Rule-based language technology for African languages

Hurskainen, Arvi

Northern European Association for Language Technology
2023-04-17

Hurskainen , A 2023 , Rule-based language technology for African languages . in A Hurskainen , K Koskenniemi & T Pirinen (eds) , Rule-Based Language Technology . NEALT Monograph Series , no. 2[1] , Northern European Association for Language Technology , Tartu , pp. 228-239 .

http://hdl.handle.net/10138/357177

Rule-based language technology for African languages[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Africa is such a language area, where rule-based language technology could have a strong influence on the status of local languages. As statistical and neural approaches require large masses of text for training the language model, rule-based methods can be applied also to languages with no traditional language resources. The development of language technology systems for minor languages would not only provide useful tools for language users. It would also contribute to the elevated status of those languages and thus help in maintaining those languages to be alive.

The chapter looks at the current situation in Africa particularly from the viewpoint of rule-based language technology.

Key Words: *African languages, language technology*

1 Introduction

When I enter into the Google search engine the search key *Language technology for African languages*, among the first hits is my article *Sustainable language technology for African languages*[2], along with *African Languages Technology Initiative* (Alt-i)[3] a Nigerian-based project. There are also many other websites, some superficial and commercially oriented. Many of them state that although there are 2,000 languages in Africa, and the number of speakers is increasing, the languages are largely out of reach of computational communication, and that something should be done to bridge the digital gap.

There is not much point in imitating in Africa what was done in Europe in language technology, because there are huge differences in terms of challenges and resources that could be allocated for solving the problems. Europe, European Union in particular, has given a unique boost to statistical machine translation (SMT), and more recently to neural machine translation (NMT), through the policy of maintaining documentation in all official EU languages. This is an unparalleled resource, a multilingual text corpus, the

---

[1] **Ref**: Hurskainen, Arvi. 2023. Rule-based language technology for African languages. In: Arvi Hurskainen, Kimmo Koskenniemi, and Tommi Pirinen (eds.), Rule-Based Language Technology. NEALT Monograph Series, 2:228-239. https://dspace.ut.ee/handle/10062/89595
[2] https://doi.org/10.4324/9781315392981-19
[3] http://www.alt-i.org/

kind of which is hardly available anywhere else. These masses of training materials have given the impression that the language translation problem can be solved through self-learning computer algorithms.

When reading the recently written articles on African language technology, one gets the impression that the statistical and neural approaches are the solution also in MT of African languages, while rule-based methods are considered too labour-intensive requiring too much high-skilled work to be competitive.

There is not much point in arguing about the excellence of either approach. We should keep in mind the bare facts.

2 Basic facts

The computer is a calculating machine, and nothing else. In certain types of calculation it outperforms human beings in speed. Even the so-called learning of computers, or computer algorithms, is simply the result of calculation procedures. If B follows A ten times, and if only once C follows A, then the conclusion is that A followed by B has special significance. This simplified example shows how statistical and neural algorithms in MT work.

Language in written form is not a sequence of characters, which as such could be fed to a computer algorithm with the hope that it could sort out from it something meaningful. It is true that texts have regularities, recurrent sequences of characters and words, and the algorithm can show those regularities to human beings, so that they can inspect them and then conclude what the regularities mean, and then give them labels. The use of machine learning algorithms has been tempting also in language technology, because it saves the human being from thinking. However, this method for understanding the language is counter-intuitive and prone to errors.

The strength of the rule-based approach is that it makes use of the calculating power of the computer, and of the innate ability of the human being to classify linguistic phenomena. Linguistic classes are not based on similar sequences of characters and words, although in some cases they seem to be. No computer algorithm can overcome the problem of classifying correctly words into all the fine-grained classes needed in language technology. A trained linguist can do it easily.

There is also difference in the expectations of performance. If a guessing algorithm reaches a 95 % accuracy, in one camp this is considered extremely high. In another camp even the 99,5 % accuracy is considered too low, and the analyser should be improved.

The often heard argument against rule-based approaches is that it is too labour-intensive. Too much human input is needed for making the approach viable. The argument is true, but the amount of human input depends very much on such preconditions as the computer skills of the developer, the level of the developer in mastering the grammar, availability of dictionaries and other language repositories.[4]

I will give an example of the time needed for developing a comprehensive language analyser. A few years ago, I decided to construct my own analyser for Finnish, a

[4] See the chapter on GiellaTekno in this book.

morphologically very complex language[5]. It took about two weeks to construct the analysis system, starting from scratch. Finnish is my mother tongue, which of course helps in the work process. My understanding of Finnish grammar dates back to school years, with no further studies thereafter. The only external help was the list of Finnish words with their inflection codes. This list was prepared by *Kotimaisten kielten keskus* (The Center for National Languages). The list was very helpful, although I had to add new words and make more sub-classifications to inflection classes.

Is two weeks for constructing a comprehensive language analyser for a morphologically complex language too long time? That time is easily wasted in waiting when the super-computer crunches a huge web-crawled corpus with lots of errors in the hope of putting out something useful.

In addition, working with rule-based approaches helps to learn such features about the language, which one can never learn by reading grammar books. Of course, the analyser needs further testing and corrections. But all this work contributes to the improved overall performance of the system.

The analysis system alone is of little use without the disambiguation system, which resolves the ambiguities on the basis of context. Ambiguity resolution must be performed at least on morphology and semantics, and also in solving other problems.

Another example of the time needed for constructing a morphological analyser for a new language is my four-month period at Makerere University in Uganda as an invited quest professor in 2008. We had a team of researchers, each working with one of the eight languages in the program. All languages, except for one, were Bantu languages. Although the students had no previous knowledge of language technology, they soon grasped the idea of how morphemes are concatenated into words, and how each feature can be linguistically described. In that course we used the Xerox tool package[6]. The advantage in the course was that most participants were staff members of the Language Department with clear understanding of the language structure. When working as a team with closely related languages, the students were able to get support from other team members. They also had access to the lexical materials compiled at the department, thus speeding up the development process.

3 Competing approaches perplex researchers

Leading African universities are targets for new approaches in language technology. Because research teams need new markets for securing funds for their work, they look for markets in such places as Africa, where there is clear need of technological innovations. As a result, salespeople of competing approaches travel in Africa trying to convince that it is just their approach that is the best in the market.

In fact, the differences between rule-based approaches are not profound. All of them address the same problems, and their methods for solving the problems are linguistically motivated. Guessing is avoided as much as possible. Programming offers many ways for

[5] http://www.njas.helsinki.fi/salama/morphological-analyzer-of-finnish.pdf
[6] https://web.stanford.edu/group/cslipublications/cslipublications/site/1575864347.shtml

solving the problems, and the differences in solving methods create the impression that the approaches are very different. It is also true that when one learns to describe a language using one method, the shift to another method is rather easy. An example of this is the shift from using the Xerox tool package[7] to Grammatical Framework[8] package in describing some South-African Bantu languages. Any approach forces the developer to classify carefully words and morphemes, which in itself is a solid background for developing the systems within any environment.

A much more serious problem is the competition between statistical/neural and rule-based approaches. Even misleading impressions on the supremacy of statistical approaches can be given. An example is the article by De Pauw and De Schryver (2009) claiming that their statistical tagger of Swahili outperforms the rule-based tagger that was used for tagging the Helsinki Corpus of Swahili 1.0 (HCS). The fact is that the writers extracted the linguistic information from the tagged corpus and made use of it in constructing their own system. In other words, the basis of their data-driven approach was the detailed linguistic information available in the tagged corpus, and their own guessing algorithm resolved some problematic cases, thus showing the supremacy of their approach. Without access to HCS the result would have been entirely different. Also, in the new larger version of HCS, the Helsinki Corpus of Swahili 2.0[9] most errors present in the earlier version have been solved.

The writers give credit to the quality of HCS, but the message is that a data-driven approach can outperform the rule-based analyser. In the concluding part, they give credit to earlier rule-based work: "This research showed how previous rule-based efforts can go hand in hand with a data-driven approach and help construct a more accurate lemmatiser that is inherently capable of analysing previously unseen word forms, even when the underlying lemma is unknown."[10]

But the more obvious possibility did not come into their mind, that the rule-based parser itself should have been corrected on points, where it made mistakes. By so doing, there would not have been any need for guessing.

4 Language technology research infrastructures in Africa

A promising initiative in language technology development is the African Languages Technology Initiative (Alt-i), started 2002 in Nigeria. This 20 years old initiative has a grand aim, as stated at their website: "ALT-I aims to take African languages - and thereby, African cultures - into the information age by developing the human and technical resources to enable African languages to engage with the widest spectrum of ICT."[11] The initiative is a network of research institutions in Nigeria, and it establishes

[7] http://users.itk.ppke.hu/~sikbo/nytech/gyak/05_morfo/xfst/book.pdf_1.pdf
[8] https://www.grammaticalframework.org/
[9] http://urn.fi/urn:nbn:fi:lb-201608301
[10] https://www.researchgate.net/publication/253005034_African_Language_Technology_The_Data-Driven_Perspective [accessed Aug 29 2022].
[11] http://www.alt-i.org/about/

partnerships with other research agencies in the same field. It is not known what measures the initiative has taken for achieving the goals.

In the web one encounters claims, such as this: "This Kenyan startup has developed technology that can translate anything into indigenous African languages. Kenyan startup Abantu AI has developed a deep learning technology in the field of natural language processing (NLP) that translates from major world languages to indigenous African languages."[12] Claims such as this are dangerous, because they may perplex many people, who are unaware of the real nature of machine translation and of its complexities. And then when it is found out that the promises have no foundation, distrust towards language technology increases.

5 Specific features of African languages

Africa south of the Sahara is lucky compared with Asia in one important feature. That is the writing system. While developers of language technology of many Asian languages must address the writing systems with thousands of different graphical symbols that convey the semantic meaning of the message, almost all African languages use the alphabetical writing system. This is a huge advantage that creates a solid foundation for language technology.

Another feature in African languages is that the writing quite closely imitates the pronunciation. The writing of African languages has a fairly short history, which is why the development of language use has not receded from the writing. The situation is very different in languages such as French and English, where pronunciation and writing often mismatch.

The alphabetical writing system does not mean, however, that there would be no problems in developing language technology. I will discuss some of them below.

5.1 Tone marking

Most Bantu languages use the traditional alphabetical characters in writing. Although almost all of them are tonal languages, they often do not mark tones in writing. This is not considered problematic in language use, because readers are able to figure out the correct tone pattern intuitively. From the viewpoint of language technology, the absence of tone marking increases ambiguity in the analysed text, but it is not an insurmountable obstacle for processing.

In developing text-based language technology, we are entirely dependent on the writing styles of the people. The most straight-forward solution is that we accept the texts as they are, and we cope with the unnecessarily big amount of ambiguity.

Another method is to convert the text into the format, where tones are marked accordingly. This can be done using context-sensitive rules. The need for tone-marking would be necessary especially in text-to-speech technology. I studied the possibility of

[12] https://disrupt-africa.com/2022/05/12/this-kenyan-startup-has-developed-technology-that-can-translate-anything-into-indigenous-african-languages/

converting text without tone marking into tone-marked text using Kinyarwanda as test language[13].

Tone marking in text is usually implemented using diacritic marks above the letters, such as á (high tone), à (low tone), and â (contour tone). In language technology, such difficult-to-handle marking is not necessary. Marking that is easier to handle can be used instead. I will discuss this problem in more detail below.

5.2 Re-writing problematic characters

The character sets that can be used in developing language technology have increased in size, and Unicode contains practically all characters found in African writing. The problem is, however, that some development environments in the processing chain do not allow the use of all characters in the language.

In practice it often turns out that although some development platforms allow larger character sets, the chain of programs may also contain such sections that allow only ASCII characters. This can be very annoying.

A safe solution, although not very elegant, is to rewrite problematic characters first into something else, for example into bi-graphs with two such characters that never co-occur in normal text, and then in the end of the process return them into original form. I have found this necessary even with such characters as Ä/ä and Ö/ö, which are common in many European languages. Problems with these characters occur especially when working in such servers, which one does not have control of. Some editor distributions, for example Emacs, can easily interpret a text with such characters as Japanese text and save the whole file with Japanese characters. This is a good example of what happens if computer guessing is automated too much. If these characters are first rewritten into bi-graphs, such problems can be avoided.

There are, especially in West Africa, writing systems with such characters that are not encoded in language technology environments, although the situation is becoming better over the years. Kwa languages are tone languages with tone marking, and with marking of some other linguistic features. All this cannot be directly handled in processing. Also here, the possibility of rewriting the text first using only acceptable characters would solve the computational problems.

5.3 Disjoining writing systems

The writing of verbs in Bantu languages has been a subject of dispute since the time, when people, often missionaries, began to put these languages on writing. The key problem was whether verb prefixes should be considered as part of the verb, or should they be considered as separate words. The problem has not been solved so far, and both types of writing occur. Examples are such languages as Northern Sotho and Zulu. The former uses disjoining writing, and the latter uses conjoining writing. Kwanyama is an interesting case, because the same language uses disjoining writing in Northern Namibia, but conjoining writing in Angola.

---

[13] http://www.njas.helsinki.fi/salama/art-tone-marks.pdf

It is understandable that both styles of writing exist, because the verb prefixes are in strict order, and they do not allow other words between the prefix and the stem. Therefore, it is natural to think that the set of prefixes and the verb stem form a single word. On the other hand, when we translate the verb, together with its prefixes, into another language, the result is more than one word. We can conclude that semantically the Bantu verb contains also other words, such as, subject, relative pronoun, and object, albeit in pronoun form. Therefore, also the proponents of disjoining writing systems have a strong founding. Nevertheless, today most Bantu languages use conjoining writing system.

From the viewpoint of language technology, the writing system has a strong effect on the manner, how the analysis system is constructed. If the verbs in disjoining writing systems are analysed as such, we get a large amount of ambiguity. The ambiguity does not concern the part-of-speech affiliation only. There is also ambiguity in relation to noun class because a marker may refer to more than one noun class.

One solution for decreasing ambiguity would be to pre-process the text so that the separate verb prefixes would first be joined to the verb stem, and the analysis would then be done on this text form. In other words, the disjoining writing system would be converted into conjoining writing system. I have, together with colleagues, implemented this on Kwanyama[14] and Northern Sotho[15]. When the verb morphemes are joined with the stem, the analysis system can be constructed in the same way as in languages with conjoining writing systems.

However, it is also possible to do the analysis directly on the disjoining writing. The added ambiguity could then be disambiguated using context sensitive rules.

It is important to note that one faces the same ambiguity problems regardless the method applied. The conversion of a disjoining writing system into a conjoining one might be useful in cases such as Kwanyama, which in Angola is written conjointly and in Namibia disjointly.

6 Rule-based approaches open possibilities for applications

Discussion on language technology tends to be too much focussed on machine translation. This must be due to the dominance of statistical and neural methods in translation, because they have been successful in translation between major languages. Such methods can only be used for mapping between sequences of characters. There is a wide range of applications, for which only rule-based methods are suitable, because they provide precise analysis and systematic tagging. Below I will list some of them.

6.1 Grammar books

The construction of a comprehensive morphological lexicon is a solid basis for writing such grammars that include all important features of the language. If all features are not

[14] https://doi.org/10.53228/njas.v10i3.578
[15] https://doi.org/10.53228/njas.v14i4.251

described in the morphological lexicon, the analyser does not function faultlessly with all texts.

There is an urgent need for good grammars for African languages. This area of research is badly neglected in many areas in Africa.

The construction of a grammar book could proceed in the following way: We first collect a text corpus of the language. The texts can be from books, newspapers, and the web. Removing typos from texts would help in the process.

Then the linguist starts constructing the morphological analyser, piece by piece, using the text corpus for checking the correctness of the morphological lexicon. When each word in the corpus is correctly analysed, the initial phases in constructing the analyser are over. In many words, there will be ambiguity, that is, the word has more than one interpretation. This is natural, but one should make sure that at least one of the readings is correct in that context.

The morphological lexicon is like a grammar book, although in a very different form than traditional books. It contains all relevant information that is needed in grammar writing. Whether this knowledge is processed into book form is another question. A grammar in the web is certainly much more useable than printed books.

6.2 Accurate information retrieval

Information retrieval in languages with rich morphology is problematic because words inflect, which makes the formulation of search keys difficult if not impossible. Particularly verbs in Bantu languages inflect to both directions, which makes accurate information retrieval problematic.[16] Information retrieval is needed wherever there is text.

Traditionally, accurate information retrieval has been possible only from linguistically tagged text corpora. Information retrieval from plain text was done using such search keys that matched directly to the text. It is also possible to convert the text into analysed format with selected linguistic tags, and then direct the search to this enriched text. Using this method, the accuracy is high.

When searching from very large text masses, it is not always feasible to convert the whole text into analysed form. The task can be split into two phases. First a rough search is done from the text, and the result includes all the wanted hits, and an arbitrary amount of non-wanted hits. This reduced text is then converted into analysed format, and accurate search is then done on this text. Using this method, also large masses of text can be used for accurate information retrieval.[17]

6.3 Web dictionaries

By web dictionaries I do not mean a printed dictionary converted to digital form and put on the web. It is an entirely different kind of concept. The search system for a dictionary

---

[16] https://www.researchgate.net/publication/249711652_Information_Retrieval_and_Two-directional_Word_Formation1

[17] http://www.njas.helsinki.fi/salama/accurate-information-retrieval-from-large-corpora-1.pdf
http://www.njas.helsinki.fi/salama/accurate-information-retrieval-from-large-corpora-2.pdf

can be made to run through a language analyser, which makes it possible to give any wordform of the word as a search key.[18] The system finds its stem and looks for its possible interpretations in the dictionary. The system can also give use examples of the word, including translation in another language. Also frequencies of the word can be given as well as cross-references to synonyms or in other ways related words.[19]

For most African languages, there are no dictionaries. Such a computational dictionary, compiled with the help of a morphological analyser and a sufficiently large text corpus, would boost the visibility and usefulness of African languages. Since the wireless web devices, such as mobile phones, are spreading rapidly in Africa, smart web dictionaries could be made available to most people in Africa.

The manually compiled dictionaries tend to have the compiler bias. That is, the compiler has his own preferences, which may distort the balance of the dictionary. The dictionary may have interesting but bizarre words, although at the same time many common words are missing. This was clearly revealed in my evaluation of *Kamusi ya Kiswahili Sanifu* (Hurskainen 1994) and the comparative study of five Swahili dictionaries[20].

Today, dictionary compilation for low-resource languages is best to combine with the development of the analysis system of the language. When the known words are described in the morphological lexicon, it is easy to control the correctness of the analysis system and, what is even more important, the absence of words in the system. The analysis system detects all words in any of their inflected forms if they are described in the lexicon. When new texts are run through the analysis system, it is easy to detect truly new words, and by inspecting them, they can be classified and added to the analysis system in correct places. This enhanced analysis system can then be used in compiling the enhanced computational dictionary. This is a much better method than a method recently proposed[21].

6.4 Self-tutored language learning applications

Since many African languages, for example Bantu languages, have a rich verb morphology, learning word-forms and linguistic structures is a tedious process. In a self-learning environment, it is often impossible to check whether the sentence and word structures, or even word structures, are correct. Even printed grammars are seldom helpful. Rule-based language technology offers a solid basis for constructing such learning applications that guide the learner interactively through the learning process. These applications can be used through the web.

Language learning applications currently on the web are often colourful, and they make use of graphical images, giving an impression of an attracting application.

---

[18] See the chapter on GiellaLT in this book for similar applications.
[19] http://www.njas.helsinki.fi/salama/salama-dictionary-compiler.pdf
[20] https://doi.org/10.53228/njas.v11i2.361
[21] https://www.researchgate.net/search.Search.html?query=towards+a+monitor+corpus&type=publication

However, they are often of little use, because they do not teach the principles of the grammar. A good morphological analyser, combined with a disambiguation component, would bring such applications to an entirely different level.[22]

Above I have described only four types of applications that can be useful for a large number of people. I have not mentioned spelling checkers, grammar checkers, hyphenators and the like, because these applications are already a standard in text processing. However, for minor languages these text checking applications are still missing, and work with those languages is needed.[23]

7 The status of language technology in Africa in general

Despite various efforts for developing language technology for Africa languages, there is no such infrastructure that would locate various projects as parts of the larger structure that covers all African language areas and language types. Yet the initiatives can learn from each other provided that the communication between them is intensive.

It would be optimal to make all basic language technology environments as well as language resources open access, so that all unnecessary barriers would be removed. The problem with the open access environments is that, in order for them to succeed, they need public and continual funding. Most of the approaches described in this book are open access, but they also are continually supported by governments. Without public funding, open access technology cannot be developed and maintained.

When we move this concept to Africa, we encounter huge problems. Although African Union has expressed its support for African indigenous languages, there is no financial plan for making these hopes effective. European Union is in a different situation because it has decided to make all official documents available on every official language within the union. And this costs a lot of money if the target is achieved using human translators. Therefore, there is need to develop translation tools for saving translation costs. The documents of African Union are translated only to the few official languages of the union, and Swahili is the only indigenous language among them. The web page of AU has a statement: *Article 11 of the Protocol on Amendments to the Constitutive Act of the African Union states: The official languages of the Union and all its institutions shall be Arabic, English, French, Portuguese, Spanish, Kiswahili and any other African language.*

The text does not make clear what *any other African language* means. It cannot mean that all up to 3,000 languages of Africa are official languages of AU. The formulation means nothing in practise. One must conclude that African Union has not addressed the position of indigenous African languages even in its own communication. On the other hand, AU does not have mandate to direct the language policy of member states. Yet some sort of language policy would be needed, so that language technology initiatives could orient themselves to the situation and plan cooperation.

---

[22] http://www.njas.helsinki.fi/salama/rule-based-language-technology-and-self-tutored-language-learning-systems.pdf

[23] See the chapter on GiellaLT for implementation of proofing tools for minor languages.

Under African Union, there is African Academy of Languages (ACALAN), the aim of which is to promote African languages in various ways. One of its projects is *African Languages and the Cyberspace (ALC*). Its aim is 'To promote African languages in the cyberspace and apply Human Language Technologies to them.' This is the only place where language technology is mentioned in the work plan.

Perhaps the most promising initiative is the South African Centre for Digital Language Resources (SADiLaR) (Roux and Ndinga-Koumba-Binza 2019). It is a consortium of a number of South African universities and research institutions, which in a coordinated way develops digital languages resources and utilities and makes them all freely available for research purposes. It can be expected that such a coordinated effort will have permanent government funding, and costs can be saved, when resources can be shared in various projects. The development of high quality language resources requires highly skilled human input, and the reuse of such resources in various projects is cost-effective. SADiLaR has also links to other relevant framework such as CLARIN, which is part of the European Strategic Framework for Research Infrastructures Roadmap. CLARIN is a research infrastructure for language as social and cultural data. Helsinki Corpus of Swahili 2.0[24] is hosted by Kielipankki, which is part of CLARIN infrastructure.

References

De Pauw, Guy and De Schryver, Gilles-Maurice, 2009:
African Language Technology: The Data-Driven Perspective. Conference: Lesser Used *Languages and Computer Linguistics* II.

De Schryver, Gilles-Maurice and Minah, Nabirye, 2022:
Towards a monitor corpus for a Bantu language. A case study of neology detection in Lusoga. Dictionaries and Society, *XX EURALEX International Congress,* 12-16 July 2022. Mannheim, Germany.

Hurskainen, A., 1994:
Kamusi ya Kiswahili Sanifu in test: A computer system for analyzing dictionaries and for retrieving lexical data. *Afrikanistische Arbeitspapiere*: Schriftenreihe des Kölner Instituts für Afrikanistik., 37, 169-179.

Hurskainen, A., 2002:
Tathmini ya Kamusi Tano za Kiswahili. *Nordic Journal of African Studies*, 11(2): 283-301. https://doi.org/10.53228/njas.v11i2.361

Hurskainen, A., & Halme, R. 2001:
Mapping between disjoining and conjoining writing systems in Bantu languages: Implementation on Kwanyama. *Nordic Journal of African Studies*, 10(3), 399 - 414. https://doi.org/10.53228/njas.v10i3.578

Hurskainen A., Louwrens L. & Poulos, G. 2005:
Computational description of verbs in disjoining writing systems. *Nordic Journal of African Studies*14(4): 438-451. DOI: https://doi.org/10.53228/njas.v14i4.251

[24] **http://urn.fi/urn:nbn:fi:lb-2016011301**

Hurskainen , A J 2018:
Sustainable Language Technology for African Languages . in A Agwuele & A Bodomo (eds) , *The Routledge Handbook of African Linguistics*. Routledge Handbooks , Routledge , Abingdon , pp. 359-375 . https://doi.org/10.4324/9781315392981-19

Hurskainen, A., & Halme, R. 2001:
Mapping between disjoining and conjoining writing systems in Bantu languages: Implementation on Kwanyama. *Nordic Journal of African Studies*, 10(3), 399 - 414. https://doi.org/10.53228/njas.v10i3.578

Roux, Justus and Steve, Ndinga-Koumba-Binza, 2019:
African languages and Human Language Technologies. In: *The Cambridge Handbook of African Linguistics*: Chapter 22. Cambridge: Cambridge University Press. https://www.researchgate.net/publication/317662118_African_languages_and_Human_Language_Technologies