An overview of Sesotho BLARK content

Sibeko, Johannes
Nelson Mandela University
johannes.sibeko@mandela.ac.za

Setaka, Mmasibidi
South African Centre for Digital Language Resources
mmasibidi.setaka@nwu.ac.za

Abstract

This article overviews digital language resources available for Sesotho, an official language of South Africa. The South African Center for Digital Language Resources (SADiLaR) repository is used as a reference as it is the official host of various language resources for South African languages. A total of 18 written resources are identified from the repository, and a further 16 spoken resources are identified. Finally, a total of 45 applications and modules were identified. Findings indicate that the majority of applications and modules available for Sesotho are in fact general resources aimed at all eleven official South African languages. Furthermore, the available resources indicate an inclination to the development of entry level, basic language resources and an absence of middle and higher resources with functionalities such as semantic analyses for written resources and prosody prediction for spoken resources. The study is hindered by the dearth of resource specific evaluations and related research and exacerbated by the absence of some of the resources on the repository.

Keywords: Sesotho, BLARKs, Written resources, Spoken Resources, Digital language resources

1 Introduction

There is a growing interest in Human Language Technologies (HLTs) for low-resourced languages (LRLs) (Strassel & Tracey 2016). Accordingly, a number of HLT audits have been conducted on South African official languages (Grover et al. 2010, 2011, Moors, Wilken, Calteaux & Gumede 2018, Moors, Wilken, Gumede & Calteaux 2018). The language audits are aimed at two objectives that are (i) determining resources that need to be developed, and (ii) opportunities for multidisciplinary research. South Africa currently recognizes eleven official languages, namely, Afrikaans, English, isiZulu, isiXhosa, Siswati, Xitsonga, Tshivenda, isiNdebele, Setswana, Sepedi and Sesotho. In this article, we pay special attention to digital language resources available for Sesotho, a Bantu language that forms part of the bigger Sotho-Tswana group with Sepedi and Setswana (Riep 2013, Van Heerden et al. 2010, Nkolola-Wakumelol et al. 2012, Mojela 2016). Additionally, Sesotho is one of the official languages in Lesotho, and an officially recognised language in Zimbabwe (Ndlovu 2011, 2013, Kadenge & Mugari 2015, Wissing & Roux 2017). Sesotho has developed as both a spoken and written language (Moeketsi 2014, Koai & Fredericks 2019), and is used in a variety of domains.

Of the eleven official languages of South Africa, two languages, namely: Afrikaans and South African English, have the most digital language resources followed by Setswana, Sepedi, isiZulu, isiXhosa and Sesotho in no particular order (Moors, Wilken, Calteaux & Gumede 2018, Moors, Wilken, Gumede & Calteaux 2018). This reality is propelled by the lack of investment into the remaining nine indigenous languages. This image was illustrated at the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (Pretorius et al. 2014). Nonetheless, although there are digital language resources in Sesotho, the language remains marginalised and under-resourced as it is yet to have enough high quality supervised data (Koai & Fredericks 2019, Magueresse et al. 2020). As a result, multiple studies identify Sesotho as a low-resourced language (Mahloane & Trausan-Matu 2015, Chiguvare & Cleghorn 2021, Hanslo 2021, Sibeko & Van Zaanen 2022*a*). Moors, Wilken, Gumede & Calteaux (2018) go as far as identifying it as a severely under-resourced language. That is, although it is taught as a home language and an additional language in both South Africa and Lesotho, it persists to be less studied for computerization, and thus remains less privileged and of low density (Cieri et al. 2016, Magueresse et al. 2020).

Although there have been language audits of all eleven official languages of South Africa, as far as we could ascertain for this article, no study has paid in depth attention to Sesotho digital language resources. Ensuingly, this article seeks to describe basic digital language resources, modules, and applications used in written and speech applications as a way of inventorising Sesotho digital language resources.

2 Background

Natural Language Processing (NLP) has been a very active area of research (Drake 2003). It is aimed at both automatically analysing and representing human language using computational technologies (Cambria & White 2014, Kang et al. 2020). These computational techniques are based on both theory and available technologies that enable them to learn, from existing human language content, understand how the human language is produced, and produce human language content independent of humans (Liddy 2001, Hirschberg & Manning 2015). NLP research gathers knowledge on how humans use language and in turn develop applications to enable computers to simulate and handle natural languages (Chowdhury 2003). As such, NLP is a discipline within Artificial Intelligence (AI) and linguistics since it is characterized by human-like language processing capabilities (Drake 2003, Nadkarni et al. 2011). Frequent applications of NLP include (i) information retrieval, (ii) information extraction, (iii) question answering, (iv) summarization, (v) machine translation, and (vi) dialogue systems (Drake 2003). Some models cater for more than one language (Castellucci et al. 2021).

Carrying out NLP tasks relies on the availability of HLT resources which are developed using digital language resources. Unfortunately, none of the South African official languages have official Basic Language Resource Kits (BLARKs) as conceptualised in Krauwer (2003). Generally, a BLARK defines the minimal required resources for performing pre-competitive research in both spoken and written language for a specific language (Arppe et al. 2010). BLARKs are categorised into three, namely, definition, specification, and content (Maegaard et al. 2006).

First, the definition category indicates what should be regarded as basic in the given language. For instance, Arppe et al. (2016) present a matrix that indicates the resources needed for Plains Cree (an indigenous language of central Canada) together with their levels of importance.

Second, the specification category prescribes the quantities of the resources defined such as the Arabic basic resource specification by Maegard et al. (2006) that lists the quantities of resources needed for Arabic. Finally, the content category describes what already exists. One example of this is a survey of applications available for Swedish (Elenius et al. 2008).

In this article, we limit our discussion of basic digital language resources to BLARK content. That is, we inventorize written and spoken digital language resources available in and for Sesotho. Although the definition of BLARK content is language-independent, the resources identified and the level of importance allotted to the resources are specific to Sesotho (Krauwer 2003, Arppe et al. 2010).

Amongst others, industry and education institutions benefit from the availability of BLARKs (Maegaard et al. 2005). In other words, since the BLARK can identify both what already exists and what should be developed, industrial and academic developers can earmark needed resources. Furthermore, BLARK content makes it easier for researchers and educators to unearth paid and freely accessible resources at their disposal. Even so, there are two main issues with BLARKS. First, availability is not a binary distinction in that some existing resources may still be inaccessible due to financial or copyright issues (Krauwer 2003). Second, the multiplicity of resources may not be equated to usability, quality output, or user-friendliness.

When carrying out BLARK content, three BLARK aspects are evaluated, namely, (i) data, (ii) modules, and (iii) applications (Arppe et al. 2010, Krauwer 2003). For instance, in their discussion, Sibeko & Van Zaanen (2022*a*) identify their target application, that is, a readability analysis application. However, they indicated that some modules such as a syllabification system needed to be developed. For the development of the syllabification system, they created a corpus of syllabification annotated corpora. In this way, the data enabled the development of the module which will then be incorporated into their application.

3 Methodology

It is recommended that the South African Centre for Digital Language Resources (SADiLaR) purposefully manages all publications of language resources in South African languages (Moors, Wilken, Gumede & Calteaux 2018). SADiLaR is supported and funded by the South African National Department of Science and Innovation (Sefara et al. 2021). SADiLaR supports development and innovation in the official languages of South Africa. According to Wilken et al. (2018), SADiLaR also aims to facilitate access to digital data and software applications. To this end, it has a publicly accessible repository [https://repo.sadilar.org].

The BLARK content presented in this article is based on digital language resources indexed in the afore mentioned repository. In the first stage of our research, we used the search functionality on SADiLaR's website to extract the indexed language resources. A very broad search query: "Sesotho" was used to yield a total of 123 resources. A summary of these results is presented in Table 1. In the second stage, we searched google scholar for literature related to the digital language resources identified from SADiLaR's repository. Our analysis includes both general language resources that were developed for other languages and can be applied to Sesotho, and resources that were specifically developed for Sesotho. We do this since overviews of this kind ought not be limited to monolingual resources (Arppe et al. 2010). Search results were independently evaluated by the researchers and the results were comparatively analysed.

4 Findings

As indicated above, we limit our discussion to BLARK content as defined by Maegaard et al. (2006). We discuss written resources, speech-based resources, modules, and applications available to and for Sesotho.

4.1 Written Resources

4.1.1 Monolingual lexicon

Few written monolingual corpora were identified. One, the National Centre for Human Language Technology (NCHLT) produced four data sets focused on plain corpora, annotated corpora (Eiselen & Puttkamer 2014), phrase chunks (Eiselen 2016*b*), and named entity (Eiselen 2016*a*). Two, a customised government domain specific dictionary was identified (Bosch & Griesel 2017). Furthermore, a genre classification corpus for Afrikaans, Sepedi, Setswana, isiXhosa, isiZulu, and Sesotho was also identified. The corpus is composed of poetry, advertisements, informational pamphlets, instructions, news, official texts like policies, and speech texts (Snyman et al. 2011). Finally, a syllable annotated word list was identified. It contains one word and its corresponding syllabified versions on each line (Sibeko & Van Zaanen 2022*b*). The word list contains a set of 1355 entries extracted from an existing bilingual Sesotho-English dictionary (Chitja 2010).

Table 1: Search Query Results

Results	Sum
Total results	123
Modules and applications	48
Spoken corpora	16
Written corpora	19
Relevant results	83

Table 2: Translated word lists available in Sesotho

word list	Size
Election	559
Parliamentary jargon	502
HIV/AIDS	586
Arts and Culture for the intermediate phase	550
Mathematics	984
Natural Sciences and Technology list	2756
Information and Communication Technology	132
Life Orientation for the intermediate phase	1628
Soccer	297
Gender Terminology List	446

4.1.2 Translated word lists

Many language pairs lack enough parallel texts (Koehn & Knight 2002). This is seemingly the same case in South African indigenous languages. Even so, there are few translations that use English as a pivot language. We identified a total of nine such word lists that are translated from English to the other official languages of South Africa. The word lists are presented in Table 2.

The word lists were commissioned by the National Language Services under South Africa's Department of Sport, Arts and Culture. The word lists present singular word translations on different subject matters such as politics, education and sports. Language and translation experts in different official languages of South Africa gather in a workshop setting and words are translated in groups. Group quality assurance workshops are then held to ensure quality translations.

Two written dictionary data sets were identified. One, the Bukantswe Sesotho-English bilingual dictionary word list contains a total of 10085 bilingual segments in Sesotho and English. Each line presents the Sesotho term followed by the English translation and the relevant part of speech where available. Two, the Sesotho custom dictionary for government domain is a word list that contains two types of words, namely, those that are exclusive to the government domain and those that do not follow official Sesotho orthographic conventions.

We also identified the Mburisano Covid-19 multilingual corpus that contains screening and triage vocabulary of Covid-19 related multilingual corpora in all official languages. The corpus was developed in response to the Covid-19 pandemic as an attempt to ensure access to information for people of different linguistic repertoires. English is used as the source text for all the other official languages. Unfortunately, the corpus is not clearly marked for specific languages. Although this resource is called a corpus, it contains a translated word list.

4.1.3 Machine Translated corpora

We identified three Machine Translation (MT) corpora. Two of the corpora were produced as part of the Autshumato MT Translation Memory (TM) project. Both the multilingual word and phrase translations and the MT evaluation set contain aligned translations from English to the other ten official languages. The third MT corpus, the Centre for Text Technology (CText) multilingual text corpora, provides document level aligned texts for MT purposes.

4.2 Spoken resources

We initially identified a total of 18 speech related resources. However, upon close scrutiny, we found that two of these resources were false results. One such instance is the Lwazi II Cross-lingual Proper Name corpus that is meant for Northern Sotho and not Sesotho (Kgampe & Davel 2010, 2011). In the end, we discuss a total of 16 spoken language resources.

4.2.1 Text to Speech and Automatic Speech Recognition

Seven resources were identified in this category. Three sets of corpora were produced by the Lwazi project. Two of the corpora are purposed for text-to-speech (TTS) while one is aimed at automatic speech recognition (ASR). The Lwazi and Lwazi II Sesotho TTS corpora contain transcriptions annotated with phonemic and orthographic information. The training sentences contain approximately equal speech sounds (Badenhorst et al. 2011). The Lwazi Sesotho ASR corpus contains audios and transcriptions used for the Lwazi speech recognition systems. The transcriptions are annotated with orthographic information for each word.

The NCHLT project also produced three TTS corpora. The speech corpus contains orthographically transcribed broadband speech data including a text suite of eight speakers (De Vries et al. 2014). We also identified the auxiliary speech corpus (Barnard et al. 2014, De Vries et al. 2014, Badenhorst et al. 2019). Finally, the inlang Pronunciation Dictionary for Sesotho contains an associated rule for generating pronunciations for unseen words (Barnard et al. 2014, Davel et al. 2013).

Lastly, we identified the Sesotho multi-speaker TTS corpus that contains audio and annotated transcriptions created for investigating the implementation of a high-quality TTS system that uses a low-cost process. The data sets were quality checked. However, the read.me file indicates that there might still be some errors. Unfortunately, accuracy results were not reported.

4.2.2 Sound-based corpora

In this category, we discuss tone-based and pronunciation-based corpora. Four tone-based corpora were identified. First, the Sesotho vowel speech data set contains a collection of words that represent five Sesotho orthographic vowels, that is, —*a e i o u*. The speech data was recorded with seven females and three male participants. Second, the metadata indicates that the intonation model for Bantu tone languages contains a model for isiZulu, Sepedi, Setswana, and Sesotho. The model is intended for theoretical linguistic intonation rules in prose. Unfortunately, the model has not been uploaded on the repository. As such, we are unable to report on its contents. Third, the Sesotho tone data set also contains male and female audio recordings. The participants were sourced from a specific region of South Africa called Qwaqwa. Fourth, the Sesotho function word speech data corpora, contains audios and annotated transcriptions aimed at studying the role of tone in *ke* and *o* as function words.

One pronunciation-based corpus was identified. The South African Multilingual Proper Names Corpus (Multipron) was developed in response to accent based variation in the pronunciation of personal names (Giwa et al. 2011). This corpus uses different speakers from isiZulu, Afrikaans, English and Sesotho. Four participants read Sesotho words. Sesotho words comprise 15% of the total data collected in this corpus. Each directory consists of the orthographic transcription in a text file, phonemic transcription containing phoneme strings and an audio file consisting of an acoustic representation of each word.

4.2.3 Dictation and telephony data

We identified a total of three digital language resources in this category. First, the South African Directory Enquiries (SADE) Name corpus (Thirion et al. 2020), uses a Sesotho home language speaker for accent, but the words used for training the voice dictation platform do not contain Sesotho names. Second, we identified two telephony speech data sets. The High quality TTS data for Afrikaans, Setswana, isiXhosa and Sesotho contains multi-speaker TTS audio data and transcription files. The African Speech Technology Sesotho speech corpus contains speech spoken by Sesotho mother tongue speakers (Roux et al. 2004). Third, the SADE municipality hotline IVR prompts corpus contains audio and corresponding transcriptions in English, isiZulu and Sesotho (Van Heerden et al. 2014). The produced recognition system recognises pronunciations with Afrikaans, English, isiZulu and Sesotho accents. The interface can also be customised to isiZulu or Sesotho.

4.3 Modules and Applications

This section discusses a total of 48 modules and applications available to Sesotho. Some applications have older and newer versions while others have basic and professional versions.

4.3.1 Non-language specific applications

Six non-language specific applications were noted. First, the Autshumato PDF Text Extractor is used for extracting texts for translation using the Autshumato automatic translation machine. It functions as a plugin for the OmegaT computer-assisted translation application. The translation system is named after Autshumato, possibly South Africa's first official translator and interpreter (Groenewald & Fourie 2009, Skosana & Mlambo 2021). Autshumato is one of South African government's initiatives for improving multilingualism through an increase in both quantity and quality of translation services (Groenewald & Fourie 2009). Unfortunately, the quality and accuracy of the machine translation web service is only optimal for government data because it was trained on this type of data (Skosana & Mlambo 2021). Nonetheless, if a translator translates similar documents, they can save the translation memories (TMs) and rely on them.

Second, we identified the Autshumato translation memory exchange (TMX) integrator which works as a utility that enables merging multiple TMs over networks through subversion (Schlemmer & Fourie 2013). Unfortunately, translation memory sharing is not yet common practice. To this end, we cannot estimate the possible contribution that individual translation projects can make towards building a bigger and more reliable translation memory for Sesotho translations. The TMX integrator only supports translation from English to Sesotho and not from Sesotho to English (Reina et al. 2013).

Third, the DictionaryMaker (Davel & Barnard 2003), has been evaluated on German, but we hope that it can also be applied on Sesotho texts. The DictionaryMaker allows the user to develop a pronunciation dictionary. When used, the human effort needed for developing such a dictionary is decreased. Moors, Wilken, Gumede & Calteaux (2018) indicate that there are three pronunciation dictionaries for Sesotho, one identified in 2009 and two identified in the 2014 audit, namely, the NCHLT-inlang, Lwazi and Lwazi II pronunciation dictionaries.

Fourth, we identified corpus related applications such as the (i) CorpusCatcher, designed to crawl the web for data using seed documents for constructing queries for document retrieval, (ii) Spelt, used in the creation of classified word lists that are used in spell checking, and (iii) TurboAnnotate1.0, used for manual creation of gold standard and annotated lists. According to Van Huyssteen & Puttkammer (2007), this application lowers human effort and improves accuracy of annotation. The TurboAnnotate1.0 application uses the Tilburg Memory-Based Learner machine learning system (*see* Daelemans et al. (2004)). It allows mother tongue speakers with limited and no experience with computational linguistics to annotate texts. Machine learning then learns from the user generated annotations.

4.3.2 South African official languages

In this section, we discuss applications that were developed for all eleven South African official languages. We identified at least 30 applications in this category.

Five NCHLT products were identified, namely, the Optical Character Recognition (OCR), language identifier, text web service, tagger and Part of Speech tagger. The OCR for South African Languages (Hocking & Puttkammer 2016) enables the user to convert scanned documents into editable texts. It can reproduce almost any character or image. The Language Identifier uses both a graphical user interface and a command line interface for automatically identifying official South African languages. The text web services provides access to tokenisers, sentence separators, POS taggers, phrase chunkers, named entity recognisers and OCR in South African languages. The tagger can be used either through the command line or a user interface. It annotates running text with either POS, named entity, noun phrase chunks, or nouns. Finally, the Part of Speech Taggers were developed using a minimum of one million government published tokens per language (Eiselen & Puttkamer 2014).

Four CText products were identified. First, the Alignment Interface and the Alignment Interface Pro are utility applications used for aligning source texts. The Pro version allows for editing the segments. These products work with the CText applications 1 that allows for automatic corpus query and manipulation for tokenisation and sentencisation, frequency and word list extraction, searching and extracting collocations. Additionally, the CText Applications 2 adds POS tagging, named entity recognition, and phrase chunking.

We also identified six independent applications, namely, (i) the AStudio, a software that incorporates a graphic interface for the developing flowcharts for speech applications, (ii) Automatic Oral Proficiency Assessment application, developed as part of the Development of Resources for Intelligent Computer-Assisted Language Learning project, (iii) the Language Identifier (LID) classifier token level classification for all official languages, (iv) a Combination Tagger that uses memory-based tagger (MBT), support vector machines (SVM), Mobotix part of speech tagger (MXPOST) and Trigrams'n'Tags (TnT) for deciding on tags, (v) the South African Fonts collection that contains fonts representing all alphabets and characters used in South African official languages, and (vi) the Format Normaliser 1.0. for normalising input files to utf8 txt, replacing smart quotes, and removing empty lines.

Four translation resources were also identified. One, the Rhonda machine translation system can handle speech to speech translation, or speech to text translations. Two, the Translate application kit 1.4.0 is a collection of applications and parsers that handles various localisable and translatable formats. It composes modules for segmentation, authentication and text enumeration. Three, the Autshumato Translation Management Systems web applications allow for capturing, editing, exporting and importing terminologies. Four, the Autshumato Text Anonymiser classifies and replaces sensitive information. For instance, if one wishes to anonymise sensitive information such as study participants' names, this application replaces them with pseudo names.

Five TTS related applications were also identified in this category. One, the Lwazi Telephony Platform combines Asterisk with MobillVR Python interface in one unified control interface. In this way, the process of developing experimental applications is accelerated. Two, the Qfrency TTS phone mappings application maps is used with Lwazi and NCHLT pronunciation dictionaries of the official languages. Three, the multilingual TTS Speect system provides a full service of decoding and encoding texts, that is, text analysis and speech synthesis with various APIs. Furthermore, it can be used for research and development of TTS system voices. Four, the Phonetic aligner contains scripts for automatic phonetic alignment of speech corpora using the hidden markov models. Finally, the Text Selection scripts for ASR/TTS, uses phonetic rules to phonetise texts, then the diphones are used for TTS and triphones are used for ASR on a per language basis.

4.3.3 Language specific resources

We identified four language specific applications that also incorporate Sesotho. One, the EtsaTrans translation system user interface is available in English, Afrikaans, isiXhosa and Sesotho. Two, the Multilingual Illustrated Dictionary with interactive games application is available for seven of official languages, namely, Afrikaans, South African English, isiXhosa, isiZulu, Sepedi, Setswana and Sesotho. Three, the NWU TransTips 1.0 is a PhP programming script that browses the web page for terms in the database. The translations appear when the user hovers over a certain word. Unfortunately, the user still has to decide on the correct translation between those presented.

Lastly, the Automated multilingual telephone access to financial services is a prototype that allows switching between isiZulu, isiXhosa, English and Sesotho.

4.3.4 Sesotho-only

We identified seven applications that are specifically for Sesotho. First, we identify two NCHLT products. One, the lemmatiser was developed using a rules-based approach (Eiselen & Puttkamer 2014). Two, the Morphological Decomposer that splits tokens to morphemes was developed after the POS tagger (Eiselen & Puttkamer 2014). Although Eiselen & Puttkamer (2014) provide examples of decomposition for isiZulu and Afrikaans, there are no examples for decomposition in other languages such as Sesotho. Second, the Lwazi Sesotho Pronunciation Dictionary includes audios for phonemes and letter-to-sound rule set based on generic words (Davel & Martirosian 2009). Although accuracy is not guaranteed, practical usability is assured. Third, the Lwazi II Sotho Pronunciation Dictionaries (Du Plessis et al. 1974), are based on the Lwazi dictionaries. Fourth, the Spelling Checker 1.0 is an application that checks spelling and provides hyphenations. It is compatible with some versions of Ms Office. CText continues to improve this application.

Finally, syllabification systems were developed as part of a project on developing a metric for measuring text readability in Sesotho. Two systems are part of the package. One, the ML-Based system produces 78.97% accurate results. Two, the rules-based system produces 99.69% accurate results (Sibeko & Van Zaanen 2022*a*).

4.3.5 Games

We identified one game, the Open Spell (v1.0), a spelling game that contains spelling exercises aimed at teaching spelling skills to school children. The source code is also freely accessible.

5 Discussion and conclusion

Many LRL's are underrepresented in NLP tasks because of insufficient corpora needed to complete NLP tasks (Mahloane & Trausan-Matu 2015). The limited availability of corpora as indicated in this article shows a great need for curating more corpora for Sesotho. This article considered basic digital language resources available to Sesotho from a BLARKs content perspective.

The listed resources were listed without in depth analysis of each resource such as considering how each resource works or the levels of accuracy and practical issues such as user-friendliness. We consulted literature relevant to the resources presented in the repository. To this end, some of these resources have been judged on issues such as accuracy and user-friendliness. However, this was not the aim of this paper. This is typical of BLARK content studies. Although this is basic, it functions as a starting point for investigations of BLARKs specification and definition. That is, determining what should be regarded as basic in the language in Sesotho and the quantities that should be developed. Nonetheless, it is clear from this article that written resources are focused only on very basic functions. For instance, there are no resources for higher level basic functionalities such as semantic analysis and term extractors. Even so, only the handwritten OCR and ontologies are missing from the written resources as identified in the BLARKs (Maegaard et al. 2005, 2006, Krauwer 2003). There have been even fewer attempts at speech technologies, especially those that are specifically aimed at Sesotho or the Sesotho language group. Even so, it is interesting how much work has been achieved.

A narrowed investigation into the current resources and an evaluation of each technology should be considered for future studies. SADiLaR is currently conducting their routine language resource audit, more resources might be added to their repository after their investigation. Perhaps we may gain more insight into digital language resources available to Sesotho.

This article was limited in three aspects. First, the resources surveyed in this review are not evaluated for their usability, accuracy, and precision. Second, some of the resources were only listed and indexed on the repository, however, the actual resources were not accessible. Third, the study only reviews resources that are indexed by SADiLaR. We recommend that future developments of digital language resources for Sesotho consider this inventory in their decisions on what resources to develop so that focus is paid to new and currently unavailable resources. Even so, we acknowledge that the current index indicates variation in the current collection.

Acknowledgements

We acknowledge Professor Menno van Zaanen for his guidance in the writing of this article and for suggesting that we collaborate on this project.

References

Arppe, A., Beck, K., Branco, A., Camilleri, V., Caselli, T., Cristea, D., Hinrichs, E., Liin, K., Nissinen, M., Parra, C., Rosner, M., Schuurman, I., Skadina, I., Quochi, V., Van Uytvanck, D. & Vogel, I. (2010), Description of the BLARK, the situation of individual languages, Report, Clarin.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L. & Moshagen, S. N. (2016), Basic language resource kits for endangered languages: A case study of plains cree, *in* 'Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016)', pp. 1–8.

Badenhorst, J., Martinus, L. & De Wet, F. (2019), Blstm harvesting of auxiliary nchlt speech data, *in* 'Proceedings of South African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/ROBMECH/PRASA 2019)', pp. 123–128.

Badenhorst, J., Van Heerden, C., Davel, M. & Barnard, E. (2011), 'Collecting and evaluating speech recognition corpora for 11 South African languages', *Language resources and evaluation* pp. 289–309.

Barnard, E., Davel, M. H., Van Heerden, C., F. De Wet, F. & Badenhorst, J. (2014), The nchlt corpus of the South African languages, *in* 'Proceedings of the 4th International Workshop Spoken Language Technologies for Under-resourced Languages', pp. 194–200.

Bosch, S. & Griesel, M. (2017), 'Strategies for building wordnets for under-resourced languages: the case of African languages', *Literator* .

Cambria, E. & White, B. (2014), 'Jumping NLP curves: A review of natural language processing research', *IEEE Computational intelligence magazine* **9**, 48–57.

Castellucci, G., Filice, S., Croce, D. & Basili, R. (2021), Learning to solve NLP tasks in an incremental number of languages, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)', pp. 837–847.

Chiguvare, P. & Cleghorn, C. W. (2021), Improving transformer model translation for low resource South African languages using bert, *in* 'IEEE Symposium Series on Computational Intelligence (SSCI)', IEEE, pp. 1–8.

Chitja, M. (2010), *Phatlamantsoe ya Sesotho ya Machaba*, Mazenod Publishers.

Chowdhury, G. (2003), 'Natural language processing', *Annual Review of Information Science and Technology* **37**, 51–89.

Cieri, C., Maxwell, M., Strassel, S. & Tracey, J. (2016), Selection criteria for low resource language programs, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 4543–4549.

Daelemans, W., Zavrel, J., Van Der Sloot, K. & Van den Bosch, A. (2004), Timbl: Tilburg memory-based learner, Tehnical report, Tilburg University.

Davel, M. & Barnard, E. (2003), Bootstrapping in language resource generation, *in* 'Fourteenth Annual Symposium of the Pattern Recognition Association of South Africa', pp. 97–100.

Davel, M., Basson, W., Charl, V. H. & Barnard, E. (2013), 'Nchlt dictionaries: Project report'.
URL: `https://sites.google.com/site/nchltspeechcorpus/home`

Davel, M. & Martirosian, O. (2009), Pronunciation dictionary development in resource-scarce environments, *in* 'Proceedings of the Interspeech', pp. 2851–2854.

De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E. & De Waal, A. (2014), 'A smartphone-based ASR data collection tool for under-resourced languages', *Speech Communication* pp. 119–131.

Drake, M. (2003), *Encyclopedia of Library and Information Science, Second Edition*, Taylor & Francis.

Du Plessis, J. A., Gildenhuys, J. G. & Moiloa, J. J. (1974), *Moiloa. Bukantswe ya malemepedi Sesotho-Seafrikanse / Tweetalige woordeboek Afrikaans-Suid-Sotho*, first edition edn, Via Afrika Beperk.

Eiselen, E. R. & Puttkamer, M. J. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the 9th International Conference on Language Resources and Evaluation', pp. 3698–3703.

Eiselen, R. (2016*a*), Government domain named entity recognition for South African languages, *in* 'Proceedings of the 10th Language Resource and Evaluation Conference'.

Eiselen, R. (2016*b*), South African language resources: phrase chunkers, *in* 'Proceedings of the 10th Language Resource and Evaluation Conference'.

Elenius, K., Forsborm, E. & Megyesi, B. (2008), Language resources and tools for swedish: A survey, *in* 'Proceedings of the LREC 2008'.

Giwa, O., Davel, M. H. & Barnard, E. (2011), A Southern African corpus for multilingual name pronunciation, *in* 'Pattern Recognition Association of South Africa and Mechatronics International Conference'.

Groenewald, H. J. & Fourie, W. (2009), Introducing the autshumato integrated translation environment, *in* 'Proceedings of the 13th Annual conference of the European Association for Machine Translation', pp. 190–196.

Grover, A. S., Van Huyssteen, G. B. & Pretorius, M. W. (2010), The South African human language technologies audit, *in* 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', pp. 2847–2850.

Grover, A. S., Van Huyssteen, G. B. & Pretorius, M. W. (2011), 'The South African human language technology audit', *Language resources and evaluation* **45**, 271–288.

Hanslo, R. (2021), Evaluation of neural network transformer models for named-entity recognition on low-resourced languages, *in* '16th Conference on Computer Science and Intelligence Systems (FedCSIS)', IEEE, pp. 115–119.

Hirschberg, J. & Manning, C. D. (2015), 'Advances in natural language processing', *Science* **349**, 261–266.

Hocking, J. & Puttkammer, M. (2016), Optical character recognition for South African languages, *in* 'Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)', pp. 1–5.

Kadenge, M. & Mugari, V. (2015), 'The current politics of African languages in zimbabwe', *Per Linguam: a Journal of Language Learning* **31**, 21–34.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q. & Liu, H. (2020), 'Natural language processing (NLP) in management research: A literature review', *Journal of Management Analytics* **7**, 139–172.

Kgampe, M. & Davel, M. H. (2010), Consistency of cross-lingual pronunciation of south-African personal names, *in* 'Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)', pp. 123–127.

Kgampe, M. & Davel, M. H. (2011), The predictability of name pronunciation errors in four South African languages, *in* 'Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)', pp. 85–90.

Koai, M. & Fredericks, B. G. (2019), 'Sesotho is still a marginalised language', *Southern African Linguistics and Applied Language Studies* **37**, 303–314.

Koehn, P. & Knight, K. (2002), Learning a translation lexicon from monolingual corpora, *in* 'ACL Special Interest Group on the Lexicon (SIGLEX)', pp. 9–16.

Krauwer, S. (2003), The basic language resource kit (BLARK) as the first milestone for the language resources roadmap, *in* 'Proceedings of SPECOM', Vol. 2003, pp. 15–22.

Liddy, E. D. (2001), *Natural Language Processing*, 2nd ed edn, Marcel Decker, Inc.

Maegaard, B., Choukri, K., Mokbel, C. & Yaseen, M. (2005), *Language technology for Arabic*, Center for Sprogteknologi, University of Copenhagen.

Maegaard, B., Krauwer, S., Choukri, K. & Jørgensen, L. D. (2006), The BLARK concept and BLARK for arabic, *in* 'LREC', pp. 773–778.

Magueresse, A., Carles, V. & Heetderks, E. (2020), 'Low-resource languages: A review of past work and future challenges', *arXiv:2006.07264* .

Mahloane, M. J. & Trausan-Matu, S. (2015), Metaphor annotation in Sesotho text corpus: towards the representation of resource-scarce languages in NLP, *in* '20th International Conference on Control Systems and Computer Science', IEEE, pp. 405–410.

Moeketsi, V. S. M. (2014), 'The demise of Sesotho language in the democratic South Africa and its impact on the socio-cultural development of the speakers', *Journal of Sociology and Social Anthropology* **5**, 217–224.

Mojela, V. (2016), Etymology & figurative: The role of etymology in the lemmatization of Sotho terminology, *in* 'The 10th International Conference of the Asian Association for Lexicography (AsiaLex2016)', IEEE, pp. 93–100.

Moors, C., Wilken, I., Calteaux, K. & Gumede, T. (2018), Human Language Technology Audit 2018: Analysing the development trends in resource availability in all South African languages, *in* 'Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists', pp. 296–304.

Moors, C., Wilken, I., Gumede, T. & Calteaux, K. (2018), 'Human Language Technology Audit 2017/18'.
URL: `https://sadilar.org/index.php/en/2-general/284-health-resources`

Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011), 'Natural language processing: an introduction', *Journal of the American Medical Informatics Association* **18**, 554–551.

Ndlovu, E. (2011), 'Mother tongue education in the official minority languages in zimbabwe', *South African Journal of African Languages* **31**, 229–242.

Ndlovu, E. (2013), Mother tongue education in official minority languages of Zimbabwe: A language management critique, Thesis, University of the Free State.

Nkolola-Wakumelol, M., Rantsoz, L. & Matlhaku, K. (2012), Syllabification of consonants in Sesotho and Setswana, *in* H. S. Nginga-Koumba-Binza & S. Bosch, eds, 'Language Science and Language technology in Africa: Festschrift for Justus C. Roux', Sun Express, Stellenbosch, South Africa, pp. 10–13.

Pretorius, L., Soria, C. & Baroni, P., eds (2014), *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, LREC.

Reina, A., Robles, G. & González-Barahona, J. M. (2013), A preliminary analysis of localization in free software: how translations are performed, *in* 'IFIP International Conference on Open Source Systems', pp. 153–167.

Riep, D. M. (2013), 'Seeing Sesotho: art, history, and the visual language of South Sotho identity', *Southern African Humanities* **25**, 217–244.

Roux, J. C., Louw, P. H. & Niesler, T. R. (2004), The African speech technology project: An assessment, *in* 'Proceedings of LREC', pp. 93–96.

Schlemmer, M. & Fourie, W. (2013), 'Autshumato tmx integrator'.
URL: https://repo.sadilar.org/handle/20.500.12185/416

Sefara, T. J., Mokgonyane, T. B. & Marivate, V. (2021), Practical approach on implementation of wordnets for South African languages, *in* 'Proceedings of the 11th Global Wordnet Conference', Global WordNet Association, pp. 20–25.

Sibeko, J. & Van Zaanen, M. (2022*a*), Developing a text readability system for Sesotho based on classical readability metrics, *in* 'Proceedings of Digital Humanities Conference: Responding to Asian diversity', Vol. 2022.

Sibeko, J. & Van Zaanen, M. (2022*b*), 'Raw and syllabified word list for Sesotho'.
URL: https://repo.sadilar.org/handle/20.500.12185/556

Skosana, N. J. & Mlambo, R. (2021), 'A brief study of the autshumato machine translation web service for South African languages', *Literator* pp. 1–7.

Snyman, D., Van Huyssteen, G. B. & Daelemans, W. (2011), Automatic genre classification for resource scarce languages, *in* 'Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa', pp. 132–137.

Strassel, S. & Tracey, J. (2016), Lorelei language packs: Data, tools, and resources for technology development in low resource languages, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 3273–3280.

Thirion, J. W., Van Heerden, C., Giwa, O. & Davel, M. H. (2020), 'The South African directory enquiries (sade) name corpus', *Language Resources and Evaluation* pp. 155–184.

Van Heerden, C., Kleynhans, N., Barnard, E. & Davel, M. (2010), Pooling ASR data for closely related languages, *in* L. Besacier & E. Castelli, eds, 'Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU 2010)', School of Computer Sciences, Universiti Sains Malaysia, pp. 17–23.

Van Heerden, E., Davel, M. & Barnard, E. (2014), Performance analysis of a multilingual directory enquiries application, *in* 'Proc. Annual Symp. Pattern Recognition Association of South Africa (PRASA)', pp. 258–263.

Van Huyssteen, G. B. & Puttkammer, M. (2007), 'Accelerating the annotation of lexical data for less-resourced languages', *Interspeech* pp. 1505–1508.

Wilken, I., Gumede, T., Moors, C. & Calteaux, K. (2018), Human Language Technology Audit 2018: Design considerations and methodology, *in* 'International Conference on Intelligent and Innovative Computing Applications (ICONIC)', IEEE, pp. 1–7.

Wissing, D. & Roux, J. C. (2017), 'The status of tone in Sesotho: a production and perception study', *Nordic Journal of African Studies* **26**, 19–19.