Proceedings of the Language Technologies for All (LT4All) , pages 97–100
Paris, UNESCO Headquarters, 5-6 December, 2019.

Preserving and Developing Indigenous Languages in the South African Context

Justus C Roux, Stellenbosch University, Stellenbosch, South Africa, jcr@sun.ac.za
Sonja E Bosch University of South Africa, Pretoria, South Africa, boschse@unisa.ac.za

Abstract

This article focuses on the development of infrastructures in the South African context that are dedicated towards the preservation and development of the eleven official languages of the country. It indicates to what extent the aims of this UNESCO conference are already being met by local activities. Specific attention is also paid to the role of language in the digital age.

Keywords: Infrastructures, African languages, digital resources

Abstrak

Hierdie artikel fokus op die ontwikkeling van infrastrukture binne die Suid-Arikaanse konteks wat toegespits is op die ontwikkeling en preservering van die elf amptelike tale van die land. Dit verwys spesifiek na die doelstellings van hierdie kongres van UNESCO en toon aan hoedanig plaaslike aktiwiteite alreeds vordering gemaak het. Spesifieke aandag word ook gegee aan die rol van taal in die digitale era.

1. Introduction

Given the main aims and objectives[1] of this conference, i.e.

(i) to identify recommendations on how to harness technology for the preservation, support and promotion of languages including lesser-used and Indigenous languages,
(ii) to provide access to information and knowledge to all language users and facilitate their inclusion and participation in building sustainable knowledge societies,
(iii) to promote the human rights and fundamental freedoms of all language users to access and create information and knowledge in language they best understand

this paper focuses on a relatively recent development in the South African context that implicitly addresses all of the above mentioned aims and objectives.

2. Background

South Africa has a population of 58,78 million[2] and eleven official languages that are used to a more or lesser extent across various regions of the country. These languages comprise four Nguni languages (isiZulu, isiXhosa, Siswati and isiNdebele), three Sotho languages (Setswana, Sesotho sa Leboa and Sesotho), Tshivenda, Xitsonga, Afrikaans and English. These African languages as well as Afrikaans are considered to be resource scarce languages (Moors et al., 2018) in a domain of digital communication.

At the turn of the century several academics lobbied national government to take the lead in setting up an infrastructure that could support the development of all official languages to keep these languages abreast of developments in the ICT and Human Language Technology (HLT) domains. This eventually led to the establishment of a Ministerial Advisory Panel in 2002 tasked to develop a strategic plan for the implementation of HLT in South Africa. In 2006 the South African Department of Arts and Culture established an HLT Unit responsible for driving the new HLT strategy which supported the research community with funds for collecting reusable resources and for developing appropriate NLP applications. One of the main recommendations of the Advisory Panel related to setting up a central repository to ensure the availability of reusable digital language resources for all official languages of the country. This recommendation eventually became a reality with the establishment of a Resource Management Agency (RMA) in 2012, amidst a growing number of research and development projects in HLT conducted by academics, see Roux and Ndinga-Koumba-Binza (2019). Through commissioned and own projects, the RMA rendered impressive results in the acquisition, enhancement and distribution of (South African) language resources and software tools. Within the first two years it had 258 registered users of which 157 were from South Africa, seven from five African countries, and 94 from 16 other countries worldwide with a total resource download at that point in time of 1 141. These resources and tools found their way into various research and development projects worldwide.

The RMA was eventually merged with the new South African Centre for Digital Language Resources (SADiLaR)[3] in 2017. The establishment of SADiLaR may

[1] https://en.iyil2019.org/events/lt4all-international-conference-on-language-technologies-for-all-enabling-linguistic-diversity-and-multilingualism-worldwide/ (accessed 13.11.2019)

[2] http://www.statssa.gov.za/ mid-year estimate 2019 (accessed 13.11.2019)

[3] http://www.sadilar.org (accessed 13.11.2019)

be regarded as fruition of numerous activities of researchers over a period of at least 15 to 20 years, eventually providing a long-term infrastructure in support of research and development. SADiLaR is a national entity sponsored by the Department of Science and Innovation (DSI) of the South African government, where it is a member of a new South African Research Infrastructure Roadmap (SARIR). This research and development entity comprises a Hub (hosted by the North-West University in Potchefstroom) linked to different Nodes such as the Department of African Languages at the University of Pretoria, the Department of African Languages at the University of South Africa (UNISA), the Centre for TexT Technology (CTexT) at the North-West University, the Council for Scientific and Industrial Research (CSIR) in Pretoria, and a consortium of universities (ICELDA). SADiLaR furthermore links up to similar international centres: currently it is a CLARIN C-centre and intends to apply for CLARIN B-centre status given the growth in activities. Furthermore, CLARIN, ELRA and the Linguistic Data Consortium (LDC) are all represented on the Scientific Advisory Board of SADiLaR, rendering valuable advice to the new entity.

Although the initial focus of SADiLaR is on the official languages of South Africa, it aspires to be of support to all major languages spoken in the sub-Saharan region. Researchers from these areas are invited to contact SADiLaR in this regard.

The main point to be made in this section relates to the establishment of appropriate infrastructures and calls for engagement with cultural organisations and governmental structures emphasising the shared responsibility towards the languages spoken in the particular country. It is necessary to keep as many indigenous languages alive in an ever-growing technological era, and to enable all speakers of indigenous languages, be they fully literate, semi-literate or even illiterate, to use their own languages for communicating both with other humans and with machines.

SADiLaR currently runs two programmes:

(i) A **digitisation programme**, which entails the systematic creation of relevant digital text, speech and multi-modal resources related to all official languages of South Africa. The development of appropriate natural language processing software tools for research and development purposes are included as part of the digitisation programme.

(ii) A **Digital Humanities programme**, which facilitates the building of research capacity by promoting and supporting the use of digital data and innovative methodological approaches within the Humanities and Social Sciences

The next three sections below will describe some of the activities of SADiLaR implicitly supporting the aims of this LT4All conference sponsored by UNESCO.

3. Harnessing technology

One of the prerequisites of 'harnessing' specific technologies relates to the availability of contents. In attempting to implement machine translation (MT) technologies it follows that ample digital text data of at least two languages should be available, firstly to train the system and then to run the translation engine, converting language A to language B. Similarly, the application of automatic speech recognition (ASR) technologies implies availability of different kinds of digitised speech depending on whose speech needs to be recognised. One of SADiLaR's aims as an enabling entity therefore is to accrue representative sets of digital text and speech corpora of the official languages of the country. Representative refers to language data acquired from as many different sources as possible. In the case of text this will imply newspaper items, literary works from different genres, instruction manuals, advertisements, formal government documents etc. In the case of speech, the data collected could be male or female speech, speech of young and old speakers, or speech regardless of age and gender, or the environment where speech is uttered, on a street corner with background noise, or in a motor vehicle listening to a GPS navigator.

The resource index[4] of SADiLaR already boasts a wide arrange of annotated text and speech resources as well as software tools which are freely available to researchers world-wide. Given these resources, and those still to be developed in effect function as living archives representing language types at different points in history and hence preserving these languages. At the same time the node at UNISA is highly involved in developing African Wordnets (AWN), an example of a multilingual knowledge resource covering all nine official African languages of South Africa. The open source nature of many wordnets (as is the case of the AWN as well) results in applications in different fields of research, making it "an ideal tool for disambiguation of meaning, semantic tagging and information retrieval" according to Morato et al. (2004:270). Concerning language learning applications, the use of wordnets to automatically generate vocabulary tests for second language acquisition has been reported on by Susanti et al. (2015). The usefulness of the current isiZulu Wordnet in a language learning application is also investigated by Bosch and Griesel (2018) who demonstrate how the unique sense identification features of wordnets can be incorporated into a language learning system thereby improving user interaction.

One of the nodes of SADiLaR, the Centre for Text Technology (CTexT) has made significant strides in developing machine translation technology for interactive translation between various official languages of South Africa. Software developed for translators includes the following tools[5] :

(i) "Integrated Translation Environment (ITE): The ITE is computer software that assists translators with their translation process, making use of translation

[4] https://repo.sadilar.org/handle/20.500.12185/1 (accessed 14.11.2019)

[5] http://humanities.nwu.ac.za/ctext/autshumato (accessed 20.11.2019)

memories and glossaries. It can be used to translate between any two natural language pairs in any of the South African languages."

(ii) "Translation Memory and Glossary Integration System (TMG): The TMG is a crowd-sourced platform from which translation resources can be provided as well as obtained. The sharing of translation resources between various translation units and freelance translators can ensure improved consistency and increase productivity throughout translation projects. These in turn can provide more access to information for everyone in their native languages. Users can rate and comment on resources, in order to give others an indication of the quality of a specific resource."

(iii) "Autshumato Terminology Management System (TMS): TMS helps with the development of terminology databases which contain terminology from different languages."

(iv) "Other resources that were developed as part of the Autshumato project include alignment software, a PDF extractor and a text anonymiser to safeguard privacy when using Autshumato ITE."

These cursory examples demonstrate how different technologies have been harnessed and applied to local languages by SADiLaR and two of its Nodes.

4. Access to information and knowledge

In its aim to provide access to information and knowledge, SADiLaR is currently involved in co-operation of the Dutch Language Union developing online Language Portal for two African languages, Setswana and isiXhosa. This two-year pilot project links up to similar portals developed for Dutch, Frisian and Afrikaans.[6] These two portals provide (in certain instances) new information through:

(i) a Grammar portal, comprising sub-domains such as Phonetics/Phonology, Morphology and Syntax,
(ii) a Dictionary portal,
(iii) a Corpus portal and
(iv) an Advice portal and forum.

This Advice portal and forum provides an infrastructure for "… language learners, language practitioners of various educational levels access to experts in the language to answer questions related to meaning, terminology and standardized spelling and grammar."[7] The web obviously provides multimodal content to be displayed, such as texts, speech and video which all contribute to new learning experiences.

5. Create information and knowledge

The nature of research in the Humanities and Social Sciences (HSS) has globally undergone a major paradigm shift over the last two decades due to:

(i) advances in the domain of Information Communication Technologies (ICT), and
(ii) increased access to digital resources.

This has given rise to an ever-expanding interdisciplinary domain of research and development, referred to as *Digital Humanities* (DH).

Given the development of large multilingual corpora, SADiLaR operates in close co-operation with the Digital Humanities Association of Southern Africa (DHASA).[8] These corpora can be searched by means of a wide range of applicable software, and as has been proven elsewhere, this more than often provides specific new knowledge. The strategic importance of SADiLaR became even more relevant as new methodological approaches toward research and development in the domain of Humanities and Social Sciences are posing new challenges to researchers. It was therefore strategically important to assist researchers not only in getting access to large corpora of authentic digital data and applicable software tools, but also to acquire skills related to the use of such data in order to render high quality research outputs nationally and internationally. This was furthermore a deliberate attempt to incubate the field of Digital Humanities in the South African context with benefits to society, academia, industry and government.

The close co-operation between SADiLaR and DHASA has led to a wide range of training programs such as for intance, three GROBID Dictionary Workshops[9] of SADiLaR across the country during 2018 related to "a machine learning infrastructure for creating structured lexicographic data from digitised dictionaries."

Given the growth of DH, SADiLaR has also become the academic home of the first professor in Digital Humanities in South Africa in 2019.

6. Conclusion

One of the first projects of SADiLaR was conducted by its node at the Council for Scientific and Industrial Research (CSIR) focusing on a national audit of HLT activities in South Africa (Moors et al., 2018:558). This audit was a follow up of a previous audit in 2009. One of the main findings was the following:

"Based on the comparisons between datasets and calculating the increase in resources, we were able to determine that there is an increase in resource availability for most South African languages. However, languages such as Xitsonga, Tshivenda, Sesotho, siSwati and isiNdebele still remain under-resourced. We were further able to deduce that more text than speech resources are currently available in South Africa. In addition to the comparison between resource types, we also determined the maturity and accessibility of the resources in all official languages in South Africa."

It appears that activities related to the preservation, and the development of indigenous languages of South Africa have

[6] Cf. www.taalportaal.org (accessed 14.07.2019)

[7] Quoted from original Project proposal: *Toward language portals for South African languages*.

[8] www.digitalhumanities.org.za (accessed 17.11.2019)

[9] http://digitalhumanities.org.za/index.php/dhasa-news (accessed 14.11.2019)

shown remarkable progress over the last decade. Implicitly supporting this view, Kaschula (2019:619) is of the opinion that "Technology is used here to capture, archive, and disseminate literary work in African languages (…). It is envisaged and suggested in this chapter that in future HLT will become the cornerstone of intellectualization of African languages.'

7. Bibliographical References

Abdullah, N. & Ibrahim, R., 2015, 'Managing information by utilizing WordNet as the database for semantic search engine', International Journal of Software Engineering and Its Applications 9(5), 193–204. https://doi.org/10.14257/ ijseia.2015.9.5.19

Bosch, Sonja and Griesel, Marissa. African Wordnet: facilitating language learning in African Languages. In: Francis Bond, Takayuki Kuribayashi, Christiane Fellbaum, Piek Vossen (eds) Proceedings of the 9th Global WordNet Conference (GWC 2018), 8-12 January, 2018, Nanyang Technological University (NTU), Singapore. Pp. 309-316.

Kaschula, R.H. and Nkomo, D. (2019). Intellectualization of African languages: Past, Present and Future, pages 601-622. The Cambridge Handbook of African Linguistics. H.E. Wolff (ed.) Cambridge. United Kingdom.

Moors, C., Calteaux K., Wilken, I and Gumede, T. 2018. Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. ACM International Conference Proceeding Series, pp. 296–304. Available from https://doi.org/10.1145/3278681.3278716 (Accessed on 2019-03-14)

Morato, J., Marzal, M.A., Lloréns, J. and Moreiro, J. 2004. WordNet Applications. In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): Global Wordnet Conference Proceedings 2004, pp. 270–278. Masaryk University, Brno.

Roux, J.C. and Ndinga-Koumba-Binza, H.S. (2019). African Languages and Human Language Technologies, pages 623-644. *The Cambridge Handbook of African Linguistics*. H.E. Wolff (ed.) Cambridge. United Kingdom.

Susanti, Y., Iida, R. & Tokunaga, T., 2015, 'Automatic generation of English vocabulary tests', in M. Helfert (ed.), Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015), Lisbon, Portugal, 23–25 May 2015, pp. 77–78.