Localising the Mozilla Common Voice platform for South Africa's official languages

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University & School of Electrical, Electronic and Computer Engineering, North-West University
fdw@sun.ac.za

Bukula, Andiswa
South African Centre for Digital Language Resources, North-West University
Andiswa.Bukula@nwu.ac.za

Karsten, Willem
School of Electrical, Electronic and Computer Engineering, North-West University
willemkarsten2308@gmail.com

Puttkammer, Martin
Centre for Text Technology, North-West University
Martin.Puttkammer@nwu.ac.za

Schillack, Erwin
School of Electrical, Electronic and Computer Engineering, North-West University
schillackerwin@gmail.com

Wierenga, Rone'
Virtuele Instituut vir Afrikaans
rone@viva-afrikaans.org

Eiselen, Roald
Centre for Text Technology, North-West University
Roald.Eiselen@nwu.ac.za

Abstract

Despite many attempts to address the situation, South Africa's official languages remain under-resourced in terms of the text and speech data required to implement state-of-the-art language technology. To ensure that *no language is left behind*[1], resource development should remain a priority until a strong digital presence has been established for all indigenous languages. This paper provides an overview of previous projects that were specifically aimed at speech resource development and introduces an ongoing initiative to launch South Africa's languages on the Mozilla Common Voice platform.

Keywords: DHASA, under-resourced languages, speech resources, Mozilla Common Voice

1 Introduction

South Africa's constitution recognizes eleven official languages: Afrikaans (Afr), South African English (Eng), isiNdebele (Nbl), isiXhosa (Xho), isiZulu (Zul), Sepedi (Nso), Sesotho (Sot), Setswana (Tsn), Siswati (Ssw), Tshivenda (Ven), and Xitsonga (Tso). With the exception of English and Afrikaans, all the official languages belong to the South-Eastern Bantu family. IsiNdebele, Siswati, isiXhosa and isiZulu are part of the Nguni group of languages and Sepedi, Sesotho and Setswana are part of the Sotho language group. The languages within each family are closely related, with similar orthographic and morphosyntactic attributes.

Most people in South Africa speak more than one Bantu language and English. As a result, English serves as *lingua franca* and is most frequently used in commerce and law. Some of the country's citizens have access to language and speech technology through English. For the ten remaining languages, much remains to be done to match the level of technology development that has already been achieved for languages like English. In this regard South Africa's indigenous languages are in the same position as the majority of the almost 7 000 languages that are spoken in the world today: usable language and speech technology is not readily available yet (Adda et al. 2019, Joshi et al. 2020).

Despite various projects aimed at addressing this situation, the pace of resource development in South Africa's languages has not kept up with the rate at which technology and the data requirements associated with state-of-the-art techniques have advanced. As a result, many of the latest technology, especially deep learning techniques, cannot be implemented effectively for South Africa's local languages due to

a lack of appropriate data.

This paper describes a recent initiative to launch South Africa's official languages on Mozilla's Common Voice[2] platform. The Common Voice project aims to make speech recognition technology open and accessible by creating open, high quality, publicly available data sets in as many languages as possible. For a language to achieve *launched* status, the Common Voice website needs to be localised and at least 5 000 sentences in the target language have to be collected and be available in the open domain under CCo licensing [3]. Once *launched* status is achieved, the sentences are used as prompts for speech data collection through the Mozilla Common Voice platform. While Mozilla makes the Common Voice platform freely available, they are not involved in localisation and data collection and do not provide any financial support to participants. The presence of a language on the platform is determined by language communities themselves.

2 Background

A number of projects have already contributed to the establishment of basic language resources as well as speech and text technology in South Africa's official languages. Many of these were supported by the South African Government[4].

One of the first attempts to develop technology in the country's indigenous languages, the *African Speech Technology* (AST) project, was funded by the Department of Science and Technology's Innovation Fund (Roux et al. 2004). One of the aims of the project was to prepare South Africa's languages for a digital future. It was also envisioned that language technology would facilitate multilingual information access to South Africa's citizens. Five telephone speech databases in isiXhosa, Sesotho, isiZulu, South African English and Afrikaans were developed during the course of the project. The data was transcribed orthographically as well as phonetically and used to develop a prototype version of a multilingual, telephone-based hotel booking system. A limited domain text-to-speech voice was also built for each of the five languages, allowing the system to provide dynamic although domain limited speech feedback.

Subsequent to the AST project, three *Lwazi*[5] projects were funded by the South African Department of Arts and Culture with the aim of extending the available telephone speech data sets to include all 11 official languages and to increase the impact of speech technologies in South Africa (Barnard et al. 2010, Kuun 2012, Calteaux et al. 2013, Titmus et al. 2016). Toward the latter aim, text-to-speech and speech-to-text systems were developed in all 11 languages and evaluated in applications including a voice-based telephone service for rural veterinarians and a multilingual, telephone-based interactive voice response system for the Department of Basic Education's National School Nutrition Programme.

isiZulu was included in the data sets that were collected to support IARPA/DARPA's[6] *Babel* and *LORELEI* (Low Resource Languages for Emergent Incidents) programs (Harper 2011, Strassel & Tracey 2016). These programs resulted in numerous investigations on the development of automatic speech recognition and spoken term detection capabilities in low-resource languages, many of which included isiZulu as an example language.

The National Centre for Human Language Technology (NCHLT) subsequently funded two projects to collect substantially larger speech and text data sets than those that were compiled during previous projects. The data collection efforts therefore went beyond the telephone-based, limited domain scope of the AST and Lwazi projects. This effort resulted in 11 speech corpora containing 50-60 hours of orthographically transcribed broadband speech per language and 11 text corpora of between 1.15 and 3.27 million words per language (Barnard et al. 2014, Eiselen & Puttkammer 2014). With the exception of a few domain specific data collection efforts (Davel et al. 2011, de Wet et al. 2011, de Wet et al. 2016), these remain the most extensive resources that are available for speech technology development in the country.

The majority of the projects mentioned in the previous paragraphs were associated with a specific institution in South Africa, whereas anybody from any language community (who adheres to Mozilla's code of conduct) can participate in the Common Voice project. Anybody who would like to can therefore contribute to the data collection effort and become involved in the local language technology community. The Common Voice project also has an international reach which means that the South African community stands to benefit from "lessons learnt" during localisation and data collection in other countries as well as the technical support provided on Common Voice community user groups. The rest of the paper describes how the Common Voice website was localised and presents the results of an initial attempt to harvest sentences from text data on the web.

3 Mozilla Common Voice platform

As was mentioned in Section 1, a language needs to meet two requirements to be launched on the Common Voice platform: 1) the website needs to be translated into the target language and 2) 5 000 sentences that are in the open domain need to be collected. Mozilla provides tools (via websites and community user groups) and guidelines to assist with both these processes. Their implementation in the current project is briefly described in the next two sections of the paper.

3.1 Translation

A number of service providers were requested to submit quotes for translating the Common Voice website from English into the 10 other official languages. A company with previous experience in localisation was identified as the best candidate to perform the translations. All translations were performed by the same company so that they could manage aspects like the standardisation of terminology between languages in the same manner for all languages.

Words and utterances were translated using Mozilla's translation tool, Pontoon[7], before being used to generate language specific web pages automatically. The resulting web pages were subsequently proofread by a second team of linguists. The feedback they provided ranged from remarks on lexical choice based on differences in intra-lingual geographical variation between the dialects of the translators and proofreaders (despite both parties being native speakers of the language) to the way in which Mozilla's technology utilized words to create automatic translations without taking the morphological makeup of a word into account. Amongst other things their comments indicated that Mozilla's tools do not make provision for the noun class agreement system in the Nguni languages, resulting in words translated in isolation not appearing correctly in sentences. In some cases the tools did not accommodate the length of words in more descriptive sentences where words are more morphologically complex.

Translators and proofreaders also found it difficult to perform the localisation because many of the languages do not currently have words for technological terminology such as *part-of-speech tagger* or *sentence builder* resulting in the choice between creating or establishing terminology which might alienate potential users of the website, or using existing English terminology which might create the perception that the entire website has not been localised.

3.2 Text collection

Although a number of curated text corpora collections already exist for all of the South African languages, there are several complications to using these corpora as example sentences for the Mozilla Common Voice project. Firstly, almost all of these corpora are distributed under CC-BY licenses, similar to those used by open source initiatives such as Wikipedia. This implies that only subsections of these data sets (typically less than 10% of the original article) qualify as CC0. Secondly, many of these data sets are either sourced from government documents, speeches, and websites, or from religious ma-

terial, such as the Bible. These texts represent very specific domains, subject matter, as well as writing style.

To mitigate these problems, we investigated the possibility of sourcing text data from other web sources that are commonly used in language technology development. For instance, Mozilla provides a set of text processing tools to harvest data from Wikipedia. The tools could not be used "as is" in this project, because the default English rule set only allows sentences with ASCII characters. However, most of the South African languages include diacritic markers encoded by UTF-8 characters. The rule set therefore had to be adapted to accept within-language UTF-8 characters but to reject irrelevant ones.

One of the Mozilla selection rules specifies that only three sentences may be copied from a Wikipedia page, but only if the article contains 10 or more sentences. Many articles in South African languages did not meet the 10-sentence limit and, as a result, no sentences could be harvested from them. Another limitation that became evident is the lack of lists of "disallowed words" in most of the languages. These lists are used to prevent possibly offensive words from appearing in the sentences. The text collected from Wikipedia was also verified using automatic Language Identification (LID) (Puttkammer et al. 2018, Hocking 2014)[8]. The verification revealed that many articles contain text in languages other than the target language. These sentences were discarded.

Although the Mozilla tools and Wikipedia could be used to obtain some data in a few languages, the current Wikipedia presence of the majority of the languages yielded less than a thousand sentences per language. Moreover, the text collection did not produce any isiNdebele sentences because, at the time of writing, the language did not have a presence on Wikipedia.

Other sources of web-based text were subsequently explored in an attempt to collect isiNdebele sentences as well as additional data for the other languages. These included the Leipzig Corpora Collection (LCC) (Goldhahn et al. 2012), OPUS (Tiedemann & Nygaard 2004) and the FLORES-200 (Goyal et al. 2022) data sets. The Mozilla tools were also used to collect sentences from these sources, but with the restriction that no more than 9.5% of any particular source was allowed to be harvested. The resulting selections were also verified using LID and the same rules for discarding unwanted characters were applied.

In addition to these pre-processing steps, the text was sentence separated and frequency lists were generated using CTexTools 2 (Puttkammer et al. 2018)[9]. Sentences or segments that did not include useful data (e.g. lines containing only telephone numbers or punctuation) as well as lines that did not constitute a well formed sentence (starting with optional punctuation or numbering, then a capital letter and ending with sentence ending punctuation) were removed from the sentence separated data. The sentences were also filtered to contain between three and fourteen words but with an absolute character limit of 99 [10]. Sentences including numerals were also removed according to the guidelines [11].

The frequency lists were then spell checked using commercially available spelling checkers [12] developed by the Centre for Text Technology at the North-West University in South Africa [13]. Using the spell checked lists, all remaining sentences were ranked according to the percentage correctly spelled words they contain and only sentences with more than 80% correctly spelled words were kept. To ensure better coverage of the languages, these sentences were then compared using the Levenshtein edit distance [14]. Only sentences with less than 70% overlap were included in the final set for each language.

After completing the above mentioned steps, only the Afrikaans and Setswana texts still contained more than 5 000 sentences. This was partially due to the fact that the initial corpora were relatively small. Another contributing factor is an overlap of up to 80% between the web-based corpora like LCC and OPUS. This observation seems to suggest that the

text collections were probably obtained form the same sources.

Searching for some of the terms in the Afrikaans list of disallowed words revealed that adding this type of filtering is essential to prevent offensive words and sentences from appearing in the sentences. A similar process also indicated that there is a strong presence of religious text in the harvested data, despite a concerted effort to avoid religious and government publications. Appropriate filters for these types of texts will therefore also have to be designed for each of the 10 languages under consideration before adding any text harvested from the web to the Common Voice platform.

4 Future work

Immediate next steps in the project will be to address the issues discussed in the previous section in order to reach the target of 5 000 CCo sentences per language. The South African governmental websites (*.gov.za) appear in all the official languages and the possibility to obtain additional text from this source will be investigated. Once the required number of sentences have been collected, the Common Voice websites will be ready for speech data collection to start. The project will be promoted as widely as possible with the aim to encourage language communities across the country to become involved. Hopefully these efforts will be successful to the extent that data collection can be followed by dedicated speech technology development workshops.

Notes

[1] https://www.undp.org/sustainable-development-goals, https://odi.org/en/publications/leave-no-one-behind-index-2019/

[2] https://commonvoice.mozilla.org/en

[3] https://creativecommons.org/share-your-work/public-domain/cc0/

[4] All resources that were generated with government grants are made freely available and are accessible via the South African Centre for Digital Language Resources' Resource Catalogue: https://repo.sadilar.org/.

[5] In the Nguni languages spoken in South Africa *lwazi* means knowledge or information.

[6] Intelligence Advanced Research Projects Activity/Defense Advanced Research Projects Agency

[7] https://pontoon.mozilla.org/projects/common-voice/

[8] https://hdl.handle.net/20.500.12185/350

[9] https://hdl.handle.net/20.500.12185/480

[10] https://discourse.mozilla.org/t/using-the-europarl-dataset-with-sentences-from-speeches-from-the-european-parliament/50184

[11] https://commonvoice.mozilla.org/sentence-collector/##/en/how-to

[12] https://spel.co.za/en/product/african_spelling_checkers/

[13] https://humanities.nwu.ac.za/ctext

[14] https://metacpan.org/pod/Text::LevenshteinXS

Acknowledgements

The localisation of the Mozilla Common Voice platform in South Africa is supported by the German Federal Ministry of Economic Cooperation and Development (BMZ), represented by the GIZ project FAIR Forward - Artificial Intelligence for All.

References

Adda, G., Choukri, K., Kasinskaite, I., Mariani, J., Mazo, H. & Sakriani, S., eds (2019), *Proceedings of the 1st International Conference on Language Technologies for All*, European Language Resources Association (ELRA), Paris, France.

Barnard, E., Davel, M. H., van Heerden, C., de Wet, F. & Badenhorst, J. (2014), The NCHLT Speech Corpus of the South African languages, *in* 'Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages', St Petersburg, Russia, pp. 194–200.

Barnard, E., Davel, M. H. & Van Huyssteen, G. B. (2010), Speech Technology for Information Access: A South African Case Study, *in* 'AAAI Spring Symposium Series', pp. 8–13.

Calteaux, K., de Wet, F., Moors, C., van Niekerk, D., McAlister, B., Sharma-Grover, A., Reid, T., Davel, M., Barnard, E. & van Heerden, C. (2013), Lwazi II Final Report: Increasing the impact of speech technologies in South Africa, Technical report, CSIR.

Davel, M. H., van Heerden, C., Kleynhans, N. & Barnard, E. (2011), Efficient harvesting of Internet audio for resource-scarce ASR, *in* 'Proceedings of Interspeech', International Speech Communication Association (ISCA), Florence, Italy, pp. 3153–3156.

de Wet, F., Badenhorst, J. & Modipa, T. (2016), 'Developing Speech Resources from Parliamentary Data for South African English', *Procedia Computer Science* **81**, 45–52.

de Wet, F., de Waal, A. & van Huyssteen, G. B. (2011), Developing a broadband automatic speech recognition system for Afrikaans, *in* 'Proceedings of Interspeech', International Speech Communication Association (ISCA), Florence, Italy, pp. 3185–3188.

Eiselen, R. & Puttkammer, M. J. (2014), Developing Text Resources for Ten South African Languages, *in* 'Proceedings of Language Resource and Evaluation (LREC)', Reykjavik, Iceland, pp. 3698–3703.

Goldhahn, D., Eckart, T. & Quasthoff, U. (2012), Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages, *in* 'Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)'.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F. & Fan, A. (2022), 'The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation', *Transactions of the Association for Computational Linguistics* **10**, 522–538.

Harper, M. P. (2011), 'Data Resources to Support the Babel Program Intelligence Advanced Research Projects Activity (IARPA)'.

Hocking, J. (2014), Language identification for South African languages., *in* 'Proceedings of the Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)', PRASA.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. (2020), The State and Fate of Linguistic Diversity and Inclusion in the NLP World, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 6282–6293.

Kuun, C. (2012), Development of a telephone-based speech-driven information service for the

South African Government, Technical report, CSIR.

Puttkammer, M., Eiselen, R., Hocking, J. & Koen, F. (2018), NLP Web Services for Resource-Scarce Languages, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)', Association for Computational Linguistics, pp. 43–49.

Roux, J. C., Louw, P. H. & Niesler, T. (2004), The African Speech Technology Project: An Assessment, *in* 'Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)', European Language Resources Association (ELRA), Lisbon, Portugal, pp. 93–96.

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)', European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3273–3280.

Tiedemann, J. & Nygaard, L. (2004), The OPUS corpus - parallel and free: `http://logos.uio.no/opus`, *in* 'Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)', European Language Resources Association (ELRA), Lisbon, Portugal.

Titmus, N., Schlünz, G. I., Louw, J. A., Moodley, A., Reid, T. & Calteaux, K. (2016), Lwazi III Project Final Report: Operational Deployment of Indigenous Text-to-Speech Systems, Technical report, CSIR.