



A brief study of the Autshumato Machine Translation Web Service for South African languages

**Authors:**

Nomsa J. Skosana¹ 
Respect Mlambo¹ 

Affiliations:

¹South African Centre for Digital Language Resources, Faculty of Humanities, North-West University, Potchefstroom, South Africa

Corresponding author:

Nomsa J. Skosana,
Nomsa.Skosana@nwu.ac.za

Dates:

Received: 08 Dec. 2020

Accepted: 21 June 2021

Published: 29 Oct. 2021

How to cite this article:

Skosana, N.J. & Mlambo, R., 2021, 'A brief study of the Autshumato Machine Translation Web Service for South African languages', *Literator* 42(1), a1766. <https://doi.org/10.4102/lit.v42i1.1766>

Copyright:

© 2021. The Authors.
Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

The scarcity of adequate resources for South African languages poses a huge challenge for their functional development in specialised fields such as science and technology. The study examines the Autshumato Machine Translation (MT) Web Service, created by the Centre for Text Technology at the North-West University. This software supports both formal and informal translations as a machine-aided human translation tool. We investigate the system in terms of its advantages and limitations and suggest possible solutions for South African languages. The results show that the system is essential as it offers high-speed translation and operates as an open-source platform. It also provides multiple translations from sentences, documents and web pages. Some South African languages were included whilst others were excluded and we find this to be a limitation of the system. We also find that the system was trained with a limited amount of data, and this has an adverse effect on the quality of the output. The study suggests that adding specialised parallel corpora from various contemporary fields for all official languages and involving language experts in the pre-editing of training data can be a major step towards improving the quality of the system's output. The study also outlines that developers should consider integrating the system with other natural language processing applications. Finally, the initiatives discussed in this study will help to improve this MT system to be a more effective translation tool for all the official languages of South Africa.

Keywords: Autshumato project; Autshumato machine translation web service; South African languages; translation output; machine translation system; parallel corpora.

Introduction

South Africa is a nation of cultural diversity, with 11 official languages spoken by various ethnic groups. Nine of these languages – Xitsonga, isiNdebele, Tshivenda, Siswati, isiXhosa, isiZulu, Sepedi, Sesotho and Setswana – are indigenous and were officially recognised by an Act of Parliament after South African independence in 1994. Ngcobo and Nomdebevana (2010:187) reported that these nine languages are under-resourced when compared with English and Afrikaans, the other two official languages, in terms of the scarcity of sufficient computational resources such as online dictionaries, machine translation (MT) systems, part-of-speech taggers, spellcheckers, named entity recognisers, morphological analysers and so on. The problem of the availability of such linguistic resources is especially acute for under-resourced languages and narrow domains. The South African government, recognising the value of these languages, and within the jurisdiction of the Pan South African Language Board, has enacted language policies aimed to promote indigenous languages that were historically marginalised. Recently the government introduced another legislation that regulates and monitors the use of official languages by public servants in state departments and other public entities and enterprises (Ralarala 2019:262).

As part of its constitutional obligation, the government has also seen fit to support projects that advance multilingualism in the country and the functional development of all official languages. One of the initiatives, established in 2007 under the auspices of the Department of Arts and Culture, was the Autshumato Project.

Autshumato was a leader of Khoikhoi who also worked as a translator and interpreter in the trade negotiations between Europeans and Khoikhoi at the Cape of Good Hope during the 17th century (Houston et al. 2013:68).

To raise awareness about this prominent South African and to perpetuate his legacy, Autshumato was honoured with the project's name as the first South African translator and interpreter (<https://mt.nwu.ac.za/>).

Read online:

Scan this QR code with your smart phone or mobile device to read online.

The facilities developed in this project include the Autshumato Integrated Translation Environment, the Autshumato Terminology Management System, the Translation Memory and Glossary Integration System, the Autshumato Machine Translation Web Service (MTWS) and various other tools, corpora and resources for South African languages. The website <http://autshumato.sourceforge.net/> can be viewed for a detailed description of the tools and resources created under this initiative. The ultimate goal of the Autshumato Project is to develop adequate resources and tools for all the official languages through concrete and constructive action in various specialised fields (Groenewald & Fourie 2009:190). One expectation is that it would lead to the use of these languages in modern fields such as science and technology.

The study focusses on the MT system (MTWS) designed for South African languages. It is conducted by reviewing the advantages and limitations. This particular facility was created by the Centre for Text Technology at the North-West University, Potchefstroom. The MTWS was designed to provide human translators and individuals with reliable translations, as it should enable them to communicate with each other without extensive knowledge of each other's languages (<https://mt.nwu.ac.za/>). The MTWS functions as a machine-aided human translation (MAHT) tool, which implies that it cooperates with human translators to complete translation tasks. However, as with other MT systems, the MTWS has limitations that can be improved to benefit users.

This study is structured into sections. The information on related work is briefly reviewed in the second section. The advantages of the MTWS are clarified in the third section. The fourth section outlines the limitations of the system. The potential suggested solutions are discussed in the fifth section. A conclusion is provided in the last section.

Related work

The literature reviewed in this section is primarily based on the development and the evaluation of MT systems, with a focus on European and Asian languages as much research has already been carried out in these languages. This section examines studies that discuss MT systems for minor languages and their challenges as resource-scarce languages. The section also explores studies that investigate methods for gathering parallel text to improve the performance of MT systems. The literature reviewed was chosen because it addresses the MT systems issues that are similar to those experienced by South African languages.

Forcada (2006) examined the open-source MT system for minor languages, by reflecting on its effects, opportunities and challenges. In this study, minor languages are those with a small number of speakers, used mainly at home rather than in official documents, having a restricted presence on the internet and lacking standardised writing, reliable spelling systems, linguistic expertise and machine-readable tools

such as linguistic data. The development of open-source MT systems can have a significant influence on minor languages by increasing literacy, affecting standardisation and raising awareness of the language community at large. It can also promote the normality of these languages being used not only in informal settings but also in state departments. In addition, the study indicates that the open-source MT systems will increase the independence, expertise and language resources for minor languages.

Forcada (2006) also showed that open-source MT provides opportunities for the promotion of minor languages, from being almost ignored to becoming recognisable languages. Nonetheless, to capitalise on these opportunities, the minor language users have to face challenges, which include standardisation, neutralising of technophobic attitudes, organising community engagement, and eliciting and simplifying the linguistic knowledge. The modularity and documentation of linguistic data formats were also identified as challenges that minor languages could overcome to build an open-source MT system.

Zakir and Nagoor (2017) conducted a study on various linguistic problems concerning MT systems translating English into Urdu. In this study, they indicate that target users worldwide, including the Urdu community, are using the MT systems to translate different texts. However, achieving better quality output from their MT system has become a real problem, with issues related to word and phrase translations, as well as syntactic and semantic translation. Concerning word translation, Zakir and Nagoor reported that the system cannot translate words with multiple meanings. They then quote the following examples where this challenge was experienced: 'please book my ticket for tomorrow' and 'please buy that book for me'. In the two sentences, the word 'book' was translated as published work even though in the first sentence it was supposed to be used as a verb meaning reserving a seat. With regard to phrase translation, it proved difficult for the Urdu MT system to deal with figurative words that contain hidden meanings.

For syntactic and semantic translation, a problem arises because English and Urdu are structured differently. Syntactically English follows 'subject-verb-object' word order whilst Urdu uses 'subject-object-verb' sentence structure. The word order difference between the two languages results in a syntactic translation problem. Semantically, Zakir and Nagoor (2017) highlight that the MT system is unable to translate pronoun resolutions. For example, translating a sentence such as 'my son dropped the glass plate and it broke into pieces', the MT system finds it challenging to detect what the pronoun resolution 'it' refers to in such a sentence. To address these problems Zakir and Nagoor propose that adding more parallel corpora from different fields should be considered and prioritised as it will be an essential resource in the improvement of the MT system, and they strongly believe that a large number of additional parallel corpora will provide better solutions.

Germann (2001) reported on building a statistical machine translation (SMT) system for Tamil, a language spoken mainly in Sri Lanka and India. Here, an effort is needed to build parallel corpora as an essential resource for the construction of an effective MT system. This kind of corpus contains a collection of large texts, where the source texts are paired with their translation in the target texts. Various Tamil online published newspapers and magazines – and the Tamil community at large – were consulted for the dissemination of corpora. The output and performance of the SMT system that was trained using small quantities of parallel corpora were poor, but the addition of a parallel corpus over time increased performance on evaluation tasks such as document retrieval and question-answering on the system. These documents were used as they contained answers to the questions set for doing the pilot run.

Koehn and Knowles (2017) examined general challenges that are experienced in neural machine translation (NMT) and provide the outcomes on how this application is presently maintained in comparison to SMT. The study reveals that, despite its recent triumphs, NMT systems still face a number of challenges. The study finds the following challenges with regard to the NMT systems:

- Neural machine translation systems have a lower-quality output of domain, to the point that they entirely disregard sufficiency in a favour of fluency.
- Neural machine translation systems have a steeper learning curve with respect to the amount of training data, resulting in worse quality in low-resource settings, but better performance in high-resource settings.
- Neural machine translation systems that operate at the sub-word level perform better than SMT systems on low-frequency words, but still show weakness in translating low-frequency words belonging to highly inflected word categories (e.g. verbs).
- Neural machine translation systems have lower translation quality on long sentences than SMT but do comparably better up to a sentence length of about 60 words.
- The attention model for NMT systems does not always fulfil the role of a word alignment model, but may dramatically diverge.
- In NMT systems, beam search decoding, unlike in SMT systems, only improves translation quality for narrow beams and deteriorates when exposed to larger search spaces.

The study went further to state that the main cause of these challenges is that NMT systems, when faced with situations that differ greatly from the training settings, do not exhibit robust behaviour because of limited exposure to training data. Koehn and Knowles (2017) concluded their study by stating that the potential solution to these challenges of NMT systems may thus lay in a broader training method that goes beyond maximising single-word predictions given precisely matching previous sequences.

Skadina et al. (2012) investigated collecting and using comparable corpora for the construction of an SMT system. In contrast to a parallel corpus, the comparable corpus consists of original texts from different languages, which originate from the same domain. This study was undertaken within the Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation (ACCURAT) project, which targeted a range of European languages such as Croatian, Estonian, Greek, Latvian, Lithuanian and Romanian. The study presents tools and resources for the collection, evaluation and alignment of comparable texts for the development of an effective SMT system for these languages.

Corpora containing published news articles, interlanguage connections to articles from Wikipedia and corpora covering domain-specific texts were the three primary sources used to collect large quantities of comparable texts. The collected corpora were characterised in terms of three categories of comparability. Firstly, comparable texts were aligned at the word level. Secondly, the strongly comparable corpora, as a corpus that describes the same event or a phenomenon, were aligned at the document level. Thirdly, the weakly comparable corpora, as a corpus which is within the same narrow domain, were only aligned on the corpus level. The conclusion was that all methods implemented to improve the output of the MT system were positive in terms of data collection, comparability of metrics precision and alignment procedures.

To our knowledge, no study has yet embarked on examining the MTWS as a translation tool. This study explores MTWS designed to serve South Africa's official languages, in terms of its advantages and limitations, and possible solutions that will help to improve this MT system to be a more effective translation tool for all the official languages of South Africa.

The advantages of the Autshumato machine translation web service

Machine translation systems are intended to accelerate the rate of translating documents with or without the intervention of human translators. For the translation task to be completed using MTWS, human intervention is required. The advantages of the MTWS are discussed next.

High-speed translation

The MTWS, as a translation tool that works together with human interaction, conducts high-speed translation, maximising the quantity of work translators and users can translate in a limited amount of time. The MTWS was meant to help human translators to speed up both the quantity and the quality of their translated work (Groenewald & Fourie 2009:191). One of the undeniable advantages of MT systems is the remarkably high translation speed that helps to save time by reusing previously translated strings (Muegge 2001:26). Therefore, the high-speed translation of MTWS does not always produce quality translation as it is a MAHT

tool. As a result, MTWS translated texts must be post-edited before usage.

Open-source platform

Forcada (2006:1) defined an open-source platform as software that may be freely executed, examined, modified, redistributed, improved and released to the public so that the whole community of users benefits. The MTWS is a cost-effective open-source translation assisted tool, which can be accessed freely at the South African Centre for Digital Language Resources (SADiLaR) website (<https://sadilar.org/index.php/en/>) by the language practice community and other language users with access to the internet, without a subscription. Groenewald and Fourie (2009:191) asserted that as an open-source MAHT tool, the MTWS will benefit not only translators but also the language community at large because the system is freely available online. The only expense that language users may incur is during the post-editing process. Furthermore, as an open-source tool, MTWS provides features that allow translators and users to improve translation in situations where the results have not been satisfactory, as well as to offer personal information and qualifications to become reviewers.

Multiple translation capacity

The MTWS is an important translation tool in the context of South African languages because of its capacity to assist language users in the translation of web pages, sentences and documents from one source language (English) into six target languages (Afrikaans, Xitsonga, Setswana, isiZulu, Sesotho and Sepedi). The MTWS includes a translation memory that aids in the storage of translated words, terms and segments with their matching source words, phrases and segments for future translation (Nemutamvuni 2018:51). Machine translation systems do not get tired or distracted. They reliably parse and translate every sentence in source documents into the target language (Muegge 2001:26). However, even in this situation, post-editing is advisable when users employ MTWS's ability to translate web pages, phrases and documents to improve translation quality.

Limitations

The ability to advance the status and encourage the fair use of all South African languages in various fields requires the existence of relevant resources such as online translation tools (Mlambo, Skosana & Matfunjwa 2021:82). However, such resources and particularly MTWS have limitations, which are discussed and reviewed below.

Exclusion of other languages

The MTWS developers have treated South African languages unequally. The system recognises English as the only source language and the other six languages as target languages out of the 11 official languages of South Africa. As already stated, MTWS only translates from English into

Xitsonga, isiZulu, Sesotho, Sepedi, Setswana and Afrikaans. The problem with English being used as a primary language in this tool is that the remaining official languages are still neglected in the government domain (Groenewald & Fourie 2009:190). Other languages such as Tshivenda, isiXhosa, Siswati and isiNdebele have not been incorporated into the system. The developers adopted a phased approach in the selection of languages based on the availability of training data. Groenewald and Fourie (2009:196) stated that the most difficult issue that MTWS developers face is that these languages are resource-scarce, and the development of MT systems requires a large number of parallel corpora.

Hence, one of the primary objectives of creating MTWS was to promote multilingualism and access to information to all South Africans (Nemutamvuni 2018:50). However, the inability to incorporate other official languages as source languages, as well as the exclusion of other languages, has hampered the advancement of multilingualism amongst South Africans. As a result, this exclusion of other languages deepens the inequalities for these languages and their speakers.

Training data

Mandal et al. (2008:1) observed that the output of any MT system relies heavily on the availability of parallel training data in terms of quality and size. In support, Bakaric and Pacelat (2019:11) stated that the collection and preparation of parallel corpora influence the performance of MT systems, as they depend on the quantity and quality of the parallel training data. The MTWS was trained using the Autshumato parallel corpora for selected languages. The corpora used were based on documents from the South African government domain, which were collected from government databases (<http://autshumato.sourceforge.net/>). Although part of the data acquired was parallel, not all the data gathered were available in all languages (Eiselen & Puttkammer 2014:3699), as the data were later translated into various selected South African languages. The corpora were created as part of the Autshumato initiative, which seeks to provide access to data sets to assist in the development of open-source translation technologies for all the South African languages (<http://mt.nwu.ac.za/#>).

According to Van Zaanen et al. (2020) the amount of data used to train all the language resources of the Autshumato Project facilities including MTWS was limited. Owing to the limited amount of training data, the quality output of the MTWS is relatively low, particularly if the source text is beyond the scope of training data. See the translation of an article published by Sowetanlive on 17 November 2020, entitled 'Zuma waits to hear outcome of Zondo recusal application' (Table 1).

From Table 1, the translation is not up to standard as some of the words are not translated into the target language, but have been kept in the original language: words such as

TABLE 1: A source text (English) and its translation in isiZulu from machine translation web service.

English	isiZulu
Zuma waits to hear outcome of Zondo recusal application.	UZuma Bekela ukuzwa umphumela Zondo sokuhoxisa isicelo.
Former president Jacob Zuma has arrived at the Zondo commission of inquiry into state capture where a ruling on his application for the inquiry's chairperson deputy chief justice Raymond Zondo to recuse himself is expected.	UJacob Zuma, owayengumongameli wafika Zondo ikhomishani yophenyo umbuso wokubamba lapho ukunquma uma isicelo ukuphenya's usihlalo nephini lejaji uRaymond Zondo kufanele azihoxise yena uqobo lwakhe ophenyweni kulindelwe.
Zuma's legal representative advocate Muzi Sikhakhane on Monday argued that Zondo's comments during and after the testimony of some witnesses fed into the narrative that Zuma is the man who destroyed our country.	UZuma's Legal isithunywa u-Advocate Muzi Sikhakhane ngoMsombuluko Sikhulumile Zondo ukuphawula komngani ngesikhathi; abanye ofakazi bese into the narrative ukuthi Zuma, uyena o ba izwe lethu.
According to Sikhakhane, the inquiry's selection of witnesses was also a cause for concern for Zuma.	Ngokuya Sikhakhane, ukuphenya's nezithombe ofakazi futhi indaba edinga umnako for Zuma.

Source: Mahlangu, I., 2020, 'Zuma waits to hear outcome of Zondo recusal application', Sowetanlive, November 17, viewed 18 November 2020, from <https://www.sowetanlive.co.za/news/south-africa/2020-11-17-zuma-waits-to-hear-outcome-of-zondo-recusal-application/>

'legal', 'advocate', 'into the narrative' and 'for' have been retained as they are in translation. The MTWS was unable to translate the quoted words from the original text. Such inabilities and other translation errors are experienced as it is extremely difficult for MTWS to provide context-sensitive translation (<http://autshumato.sourceforge.net/>). Hence, the development of effective MT systems for all languages requires quality training data from different spheres (Doğru, Martín & Aguilar-Amat 2018:12). Such data are most likely to contain a variety of words used in different contexts and that will enable the MT systems in general and MTWS in particular to deliver high-quality output.

In the first sentence, the source sentence reads, 'Zuma waits to hear outcome of Zondo recusal application', and the translated target text reads, 'UZuma Bekela ukuzwa umphumela Zondo sokuhoxisa isicelo'. The target language seems to have been translated correctly but the incorrect use of the capital letter 'B' in the word 'Bekela' shows the capitalisation shortcomings of the system. The whole sentence in the target language is also syntactically incorrect, as the system translated it from the source language using a word-to-word strategy without rephrasing it to conform to the norms of the target language. This issue may be caused by the fact that isiZulu is a morphologically rich language with a conjunctive writing system, which will likely pose significant challenges in the development of MT systems (Van Huyssteen & Griesel 2016:332). In this case, human intervention for post-editing was needed for the sentence to be arranged correctly as 'UZuma ulindele ukuzwa umphumela kaZondo wesicelo sokuhoxisa'.

The second sentence, that reads as:

Former president Jacob Zuma has arrived at the Zondo commission of inquiry into state capture where a ruling on his application for the inquiry's chairperson deputy chief justice Raymond Zondo to recuse himself is expected.

was translated into the target language as 'UJacob Zuma, owayengumongameli wafika Zondo ikhomishani yophenyo umbuso wokubamba lapho ukunquma uma isicelo ukuphenya's usihlalo

nephini lejaji uRaymond Zondo kufanele azihoxise yena uqobo lwakhe ophenyweni kulindelwe'. In this translation, the meaning was lost because of the omission and incorrect use of conjunctions. To detect these inconsistencies, the direct back translation would read as:

Jacob Zuma, the former president arrived Zondo the commission of inquiry the state capture where a ruling for application for the inquiry's chairperson and the deputy chief justice Raymond Zondo is expected to recuse himself.

When compared with the source text, the back translation indicates that conjunctions such as 'has' and 'at the' were omitted, whilst conjunctions such as 'into', 'on' and 'and' were improperly used. Therefore, the omission and incorrect use of these conjunctions in the target language resulted in the loss of meaning in the translation.

The third sentence in the source language that reads as:

Zuma's legal representative advocate Muzi Sikhakhane on Monday argued that Zondo's comments during and after the testimony of some witnesses fed into the narrative that Zuma is the man who destroyed our country.

was translated as 'UZuma's Legal isithunywa u-Advocate Muzi Sikhakhane ngoMsombuluko Sikhulumile Zondo ukuphawula komngani ngesikhathi; abanye ofakazi bese into the narrative ukuthi Zuma, uyena o ba izwe lethu' in the target language. The word 'representative' in the source language, in this case, means the legal representative and also has multiple meanings outside this context, hence in the target language it was translated as 'isithunywa' meaning the messenger, which is not what is meant in the source language.

The same sentence mentioned here also contains spelling errors and/or unreadable words such as 'o ba', which does not have a meaning in the target language. The final translated sentence in the target language, 'Ngokuya Sikhakhane, ukuphenya's nezithombe ofakazi futhi indaba edinga umnako for Zuma', also has a few challenges as the source text reads as 'According to Sikhakhane, the inquiry's selection of witnesses was also a cause for concern for Zuma'. The incorrect use of words and punctuation symbols such as the apostrophe in the word 'ukuphenya's' in the target language makes the whole paragraph lose its meaning. The translation errors discussed show the greatest limitation of MTWS, that it cannot relate words to the context, particularly those with multiple meanings because those words must be linked to the context to better establish their true meaning.

Possible solutions

Although some of the South African languages are under-resourced, it does not mean that nothing can be carried out to help improve the resources that were built to cater for them. This section discusses a few solutions suggested, which could further improve the MTWS's effectiveness.

Addition of data

To enhance the quality output of an MT system, high-quality training data are required. Megyesi, Hein and Johanson (2006:2130) stated that parallel corpora are an essential resource for developing NLP applications such as MT systems. As stated, the output of MTWS is fairly poor owing to the limited amount of data used to train it. Several studies such as those of Germann (2001), Skadina et al. (2012), Zakir and Nagoor (2017) and Koehn and Knowles (2017) have demonstrated the validity of using parallel corpora from various fields to improve the quality output of MT systems. To improve the MTWS, high-quality parallel training corpora from various contemporary fields should also be added for all the South African official languages. As Lü, Huang and Liu (2007) pointed out:

[T]he more parallel data is used to estimate the parameters of the MT model, the better it can approximate the true translation probabilities, which will lead to a higher translation performance. (p. 343)

Typically, the addition of new parallel corpora from various contemporary fields to MTWS could play a significant role in improving the quality output. Unfortunately, unlike monolingual corpora, parallel corpora are scarce resources in the context of South African languages. However, to address this challenge of scarce resource, the MTWS developers should consider cooperating with different language sectors that mainly deal with translation, such as lexicography units, private translation agencies, university translation departments, language standardisation sectors and state departments to share their parallel corpora. The MTWS as an open-source platform can be accessed freely online; therefore, to overcome this scarcity of parallel corpora, the developers should explain to these sectors how they will benefit from sharing their data.

Language specialists' involvement

The involvement of linguists as language experts in the development of MT systems can also play a major role in addressing some of the language-related shortcomings that inhibit the machine translator's output quality. Rohrer (1986:353) made an observation that MT initiatives are mostly driven by computer scientists who have neglected the complexities of natural language and this leads to MT systems' difficulties in producing high-quality output. The MTWS is also characterised by providing inaccurate words, spelling, tenses, sentence structuring and other language deficiencies. Some of these shortcomings can be overcome by including language experts in the development and the pre-editing of training data. Rohrer (1986:354) proposed that linguists and computer scientists must collaborate to address some of the language-related difficulties, where linguists compose their linguistic texts to be used as training data in a

formalism that meets the standards of a decent programming language. Furthermore, Muegge (2001:27) highlighted that linguists should be allowed to pre-edit training resources to eliminate translation and linguistic errors in MT systems. The collaborative efforts between language experts and computer scientists will lead to an effective MT system for South African languages.

Integration of tools

The human language technology (HLT) industry in South Africa has been one of the key contributors in addressing some of the needs required to bridge the gap that exists in resource-scarce languages, even though the HLT industry for South African languages is still relatively new. The industry consists of a few NLP applications such as tokeniser, sentence separator, phrase chunker, part of speech tagger, named entity recogniser, language identifier, optical character recognition and spelling checker for all official languages (Van Huyssteen & Griesel 2016:329). These resources are freely accessible at the SADiLaR's website (<https://sadilar.org/index.php/en/>). The integration of such applications with MTWS can help to overcome some of the language-related issues associated with the current version of MTWS. According to Van Huyssteen and Griesel (2016:329), integrating spelling checkers and MTWS can enhance the system's translation quality output, particularly for conjunctive languages, by offering spelling variants and/or validating the correctness generated from sentence constructions.

Conclusion

In this study, we reviewed MTWS as an online translation tool for South African languages and highlighted its advantages and limitations and possible improvement solutions. We established that MTWS has the advantage of being an open-source platform, performing high-speed translation and multiple translations from web pages, sentences and documents from English into selected languages. As a freely accessible MAHT tool, MTWS can be used by human translators to maximise and execute translation tasks efficiently and successfully. However, these advantages have resulted in limitations faced by the users. Of South Africa's 11 official languages, only seven have been included in the development of this system. The MTWS only performs translations from English into six target languages (Xitsonga, isiZulu, Sesotho, Sepedi, Setswana and Afrikaans). As a result of a lack of training data, other languages such as Tshivenda, isiXhosa, Siswati and isiNdebele have not been incorporated into the system. The study also found that MTWS cannot relate words to the context, particularly those with multiple meanings because the system was trained with a limited amount of data and this has an adverse effect on the quality of the output.

Therefore, improvements such as adding various specialised quality parallel corpora – and involving language specialists in the development and pre-editing of training data – could

play a major role in addressing some of the language-related challenges that are encountered in the output of MTWS. The integration of MTWS and other NLP applications could also help to address some of the translation challenges that users are encountering with the system. The initiatives discussed in this study will help to improve MTWS to be a more efficient and effective translation tool for all the official languages of South Africa.

Acknowledgements

The authors would like to thank the following people for their helpful comments and feedback during the writing of this article: Kerlick Workshop coordinators, Menno van Zaanen, Juan Steyn, Vusi Msiza, and Rootheier Mabuya.

Competing interests

The authors declare that they do not have any financial or personal advantages that could have influenced them in writing this article.

Authors' contributions

As a main author N.J.S. came with the idea of evaluating the MT system for South African languages. She also discussed advantages and limitations of the system as part of formal analysis. Lastly, she was also responsible for asking the affiliated organisation to assist with funding. As a second author R.M. contributed in the formal analysis mainly focussing on related work and suggested solutions for the system that the article was evaluating. He also did the paper layout and editing.

Ethical considerations

This article followed all ethical standards for research without direct contact with human or animal subjects.

Funding information

This publication was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). The South African Centre for Digital Language Resources is a research infrastructure established by the Department of Science and Innovation of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

Disclaimer

The views shared in this article are those of the authors and do not represent the opinions of others or the associated authors' organisation.

References

- Bakaric, M.B. & Pacelat, I.L., 2019, 'Parallel corpus of Croatian-Italian administrative texts', in *2nd International workshop, Human-Informed Translation and Interpreting Technology (HiT-IT) 2019 proceedings*, September 5–6, 2019, pp. 11–18, Varna, Bulgaria.
- Doğru, G., Martín, A. & Aguilar-Amat, A., 2018, 'Parallel corpora preparation for machine translation of low-resource languages: Turkish to English cardiology corpora', in *11th International Language Resources and Evaluation Conference (LREC) 2018 proceedings*, May 7–12, 2018, pp. 12–15, Paris, France.
- Eiselen, E. & Puttkammer, M., 2014, 'Developing text resources for ten South African languages', in *9th International conference, LREC 2014 proceedings*, May 26–31, 2014, pp. 3698–3703, Reykjavik, Iceland.
- Forcada, M.L., 2006, 'Open-source machine translation: An opportunity for minor languages', in *5th International LREC 2006 proceedings*, May 22–28, 2018, pp. 1–6, Genoa, Italy.
- Germann, U., 2001, 'Building a statistical machine translation system from scratch: How much bang for the buck can we expect?', in *International workshop, Association for Computational Linguistics (ACL) 2001 proceedings*, July 6–7, 2001, Toulouse, France.
- Groenewald, H.J. & Fourie, W., 2009, 'Introducing the Autshumato integrated translation environment', in *13th Annual conference, European Association for Machine Translation (EAMT) proceedings 2009*, May 14–15, 2009, pp. 190–196, Barcelona, Spain.
- Houston, G., Mati, S., Seabe, D., Peires, J., Webb, D., Dumisa, S. et al., 2013, *The liberation struggle and liberation heritage sites in South Africa*, Report, Human Sciences Research Council, Cape Town, viewed 25 May 2021, from <http://hdl.handle.net/20.500.11910/2491>.
- Koehn, P. & Knowles, R., 2017, 'Six challenges for neural machine translation', in *1st workshop on neural machine translation, ACL proceedings 2017*, July 30 – August 04, 2017, pp. 28–39, Vancouver, Canada.
- Lü, Y., Huang, J. & Liu, Q., 2007, 'Improving statistical machine translation performance by training data selection and optimization', in *4th joint International conference, Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) 2007 proceedings*, June 28–30, 2007, pp. 343–350, Prague, Czech Republic.
- Mahlangu, I., 2020, 'Zuma waits to hear outcome of Zondo recusal application', *Sowetanlive*, November 17, viewed 18 November 2020, from <https://www.sowetanlive.co.za/news/south-africa/2020-11-17-zuma-waits-to-hear-outcome-of-zondo-recusal-application/>.
- Mandal, A., Vergyri, D., Wang, W., Zheng, J., Stolcke, A., Tur, G. et al., 2008, 'Efficient data selection for machine translation', in *workshop on spoken language technologies (SLT), Institute of Electrical and Electronics Engineers (IEEE)/ACL proceedings 2008*, December 15–19, 2008, pp. 261–264, Goa, India.
- Megyesi, B.B., Hein, A.S. & Johanson, E.C., 2006, 'Building a Swedish Turkish parallel corpus', in *5th International LREC proceedings 2006*, May 22–28, 2006, pp. 2130–2133, Genoa, Italy.
- Mlambo, R., Skosana, N. & Matfunjwa, M., 2021, 'The extraction of terminology list using ParaConc for creating a quadrilingual dictionary', *Southern African Linguistics and Applied Language Studies* 39(1), 82–91. <https://doi.org/10.2989/16073614.2021.1896971>
- Muegge, U., 2001, 'The best of two worlds: Integrating machine translation into standard translation memories: A universal approach based on the TMX standard', *Language International* 13(6), 26–28.
- Ṁemutamvuni, M.E., 2018, 'Investigating the effectiveness of available tools for translating into Tshivenda', M.A. dissertation, Department of African Languages, University of South Africa, Pretoria.
- Ngcobo, M.N. & Nomdebevana, N., 2010, 'The role of spoken language corpora in the intellectualisation of indigenous languages in South Africa', *Alternation* 17(1), 186–206.
- Ralarala, M., 2019, 'Policy analysis as "text" in higher education: Challenging South Africa's "use of official languages act": A case-based approach', *South African Journal of Higher Education* 33(4), 253–270. <https://doi.org/10.20853/33-4-3030>
- Rohrer, C., 1986, 'Linguistic bases for machine translation', in *11th International Conference on Computational Linguistics (COLING) 86 proceedings 1986*, August 25–29, 1986, pp. 353–355, Bonn, Germany.
- Skadina, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M. et al., 2012, 'Collecting and using comparable corpora for statistical machine translation', in *8th International LREC proceedings 2012*, May 23–25, 2012, pp. 438–445, Istanbul, Turkey.
- The Autshumato MT Web Service, viewed 20 October 2020, from <https://mt.nwu.ac.za/#>.
- The Autshumato Project, viewed 20 October 2020, from <http://autshumato.sourceforge.net/>.
- The SADiLaR Website, viewed 20 October 2020, from <https://sadir.org/index.php/en/>.
- Van Huyssteen, G.B. & Griesel, M., 2016, 'Translation technology in South Africa', in C. Sin-Wai (ed.), *The Routledge encyclopedia of translation technology*, pp. 327–336, Routledge, New York, NY.
- Van Zaanen, M., Trollip, B., Ramukhadi, P.M. & Mlambo, R., 2020, 'Identifying relations between characters in Afrikaans, Tshivenda, and Xitsonga books', in *annual conference of the Alliance of Digital Humanities Organizations (ADHO)*, July 20–25, 2020, Ottawa, Canada, viewed 24 September 2020, from <https://hcommons.org/deposits/item/hc:32053/>.
- Zakir, H.M. & Nagoor, M.S., 2017, 'A brief study of challenges in machine translation', *UCSI International Journal of Computer Science Issues* 14(2), 54–57. <https://doi.org/10.20943/01201702.5457>