Evaluation of the Effects of a Spellchecker on the Intellectualisation of IsiZulu

C. Maria Keet
Langa Khumalo

Abstract

Through its bilingual language policy and plan that recognises English and isiZulu as official languages of the University of KwaZulu-Natal (UKZN), UKZN has aggressively promoted the intellectualisation of isiZulu as an effective strategy in advancing indigenous, under-resourced African languages as vehicles for innovation, science, and technology research in Higher Education and Training institutions. UKZN recently launched human language technologies (HLTs) in isiZulu as enablers towards the intellectualisation of the language. One of these is an isiZulu spellchecker, which was trained on an organic isiZulu National Corpus. We evaluate the isiZulu spellchecker's effects on the intellectualisation of isiZulu. Two surveys were conducted with the target end-users, consisting of relevant questions and the System Usability Scale, and an analysis of words added to the spellchecker. It is evident that the spellchecker has had a positive impact on the work of target end-users, who also perceive it as an enabler in the intellectualisation of isiZulu. The survey responses show modest success for a first version of the tool. The analysis of the words added to the spellchecker indicates that new words are being added to the isiZulu lexicon.

Keywords: spellchecker; intellectualisation; HLTs; survey; evaluation; lexicon

Electronic ISSN: 2519-5476; DOI: https://doi.org/10.29086/2519-5476/2017/v24n2a5

Introduction

The launch of the isiZulu spellchecker is part of UKZN's broad programme of advancing the isiZulu language to be a language of science, research, teaching and learning. A screenshot of the tool is presented in Figure 1. Its launch was part of UKZN's strategy of launching other technologies such as the Zulu Lexicon mobile-compatible application (Android and iPhone); the isiZulu Term Bank; the isiZulu National Corpus with 20.5 million tokens, and two isiZulu books, an anthology of short-stories and the first bilingual (English-isiZulu) illustrated glossary of Architectural Terms. The launch of the isiZulu spellchecker in particular raised interest among the end-user target group comprising journalists, newspaper editors, and academics. We investigate and evaluate its impact, noting that its accuracy and comparison with other spellcheckers have been assessed elsewhere (Ndaba *et al.* 2016). The evaluation seeks to answer several questions specifically related to the spellchecker itself as well as its potential to contribute to the intellectualisation of an under-resourced language such as isiZulu. In particular, we seek to answer the following high-level questions:

1. Is the spellchecker meeting end-user needs and expectations?
2. Is the spellchecker enabling the intellectualisation of the language?
3. Is the lexicon growing upon using the spellchecker?

The questions will be answered with a two-pronged approach, using data from questionnaires among the expected user base of the current version of the isiZulu spellchecker, which broadly includes academia and industry, and a linguistic analysis of its use regarding the words added to the spellchecker by the users as a possible proxy for intellectualisation. The main outcomes are that the isiZulu spellchecker is perceived to have a positive effect on the intellectualisation of this language, which is also supported by the analysis of the user-added words. The tool has been received positively by the target audience, with suggestions made for more functionalities. Its use also indicated that a few additional rules could increase its accuracy.

In the remainder of the paper, we first discuss related works, after which we describe the set-up of the evaluation and present the results and discuss them. We conclude in Section 6.

Related Works

As the scope of the paper is intellectualisation of a language through human language technologies, we discuss the state of the art of both components in this section.

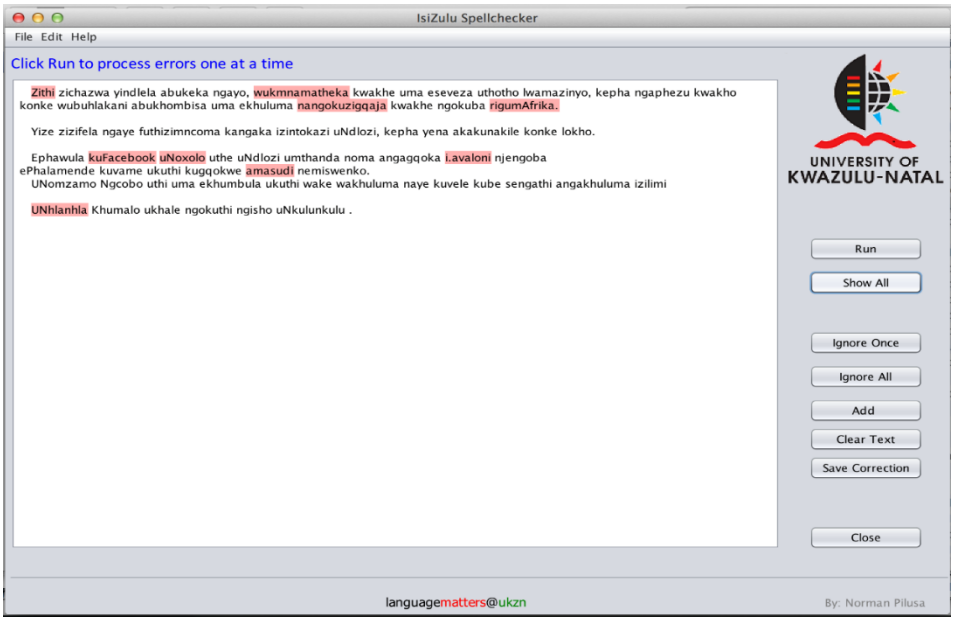


Figure 1: Screenshot of the isiZulu spellchecker, highlighting all words that it deems likely to be misspelt. It also has an isiZulu interface, which is included in Appendix B.

Intellectualisation of Languages

Intellectualisation is a term originally used by Havránek (1932), a linguist from the Prague School, to characterise a process that a language undergoes in its advancement.

> By the intellectualization of the standard language, which we could also call its *rationalization*, we understand its adaptation to the goal of making possible precise and rigorous, if necessary abstract,

> statements, capable of expressing the *continuity* and *complexity of thought*, that is, to reinforce the intellectual side of speech. This intellectualization culminates in *scientific* (theoretical) *speech*, determined by the attempt to be as *precise in expression* as possible, to make statements, which reflect the rigor of *objective* (scientific) *thinking* in which the terms approximate concepts and the sentences approximate *logical judgements* (e.a.) (Havránek 1932: 32).

Intellectualisation is thus a clear process of (functionally) cultivating a language so that its terminology can carry the full weight of scientific rigour and precision, and its sentences can accurately express logical judgements, resulting in a language that has the capacity to function in all domains. As the direct consequence of intellectualisation, speakers of the language derive pride, self-assurance and resourcefulness from their (new) ability to discuss the most complex of issues ranging from the mundane to the academic and beyond (Khumalo 2017).

Intellectualisation has been famously associated with the development of Tagalog in the Philippines. The cultivation process involved Tagalog's lexical enrichment through terminology to enable its use in academia. Philippine linguists and sociolinguists are recognised by Neville Alexander (in Busch, Busch and Press 2014) as the doyens in the scholarship of intellectualisation. Sibayan (1999: 229) characterises an intellectualised language as one '[…] which can be used for educating a person in any field of knowledge from kindergarten to the university and beyond' (Sibayan 1999: 229). Thus, an intellectualised language has the capacity to discuss any issue regardless of its complexity. According to Finlayson and Madiba (2002), in the South African context intellectualisation is a meticulous procedure aimed at expediting the growth and development of hitherto underdeveloped African languages to augment their capacity to effectively interface with modern developments, theories, and concepts. It is imperative to note that crucial to this process is the capacity to interface with technology and the general digital visibility of these under-resourced indigenous languages. The paucity of such technology and technical terminology is often cited as the reason why African languages cannot be used as languages of teaching and learning; hence their discernment as shallow and inadequate (cf. Shizha 2012).

Intellectualisation in our context thus means the radical transformation of the capacity, role, and digital and/or technological presence of indigenous

African languages in carrying and conveying all forms of knowledge in all spheres of life. While the government through the Constitution of South Africa (RSA 1996: section 6) has expressed commitment '*to elevate the status and advance the use of'* these hitherto underdeveloped and under-resourced languages, very little has actually been done to improve their status and role in Higher Education (cf. Olivier 2014). The debate on the status and role of these languages has been sharply brought back to the centre of South African Higher Education through the #FeesMustFall campaign. UKZN has thus taken the lead in the intellectualisation of isiZulu through the development of HLTs such as the isiZulu spellchecker.

Human Language Technologies

Human language technologies for isiZulu are sparse and mostly remain in the realm of theory and academic proof-of-concept tools, such as a morphological analyser (Pretorius and Bosch, 2003), machine translation (Kotzé & Wolff 2015), search engines (Malumba *et al.* 2015), and knowledge-to-text natural language generation (Keet & Khumalo 2017). The main drivers for end-user tools at present are large multinational companies, such as Google Inc. with its rudimentary GoogleTranslate for isiZulu and the localisation efforts of its search engine interface (at no monetary cost), and Microsoft's isiZulu localisation as a for-payment localisation extension/plugin. To the best of our knowledge, there are no isiZulu equivalents of widely-used end-user features such as autocomplete, spellcheckers and grammar checkers, or an isiZulu language-sensitive 'desktop document search' such as Apple's 'spotlight'.

While efforts have been documented to develop spellcheckers (Prinsloo & de Schryver 2004; de Schryver & Prinsloo 2004; Bosch & Eiselen 2005), these tools are not available. The plugins for OpenOffice, Firefox, and Thunderbird – developed by translate.org.za in 2008 – are freely available, but they have not been updated so they no longer work with the latest versions (since OpenOffice v4.x). To the best of our knowledge, no user studies on the usability or impact of isiZulu spellcheckers have been conducted.

The isiZulu spellchecker used for the experimental evaluation takes a different approach from those earlier works that relied on word lists and grammar rules. Instead, this spellchecker is based on a statistical language model learnt from a sample of the isiZulu National Corpus (INC) (Khumalo 2015) and reports (or not) a word as misspelt based on the *probability* of it

being a mistake (Ndaba *et al.* 2016). Let us illustrate the idea with a small example, as the underlying technology may affect user satisfaction in either direction. Let us assume that the spellchecker's model is trained on three words only: *sivela*, *ngihamba*, and *uvelaphi*. The algorithm first produces 16 trigrams *siv*, *ive*, *vel*, *ela*, *ngi*, *gih*, *iha*, *ham*, *amb*, …, *phi*, of which 14 are unique. It then includes those trigrams that are used most often and discards the others that are assumed to be erroneous. For this example, *vel* and *ela* are used most often yet none is hardly used (say, less than 1%), so let's assume our statistical model includes all the different trigrams of these three words. If a user were to type *ngivela* in the spellchecker, i.e., a string that it has not been trained with, it will compute it to be a very probably correct word, because all of *ngivela*'s trigrams are in the list of valid trigrams. If a user were to type *ngivella*, the spellchecker would flag it as incorrect, because there is no trigram *ell* or *lla*. The actual statistical language model of the isiZulu spellchecker was trained not with three words but with a sample of the INC and uses a cut-off threshold of 0.0003 for valid trigrams; i.e., any trigram that has a lower probability of occurring than the threshold is discarded as being invalid. This means that such three consecutive characters are so unusual in the training texts, that it is assumed to be violating isiZulu orthography rules and would thus be wrong. Thus, the spellchecker flags or accepts a word as (in)correct based on *probabilities* of correctness, *not* on *certainty* of encoded grammar or curated word list. The quality of the language model, and thus the spellchecker's performance, depends on the size and quality of the corpus it is trained on and likely datedness and genre as well, as observed in Ndaba *et al*. (2016). The sample of the INC that was used for training the model included both novels and news articles.

Human language technologies for isiZulu are sparse and mostly remain in the realm of theory and academic proof-of-concept tools, such as a morphological analyser (Pretorius and Bosch, 2003), machine translation (Kotzé & Wolff 2015), search engines (Malumba *et al.* 2015), and knowledge-to-text natural language generation (Keet & Khumalo 2017). The main drivers for end-user tools at present are large multinational companies, such as Google Inc. with its rudimentary GoogleTranslate for isiZulu and the localisation efforts of its search engine interface (at no monetary cost), and Microsoft's isiZulu localisation as a for-payment localisation extension/plugin. To the best of our knowledge, there are no isiZulu equivalents of widely-used end-user features such as autocomplete, spellcheckers and grammar checkers, or an

isiZulu language-sensitive 'desktop document search' such as Apple's 'spotlight'.

While efforts have been documented to develop spellcheckers (please see especially Prinsloo & de Schryver 2004; de Schryver & Prinsloo 2004; Bosch & Eiselen 2005), these tools are not available. The plugins for OpenOffice, Firefox, and Thunderbird – developed by translate.org.za in 2008 – are freely available, but they have not been updated so they no longer work with the latest versions (since OpenOffice v4.x). To the best of our knowledge, no user studies on the usability or impact of isiZulu spellcheckers have been conducted.

The isiZulu spellchecker used for the experimental evaluation takes a different approach from those earlier works that relied on word lists and grammar rules. Instead, this spellchecker is based on a statistical language model learnt from a sample of the isiZulu National Corpus (INC) (Khumalo 2015) and reports (or not) a word as misspelt based on the *probability* of it being a mistake (Ndaba *et al.* 2016). Let us illustrate the idea with a small example, as the underlying technology may affect user satisfaction in either direction. Let us assume that the spellchecker's model is trained on three words only: *sivela*, *ngihamba*, and *uvelaphi*. The algorithm first produces 16 trigrams *siv*, *ive*, *vel*, *ela*, *ngi*, *gih*, *iha*, *ham*, *amb*, …, *phi*, of which 14 are unique. It then includes those trigrams that are used most often and discards the others that are assumed to be erroneous. For this example, *vel* and *ela* are used most often yet none is hardly used (say, less than 1%), so let's assume our statistical model includes all the different trigrams of these three words. If a user were to type *ngivela* in the spellchecker, i.e., a string that it has not been trained with, it will compute it to be a very probably correct word, because all of *ngivela*'s trigrams are in the list of valid trigrams. If a user were to type *ngivella*, the spellchecker would flag it as incorrect, because there is no trigram *ell* or *lla*. The actual statistical language model of the isiZulu spellchecker was trained not with three words but with a sample of the INC and uses a cut-off threshold of 0.0003 for valid trigrams; i.e., any trigram that has a lower probability of occurring than the threshold is discarded as being invalid. This means that such three consecutive characters are so unusual in the training texts, that it is assumed to be violating isiZulu orthography rules and would thus be wrong. Thus, the spellchecker flags or accepts a word as (in)correct based on *probabilities* of correctness, *not* on *certainty* of encoded grammar or curated word list. The quality of the language model, and thus the spellchecker's

performance, depends on the size and quality of the corpus it is trained on and likely datedness and genre as well, as observed in Ndaba *et al*. (2016). The sample of the INC that was used for training the model included both novels and news articles.

Materials and Methods

The aim of the evaluation is to seek an answer to the main questions posed in the Introduction: *Is the spellchecker enabling the intellectualisation of the language?* This had two sub-questions that address linguists and users' opinions about the spellchecker, and the materials and methods for the evaluations are split accordingly.

Methods

The method for obtaining data to answer the first two questions posed in the introduction is, by design, mostly quantitative, with a further qualitative follow-up, depending on the results of the quantitative part. First, we devise a questionnaire that also includes open questions (i.e., not 'yes/no') so as to obtain as much open-ended feedback as possible, and administer the System Usability Scale (SUS) questionnaire (Brooke 1996) in the same survey. The questions for the first part of the questionnaire are included in Appendix A and mainly focus on feature usage and wishes, use, and opinions on intellectualisation. The SUS questionnaire (Brooke 1996) is a widely-used quick survey consisting of 10 questions to be answered on a 5-point Likert scale. The values are added up by even and odd numbered questions, and multiplied by 2.5 to obtain a value between 1 and 100. This value is a rough indicator of user-friendliness and the usability of a system's interface and enables determination of whether the usability of the tool might have had an adverse effect on its use and users' perceptions of the tool. The questions are of the type 'I think that I would like to use this system frequently' and 'I found the system unnecessarily complex'. In line with the context of the evaluation, these questions have been translated into isiZulu and included in Appendix A as a record for future use.

Based on the results obtained, a follow-up in-depth qualitative evaluation is designed in the form of a semi-structured interview with industry

stakeholders, which may reveal further contextual information about the effects of the spellchecker. The prepared interview questions are included in Appendix A. The qualitative analysis was done by means of a manual assessment of the responses.

The method pursued in order to obtain results so as to answer Question 3—*Is the lexicon growing upon using the tool?*—can be refined into two parts, where each aims to answer a sub-question, being:

i. What is the percentage of user-added words that are 'normal' (already in the dictionary) words, cf. the new words?
ii. If words are added that are not in an isiZulu dictionary, do those user-added new words follow the canonical structure or are they import words?

The method used to answer this sub-question is principally from a qualitative and linguistic perspective. First, we obtain the 'user dictionary' file of a set of participants. This plaintext file is absent upon downloading the tool, but is created in the same directory once a first word is added and is appended to each time the user clicks the 'Add' [to dictionary] button. These user dictionary files are analysed by first gathering basic descriptive data, such as aggregate data by recording how many words have been added per user, the average and median number of words added by users. Subsequently, the words will be annotated to identify what type of words are being added that were not recognised by the spellchecker, including, but not limited to: a normal isiZulu word (that the spellchecker ought to have recognised) or a new word with deviating orthography and whether it is a proper noun (a named entity, like 'Facebook'), a current abbreviation (e.g., 'EFF'), and so on. The 'new' words, if any, are analysed in terms of whether they are canonical or import, and similar.

Target Demographic and Recruitment

The target audience of version 1 of the spellchecker was people who may write isiZulu regularly or on a daily basis and do so on their desktop or laptop computer for work or study purposes. This entails that participants in the evaluation are all adults and likely will have enjoyed at least a medium-level (secondary school), if not higher (university) level, of education. While gender

and age is relevant for the evaluation of some software applications[1], spellcheckers are generally widely used; therefore, these variables are not taken into consideration as a relevant dimension of analysis.

Participant recruitment was planned as follows. It would occur in a group email invitation by one of the authors, which includes students, administrators, and academics at UKZN, and the newspaper editors and journalists of isiZulu newspapers that contributed to the isiZulu National Corpus. If, after one week, less than 25% of the invitees would have filled in the survey, a reminder email would be sent.

Participation is anonymous and on a voluntary basis without remuneration or thank-you vouchers.

Materials

The materials consist of version 1 of the isiZulu spellchecker that was launched on 10 November 2016, the user dictionary files collected from participants, the questionnaires in isiZulu, and a partially localised version of the Limesurvey software. While the Limesurvey localisation is ongoing and has a few typographical errors, this is nonetheless preferred, so that the participant experience is in the facilitating context of enabling technologies for the language of focus. The survey is accessible at http://limesurvey.cs.ukzn.ac.za/index.php?sid=38664&lang=zu. Analysis was carried out in Microsoft Excel.

Results

The isiZulu spellchecker has been downloaded 159 times (as at 14th February 2017) since the 10th November 2016. The authors have received some questions regarding its installation, especially from industry, due to restrictive security setting on installing software downloaded from the Web. This issue is a general one regardless of the technology that has been used to implement the spellchecker. Nonetheless, it may have affected its successful deployment.

[1] For instance, when assessing social media use or games.

Questionnaire Results and Discussion

The survey was open for data collection for 2.5 weeks in late January/early February 2017. A total of 59 people had been invited to participate in the survey, of whom 34 are students and staff (administrators and academics) from UKZN, and the other 25 are from industry. After 1.5 weeks, there were five completed surveys, so a reminder was sent to all original invitees. At 2.5 weeks, there were 11 completed surveys, which were used in the analysis, noting an additional 26 'incomplete' surveys that were fully empty, i.e., the webpage was only opened, and therefore not further considered. These figures amount to an RR1 of 19% and an RR2 of 63%, which is roughly within the expectations of survey response rates with respect to the invitees.

Survey Responses on Features

A brief summary of the responses to the open questions is presented in Table 1 and these are discussed in more detail in the remainder of this section. The answers should be seen in light of the fact that five participants said that they rarely used the tool, while two said that they used it weekly, another two said that they tried it once, and one participant said that s/he uses it every day.

A clear majority of the survey participants indicated that the entire tool was helpful, in that it checks isiZulu words for correctness and allows one to add new isiZulu words that the tool does not recognise. Six participants were of the view that the spellchecker assists in checking, editing and validating spelling in isiZulu, two indicated that it assists in highlighting words that are not acceptable in isiZulu, one indicated that it helps in not only editing his/her work but also in adding words that are not currently in the lexicon, and another participant indicated that all the different functionalities of the tool are useful. Those who answered the second part of the question, i.e., to give examples of how the tool is helpful to them, responded as follows (answers translated by authors):

1) 'The tool helps one a lot when editing one's work.'
2) 'The tool helps in highlighting words that are not acceptable in isiZulu.'
3) 'The tool validates one's spelling.'
4) 'The tool is easy to use as navigation is easier.'

5) 'The tool helps the user to check if the work that one has just completed does not have errors.'

Note that two respondents (1 and 5) reveal *where* in the work activity the tool is making a difference.

Question	Top choice among options	Nr top choice
1 – most useful feature	'Entire tool helpful'	7
2 – useless feature	'None' (see text for details)	4
3 – add features?	'None' (see text for feature requests)	4
4 – remove features?	'None'	10
5 – intellect. enabler?	'Yes'	11
6 – enhanced work?	'Yes'	8
7 – plugin to which tool?	Chrome plugin was rated highest	N/A
8 – usage frequency	'Rarely'	5 (i.e., 6 use it more frequently)

Table 1: Summary of responses to the first part of the survey questions; only the number (n) of the top-choices are listed; see Appendix A for the full formulation in English and isiZulu

While four respondents indicated that there is no 'useless' feature (Question 2), other useful feedback was obtained in this comment field. One participant indicated that the spellchecker did not find some real and authentic isiZulu words and another indicated that the tool indicates that some isiZulu words are in fact not correct. This is expected as the spellchecker has around 90% accuracy (Ndaba *et al.* 2016); no spellchecker achieves 100% accuracy. Another participant indicated that the tool does not provide any suggestions for possible words that they were trying to spell. Spelling correction is indeed not a feature of v1, because there is as yet no algorithm for this function. It was also reported that there is no functionality for saving and storing the corrected word after correcting the misspelt one (although that functionality is available). The last participant indicated that some instructions are confusing; for instance, the instruction to switch language from isiZulu to English is vague.

In terms of whether there are certain functionalities that users want added to the tool, four survey participants indicated that they did not feel

anything needed to be added. The feature requests by the other respondents were that:

1) the tool should be made compatible with MS Word and other mobile phone applications, and also have predictive text functionality and autocorrect;
2) voice recognition for a voice search (two respondents);
3) recognise antonyms and synonyms;
4) the tool must be populated with more isiZulu words so that it recognises most words in this language;
5) the F1 help function should be translated to isiZulu, as it is currently only in English.

From a purely scientific and technological viewpoint on the state of the art of HLTs for isiZulu, autocorrect, predictive text, voice recognition, voice search, and recognising antonyms and synonyms either have yet to be investigated or are not deployment-ready. Points 4 and 5 can be achieved. Compatibility with MS Word is problematic, because that software is 'closed source', i.e., it depends on Microsoft's willingness to add a spellchecker for isiZulu to their software.

In responding to the question on whether there is any functionality that the users wish to be removed from the tool, 10 of the 11 survey participants indicated that there are none. One participant indicated that the tool's accuracy should be enhanced (i.e., the capacity to accept (or reject) isiZulu words, because it sometimes accepts a misspelt word and sometimes rejects as incorrect a correctly spelt word).

All the survey participants felt strongly that the spellchecker has the effect of developing isiZulu as a language of teaching and learning. In response to the related question, how the tool has improved one's work in language, eight of the 11 survey participants indicated that it has improved their work, particularly in translation work, editing, and in validating spellings. One participant indicated that the tool needs improvement, and another indicated that they had not used the tool sufficiently to respond adequately.

Because this is now a standalone tool and we were not sure whether or not to develop a plugin, we asked the survey participants to rank their preference in the order of 1-4 on a plugin for OpenOffice (free, and open source office applications), Thunderbird (open source email programme), Firefox (open

source Web browser, several platforms), and Chrome (Google's Web browser) (question 7 of the survey). Six respondents did not adhere to the ranking instructions, such as allocating 10 points twice and 0 twice, or did not provide a strict order (e.g., 4 four times, or 1, 1, 3, 4). Therefore, instead of calculating on the basis of the supposed theoretical maximum of 44 an option could receive, we added up all the values (124) and computed the ratio of the total points assigned to an option. The Chrome plugin was rated highest (0.37), followed by OpenOffice (0.29), Firefox (0.20), and finally Thunderbird (0.14).

SUS Evaluation

The SUS score was computed over the 11 completed surveys of the tool and averages 75, with a median of 82.5 (in a range of 45 and 100). Considering the natural language interpretations of that (Bangor *et al.* 2009), also depicted in Figure 2, the spellchecker tool's usability is considered 'good'. This suggests that if there is any limited impact on the intellectualisation of isiZulu by means of the spellchecker, this is not attributable to the tool's interface design and, vice versa, if there is a relatively major impact on intellectualisation, this would also not be fully attributable to the tooling. Put differently, any effect observed—as described in the previous section—is an effect of the spellchecking feature, not the particular implementation.

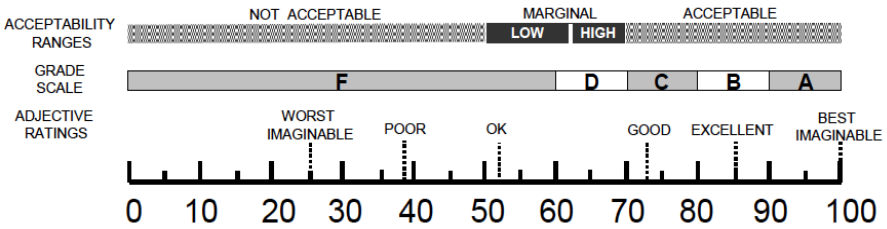


Figure 2: Descriptive interpretations of SUS scores (Source: Bangor *et al.* 2009).

Follow-up Interviews with Stakeholders

We intended to interview two industry respondents who work as editors of the two leading isiZulu newspapers in Durban. We could only interview one. The editor indicated that he uses the spellchecker every day in his line of work and

that, he finds it easy to use, and uses the spellchecker independently. However, he is not able to add words that are highlighted as not recognised to the checker's lexicon. He said that he did not have sufficient experience to comment on whether he has confidence in using the spellchecker. In responding to the question on whether there are any features that he wants to be added to the tool, he indicated that he would want grammar correction to be part of the spellchecker. He indicated that he would highly recommend the spellchecker to people in the same line of work. He noted that the spellchecker saves him time and makes his work as an editor much easier. Notably, he could not comment on whether the isiZulu spellchecker is an enabler of the intellectualisation of isiZulu, indicating that he does not fully understand what intellectualisation entails, despite having been provided with a brief explanation.

The Spellchecker's Trigrams and Lexicon

We managed to receive three user dictionaries, of which one was from an industry participant and two from within UKZN. This number is lower than anticipated, as it required more instructions to users who are not technologists. The three dictionaries had 1255, 4, and 198 entries. These entries were split successively into types, which are summarised in Table 2 and discussed in the remainder of this section.

Additions of the first type are those that the spellchecker did not recognise as correct due to 'oversensitivity' of surrounding text and context, such as the string '*ibidlala*.', whereas the tool should process it without the period, and the capitalisation at the start of the sentence, which could be lowercased internally for processing so that it is not flagged as incorrect. These are relatively simple to correct in the tool. The second type of addition is, in a way, also a 'mechanics' issue, as the statistical language model needs to detect sufficient trigrams from terms that have a within-word capital letter, such as *amaZulu*, i.e., that the trigrams *maZ*, *aZu*, and *Zul* make it above the threshold, and likewise for the other cases. Currently, it does recognise correctly a subset of such patterns only (e.g. *eGoli*). Related to these are capitalised words, such as *ABANTU*. Treating all valid and invalid capitalisations will require additional rules to augment the statistical language model. Words with dashes (e.g. *ze-PHD*, and numbers like *ngu-40*) will also be hard to learn from a dataset and may be better served by an additional rule.

Type	Examples	Cause
1. Tool (basics)	*emgonqweni*., *Iqhikiza*	The spellchecker takes whole strings, including '.' and ',' rather than without, and does not process sentence beginning (word with first letter capitalised)
2.Recognition – rule (Capitalisation)	*amaZulu, uKhisimuzi*	The training set was too small to learn all the valid within-word capital letters in trigrams
3.Recognition – rule (Dashed words)	*wase-UKZN, abanga-22*	Arbitrary compounds and numbers
4. Font size	*iigiye, abazaii, Iiphuma, aienge, namacansl*	All from one user_dictionary, which confuses *i* with *l*; (*ligiye*, *abazali, liphuma* etc. are the correct words)
5. Morphology – compound nouns	*isekelashansela, inkulumo-mpikiswano*	Different rules have been applied for compound words (as one word, dashed, two tokens)
6. Morphology – derivation expert names	*ususayensi*	Different rules for imports have been used, noting *uso-* vs *uno-* (in, e.g., *unompilo*)
7. Proper names	*Karim, ku-Andrew*	Multicultural society

Table 2: Summary of reasons why a word was added to the user dictionary.

Seen from a linguistic perspective, the data gleaned from the dictionary provides interesting questions in morphological theory. For instance, the compound word formation process for words derived from English are structurally represented differently in isiZulu. Examples from the user dictionary files are *LikaSekelaShansela* (an inflected form of the one recognised as correct *isekelashansela* 'Vice Chancellor') and *inkulumo-mpikiswano* 'debate', cf. e.g., *Umeluleki wezengqondo* 'Psychologist' that takes a different form with two lexical items in juxtaposition. Interesting theoretical questions, such as whether isiZulu has endocentric and exocentric compounds, need to be explored. Other interesting morphological observations from the data are the structure of the noun *Ususayensi* 'scientist' vs, e.g., *unompilo* 'nurse'. If one were to follow the word formation process for various

experts in isiZulu (e.g., *ususayensi* 'scientist', *usolwazi* 'professor', *usomahlaya* 'comedian', etc.) one would expert *unompilo* (nurse) to be **usompilo*. It would be interesting linguistically to shed more light on this word derivational process.

The user-added words also show that words are being added to the isiZulu lexicon that are not included in isiZulu dictionaries; e.g., *Osemnkant-shubomvu* 'experienced' of which the trigram *mnk* is below the threshold and *kan*, *ant*, *ubo* and *mvu* do not appear, i.e., this is, relatively, not fitting in the common orthography.

Finally, we took a random sample of 15 words from the set of words of the three user dictionaries minus those exhibiting one of the previously mentioned issues, so as to obtain an indication of why they were added from a technical viewpoint regarding the language model. The list of selected terms is included in Table 3. One word was actually recognised as probably correct, and the remainder was mainly due to the language model design decision being case-sensitive.

Word in user_dictionary	**English (base form)**	**Trigram analysis**
Imqomile	to date	*imq* not in trigram list
avele	to show up	Was recognised correctly
lsuke	to leave	*suk* not present as all-lowercase
sebeviyoza	dancing and singing	*ebe* not present as all-lowercase
koqobo	of self	*koq* not in trigram list
yokucobelelana	of sharing	*yok* not present as all-lowercase
yobucayi	that is sensitive	*yob* not present as all-lowercase
mashifoni	Chiffons	*oni* not present as all-lowercase
imifece	type of plant	*fec* not in trigram list
umkhonlo	a lead	*onl* below threshold; more common *umkhondo* is recognised
endlini	*in the house*	*end* not present as all-lowercase
kuieyo	*kuleyo* 'at that'	*uie* not in trigram list (interestingly, *kui* is [n=378] and *iey* is [n=8]); typo added by user
ogcagcayo	one who is wedding	*agc* not present as all-lowercase
ziyojika	will be turning	*oji* not present as all-lowercase

Table 3: Trigram analysis of a random selection of user-added words

Discussion

Overall, the isiZulu spellchecker as a tool has been positively received and it is regarded as an enabler of the intellectualisation of the language. Whether the number of downloads in a two-month period from its launch (n=159) is deemed 'high' or 'low' depends on one's expectations. What is important to note is that it has not been marketed explicitly and the two months fell in the summer holiday period. Thus, its download and usage was largely based on word-of-mouth among a reduced number of the target audience, and in this light, it can be considered a modest success for a first version of the tool and an encouragement to investigate the requested new features. The technical analysis of the user dictionaries mainly revealed that a larger training corpus may be better and that the accuracy can be improved further by adding a few simple string analysis (cf. grammar) rules on top of the statistical language model. Interestingly, the list of user-added words also contained words that are, also in their base form, not available in all dictionaries, such as *inqubekelaphambili* 'development/progress', yet they are still recognised correctly by the spellchecker. This being the case, the 'add word' feature of the spellchecker is an exciting avenue for further investigation into new words that are being added to the lexicon, and thus may soon provide a wealth of evidence on the intellectualisation of the language. The 'add word' feature is unique to this isiZulu spellchecker compared to earlier attempts (Bosch & Eiselen 2005; Prinsloo & de Schryver 2004), which is providing a wealth of information for spellchecker development as well as linguistic analyses. A controlled test setting with university students might assist in obtaining such results. Here, however, we focused on obtaining data from the broader society with its daily activities so as to assess broader impact.

Returning to the three core questions posed in the introduction, they can all be answered in the affirmative. The spellchecker does meet end-user needs and expectations, although we note the suggestions for further improving its functionality, such as suggesting corrections and voice, both of which require prior research. Users perceive that the spellchecker enables the intellectualisation of the language, which is further supported by the analysis of the words added to the dictionary, not all of which are in current dictionaries (i.e., the lexicon is indeed growing).

Conclusions

The evaluation of the isiZulu spellchecker has shown that it has a positive effect on the intellectualisation of isiZulu. The tool has been perceived by the target audience as positive, generating interest in more functionalities. Its use also indicated that a few additional rules will increase its accuracy. The isiZulu spellchecker's 'add dictionary' features proved very useful in suggesting improvements to the language model from a computational viewpoint as well as a linguistic one, to examine emerging words and orthography. The new words that were added to the lexicon are testimony to the fact that the intellectualisation of the language is taking effect. The tool is being actively used in technical spaces such as administration work (formal language), translation work and editing.

Acknowledgments: We would like to thank the survey participants and the users who provided the user dictionaries.

References

Bangor, A., P. Kortum & J. Miller 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies* 4,3:114-123.

Bosch, S.E. & R. Eiselen 2005. The Effectiveness of Morphological Rules for an IsiZulu Spelling Checker. *South African Journal of African Languages* 1:25-36.

Brooke, J. 1996. SUS – A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189,194:4-7.

Bush, B., L. Busch & K. Press 2014. *Interviews with Neville Alexander: The Power of Languages against the Language of Power*. Pietermaritzburg: UKZN Press.

De Schryver, G-M. & D.J. Prinsloo 2004. Spellcheckers for the South African Languages, Part 1: The Status Quo and Options for Improvement. *South African Journal of African Languages* 1:57-82.

Finlayson, R. & M. Madiba 2002. The Intellectualization of the Indigenous Languages of South Africa: Challenges and Prospects. *Current Issues in Language Planning* 3,1:40-61.

Havránek, B. 1932. The Functions of Literary Language and its Cultivation. In

Havránek, B. & M. Weingart (eds.): *A Prague School Reader on Esthetics, Literary Structure and Style*. Prague: Melantrich.

Keet, C.M. & L. Khumalo 2017. Toward a Knowledge-to-text Controlled Natural Language of IsiZulu. *Language Resources and Evaluation* 51,1:131-157

Khumalo, L. 2017. Intellectualization Through Terminology Development. *Lexikos* 27: (in press).

Khumalo, L. 2015. Advances in Developing Corpora in African languages. *Kuwala* 1,2:21-30.

Kotzé, G. & F. Wolff 2015. Syllabification and Parameter Optimization in Zulu to English Machine Translation. *Southern African Computer Journal* 57:23.

Malumba, N., K. Moukangwe & H. Suleman 2015. AfriWeb: A Web Search Engine for a Marginalized Language. In Proceedings of 2015 Asian Digital Library Conference, Seoul, South Korea, 9-12 December 2015.

Ndaba, B., H. Suleman, C.M. Keet & L. Khumalo 2016. The Effects of a Corpus on isiZulu Spellcheckers Based on N-grams. In Cunningham, P. & M. Cunningham (eds.): *IST-Africa 2016*. IIMC International Information Management Corporation. May 11-13, Durban, South Africa.

Olivier, J. 2014. Compulsory African Languages in Tertiary Education: Prejudices from News Website Commentary. *Southern African Linguistics and Applied Language Studies* 32,4:483-498.

Pretorius, L. & E.S. Bosch 2003. Finite-State Computational Morphology: An analyzer Prototype for Zulu. *Machine Translation* 18,3:195–216.

Prinsloo, D.J. & G-M. de Schryver 2004. Spellcheckers for the South African Languages, Part 2: The Utilisation of Clusters of Circumfixes. *South African Journal of African Languages* 1:83-94.

RSA (Republic of South Africa). 1996. *Constitution of the Republic of South Africa*. Pretoria: Government Printer.

Shizha, E. 2012. Reclaiming and Re-visioning Indigenous Voices: The Case of the Language of Instruction in Science Education in Zimbabwean Primary Schools. *Literacy Information and Computer Education Journal (LICEJ) Special Issue* 1,1:785-793.

Sibayan, B. 1999. *The Intellectualization of Filipino*. Manila: The Linguistic Society of the Philippines.

Appendix A – Questionnaires

General Questions

1. Which feature do you find most useful? Please provide a brief reason. *Iyiphi ingxenye yaleli thuluzi lokupela oyithole iwusizo kakhulu? Sicela usinike isizathu.*
2. Which feature do you find least useful? Please provide a brief reason. *Iyiphi ingxenye yaleli thuluzi lokupela oyithole ingenalusizo? Sicela usinike isizathu.*
3. Are there any features you would like to see added to the spellchecker? *Zikhona ezinye izinto ofisa zifakwe ethuluzini lokupela?*
4. Are there any features you would like removed from the spellchecker? *Ingabe kukhona ofisa kususwe kuleli thuluzi lokupela?*
5. Do you think that the spellchecker is an enabler for the intellectualization of isiZulu? *Ucabanga ukuthi leli thuluzi lokupela linomthelela omuhle ekuthuthukisweni kwesiZulu njengolimi lokufunda? (Sicela uphendule ngo yebo, kakhulu, kancane noma cha).*
6. How has the spellchecker enhanced your work as a language practitioner? *Ingabe leli thuluzi likuthuthukise kanjani ukusebenza kwakho njengosozilimi?*
7. If you want to have the spellchecker integrated in another application, in which application would you prefer to have it integrated the most? (Please indicate order of preference, with 4 the highest and 1 lowest): OpenOffice, Thunderbird, Firefox, Chrome. *Uma ufuna ithuluzi lokupela lifakwe kwenye indawo osebenza ngayo, ungalifaka kuliphi? (Sicela ukhethe ngokulandelana kwazo lapho eye-4 iphezulu ne-1 iphansi): OpenOffice, Thunderbird, Firefox, Chrome.*
8. How often do you use the spellchecker? *Awulinganise ukuthi ulisebenzisa kangaki leli thuluzi: nsukuzonke, ngesonto, qabukela, ngike ngalizama kanye, angikaze.*

SUS in English

1. I think that I would like to use this system frequently
2. I found the system unnecessarily complex

3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people would learn to use this system very quickly
8. I found the system very cumbersome to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system

SUS in isiZulu

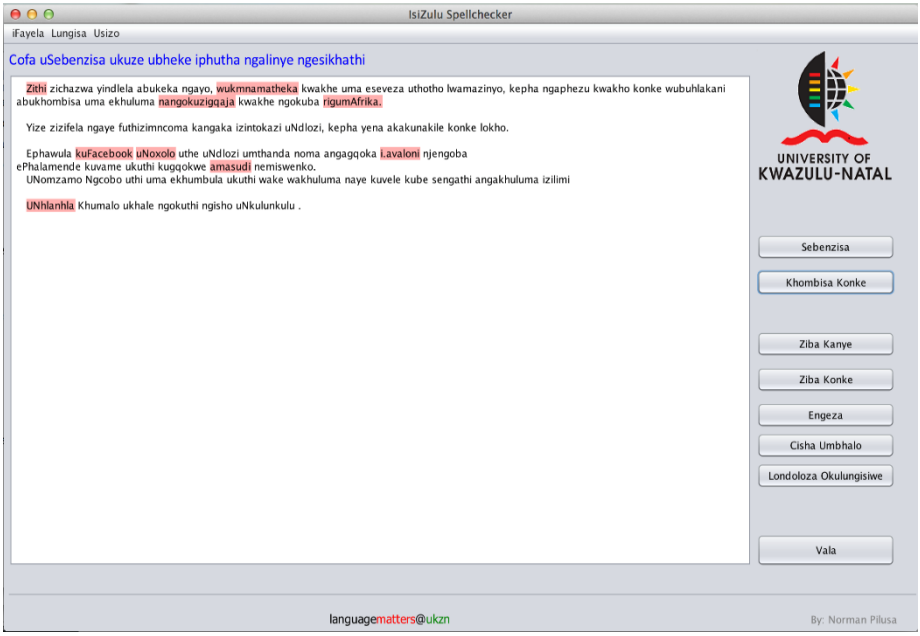
1. Ngicabanga ukuthi ngingathanda ukusebenzisa loluhlelo njalo
2. Ngiluthole ludida ngokungadingekile loluhlelo
3. Ngicabanga ukuthi kulula ukusebenzisa loluhlelo
4. Ngingadinga ukulekelelwa ngumuntu onobuchwepheshe ukuze ngikwazi ukusebenzisa loluhlelo
5. Ngithole ukusebenza kwaloluhlelo okunhlobonhlobo kudidiyelwe kahle
6. Ngicabanga ukuthi kunokuningi okungahambisani kulolu hlelo
7. Ngicabanga ukuthi abantu abaningi bazofunda ukusebenzisa loluhlelo ngokushesha
8. Ngithole kunzima ukusebenzisa loluhlelo
9. Ngibenokuzethemba ngisebenzisa loluhlelo
10. Ngidinge ukufunda izinto eziningi ngaphambi kokusebenzisa loluhlelo

Additional Questions for the interview with Industry Participants (IsiZulu Newspaper Editors)

1. How often do you use the isiZulu spellchecker in your line of work?
2. Is it easy to use the isiZulu spellchecker?
3. If not, what do you find to be the complication in the use of the isiZulu spellchecker?
4. Do you use it independently, with the aid of a technician, or practically collaboratively with a colleague?

5. Do you have any function(s) of the isiZulu spellchecker that you are unable or find difficult to use?
6. Do you have confidence in using the isiZulu spellchecker?
7. Did you need to learn something in order to be able to use the isiZulu spellchecker?
8. Are there any features that you would want added onto the isiZulu spellchecker?
9. Would you recommend it to other people in your line of work?
10. How has the isiZulu spellchecker enhanced your work?
11. Do you think that the isiZulu spellchecker is an enabler for the intellectualization of isiZulu?
12. Do you have any other opinion on the isiZulu spellchecker?

Appendix B – A Screenshot of the isiZulu spellchecker with the isiZulu interface



C. Maria Keet
Department of Computer Science
University of Cape Town
mkeet@cs.uct.ac.za

Langa Khumalo
Linguistics Program
University of KwaZulu-Natal
khumalol@ukzn.ac.za