Digitising Swahili in Arabic Script With *Andika!*

Kevin Donnelly

Abstract. In order for traditional culture, as reflected in manuscripts, to make the transition to the digital age, there is a need to use modern technology to make them available. This means more than simply making scans of the manuscripts—it means storing the manuscripts in a digital format which will allow them to be searched, to have concordances and frequency lists compiled, and so on. Where the script used for the traditional material is the same as the current script, this may present few difficulties, but in cases where the traditional material uses a script that is no longer used for the language, this may present difficulties. This paper presents free (GPL3) tools to address these issues for the Swahili language of East Africa (though the general principles are applicable elsewhere), so that heritage material written in a displaced (Arabic) script (S1) can be easily converted to digital form and automatically transliterated to the contemporary (Roman) script (S2).

1. Introduction

This paper addresses ways in which cultural material in a displaced script can be transitioned to the modern digital age. The paper is in three parts.

- The reasons why digitising the *actual text* (as opposed to providing only scans or transcriptions) is essential.
- Tools to do this for Swahili in Arabic script.
- The multitude of ways in which manuscript poetry digitised in this way can be presented.

Kevin Donnelly 0000-0002-0871-6180
Independent reseracher, 7 Ty'n Cae, Llanfairpwll, Ynys Môn, Wales, LL61 6UX, UK
kevin@dotmon.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2020. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 5.
Fluxus Editions, Brest, 2021, pp. 963–984. https://doi.org/10.36824/2020-graf-donn
ISBN: 978-2-9570549-7-8, e-ISBN: 978-2-9570549-9-2

2. Why Do We Need "Full" Digitisation?

2.1. Script Displacement

The loss of cultural capital due to language displacement is now well-recognised,[1] but a similar loss is caused by script displacement.

In many parts of the world, the scripts formerly used to write particular languages have been superseded by other scripts. This is especially the case for African languages such as Swahili, where over the last century the Roman script has displaced the Arabic script formerly used by literate individuals on the East African coast. Adapting the well-established usage of L1 and L2 to denote "first language" and "second language," we might refer to Swahili in Arabic script as S1, and to Swahili in Roman script as S2.[2]

When historical scripts (S1) are displaced by newer scripts (S2), either as a result of past colonial policies or more recent national policies enforcing orthographic change, a phenomenon of progressive "S1 deliteracy" may occur. This can be defined as a situation where modern-day speakers (especially younger ones) are increasingly unable to read documents that may encode significant amounts of cultural heritage. A wealth of traditional linguistic and cultural material (e.g., poetry, histories, religious tracts) may therefore become increasingly inaccessible to speakers of that language.

Script displacement receives less attention than language displacement, perhaps because it is assumed that S1 cultural material can be conserved via digital scanning of the manuscript, or by creating a digital transliteration (more or less phonetic as the case may be) into S2. However, there are issues with both of these.

2.2. Digital Scans

Digital scans are just so many pictures—they cannot be searched unless they are transcribed. You can change the resolution of the scan on-screen, you can move the page around, you can leaf through the document, but that's about it. They are a great resource for librarians and archivists, in that they allow easier access to manuscripts considered as objects, but they have limited value to scholars of history, language or literature who may be more interested in the content, because they lack the scope for unpacking that content rather than simply looking

[1] eldp.net, endangeredlanguages.com

[2] This usage could of course be extended if the language involved has ever been written in more than two scripts.

at it. Moreover, their large size makes them difficult to transfer, especially where internet access is limited, and frustratingly slow to navigate through, particularly on older computers.

These problems can to some extent be resolved by converting S1 scans to pdf and enriching them with additional text layers, as Thilo Schadeberg and Ridder Samsom have done for Sacleux (1939). However, selection of text on such pdfs can be haphazard. Moreover, if you add a text layer, which transcription does it use? Standard (modern), or that used in the manuscript? Or both (one text layer for each)? It is also difficult to do any sort of computer-based analysis (eg list all words occurring at the end of a line of poetry), unless you work solely on the text layer. Arguments for creating a text layer are in effect arguments for a stand-alone digitisation.

Another option with digital scans is to create an interface to them that allows annotations to be made on the image, but this raises questions about how should such annotations should be stored, whether they should be an adjunct to the scan or somehow integrated with it, and how they might be searched and compared.

Where S1 has been maintained, another option is to provide an S1 digital version of the text alongside the S1 scan.[3] Sometimes the scan is omitted in favour of a close (diplomatic) S1 transcription of the manuscript, with the interface allowing round-tripping between the transcription and the manuscript.[4] It is at this point that the text leaps, as it were, off the page and into the computer, out of the past and into the present or future—we have the potential to handle the text in the way we handle a modern computer-generated document, but it is still grounded in the original manuscript.

2.3. Transcription Only

Close S1 transcription for S1 originals requires the conventions used in the transfer from page to screen to be defined in detail. But where we have an S1 original and an S2 transliteration, this is even more important. This goes beyond the transliterations of individual letters (e.g., to transliterate the Arabic letter *khah* خ and *shin* ش, German scholars prefer ḫ and š respectively, while English scholars prefer *kh* and *sh*). More substantively, it raises questions such as:

- How much silent emendation of the text has been done?
- Have sections of the text been omitted, and why?
- Have ambiguous readings been flagged, or simply ignored?

[3] ctext.org, beowulf.uky.edu

[4] rhyddiaithganoloesol.caerdydd.ac.uk, chaucermss.org

As a thought-experiment, consider whether any linguist, literary scholar, or historian would seriously suggest studying Chinese, Greek, Arabic, Egyptian hieroglyphic texts solely via transliteration. S2 transliteration involves decisions that make the transcriber perforce an editor whose decisions the reader must take on trust. In the case of Swahili, we have in the past often ended up with what looks like an overly "tidy" text, with all lines exactly fitting the metre, all rhymes perfect, and so on. A transliteration-only approach not only wrenches the contents from the context in which they were written, it devalues S1 further, and it balkanises the material (it is scattered over various publications, may use a variety of transliterations, may reflect more or less standardisation, and so on).

It might be argued that combining both of the above methods, by presenting an S2 transliteration alongside an S1 scan, is a viable solution to the shortcomings identified. But this solution is only a partial remedy, because it decouples the medium from the message, losing part of what makes the material a cultural resource. The S1 scan now stands apart from the S2 text, and needs to be periodically reintegrated with it for reading purposes.

In fact, however valuable, all these options (S1 scan alone, S2 transliteration alone, S1 scan + S2 transliteration) tend to suggest that S1 belongs to the past, and has little to contribute to the modern culture. Moreover, a judgement is being made on the "value" of the language, such that peripheral languages (minority languages either in terms of the number of speakers or the political "heft" of those speakers) get downgraded. The implication is that some languages do not "deserve" the resources available to others. As noted above, how many scholars would consider studying Chinese or Arabic solely in Roman transliteration?

2.4. Full Digitisation

In the past, scans were expensive and impractical. Transcription, with all its shortcomings and value judgements, was therefore seen as the only viable option, even if it was "lossy" when compared with the original manuscript. This is no longer the case: most mobile phones can take high-quality photos, and the ongoing expansion of the Unicode encoding standard[5] makes it possible for virtually any script to be represented by modern computers. There are therefore few reasons nowadays for not producing "full" digitisations (where the text can be fully processed by a computer to allow searches, the creation of wordlists and so on), backing them up where possible with photographs of the manuscript.

[5] home.unicode.org

This cultural material can then transition fully to the modern, digital world, instead of being viewed as an "object" in a museum collection.

The next section of the paper looks at tools which enable this for S1 Swahili (Swahili in Arabic script). The tools are called *Andika!*, the Swahili word for "write!," which often occurs at the beginning of a poem as a command to the scribe to take up his pen and write down the words of the poem.[6] The general principles behind the toolset can be applied to any language where script displacement is an issue.

3. A Toolset to Digitise S1 Swahili

3.1. Swahili

Swahili is possibly the most widely-spoken Bantu language, in terms of both geographical area and number of speakers. It is widely used as L2 by some 90m people in Kenya, Tanzania, Uganda and the DRC, but it is spoken as L1 by perhaps only 2m people (Hinnebusch, 2003) on the East African coast, from Brava in Somalia down to the Comoro Islands off the coast of Mozambique.

The location of the Swahili meant that they became part of the Indian Ocean trading networks from an early period, and in turn this led to their becoming Islamised. The spread of literacy based on the Arabic script led to the writing of their own language in that script, and Swahili has the longest written heritage of any sub-Saharan African language—poetry survives from the late 1600s onwards (Knappert, 1967; 1972; 1982). The greatest flowering of "classical" Swahili literature was in the 1800s, when poets played a role in many of the "city-states" along the coast (Lamu, Pate, Mombasa, Zanzibar, etc). In the late 1800s European missionaries produced Christian material in Arabic script, but under the British colonial administration the language was "standardised" in a Roman orthography from the 1930s on, and since then the use of Arabic script has declined drastically. That does not mean, however, that S1 Swahili has disappeared—it is still used extensively in religious contexts (e.g., mosque schools), and in particular areas. For instance, Ottenheimer (2012, p. 2) notes that "Arabic script is widely used for Shinzwani [a Swahili dialect in the Comoro Islands], with a literacy rate over 90%".

A wide variety of Swahili poetry in different metres has been published, from religious meditations to ballads to love-songs, but there is also a body of prose work that has been less frequently published. Much

[6] The tools are available under a free (GPL3) license at `kevindonnelly.org.uk/swahili`. The site also includes a manual, and a converter to round-trip between S1 and S2 Swahili—see 3.5 below.

of this S1 material exists in manuscripts, either originals or copies of originals, in Western libraries, and this is probably only a fraction of the extant total—manuscripts are handed down through the generations as family heirlooms.

At present, the only viable way of preserving these S1 (Arabic script) manuscripts is to scan them, or to transcribe them into S2 (Roman script), because the tools available to handle S1 Swahili are limited. Although a word-processor can be set up to use Arabic script, most Arabic fonts do not contain all the glyphs (e.g., *p* /p/, *ng'* /ŋ/) necessary to write Swahili.[7] An additional factor is that the standard Arabic keyboard has a different layout from the standard English (US or UK) keyboard, so using an English keyboard to type Arabic, or vice versa, means trying to mentally translate between the two layouts.

Modern computing platforms give us a viable way to address these issues relatively easily, so that we can type S1 Swahili directly into a computer.

3.2. Characters for Swahili Sounds

The Unicode Consortium, formed in 1991, has the goal of "support[ing] the writing systems used by all the world's languages [by] provid[ing] a unique code for every character, in every language, in every program, on every platform."[8] As of March 2020, the Unicode Standard encompasses 154 scripts and over 143,000 characters,[9] meaning that a great many characters from Arabic-based scripts are covered.[10] However, even if a character has been recognised in Unicode, it does not follow that computer fonts will contain that character, and this is the case for most Arabic fonts, which do not contain all the characters necessary to write Swahili.[11] The most commonly missing sounds, with the *Andika!* character, for them are set out in Table 1.[12]

[7] In earlier times, scribes dealt with this deficiency either by using a character that represented a similar sound, or borrowing a character from another Arabic-script language that had a similar sound. So /p/ might be represented by Arabic ب /b/ or Persian پ /p/.

[8] `home.unicode.org/basic-info/overview`

[9] `unicode.org/versions/\index{Unicode}Unicode13.0.0`

[10] `unicode.org/charts/PDF/U0600.pdf`

[11] ISESCO (Islamic Educational, Scientific and Cultural Organization) has proposed a standard Arabic script that would cater for all African languages (Chtatou, 2010), but this tends to ignore local writing traditions (Warren-Rothlin, 2014)—for example, the proposed vowel for *e* seems to be used only in Fulfulde.

[12] The last three are for representation of northern dialects.

TABLE 1. Swahili characters missing from most Arabic fonts

p	ch	g	ng’	v	t^r	d^r	zh
پ	خ	ڠ	نڠ	ڤ	ٹ	ڈ	ژ

In such a case, there are then two options. One is to add that character to the desired font using a font editor.[13] But the simpler option is to use a comprehensive Arabic font that contains that character. *Andika!* uses SIL’s Scheherazade font.[14] One important point is that since S1 Swahili is usually vocalised, it is best to avoid fonts which use Arabic ligations extensively, since these can cause problems with placement of the vowel signs. Even a font like Scheherazade, though, is still missing characters used by some scribes (e.g., *noon with teh above*, as used in Chimwiini in the most northerly part of the Swahili littoral). In that case, the most workable option in the short term is to add the character by hand using a font editor, and seek in the longer term to have the codepoint added to the Unicode Standard, and then to the font.

3.3. Accessing the Characters

Having chosen a font which contains all the characters needed to represent Swahili, the next requirement is a way to access those characters via the computer keyboard. Since the standard Arabic keyboard has a different layout from the standard English (US or UK) keyboard, switching between them means memorising different keys for the same sounds for each script. For example, *t* is on the top row under the left hand on an English keyboard, but *teh* ت is in the middle row under the right hand on an Arabic keyboard.

To avoid this, *Andika!* uses a key layout for the Arabic characters (Figure 1) that maps to the layout of the English keyboard, meaning that typists can leverage what they already know from typing S2 standard Swahili. The characters are grouped as logically as possible, using either sound or character likeness. For instance, *sukun* is on the full stop key, and short vowels, long vowels, and vowel carriers are all on the same key. Related Arabic characters are mostly on the same keys as the English characters. For instance (Figure 2), *dal* د is on the *D* key, *dhal* ذ is accessed using *Shift+D*, and *dad* ض using *AltGr+D*. A character repre-

[13] designwithfontforge.com/en-US/Adding_Glyphs_to_an_\index{Arabic}Arabic_Font.html If there is not already a Unicode codepoint for that character, a codepoint in a Private Use Area can be used: en.wikipedia.org/wiki/Private_Use_Areas.

[14] software.sil.org/scheherazade. Other possibilities are Khaled Hosny’s Amiri font and the PakType fonts.

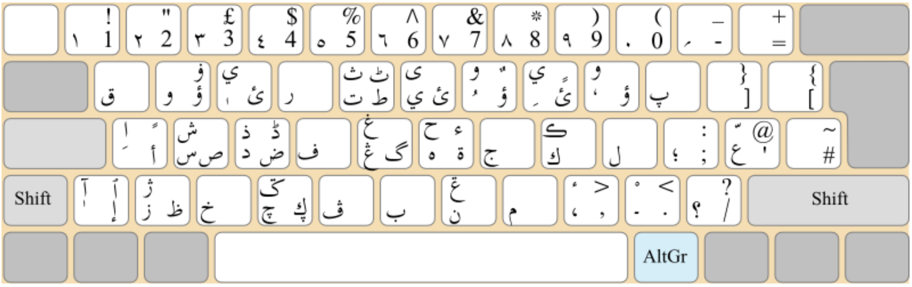
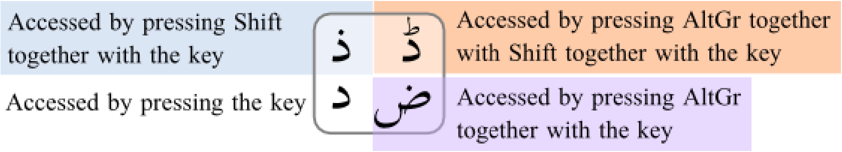


FIGURE 1. The Swahili keyboard layout in *Andika!*

senting the alveolar *d* as used in Mombasa, *dal with tah above* ڈ, borrowed from Urdu, can be accessed using *AltGr+Shift+D*.


Accessed by pressing Shift together with the key

Accessed by pressing AltGr together with Shift together with the key

Accessed by pressing the key

Accessed by pressing AltGr together with the key

FIGURE 2. Character cluster on the *D* key

The result is that S1 (Arabic) Swahili can be typed as quickly and easily as S2 (Roman) Swahili, and on the same keyboard. The same approach could be used for any language to map S1 characters to the relevant S2 keyboard.

3.4. Using S1 Swahili

Now that we have a means of representing Swahili in Arabic script, two forms of use are possible: using it to write contemporary Swahili, and using it to replicate historical Swahili manuscripts in a digital format. The remainder of this section discusses the first use-case, and the next section discusses the second.

Before that, it may be worth emphasising that Arabic script is just as capable as Roman script at representing any language. There are no purely linguistic or graphemic reasons for favouring Roman script over Arabic script for the representation of Swahili: the fact that Swahili is overwhelmingly written in Roman script is due purely to political and historical developments, and not to any shortcoming in the Arabic script. As the British Library's Endangered Archives Programme says, "Ajami, the modified Arabic scripts used in writing African languages, have been deeply embedded in the history and culture of many Islamized societies of Africa,"[15] and Mumin (2014, p. 44) notes that the use of Arabic script has been attested for at least 80 African languages. As with Roman, Cyrillic and other scripts, Arabic script has added additional characters or diacritics when necessary to cater for languages as different as Persian, Turkish, Kurdish, Pashto, Urdu, Malay, Hausa, etc., as Warren-Rothlin (2014, 269ff) points out. In light of this, comments about "the incongruence between Swahili and Arabic and the resulting incompatibility of the Arabic script to write Swahili" (Vierke, 2014, p. 326) are misleading, and indeed seem to be a manifestation of the point made in the introduction about S1 being seen as belonging to the past.

The script itself is not the problem. Rather, its usage has been hampered by a lack of standardisation, where writing conventions tended to be ad hoc responses to recording speech (for a similar issue in West Africa, see Warren-Rothlin (2014, 269ff)). S2 is more likely to have such conventions than S1, given that in many cases S2 may have been expressly designed to handle the language, perhaps via a committee, whereas S1 is more likely to have been progressively developed and adapted over a period of time by individuals unable to do anything but promote good practice as they see it.

This issue of standardised spelling is relevant when discussing contemporary use of S1 Swahili. Although S2 Swahili is now the standard used by millions of speakers, and that will not change, the ability to use S1 may be useful in domains (e.g., mosque schools) where that script is still used, or in places (e.g., the Comoros) where S1 literacy is still high. The key point is to allow the *option* of using S1 for cultural heritage purposes. A key practical issue for contemporary use, however, is avoiding the additional work involved in typing the text twice, once in each script. Preferably, it should be possible to get either S1 or S2 "for free" from the other, meaning that the same text can be created in either script, and converted to the other as required This also provides an easy way of increasing the amount of modern S1 text available, for instance, by converting S2 webpages or other documents to S1. Crucially, however, conversion is only possible if both scripts use standardised spelling.

Since there is currently no standard for S1 Swahili spelling, *Andika!* uses the proposed system set out in Omar and Frankl (1997).[16] Combined with the keyboard layout described above, this means that text

[15] `eap.bl.uk/project/EAP1042`

[16] Some slight modifications have been made. For instance, the authors suggest rules for omitting short vowels, but it is actually quicker to type them, and this also simplifies conversion.

can be typed into a word-processor at around the same speed in either script using almost exactly the same keys (Figure 3)—the only difference is the need to type a capital letter to get a long vowel in the penultimate syllable in S1.

S2 standard:	ninakwenda nyumbani sasa
S1 typing:	ninakweEnda nyumbaAni saAsa
S1:	نِنَكوِيندَ نيُمبَانِ سَاسَ
English:	I am going home now

FIGURE 3. Typing S1 Swahili

3.5. Converting Between S1 and S2

In the S1 → S2 direction (Figure 4), S1 is converted first to a Romanised abstraction, and then converted to a standard S2 transliteration.

The reason for the intermediate abstraction is to offer scope for multiple transliterations. For instance, when dealing with older manuscripts (see 4.2 below) we may wish to have a close transliteration of the Arabic script as well as a standard transliteration. Alternatively, it may be appropriate to replace a standard transliteration with one that reflects dialectal features. For instance, if a scribe has written ذِيچَ, dhīcha, the equivalent in the northern Bajuni dialect to standard Swahili *vita*, 'war', a transliteration such as *z̲it*^{j}*a* might be preferred, in order to come as close as possible to standard S1 while giving an indication of the dialectal pronunciation.[17] Another option would be to add a transliteration for Arabic, to handle bilingual Arabic/Swahili text (e.g., *Qasida Hamziyya* poems—see 4.2 below). Currently, Arabic text is transliterated using the close transliteration conventions for Swahili, and some features of the Arabic language are not optimally handled in this.

In the S2 → S1 direction, no intermediate abstraction is currently used, because there is only output at present—the proposed standard spelling in Omar and Frankl (1997). Nevertheless, the same approach could be used, so that different S1 spellings are supported.

Virtually no editing is required in either direction, with the exception that in S1 → S2 capital letters need to be added where appropriate, since Arabic has no concept of capital letters. The website[18] gives an example of a browser-based frontend to the converter code, where text can be copied and pasted into a box, and converted to the other script.

[17] *zi-* is a northern variant of the standard class 8 marker *vi-*.

[18] kevindonnelly.org.uk/swahili

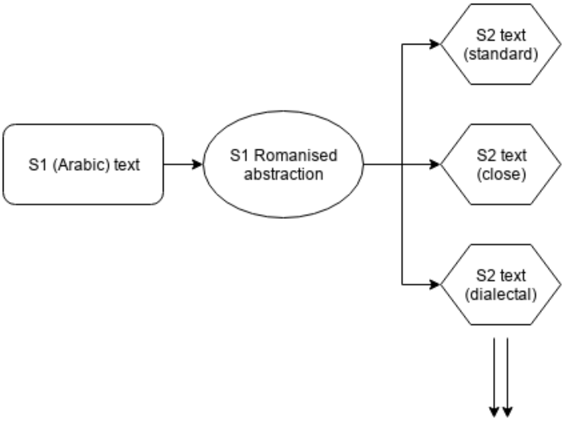


FIGURE 4. Conversion from S1 to S2

4. Digitising Heritage Manuscripts

With tools in place to type S1 Swahili into a computer, and to convert it to a variety of S2 Swahili transliterations, we now have a means of digitising S1 manuscripts, particularly those containing traditional poetry. This section demonstrates replicating the content of the manuscript in digital form, and gives examples of various types of enriched output.

The digitisation of heritage manuscripts has additional requirements compared to the digitisation of contemporary texts. For instance, we may want to reflect the layout on the physical page, or add alternate readings, or draw attention to scribal errors, or add contextual or etymological notes on individual words, or add a translation. Likewise, we are almost certain to need a list of the words in the manuscripts, so that concordances, indexes or other editorial matter can be prepared. The solution proposed by *Andika!* is to insert each word of the manuscript text into a database, so that additional material like this can be added at word-level. Subsets of the material can then be retrieved from the database as required, in whatever format is appropriate.

4.1. The Digitisation Process

A key difference between traditional transliteration and the *Andika!* approach is that the process begins with typing out the manuscript itself instead of typing out a transliteration of the manuscript. The latter step is not required, because the transliteration (indeed, several transliterations—see 3.5) can be generated automatically from the retyped manuscripts. The process is summarised in Figure 5.

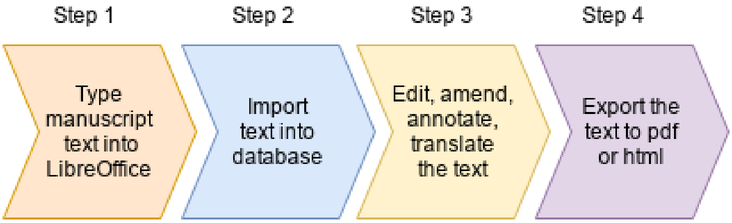
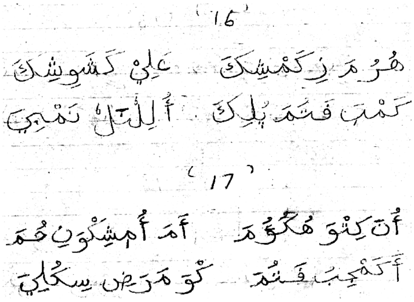

Step 1
Type manuscript text into LibreOffice
Step 2
Import text into database
Step 3
Edit, amend, annotate, translate the text
Step 4
Export the text to pdf or html

FIGURE 5. Digitisation process for S1 manuscripts

In Step 1, the manuscript is typed out as if it were a piece of contemporary text, but instead of trying to follow a spelling standard, we type only what the scribe wrote in the manuscript. To simplify import into the database, the typed text should follow a specific format—each rhymed stretch of the poem should appear on a line by itself, blank lines should be inserted after stanzas, etc. Figure 6a shows stanzas 16 and 17 from an original manuscript version of the Swahili *Ballad of Jaʾfar*. Figure 6b shows the same stanzas typed out and ready for import into the database.



(a)

#17	#16
هُرُمَ زِكَسْشِكَ	أُنَ كِتْوَ هُكُؤُمَ
عَلِيْ كَشَوِشِكَ	أَمَ اُمِشِكْوَ نِ حُمَ
كَسْبَ فَتُمَ پُلِكَ	أَكَسْجِبُ فَتُمَ
أُلِلْنَلَ نَسْبِيِ	كْوَ مَرَضِ سِكُلِيَ

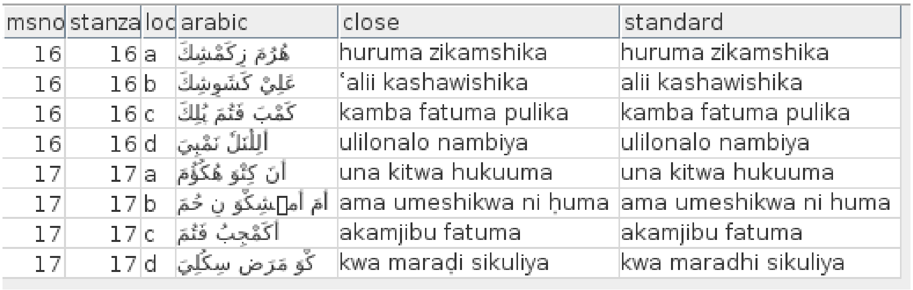
(b)

FIGURE 6. (a) Manuscript stanzas from the Ballad of Jaʾfar. (b) The stanzas as typed

In Step 2, *Andika!* parses the typed copy of the manuscript, and imports the lines of the poem into a database table. In the process both a standard and a close S2 transcription for the S1 text is automatically generated (Figure 7). Then each line is split into words, which are imported into another database table (Figure 8a).

In Step 3, each word in the database can be inspected to correct the automatic transcription if necessary. Individual words can then be annotated with notes, alternate readings, corrected transliterations, etc., and a translation can be added, as in Figure 8a, keyed to the first word of the line. This allows the development of a full critical apparatus for the text.

In Step 4, the textual material in the database can be exported in PDF format, with various options for text layout, text colouring, line numbering, translation position, etc. Figure 8b shows the two stanzas output as single lines, with S1 Swahili in green, stanza numbering, an English translation, and a generated S2 Swahili transcription. A generated close transcription (in green) is also included, output right-to-left by word so that it matches directly with the S1 script above it.



msno	stanza	loc	arabic	close	standard
16	16	a	هُرُمَ زِكَمْشِكَ	huruma zikamshika	huruma zikamshika
16	16	b	عَلِيْ كَشَوِشِكَ	ʿalii kashawishika	alii kashawishika
16	16	c	كَمْبَ فَتُمَ پُلِكَ	kamba fatuma pulika	kamba fatuma pulika
16	16	d	اُلِلُنَلُ نَمْبِيَ	ulilonalo nambiya	ulilonalo nambiya
17	17	a	اُنَ كِتْوَ هُكُوُمَ	una kitwa hukuuma	una kitwa hukuuma
17	17	b	اَمَ اُمِ□شِكْوَ نِ حُمَ	ama umeshikwa ni ḥuma	ama umeshikwa ni huma
17	17	c	اَكَمْجِبُ فَتُمَ	akamjibu fatuma	akamjibu fatuma
17	17	d	كْوَ مَرَضِ سِكُلِيَ	kwa maraḍi sikuliya	kwa maradhi sikuliya

FIGURE 7. The stanzas from Figure 6 imported as lines

4.2. Other Examples

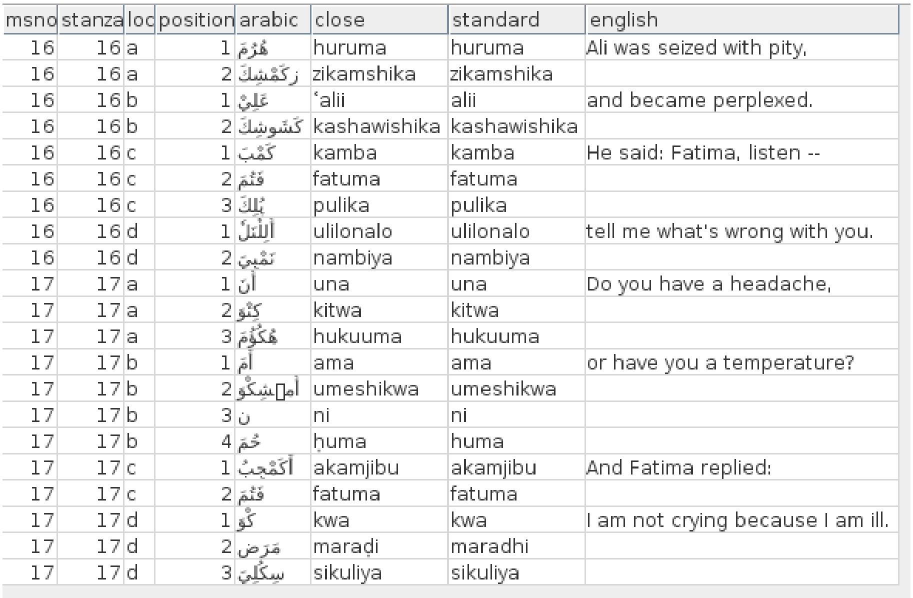
Mwana Kupona is one of the few female Swahili poets whose work has come down to us. In 1858 she wrote a poem containing advice for her daughter, and stanza 6 reads (in standard S2 Swahili):

> mwana adamu si kitu, na ulimwengu si wetu,
> walau hakuna mtu ambao atasaliya
>
> *mankind is as nothing, and the world does not belong to us,*
> *and there is no person who will live forever*

Below is an S1 transliteration and an automatically generated close transcription of this stanza from the first of two manuscripts (Figure 9).

مَانَ اَدَامُ سِكِتُ ∗ نَوُلِمِغُ سِوِتُ ∗ وَلَوُ هَكُوْنَ مْتُ ∗ اَبَوُ اَتَسَلِيَ

māna aḏāmu sikiṯu ∗ nawulimiğu siwiṯu ∗ walawu hakūna mṯu ∗ abawu aṯasaliya



msno	stanza	loc	position	arabic	close	standard	english
16	16	a	1	هُرُمَ	huruma	huruma	Ali was seized with pity,
16	16	a	2	زِكَمْشِكَ	zikamshika	zikamshika	
16	16	b	1	عَلِيْ	ʿalii	alii	and became perplexed.
16	16	b	2	كَشَوِشِكَ	kashawishika	kashawishika	
16	16	c	1	كَمْبَ	kamba	kamba	He said: Fatima, listen --
16	16	c	2	فَتُمَ	fatuma	fatuma	
16	16	c	3	پُلِكَ	pulika	pulika	
16	16	d	1	أُلِلْنَلْ	ulilonalo	ulilonalo	tell me what's wrong with you.
16	16	d	2	نَمْبِيَ	nambiya	nambiya	
17	17	a	1	أُنَ	una	una	Do you have a headache,
17	17	a	2	كِتْوَ	kitwa	kitwa	
17	17	a	3	هُكُؤُمَ	hukuuma	hukuuma	
17	17	b	1	أَمَ	ama	ama	or have you a temperature?
17	17	b	2	أُمِ□شِكْوَ	umeshikwa	umeshikwa	
17	17	b	3	نِ	ni	ni	
17	17	b	4	حُمَ	ḥuma	huma	
17	17	c	1	أَكَمْجِبُ	akamjibu	akamjibu	And Fatima replied:
17	17	c	2	فَتُمَ	fatuma	fatuma	
17	17	d	1	كْوَ	kwa	kwa	I am not crying because I am ill.
17	17	d	2	مَرَضِ	maraḍi	maradhi	
17	17	d	3	سِكُلِيَ	sikuliya	sikuliya	

(a)

(١٧) هُرُمَ زِكَمْشِكَ * عَلِيْ كَشَوِشِكَ * كَمْبَ فَتُمَ پُلِكَ * أُلِلْنَلْ نَمْبِيَ

nambiya ulilonalo ∗ pulika fatuma kamba ∗ kashawishika ʿalii ∗ zikamshika huruma

(**16**) huruma zikamshika ∗ Aliyi kashawishika ∗ kamba Fatuma pulika ∗ ulilo nalo nambiya

Ali was seized with pity, and became perplexed. He said: Fatima, listen—tell me what's wrong with you.

(١٨) أُنَ كِتْوَ هُكُؤُمَ * أَمَ أُمِشِكْوَ نِ حُمَ * أَكَمْجِبُ فَتُمَ * كْوَ مَرَضِ سِكُلِيَ

sikuliya maraḍi kwa ∗ fatuma akamjibu ∗ ḥuma ni umeshikwa ama ∗ hukuuma kitwa una

(**17**) una kitwa hukuuma ∗ ama umeshikwa na huma ∗ akamjibu Fatuma ∗ kwa maradhi sikuliya

Do you have a headache, or have you a temperature? And Fatima replied: I am not crying because I am ill.

(b)

FIGURE 8. (a) The stanzas from Figure 6 imported as words. (b) The stanzas from Figure 6 exported as a fully digital text

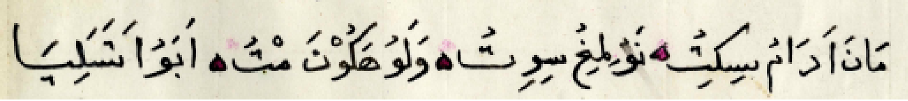


FIGURE 9. Utenzi wa Mwana Kupona, stanza 6, first manuscript.

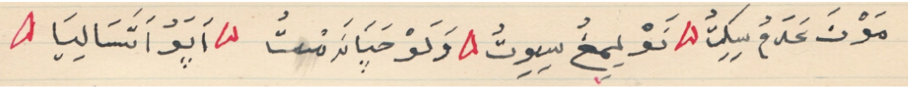


FIGURE 10. Utenzi wa Mwana Kupona, stanza 6, second manuscript.

Figure 10 shows the same stanza from a second manuscript. The S1 transliteration and transcription below show notable spelling differences compared to the first manuscript (e.g., the use of *ʿayn* ع and *ha* ح), and a typo in the first word, where *fatha* and *sukun* have been reversed.

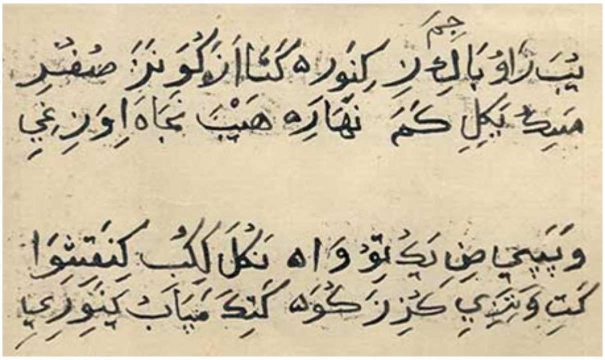
مَوْنَ عَدَمُ سِكِتُ ∗ نَوْلِمِغُ سِوِتُ ∗ وَلَوْ حَپَانَ مْتُ ∗ اَپَوُ اَتَسَالِيَا

mawna ʿaḏamu sikiṯu ∗ nawlimiḡu siwiṯu ∗ walaw ḥapāna mṯu ∗ apawu aṯasāliyā

Sayidi Abudallah (Hichens, 1939) wrote a lament in 1853 about the declining fortunes of the coastal city-state of Pate. The two stanzas shown in Figure 11, given here in S2 Swahili, describe the opulence of the town in its better days:

> Nyumba zao mbake zikinawiri kwa taa za kowa na za sufuri;
>   masiku yakele kama nahari, haiba na jaha iwazingiye.
>
> Wapambiye swini ya kuteuwa, na kulla kikombe kinakishiwa;
>   kati watiziye kuzi za kowa katika mapambo yanawiriye.
>
> Their homes were brightly lit with lamps of mother-of-pearl and copper;
>   the nights stayed bright as day, beauty and privilege surrounded them.
>
> They decorated their fine porcelain, and every goblet was engraved;
>   in the centre they placed mother-of-pearl carafes, to glitter amongst the fine things.

The transcription and close transliteration below show how the Arabic script is only a partial representation of the sounds of Swahili. Three vowel glyphs are used to represent Swahili's five vowels (e.g., *yakili* for *yakele*, 'stayed', *kuwa* for *kowa*, 'shell'), and prenasalised consonants are not distinguished (*yuba* for *nyumba*, 'house', *mapabu* for *mapambo*, 'decorative objects').



جم
يب زاو باك زكنوره كتاا زكو نز صفر
مسك يكل كم نهاره هيب نجاه اوزغي

وپپي صن يكتووا ه نكل ككب كنقشوا
كت وتزي كز زكوه كتك مپاب ينوريا

FIGURE 11. Two stanzas from al-Inkishafi.

يُبَ زَاوُ بَاكِ زِكِنَوِرِ ⁎ كَتَاَا زَكُوَ نَزَضُفُرِ
yuba zāwu bāki zikinawiri ⁎ katāa zakuwa nazaḍufuri
مَسِكُ يَكِلِ كَمَ نَهَارِ ⁎ هَيْبَ نَجَاهَ اِوَزِغِيِ
masiku yakili kama nahāri ⁎ hayba najāha iwazighiyi

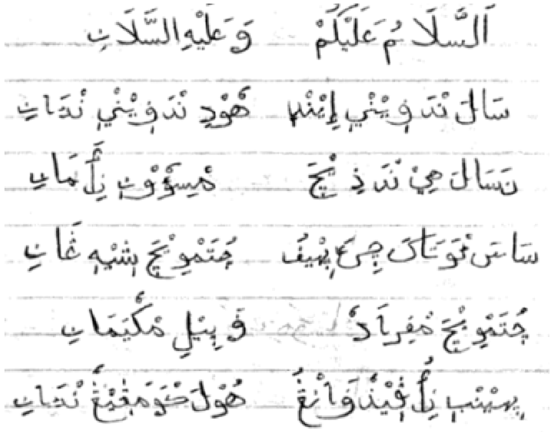
وَپَپِيِ صِنِ يَكُتِوُوَا ⁎ نَكُلَ كِكُبِ كِنَقِشِوَا
wapapiyi ṣini yakutiwuwā ⁎ nakula kikubi kinaqishiwā
كَتِ وَتِزِيِ كُزِ زَكُوَ ⁎ كَتِكَ مَپَابُ يَنَوِرِيِ
kati watiziyi kuzi zakuwa ⁎ katika mapābu yanawiriyi

Figure 12 shows part of a manuscript by the late Sh. Yahya Ali Omar recording fishing songs in kiBajuni or kiTikuu, a northern Swahili dialect (Donnelly and Omar, 1982).

The automatic S2 transliteration uses a variant of the close trancription, so that it stays as close as possible to standard Swahili orthography, while still representing the distinctive sounds of kiTikuu (see 3.5).

The *Ballad of Mkunumbi* (Harries, 1967) is one of the few books of Swahili poetry to include the S1 text of the manuscript.

This digitisation of the stanza in Figure 13 shows a close transcription only, keyed to the half-line rather than the full line. It also adds stanza numbers in eastern Arabic numerals for the S1 transcription, and in western Arabic numerals for the transcription, where the half-lines are also indicated.



اَلسَّلَامُ عَلَيْكُمْ * وَ عَلَيْهِ السَّلَانِ
assalāmu ʾalaykum * wa ʾalayhi assalāni
سَالَ نْدَ وِيْنْيِ إِيْنْدِ * هٗوْدِ نْدَ وِيْنْيِ نْدَانِ
sāla nḏa wēnye īnḏe * hōḏi nḏa wēnye nḏāni
نَ سَالَ هِيْ نْدَ ذِيْچَ * مْسِوٗوْنِ نِأَمَانِ
na sāla hii nḏa ẕītʲa * msiwōne niamāni
سَاسَ ٹْوَتَاكَچِئَ پِيْفُ * چُتَمْوِيْچَ شِيْهِ غَانِ
sāsa ṭwaṯākatʲia pēfu * tʲuṯamwītʲa shēhe gāni
چُتَمْوِيْچَ مْفِرَادٗ * وَ پِيْلِ مْكُيَمَانِ
tʲuṯamwītʲa mfiraḏo * wa pīli mkoyamāni
پِيْنْبِ نِ أُڤِيْذٗ وَانْغُ * هُوْلَ كْوَ مَغِيْغٗ نْدٰانِ
pēm̱be ni uw̱ēẕo wāngu * hūla kwa magēgo nḓāni

FIGURE 12. Stanza from a Bajuni fishing song.

١ ذٗوْلَ مْبِلِ زِلِوَانَ   شِكُوِ نَاسِمْبَ مْبَوَانَ
1b/a ḏōla mbili ziliwāna   shikuwe nāsimba mbawāna
كَمَتِزٗ كُشِنْدَانَ   مْتَانَ نَلَيْلِيَ
1d/c kamaṯezo kushinḏāna   mṯāna nalayliya

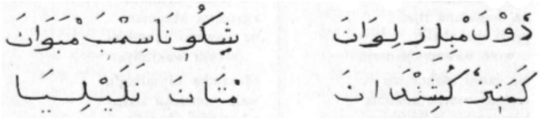


FIGURE 13. A stanza from the Ballad of Mkunumbi

The *Ballad of Rasi ʾlGhuli* was written around 1850 by Mgeni bin Faqihi, and at over 4,500 stanzas is the longest Swahili ballad in existence (van Kessel, 1979). This digitisation of stanza 2,280 (Figure 14) has a close transcription keyed to the half-line, and a standard transcription keyed to the full line.

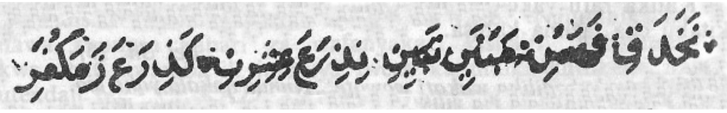


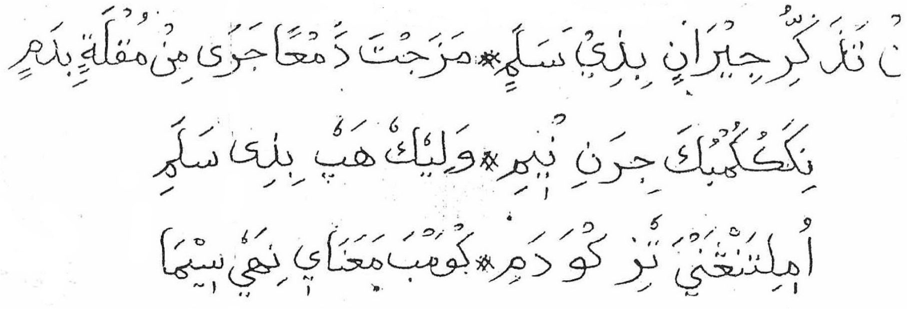
FIGURE 14. Stanza 2,280 from the Ballad of Rasi ʾlGhuli

٢٢٨٠ نَخَدَقِ فَهَمُنِ مَبَنَيِ تَبَيِنِ
mabanayi tabayini nakhadaqi fahamuni 2280b/a
na khandaqi fahamuni ∗ mpanaye tabaini 2280a/b

نِذِرَعَ عِشِرِنِ كَذِرَعَ زَمَكُفَرِ
kadhiraʾa zamakufari nidhiraʾa ʾishirini 2280d/c
ni dhiraa ishirini ∗ kwa dhiraa za makufari 2280c/d

Qasidas are panegyric poems in Arabic eulogising the Prophet (Knappert, 1971; Sperl and Shackl, 1995). The *Qasida ya Burda* was composed in Arabic by Muhammad bin Saʿidi al-Busiri in the 1300s, and rendered into Swahili verse by Muhammad bin Athumani Hajji al-Hilali Mshela, 1840-1930 (wa Mutiso, 1996). The digitisation of the manuscript in Figure 15 has been set to show the original text in Arabic in blue.

The *Qasida Hamziyya* (so called because it rhymes in *hamza* ء) was composed in Arabic by Muhammad bin Saʿidi al-Busiri (1212-1294), and rendered into Swahili verse by Aidarus bin Athumani bin Ali bin Sheikh Abubakar bin Salim in (probably) 1749 (Knappert, 1968; Parkar, 2020). The digitisation of the manuscript in Figure 15 show the original text in Arabic in blue, and no transcription.



١ أَمِنْ نَذَكُّرِ جِيْرَانٍ نِذِيْ سَلَمٍ * مَزَجْتَ دَمْعًاجَرَى مِنْ مُقْلَةٍ بِدَمٍ
نِكَكُكُمْبُكَ جِرَنِ نْيمَ * وَلِيْكْ هَبٌّ نِذِى سَلَمِ
اُمِلِتَنْغَنْيَ تِزِ كْوَ دَمِ * كْوَمْبَ مَعَنَاي نِهَيٌ سِيْمَا

FIGURE 15. First stanza of the Qasida ya Burda

لَمْ يُسَاوُكَ فِيْ عُلَاكَ وَقَدْ حَالَ * سَنَي مِنْكَ دُوْنَهُمْ وَ سَنَاءُ
lam yusāwuka fi ʿulāka waqad ḥāla ∗ sanay minka dūnahum wa sanāʾu

كَوَفَنِ نَوِ رِفْعَانِ بَحَجِزِلِ * نُوْرُ نَرُفْعَةَ كَتِكِنُ كُلُ عَظِيْمَ
kawafani nawi rifʿāni baḥajizili ∗ nūru naruʿfaẗa katikinu kulu ʿaẓīma

They are not equal to you in your elevated status,
the light and sublimity in you is great (in all respect).

FIGURE 16. Second stanza of the Qasida Hamziyya

Figure 16 shows a digitisation of second stanza of the poem. Here again the Arabic original is in blue, and there is an English translation. Both have a close transcription (though the Arabic one, as noted in 3.5, is less than optimal).

Andika! also allows multiple copies of the same poem to be presented in parallel. Below are S1 digitisations of two manuscript versions of the *Ballad of Jaʾfar* (see 4.1 above), each coloured differently, so that they can be compared (in this example, the third line of the stanza differs in each version). Each version has an S2 standard transliteration, keyed to the stanza, and a close transcription (coloured to match the S1 text) keyed to the S2 word. An English translation is attached to the first (Y) manuscript version.

(٢٦) كَمْجِبُ كْوَ لِسَنِ * مْٹُي سِمْبَئِنِ * پِٹ أُمْپِيْ نَنِ * أُنِيَپْ تهَرِضِيَ

tʰariḍiya unipapo ∗ nani umpee peṭe ∗ simbaini mṭuye ∗ lisani kwa kamjibu

Y 25 [23] (**26**) kamjibu kwa lisani ∗ mtuye simbaini *pete umpee nani ∗ unipapo taridhiya
She replied forcefully: I will not disclose that person.
Who have you given the ring to? [Only] when you give [it to me] will I be satisfied.

كَمْجِبُ كْوَا لِسَنِ * مْتُي سِمْبَاءِنِ * پِتِ يَكُ يَكَنْدَانِ * أُنِيَپْ تَرِظِيْيَا

tariẓīyā unipapo ∗ yakʲandāni yaku piti ∗ simbaini mtuyi ∗ lisani kwā kamjibu

R 26 [26] kamjibu kwa lisani ∗ mtuye simbaini ∗ pete yako ya chandani ∗ unipapo taridhiya

4.3. Beyond the Manuscript

Storing manuscript text in a database, as *Andika!* does, opens up some interesting possibilities. As noted earlier, one important side-effect is the easy creation of concordances and indexes. If the wordstores for a number of different manuscripts are combined (in effect, creating a searchable literary corpus), we get a multiplier effect: we are working across manuscripts instead of within them. Such a corpus would, for instance, allow scholars to:

- study character usage and spelling conventions, which may help clarify the genealogy of particular manuscripts;
- trace the occurrence of particular words and examine vocabulary usage in general, which may identify particular schools or authors;
- analyse textual features such as syntactic structure, which could be useful in researching diachronic and synchronic variation;
- consider the usage of fixed expressions (formulae), which may give clues about the process of composition and recital.

As an example, studying the wordstore for the *Ballad of Jaʾfar* referred to several times above allows some significant conclusions to be drawn:

- the verbal consecutive (non-time specific) marker occurs in 30% of all verbforms, reflecting the emphasis on timeless action in the ballad;
- around a quarter of the words in the ballad are derived from Arabic;

- Arabic words are more likely to occur in rhyming positions in stanza-internal lines, suggesting that considerations of rhyme and metre are their main rationale;
- one in five of the verbs used relate to speaking (*say, reply, speak, greet*, etc.);
- almost half of the stanza-internal lines use just three rhymes (*-ni, -ri, -ka*);
- ready-made rhyme-sets seem to be available that will allow the reciter to refer to one of the characters saying something, and bring in a reference to God if appropriate.

5. Conclusions

This paper has argued that "full" digitisation of S1 heritage manuscripts is the only approach that will liberate the cultural riches locked in them, and avoid them being seen as museum objects that belong in the past, defined solely by type of paper, ink composition, and layout. Such manuscripts are more than just scanned images—they are unique snapshots of a nexus of cultural ideas that still speak to the present and future, even if they were produced in the past. To engage with these ideas, we need to pay these manuscripts the courtesy of transitioning them fully to the digital age, so that we can bring to bear all the tools now available to us in unlocking their meaning.

The *Andika!* toolset described here is one possible way in which this concept could be executed. It is still a work-in-progress, but the concept could be adapted to any language where cultural material is available in a displaced script. Funders of humanities research might also consider the benefits of generating local employment in the "knowledge industry" by paying local people to type heritage manuscripts into a computer alongside the work already being done to scan manuscripts.

References

Chtatou, Mohamed (2010). *Using Arabic script in writing African languages, revisiting ISESCO's experience 25 years later: Field successes and shortcomings*. Paper presented at the workshop "The Arabic Script In Africa: Diffusion, Usage, Diversity And Dynamics Of A Writing System," University Of Cologne.

Donnelly, Kevin and Yahya Ali Omar (1982). "Structure and association in Bajuni fishing songs." In: *Genres, Forms, Meanings: Essays in African Oral Literature*. Ed. by Veronika Görög-Karady. Vol. 1. JASO Occasional Papers, pp. 109–122.

Harries, Lyndon (1967). *Utenzi wa Mkunumbi. A Swahili Potlatch—The Poem about Mkunumbi*. Nairobi: East African Literature Bureau.

Hichens, William (1939). *Al-Inkishafi: The Soul's Awakening*. London: Sheldon Press.

Hinnebusch, Thomas J. (2003). "Swahili." In: *International Encyclopedia of Linguistics*. Ed. by William J. Frawley. Oxford: Oxford University Press, pp. 99–106.

Knappert, Jan (1967). *Traditional Swahili Poetry*. Leiden: Brill.

——— (1968). "The Hamziya deciphered." In: *African Language Studies* 9, pp. 52–81.

——— (1971). *Swahili Islamic Poetry*. Leiden: Brill.

——— (1972). *A Choice of Flowers: Swahili Songs of Love and Passion*. Portsmouth, NH: Heinemann.

——— (1982). *Four Centuries of Swahili Verse: A Literary History and Anthology*. Portsmouth, NH: Heinemann.

Mumin, Meikal (2014). "The Arabic script in Africa: Understudied literacy." In: *The Arabic Script in Africa: Studies in the Use of a Writing System.* Ed. by Meikal Mumin and Kees Versteegh. Leiden: Brill, pp. 41–76.

Omar, Yahya Ali and P. J. L. Frankl (1997). "An historical review of the Arabic rendering of Swahili, together with proposals for the development of a Swahili writing system in Arabic script." In: *Journal of the Royal Asiatic Society*. 3rd ser. 7.1, pp. 55–71.

Ottenheimer, Harriet J. (2012). "Ideology and orthography. Dictionary construction and spelling choice in the Comoro Islands." In: *Études océan Indien* 48. DOI: https://doi.org/10.4000/oceanindien.1521.

Parkar, Ahmed (2020). "Manuscripts and Transmission of Knowledge in Swahili Society: A Comparative Analysis of Form and Usage of Qaṣīda al-Hamziyya." PhD thesis. University of Hamburg.

Sacleux, Charles (1939). *Dictionnaire swahili-français*. Vol. 36. Travaux et mémoires de l'Institut d'Ethnologie. Paris: Institut d'Ethnologie.

Sperl, Stefan and Christopher Shackl, eds. (1995). *Qasida Poetry in Islamic Asia and Africa*. Leiden: Brill.

van Kessel, Leo (1979). *Utenzi wa Rasi ˀlGhuli*. Dar-es-Salaam: Tanzania Publishing House.

Vierke, Clarissa (2014). "Akhi patia kalamu: Writing Swahili poetry in Arabic script." In: *The Arabic Script in Africa: Studies in the Use of a Writing System*. Leiden: Brill, pp. 319–339.

wa Mutiso, Kineene (1996). "Archetypal Motifs in Swahili Islamic Poetry: Kasida ya Burudai." PhD thesis. University of Nairobi.

Warren-Rothlin, Andy (2014). "West African scripts and Arabic-script orthographies in socio-political context." In: *The Arabic Script in Africa: Studies in the Use of a Writing System*. Leiden: Brill, pp. 261–289.