UNIVERSITY OF KWAZULU-NATAL™
INYUVESI YAKWAZULU-NATALI

languagematters@UKZN

University Language Planning and Development Office
Third Quarterly Report: 2017

INSPIRING GREATNESS

Langa Khumalo, PhD Linguistics.
Director: Language Planning and Development Office.

The highlight of this quarter is, without doubt, the Corpus Field work, which was conducted in July 2017. Following the University Language Planning and Development's effort to create arguably the biggest African Language Corpus, and in an effort to have a balanced, organic, monitor, monolingual corpus of isiZulu, the Office embarked on a fieldwork to collect natural language speech samples across the KwaZulu-Natal Province. The geographical distribution of selected districts and thematic coverage was carefully sampled to achieve balance. These oral corpus materials are going to be processed and added onto the isiZulu National Corpus, which now stands at an impressive 23 million tokens.

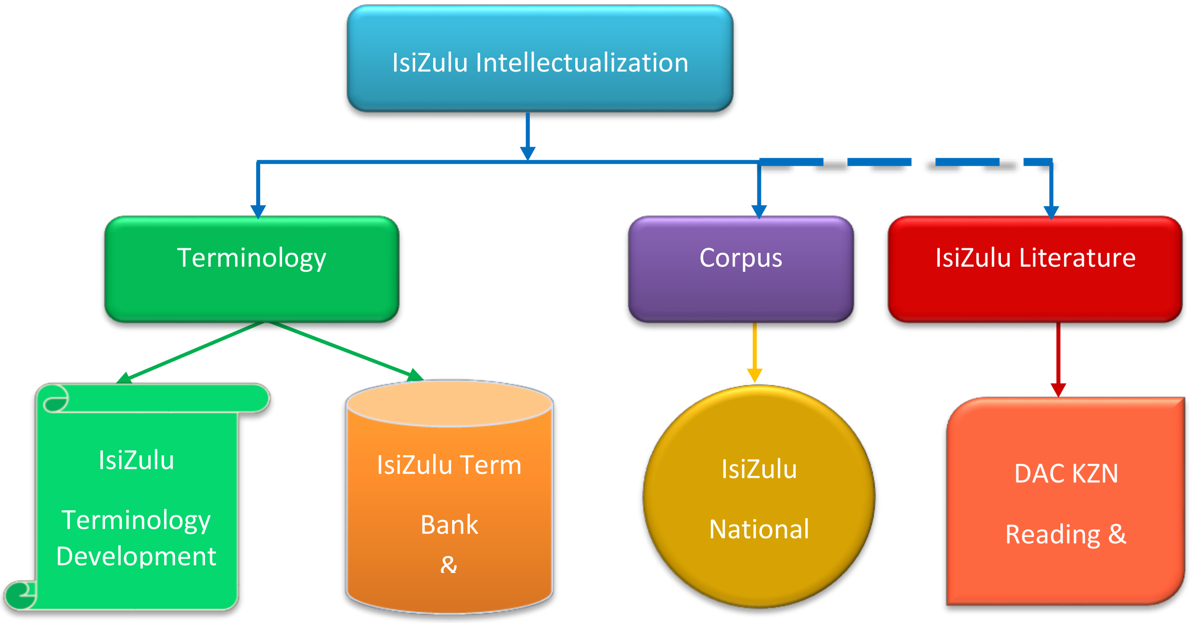
Third Quarterly Report
June-Sept 2017.

Contents

1. Introduction

The University Language Board implements and monitors the provisions of the University Language Policy as set out in the ULB Charter. The program to intellectualize isiZulu so that it (ultimately) functions as both an academic language, and a language of administration across the University alongside English is initiated by the ULPDO, and is reported quarterly at the ULB. The major thrust of the language program (*as seem in the figure below*) as approved by the ULB is the creation of discipline specific terminology in isiZulu, the building of an isiZulu National Corpus, and the development of a contemporary body of literature in isiZulu. Other language activities include the provision of training workshops, translation and (simultaneous) interpreting services, Sign Language advocacy, the Sesotho *Bua Le Nna* Program, language research, and the development of computational tools.

Figure showing the Language Program



The major thrust in this quarter was the expansion of the IsiZulu National Corpus (INC). As reported in 5.1 the INC has surpassed the 23 million token landmark and continues to grow. The major development in this period under review is the collection of natural language speech samples of isiZulu data under the auspices of the first ULPDO corpus fieldwork undertaken in July 2017. The ULPDO is in the process of processing these oral speech materials for addition in the ever-growing INC. This will result in a new milestone, the INC will be one of the few corpora in the world that would have achieved the balance of both written and spoken data. We are very excited to have achieved this. The ULPDO also hosted an international scholar who conducted training on computational linguistics and digital humanities.

Dr Langa Khumalo
ULPDO

2. Stakeholder meeting/workshops

2.1. Computer Aided Text Mark-up Analysis workshop

The ULPDO in collaboration with the South African Centre for Digital Language Resources (SADiLaR) successfully hosted a Computer Aided Text Mark-Up Analysis (CATMA) workshop on the 25th of August 2017 at the Howard College Language Laboratory. The CATMA training workshop was led by Professor Dr Jan Christoph Meister from the University of Hamburg, Germany. The ULPDO Director is a member of the Scientific Advisory Committee (SAC) for SADiLaR, and was instrumental in partnering SADiLaR to bring Professor Meister to UKZN to share his expertise in the area of Computational Linguistics and Digital Humanities.

This hands-on workshop introduced computational linguists, language experts, and media experts working in Digital Humanities to the CATMA.50 tool, which was developed at the University of Hamburg, and is currently used by over 60 research projects worldwide. CATMA offers a unique combination of three main features found in no other text analysis tool:

1) CATMA supports collaborative annotation and analysis - a text or text based corpus can be investigated individually, but also jointly by a group of students or researchers.

2) CATMA supports explorative, non-deterministic practices of text annotation - a discursive, debate-oriented approach to text annotation based on the research practices of hermeneutic disciplines is the underlying conceptual model.

3) CATMA integrates text annotation and text analysis in single web based working environment - which makes it possible to combine the identification of textual phenomena with their investigation in a seamless, iterative fashion.

It is a tool that can be innovatively used in research and teaching & learning across all disciplines. The workshop was an important precursor to the processing and annotation of the ULPDO's massive collection of natural language isiZulu oral speech data. The CATMA workshop was a huge success and participants from UKZN and Durban University of Technology (DUT) were very grateful to have been part of it.

3 Fieldwork Corpus

The IsiZulu National Corpus (INC) has hitherto been made up of isiZulu written digital text data. In order to achieve balance in the INC, the University Language Planning and Development Office (ULPDO) embarked on a corpus fieldwork to collect oral speech samples from isiZulu speaking communities across the length and breadth of the KwaZulu-Natal Province.

The field work began on the 1st - 14th of July 2017. ULPDO advertised for and appointed 30 Student Research Assistants who were trained before being sent to all the eleven districts of the KwaZulu-Natal (KZN) Province to collect speech samples on carefully selected thematic topics. ULPDO Research Team leaders travelled to designated districts of the North and South regions of KZN to monitor, collate and store the work done by the Research Assistants.

The speech samples covered various pre-determined thematic areas in order to be as reflective of the Zulu people's human experiences as possible. The oral speech samples are currently being processed and added onto the existing INC in order to have a rich balance of both written and oral corpus materials. The oral corpus target is 1 million running words.

The fieldwork was a huge success, and an important training for both the student research assistants who participated, and the ULPDO staff members. The appointed student research assistants together with ULPDO staff members did an excellent job. To this end, the oral corpus files have been received from the student research assistants and are being processed. The digital recordings are going to be stored for posterity.

4. Running Projects

4.1. Bua Le Nna Sessions

The ULPDO successfully held four Bua Le Nna sessions in June, July and August 2017 respectively.

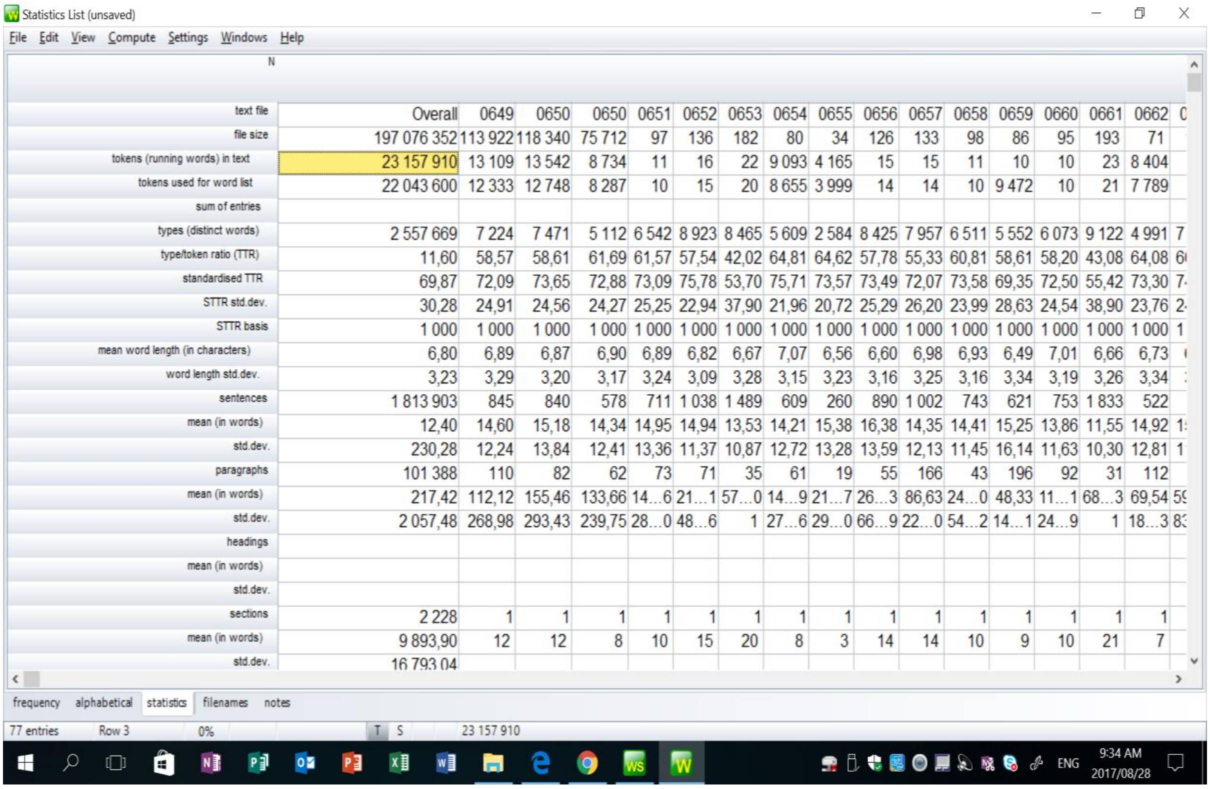
The picture shows the students who participated in this session including the language champions and ULPDO staff members.

This project continues to contribute to the imperative of promoting multilingualism in Higher Education and social cohesion. Multilingualism is articulated in the country's constitution, and accentuated by the UKZN language policy and plan. The students at UKZN continue to support this project by attending the Bua Le Nna sessions in impressive numbers. The numbers of students attending these sessions are increasing every term. This has certainly boosted the social cohesion project of the university, which is also central actualizing the culture of inclusivity, respect for other cultures in the university community. ULPDO is now proud to report that students are confidently conversing in Sesotho language during these sessions.

(Computer Aided

5. ULPDO Human Language Technologies (HLTs)

5.1. Corpus Collection and cleaning



text file	Overall	0649	0650	0650	0651	0652	0653	0654	0655	0656	0657	0658	0659	0660	0661	0662
file size	197 076 352	113 922	118 340	75 712	97	136	182	80	34	126	133	98	86	95	193	71
tokens (running words) in text	23 157 910	13 109	13 542	8 734	11	16	22	9 093	4 165	15	15	11	10	10	23	8 404
tokens used for word list	22 043 600	12 333	12 748	8 287	10	15	20	8 655	3 999	14	14	10	9 472	10	21	7 789
sum of entries																
types (distinct words)	2 557 669	7 224	7 471	5 112	6 542	8 923	8 465	5 609	2 584	8 425	7 957	6 511	5 552	6 073	9 122	4 991
type/token ratio (TTR)	11,60	58,57	58,61	61,69	61,57	57,54	42,02	64,81	64,62	57,78	55,33	60,81	58,61	58,20	43,08	64,08
standardised TTR	69,87	72,09	73,65	72,88	73,09	75,78	53,70	75,71	73,57	73,49	72,07	73,58	69,35	72,50	55,42	73,30
STTR std.dev.	30,28	24,91	24,56	24,27	25,25	22,94	37,90	21,96	20,72	25,29	26,20	23,99	28,63	24,54	38,90	23,76
STTR basis	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
mean word length (in characters)	6,80	6,89	6,87	6,90	6,89	6,82	6,67	7,07	6,56	6,60	6,98	6,93	6,49	7,01	6,66	6,73
word length std.dev.	3,23	3,29	3,20	3,17	3,24	3,09	3,28	3,15	3,23	3,16	3,25	3,16	3,34	3,19	3,26	3,34
sentences	1 813 903	845	840	578	711	1 038	1 489	609	260	890	1 002	743	621	753	1 833	522
mean (in words)	12,40	14,60	15,18	14,34	14,95	14,94	13,53	14,21	15,38	16,38	14,35	14,41	15,25	13,86	11,55	14,92
std.dev.	230,28	12,24	13,84	12,41	13,36	11,37	10,87	12,72	13,28	13,59	12,13	11,45	16,14	11,63	10,30	12,81
paragraphs	101 388	110	82	62	73	71	35	61	19	55	166	43	196	92	31	112
mean (in words)	217,42	112,12	155,46	133,66	14...6	21...1	57...0	14...9	21...7	26...3	86,63	24...0	48,33	11...1	68...3	69,54
std.dev.	2 057,48	268,98	293,43	239,75	28...0	48...6	1	27...6	29...0	66...9	22...0	54...2	14...1	24...9	1	18...3
headings																
mean (in words)																
std.dev.																
sections	2 228	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mean (in words)	9 893,90	12	12	8	10	15	20	8	3	14	14	10	9	10	21	7
std.dev.	16 793,04															

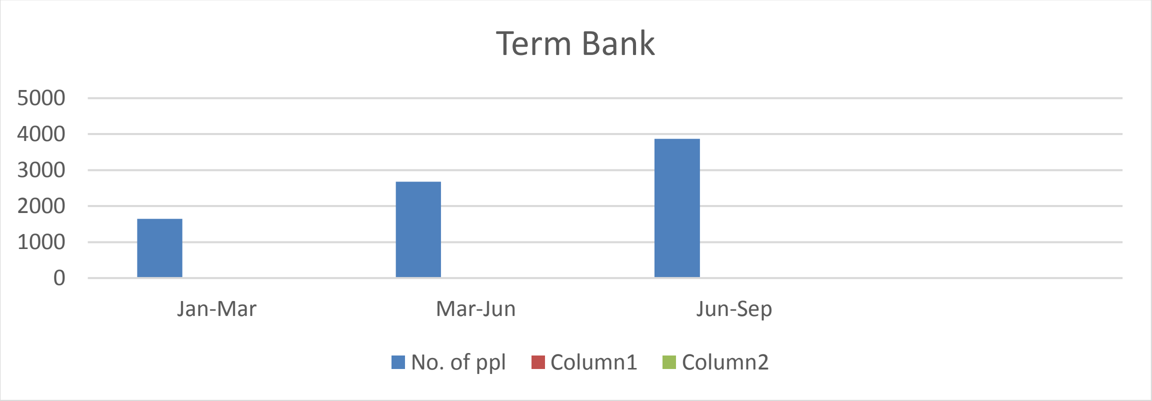
As shown in the Figure above, a total of 423 842 tokens were collected during this third quarter of 2017, which contribute to a current total of 23,157,910 tokens (roughly 23.1 million tokens). The office is currently seized with the task of corpus cleaning, which is one of the important stages in the corpus building. The office continues with its collaborative partnership with Kellie Campbell for corpus material collection. The office continues to unlock new relationships, which are going to contribute immensely in the corpus building effort. In this regard, the office is currently engaging the Mazisi Kunene Foundation in order to unlock a synergy that will mutually profit the two institutions, particularly access to isiZulu corpus materials.

5.2 Term Bank, Spellchecker and Zululex

The ULPDO continues to monitor the use of human language technologies (HLTs) that were launched in November 2016. The records are indicating that targeted end-users and the general public are indeed using these HLTs and their usage is gradually increasing every day. So far, the IsiZulu Term Bank has attracted 3869 people who have visited the electronic term bank.

Different institutions and stakeholders continue to download the Zulu Lexicon, a mobile application compatible with android and iOS phones, and the IsiZulu spell checker. Overall, the software tools are being used as expected and their utility is gradually increasing. As a result of this, the office has received compliments relating to the quality of these HLTs and their functionality. This affirms UKZN as increasingly becoming an epicenter for language development and innovation. The Figure below illustrates the uptake of the isiZulu Term Bank based on visitors to the electronic platform.

People accessing the Term Bank



The Figure above shows that following the launch of the Term Bank late last year, in the first quarter of this year 1642 people had visited the term bank. The number grew in the second quarter with more than 1000 people, with a total of 2678 people having accessed the term bank. In this quarter the figure grew to 3869. The isiZulu Term Bank is the first multidisciplinary bilingual English-IsiZulu electronic resource that is available as an open source. The office expects that this tool will contribute towards the active intellectualization of isiZulu.

6. Workshops and Conference attendance

6.1 Director's conference and workshops attendance

The ULPDO director was invited to the following important functions that recognize UKZN's leading role in language development/intellectualization.

1. To travel as part of KZN delegation to China on 15-25 June 2017
2. To address the NWU Senate on 27 June 2017
3. To attend DHET Ministerial Committee Meeting on 3 July 2017
4. To lead the UNISA Council Workshop on Transformation on 1-2 August 2017

7. The DR9 Rule

The new Doctoral Rule (DR9b) as approved by Senate in November 2016 states that every thesis submitted shall be in such format as prescribed by the Senate and the rules of the relevant college; provided that each thesis shall include an abstract in both English and isiZulu. Each English and isiZulu abstract shall not exceed 350 words. At the last ULB the operationazation of the DR9 Rule was referred to the Research Office. The current status is that following a meeting with the DVC Research and DVC Teaching & Learning, and the Director ULPDO, the matter was referred to Executive Management Committee (EMC) for the EMC to consider and approve the abstract submission process and the annual budget for the translation of abstracts.

8. Publications

1) **Khumalo, L.** 2017. Intellectualization through terminology development. Lexikos 27 (Accepted).
2) Keet, C.M., **Khumalo, L.** 2017. Toward a knowledge-to-text controlled natural language of isiZulu. Language Resources and Evaluation, 51(1): 131-157.
3) Keet, C.M., **Khumalo, L.** 2017. Evaluation of the effects of a spellchecker on the intellectualization of isiZulu. Alternations. (Accepted).