Investigating the extent and usability of webtext available in South Africa's official languages*

Febe de Wet[1][0000−0003−3495−9802], Roald Eiselen[2][0000−0002−8612−5175], Erwin Schillack[1][0009−0006−2289−4131], and Martin Puttkammer[2][0000−0002−0493−3671]

[1] Department of Electrical and Electronic Engineering, Stellenbosch University, Stellenbosch, South Africa
{fdw, 22145540}@sun.ac.za

[2] Centre for Text Technology, North-West University, Potchefstroom, South Africa
{Roald.Eiselen, Martin.Puttkammer}@nwu.ac.za

Abstract. Large collections of text data are a central part of many aspects of natural language processing (NLP) development, and the availability of such data is a prerequisite for advances in the field. With the proliferation of very large, multi-lingual, web-based data sets over the last decade, it is often overlooked that there are still many languages that are not well represented in these data sets, and for which collecting even moderately sized data sets remains a challenge. Furthermore, a lack of data for a language in one or more of the core data sources, such as Wikipedia, often leads to such languages being further excluded in other collections of web-based data or having a detrimental effect on the quality of the data collected. The systematic review and investigation of the quality of these data sets are relatively limited, and the quality of the data is primarily extrinsically evaluated by measuring the improvements on downstream tasks, rather than implicitly evaluating the data. This paper reports on some of the text data that is currently available for South Africa's official languages (except for English and South African Sign Language) in various widely available web-derived corpora. The aim of the study was to harvest text from the web that could be used as prompts on the Mozilla Common Voice Platform. Towards this aim the extent of the resources available in each language as well as the degree of overlap between different sources were quantified. Results show that there remain several South African languages for which web-based corpora are still severely limited and that, for languages with at least some web presence, the majority of the text is of disputable quality.

Keywords: Under-resourced languages · text data · web text · South African languages · Mozilla Common Voice Platform.

* Supported by the German Federal Ministry of Economic Cooperation and Development (BMZ), represented by the GIZ project FAIR Forward - Artificial Intelligence for All.