Bootstrapping NLP tools across low-resourced African languages: an overview and prospects

C. Maria Keet
Department of Computer Science
University of Cape Town
South Africa
mkeet@cs.uct.ac.za

Abstract

Computing and Internet access are substantially growing markets in Southern Africa, which brings with it increasing demands for local content and tools in indigenous African languages. Since most of these languages are low-resourced, efforts have gone into the notion of bootstrapping tools for one African language from another. This paper provides an overview of these efforts for Niger-Congo B (‘Bantu’) languages. Bootstrapping grammars for geographically distant languages has been shown to still have positive outcomes for morphology and rules or grammar-based natural language generation. Bootstrapping with data-driven approaches to NLP tasks is difficult to use meaningfully regardless geographic proximity, which is largely due to lexical diversity due to both orthography and vocabulary. Cladistic approaches in comparative linguistics may inform bootstrapping strategies and similarity measures might serve as proxy for bootstrapping potential as well, with both fertile ground for further research.

1 Introduction

Nearly 1.5 billion people live in Africa, of which many who speak multiple languages other than the relatively well-resourced languages English, French, and Arabic, among an estimated 1441 to 2169 African languages (Hammarström, 2018). Notably, by first language speakers, Swahili of the Niger-Congo family is the next-largest language (around 50 million L1 speakers, some 100-200 million overall), with as close second the Afroasiatic Hausa and then Oromo, and then Yoruba and Igbo in the Niger-Congo family with each around 28 million L1 speakers. While illiteracy exists in Sub-Saharan Africa, there are still very many people who can read and write in indigenous languages, some of which having official status in one or more countries where they are used in education, work, and social life. This entails a need for language support for task such as spelling and grammar checking, translation, and natural language generation. Within the context of the United Nations’ Sustainable Development Goals, they include language technologies to ameliorate the language gap (and consequent lower service (Hussey, 2012/2013)) in healthcare (Byamugisha et al., 2017; Marais et al., 2020), educational digital assistants to ease the burden of overworked teachers in crowded classrooms of up to 100 learners (Keet, 2021), and many other supportive tasks for society, such as machine translation for humanitarian response (Öktem et al., 2021). In this paper, we zoom in on the Niger-Congo B (‘Bantu’), or NCB, family of languages. In Joshi et al. (2020)’s classification of NLP support for languages, the NCB languages fall into three of the six categories: the “left behinds”, the “scraping-bys”, and the “hopefuls”, with isiZulu (spoken in South Africa) and Swahili (spoken primarily in Tanzania and Kenya) in the latter group.

A selection of the usual NLP tasks have been taken up for a few languages of the Niger-Congo family, notably indeed Swahili and isiZulu, and to a lesser extent Yoruba, Igbo, isiXhosa, and Runyankore. Examples are diverse. They range from corpus creation for data-driven NLP, such as the IsiZulu National Corpus (Khumalo, 2015) that was used for a statistical language model for a spellchecker (Ndaba et al., 2016), the Mashakane grassroots initiative that focuses on data-driven machine translation for multiple African languages (Nekoto et al., 2020), to data-driven text-to-speech (Marais et al., 2020) based on Qfency, and other language modelling and data augmentation (e.g., (Byamugisha, 2020; Mesham et al., 2021); see numbers from various open and paywalled sources, collated at https://en.wikipedia.org/wiki/Languages_of_Africa.
also Kambarami et al. (2021) for an overview). The main knowledge-driven approaches include terminology development in general (Khumalo, 2017) and domain-specific (e.g., (Engelbrecht et al., 2010)), and rule-based morphological analysers (Pretorius and Bosch, 2003; Bosch and Pretorius, 2017), grammars (Bamutura et al., 2020; Pretorius et al., 2017), and natural language generation (Keet and Khumalo, 2017; Byamugisha, 2019; Mahlaza and Keet, 2020). Most of the research has taken place over the past 5-10 years and is gaining pace, albeit still for only a slowly increasing number of NCB languages.

The low-resourced and very low-resourced languages\(^5\) face a ‘catch-22’, however: there are few language resources but one needs language and linguistics resources to increase the language resources. A well-known idea is to try to ‘bootstrap’ resources for a very low-resourced language from a low-resourced one; e.g., to bootstrap a spellchecker for isiNdebele from an isiZulu spellchecker, which are both languages in the Nguni group of the NCB languages in South Africa.

Theoretically this makes sense, but practically it is nontrivial to figure out bootstrapping potential and strategies. In this paper, we report on a preliminary review of published research on bootstrapping for the NCB languages to provide better insight into it such that it can better inform NLP tasks for NCB languages. Observing that it remains imprecise as to what should be assessed quantitatively to gauge bootstrapping potential, the research demonstrates that bootstrapping with rules and grammars extends to more languages than initially assumed and with data-driven approaches to fewer languages due to limited lexical proximity due to variations in vocabulary and orthography.

In the remainder of the paper, we first summarise NLP-relevant features of NCB languages and linguistics-focussed categorisations of NCB languages in Section 2. We then proceed to the key questions for bootstrapping and the review in Section 3, discuss in Section 4, and close in Section 5.

2 NCB languages: some key features

This section describes language and linguistic of NCB languages insofar as they are relevant to comp-

\(^5\)While there is no crisp demarcation of ‘low’ in low-resourced languages, it is to be understood as having only small (e.g., 20K tokens) or no curated monolingual or parallel corpora, limited (including outdated) or no grammar books, and typically also comparatively few researchers and funding.

utation thus far, with first key language features and then the grouping of subsets of NCB languages.

2.1 Grammar and orthography

The system of noun classes is emblematic for the NCB languages. Each noun belongs to a noun class and there are up to 23 noun classes; see Table 1 for a summarised overview. All NCB languages retained the lower numbers and are fairly similar up to noun class 11, in that the odd numbered classes contain nouns in the singular and even number classes nouns in the plural, they pair up mostly in the same way, and they have a large overlap in the kind of things that can be found in each pair of noun classes. After that, the NCB languages diverge on which noun classes are retained in the language and there may not be a singular/plural pairing. For instance, isiZulu’s ubuntu ‘humanity’ is an abstract concept in noun class 14 for which there is no singular or plural, whereas in Chichewa a noun in noun class 14 may have a plural in noun class 6. The nouns with their noun classes govern a rich system of concordial agreement across sentence constituents, ranging from verb conjugation to modifying adjectives, possessives, and other relational notions.

The verbs have a so-called “slot system” where each slot fulfils a specific function, if used (Khumalo, 2007): there are eight ordered slots, being the pre-initial, initial, post-initial, pre-radical, (verb) radical, pre-final, final, and post-final slot. The pre-initial and post-initial can take tense, aspect, mood and negation, and the pre-final can take tense, aspect, mood and valence change (causative, accusative, reciprocative, and passive). The initial is for the subject concord to conjugate the verb depending on the subject in the sentence and the pre-radical slot is for the object concord. The final slot is for the final vowel (e.g., default /a/ in isiZulu, but /i/ if the verb is negated) and post-final is used for extensions including the wh-questions and locative suffix.

Many of the NCB language are agglutinating and thus have a substantial set of phonological conditioning rules especially for vowel coalescence and vowel elision. For instance, for ‘(located) in the envelope’ in isiZulu, one has to modify invilophu ‘envelope’ with phonologically conditioned locative prefix (e-) and suffix (-ini) to result in emvilophini, and the enumerative ‘and’ na- merges with the successive noun, as in, e.g., (na- + umfana =) nomfana
Table 1: Generalisation of the semantics of the kinds of objects that the nouns in the respective noun classes (NCs) refer to. Examples from isiZulu (1-11, 14, 15), Chichewa (12,13,16-18), Hunde (19), Runyankore (20,21), and Luganda (22,23). (Source: adapted from (Byamugisha et al., 2018).)

| NCs          | Semantics (generalised) | Examples                  |
|--------------|-------------------------|---------------------------|
| 1            | People and kinship      | *umfana* (nc1) ‘boy’      |
| 2            |                         | *abafana* (nc2) ‘boys’    |
| 3            | Plants, nature, some parts of the body | *umuthi* (nc3) ‘tree’     |
| 4            |                         | *inithi* (nc4) ‘trees’    |
| 5            | Fruits, liquids, parts of the body, loan words, paired things | *ikhala* ‘nose’            |
| 6            |                         | *amakhala* ‘noses’        |
| 7            | Inanimate objects       | *isihlalo* ‘chair’         |
| 8            |                         | *izihlalo* ‘chairs’        |
| 9            | Loan words, tools, and animals | *inja* ‘dog’              |
| 10           |                         | *izinja* ‘dogs’            |
| 11           | Long thin stringy objects, languages, inanimate objects | *ucingo* ‘wire’           |
| 12           |                         | *izingcingo* ‘wires’       |
| 13           | Diminutives             | *kagalimoto* ‘small car’   |
| 14           | Abstract concepts       | *ubuhle* ‘beauty’          |
| 15           | Infinitive nouns        | *ukucula* ‘to sing’        |
| 16           | Locative classes        | *pamsika* ‘round the market’ |
| 17           |                         | *kumsika* ‘at the market’ |
| 18           |                         | *mumsika* ‘in the market’ |
| 19           | Diminutives             | *hyündü* ‘a little bit of porridge’ |
| 20           | Augmentative and pejorative | *ogusajja* ‘big ugly man’ |
| 21           |                         | *agasajja* ‘big ugly men’ |
| 22           |                         | *gubwa* ‘mutt’ (pejorative of dog) |
| 23           | Locative class          | *eka* ‘at home’            |

‘and the boy’ and (na- + inja =) *nenja* ‘and the dog’.

They overwhelmingly use Latin script, with some also Arabic script and modern indigenous writing systems. Among the ones that use Latin script, there can be language-specific variations; e.g., isiZulu typically does not have words with an /r/ and Swahili no /q/, Mboshi has an ε variant of /e/ in addition to the /e/, and there are letter combinations to stand in for more consonants, such as a ‘hard b’ and a ‘soft b’ (/bh/ and /b/, respectively), and for ‘click sounds’, such as a nasalised click that may written as /nc/. Many NCB languages are tonal, although this may not be reflected in the orthography (Maddieson and Sands, 2019).

2.2 Categorising NCB languages

There have been multiple attempts at grouping the NCB languages according to various parameters. The one most well-known is based on geographic regions devised by Guthrie (1971), which has been updated in (Moho, 2003) and again informally in 2009 with detailed maps and many references. The system counts from A to S, from top-left in Cameroon to down-right in South Africa. Each zone has groups, indicated by increments of 10 (e.g., A10), where all languages within the group have arbitrarily ordered increments of 1 (e.g., A11 and A12), and possibly further minor increments, such as A111 and A12a. A map overlaid with the NCB languages mentioned in this paper is shown in Fig. 1 (the non-NCB Niger-Congo languages Yoruba and Igbo are located left of A86c, in Nigeria); see also Table 2.

In addition to descriptive linguistics-based overviews, comparisons, and high-level groupings (Güldemann, 2018), some of those research efforts also have a computational component that also may inform potential for bootstrapping language resources. Petzell and Hammarström (2013) com-

---

6The document is available at [https://brill.com/fileasset/downloads_products/35125_Bantu-New-updated-Guthrie-List.pdf](https://brill.com/fileasset/downloads_products/35125_Bantu-New-updated-Guthrie-List.pdf) (last accessed 3 Sep 2022), but does not seem to have been published.
pared 8 languages in Tanzania on 27 morphophonological and morphosyntactic parameters, which showed that the neighbour-joining trees for lexical and for grammatical distance are distinct to the extent that they indicate different language clusters, and, consequently, a merger of the two can lead to different clusters. This feature was also observed in the cladistic analyses of 87 languages by Rexová et al. (2006). They merged 92 lexical and 52 grammatical characters and tried two different classification computations (parsimony with Bremer indices and Bayes), which revealed different trees and putative percentages for similarity. Notably, it also showed that, while the clades largely cluster by the high-level Guthrie zones, this does not hold in all cases (e.g., Pokomo (E71) and Swahili (G40 (G41-43)) are siblings in both trees), and multiple times not so for groups (e.g., Sanga (L35) and Kaonde (L41) are siblings). Similar works exist, varying input and algorithms; e.g., also sound files and an algorithm based on DNA sequence optimisation to avoid orthographic peculiarities in computational comparisons (Whiteley et al., 2019). Such works could provide guidance for devising a bootstrapping strategy for a particular NLP task.

3 Bootstrapping: questions and related work

The notion of bootstrapping resources from one language for another applies to any set of languages. General questions that direct the research that surface include, mainly:

1. Bootstrapping for which NLP task?
   (a) Does the type of task (e.g., machine translation (MT) versus natural language generation (NLG)) and approach (corpus-based, rules-based, class of algorithm) matter?
   (b) If so: how and to what extent? If no: why not?

2. How to measure language similarity, as a proxy for bootstrappability potential?
   (a) Which metrics to use or devise?
   (b) For a given metric, what is the threshold (or fuzzy function) for ‘similar’ and for determining ‘similar enough for bootstrapping anyhow’ and for ‘similar enough for bootstrapping for NLP task X’?
   (c) In which way do the similarity metric and bootstrapping ease relate, if at all?
   (d) Can metrics for distinguishing (separating) languages automatically be flipped and be used for similarity for bootstrapping, and if so, how?

3. Are bootstrapping gains reliably quantifiable beyond a ‘many hours saved’ or ‘any amount of a better BLEU score’ in carrying out NLP task X?

4. To what extent, if at all, does some success of cross-language learning (e.g., using a L1 corpus for training an L2 spell checker) shed any useful light on bootstrappability prospects of L2 from L1?

5. If bootstrapping from L1 to L2 works well, would it work just as well from L2 to L1? Or: what is the ‘fastest’ route from L_i to L_j?

These broad questions are hard to answer based on the available research on NCB languages, but progress is being made.

The literature review was necessarily of an exploratory nature rather than that of, e.g., a meta review, since only limited research on bootstrapping
Table 2: Listing of the languages mentioned in the paper, with Guthrie zone and ISO 639-3 codes for at least some of the languages grouped within it, the NLP task(s) where the language was mentioned (if applicable) and whether it was used in some way of bootstrapping (“Boot.”) efforts. MT: machine translation; NLG: natural language generation; POS: part-of-speech; IR: information retrieval.

| Language       | Guthrie code | ISO 639-3               | Task mentioned                                      | Boot. |
|----------------|--------------|-------------------------|-----------------------------------------------------|-------|
| Chichewa       | N31          | nya                     | IR, similarity                                      | –     |
| Cinyanja       | N31a?        | nya                     | IR, similarity                                      | –     |
| Cisena         | N44          | seh                     | IR, similarity                                      | –     |
| Citonga        | N15          | tog                     | IR, similarity                                      | –     |
| Citumbuka      | N21          | tum                     | IR, similarity                                      | –     |
| Gĩkũũũ         | E51          | kik                     | prefix extraction                                   | +     |
| Hunde          | JD51         | hke                     | –                                                   | –     |
| Kaonde         | L41          | kqn                     | (computational cladistics)                           | –     |
| Kimbundu       | H21          | kmb                     | similarity                                          | –     |
| Kinyarwanda    | JD61         | kin                     | noun class prediction                               | –     |
| Kwangali       | K33          | kwn                     | (computational cladistics)                           | –     |
| Luganda        | JE15         | lug                     | noun class prediction                               | –     |
| isiNdebele (ZW)| S44          | nde                     | morphological analysers                             | +     |
| isiNdebele (ZA)| S408         | nbl                     | pronunciation dictionary                            | +     |
| isiXhosa       | S41          | xho                     | NLG, MT, morphological analysers, pronunciation dictionary, similarity | +     |
| isiZulu        | S42          | zul                     | NLG, morphological analysers, pronunciation dictionary, similarity | +     |
| Mbosho         | C25          | mdw                     | –                                                   | –     |
| Mpiemo         | A86c         | mcx                     | POS tagger                                         | +     |
| Ngoni          | N12          | ngo                     | morphological analysers                             | +     |
| Pokomo         | E71          | pkb                     | (computational cladistics)                           | –     |
| Runyankore     | JE13         | nyn                     | NLG, similarity, (computational cladistics), noun class prediction | +     |
| Sanga          | L35          | sng                     | (computational cladistics)                           | –     |
| Sepedi         | S32          | nso                     | POS tagger, pronunciation dictionary, similarity    | +     |
| Setswana       | S31          | tsn                     | morphological analysers, pronunciation dictionary, similarity | +     |
| Shona          | S10 (S11-15)| sn (twl, mxc, twx, ndc)| MT, similarity                                      | +     |
| siSwati        | S43          | ssw                     | morphological analysers                             | +     |
| Swahili        | G40 (G41-43)| swa, swh (ccl, sta)     | MT, POS tagging, pronunciation dictionary, news item monitoring, similarity, (computational cladistics) | +     |
| Swahili (Congolese)| G40g     | swc                     | MT                                                  | +     |

for NCB languages has been carried out (discussed in Section 4). To cast the net wide, we searched the scientific literature by

- multiple terminology variants: including bootstrapping, bootstrappability, bootstrapping NLP tools, cross-language;
- broad language terms, such as African languages, Bantu languages, Niger-Congo languages;
- specific language names with variant spellings, such as both Swahili and Kiswahili, Zulu and isiZulu; and
- a number of categories of NLP tasks, including machine translation, morphological analyser.

They were combined in multiple combinations of
these dimensions, such as “bootstrapping morphological analyser Swahili”. Of the papers found, the search was then augmented with ‘forward’ searches in time to find other papers that cited them, which were then assessed manually on whether those papers were also about bootstrapping for NCB languages.

The literature we found describe few and largely disparate efforts, which we nudged to attendant topics that may assist with bootstrapping across NCB languages: bootstrapping within a single language and similarity among NCB languages. We grouped them into two approaches: qualitative with estimates on time saved and quantitative with similarity measurements and output assessments. We describe and discuss them in the next two subsections, and subsequently synthesize and discuss them in Section 4.

3.1 Qualitative assessments

Qualitative assessments can be grouped into bootstrapping across language families, cross-language rules- or grammar-based approaches among the NCB family, and within-language bootstrapping. We’ll discuss each in turn.

3.1.1 Experiments across language families

Reusing tools for resource-rich languages for bootstrapping tools for low-resourced languages in another language family is a tempting idea—that one’s NLP tool would be that much generalisable. Among others, Baumann and Pierrehumbert (2014) tried to find isiZulu prefixes based on an English wordlist under the assumption of loanword adoption and Pauw and Wagacha (2007) tried prefix extraction for Gikuya with the Indo-European languages-tailored AutoMorphology. The latter compared AutoMorphology to a new maximum entropy modeling (maxent) algorithm, which was much better suited to the agglutinating languages than AutoMorphology. Cross-family efforts have obtained only limited results; rather, they fed into the drive to develop NLP tools for the NCB languages proper and the notion of learning from other NCB languages that are relatively more closely related. For instance, common spellcheckers for languages such as English are dictionary-based, but it is infeasible to use that approach for highly inflectional agglutinating languages—or: it was easily bested with character 3-grams extracted from a small corpus (Ndaba et al., 2016). For the Europe Media Monitor’s inclusion of Swahili, the bootstrapping amounted to only determining which NLP sub-tasks needed to be carried out, but tools had to be sourced, adapted, and developed as well (in about 3 person-months) (Steinberger et al., 2011), as was the case for word alignment in an English-Swahili parallel corpus (De Pauw et al., 2011). Once De Pauw et al. (2011) had created that word alignment, however, POS tagging for Swahili was bootstrapped from the POS tagged aligned English with 74.8% accuracy.

3.1.2 Rules and grammar bootstrapping

The qualitative experiments on bootstrapping across NCB languages focus on the knowledge-based approach. This was first reported for morphological analysers (Bosch et al., 2008; Pretorius and Bosch, 2009) on similarity among languages in the Nguni language cluster vs disjunctive orthography of the Tswana cluster in South Africa, where the within-cluster was substantially faster for bootstrapping. For isiZulu, it was estimated to cost around 3000 highly skilled person hours, but subsequent analyser development for the three other Nguni languages (isiXhosa, isiSwati and isiNdebele) based on ZulMorph took only about 300 hours in total (Bosch et al., 2008). At the sentence level, bootstrapping natural language generation was experimented with for rules for Runyankore (JE13, spoken in Uganda) from rules for isiZulu (S42, South Africa, about 4800km further south). This was successful, albeit indicating only “a lot easier” (Byamugisha et al., 2016). Their later work further demonstrated that Guthrie zones indeed are not a reliable predictor for bootstrapping advantages (Byamugisha, 2019).

A linguistic analysis across Guthrie zones, in this case for bootstrapping morphological analysers for “Zimbabwe Ndebele” (S44, isiNdebele) and Ngoni (N12, Tanzania) from the isiZulu (S42) Zulmorph tool also concluded that bootstrapping across zone is feasible, albeit easier for the one closer-by (isiNdebele) than geographically further away (Ngoni), due tot influences of more disparate NCB languages on morphemes and noun classes (Bosch, 2012), as is the case for Runyankore cf. isiZulu as well. But, like with Byamugisha et al. (2016)’s outcomes, also for these language comparisons, it appeared that word categories and the order of morphemes agree largely (Bosch, 2012).

In all these experiments, precise measurements beyond rough estimates are inherently difficult to obtain, because the development task for the lan-
guage to bootstrap from has too many variables to make it reliable to devise one final precise value, be this for how to compute grammar similarity or the time saved. The time saved also depends on linguistic knowledge of the developers and quality and recency of the grammar documentation, and the bootstrap values are sensitive to how one translates qualitative input into a numerical value or vector. We illustrate this in the next example.

Example 1 Consider ontology or knowledge graph verbalisation of the all-some axiom pattern \( (C \sqsubseteq \exists R.D) \); the template ‘pattern’ for isiZulu and Runyankore are, respectively:

Zu: [QCall] \([C]\) [SC]\[R\] [D] [RC][QC]dwa
Nyn: [QCall] \([C]\) [SC]\[R\] hakiri \([D]\) [RC]mwe

as generalised from the example and algorithm of Byamugisha et al. (2016), with [QCall] the quantitative concord for the universal quantifier, [SC] the subject concord, [RC] the relative concord, and [QC] the quantitative concord for the existential quantifier. QCall, SC, RC, and QC are clitics governed by the noun class of the respective noun.

The structure of the sentences and the components are mostly the same, with just variation in vocabulary. The only difference is in part of the verbalisation of “\(\exists\)”: a fixed hakiri versus a noun class-dependent [QC], whilst [RC], dwa and mwe are the same elements. A sample calculation for syntax could be a percentage for each element in common (100%) and the ordering similarity computed with a Levenshtein distance for elements or some other measure (5 of the 8 are in the same position, or 62.5%). It would thus score very low on lexical proximity, possibly high depending on which metric is chosen for the grammar, and a merged value based on equal weighting could ever be only at most 50%. That was still enough for bootstrapping advantages.

A high grammar similarity clearly can assist bootstrapping rules-based NLP approaches and tasks, whereas the substantive lexical distance due to entirely different vocabulary will negatively affect any cross-lingual data-driven NLP task.

3.1.3 Bootstrapping within a single language

Although this sense of bootstrapping is not the scope of the paper, it is worth mentioning that it has been done in the comparatively early years of language technologies for South African languages, upon which some of the later techniques build at least in part. An illustrative example concerns data-driven text-to-speech applications with minimal human-in-the-loop intervention, where first a pronunciation dictionary for German was attempted (Davel and Barnard, 2003) that was then extended to experiments with South African languages (Davel and Martirosian, 2009) that evolved into afore-mentioned Qfrency that, in turn, was embedded in the Awezamed app (Marais et al., 2020). They commenced with little data and expanded it through a continuous feedback loop rather than relying on similarities of grapheme-to-phoneme of various subsets of the 11 official South African languages and rather than resource-consuming corpus development. The term ‘bootstrapping’ is used likewise elsewhere for, among others, Sepedi, isiZulu, Swahili, and Mpiemo NLP resources (Prinsloo and Heid, 2005; Spiegler et al., 2010; Hammarström et al., 2008; Schlippe et al., 2014). Each commenced with limited data and then used rules-based and/or statistical approaches with a human-in-the-loop for curation in the iterations. Neither reported on relative gains after each step in the procedure nor did they examine the effect of the input type or size (e.g., varying the amount of seed words or rules) on the quality of the output. Hence, determining a sweet spot for an ideal, or at least minimum, amount of seed data to obtain a certain level of quality remains an open question.

3.2 Quantitative measurements

With the increasing popularity of data-driven approaches, measurements have gained attention, but quantitative measurements are scarce and even then with different foci and mostly on similarity rather than assessments on the bootstrapping process. We’ll discuss both.

3.2.1 Data-based bootstrapping

A ‘small data’ data-based approach informed by grammar was taken by (Byamugisha, 2022) in the task of predicting the noun class of nouns for Runyankore, Luganda, and Kinyarwanda. Sub-word, word-level, and syntax level with nearest neighbour strategies were used, which showed a 10% improvement in noun class prediction in the case of cross-language resource reuse from a Runyankore resource to Luganda. This was attributed to noun class prefix similarities: 15 Luganda noun class prefixes out of 21 are the same as for Runyankore. For Kinyarwanda, which has only 5 exact matches in prefixes with the Runyankore ones, combining the resource had a detrimental effect of about 8%
were often still in the single digit range. Most of the other texts of the other languages), including Swahili, Kimbundu-Mbundu, Runyankore, Shona, and several South African languages. A key insight with just the cumulative frequency distributions of the words in the Universal Declaration of Human Rights was that not all NCB languages are agglutinative and three groups can be identified. There are those that indeed are highly agglutinative, such as Runyankore, Shona, and isiZulu, those that are disjunctive as English, such as Swahili, and those that are highly disjunctive, such as Sepedi—and statistically significantly so. Based on the morphological analysis reported by Pretorius and Bosch (2009) and the comparison table by Kambarami et al. (2021), the latter two clusters may still have a similar enough grammar for bootstrapping compared to starting from scratch, but they mainly will have fewer phonological conditioning rules due to the disjunctive orthography. This is illustrated and elaborated on in the following example.

Example 2 Consider the examples in Table 3, where agglutinating Shona and isiZulu have 4 types and 4 tokens in the examples, and disjunctive Setswana and English have 12 tokens and 6 types. Let’s assume a similarity calculation with the Levenshtein distance: from *ndinomutya* to *ngimsaba* it is already 6 and from *kea mo tsa to* *ngimsaba* it is 7; it is as bad or downhill from there with the other examples up to a distance of 10. An agglutinating *keamoisaba* to *ngimsaba* distance (or, vice versa, the disjunctive *ngi m saba*) reduces it to 5 and *keamothusa* to *ngimisiza* from 10 to 8, which are still not good, but it demonstrates that orthographic conventions matter: a 20-30% improvement in similarity.\(^8\)

---

\(^7\)We did not find any research, but this may be due to the search terms, which zoomed in on bootstrapping.

\(^8\)In contrast to the large lexical distances observed, the respective grammar components are the same: SC-OC-VerbStem, i.e., first the subject concord (verb conjugation) for ‘I’, then object concord for ‘him/her’ (the OC does not...
Conversely, a quick check on amount of agglutination may provide a rough idea of good bootstrapping potential, as supported also by the Runyankore, isiNdebele, and Ngoni resources bootstrapping from isiZulu (Byamugisha et al., 2016; Bosch, 2012). Whether this simple proxy holds generally is yet to be investigated.

The phonological conditioning does have an effect on the ruleset for understanding and generating text. This was examined in the context of similarity measures for verbs in isiZulu and isiXhosa (Mahlaza and Keet, 2019). For the domain of verbs for weather forecasts, the 52 isXhosa and 49 isiZulu rules have 42 rules in common, but it reached only 59.5% overall similarity on an adapted Driver-Kroeber metric as morphosyntactic similarity measure, although 99.5% for the past tense only (Mahlaza and Keet, 2019). In other words: it fluctuates substantially to either not be a reliable indicator or one where the notion of ‘sufficient similarity’ has to have a low threshold value.

Around the same time, Chavula and Suleman (2017) tried a weighted metric that uses the Ordered Weighted Aggregator, on geographically neighbouring languages Chichewa (N31) and Citumbuka (N21), which outperformed Dice on morpheme segmentation. This was then extended for five languages in Malawi (Cinyanja, Citumbuka, Chichewa, Citonga and Cisena) and zooming in on predicting intelligibility to serve reranking of results for better information retrieval (Chavula and Suleman, 2021). Those intelligibility prediction values using Random Forest classifiers achieved the best results with a prediction accuracy up to nearly 0.75 for the largest data set size, for selected topics. Intelligibility entails similarity of vocabulary and grammar and therewith bootstrapping for any NLP task and cross-lingual learning in data-driven NLP. Also here, the meaning of the actual value of prediction accuracy is not fully clear. That is, it leaves questions regarding the grounding, in the sense that 0.75 is good enough for understanding text sufficiently or with difficulty or ease, and whether a few percentage points more or less would have any effect on, e.g., statistical language modelling and resulting performance of, say, data-driven spellcheckers or machine translation.

4 Discussion

Revisiting the typical sort of bootstrapping questions from the beginning of Section 3 in the light of the related work, one has to concede that very few questions can be answered conclusively. There are many different NLP tasks and a wide variety of these tasks have been attempted for one or more NCB languages over the years with both data-driven and rules-based approaches or a combination thereof. And for each of those tasks, limited bootstrapping has been experimented with to a limited extent. We tried to be inclusive in the review of papers on NCB languages, but since there are very many NCB languages known under even more names that are not always similar strings\(^9\) (and many NLP tasks) that are sometimes only mentioned in a paper’s section on data collection, this overview on bootstrapping for NCB languages is likely to be incomplete. Nonetheless, several insights can be distilled from it already.

The rules-based bootstrapping experiments focussing on grammar and morphology demonstrated that many of the key grammar components and their use are similar across NCB languages in different Guthrie zones thousands of kilometres apart, as demonstrated for isiZulu on the one hand and Runyankore and Ngoni on the other. Their distance is visualised in the cladogram-like diagram of Rex-

---

\(^9\)e.g., Hima and Runyankore or Nyanya and Chichewa, rather than only the common prefix addition or skipping like with isiZulu and Zulu or ciShona and Shona.
Figure 2: Section of Rexová et al. (2006)'s cladogram-like binary tree with bootstrapping attempts encircled for the overlap in language inclusion: success of grammar from S42 (isiZulu) to J13 (Runyankore (Hima)), mixed syntax and data from J13 to J15 (Luganda), and limited benefits of enriching S42 data with S10 (Shona) data.

ová et al. (2006) in Fig. 2, suggesting that at least the languages in this sub-branch are promising targets for bootstrapping for grammar-based tasks and approaches. Conversely, the negligible advantage of cross-language learning of adding Shona data for English-isiZulu MT, yet being relatively nearby, is not at all promising. Data augmentation with an arbitrary other African language is even less likely to boost BLEU scores. An illustration that suggests why, was seen in the small example in Example 2 and more careful selection of a neighbouring language likely will yield better results. Other issues (see, e.g., (Hovy and Spruit, 2016)) are yet to be dealt with as well, such as which data is used for bootstrapping—bible translations—bible translations and old texts are not representative for current language use, as shown by (Ndaba et al., 2016).

Clade-based approaches to grouping NCB languages, such as in Fig 2 and mentioned in Section 2.2, may further guide the ‘guessing game’ to better informed experimenting of bootstrapping. For instance, whether it is indeed harder, and if so how much so, to bootstrap an NLP task for, say, Kwangali (K33), which is located in a separate branch from the one shown in Fig 2, from isiZulu as compared to all the current reported isiZulu-to-LanguageX bootstrapping results.

Last, language similarity computed with the various measures as a proxy for bootstrappability did not reveal more than that a low similarity value on lexical proximity may still yield good results for rules-based bootstrapping. This is invariant of the similarity measures that have been used. Better grounding of similarity measures to the quality of executing of a specific NLP task may give more insight into the meaning of the values obtained. The notions of similarity for bootstrapping may avail of the concept of intelligibility for which many methods exist (Gooskens, 2013). While of interest, it was deemed beyond the scope, because it requires the existence of substantial resources.

5 Closing remarks

The review of bootstrapping efforts for Niger-Congo B (‘Bantu’) languages showed that a variety of bootstrapping efforts have been investigated to a very limited extent. The most promising for bootstrapping are rules/grammars, even if the languages spoken are geographically distant. Lexical diversity due to orthographic and vocabulary differences indicated that geographic proximity cannot be relied on for enhancing data-driven approaches to NLP tasks. The results on measuring similarity as proxy for estimating bootstrapping leaves ample room for future research, both on the metrics and the grounding thereof for reliable estimation. Also, more cross-fertilisation between linguistic theory and comparative linguistics and computational experiments on bootstrapping may have an added value for guiding what to bootstrap from what.

Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 120852).

References

David Bamutura, Peter Ljunglöf, and Peter Nebende. 2020. Towards computational resource grammars for runyankore and rukiga. In Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16,
Peter Baumann and Janet Pierrerhumbert. 2014. Using resource-rich languages to improve morphological analysis of under-resourced languages. In In Proc. of 9th International Conference on Language Resources and Evaluation (LREC’14). European Language Resources Association (ELRA). Reykjavik, Iceland, May 26-31, 2014.

Sonja E. Bosch, Laurette Pretorius, and Axel Fleisch. 2008. Experimental bootstrapping of morphological analysers for nguni languages. Nordic Journal of African Studies, 17(2):66–88.

Sonja E. Bosch. 2012. Bootstrapping the development of morphological analysers for "dispersed" nguni languages - a linguistic investigation. In Hugues SteveNdiga-Koumba-Binza and Sonja E. Bosch, editors, Language Science and Language Technology in Africa: Festschrift for Justus C. Roux, chapter 8, pages 123–144. SUN Press.

Sonja Bosch and Laurette Pretorius. 2017. A computational approach to zulu verb morphology within the context of lexical semantics. Lexikos, 27:152–182.

Joan Byamugisha. 2019. Ontology Verbalization in Agglutinating Bantu Languages: A Study of Runyankore and Its Generalizability. Phd thesis, Department of Computer Science, University of Cape Town, South Africa.

Joan Byamugisha. 2020. Generating varied training corpora in Runyankore using a combined semantic and syntactic, pattern-grammar-based approach. In Proceedings of the 29th International Conference on Computational Linguistics, pages 4350–4359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joan Byamugisha. 2022. Noun class disambiguation in Runyankore and related languages. In Proceedings of the 29th International Conference on Computational Linguistics, pages 4350–4359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2018. Pluralizing nouns across agglutinating Bantu languages. In 27th International Conference on Computational Linguistics (COLING’18), pages 2633–2643. ACL. 20-26 August, 2018, Santa Fe, New Mexico, USA.

Catherine Chavula and Hussein Suleman. 2017. Morphological cluster induction of bantu words using a weighted similarity measure. In Proceedings of the South African Institute of Computer Scientists and Information Technologists, SAICSIT 2017. pages 6:1–6:9. ACM.

Catherine Chavula and Hussein Suleman. 2021. Ranking by language similarity for resource scarce southern bantu languages. In ICTIR ’21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, pages 137–147. ACM.

Marelie Davel and Etienne Barnard. 2003. Bootstrapping in language resource generation. In PRASA Conference 2003, pages 97–100.

Marelie Davel and Olga Martirosian. 2009. Pronunciation dictionary development in resource-scarce environments. In 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), pages 2851–2854. ISCA. Brighton, UK, 6-10 September, 2009.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2011. Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources & Evaluation, 45:331–344.

C. Engelbrecht, N. C. Shangase, S. J. Majeleke, S. Z. Mthembu, and Z. M. Zondi. 2010. Isizulu terminology development in nursing and midwifery. Alternation, 16(1):249–272.

Charlotte Gooskens. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In Robert Bayley, Richard Cameron, and Ceil Lucas, editors, The Oxford Handbook of Sociolinguistics, chapter 10, page 195?213. Oxford University Press.

Tom Güldemann. 2018. 2. historical linguistics and genealogical language classification in africa. In Tom Güldemann, editor, The Languages and Linguistics of Africa, pages 58–444. De Gruyter Mouton, Berlin, Boston.

Malcolm Guthrie. 1971. Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. (An outline of Bantu prehistory: an inventory of the Bantu languages). v. 1-2. Gregg.

Harald Hammarström. 2018. A survey of african languages. In Tom Güldemann, editor, The Languages and Linguistics of Africa, pages 1–57. De Gruyter Mouton, Berlin, Boston.
Harald Hammarström, Christina Thornell, Malin Petzell, and Torbjörn Westerlund. 2008. Bootstrapping language description: the case of Mpiemo (Bantu A, Central African Republic). In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), European Language Resources Association (ELRA).

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 591–598, Berlin, Germany. Association for Computational Linguistics.

N. Hussey. 2012/2013. The language barrier: The overlooked challenge to equitable health care. SAHR, pages 189–195.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. Association for Computational Linguistics.

Farayi Kambarami, Scott McLachlan, Bojan Bozic, Kudakwashe Dube, and Herbert Chimhundu. 2021. Computational modeling of agglutinative languages: The challenge for southern Bantu languages. Arusha Working Papers in African Linguistics, 3(1):52–81.

C. Maria Keet. 2016. An assessment of orthographic similarity measures for several african languages. Technical Report Arxiv.org 1608.03065, University of Cape Town.

C. Maria Keet. 2021. Natural language generation requirements for social robots in subsaharan africa. In IST-Africa 2021, page 8p. IIMC International Information Management Corporation. 9-11 May, online.

C. Maria Keet and Langa Khumalo. 2017. Toward a knowledge-to-text controlled natural language of isiZulu. Language Resources and Evaluation, 51(1):131–157.

Langa Khumalo. 2007. An analysis of the Ndebele passive construction. PhD thesis, University of Oslo, Norway.

Langa Khumalo. 2015. Advances in developing corpora in African languages. Kauwaia, 1(2):21–30.

Langa Khumalo. 2017. Intellectualization through terminology development. Lexicos, 27:252–264.

Ian Maddieson and Bonny Sands. 2019. The sounds of the bantu languages. In Mark Van de Velde, Koen Bostoen, Derek Nurse, and Gérard Philipsson, editors, The Bantu Languages, 2 edition, chapter 2, pages 79–127. Routledge.

Zola Mahlaza and C. Maria Keet. 2019. A method for measuring verb similarity for two closely related languages with application to zulu and xhosa. South African Computer Journal, 31(2):34–56.

Zola Mahlaza and C. Maria Keet. 2020. OWLSIZ: An isiZulu CNL for structured knowledge validation. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), pages 15–25, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Laurette Marais, Johannes A. Louw, Jaco Badenhorst, Karen Calteaux, Ilana Wilken, Nina van Niekerk, and Glenn Stein. 2020. Awezamed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In IEEE 23rd International Conference on Information Fusion (FUSION’20), pages 1–8. IEEE.

Stuart Mesham, Luc Hayward, Jared Shapiro, and Jan Buys. 2021. Low-resource language modelling of south african languages. CoRR, abs/2104.00772.

Jouni Moho. 2003. A classification of the bantu languages: an update of guthrie’s referential system. In Derek Nurse and Gérard Philipsson, editors, The Bantu languages, chapter 31, pages 639–651. Routledge.

Balone Ndaba, Hussein Suleman, C. Maria Keet, and Langa Khumalo. 2016. The effects of a corpus on isizulu spellcheckers based on n-grams. In IST-Africa 2016. IIMC International Information Management Corporation. 11-13 May, 2016, Durban, South Africa.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsudeen Muhammad, Salomon Kabongo Kabenamuul, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Oreevahgene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogreui, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kiko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefulu, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefulu, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2144–2160, Online. Association for Computational Linguistics.

Evander Nyoni and Bruce A Bassett. 2021. Low-resource neural machine translation for southern african languages. arXiv preprint arXiv:2104.00366.
Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending “small data” pretraining approaches to sequence-to-sequence models. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 126–135, Hybrid. Association for Computational Linguistics.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, and Grace Tang. 2021. Congolese swahili machine translation for humanitarian response. CoRR, abs/2103.10734.

Guy De Pauw and Peter Waiganjo Wagacha. 2007. Bootstrapping morphological analysis of gikuyu using unsupervised maximum entropy learning. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH’07), pages 1517–1520. ISCA.

Malin Petzell and Harald Hammarström. 2013. Grammatical and lexical comparison of the Greater Ruvu Bantu languages. Nordic Journal of African Studies, 22(3):129–157.

Laurette Pretorius and Sonja E. Bosch. 2003. Enabling computer interaction in the indigenous languages of South Africa: The central role of computational morphology. ACM Interactions, March + April:56.

Laurette Pretorius and Sonja E. Bosch. 2009. Exploiting cross-linguistic similarities in zulu and xhosa computational morphology: Facing the challenge of a disjunctive orthography. In Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages - AFLaT 2009, pages 96–103.

Laurette Pretorius, Laurette Marais, and Ansu Berg. 2017. A GF miniature resource grammar for tswanaga: modelling the proper verb. Language Resources and Evaluation, 51(1):159–189.

Danie J. Prinsloo and Ulrich Heid. 2005. Creating word class tagged corpora for northern sotho by linguistically informed bootstrapping. In Proceedings of the Lesser Used Languages and Computer Linguistics Conference (LULCL’05), pages 97–113. Eurac.

Katerina Rexová, Yvonne Bastin, and Daniel Frynta. 2006. Cladistic analysis of bantu languages: a new tree based on combined lexical and grammatical data. Naturwissenschaften, 93(3):189–194.

Tim Schlippe, Matthias Merz, and Tanja Schultz. 2014. "methods for efficient semi-automatic pronunciation dictionary bootstrapping. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH’14), pages 2867–2871. ISCA. 14 - 18 September 2014, Singapore.

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach. 2010. Ukwabelana – an open-source morphological zulu corpus. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING’10), pages 1020–1028. Association for Computational Linguistics. Beijing.

Ralf Steinberger, Sylvia Ombuya, Mijail Kabadjov, Bruno Pouliquen, Leo Della Rocca, Jenya Belyaeva, Monica de Paola, Camelia Ignat, and Erik van der Goot. 2011. Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. Language Resources & Evaluation, 45:311–330.

Peter M. Whiteley, Ming Xue, and Ward C. Wheeler. 2019. Revising the bantu tree. Cladistics, 35(3):329–348.