Technical Reports on Language Technology
Report No 60, 20020
http://www.njas.helsinki.fi/salama

Printed text into machine-readable form[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Currently, large projects are going on for digitizing old printed materials. It is assumed that when the text is in digital form, the search and retrieval of information from text becomes easy. However, the accuracy and coverage of information retrieval depends on the digital form of text. The usual solution is to scan the text into such a format that it looks similar with the original outlay. It is a kind of photograph, which then can be read as the original paper version. The scanning result is usually saved into pdf-format. It is currently also possible to enhance the result, so that individual words can be identified, and a search engine can be used for finding words in context. This is already a significant improvement to the simple bitmat format, which earlier did not allow any intelligent search. Even this advanced search method has, however, a major problem, because the text cannot be edited. It includes all mistakes made by the OCR system. If the original paper text has high quality, the scanning result will be almost perfect. But this is seldom the case, especially with old texts, which were printed on weak paper, and which have already detoriated due to long age.

In this report, I will suggest a method for converting the printed texts into true digital text format, which can be corrected, edited, and manipulated using any methods and tools developed for digital text. The test material is a Swahili text *Msimulizi*. The text is the oldest Swahili language periodical, printed in Zanzibar since 1888. SOAS (School of Oriental and African Studies) has scanned the text and made it available.

Key Words: *scanning, information retrieval, morphological analysis.*

1 Introduction

SOAS (School of Oriental and African Studies) has digitised old printer materials from Asia and Africa and made them available for the public. The archives include texts as well as images. I am interested especially of text, because it includes invaluable information on many aspects of life in the past. It is already a major step forward to digitize the materials and make them available over the net. Part of the material can also be downloaded and used locally, observing the licensing conditions, of course.

In this report I will study the possibility and feasibility to convert digitised text materials into true text format, so that it can be handled as the normal digital text. I shall use the 14 first printed issues of Msimulizi (years 1888-1890), the first Swahili language periodical, published originally in Zanzibar. Msimulizi, which appeared six times per year, was an information channel between various mission stations of the Universities' Mission to Central Africa, in the area of current Tanzania and Malawi. The misson organisation was established after hearing the horrible reports of massive slave trade in Central Africa. Especially the travel reports of David Livingstone prompted action in Britain for abolishing slave trade by introducing Christian education in the area. *Msimulizi* describes vividly the progress of the work from the grassroot viewpoint. Also the clashes between English, German, Portuguese and Arab forces in the area have been described.

In all, *Msimulizi* contains a local contemporary description of the history in East Africa during the period beginning immediately after the division of Africa in 1888 in Berlin.

2 From scanned version to true text form

SOAS has scanned the pages of *Msimulizi* so that, when they are saved in pdf-format, they look precisely as the original pages, including colour and dirt. This format can be read quite easily, and some elementary searches can also be made to the text. The search succeeds, if the search string matches with the string in text. However, one cannot be sure that the string in scanned text is interpreted internally in the way it looks when reading it. The internal OCR system makes frequently mistakes, and one can seldom guess how the system interprets the string. As a result, only part of hits will be found.

The only solution for solving the problem is to convert the whole text into real text format, so that it can be edited, and the OCR mistakes can be corrected. This can be done with a copy/paste system, where the text is copied from the pdf-version and pasted to a file in a text editor. On the editor, we can see character by character how each character in original text was interpreted.

When looking at the text in the text editor, it soon becomes clear that if the original text was clean, also the result has only a few mistakes to be corrected. Less clean original texts produce a lot of mistakes and correcting them is a slow process. Below is an extract from a clean section of scanned text (1).

(1)
Assuboi na mapema kiisha kutoka kanisani tukatengeneza uwani, na tukalima mahali vizuri pa kuwekea gari la farasi la Sayid, kwani tukasikia kama Sayid Khalifa Mfalmewa Unguja amekubali kuja kututazama. Tulipokwisha tukucheza mpaka athunri, tukaenda Chuoni kununua zawadi pamoja na Wamjini, na Wambweni, nasisi wa Kiungani, kiisha tukacheza hatta saa kumi.

Sections that need editing are marked with red colour. This text is quite easy to edit.

In (2) we have two extracts from less clean sections. Sections to be edited are marked with red.

(2)
*Katika June 18 ikaja habari kusema kama Wamasai wanakuja, wapo pa Ntunuritu, ndio karibu na Misozwe. Lakirii fukangoja wee/ nao hawaji. Lakini tukasikia tena kama wapo twa Kibanga, wanajenga rriji na bomay wakiisha ndio ' walete vita; wakifukuzwa waende bomani pao^ lakihi k'ut'bka bomani hdib hawatakubali hatta kidogo^*

*Ah ! bwana, hawawezi rnoyo hauwapi, kwa kustaajabu sana, naswi tulikufa kwa kucheka, bwana, k'wa ginsi wanavyofikiria tnambo ya mbali ya- siokuwamo hunro. Mar*ra inmoja akasema, akaiilfea, Ah! btoana, hivyo usiku je watu hawa htifanya kaii vile vife wala hawalali nini ? Tukamjibu, ndio huenda mchan& rfa tisiku haisitnami. Akajibu,— ajabu hiyo, hapana teni ■&atu wenyi maarifa kupita Wazungu !*

These sections need so much editing that one could consider typing such sections anew.

Is it feasible to manually correct texts, which partly are fairly clean and partly full of scanning errors? The answer to this question depends on the importance of the text. We can also approach the question by working with the text in phases. It is possible, for example, to mark the dirty sections in the way that they will be ignored in the search process and only the edited sections will be the subject of search. If time is in short supply, the easily edited sections can be worked on first, and when time allows (if it allows!), the dirty sections will be edited and included into the clean text corpus.

3 Using morphological analyzer for finding scanning errors

When the text is manually edited and the scanning mistakes are corrected, there will still be mistakes left. The human eye is not faultless, and the thought wanders its own ways. Using a morphological analyzer we can detect each remaining error. Using a list of still erroneous words, we can correct the remaining mistakes in text. As a result, we have a text free of scanning errors.

4 Searching methods

When we have the corrected and error free text, we can prodeed in two ways for constructing information retrieval systems. We can use direct and concrete string search, where we search for words in context using concrete strings as search keys. We can also construct an advanced search system, where we first analyze and disambiguate the text, and then we can direct the search to the enriched text form. This method was described in Technical Report No. 45 (2019).

4.1 Direct string search

Several search systems are available for direct string search. There is no need to describe them here. They range from single word search to combinations of words, and also boolean operators can be used for constructing the search key. Also, the amount of context can be defined in many of them. Findings can be sorted according to the search key.

For languages such as Swahili, the string search methods are helplessly inaccurate. It is sometimes impossible to formulate the search key, because the word may have only one single consonant as a stable element, and everything before and after is inflection or derivation. In fact, for morphologically complex languages there is need for a more sophisticated search system, where each inflected word is attached to its base form and selected morphological information. I will describe this system below.

4.2 Advanced search system using analysed text form

It is possible to enrich the text so that each word has also its base form plus additional grammatical information. Morphological analysis and disambiguation (including syntactic mapping) produces a rich set of linguistic information, which can be used for making search more accurate. If we know for each word its lemma and POS category, we have already taken a big step towards accuracy and coverage of search.

Various language types have very different problems in constructing search systems. English, for example, has only limited inflection, and the words can often be found by using the direct string search. But if we need also to know their POS kategory in each context, we must first analyse and disambiguate the text. Disambiguation is particularly importnt in English for defining the POS category. If our language is a Bantu language such as Swahili, the main problem is to find the lemma of the word, because words, especially verbs, inflect to both directions. On the other hand, defining the POS category in such languages is not a big problem, because the combination of morphemes, which can be attached to the stem, is often unique for each POS category, and disambiguation has a less central role.

Regardless the language type, morphological analysis and disambiguation enhance greatly the accuracy and coverage of search.

Below are examples of enriched text in various languages.

(3) **Swahili**: Siku {siku_N} hizi {h_PRO} habari {habari_N} za {a_GCON} miji {mji_N} ya {a_GCON} pwani {pwani_N} zimechafuka {chafuka_V} sana {sana_ADV} tangu {tangu_PREP} Tanga {Tanga_PROPN} hatta {hatta_PREP} Lindi {Lindi_PROPN} na {na_CC} Mikandani {Mikindani_PROPN}.

(4) **Finnish**: 1_§_2 Suomen {Suomi_PROPN} valtiosääntö {valtiosääntö_N} on {olla_V} vahvistettu {vahvistaa_V} tässä {tämä_PRON} perustuslaissa {perustuslaki_N}.

(5) **English**: Various {various_A} languages {language_N} types {type_N} have {have_V} very {very_ADV} different {different_A} problems {problem_N} in {in_PREP} constructing {construct_V} search {search_N} systems {system_N}.

In all above examples, each surface word form has a lemma and its disambiguated POS category. Therefore, regardless of language, the same search system can be used in all of them. When the text is in two formats - the text itself and its analysed form - search can be directed to either form.

5 Advanced search from Msimulizi

In (6) is an example of what the enriched Swahili text looks like. The extract is from *Msimulizi*.

(6)
Siku_hizi {siku_hizi_ADV} habari {habari_N} za {za_GEN-CON} miji {mji_N} ya {ya_GEN-CON} pwani {pwani_ADV} zimechafuka {chafuka_V} sana {sana_AD-ADJ} tangu {tangu_PREP} Tanga {Tanga_PROPN} hatta {hatta_N} Lindi {Lindi_PROPN} na {na_CC} Mikindani {Mikindani_PROPN}.
Labda {labda_ADV} Wadoicha {Wadoicha_PROPN} wamekosa {kosa_V} utaratibu {utaratibu_N} katika {katika_PREP} kuanza {anza_N} shughuli {shughuli_N} yao {yao_POSS} ya {ya_GEN-CON} biashara {biashara_N}, {,_COMMA} hatta {hatta_N} wamethaniwa {amethaniwa_N} wanakuja {ja_V} kutwaa {twaa_V} inchi {inchi_N} ya {ya_GEN-CON} watu {mtu_N} wa {wa_GEN-CON} pwani {pwani_ADV}.

Msimulizi contains many such spelling conventions, which are not in current use any more. It would be necessary to enhance the analyzer, so that it recognizes also those deviant forms. In the current implementation this is not done.

Therefore, we see that the word *Wadoicha* has been guessed correctly as a proper name, but its stem form should be *Mdoicha*. The words *hatta* (correctly: hata_ADV) and *wamethaniwa* (correctly: dhani_V) were guessed as nouns, and the surface word form was copied as stem. The inclusion of these rare forms into the morphological lexicon would solve the problem.

We can direct the search to this enriched text form. We test with the string mtumwa_N. The hit will be surrounded with '[' and ']' instead of '{' and '}' (7).

(7)
$echo mtumwa_N | /home/arvi/msimulizi-2/find-msimulizi-pro
Twalikuwa {Twalikuwa_PROPN} na {na_CC} habari {bari_V} ya_huzuni {ya_huzuni_ADJ} nzito {zito_ADJ}, {,_COMMA} sasa {sasa_ADV} imetengenea V ndiyo {ndiyo_V-BE} ya {ya_GEN-CON} ndugu {ndugu_N} yetu {etu_PRON} mmoja {moja_NUM} aliyetumbukia {tumbukia_V} katika {katika_PREP} habari {bari_V} ya {ya_GEN-CON} kuuza {uza_N} na {na_CC} kununua {nunua_N} watumwa [mtumwa_N], {,_COMMA} sasa {sasa_ADV} tumempata {pata_V} tena {tena_ADV}, {,_COMMA} kwani {kwani_CONJ} amelirejea {rejea_V} kundi {kundi_N} lake {ake_PRON} la {la_GEN-CON} Umasihiya {Umasihiya_PROPN}.
Hatimaye {hatimaye_ADV} yule {yule_PRON} Mwarabu {Mwarabu_N} aliyemnunua {nunua_V} alimwaminia {aminia_V} vyote {ote_PRON}, {,_COMMA} hatta {hatta_N} alimpa {pa_V} shauri {shauri_N} afuatane {fuatana_V} naye {naye_CC} kwenda {kwenda_V} Mrima {mrima_N} kununua {nunua_V} watumwa [mtumwa_N], {,_COMMA} bwana {bwana_N} akasema {sema_V}, {,_COMMA} twende {enda_V} kwanza {kwanza_ADV} Zingibari {Zingibari_PROPN} baadaye {baadaye_ADV} twende {enda_V} Mrima {mrima_N}.

Bwanawe {Bwanawe_PROPN} alitangulia {tangulia_V} kutoka {kutoka_PREP} Pemba {Pemba_N} kuja {ja_V} Zingibari {Zingibari_PROPN} akitumaini {tumaini_V} mtumwa [mtumwa_N] atafuata {fuata_V}.
Ikawa {wa_V} siku {siku_N} ile_ile {ile_ile_PRON} ya {ya_GEN-CON} harusi {harusi_N} ya {ya_GEN-CON} C. {C._N} Singano {Singano_PROPN} na {na_CC} Emily {Emily_PROPN} Beza {beza_V} na {na_CC} A. {A._N} Yakuti {yakuti_N} na {na_CC} M. {M._N} Shantu {Shantu_PROPN}, {,_COMMA} mmoja {moja_NUM} wa {wa_GEN-CON} watu {mtu_N} wa {wa_GEN-CON} Mbweni {mbwe_N}, {,_COMMA} jina {jina_N} lake {ake_PRON} Nicholas {Nicholas_PROPN} Neiluwa {Neiluwa_PROPN} alikilima {lima_V} katika {katika_PREP} kikonde {konda_V} chake {ake_PRON}, {,_COMMA} alipoinuka {inuka_V} aona {ona_V} mtoto {mtoto_N} kama {kama_ADV} wa {wa_GEN-CON} umri {umri_N} wa {wa_GEN-CON} miaka {mwaka_N} labda {labda_ADV} 6 {6_NUM} ao {ao_N} 8 {8_NUM}, {,_COMMA} naye {naye_CC} hajui {jua_V} Kiswahili {Kiswahili_PROPN} akamwuliza {uliza_V}, {,_COMMA} habari {bari_V} zake {ake_PRON}, {,_COMMA} mtoto {mtoto_N} akajibu {jibu_V}, {,_COMMA} Mimi {mimi_PRON} ni {ni_V} mtumwa [mtumwa_N], {,_COMMA} nalikuwa {wa_V} katika {katika_PREP} chombo {chombo_N} tukienda {enda_V} Pemba {Pemba_N} tukafika {fika_V} karibu_ya {karibu_ya_PREP} pwani {pwani_ADV} upande {upande_N} wa_Kusini {wa_Kusini_ADJ} Zanzibar {Zanzibar_N} ndio {ndio_V-BE} Chukwani {Chukwani_PROPN}, {,_COMMA} ikawa {wa_V} usiku {usiku_N} mimi {mimi_PRON} nikaruka {ruka_V} majini {jini_N} nikaogelea {ogelea_V} hatta {hatta_N} pwani {pwani_ADV}, {,_COMMA} Chombo {chombo_N} kikaenda {enda_V} joshi {joshi_N}, {,_COMMA} na {na_CC} sasa {sasa_ADV} natafuta {tafuta_V} ulindo V.
Ilikuwa {wa_V} siku {siku_N} ya {ya_GEN-CON} 11 {11_NUM} ya {ya_GEN-CON} March {March_PROPN}, {,_COMMA} nalienda {enda_V} tembea {tembea_V} nionane {onana_V} na {na_PREP} watu {mtu_N} wa {wa_GEN-CON} Mpakani {mpaka_N}, {,_COMMA} mimi {mimi_PRON} na {na_CC} Nicholas {Nicholas_PROPN} Faraji {faraji_N}, {,_COMMA} tukirejea {rejea_V} tukapewa {pewa_V} habari {habari_N} ya_kuwa {ya_kuwa_CONJ} kama {kama_CONJ} mtu {mtu_N} mwanamke {mwanamke_N} kaja {ja_V} kwa {kwa_PREP} Bwana {Bwana_N} G. {G._N} Dale {Dale_PROPN}, {,_COMMA} tukifika {fika_V} kwangu {angu_PRON} tukamwona {ona_V}, {,_COMMA} tukamwuliza {uliza_V} habari {bari_V}, {,_COMMA} akasema {sema_V}, {,_COMMA} Mimi {mimi_PRON} ni {ni_V} mtumwa [mtumwa_N] nimetoka {toka_V} pwani {pwani_ADV}, {,_COMMA} bwana {bwana_N} wangu {angu_PRON} na {na_CC} bibi {bibi_N} yangu {angu_PRON} amekufa {fa_V}, {,_COMMA} nami {nami_CC} sina {sina_V} ulindo V sasa {sasa_ADV}, {,_COMMA} natafuta {tafuta_V} kwa {kwa_PREP} Mzungu {Mzungu_N} huenda {enda_V} nikapata {pata_V}.
Tena {tena_ADV} maneno {neno_N} mengi {ingi_PRON} yako {yako_V-BE} Ulaya {Ulaya_N} juu_ya {juu_ya_PREP} biashara {biashara_N} ya {ya_GEN-CON} watumwa [mtumwa_N] na {na_CC} ya {ya_GEN-CON} mvinyo {mvinyo_N} mbaya {baya_ADJ}, {,_COMMA} zipate {pata_V} kuzuiliwa {zuiliwa_V}.

When this is done, the strings surrounded with '{' and '}' will be removed (8).

(8)
Twalikuwa na habari ya_huzuni nzito, sasa imetengenea V ndiyo ya ndugu yetu mmoja aliyetumbukia katika habari ya kuuza na kununua watumwa [mtumwa_N], sasa tumempata tena, kwani amelirejea kundi lake la Umasihiya.
Hatimaye yule Mwarabu aliyemnunua alimwaminia vyote, hatta alimpa shauri afuatane naye kwenda Mrima kununua watumwa [mtumwa_N], bwana akasema, twende kwanza Zingibari baadaye twende Mrima.
Bwanawe alitangulia kutoka Pemba kuja Zingibari akitumaini mtumwa [mtumwa_N] atafuata.
Ikawa siku ile_ile ya harusi ya C. Singano na Emily Beza na A. Yakuti na M. Shantu, mmoja wa watu wa Mbweni, jina lake Nicholas Neiluwa alikilima katika kikonde chake, alipoinuka aona mtoto kama wa umri wa miaka labda 6 ao 8, naye hajui Kiswahili akamwuliza, habari zake, mtoto akajibu, Mimi ni mtumwa [mtumwa_N], nalikuwa katika chombo tukienda Pemba tukafika karibu_ya pwani upande wa_Kusini Zanzibar ndio Chukwani, ikawa usiku mimi nikaruka majini nikaogelea hatta pwani, Chombo kikaenda joshi, na sasa natafuta ulindo.
Ilikuwa siku ya 11 ya March, nalienda tembea nionane na watu wa Mpakani, mimi na Nicholas Faraji, tukirejea tukapewa habari ya_kuwa kama mtu mwanamke kaja kwa Bwana G. Dale, tukifika kwangu tukamwona, tukamwuliza habari, akasema, Mimi ni mtumwa [mtumwa_N] nimetoka pwani, bwana wangu na bibi yangu amekufa, nami sina ulindo sasa, natafuta kwa Mzungu huenda nikapata.
Tena maneno mengi yako Ulaya juu_ya biashara ya watumwa [mtumwa_N] na ya

The hits are now surrounded with '[' and ']'.

We can also test with a verb. Our search key is pata_V. There are so many hits that only a few of them are in (9).

(9)
Watson wakaenda Rovuma pamoja nao watoto wawili, wakalala siku tatu wakarudi hawakupata [pata_V] kitu.
Bassi Bwana Hainsworth tu akasalisha, kwani tumepata [pata_V] makasisi wawili.
Mwalimu Paul Kasese alikwenda awinde baada_ya siku kuu ya Pasaka, hakupata [pata_V] kitu.
Bassi kule_kule Msumba twalipata [pata_V] habari za_kweli za mwenzetu Stefano Rehani, kama ameuawa kwa amri ya Mkaliwili mwenyi mji ule_ule alipokamatwa yeye na alipouthika V Bwana Johnson.
Tena maneno aliyotuandikia naona hatutayasahau, kwa ginsi yapate [pata_V] kututia_moyo tuzidi kuendelea mbele.
Jambo hili lipate [pata_V] kutimizwa kwa_kweli yatakiwa imani hayi na nia imara kwenu.
Ndio hii Muungu anayowaitia ninyi kwa ufathili wake mmtolee, apate [pata_V] kutimiza nayo shauri lake yee.
Johnson alikamatwa na Amdoka, ndio yaliyompata [pata_V].

Werbs have typically many forms. Not only prefixes complicate search, also suffixes cause problems. When we use the method described here, such problems do not occur,

provided that the wordform is described properly in the lexicon, and that ambiguity is solved.

6 Conclusion

I have shown above how scanned old texts can be converted into text form, edited, and then converted into enriched thest form using morphological analysis and disambiguation. This text form enhances almost faultless information retrieval, which greatly improves and speeds up research work with such texts. The report includes only very simple search examples. However, all kinds search methods can be used, including surface form search, partial word search, POS search, search of multiple words simultaneously, and counting of found hits.

References

Hurskainen Arvi, 2019. Intelligent search engines. *Technical Reports on Language Technology*. Report No. 45. www.njas.helsinki.fi/salama/intelligent-search-engines.pdf