forms) on both vowels, i.e. < ị>, <ẹ>, <ọ>, <ụ> and consonants <ṣ>
  - Igbo (Benue-Congo, Nigeria) for example uses a combining mark above to distinguish nasals, i.e. <n> /n/ vs. <ṅ> /ŋ/
  - Berber Latin Alphabet used superscript <ʷ> as in IPA to represent labialized letters, e.g. <bʷ>, <gʷ>, or <qʷ>
  - Supposedly, even combinations of two combining marks may be used in African orthographies to represent consonants, however it is difficult to find supporting evidence

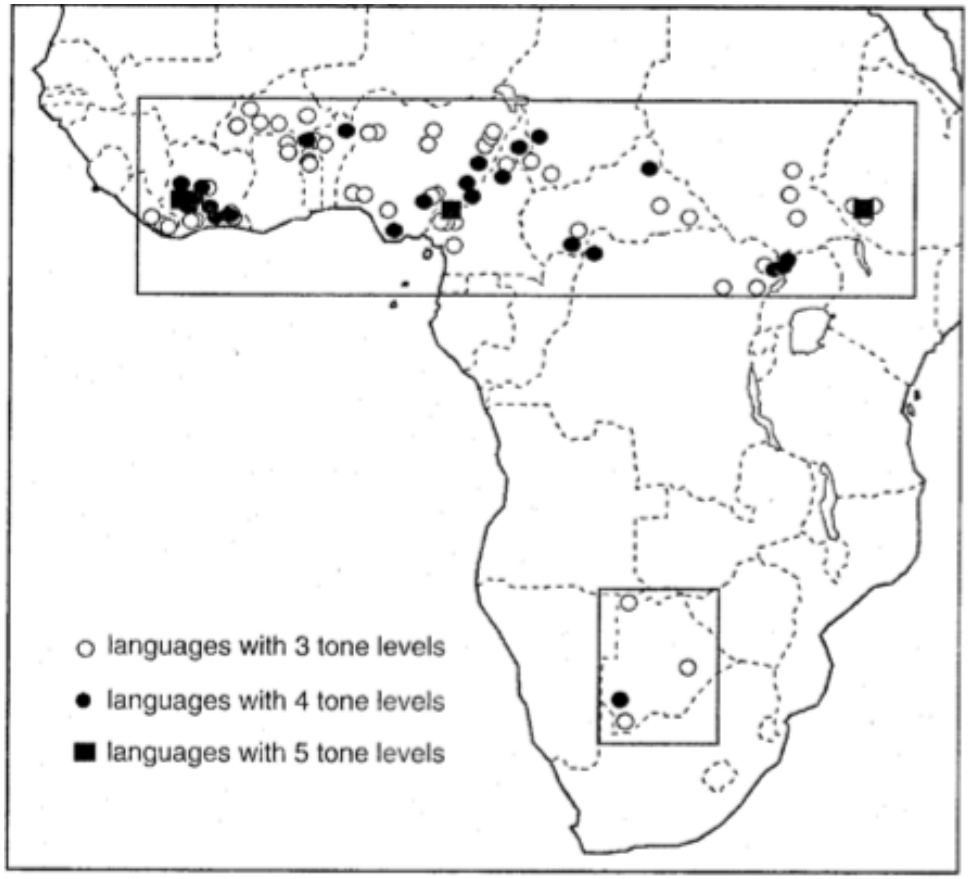
Representing Clicks in African Orthographies

- Clicks are often represented using the IPA symbols <ʘ> for labial, <ǀ> for dental, <ǂ> for palato-alveolar, <ǃ> for (post)alveolar and <ǁ> for lateral, as for example in Khoekhoe (Khoe, Namibia), e.g. <lhubu> (or <lhuwu>) /lʰùwú/ 'to stop hurting'
- In most other cases, sequences of several letters (multigraphs) are used to represent clicks orthographically, as for example in Zulu (Nguni, South Africa), e.g. <gc> for /ᵍǀʰ/ in <isigcino> /isiᵍǀʰiːno/ ‘end’, or <gq> for /ᵍǃʰ/ as in <uMgqibelo> /umᵍǃʰibeːlo/ ‘Saturday’

Representing Vowels and Tone in RLS

- Most African languages have between 5-7 vowels (Gregersen 1977) and are tone languages (Clements & Rialland 2008)
  - For example, Ịjọ languages of Southern Nigeria frequently have a 9 vowel system typically rendered as <i, ɪ, e, ɛ, a, ɔ, o, ʊ, u> (Williamson 2009: 128), while Ngwe (Grassfields, Cameroon) has at least 13 vowels
- In orthographies, tone is usually represented using combining marks, such as accents on letters, such as á, à, â, ǎ, or ā
  - Expressing tones plus other features has often led to a large number of vowel graphemes including combinations of several combining marks, as in Yoruba (Volta-Niger, Nigeria), which has up to 31 vowel graphemes, i.e. <á, à, ā, é, è, ē, ẹ, ẹ, ẹ́, ẹ́, ẹ̀, ẹ̀, ẹ̄, ẹ̄, í, ì, ī, ó, ò, ō, ọ, ọ, ọ́, ọ́, ọ̀, ọ̀, ọ̄, ọ̄, ú, ù, ū>
- Apart from vowels and tones, numerous other suprasegmental features such as nasalization of vowels, or vowel harmony based on the tongue root, are frequently encountered in African languages

Representing Vowels and Tone in RLS



Map 3.6 Distribution of 76 African languages with three or more distinctive tone levels. The two major concentrations are enclosed in rectangles. Languages with three, four, and five tone levels are indicated by white circles, black circles, and black squares, respectively

Distribution of Tone in Africa (Rialland & Clements 2008:73)

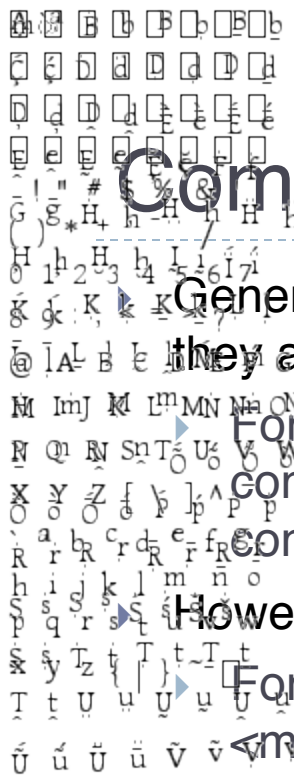
Further strategies of representing Tone

- Occasionally, also other strategies are used, including
  - Double consonants as in Jur Modo (Nilo-Saharan, Sudan; Roberts 2013: 7), e.g. <nï> [nì] ‘her’ vs. <nnï> [ní] ‘their’
  - Punctuation marks as in Budu (Bantu D.30, Democratic Republic of Congo; Roberts 2013: 9) e.g. <wabenda>[wàɓɛ́n͡dà] ‘You hit’, <wa=benda>[wàɓɛ́n͡dā] ‘You will hit’ <wa:benda> [wǎɓɛ̀n͡dà] ‘You have hit’
  - Consonants as in Ngangam (Gur, Togo; Roberts 2013: 7)<bere> [beɻre˥] ‘destroy (imperative)’ <bereh> [beɻre˥] ‘destroy (imperfective)’
- While the use of sequences of codepoints, such as <nn> or <eh> pose no problem, the use of ‚punctuation marks‘ such as Budus <wa=benda> are blocked by the LDH principle of IDNA

Representing African orthographies in IDNs

Representing consonant graphemes in IDNs

- Generally, such graphemes used to represent consonants can be employed in IDNs
  - Sequences of code points (multigraphs) such as
    - digraphs (2 elements), such as <ny> representing /ɲ/ in Swahili (Bantu, East Africa)
    - trigraphs (3 elements), such as <ngq> representing /ᵑǃʰ/ in Xhosa (Nguni, South Africa)
    - tetragraphs (4 elements), such as ⟨thsh⟩ representing /tʃʰ/ in Xhosa (Nguni, South Africa)
  - are generally not a problem in the context of IDNs
    - Hypothetically, the maximum string length of labels of 255 in ACE-encoding (A-Labels), and by consequence in UTF-8-encoding (U-labels) to 59 characters, may pose a problem with agglutinating languages or orthographies making use of tri- and tetragraphs



Combining marks in IDNs

- Generally, most combinining marks are encoded together with those letters they are commonly used with
  - For example, Berber Latin Alphabet represents /ðˤ/ as <ḍ>, which is a combination of <̣> (U+0323) and <d> (U+0064) existing also in a pre-composed form (U+1E0D)
- However, some sequences may not exist (as of yet) in a pre-composed form
  - For example, Jukun (Benue-Congo, Cameroon), uses a combination of <m> and <̀>, i.e.<m̀>, which does not exist in pre-composed forms
    - In the case of vowels, several combinations of two (and possibly three) combining marks are encountered, such as <ɛ̃> used in in Berom (Plateau, Nigeria), many of which are also not yet encoded in pre-composed forms
- While pre-composed forms are probably safe to use in IDNs, the use of combining marks is considered a problem for the safey and stability of the Domain Name System

Consonant graphemes which cannot be represented in IDNs

- MSR-2s so-called ‘Letter Principle’ poses significant problems for languages making use of clicks
  - Symbols such as <ʘ> are considered IPA Extensions for specialist use, or punctuation marks such as <!>
  - In languages, which use these symbols to represent clicks these are simple letters, which is why the ‘Letter Principle’ is based on a Eurocentric perspective
- Since MSR-2 and IDNA 2008 are based on Unicode 6.3, any Unicode code points added in subsequent revisions are excluded from use in IDNs
  - Examples include LATIN SMALL LETTER BETA' (U+A7B5) encoded in Unicode 8.0, which is used in the writing of some languages of Gabon (cf. ISO/IEC JTC1/SC2/WG2 N4297 L2/12-270)
  - Both standards would have to be updated to newer revisions to include such code points

Variant relationships in IDNs

- In Label Generation Rules, variant relationships describe variation in between the use of code points
  - Given the significant number of competing orthographies, variation is frequently encountered in RLS orthographies, such as <ġ> and <ɣ> used in the Berber Latin Alphabet to represent /ɣ/
  - Such variation should be encoded as either blocked or allocatable variants at the second level
  - For the root zone, IP does not expect the use of variants, and by tendency the number of variants should be kept low based on the Conservatism Principles of the Procedure, which means variants should be rather blocked than allocatable

Contexuality in LGRs: Whole Label Evaluation Rules

- To further control such variation, the mechanism of Whole Label Evaluation rules (WLE rules) can be used in LGRs, which „determine the validity of a label based on whether its code points appear in permissible contexts„
  - Such context maybe be the immediate preceeding or following code point, or the whole label
  - For example, in Somali orthography <‘> is used to represent <?> word-internally and word-finally, but not word-initially, e.g. <afka> [?afka] ‚mouth, tongue‘
  - In the context of an LGR for Somali at the second level, an WLE rule could block the use of <‘> in the beginning of a label
  - At the root zone leve, <‘> would obviously not fulfil the ‚Letter Principle‘ of MSR, which is why it can not be included in an LGR

Thank you for your kind attention

Questions? Questions!

References

- Bendor-Samuel, John (1996). "African Languages". In *The World's Writing Systems*, Daniels, Peter T. & Bright, William (eds), pp. 689-691. New York, Oxford: Oxford University Press.
- Clements, G. N. & Annie Rialland. 2008. Africa as a phonological area. In Bernd Heine & Derek Nurse (eds.), A *Linguistic Geography of Africa* (Cambridge approaches to language contact), 36–85. Cambridge: Cambridge University Press.
- Gregersen, Edgar A. 1977. *Language in Africa: An Introductory Survey*. New York, Paris, London: Gordon and Breach.
Integration Panel. 2015. Maximal Starting Repertoire: MSR-2 Overview and Rationale.
- Pasch, Helma. 2008. Competing scripts: The introduction of the Roman alphabet in Africa. *International Journal of the Sociology of Language* 2008(191). 65–109.
- Roberts, David. 2013. A tone orthography typology. In Susanne R. Borgwaldt & Terry Joyce (eds.), *Typology of Writing Systems* (Benjamins Current Topics 51), 85–112. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Williamson, Kay (2004). „Typical Vowel Systems and Processes in West African Niger-congo Languages". In *Journal of West African Languages* XXX.2: 127-142.