Human Language Technology Audit 2017/8

Final report

1 July 2017 – 31 August 2018

PROJECT INFORMATION	
PROJECT	Human Language Technology Audit 2017/8
BENEFICIARY	MERAKA INSTITUTE, CSIR
CONTRIBUTORS	Carmen Moors, Karen Calteaux, Ilana Wilken, Tebogo Gumede
	HLT Research Group, CSIR Meraka Institute
PRESENTED TO	SADiLaR, North-West University
REPORTING PERIOD	1 July 2017 – 31 August 2018
VERSION	1.7
DATE SUBMITTED	31 August 2018

EXECUTIVE SUMMARY

This document reports on all work conducted in the 2017/8 Audit of human language technology (HLT) resources available in South Africa project. The purpose of conducting the HLT Audit 2017/8 is to update the previous HLT Audit conducted by Sharma Grover in 2009 as part of her research thesis. Increased activity in the HLT field in South Africa, an increase in the number of institutions undertaking HLT research and development (R&D) in the country, the increased local interest in digital humanities as a field of study and the recent establishment of the South African Centre for Digital Language Resources, form part of the rationale for undertaking this project - eight years after the previous Audit.

The HLT Audit 2017/8 project consists of five work packages encompassing the Audit design, instrument development, execution, and analysis, as well as a work package aimed at designing a dynamic Audit update system which will enable entities involved in HLT R&D to upload the outputs of their work as this becomes available.

The HLT Research Group at the CSIR Meraka Institute collaborated with the South African Centre for Digital Language Resources (SADiLaR) and the Resource Management Agency (RMA) to achieve the desired outcomes of the project.

This final report provides detailed information on the work completed from 1 July 2017 to 31 August 2018.

ACRONYMS

BLaRK	Basic Language Resource Kit
BLEU	Bilingual Evaluation Understudy
CLARIN	Common Language Resources and Technology Infrastructure
CSIR	Council for Scientific and Industrial Research
CTexT	Centre for Text Technology
ELRA	European Language Resources Association
HLT	Human language technology
HLTRG	Human Language Technology Research Group
LRE(C)	Language Resources and Evaluation (Conference)
LRE Map	Language Resources and Evaluation Map
MuST	Multilingual Speech Technologies
NIST	National Institute of Standards and Technology
NWU	North-West University
RMA	Resource Management Agency
R&D	Research and development
SADiLaR	South African Centre for Digital Language Resources
SET	Science, engineering and technology
SU	Stellenbosch University
VLO	Virtual Language Observatory
WER	Word-error-rate
UNISA	University of South Africa

CONTENTS

FIGURES

TABLES

1. INTRODUCTION

South Africa is a highly multilingual country in which communication barriers and the digital divide are still largely prevalent. Research and development (R&D) activities on language form a fertile field of study which can contribute to nation-building, as well as regional growth and economic development.

The South African government has realised the role which technology, and in particular human language technologies (HLTs), can play in bridging communication barriers and addressing the digital divide. HLTs fall roughly into two main categories - text technologies and speech technologies. Through its custodian of language matters, the Department of Arts and Culture, the South African government has contributed significantly to HLT R&D and the development of human language technologies, tools, resources, and applications in terms of both text and speech technology. In addition, there have been several industry-funded initiatives to develop and make available HLT resources, tools and applications in the South African languages over the last approximately 15 years.

In order to make decisions about future investment in ICT R&D (including HLT R&D), digital humanities-related R&D, and language resource management and infrastructure development, a current unified picture of the South African language technology domain (among others) is required.

Technology audits aim to identify, assess and catalogue technologies according to different criteria, ranging from categories of technologies, maturity of technologies, competitive position, location in the supply chain, and levels of competencies, through to impact of technologies [1]. By employing a mapping technique known as a technology matrix, technology audits can provide an overview of the related technology landscape in a company, market or country. This information can then be used in various decision-making scenarios, among others, investment decisions to be taken by the newly-established South African Centre for Digital Language Resources (SADiLaR).

Figure 1 below[1] illustrates the link between technology and business planning and the role of the technology audit in that process:

[1] Image courtesy of the Centre for Technology Management at the University of Cambridge. Accessed at https://www.google.co.za/search?q=technology+audit&espv=2&biw=1396&bih=669&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiFw52z7arRAhWIIsAKHXoTC5YQ_AUIBigB#imgrc=yeVtR3FB1mICgM%3A on 5 January 2017.

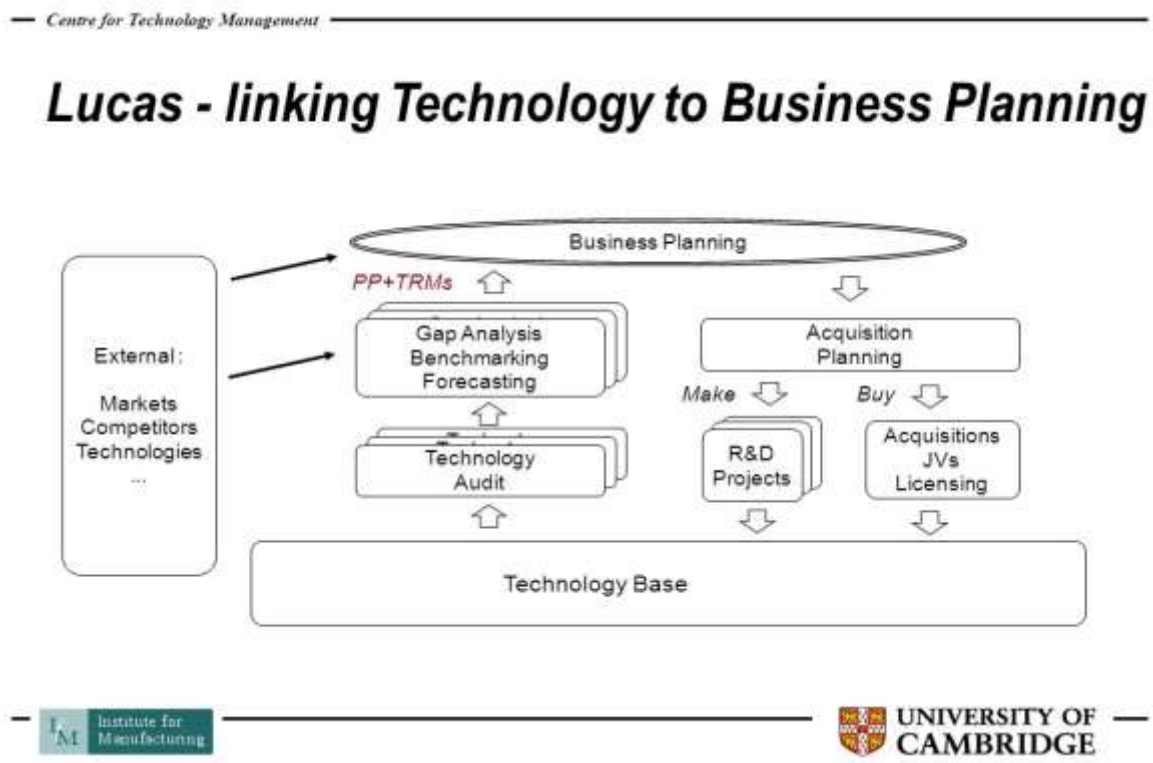


FIGURE 1: LINKING TECHNOLOGY TO BUSINESS PLANNING

2. PROJECT BACKGROUND

A technology audit on the state of human language technologies (HLT) research and development (R&D) in South Africa, undertaken and published by A Sharma Grover in 2009, revealed that the HLT components available at the time, were of a basic and exploratory nature and required substantial investment to be developed into fully-fledged language resources. Comparatively, it was found that HLT development and resources in languages such as Afrikaans and South African English were better developed than in the African languages, but also that technologies and resources in some of the African languages, notably isiZulu, isiXhosa, Setswana, Sesotho and Sepedi were better developed than in the other official African languages. The study in question analysed the South African HLT landscape using a number of complementary methods and made recommendations on how to accelerate HLT development in South Africa.

The relatively recent growth in interest in the field of digital humanities provides further context to the current proposal. Digital humanities has emerged as a field of scholarly activity at the intersection of computing or digital technologies and the humanities. An audit of the HLT R&D landscape in South Africa provides insight into the digital tools, resources and applications available to researchers in the humanities. These tools, resources and applications make new kinds of knowledge generation possible, indicating the importance of up-to-date information on what is and is not available.

Lastly, the growth in interest in managing data emanating from research in order to ensure availability to other researchers and contribution to the body of knowledge on a topic provides a further rationale for reviewing the landscape of HLT R&D in South Africa, eight years after the first HLT audit. In the era of data becoming available at a rapid rate, it is not only important that information on the state of HLT R&D be available to researchers, but also that such data be current and be continually updated, preferably in an automated manner.

3. PROJECT OBJECTIVES

This project aims to address several of the issues mentioned above and provide information on the current state of HLT R&D in South Africa. Specifically, the proposed HLT audit aims to replicate the one completed in 2009 and update the information on the various HLT tools, resources and applications identified in that audit. In addition, tools, resources and applications developed since 2009, will be identified and categorised using a slightly updated version of the technology matrix previously employed.

As the audit will replicate the 2009 survey to the extent possible, the objectives of the project will be similar to the research questions posed in the 2009 study, namely -

1. to provide a systematic and detailed inventory of the current HLT components in the official South African languages;
2. to set out the most important dimensions/criteria for the documentation of the HLT components;
3. to describe the status of the HLT components in the SA R&D environment for the 11 official languages:
4. Which components exist and are freely available?
5. Which components are most important for the SA context?
6. What differences exist in components across the 11 official languages? and
7. to indicate the gaps between available components and the most important components for the South African context.

The information gathered through the audit will be made available in the form of a report, which can be accessed through a public website such as that of the Resource Management Agency (RMA) or, if available by then, that of the new South African Centre for Digital Language Resources (SADiLaR).

Ideally, the information from the audit should be updated regularly. The project will therefore also deliver architecture for a system which will enable the automated updating of the information. This system will be designed in collaboration with the RMA and SADiLaR.

4. PROJECT SCOPE AND ASSUMPTIONS

The CSIR will undertake an audit of text technology and speech technology tools, resources and applications developed between 2009 and 2017. Where required, previous data will be updated and the current status of the technology category will be indicated. Where no category existed before a new category will be added. These categories will be identified as such and the current status will be indicated.

Non-HLT-related tools, resources and applications, i.e. tools, resources and applications which do not relate to either text-based technology or speech-based technology, will not be surveyed. An example would be purely linguistic resources such as formal grammars, or literary texts which are not directly related to HLT development.

The audit will be limited to the HLT R&D community. Companies and institutions that perform in-house development of any kind of HLTs may be considered secondarily - especially if they were included in the previous survey. Entities that do not perform HLT R&D, e.g. companies that are service providers of HLT products purchased off-the-shelf, or that only perform integration and/or general maintenance activities relating to these technologies, e.g. call centres, will not be included in the audit.

In-depth discussions on HLT, beyond an overview of what is understood as text and speech technology, and on technology audits per se, also lie outside the scope of this project. Background information on these topics is available in the original audit report.

Lastly, development and implementation of the proposed automated/dynamic audit update system, falls outside the scope of this project. This project will only deliver an architecture design for such a system which will be documented by means of solution architecture diagrams and functional specifications.

Project assumptions:

- Assumption 1: The project will apply the same technology classification/categories and methodology as used in the 2009 audit.
- Assumption 2: All stakeholders are available in the required timeframe to participate in the interviews for the audit.
- Assumption 3: Representatives from SADiLaR and the RMA are available for discussions and prepared to be consulted in conceptualising the dynamic audit update solution.
- Assumption 4: There will be no significant increase in the costs of travel to interview the stakeholders.
- Assumption 5: All HLT researchers will be available to conduct the audit within the required timeframe.
- Assumption 6: There will be limited delays in signing the project agreement.

- Assumption 7: The latest date for starting the project will be 1 April 2017.
- Assumption 8: It will be possible to roll over funding from one financial year to another, if required and by means of a written motivation.

The project would take place over 12 months and include five main work packages (WPs) as follows:

- Work Package 1: Audit design
- Work package 2: Instrument development
- Work package 3: Audit execution
- Work Package 4: Results consolidation, reporting, dissemination and transfer
- Work Package 5: Architecture design.

5. PROJECT APPROACH

The project proposes to conduct a second HLT R&D audit based on work done by A Sharma Grover in 2009. The methodology to be used will follow the same concepts and structure as the first audit. The instruments will be modified and adapted based on current requirements.

The 2009 audit focused on the profile of HLT technologies in South African official languages [3, 4]. Sharma-Grover (2009:44) indicates that a technology audit may "*address specific technological components, or take a higher macro-level view of the bigger picture or value chain around the technology depending on the goals of the audit*" [2].

As the purpose of this project is to undertake an HLT R&D audit from a national science and technology sector level, the approach we propose is to apply the Basic Language Resource Kit (BLaRK) concept [5, 6] to audit each HLT component, i.e. data, modules and applications. Sharma Grover describes these as follows:

- **Data**: The linguistic data sets or collections, which may be either speech or text, in a machine readable form that are used to create, evaluate and improve HLT technology modules. Data includes items such as corpora, lexicons and grammars.
- **Modules**: The basic software units or processes that are usually required to create HLT applications and products. This includes items such as part-of-speech taggers, sentence tokenisers, language models and acoustic models.
- **Applications**: The categories of different application areas where HLT is used. It includes application domains such as speech input, document production, proofing/authoring tools and translation." [2]

The respondents will be drawn from the HLT R&D landscape in South Africa which has the following components: role players, stakeholders and clients. As part of the Audit Design work package (WP1), the researchers will update the list of all current role players, stakeholders and clients. An audit plan and an interview plan will guide the data collection process.

The proposed outputs are presented as WP deliverables. WP6 entails project management, coordination and communications activities. Tasks and activities for each WP are described in the Work Breakdown Structure, together with time-schedules and deliverables. An appropriate project methodology will be adopted; tasks will be completed as specified in the Task Schedule. A series of progress meetings will be scheduled to report and monitor progress and delivery.

6. SCOPE OF WORK

6.1 WORK PACKAGE 1 - DESIGN AUDIT

Since this project involves replicating and updating the 2009 technology audit, the previous audit (published in various conference papers [3, 4, 7, and 9] and Aditi Sharma Grover's Masters Dissertation [2]) will serve as a foundation for this audit. The researchers will be tasked with familiarising themselves with the technology audit methodologies applied in the 2009 audit, as well as the BLaRK concept utilised.

Additional metrics such as NIST and BLEU scores and word-error-rate (WER), will be considered to determine the maturity of the applicable technologies audited, for example in evaluating machine translation and text processing technologies (NIST & BLEU), or speech processing technologies (WER).

Additional methods of collecting information on available language resources, such as that employed by the Language Resources and Evaluation (LRE) Map, namely to consult conference proceedings as a supplementary source of information, will also be considered.

The implications of the country-wide research data management initiative for this project will also be considered and taken into account in the audit design (if required).

Finally, the proposed audit design and methodology will also be discussed with SADiLaR and RMA representatives to ascertain specific requirements and understand alternative methodologies which may be relevant, and the design will be adapted accordingly.

The outputs of this WP include an audit design document which sets out the categories of technologies to be audited and the methodology to be followed, as well as an audit plan (i.e. audit/interview time schedule). It is foreseen that the categories from the 2009 audit will be used as far as possible, but that new categories may be added if required by the current status of technologies developed since 2009.

6.2 WORK PACKAGE 2 - DEVELOP AND TEST AN INSTRUMENT TO CAPTURE UPDATED INFORMATION

Once the familiarisation of current work has been completed as described in WP1, the instrument development and stakeholder identification will commence. The instrument development will be based on the previous work done in this field; however, the design will be customised based on current requirements.

The stakeholders will be identified from the HLT community within South Africa.

The audit instruments such as the questionnaires and interview outlines will be determined, drafted and tested within a small group before finalisation. Two possible methods for the interviews will be tested during the pilot period. These methods are remote interviews (e.g. via tele- or video-conferencing technology) and *in situ* interviews. The reason for this is to test whether the questionnaire can be successfully administered remotely.

Once finalised, the methodology and instrument will be shared with the SADiLaR and RMA representatives for final sign-off before commencement of the audit with confirmed stakeholders.

6.3 WORK PACKAGE 3 - EXECUTE AUDIT

The methodology and instruments developed and signed-off on in the previous WP will be used to conduct the actual audit. The stakeholders identified in the previous WP will be contacted to set up interviews for the audit, based on the final interview method determined by the outcome of the pilot in WP2.

The interview session will be conducted according to the interview methodology in terms of time and structure. The results of the interviews will be captured in an appropriate database which will be developed in line with the questionnaire/interview questions.

6.4 WORK PACKAGE 4 - CONSOLIDATE RESULTS OF INTERVIEWS, REPORT AND DISSEMINATE FINDINGS AND TRANSFER DATA

The interview results captured in the previous WP will be analysed according to the audit design criteria and recorded in an audit report. The audit report will be presented to SADiLaR and RMA representatives and will include a narrative summary of the methodology and instruments used the actual interviews and the results obtained from these interviews.

A one-day workshop will be held to disseminate the findings of the audit to the HLT R&D community and other interested stakeholders, including the HLT Expert Panel. A research paper on the audit and the findings will be written in order to be submitted for presentation at a relevant conference.

6.5 WORK PACKAGE 5 - DYNAMIC AUDIT UPDATE SOLUTION DESIGN (ARCHITECTURE DESIGN)

In the modern era of big data (collection, fusion and visualisation), distributed networking, and large infrastructure programmes, automating the updating of information on HLT R&D outputs using a distributed platform is imperative. Undertaking technology audits every few years should no longer be the only way to obtain an overview of HLT R&D in the country, as the information should be continuously updated by the developers thereof.

This work package aims to design a solution architecture for a system that will enable entities involved in HLT R&D to upload the outputs of their work as this becomes available. The envisaged outcome is a secure distributed platform which provides for quality control over the uploaded content.

The WP will involve familiarisation with the information management architecture of CLARIN and the RMA, as a point of departure.

The outputs will be solution architecture design diagrams, functional specifications, and a recommendation report on how to implement the proposed solution.

7. DELIVERABLES

TABLE 1: PROJECT DELIVERABLES

WP	Description	Deliverable
WP1	Audit design	● Audit design workshop ● Audit designs ● Deliverable report
WP2	Audit instrument development	● Online audit instrument ● Deliverable report
WP3	Audit execution	● Completed online audit ● Deliverable report
WP4	Audit results analysis, consolidation, reporting and dissemination of findings	● Raw data ● Analysed data and representation of the data ● Deliverable report ● Dissemination workshop ● Conference papers
WP5	Architecture design and recommendations	● Deliverable report

8. WORK PACKAGES

The progress on each of the work packages will be reported on in the following 5 sections.

8.1 WP1: 2017/8 HLT AUDIT DESIGN

8.1.1 INTRODUCTION

The HLT Audit design work package was planned to include the following activities:

- Familiarise with Aditi Sharma Grover's thesis entitled "Technology Audit: the state of human language technologies R&D in South Africa".
- Familiarise with formal literature on technology Audits.
- Familiarise with and determine implications of research data management policies and requirements.
- Updates and integration discussion with SADiLaR and the RMA.
- Draft Audit design document – capture technology categories to be audited and describe methodology.

The work was conducted as planned and is described in more detail below.

8.1.2 FAMILIARISATION WITH (HUMAN LANGUAGE) TECHNOLOGY AUDITS

Since this project involves replicating and updating the 2009 HLT Audit, the researchers familiarised themselves with the technology audit methodologies applied in the 2009 HLT Audit, as well as the BLaRK concept utilised there.

Literature on the metrics used to determine the maturity of the applicable technologies audited, for example NIST and BLEU scores for evaluating machine translation and text technologies, and word-error-rate (WER) for evaluating speech technologies, was also perused in order to obtain a better understanding of these techniques.

Additional methods of collecting information on available language resources, such as that employed by the Language Resources and Evaluation (LRE) Map, namely to consult conference proceedings as a supplementary source of information, were also considered and shared with the RMA in a meeting on 7 February 2018. This information was documented in a recommendations report and was submitted to SADiLaR by 30 June 2018.

Finally, the proposed Audit design and methodology was shared with SADiLaR and RMA representatives via email to ascertain specific requirements and understand alternative methodologies which were relevant, and the design was adapted accordingly. SADiLaR agreed on the methodology adopted in preparation for an Audit design workshop.

8.1.3 AUDIT DESIGN PROCESS

8.1.3.1 Approach

The Audit design process commenced with the identification and understanding of current frameworks available to conduct HLT Audits. This investigation yielded that there are not many frameworks available; however, there was substantial reference in our sources to the Language Resources and Evaluation Conference (LREC) and Basic Language Resource Kit (BLaRK). After a further familiarisation with the BLaRK concept used in the South African HLT Audit conducted in 2009 [8], a decision was made to use the BlaRK framework and adapt it with inputs from experts in the field.

An important characteristic of BLaRKs is that they are able to leverage on knowledge within HLT communities and are language independent.

The BLaRK concept process is outlined in Figure 2 below.

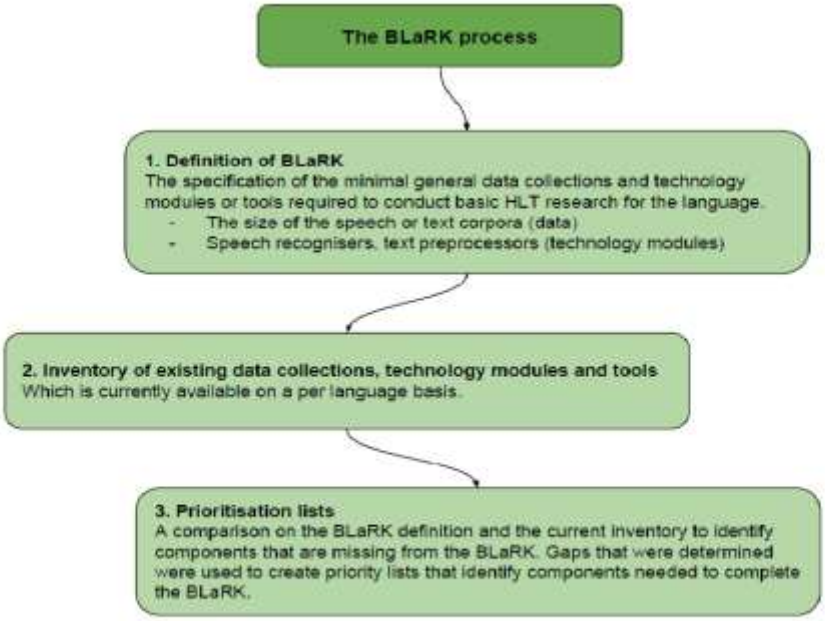


FIGURE 2: BLARK PROCESS

The 2009 HLT Audit classified HLT resources into the following categories, which are defined in the BLaRK concept:

Data

- Linguistic data sets or collections (speech or text), in a machine-readable form, used to create, evaluate and improve HLT modules.
- Includes corpora, lexicas and grammars.

Modules

- Basic software units or processes usually required to create HLT applications and products.
- Includes part-of-speech taggers, sentence tokenisers, language models, acoustic models.

Applications

- Categories of different application areas where HLT is used.
- Includes application domains such as speech input, document production, proofing/authoring tools, and translation.

The reasons for selecting the BLaRK concept as framework for the 2017/8 Audit were the following:

- **Consistency** with the 2009 Audit was important, since the 2017/8 Audit is an update of the previous one.
- The BLaRK process is **relevant** to the various HLT components (**data, modules and applications**).
- The BLaRK process follows a **logical** and **all-inclusive** approach to conducting an audit.

8.1.3.2 Process

The 2017/8 Audit followed a similar design process to that of the 2009 Audit [2]. Figure 3 below outlines the 2009 HLT Audit design process.

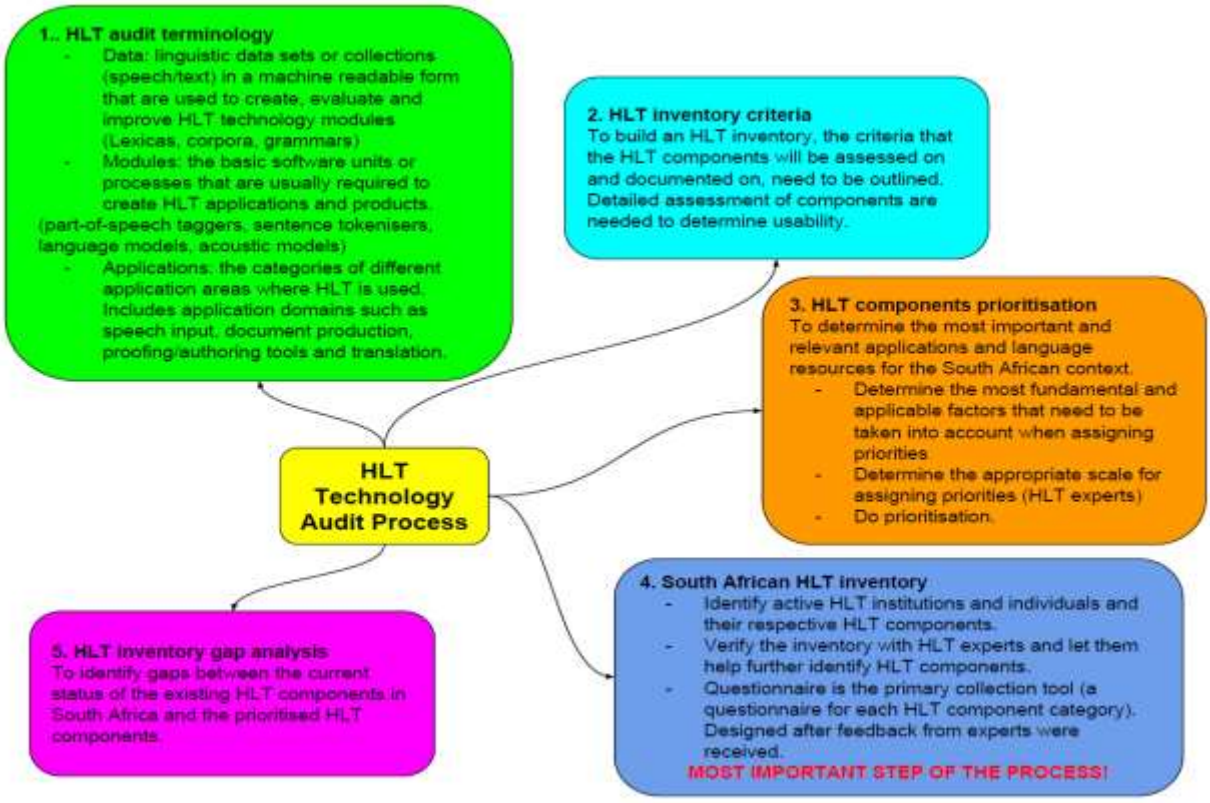


FIGURE 3: HLT AUDIT PROCESS

The 2018 Audit design process involved the following steps:

- Understanding the 2009 Audit design process
- Determining the process to be followed for the 2017/8 Audit, including
    - HLT Audit terminology development
    - HLT inventory criteria selection
    - HLT components (and priorities) definition
    - HLT Audit execution
    - HLT inventory gap analysis
- Deciding on the resource categories to be included in the design
- Reviewing the 2009 HLT Audit tool (questionnaire) and determining fit-for-purpose for new Audit design
- Obtaining a thorough understanding of the data analysis techniques used in the 2009 HLT Audit.

8.1.4 AUDIT DESIGN WORKSHOP WITH INITIAL EXPERTS

An Audit Design workshop was held on 30 and 31 August 2017 at the CSIR in Pretoria.

8.1.4.1 2017/8 HLT Audit design workshop preparation

In designing the 2017/8 Audit, we consulted experts in the field of HLT to assist with modernising the previous design. We hoped this would have the added benefit of creating buy-in into the process. We opted to host an HLT Audit design workshop with these experts.

The aim of the workshop was to update the list of component categories (taken from the 2009 Audit), as well as to update the survey tool. This was needed, as new types of resources have been developed since the previous Audit.

The decision to host the workshop was made because during the 2009 HLT Audit, discussions between experts proved to be effective when they had to decide on the component categories and the questions for the questionnaire. These discussions took place during a similar workshop in 2009.

In preparation for the workshop, a preliminary list of experts was identified and they were assigned to relevant categories. The purpose of this exercise was to ensure that there would be representation of each category present at the workshop.

8.1.4.2 HLT Audit design workshop proceedings

The workshop participants included experts from the University of Pretoria, North-West University, University of KwaZulu-Natal, MuST, CTexT, University of South Africa, Department of Arts and Culture, University of Witwatersrand, Limpopo University and the HLT Research Group of the CSIR Meraka Institute. The programme of this workshop is attached as Annexure A.

The purpose of this workshop was to formulate and update the following in preparation for the Audit:

- Update the component categories which form the basis of the Audit:
    - **Data** - linguistic data sets or collections (speech/text) in a machine readable form that are used to create, evaluate and improve HLT technology modules.
    - **Modules -** the basic software units or processes such as lexica, corpora and grammars that are usually required to create HLT applications and products (POS, tokenisers, language models, acoustic models).
    - **Applications -** categories of different application areas where HLT is used (in domains such as speech input, document production, proofing/authoring tools and translation).
- Prioritise text and speech components to be included in the Audit.
- Obtain inputs into the Audit questionnaire.
- Obtain inputs into the list of institutions which would be included in the Audit.

The workshop commenced with the CSIR providing an overview on the Audit and the purpose of the participation of HLT experts in the Audit design process. Prof Justus Roux also provided an overview on SADiLaR.

8.1.4.2.1 Working groups

The HLT experts participating in the workshop were allocated into *text* and *speech* working groups according to their areas of expertise, and further tasked with providing inputs in the *data*, *module* and *applications* subcategories. The purpose of allocating these experts into groups was to ensure that consensus could be reached on the subcategories of components and that all possible components had been accounted for.

8.1.4.2.2 Updating and prioritising component categories

The component categories were taken from the Audit of 2009 and updated by initial experts prior to the HLT Audit design workshop. These component categories were then distributed in advance to the experts to review prior to the workshop. During the workshop, the 2009 component categories outlined as listed below were modified as set out in Annexure B.

- Current components for text data
- Current components for speech data
- Current components for data in the form of corpora
- Current components for text modules
- Current components for speech modules
- Current components for text applications
- Current components for speech applications

The working groups were tasked with the following:

- Reviewing the 2009 components
    - Determining what is still relevant
    - Determining what needs to be changed, added or deleted
- Ensuring that components pertaining to all languages are covered.

The working groups agreed that the Modules and Applications categories are no longer applicable. A decision was made to only include the *data* category and then combine the Modules and Applications categories into a Software category. A Model category was added for Speech components.

The Data, Model and Software categories were then updated and relevant metadata was added. The final version of the component categories with metadata is attached to this report as Annexure B.

8.1.4.2.3 Obtain inputs into the Audit questionnaire

The various sections of the questionnaire were made available and all experts were provided with an opportunity to give inputs. The previous questionnaires used in the 2009 HLT Audit were proposed as a starting point for the discussions and inputs. These questionnaires are attached as Annexure C.

There were a number of modifications to the questionnaire, including the following:

- There would be a **Your Information** page at the start of the questionnaire. This section would include:
    - Contact details (name, surname, email address and contact number)
    - Affiliation of the person submitting the resource (the user)
    - A question on whether or not the user would like SADiLaR to contact them to add this resource as a catalogue item on the RMA.
    - A question on whether or not the user would like to be kept informed of the publication of the Audit report and further developments.
- There would be a **Resource Type** page, where the type of resource being entered can be classified as Text, Speech or Multimodal. The resource can also be classified as a Data, Model or Software resource. After classifying the resource, the user has to select the category in which their resource falls.
- There would be **Required Information** pages for Data, Model and Software. The minimum requirements section would include:
    - The resource name
    - A short description of the resource
    - Keywords to easily search for the resource
    - Language(s) of the resource
    - Availability of the resource (research, commercial or open/freely-available)
    - The cost of the resource (if any).
- There would be separate **Technical Description** pages for Data, Model and Software. These pages were then updated based on the technical description information required for Data, Model and Software.
- There would be Availability, Quality and Documentation pages for Data, Model and Software.

8.1.4.2.4 Obtain inputs into the list of institutions

A list of proposed institutions to be invited to participate in the 2017/8 Audit was prepared in advance. This session included an update of the proposed institutions and to confirm their representative details.

All experts were requested to provide inputs into the list of institutions and their representatives.

A decision was made to split the institutions list into an A list and a B list. The A list would be for the experts who are active in the HLT community. The B list would be for representatives from libraries, National Language Units, private companies and other departments not yet identified from all universities in South Africa. As time was not sufficient in the session, a Google document was shared with the experts to populate with their inputs.

Three representatives from each working group (Text and Speech) were nominated to form a smaller task team which would assist the project team to finalise the component categories and give feedback on the tool. These representatives were:

- **Text**: Roald Eiselen (CTexT), Laurette Pretorius (UNISA) and Marissa Griesel (UNISA)
- **Speech**: Daniel van Niekerk (MuST), Jaco Badenhorst (CSIR) and Georg Schlunz (CSIR).

8.1.5 MINI-WORKSHOPS WITH EXPERTS

Mini-workshops were scheduled with these representatives on 16 and 17 October 2017 at the CSIR. The component categories were finalised and a workflow design for the Audit was proposed and revised by the representatives. They also provided valuable inputs into the structure of the Audit questions.

8.1.6 AUDIT DESIGN WORKFLOW

The questionnaire used for the 2009 HLT Audit was in the format of a Microsoft Excel spreadsheet. The first column contained the questions and the subsequent columns contained the different official South African languages. When entering a resource, the participant then had to fill in all the details of the resource in the cells provided. Some of the HLT experts who attended the 2017/8 HLT Audit workshop also participated in the 2009 HLT Audit. They commented that the format of the 2009 Audit had not been user-friendly and the spreadsheet had been difficult to complete. Navigating through the spreadsheet became cumbersome when large amounts of information needed to be entered. A suggestion was made to rather make use of an online questionnaire or something similar.

At the workshop, the questionnaire from the 2009 HLT Audit was studied and discussed. New questions were added and the existing fields were updated. Based on these discussions, the design team developed the Audit design workflow. The workflow was presented to the identified representatives for Text and Speech at the mini-workshop that was held.

The Audit design workflow was then used to design an online questionnaire using LimeSurvey. Detailed information on LimeSurvey and its selection is provided in the section below.

In the following sections, we provide an overview of the design of the 2017/8 Audit.

8.1.6.1 High-level workflow for 2017/8 Audit

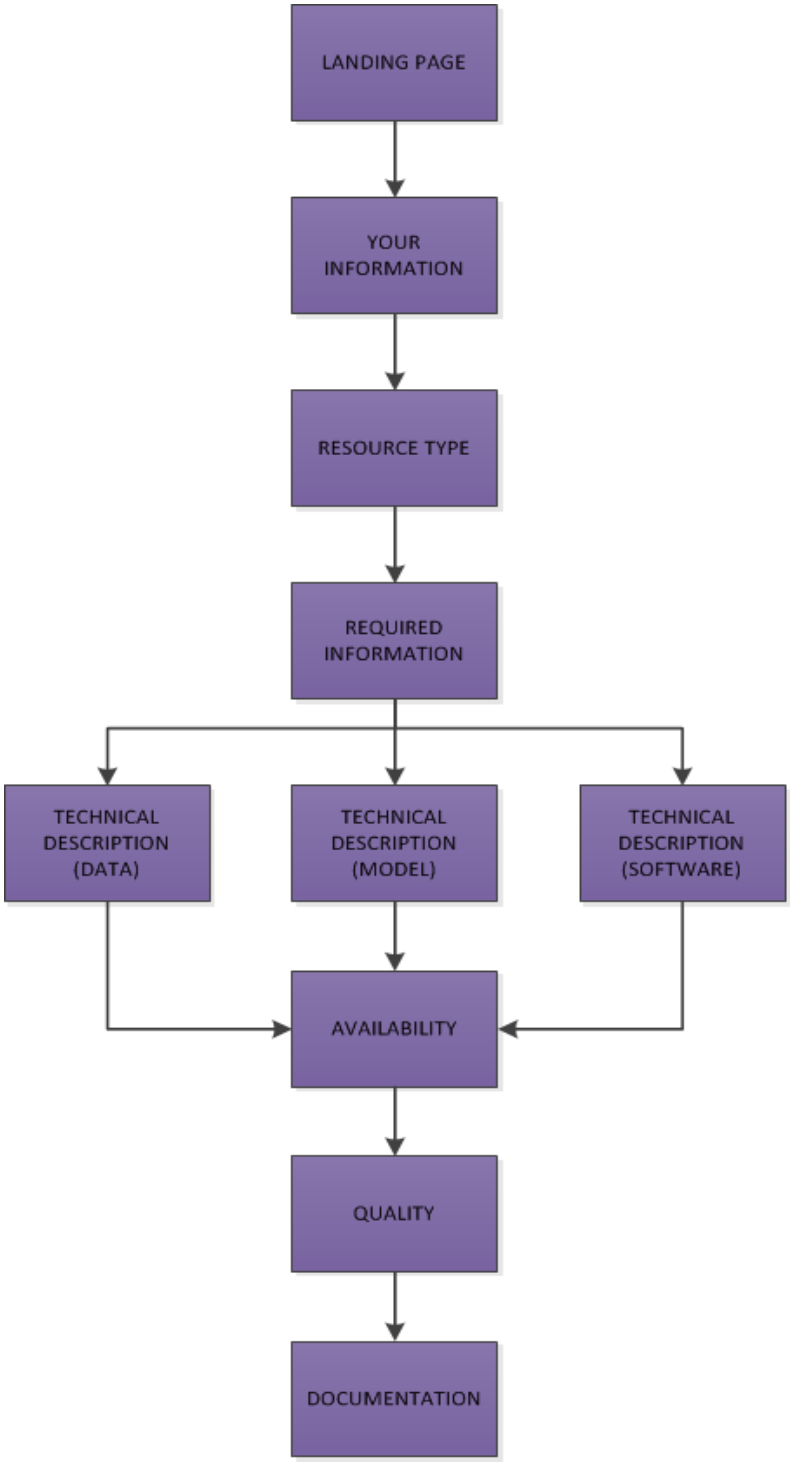


FIGURE 4: HIGH-LEVEL WORKFLOW DESIGN

8.1.6.2 Audit landing page

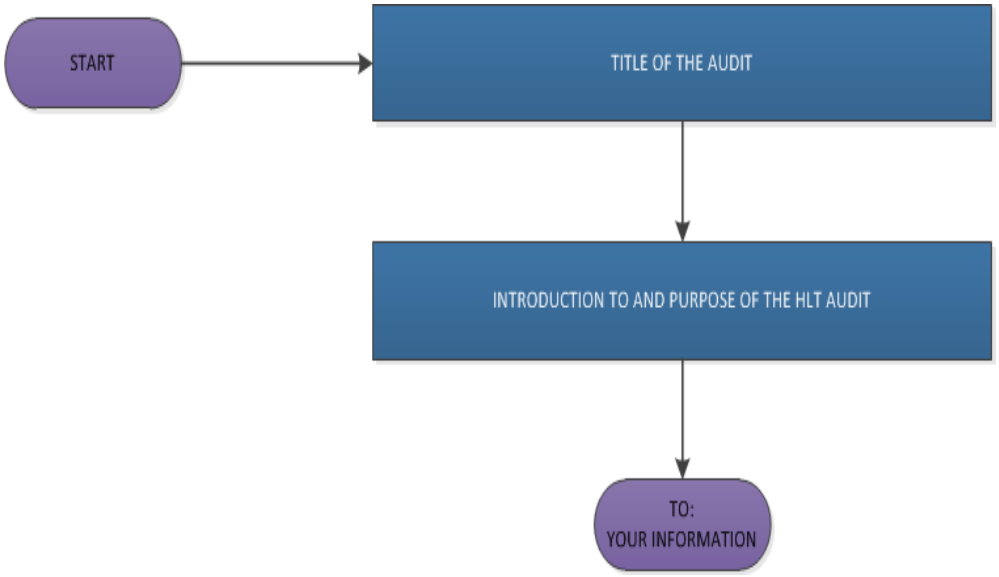


FIGURE 5: AUDIT LANDING PAGE DESIGN

8.1.6.3 Participant information page

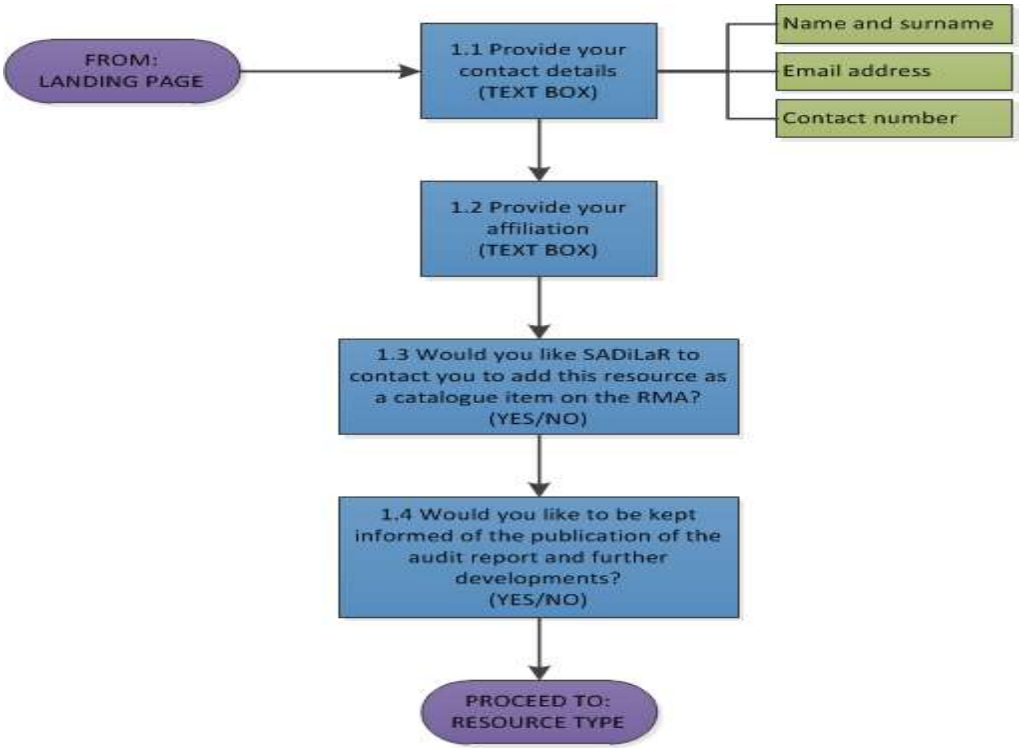


FIGURE 6: PARTICIPANT INFORMATION PAGE DESIGN

8.1.6.4 Resource type page

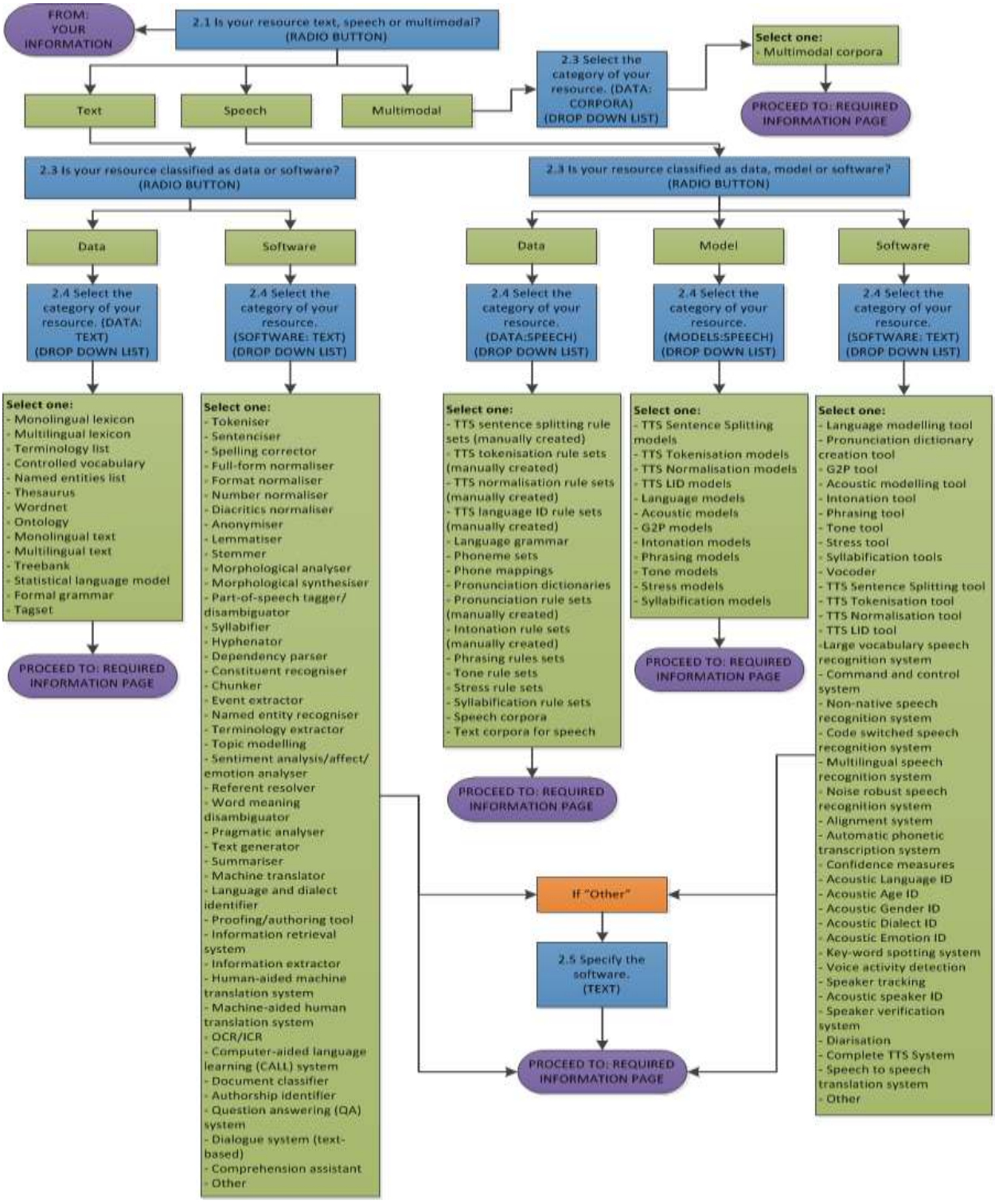


FIGURE 7: RESOURCE TYPE PAGE DESIGN

8.1.6.5 Required information page

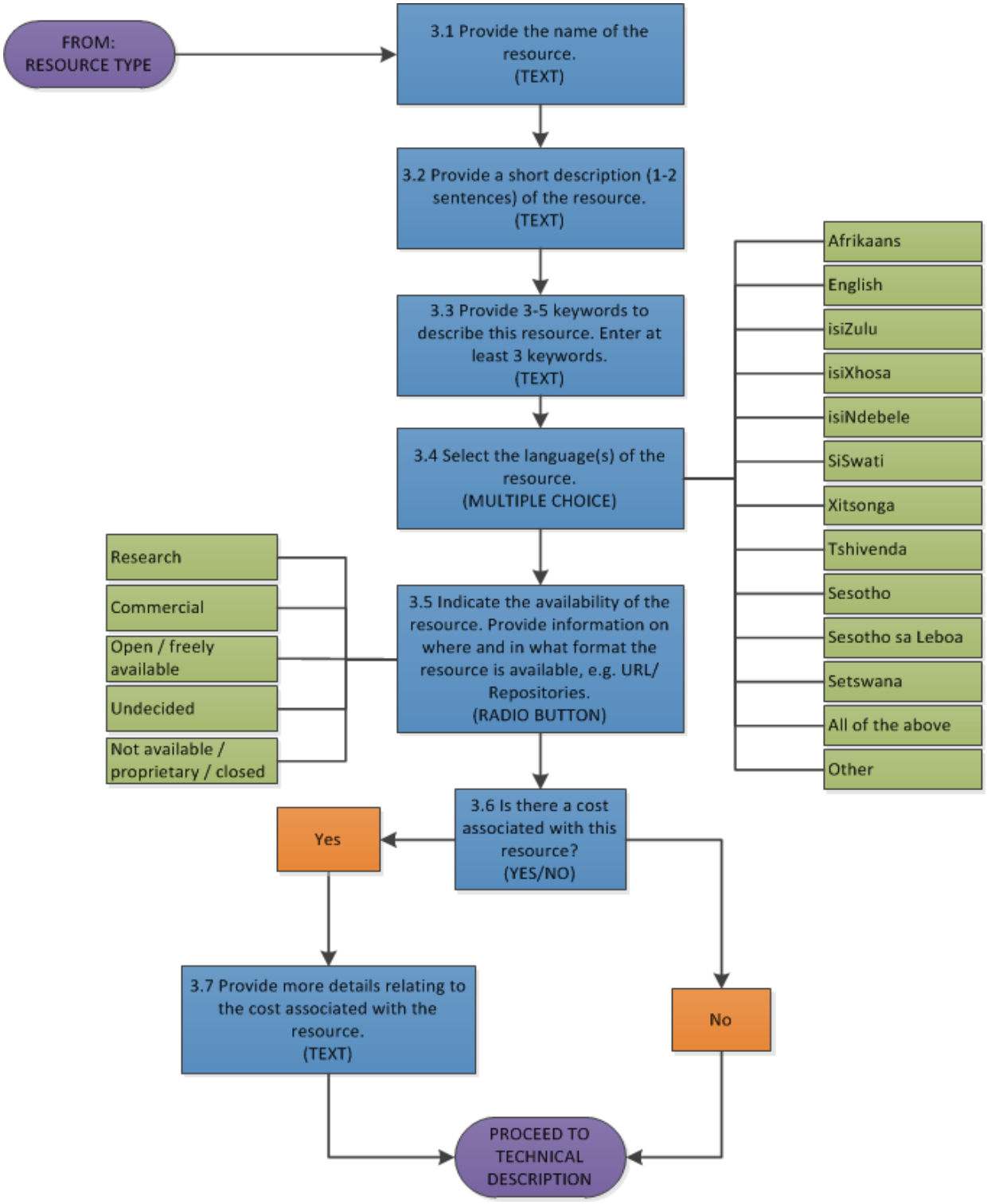


FIGURE 8: REQUIRED INFORMATION PAGE DESIGN

8.1.6.6 Technical description (data) page

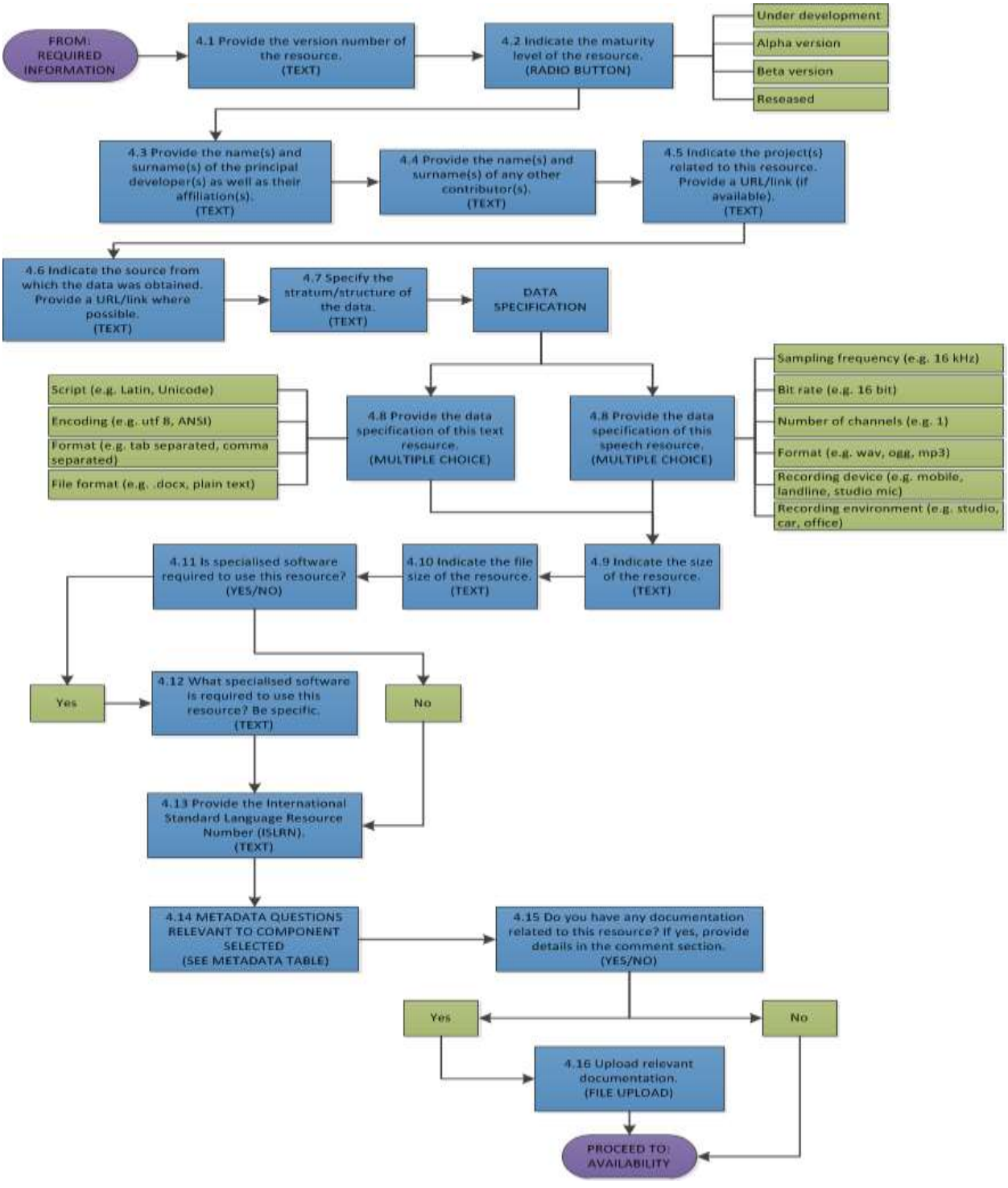


FIGURE 9: TECHNICAL DESCRIPTION (DATA) PAGE DESIGN

8.1.6.7 Technical description (model) page

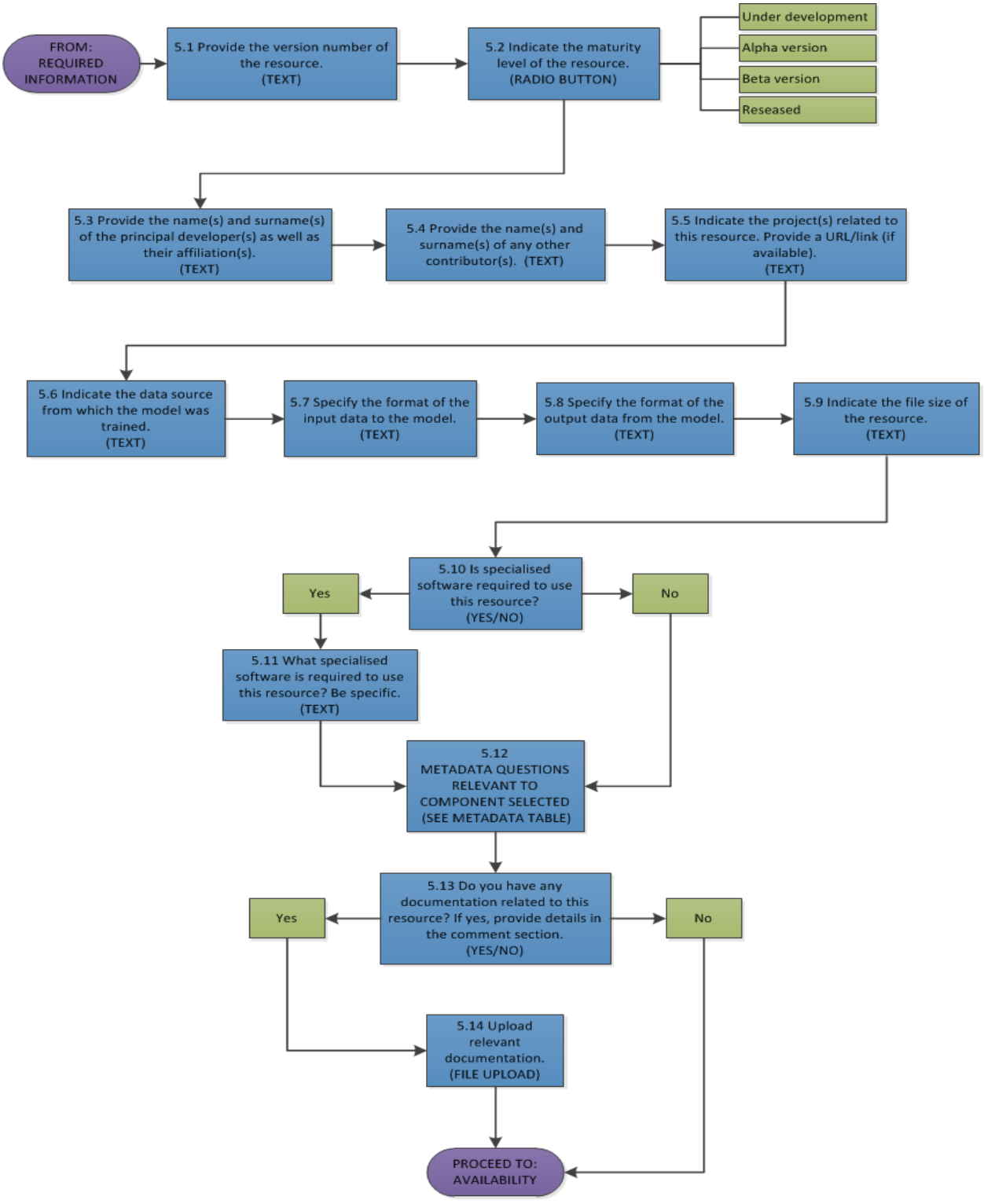


FIGURE 10: TECHNICAL DESCRIPTION (MODEL) PAGE DESIGN

8.1.6.8 Technical description (software) page

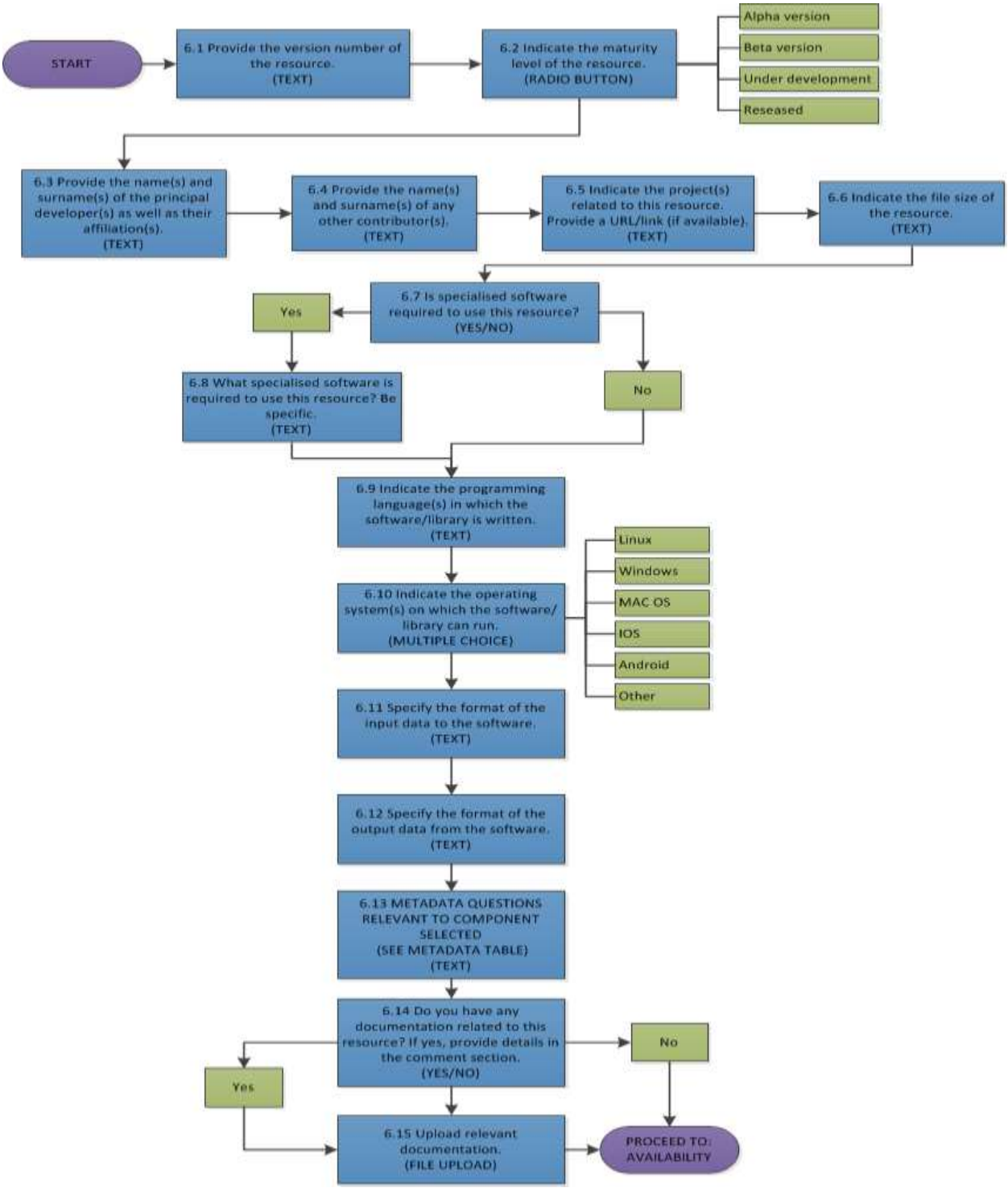


FIGURE 11: TECHNICAL DESCRIPTION (SOFTWARE) PAGE DESIGN

8.1.6.9 Availability page

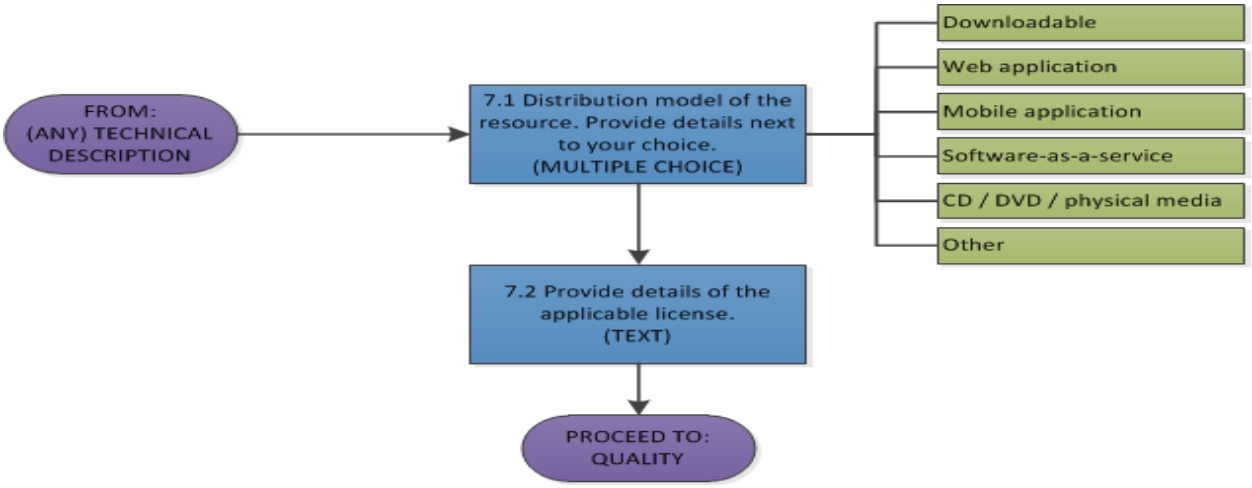


FIGURE 12: AVAILABILITY PAGE DESIGN

8.1.6.10 Quality page

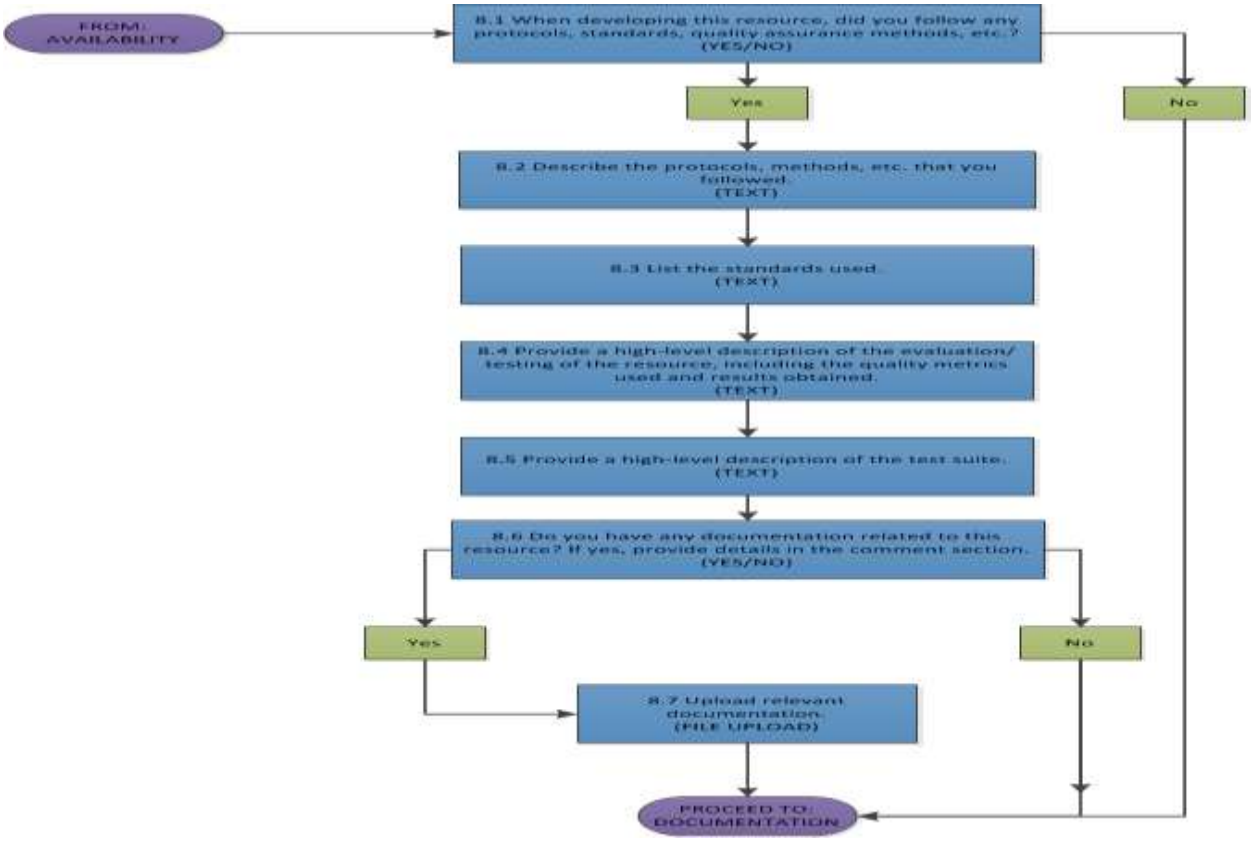


FIGURE 13: QUALITY PAGE DESIGN

8.1.6.11 Documentation page

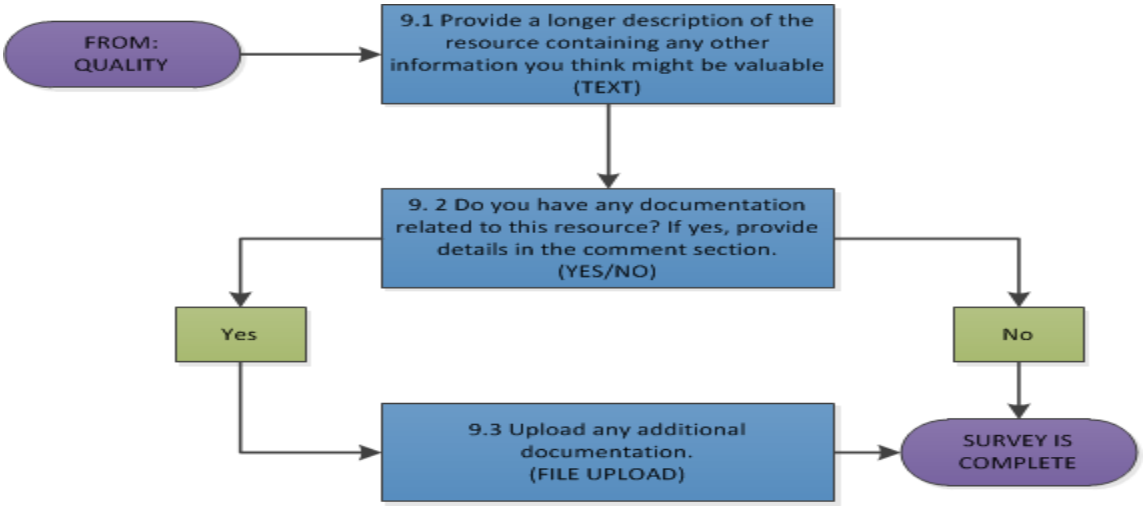


FIGURE 14: DOCUMENTATION PAGE DESIGN

8.2 WP2: 2017/8 HLT AUDIT INSTRUMENT DEVELOPMENT

8.2.1 INTRODUCTION

Based on the design sketched out above and using the online tool called LimeSurvey, the Audit instrument was then developed. Careful consideration was given to the types of questions to be used for each piece of information required. Usability and user-experience further guided decisions on layout and wording.

A beta version of the Audit instrument was tested with a small group of beta testers and the feedback was incorporated to the extent possible given the constraints of the online tool. In parallel, the content for the email eliciting participation in the Audit was carefully crafted.

Once finalised, the methodology and instrument were shared with the SADiLaR and RMA representatives for final sign-off before commencement of the Audit with confirmed stakeholders.

The HLT Audit instrument development work package was planned to include the following activities:

- Design methodology to capture updated information.
- Identify stakeholders to participate in the Audit.
- Draft instrument for the Audit (questionnaire and interview questions).
- Test instrument and methodology on small group.
- Finalise the instrument and methodology.
- Confirm stakeholders to participate in the Audit.
- Sign-off on methodology and instrument by SADiLaR and the RMA.

The work was conducted as planned and is described in more detail below.

8.2.2 METHODOLOGY AND AUDIT TOOL REQUIREMENTS

The previous Audit conducted in 2009 distributed the questionnaire by email. The questionnaire was in the form of a Microsoft Excel spreadsheet, which included a detailed explanation of how to complete the questionnaire.

Some of the 2017/8 Audit workshop participants also participated in the 2009 HLT Audit and were of the opinion that the method in which the questionnaire was presented was not user-friendly. The Microsoft Excel version was also admin intensive to capture the results. The participants then suggested that another tool be used to conduct this Audit.

The workshop participants provided us with a number of factors that we needed to take into consideration when investigating available Audit tools, such as cost, functionality and where and how to host the tool. We used the following requirements as a basis for selecting an Audit tool:

8.2.2.1 Workflow

- Generic minimum requirements for data, model and software
- Data, model and software questionnaires

8.2.2.2 Client/user requirements

- Online tool
- Attractive to the user (modern look and feel)
- Clear and easy to use
- Logical flow
- Questionnaire functionality:
    - Drop down menu
    - Multiple choice
    - Yes/No questions
    - Short narrative description
    - Document/file upload

8.2.2.3 Technical requirements

- Accessible to institutions nationally (cloud-based or open platform)
- Only accessible to institutions invited (not available to public)
- Ability to for many users to complete the survey simultaneously
- Ability to store large documents (in specific format(s))
- Ability to export to a database

- Ability to convert raw data into Excel format

8.2.2.4 Success criteria

- Completeness of information received
- Scalability

8.2.2.5 Outputs

- Export raw data to Excel format (required)
- Dashboard with a consolidated view of the Audit outcome (optional)
- Transfer to client website/database (required)

8.2.3 Selection of Audit tool

We conducted an Internet search for online questionnaire/survey tools which would suit our needs. Unfortunately, some tools had the appropriate functionality, but could not be used due to cost. We also contacted two companies and obtained quotes for the development of a bespoke Audit tool. This too proved to be too costly. It was then decided to opt for an online tool called LimeSurvey.

LimeSurvey (https://www.limesurvey.org/) is worldwide leading open source survey software which is available as a Software-as-a-Service or as a self-hosted Community Edition. LimeSurvey is a powerful survey tool which is highly customisable to suit the user's needs.

Due to a number of reasons, we opted to host the **Community Edition** version on a local server. These reasons included:

- Affordability (LimeSurvey is open source software and the Community Edition is free of charge)
- Customisable (this version is easy to customise to the user needs and can be set up by a non-technical person)
- Fit-for purpose
- Compatibility (this version could easily be hosted on our server/platform)
- Accessibility (this version is accessible using a screen reader).

We then used the user manual and the community forum to learn the functionalities offered by LimeSurvey.

8.2.4 Configuration of the Audit tool

LimeSurvey offers the functionality of creating a questionnaire using an existing template, or completely from scratch. Since none of the existing templates met our needs, we opted to develop our own survey.

8.2.5 QUESTIONNAIRE PROPERTIES

The properties for every questionnaire created can be changed to suit specific needs. To create a new questionnaire, the following was required:

- Questionnaire title
- Description
- Welcome message
- End message.

There are general settings for each created questionnaire which can be changed as needed. These include:

- Administrator contact details
- How the questions are displayed (question by question vs question group by question group vs all-in-one)
- Navigation settings (will the user be allowed to navigate backwards or not)
- Displaying the number of questions
- Displaying the progress a user is making
- Access to the questionnaire (open to everyone vs open to anyone who has a token).

The properties page is shown in Figure 15 below.

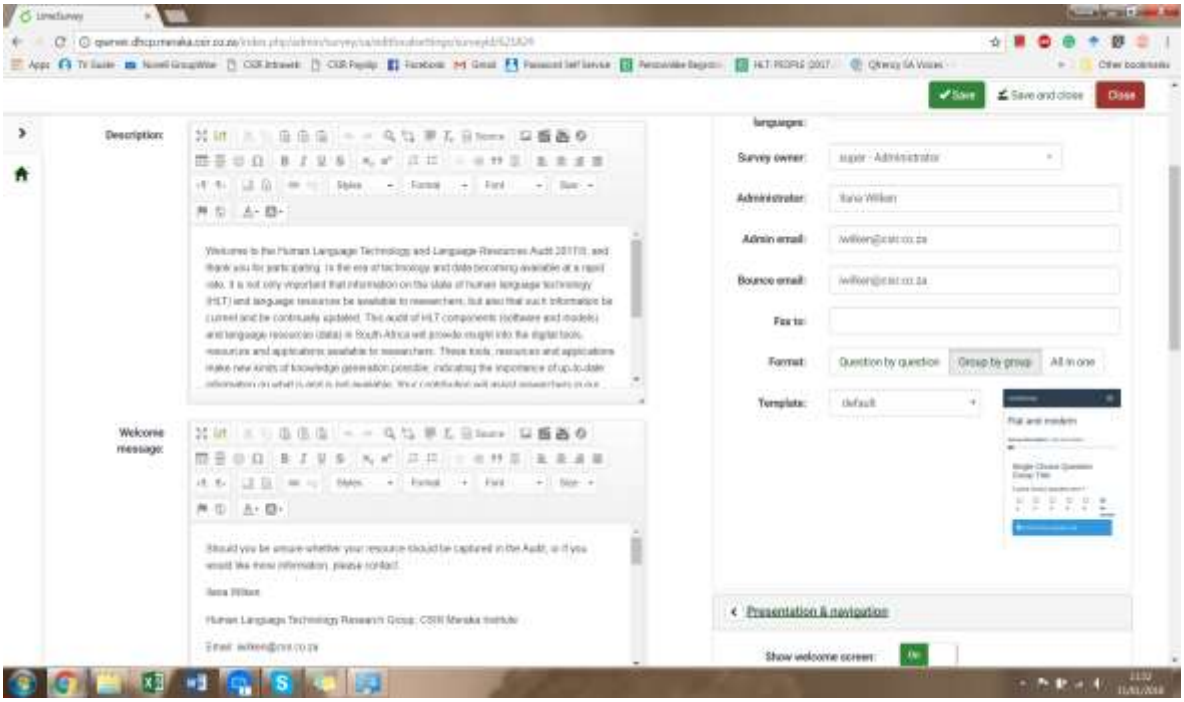


FIGURE 15: GENERAL SETTINGS AS DISPLAYED IN BACK-END

8.2.6 PARTICIPANTS

We decided to only grant access to the questionnaire to anyone who has a token, for security purposes[2]. A participants table was created by uploading a list of potential participants, including their names, surnames and email addresses. Each token is valid for a certain number of uses. In this instance we opted for 100 uses (i.e. to upload 100 resources). After the participants table is created, the tokens are generated, after which each participant is sent a personalised email containing the link to the questionnaire as well as their unique token.

8.2.7 THE QUESTIONNAIRE

The development of the questionnaire consists of 2 sections, namely the back-end development, and the front-end interface.

8.2.7.1 Back-end development

The questionnaire workflow - which was finalised at the mini-workshops - was used as the basis for the online questionnaire populated in LimeSurvey.

Every question was manually created. This entailed:

- Typing the question
- Defining the question type
    - short text, long text, multiple choice, multiple choice with comments, radio list, radio list with comments, drop down lists, yes/no questions, file upload questions, etc.
- Adding the predetermined answer options (for the multiple choice and radio list type questions)
- Creating conditions for certain questions (for example, ask Question 3 if the answer to Question 1 is “blue”).

The summary of a text input question and sub-questions to a question as displayed in the back-end are shown in Figures 16 and 17 below.

[2] It is possible to open the survey to anyone who has the link. However, this would pose a security risk to whoever is hosting the survey should they not have adequate security on their hosting platforms. An option would be to configure the survey to allow automatic token generation. This will also be investigated and put forward in the recommendations report under WP5.

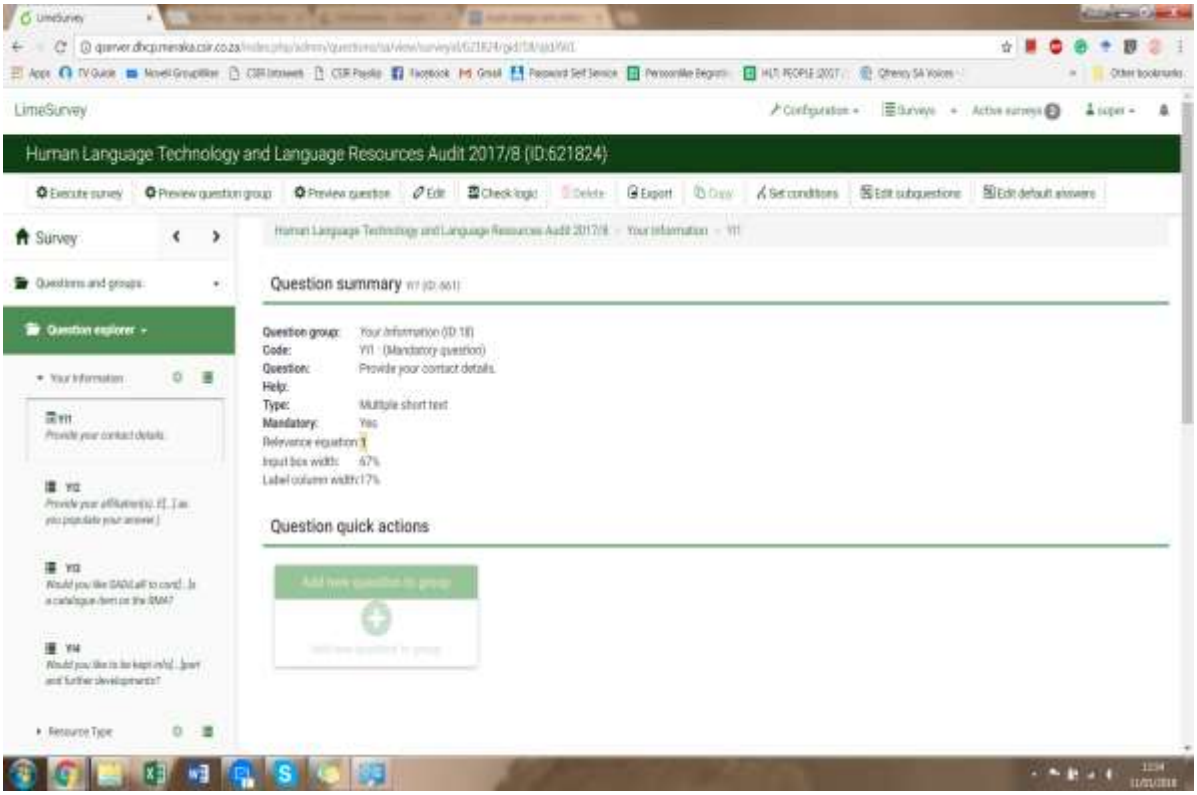


FIGURE 16: SUMMARY OF A TEXT INPUT AS DISPLAYED IN THE BACK-END

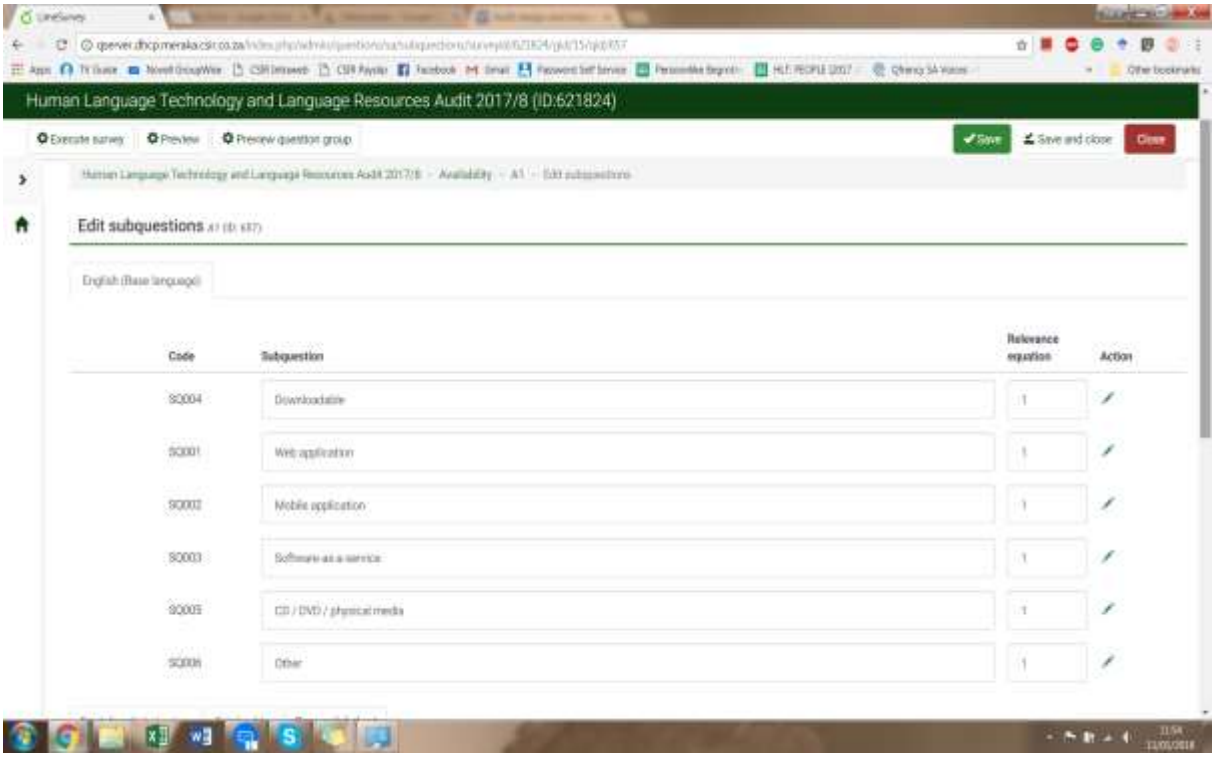


FIGURE 17: SUB-QUESTIONS TO A QUESTION AS DISPLAYED IN THE BACK-END

8.2.7.2 Front-end interface

The front-end of the questionnaire is the actual survey in which users will participate. The front-end pages are attached as Annexure D.

The front-end of the survey consists of the following pages:

- The ***Landing page*** that provides a brief introduction on the HLT Audit, including an overview of how the HLT Audit will work.
- The ***Your Information page*** allows users to complete their general information such as name, contact, affiliation and cataloguing resources under the RMA.
- The ***Resource type page*** allows users to select the type of resource that they are uploading, such as text, speech or multimodal.
    - The ***Resource type page – text selection.*** The user then has to select whether their resources are Data or Software. Finally, under either Data or Software, the user may then select a category into which their resource will be classified.
    - The ***Resource type page – speech selection.*** The user then has to select whether their resources are Data, Model or Software. Finally, under Data, Model or Software, the user may then select a category into which their resource will be classified.
    - The ***Resource type page – multimodal selection.*** The user then has to select Multimodal corpora.
- The ***Required information page*** allows user to complete information on the resource they will be uploading. This information includes the name, description and keywords associated with the resource, language, availability and cost of the resource.
- The ***Technical description page*** allows users to complete further technical information on the resource under a ***Data, Model*** and ***Software*** pages – this is dependent on the resource type selected earlier in the survey.
- The ***Availability page*** allows users to indicate the models of distribution and licenses associated with their resources.
- The ***Quality page*** allows users to select and complete any protocols, standards and QA methods followed in their resources. Should users select **YES** to this question, they will be prompted to move to the next page which requires detailed input.
- The ***Documentation page*** allows users to include a more detailed description on their resource which may not have been covered earlier in the survey.
- The ***End page*** thanks users for their participation in the HLT Audit and acknowledges partners.

8.2.8 BETA TESTING OF THE AUDIT TOOL

The beta version of the questionnaire was tested by members of the HLTRG, as well as a representative from the RMA. Valuable feedback was received and this assisted in finalising the questionnaire for Audit execution in early December 2017.

Some of the changes required based on the feedback received included:

- Refining/rewording questions
- Changing conditions on certain questions (and only show them when necessary)
- Adding an "other" option to some multiple choice questions
- Adding a list of definitions for the components (see Figure 18).

An aspect which was raised as a concern during beta testing and for which we have not yet found a resolution, is that of enabling duplication of information if similar resources are being submitted to the Audit, e.g. similar speech corpora in different languages, where much of the information is the same and only some details are changed. This is something that we will need to address in a future version of the tool.

The beta testers also assisted with checking consistency in the questions as well as error-checking.

A separate website (https://sites.google.com/view/hlt-Audit-definitions/home) was created for the list of definitions for the components, as adding all the definitions to the questionnaire would have cluttered the layout and overwhelmed the participant. The URL to the website was included in the invitation email sent to the participants, as discussed below.

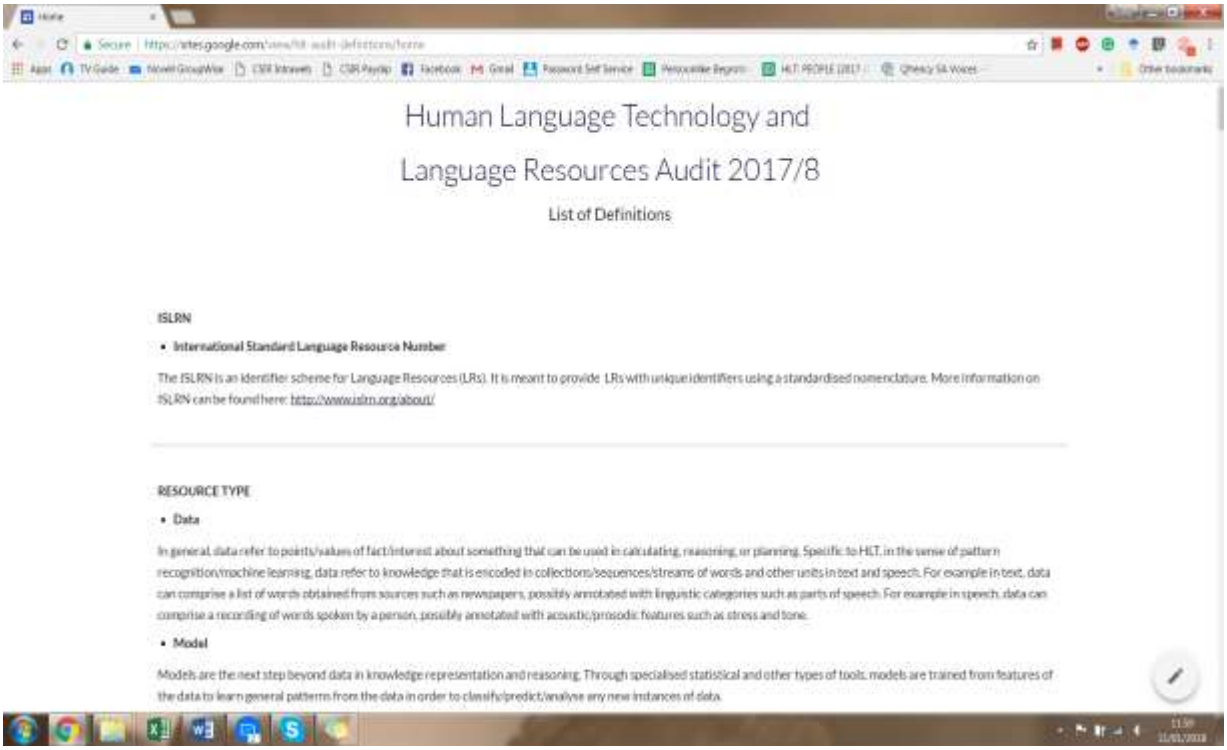


FIGURE 18: LIST OF DEFINITIONS OF THE COMPONENTS AS DISPLAYED ON THE PAGE OF THE SURVEY

8.3 WP3: 2017/8 HLT AUDIT EXECUTION

8.3.1 INTRODUCTION

The methodology and instrument developed and signed-off on in WP2 were used to conduct the actual Audit (WP3). The Audit Execution WP had been planned to entail interviews but the change in methodology to an online survey, meant that the work became more focused on engaging with stakeholders to facilitate and ensure their participation.

The results of the online Audit will be captured and analysed in WP4 and will be reported on in a project report, and presented at a dissemination workshop and via publications.

8.3.2 INVITATION TO PARTICIPATE

During the Audit design workshop, a decision was made to extend the Audit beyond HLT resources, to include generic language resources as well. This was communicated in the email (Annexure E) informing potential participants that the Audit is taking place (Audit notification email). This email was sent to known members of the HLT community, as well as government departments, national lexicography units, publishers, private companies, professional associations, tertiary institutions and the mailing lists of the National HLT Network (NHN) and the Resource Management Agency.

The email was sent out on 5 and 6 December 2017. This email provides background information to the Audit, and also requested the potential participants to provide us with their contact details should they wish to participate. The email also requested them to forward the email to other potential participants that we might not know of.

A formal automated invitation email (Annexure F) to participate in the Audit was sent out on 6 December 2017 (and is continually being sent as and when feedback to the first email is received). This Audit invitation email contains a link to the questionnaire titled “Human Language Technology and Language Resources Audit 2017/8”, the participant’s unique token, which is valid to enter up to 100 resources, as well as a link to the list of the definitions of the components.

8.3.3 SURVEY RESPONSES

Since commencing with the Audit, the responses are being monitored at regular intervals. Weekly reminders are also being sent to participants to submit their information online before the Audit deadline of 28 February 2018.

In order to prevent spamming the participants on the mailing list with regular reminders, participants were asked to give an indication of whether or not they intend to participate, and the reminders were then only sent out to those participants.

8.4 WP4: 2017/8 HLT AUDIT DATA ANALYSIS, CONSOLIDATION, REPORTING AND DISSEMINATION OF RESULTS

8.4.1 INTRODUCTION

Once the Audit was completed and reported on, we commenced with the analysis of the data, the consolidation of the data, reporting and dissemination of results.

The data on all resources included in the RMA (now SADiLaR) database (catalogue and index) was shared with our team. This data was split into two datasets: the 2009 Audit data (referred to as the **2009 data** in this report although it was only uploaded onto the RMA in 2013); and the resources uploaded onto the RMA from 2014 to 2017 (referred to as the **2014 data** in the report). The third dataset is the data from the 2018 Audit (referred to as the **2018 data** in the report).

All data categories (from all three datasets) were mapped to one another and the results analysed to determine both resource development trends and gaps in resources available in all South African languages. At least one research paper on the Audit will be submitted to a conference, and a one-day workshop will be held in July 2018 to disseminate the findings of the Audit to the HLT R&D community and other interested stakeholders, including the DAC HLT Expert Panel.

8.4.2 DATA CATEGORIES IN THE ANALYSIS

The 2009 and 2014 datasets divided the resources into the following categories:

- **DATA -** split between **text** and **speech** resources
- **MODULES** - split between **text** and **speech** resources
- **APPLICATIONS** - split between **text** and **speech** resources
- **TOOLS** - split between **text** and **speech** resources.

The 2018 dataset divided the resources into the following three categories based on the modernised design emanating from the Audit design workshop held in August 2017:

- **DATA** - split between **text** and **speech** resources
- **MODELS** - only includes **speech** resources
- **SOFTWARE** - split between **text** and **speech** resources and consolidates the previous MODULES, APPLICATIONS AND TOOLS categories.

8.4.3 ANALYSIS OF THE 2018 AUDIT RESULTS

Below, we report on a detailed analysis of the data received in the 2018 Audit. The detailed analysis starts with an overview of the resources submitted and the institutions which submitted these resources, and then progresses into a more detailed analysis of the data, focusing on the details per resource category in terms of the maturity and availability of the resources.

8.4.3.1 Participation in the 2018 Audit

Experts from the HLT community proposed relevant institutions to invite to participate in the 2018 Audit. These institutions included: universities, government departments, research institutions, language technology companies, publishers, national lexicography units, and professional associations.

A total of **53 invitations were sent** to these institutions. We also requested that these institutions share the information about the Audit with others who may request tokens should they wish to participate. In our initial communication to participating institutions we had indicated that only resources that are not yet included in the RMA (now SADiLaR) database should be submitted. From the 53 institutions invited, **only eight participated** in the Audit. We followed up with the institutions that did not participate, and it was confirmed that they did not have any (new) resources to contribute to this audit.

8.4.3.2 Resources submitted per institution, category and language

Eight institutions participated in the 2018 Audit. These institutions submitted resources across the data and software categories, including speech and text. Figure 19 below provides an overview of the number of resources submitted per participating institution.

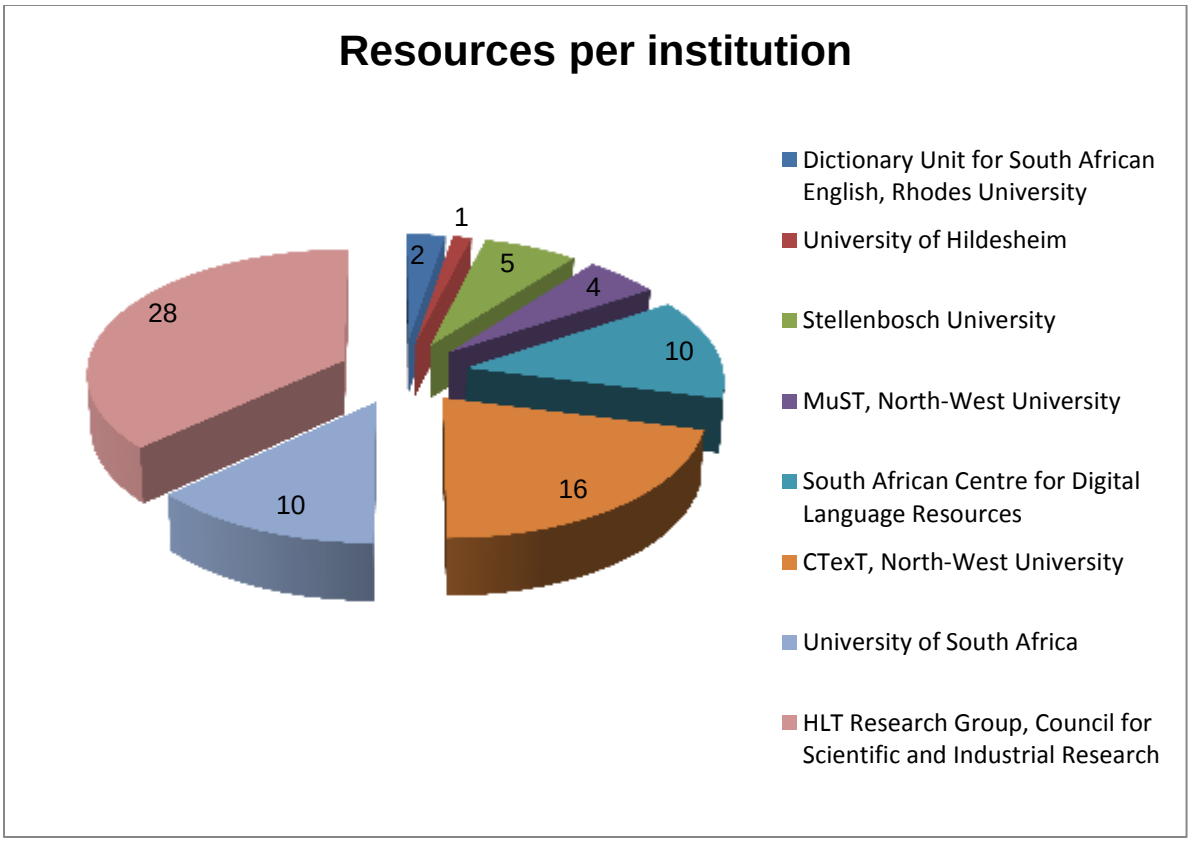


FIGURE 19: NUMBER OF RESOURCES PER INSTITUTION

In Table 2 below, we have broken down the submission of resources into resources submitted per institution, category and language. A total of 76 additional resources (75 language-specific and 1 language independent resources) were submitted during the 2018 Audit. Although only 26 submissions were made, a number of submissions covered multiple languages over a number of categories and resource types.

TABLE 2: RESOURCES SUBMITTED PER INSTITUTION, CATEGORY AND LANGUAGE

Name	No	Data	Software	Eng	Afr	Zul	Xho	Nbl	Ssw	Tso	Tsn	Sot	Nso	Ven
Dictionary Unit for SA English	2	Text		2										
University of Hildesheim	1	Text											1	
SU	2	Speech		1	1									
	3	Text		1	1		1							
MuST, NWU	4	Speech			1		1					1		1
SADiLaR	10		Text		1	1	1	1	1	1	1	1	1	1
CTexT, NWU	16	Text		1	2	2	1	1	1	2	1	1	2	2
UNISA	2		Text			1								1
	8	Text		2	1	1	1							3
HLT, CSIR	1		Speech	1	1	1	1	1	1	1	1	1	1	1
	12		Text	1	1	1	1	1	1	1	1	1	1	2
	15	Speech		3	3	1	1	1	1	1	1	1	1	1
TOTAL	**76**			**12**	**12**	**8**	**8**	**5**	**5**	**6**	**4**	**6**	**7**	**12**

8.4.4 RESOURCES PER CATEGORY, LANGUAGE AND RESOURCE TYPE - OVERVIEW

Fourteen resources were submitted for Data (Text) and 21 for Data (Speech). Forty resources were submitted for Software (Text) and one for Software (Speech). No resources were submitted for the Models category.

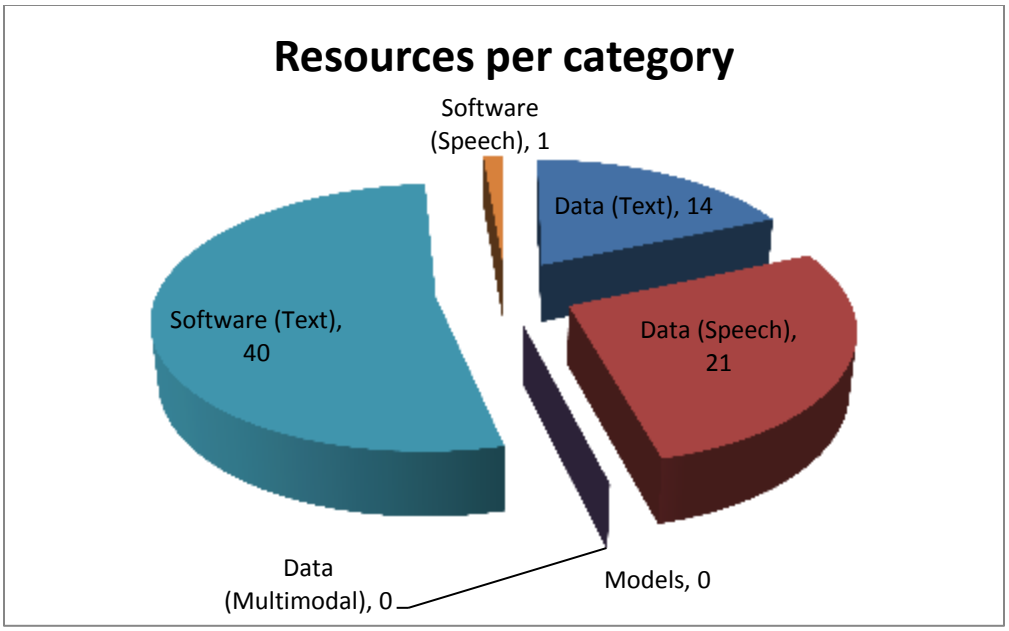


FIGURE 20: REPRESENTATION OF RESOURCES PER CATEGORY

As represented below in Figure 21, of the 76 resources submitted 12 were for English[3], Afrikaans and Setswana, eight for isiZulu, isiXhosa and Sepedi, five for isiNdebele, siSwati and Tshivenda, six for Sesotho and Xitsonga and one was a language independent resource, i.e. it can be applied to any language[4].

[3] English refers to South African English throughout.

[4] Note that language independent (LI) resources have been kept separate in the dataset and not aggregated across all languages. This is in line with the approach in the 2009 Audit and the preference of SADiLaR.

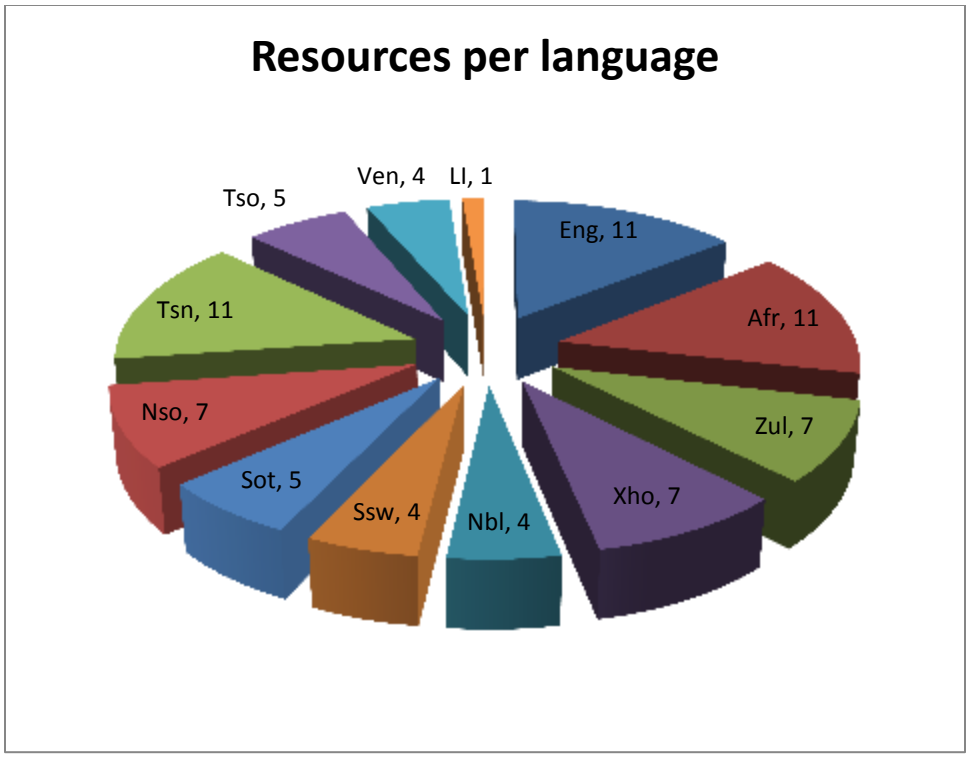


FIGURE 21: REPRESENTATION OF RESOURCES PER LANGUAGE

A more detailed view of the **resources submitted per language and per category** is provided in Table 3 below. Of the 76 resources submitted, the largest numbers were for English, Afrikaans, Setswana, Sepedi, isiZulu and isiXhosa.

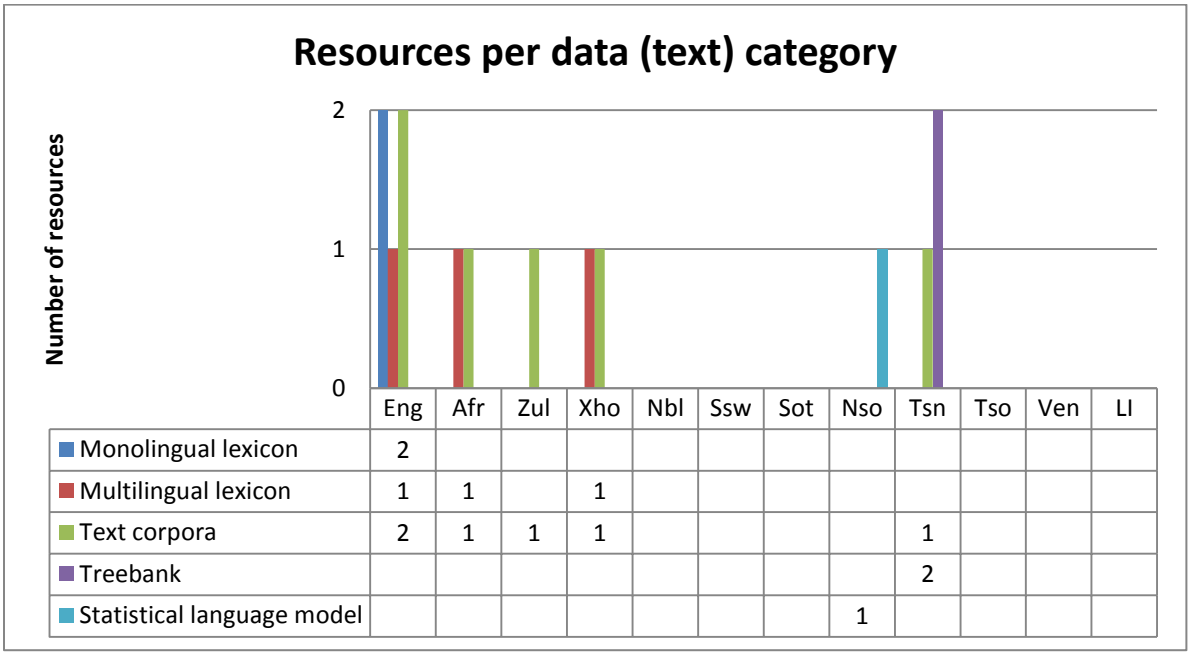
TABLE 3: SUMMARY OF RESOURCES PER LANGUAGE AND CATEGORY

Language	Language code	Language categories		
		Data	Models	Software
English	Eng	9	0	3
Afrikaans	Afr	7	0	5
isiZulu	Zul	2	0	6
isiXhosa	Xho	4	0	4
isiNdebele	Nbl	1	0	4
siSwati	Ssw	1	0	4
Sesotho	Sot	2	0	4
Sepedi	Nso	2	0	6
Setswana	Tsn	5	0	6
Xitsonga	Tso	1	0	5
Tshivenda	Ven	1	0	4

8.4.5 RESOURCES PER CATEGORY, LANGUAGE AND RESOURCE TYPE - DETAILED ANALYSIS

8.4.5.1 Data (Text)

Resources were submitted in six of the 14 resource types in the Data (Text) category, and mainly for the "better resourced" languages[5] (English, Afrikaans, isiZulu, isiXhosa, Sepedi and Setswana).



	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Monolingual lexicon	2											
Multilingual lexicon	1	1		1								
Text corpora	2	1	1	1					1			
Treebank									2			
Statistical language model								1				

FIGURE 22: REPRESENTATION OF RESOURCES PER DATA (TEXT) CATEGORY

8.4.5.2 Data (Speech)

Speech data resources were only submitted in two of the 15 resource types in the Data (speech) category. One of these resources was for all South African official languages, while the other focused on the better resources languages.

[5] By "better resourced" languages, we mean those languages for which more resources have been developed over time, typically the well-resourced English, Afrikaans, and the larger of the African languages, namely isiZulu, isiXhosa, Sepedi and Setswana.

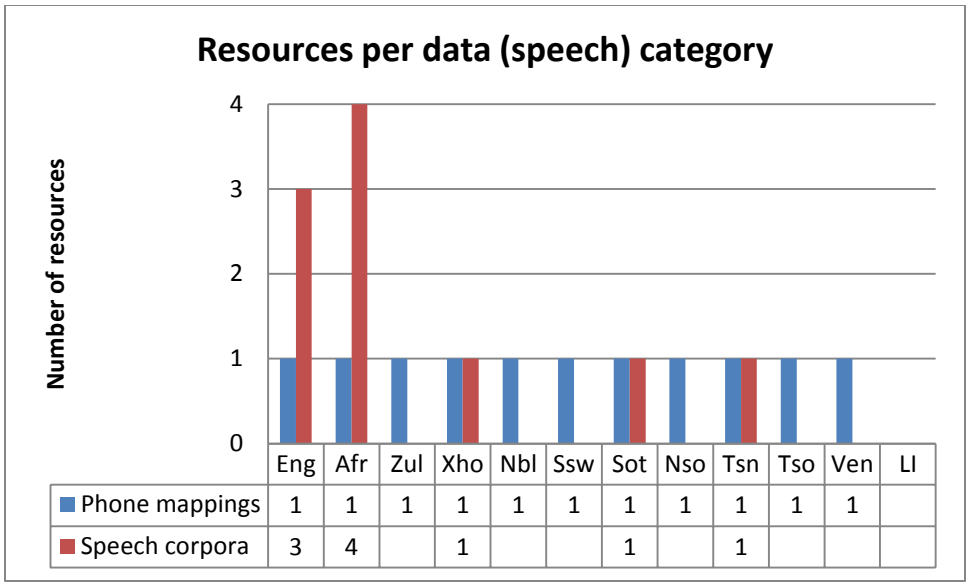


FIGURE 23: REPRESENTATION OF RESOURCES PER DATA (SPEECH) CATEGORY

8.4.5.3 Software (Text)

Resources were only submitted in six of the 23 resource types in the Software (Text) category. Resources submitted were fairly evenly distributed over the 11 South African official languages, with more focus on the indigenous languages than on English.

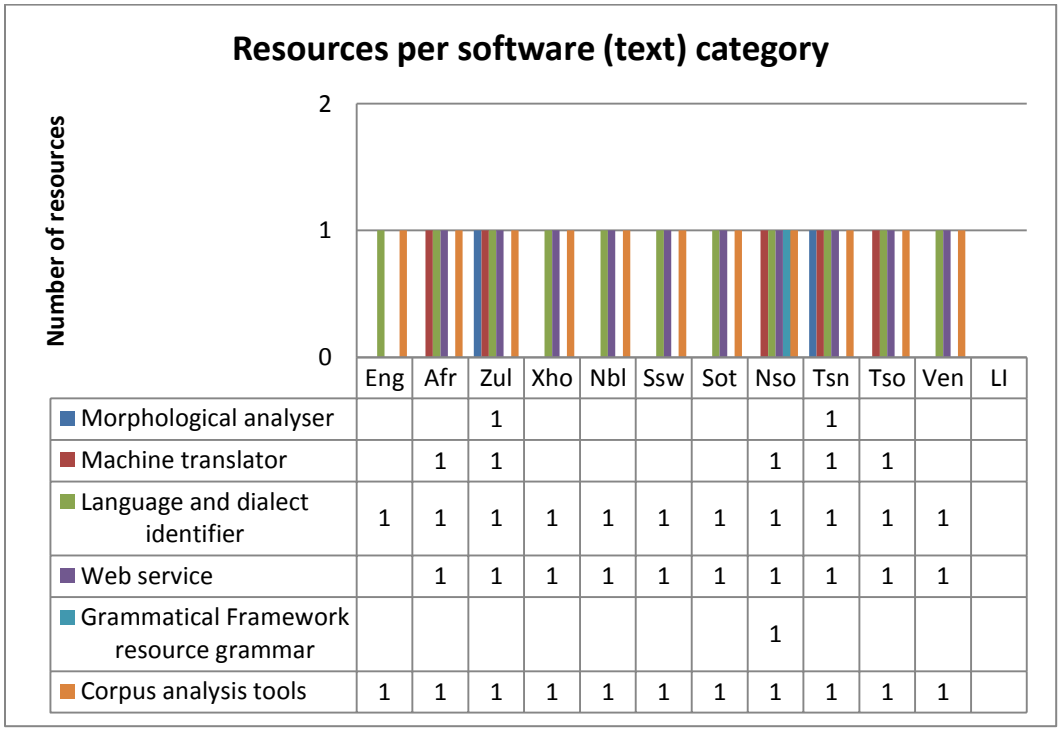


FIGURE 24: REPRESENTATION OF RESOURCES PER SOFTWARE (TEXT) CATEGORY

8.4.5.4 Software (Speech)

Resources were only submitted in one of the 36 types in the Software (Speech) category. This resource type submitted was language independent.

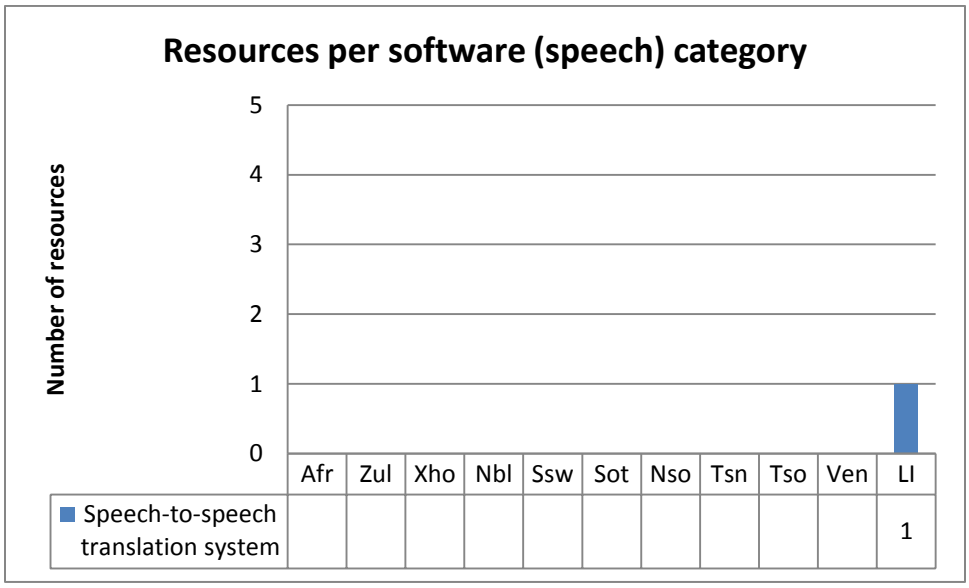


FIGURE 25: REPRESENTATION OF RESOURCES PER SOFTWARE (SPEECH) CATEGORY

8.4.6 MATURITY OF RESOURCES

The maturity of a resource is defined by its stage of development. As in the 2009 Audit, the maturity was measured in terms of the following:

- Under development (classified with a ◑ (half circle)/H in the 2009 Audit report)
- Alpha version (classified with a ◑ (half circle)/H)
- Beta version (classified with a ◑ (half circle)/H)
- Released (classified with a ● (full circle)/F).

Participants in the 2018 Audit were requested to indicate the maturity of the resources they submitted. An overview of the maturity of the 2018 resources submitted is provided in Table 4 below.

TABLE 4: SUMMARY OF MATURITY OF RESOURCES

Maturity level	Category	Resource type	Language/s	Number
Under development ◑	No resources submitted that are under development			0
Alpha version ◑	Data (Text)	Multilingual text	Eng, Zul, Afri, Tsn	5
		Treebanks	Tsn	2
Beta version ◑	Data (Speech)	Speech corpora	Eng, Afr	2
		Phone mappings	Eng, Afr, Zul, Xho, Ven, Nbl, Ssw, Sot, Nso, Tsn, Tso	11
Released ●	Data (Text)	Monolingual lexicon	Eng	1
		Statistical language model	Sot	1
		Monolingual text	Xho	1
	Data (Speech)	Speech corpora	Xho, Afr, Tsn, Sot, Eng	8

8.4.7 AVAILABILITY OF RESOURCES

The availability of a resource is defined as the permitted use of the resource by researchers, academics or companies who access the resource on the SADiLaR database. As in the 2009 Audit, the availability was measured in terms of the following:

- For research use only (classified with a ◑ (half circle)/H in the 2009 Audit report)
- For commercial use (classified with a ●(full circle)/F)
- The resource is open and freely available for use (classified with a ●(full circle)/F)
- The availability of the resource is undecided (classified with a ◑ (half circle)/H)
- The resource is not available, proprietary or closed for use (classified with a ◑ (half circle)/H).

Participants in the 2018 Audit were requested to indicate the availability of the resources they submitted. An overview of the availability of the 2018 resources submitted is provided in Table 5 below.

TABLE 5: SUMMARY OF AVAILABILITY OF RESOURCES

Availability	Category	Resource type	Language/s	Number
Commercial ●	Software (Speech)	Speech-to-speech translation system	All	11
	Software (Text)	Language and dialect ID	All	11
Not available/ proprietary/ closed ◑	Software (Text)	Morphological analyser	Zul, Tsn	2
Open/ freely available ●	Data (Text)	Monolingual lexicon	Eng	1
		Multilingual lexicon	Eng, Afr, Xho	3
		Treebank	Tsn	1
	Data (Speech)	Speech corpora	Afr, Xho, Sot, Tsn	4
	Software (Text)	Other (web service)	Afr, Xho, Zul, Ven, Nbl, Ssw, Sot, Nso, Tsn, Tso	10
		Other (corpus analysis tools)	All	11
		Machine translation	Afr, Zul, Nso, Tsn, Tso	5
		Other (grammatical framework resource grammar)	Tsn	1
Research ◑	Data (Text)	Monolingual lexicon	Eng	1
		Statistical language model	Sot	1
	Data (Speech)	Phone mappings	All	11
		Speech corpora	Afr, Eng	4
Undecided ◑	Data (Speech)	Speech corpora	Eng, Afr	2
	Data (Text)	Text corpora (Multilingual)	Eng, Zul, Afr, Tsn	5
		Monolingual text (Monolingual)	Xho	1
		Treebank	Tsn	1

8.4.8 SUMMARY OF THE 2018 AUDIT RESULTS

From the above analysis, we deduce that the resources available for **English, Afrikaans, isiZulu, isiXhosa, Sepedi and Setswana** are the **most mature** and that these languages are the **most resourced languages** (languages for which resources are most available). The **software (text) category received the most submissions**, which could also be attributed to the nature (more text focused), of the submitting institutions.

8.4.9 COMPARISON OF THE 2009, 2014 AND 2018 DATASETS

8.4.9.1 Process

In order to compare the three datasets, we matched (with the assistance of an expert in the field) the resource types across all three datasets to one another. As the matching could not be done on a per resource category basis, due to the changes made to the 2018 Audit design, the resource types were matched. Only matched resource types, for which there was at least one entry in one dataset, are included in the analysis and representation in this section of the report[6]. This resulted in three sets of comparisons:

- A comparison of resource types which matched across **all datasets**
- A comparison of resource types which matched across **two datasets** (2009 and 2014), but did not match those in the 2018 data
- A representation of the resource types from the 2018 data that could not be matched to data in the 2009 and 2014 data.

We then indicated the maturity and accessibility of each resource type across all the data, as follows:

[6] Matched resource categories with no resource entries, will be indicated in the Data analysis under section 5 of the report.

TABLE 6: OVERVIEW OF REPRESENTATION OF MATURITY AND ACCESSIBILITY

Maturity		Availability	
Level	**Representation**	**Level**	**Representation**
Under development	◑	Not available/ proprietary/ closed	◑
Alpha version	◑	Undecided	◑
Beta version	◑	Research	◑
Released	●	Commercial	●
		Open/ freely available	●

Based on a combination of its maturity and availability, each resource was then categorised as either a full resource (F), or a partial resource (P)[7]:

- A full resource is a resource which is fully mature (released) and is fully available (commercially, openly/freely), i.e. has full circles in terms of maturity and availability.
- A partial resource is a resource which is only partially mature (under development, alpha version, beta version) and only partially available (for research purposes, undecided, not available/proprietary/closed), i.e. features half circles in terms of maturity and availability.

The spreadsheets per dataset were populated with these measures.

8.4.10 MATCHED RESOURCE TYPES ACROSS ALL THREE DATASETS

After matching the resource types, we determined that the following were comparable across all **three datasets** (2009, 2014 and 2018).

8.4.10.1 Data (Speech and Text)

TABLE 7: SUMMARY OF MATCHED RESOURCE TYPES ACROSS THREE DATASETS: DATA

Resource type	Full	Partial
Text corpora	X	X
Monolingual lexicons	X	X
Multilingual lexicons	X	X

[7] Full resources are available for download via the SADiLaR Catalogue, while partial resources are visible in the SADiLaR Index and not yet available for download.

Resource type	Full	Partial
Wordnets	X	X
Treebanks	X	X
Speech corpora	X	X
Pronunciation dictionaries	X	X
Terminology lists		X
Intonation models		X

8.4.10.2 Software (Modules, tools and applications)

TABLE 8: SUMMARY OF MATCHED RESOURCE TYPES ACROSS THREE DATASETS: SOFTWARE

Resource type	Full	Partial
Lemmatisers	X	X
Morphological analysers	X	X
POS Taggers / disambiguators		X
Speech-based tools	X	X
Speech recognition systems	X	X
Machine translators	X	
Comprehension assistants		X
Language and dialect identifiers	X	
Machine-aided human translations	X	
Human-aided machine translations		X
Format normalisers		X
Corpus analysis tools	X	
Acoustic analysis tools	X	
Chunkers	X	X
Automatic phonetic transcriptions		X
Tokenisers	X	X
Limited domain TTS (Complete TTS)		X
Domain independent TTS (Complete TTS)		X
Hyphenators		X
Proofing/authoring tools		X
Speech-to-speech translation systems	X	X
Named entity recognisers	X	
OCR/ICRs	X	
Integrated automatic annotations	X	

8.4.11 MATCHED RESOURCE TYPES ACROSS TWO DATASETS

After matching the resource types, we determined that the following were comparable across all **two datasets** (2009 and 2014).

8.4.11.1 Data (Speech and Text)

TABLE 9: SUMMARY OF MATCHED RESOURCE TYPES ACROSS TWO DATASETS: DATA

Resource type	Full	Partial
Lexical databases		X
Other text resources		X
Test suites and test corpora		X
Multimedia corpora (Multimodal corpora)		X

8.4.11.2 Software (Modules, tools and applications)

TABLE 10: SUMMARY OF MATCHED RESOURCE TYPES ACROSS TWO DATASETS: SOFTWARE

Resource type	Full	Partial
G2P convertors		X
Compound analysers	X	X
Computer-assisted (aided) Language Learning		X
Audio search		X
Access control		X
Speaking devices		X
Telephony applications	X	X
Text selection tools		X
Parameter search		X
Web crawlers		X
Accessibility		X
Multimodal information access		
Annotations	X	X
PDF converters	X	
Anonymisers	X	
Terminology integration texts	X	
Text aligners	X	

8.4.12 UNMATCHED RESOURCE TYPES

We could **not match these resource types from 2018** to the 2009 and 2014 resource types.

8.4.12.1 Data (Speech and Text)

TABLE 11: SUMMARY OF UNMATCHED RESOURCE TYPES FROM 2018 DATA

Resource type	Full	Partial
Phone mappings		X
Statistical language models		X

8.4.12.2 Software (Speech and Text)

TABLE 12: SUMMARY OF UNMATCHED RESOURCE TYPES FROM 2018 DATA

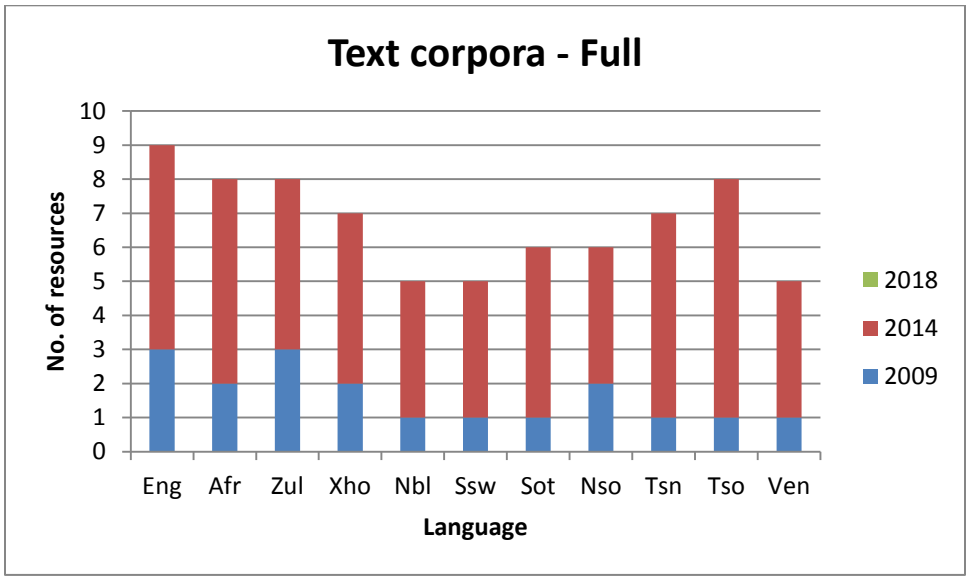
Resource type	Full	Partial
Web services	X	
Grammatical framework resource grammars	X	

8.4.13 Results of the Data Category comparison

Comparison over THREE datasets

8.4.13.1 Text corpora

The text corpora resource type indicated below includes both annotated and unannotated monolingual text corpora as well as aligned and unaligned multilingual text corpora. The comparison indicates that there was an increase in full text corpora resources over the languages from 2009 to 2014; however, no additional resources were submitted in 2018. In 2009, no partial text corpora resources were available, with a few added in 2014 and resources for isiXhosa added in 2018.



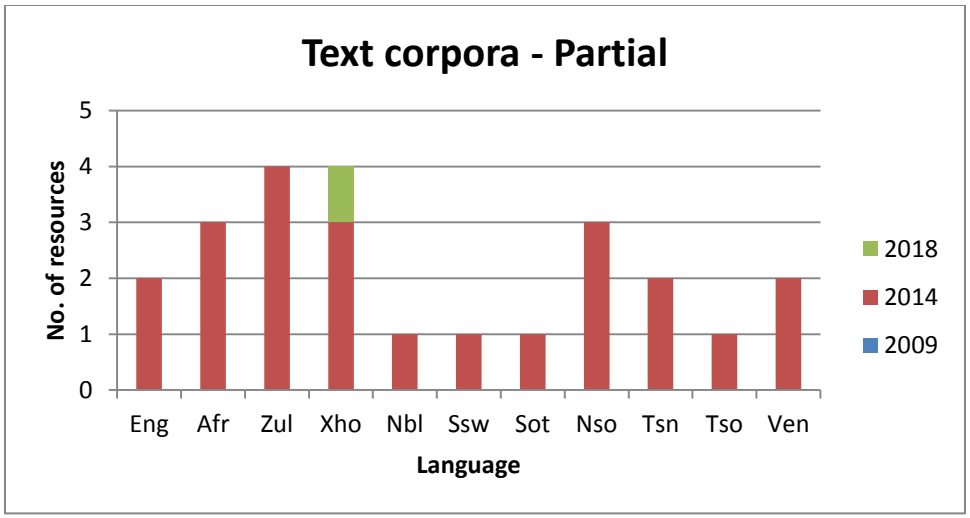
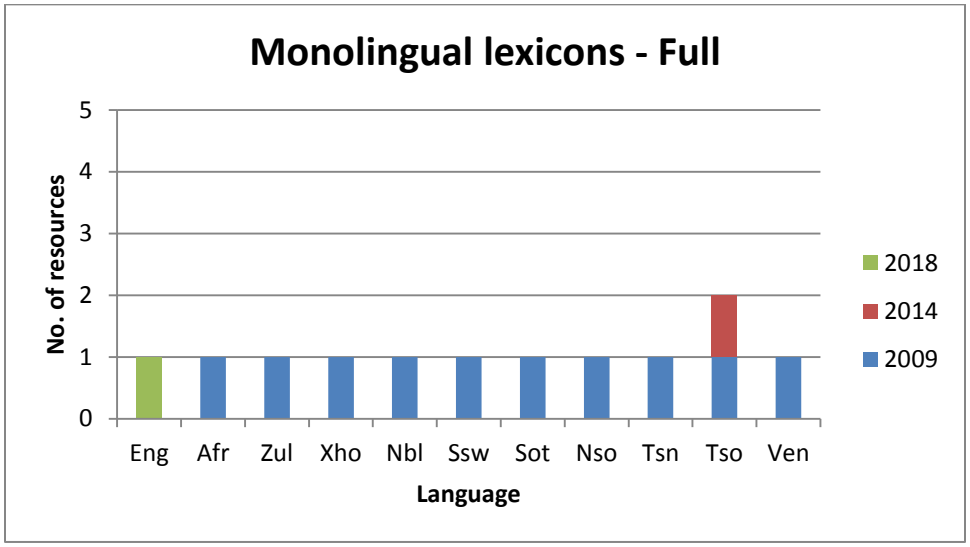


FIGURE 26: REPRESENTATION OF TEXT CORPORA

8.4.13.2 Monolingual lexicons

One monolingual lexicon resource per language (except English) has been available since 2009, with two additions (one for Xitsonga and one for English) added in 2014 and 2018 respectively. No partial monolingual lexicon resources were available prior to the one added in 2018.



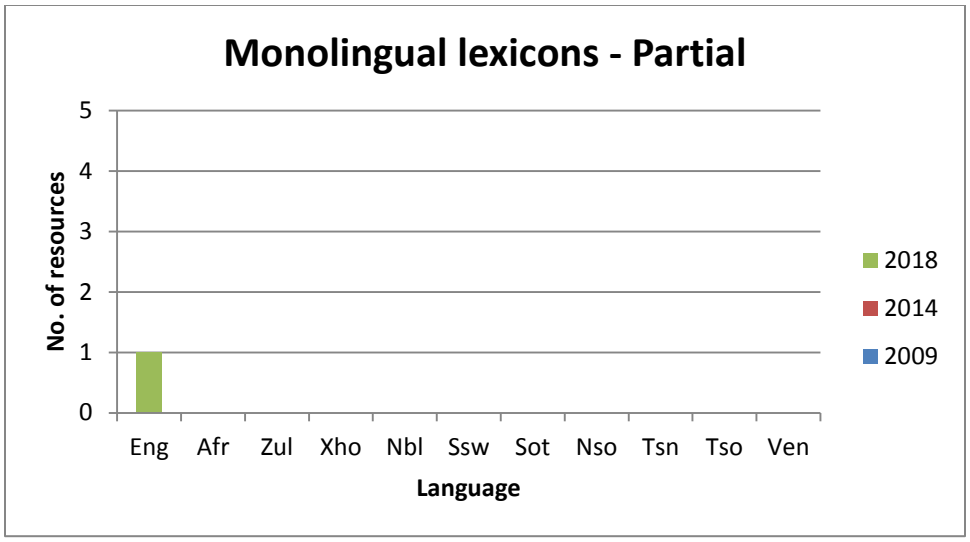
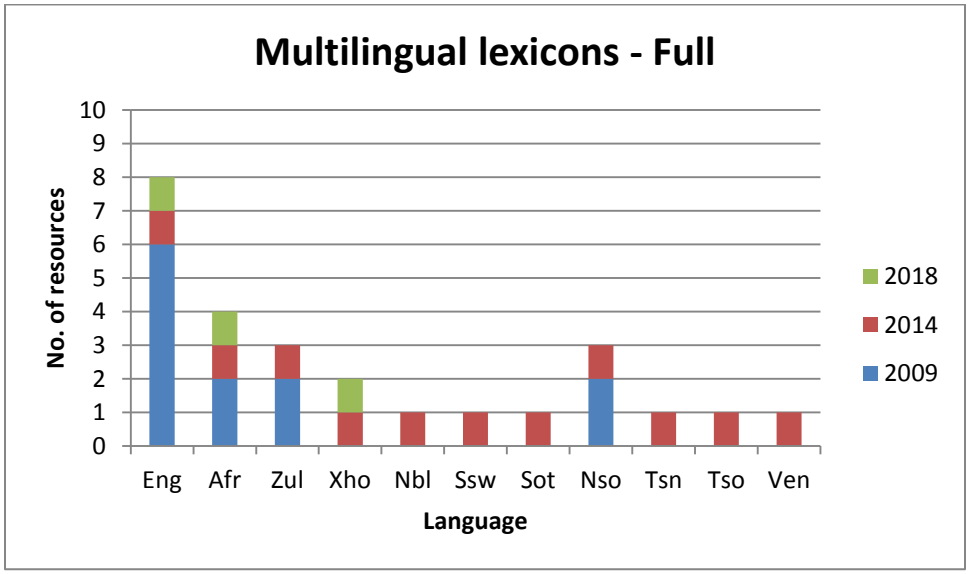


FIGURE 27: REPRESENTATION OF MONOLINGUAL LEXICONS

8.4.13.3 Multilingual lexicons

Multilingual lexicons, while available in limited number for all languages, are more prevalent for English, Afrikaans, isiZulu and Sepedi.



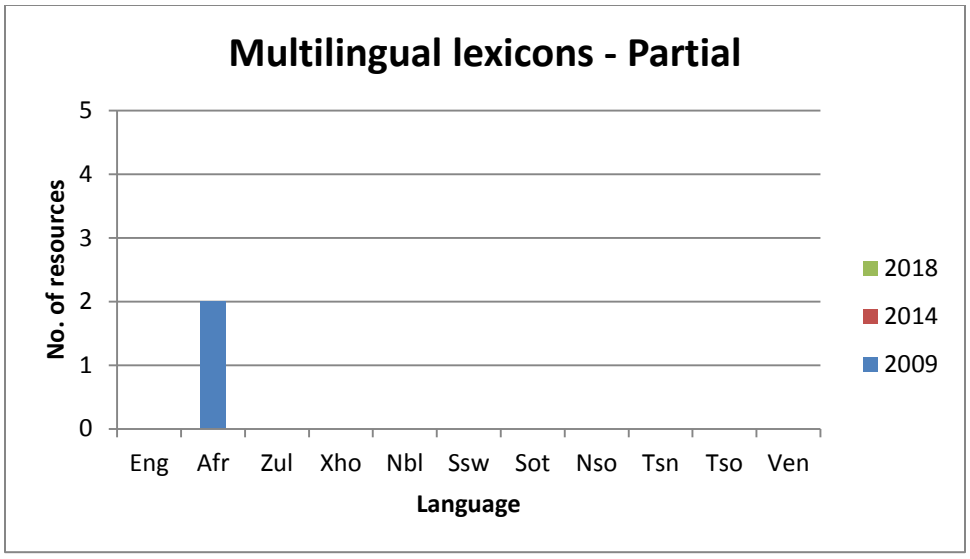
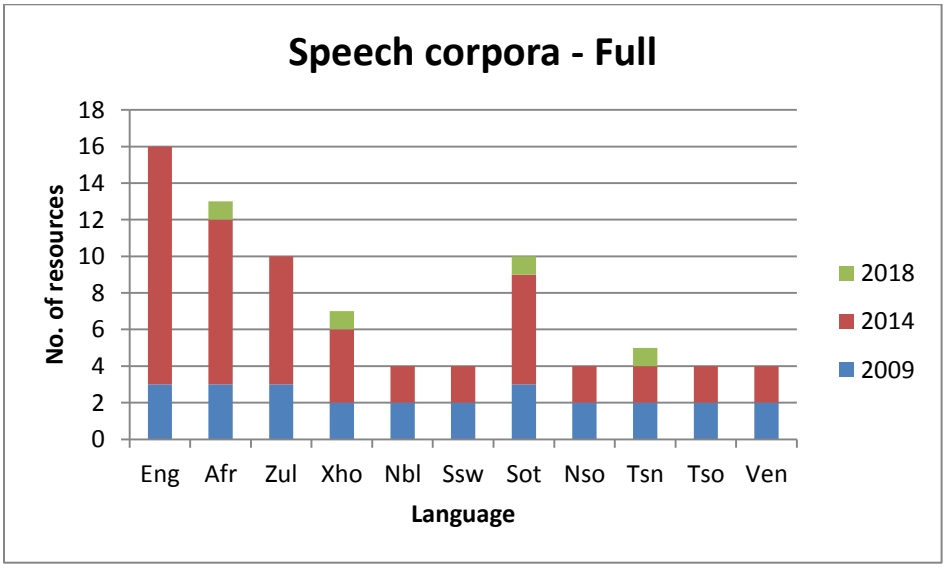


FIGURE 28: REPRESENTATION OF MULTILINGUAL LEXICONS

8.4.13.4 Speech corpora

The speech corpora resource type indicated below includes both annotated and unannotated monolingual speech corpora as well as aligned and unaligned multilingual speech corpora. Full speech corpora exist for all languages, with English, Afrikaans, isiZulu and Sesotho featuring more such resources than the other languages.



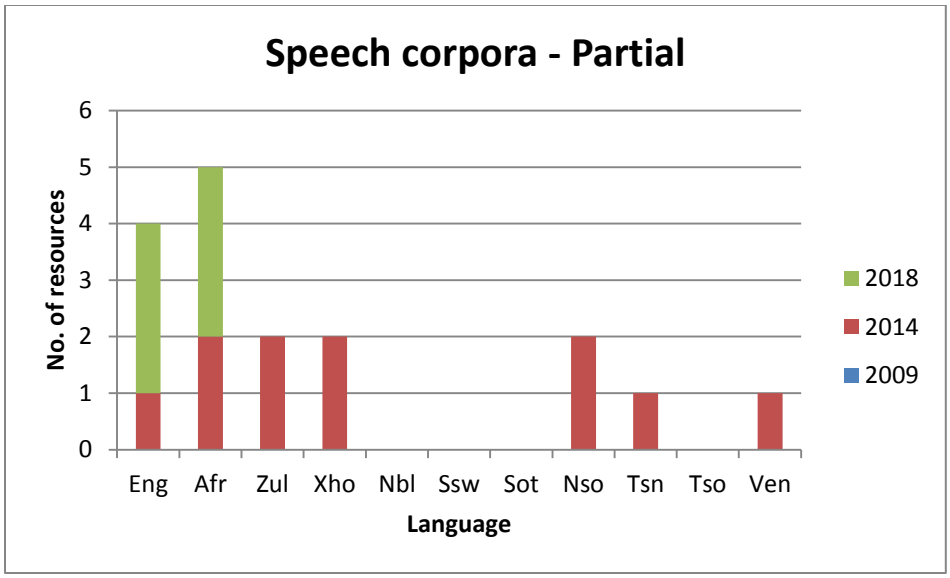
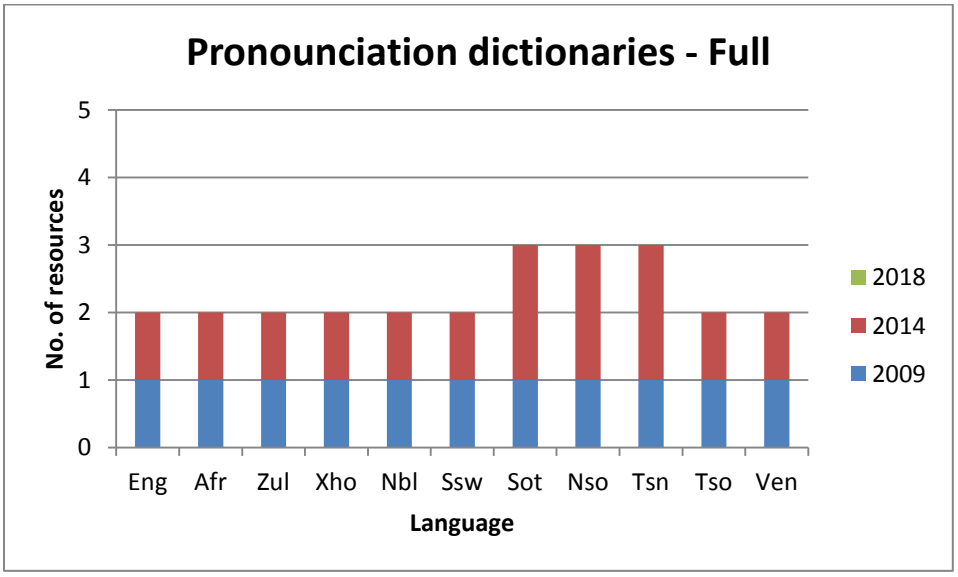


FIGURE 29: REPRESENTATION OF SPEECH CORPORA

8.4.13.5 Pronunciation dictionaries

Pronunciation dictionary resources were made available from 2009 to 2014. In 2009 and in 2014, full pronunciation dictionaries were available for all 11 South African official languages. In 2014, additional full dictionaries for Sesotho, Sepedi and Setswana were released. In addition, there were partial resources for English, Afrikaans, Sepedi and Tshivenda in the 2014 data. No pronunciation dictionary resources were added in 2018.



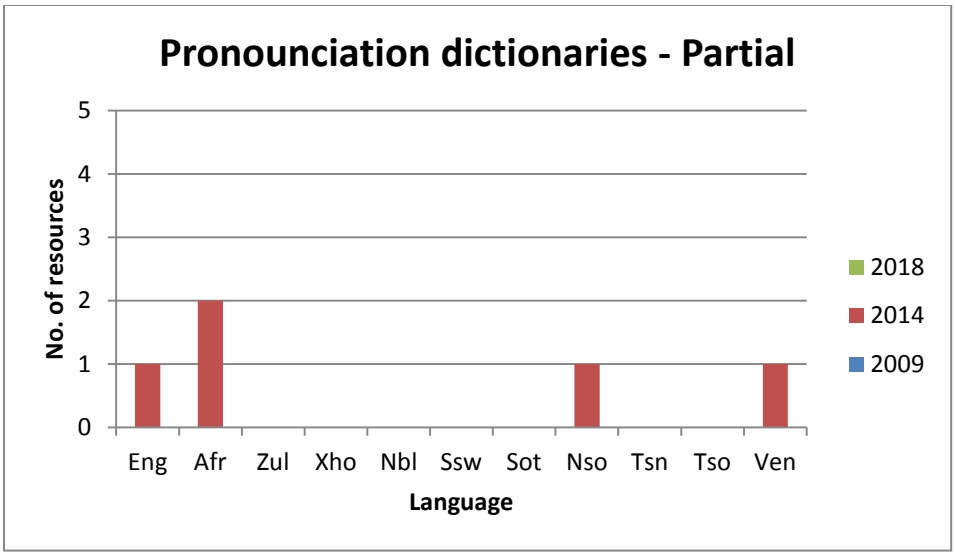
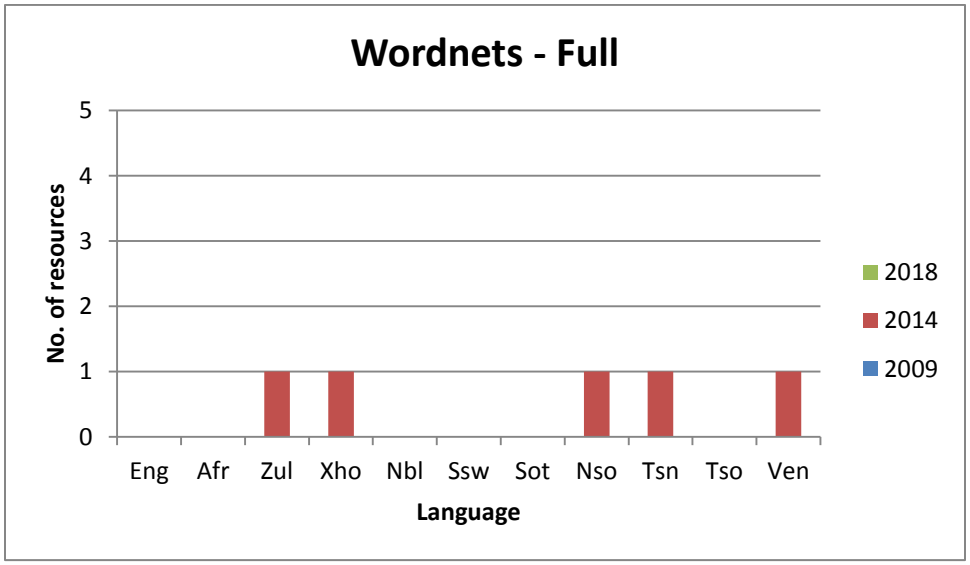


FIGURE 28: REPRESENTATION OF PRONUNCIATION DICTIONARIES

8.4.13.6 Wordnets

Full and partial wordnet resources were only released since 2014 for isiZulu, isiXhosa, Sepedi, Setswana, Afrikaans and Tshivenda. No full wordnet resources were added in 2018.



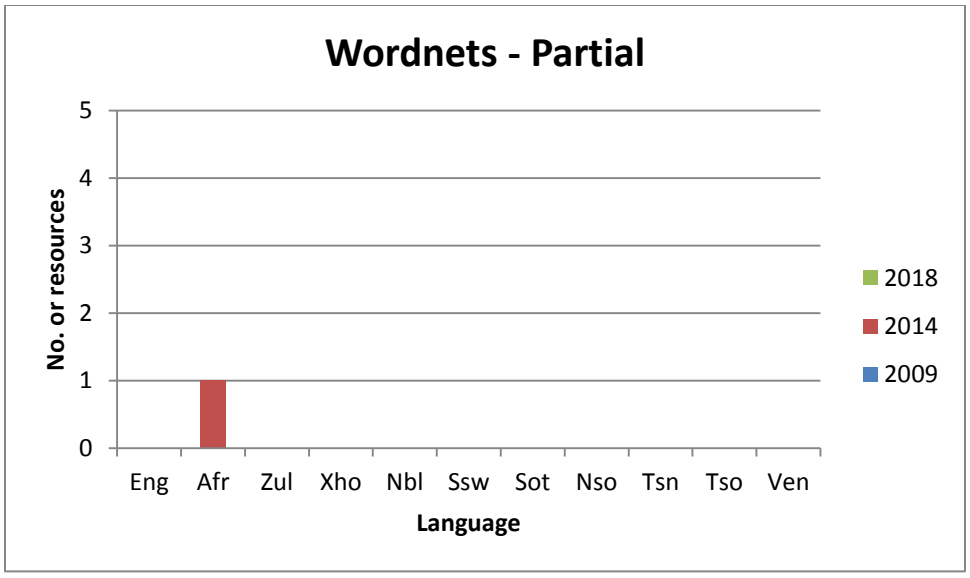


FIGURE 29: REPRESENTATION OF WORDNETS

8.4.13.7 Terminology lists

Partial terminology lists have been available since 2014 for all 11 South African official languages. No full resources are available.

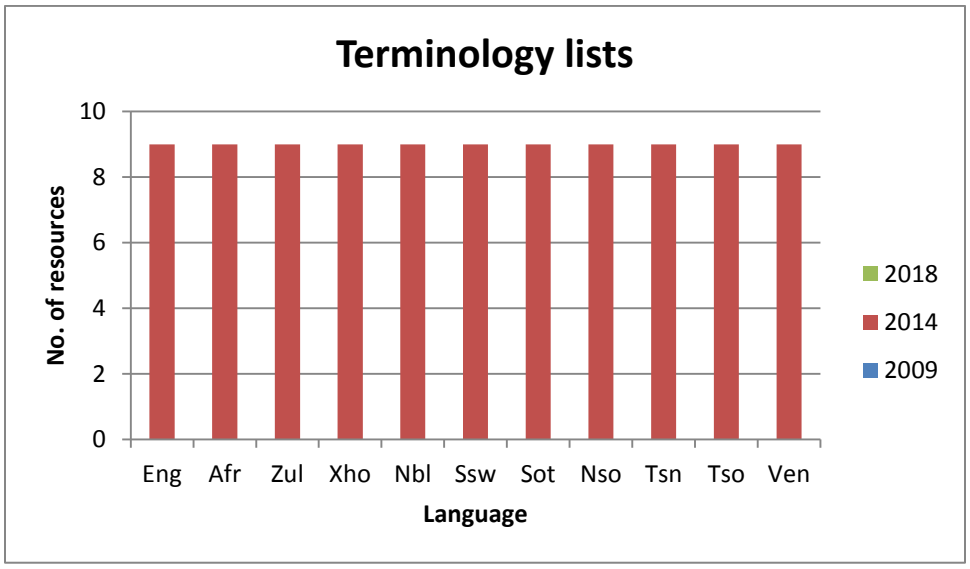
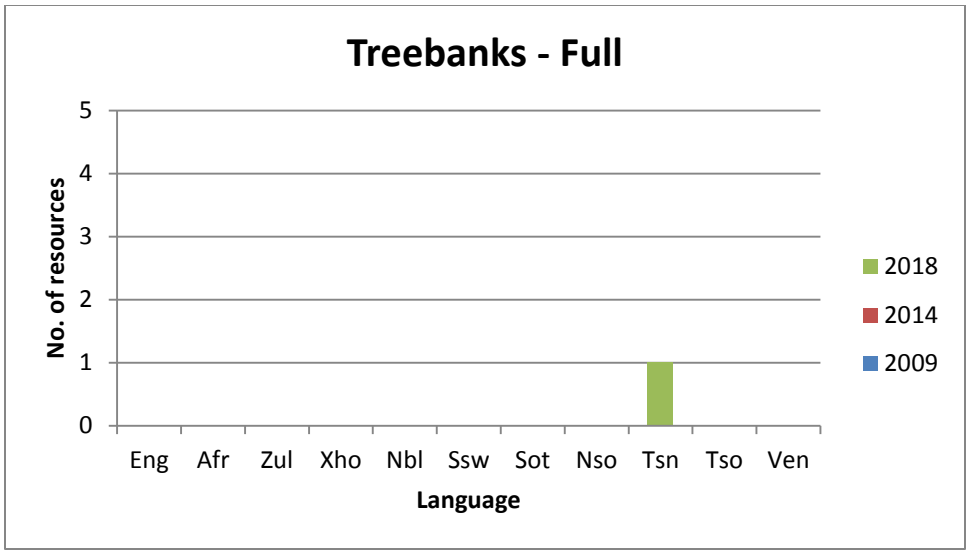


FIGURE 30: REPRESENTATION OF TERMINOLOGY LISTS

8.4.13.8 Treebanks

A full treebank resource was added in 2018 for Setswana only and a partial resource was added for Sepedi.



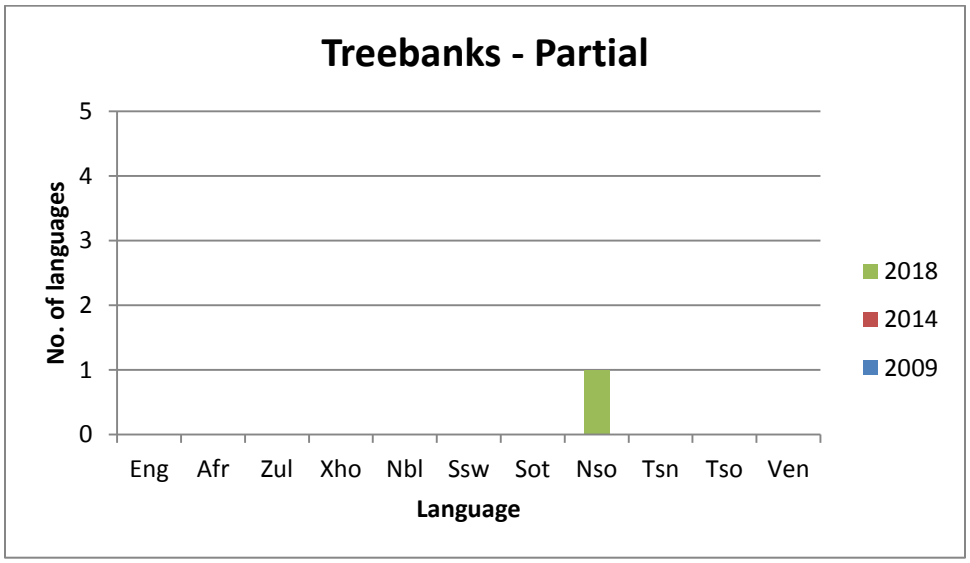


FIGURE 31: REPRESENTATION OF TREEBANKS

8.4.13.9 Intonation models

Partial intonation models have been available since 2014 for isiZulu, Sesotho, Sepedi and Setswana. No full resources are available.

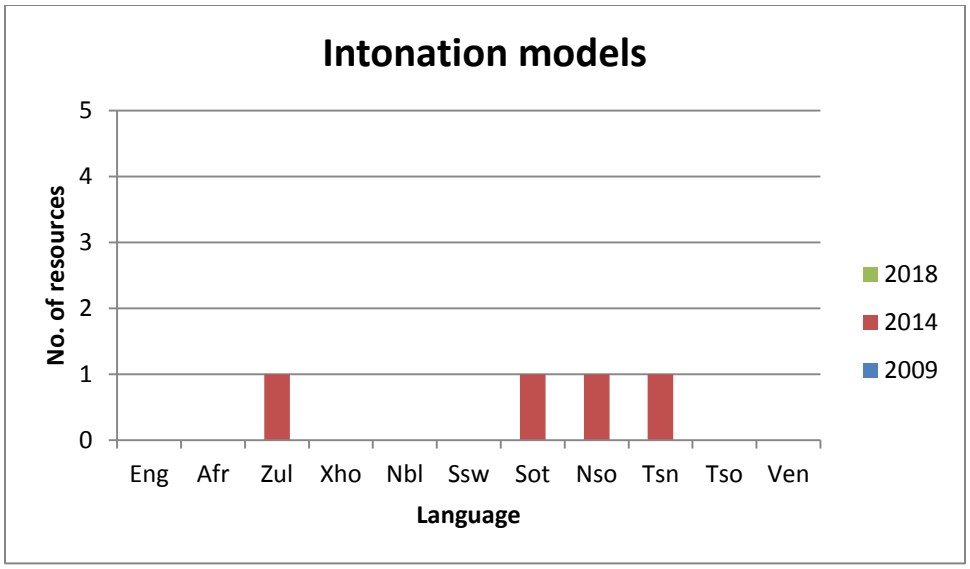


FIGURE 32: REPRESENTATION OF INTONATION MODELS

COMPARISON OVER TWO DATASETS

8.4.13.10 Lexical databases

Partial lexical databases have been available since 2014 for English and Afrikaans. No full resources are available.

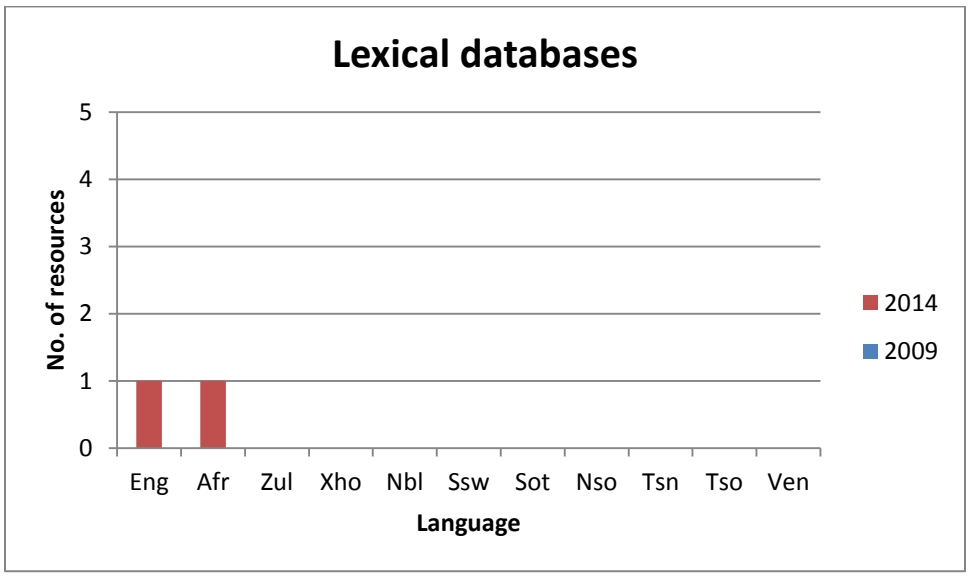


FIGURE 33: REPRESENTATION OF LEXICAL DATABASES

8.4.13.11 Other text resources

Partial “other text resources” which have not been classified have been available since 2009 for all 11 South African official languages and in 2014 an addition was made for Afrikaans only. No full resources are available.

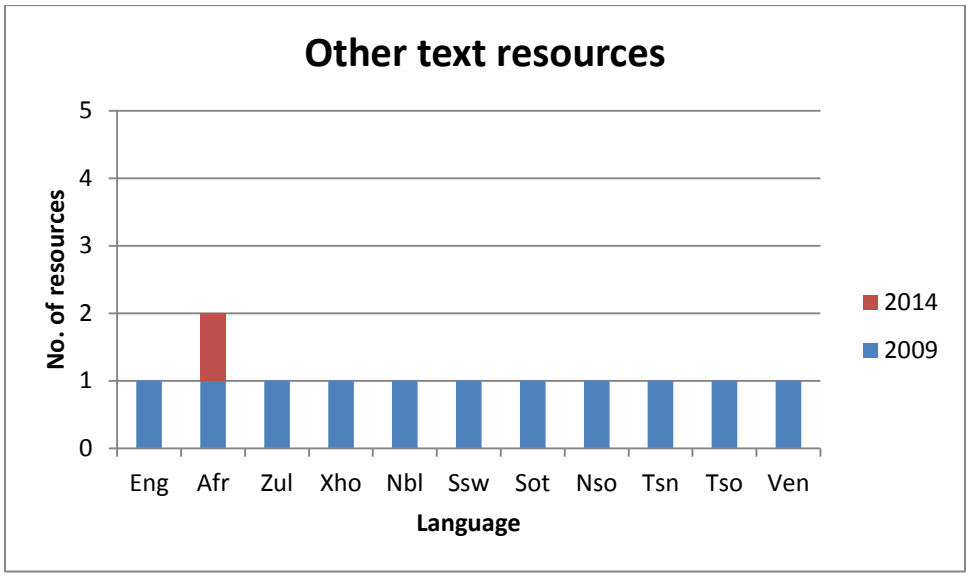


FIGURE 34: REPRESENTATION OF "OTHER TEXT RESOURCES"

8.4.13.12 Test suites and test corpora

Only partial test suites and test corpora are available, and these are only for Sepedi.

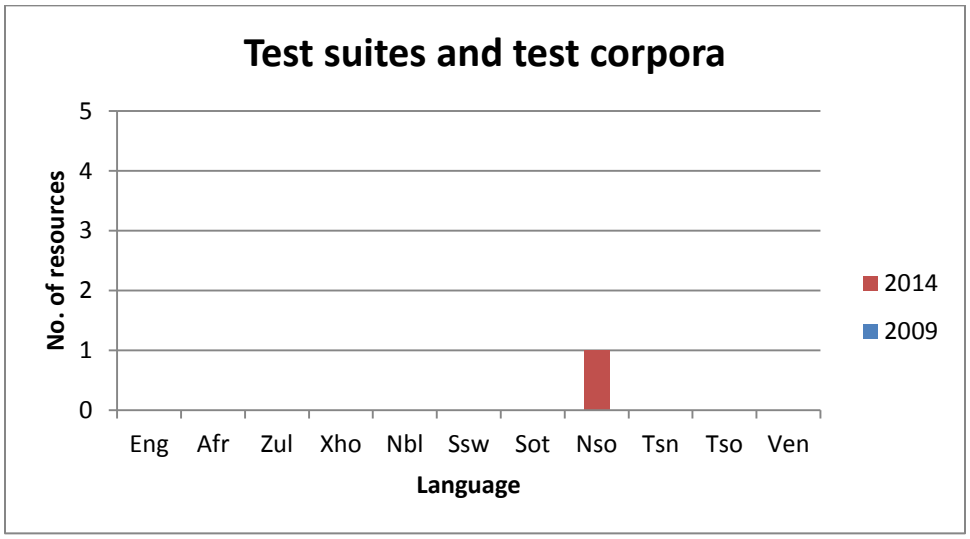


FIGURE 35: REPRESENTATION OF TEST SUITES AND TEST CORPORA

8.4.13.13 Multimedia corpora

Partial multimedia corpora were made available in 2014 for isiZulu and isiXhosa only.

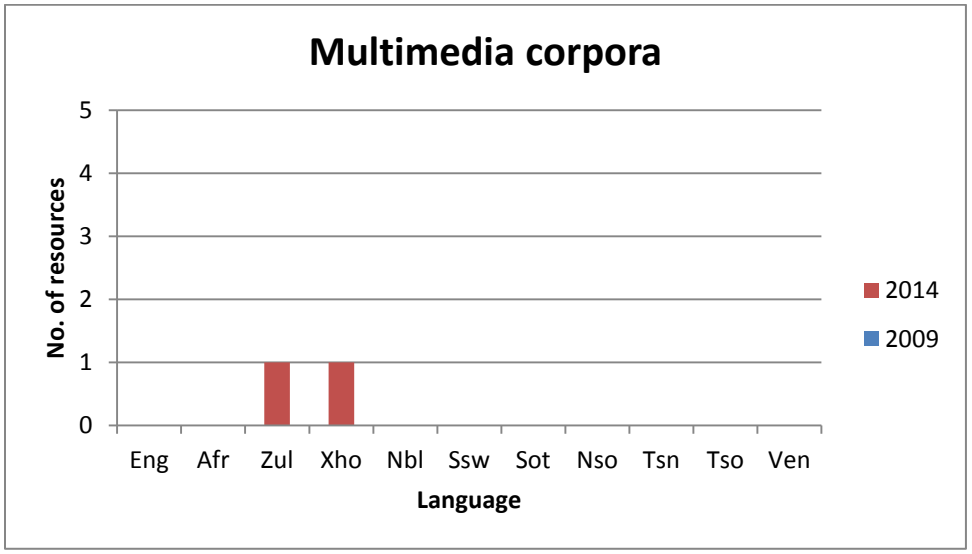


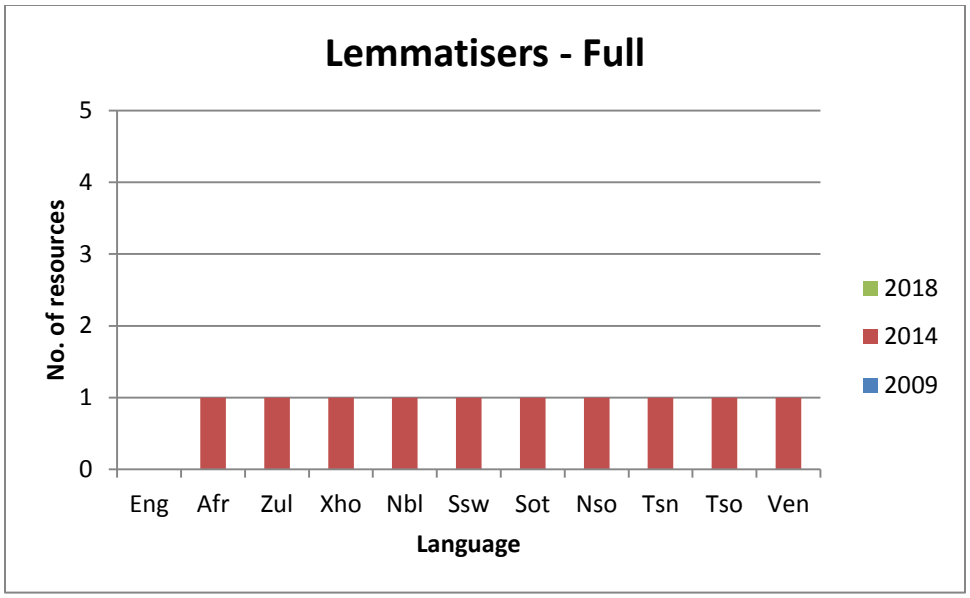
FIGURE 36: REPRESENTATION OF MULTIMEDIA CORPORA

8.4.14 RESULTS OF THE SOFTWARE (MODULES/TOOLS/APPLICATIONS) CATEGORY COMPARISON

COMPARISON OVER THREE DATASETS

8.4.14.1 Lemmatisers

One full lemmatiser resource is available for 10 of the 11 South African official languages. An additional partial resource is available for Afrikaans.



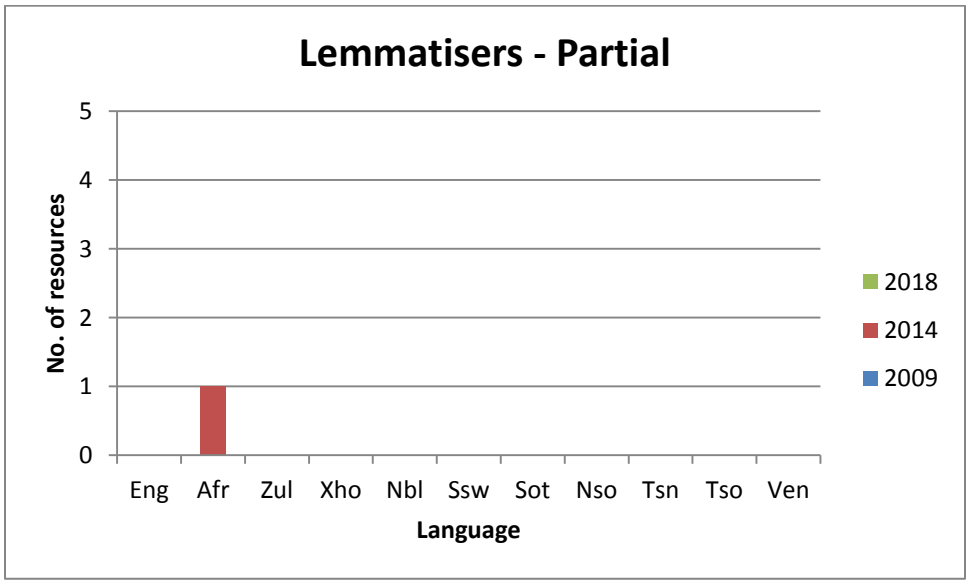


FIGURE 37: REPRESENTATION OF LEMMATISERS

8.4.14.2 Morphological analysers

Full morphological analysers are available for 10 of the 11 South African official languages. Partial morphological analysers were added for Afrikaans and isiZulu in 2014 and most recently, one partial resource each for isiZulu and Setswana have also been added.

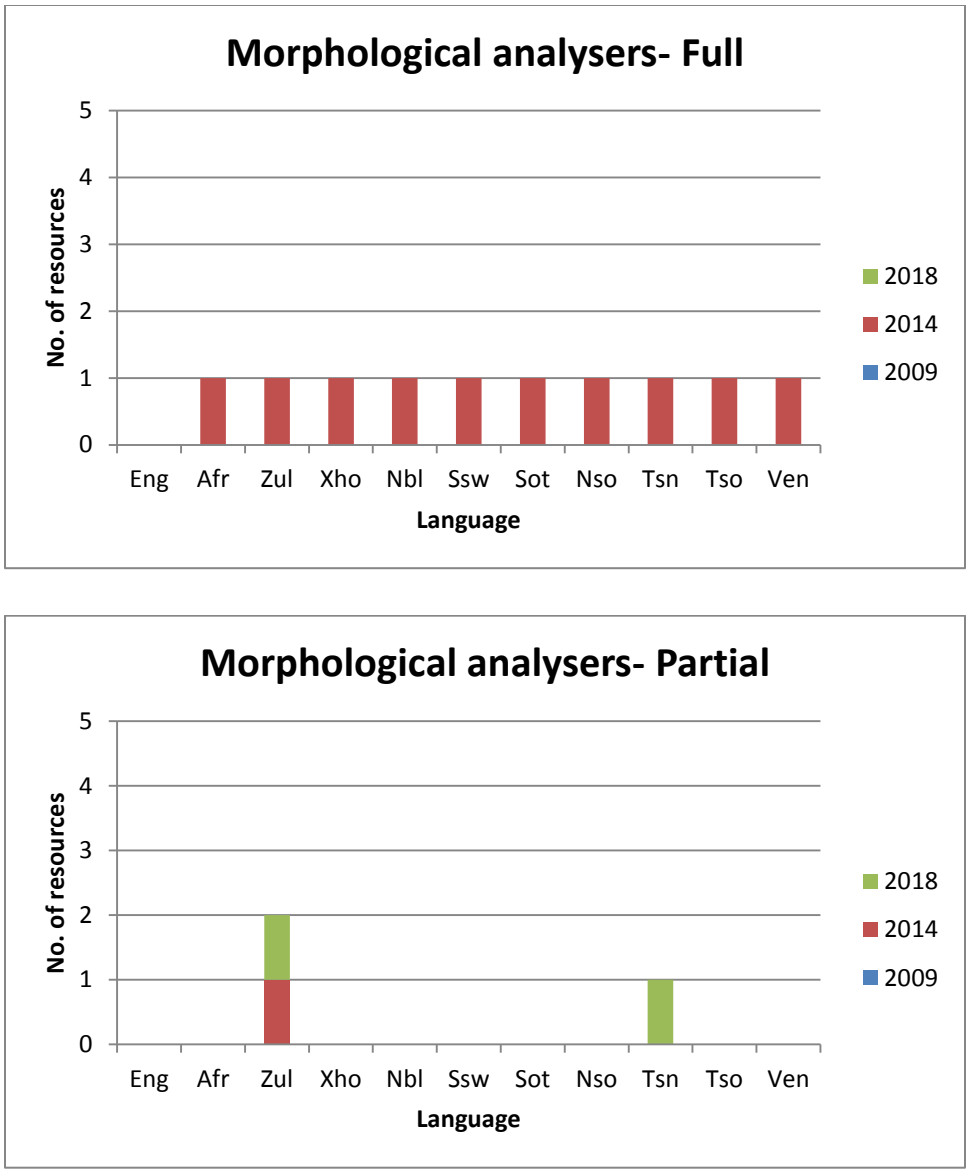


FIGURE39: REPRESENTATION OF MORPHOLOGICAL ANALYSERS

8.4.14.3 POS taggers/disambiguators

Partial POS taggers/disambiguators are available Afrikaans, isiXhosa and Sepedi. 1 POS tagger existed in 2009 for Afrikaans and the POS taggers for isiXhosa and Sepedi were added in 2014.

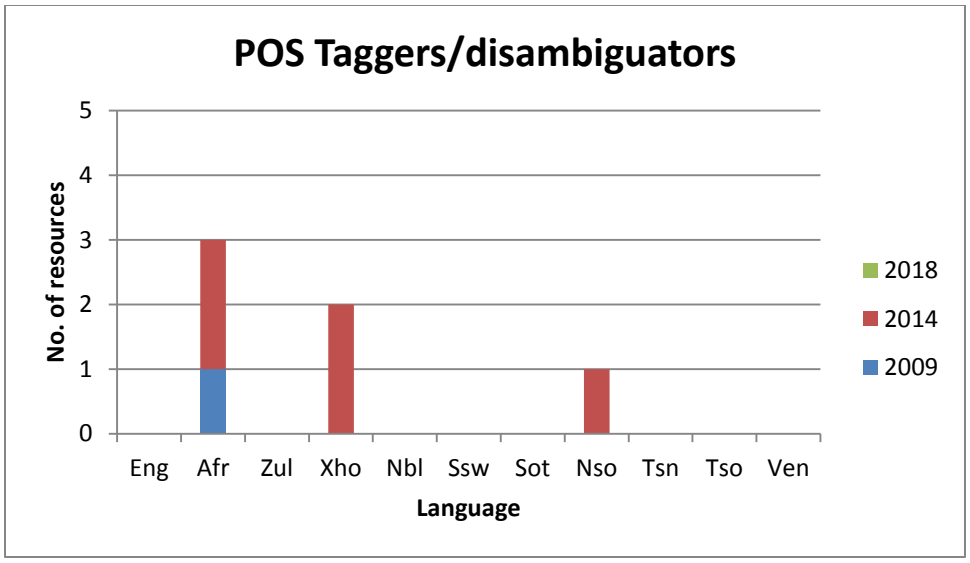
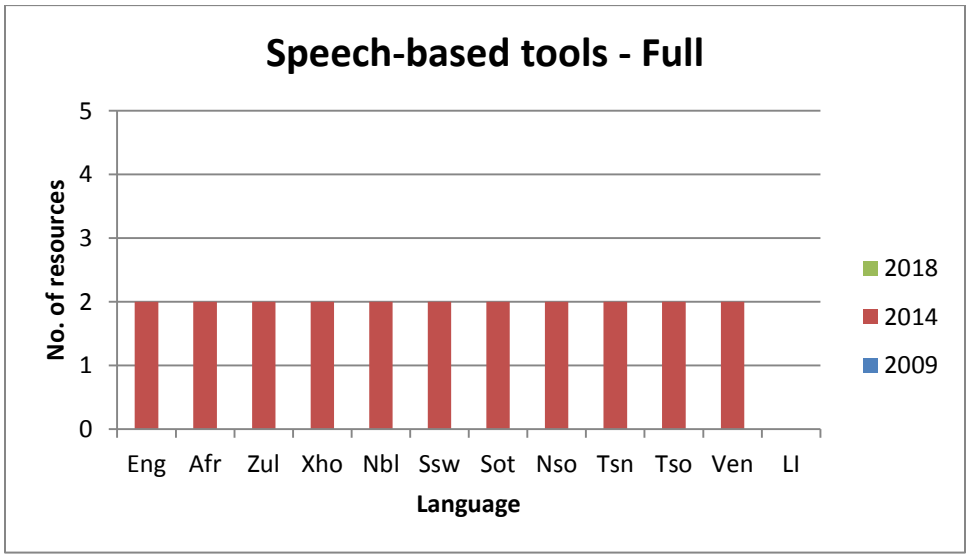


FIGURE 38: REPRESENTATION OF POS TAGGERS/DISAMBIGUATORS

8.4.14.4 Speech-based tools

Two full and one partial speech-based tool for all of the South African official languages have been available since 2014. One partial, language independent, speech-based tool has been available since 2014. No speech-based tools were submitted in 2018.



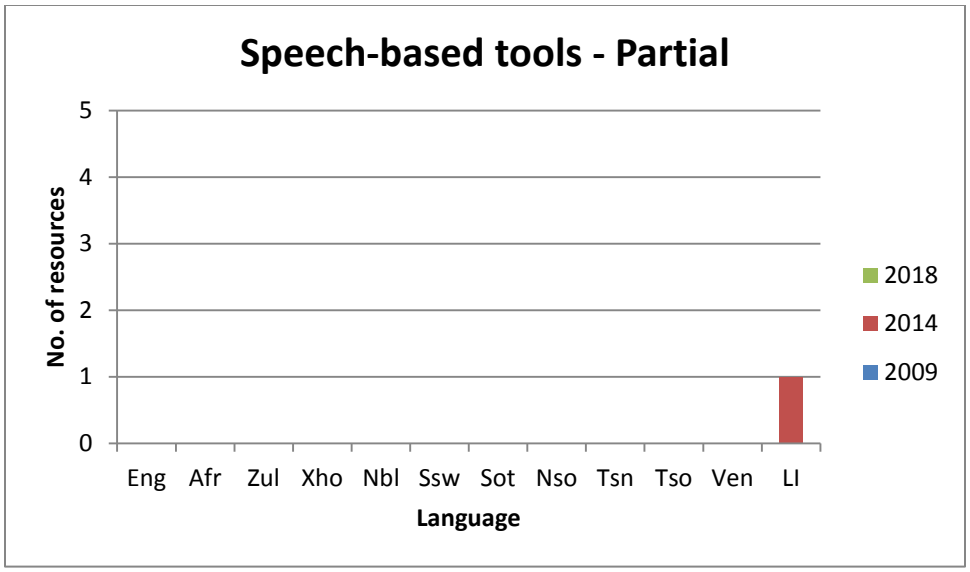
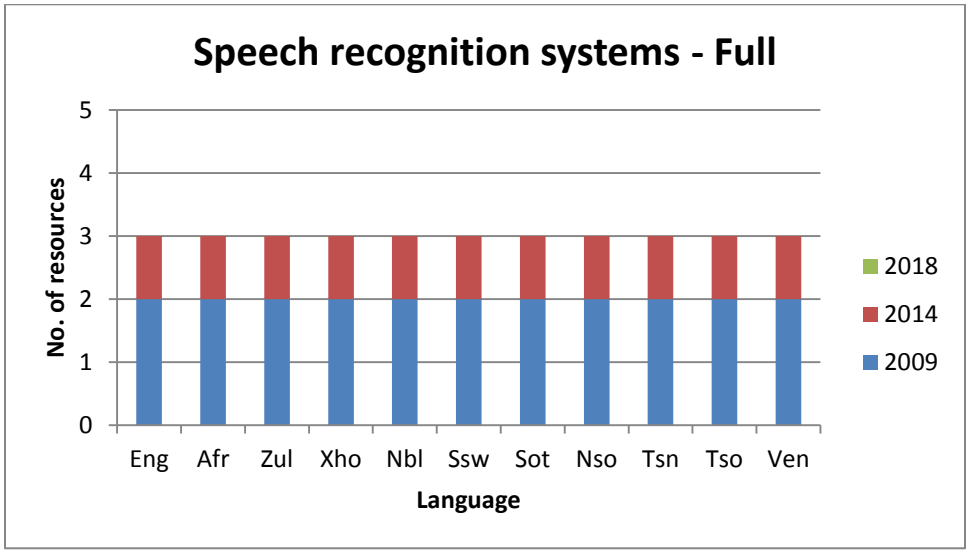


FIGURE 39: REPRESENTATION OF SPEECH-BASED TOOLS

8.4.14.5 Speech recognition systems (Complete/embedded)

Full speech recognition systems for all 11 South African official languages have been available since 2009, with resources added in 2014. Partial, speech recognition systems have been available since 2009 for English and isiXhosa and since 2014 for Afrikaans. No speech recognition systems were added in 2018.



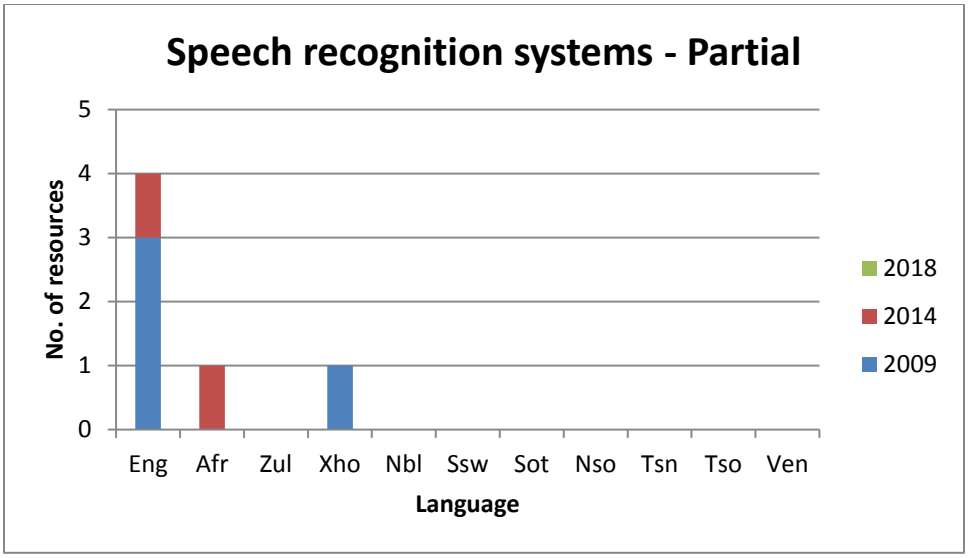


FIGURE 40: REPRESENTATION OF SPEECH RECOGNITION SYSTEMS

8.4.14.6 Machine translators

Full machine translators for Afrikaans, isiZulu, Sepedi, Setswana and Xitsonga were added for the first time in 2018.

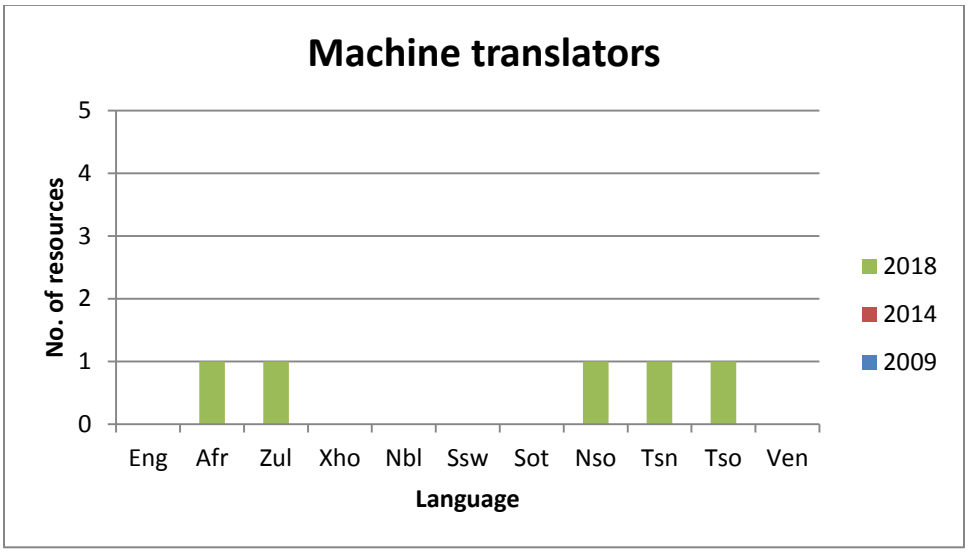


FIGURE 41: REPRESENTATION OF MACHINE TRANSLATORS

8.4.14.7 Language and dialect identifiers

Full language and dialect identifiers for all 11 South African official languages were made available in 2014 and 2018.

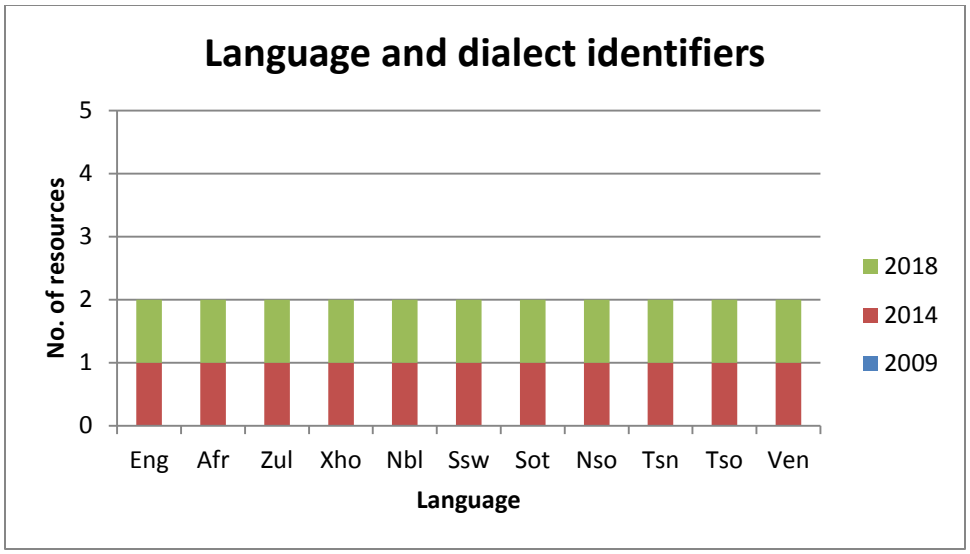


FIGURE 42: REPRESENTATION OF LANGUAGE AND DIALECT IDENTIFIERS

8.4.14.8 Comprehension assistants

Partial comprehension assistants were available for Afrikaans only in 2009 however additional resources were added for Afrikaans, isiXhosa and Sepedi in 2014.

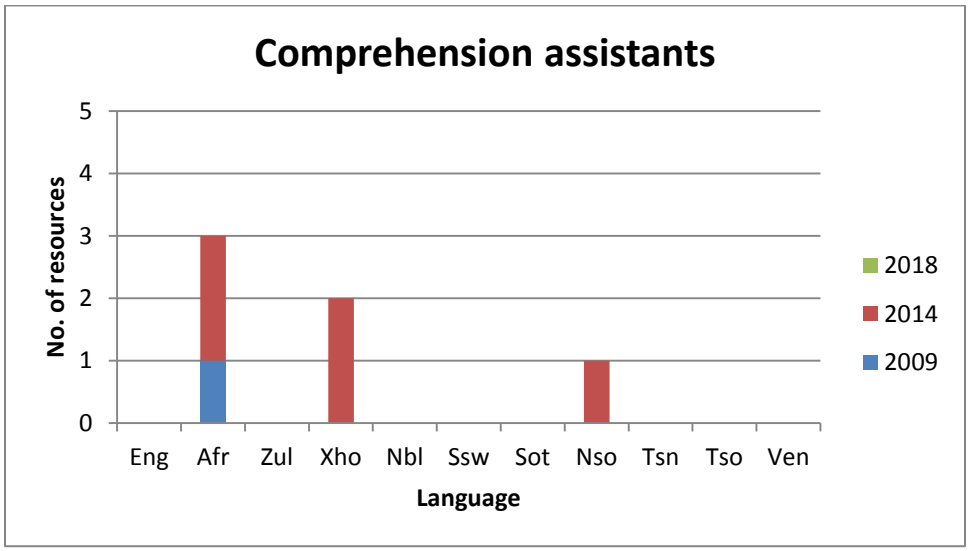


FIGURE 43: REPRESENTATION OF COMPREHENSION ASSISTANTS

8.4.14.9 Machine-aided human translation systems

Full machine-aided human translation resources for all 11 South African official languages have been available since 2009.

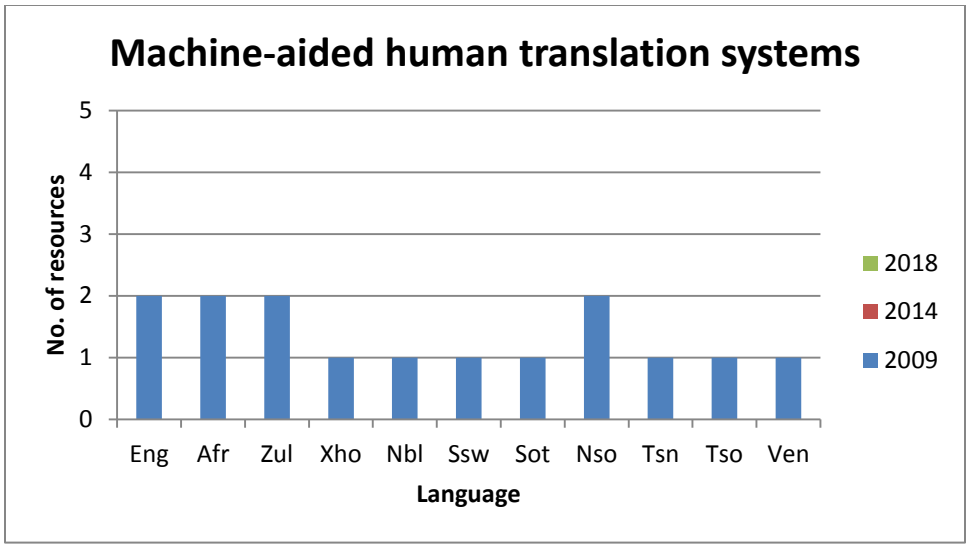


FIGURE 44: REPRESENTATION OF MACHINE-AIDED HUMAN TRANSLATION SYSTEM

8.4.14.10 Human-aided machine translation systems

Partial human-aided machine translation systems have been available since 2009 for English, Afrikaans, isiXhosa and Setswana.

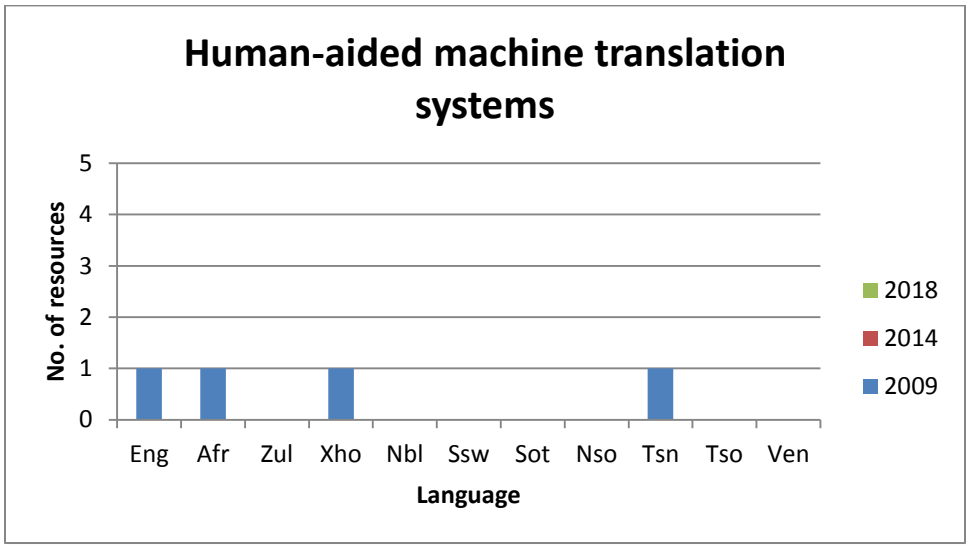


FIGURE 45: REPRESENTATION OF HUMAN-AIDED MACHINE TRANSLATION SYSTEM

8.4.14.11 Format normalisers

Partial format normalisers for all 11 South African official languages have been available since 2014.

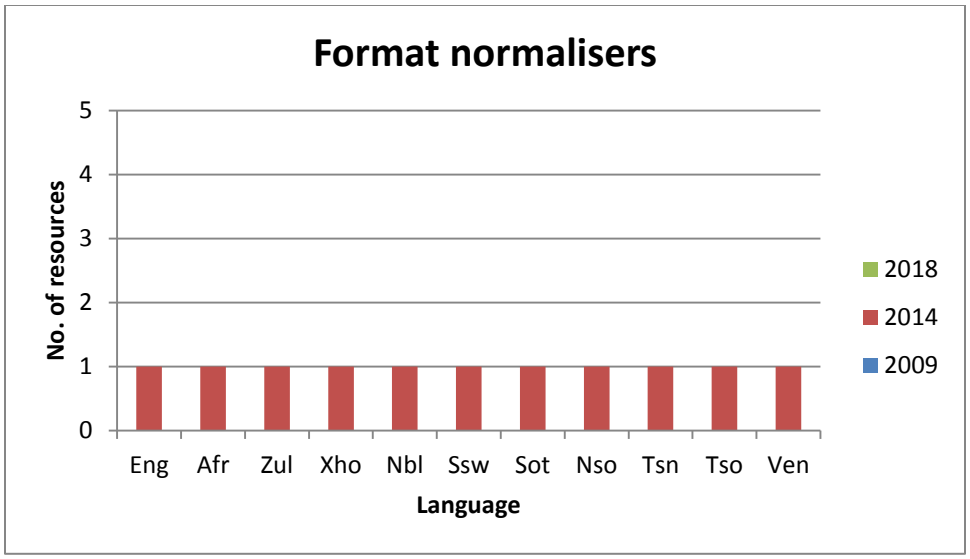
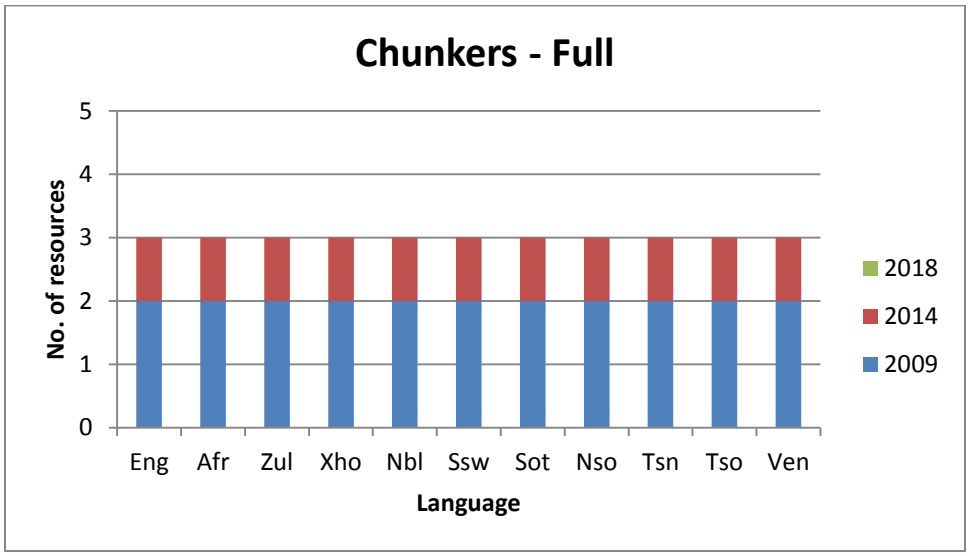


FIGURE 46: REPRESENTATION OF FORMAT NORMALISERS

8.4.14.12 Chunkers

Full chunkers have been available since 2009 and 2014 for all official South African languages. One partial chunker for Afrikaans only has been available since 2014.



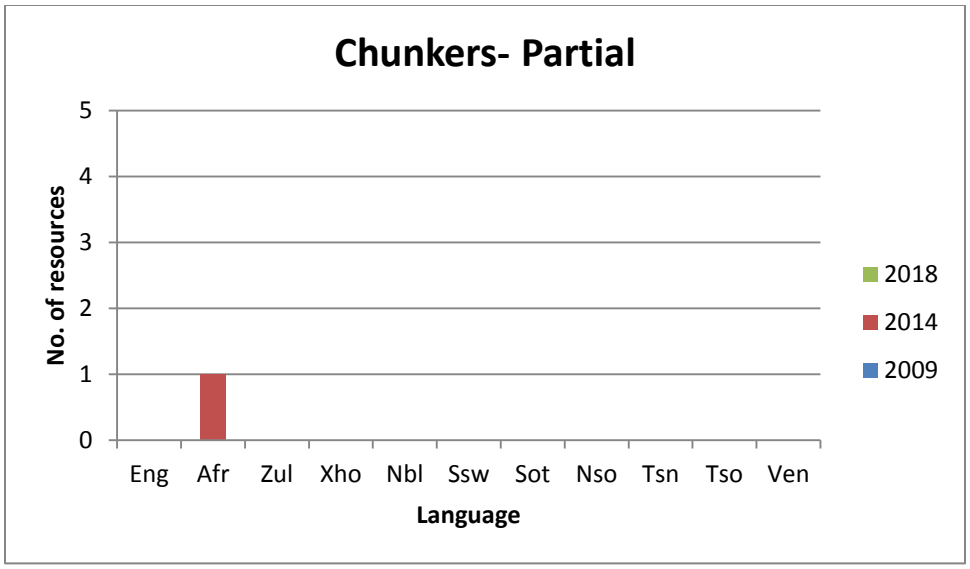


FIGURE 47: REPRESENTATION OF CHUNKERS

8.4.14.13 Automatic phonetic transcriptions

One partial, language independent, automatic phonetic transcription resource has been available since 2009.

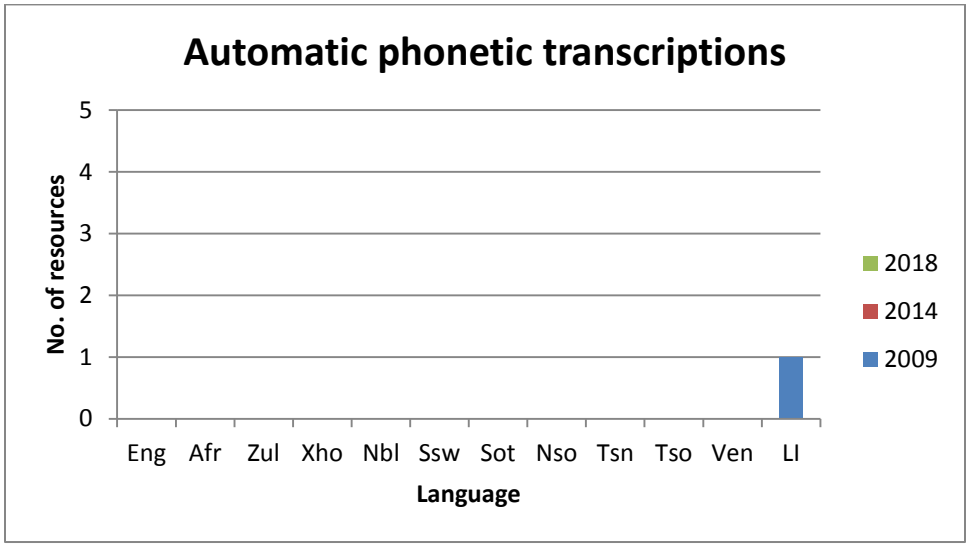
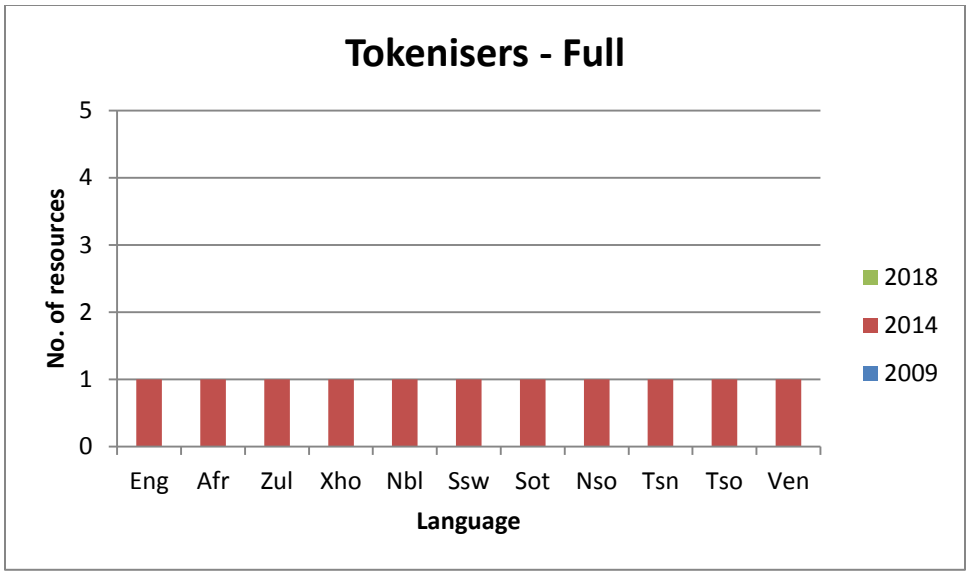


FIGURE 48: REPRESENTATION OF AUTOMATIC PHONETIC TRANSCRIPTIONS

8.4.14.14 Tokenisers

Full tokenisers for all of the 11 official South African languages were made available in 2014. In addition, partial tokenisers for Sepedi and Setswana were also made available since 2014.



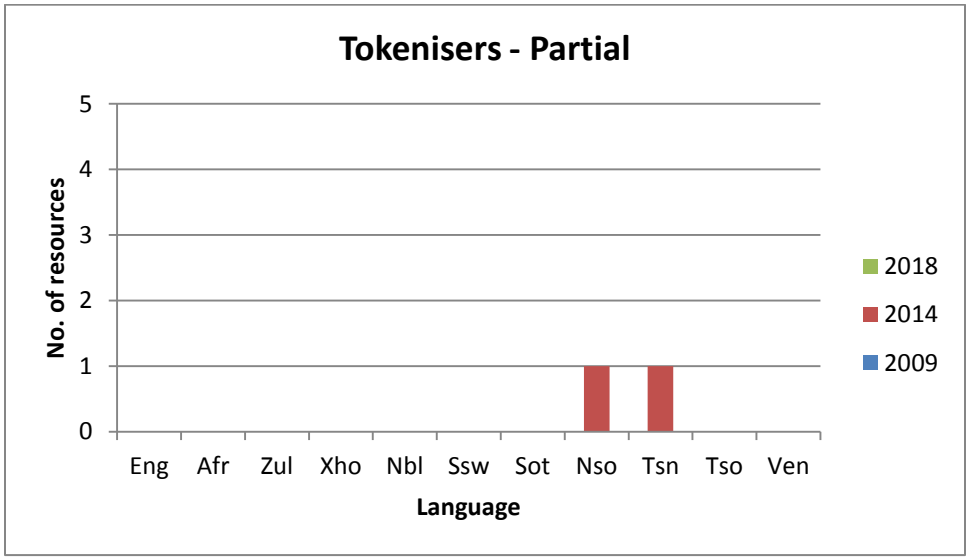


FIGURE 49: REPRESENTATION OF TOKENISERS

8.4.14.15 Limited domain TTS resources (Complete TTS)

Partial limited domain TTS resources exist for all 11 South African official languages and more than one resource exists for English, Afrikaans and isiXhosa.

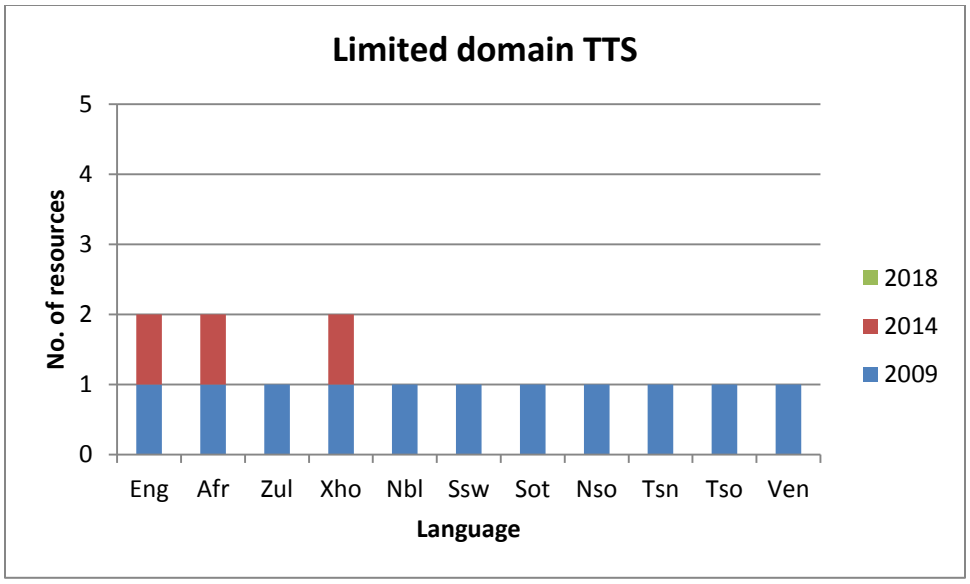


FIGURE 50: REPRESENTATION OF LIMITED DOMAIN TTS

8.4.14.16 Domain independent TTS resources (Complete TTS)

Partial domain independent TTS resources exist for English, isiZulu and isiXhosa.

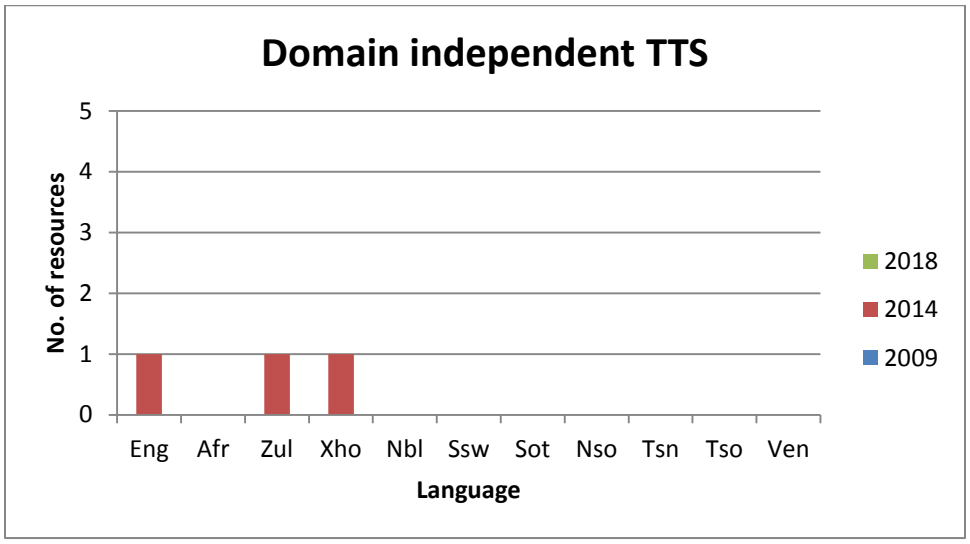


FIGURE 51: REPRESENTATION OF DOMAIN INDEPENDENT TTS

8.4.14.17 Hyphenators

A partial hyphenator is available for Afrikaans only.

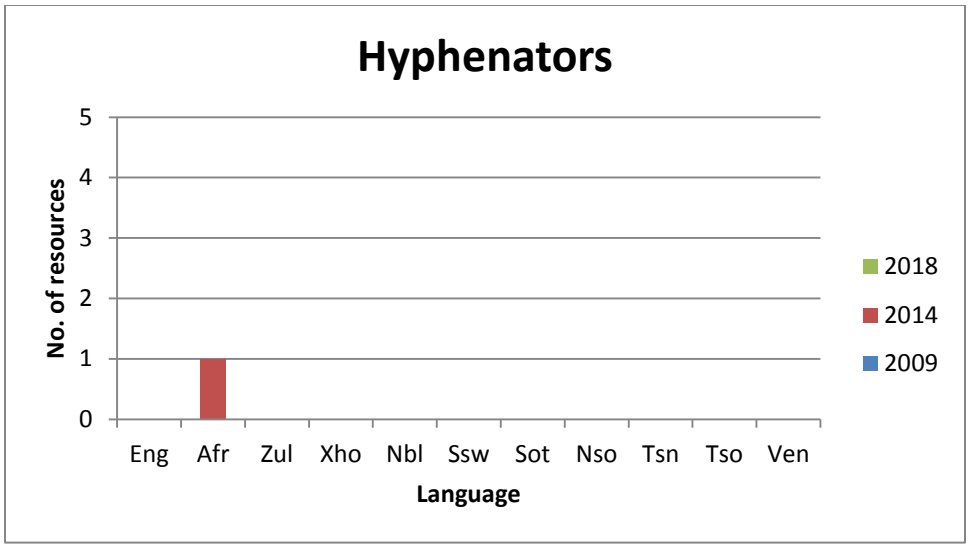


FIGURE 52: REPRESENTATION OF HYPHENATORS

8.4.14.18 Proofing/authoring tools

Partial proofing/authoring tools have been available from 2009 in all 11 South African official languages, with more such tools available for English and Afrikaans than for the African languages. Two language independent proofing/authoring tools were also made available from 2009.

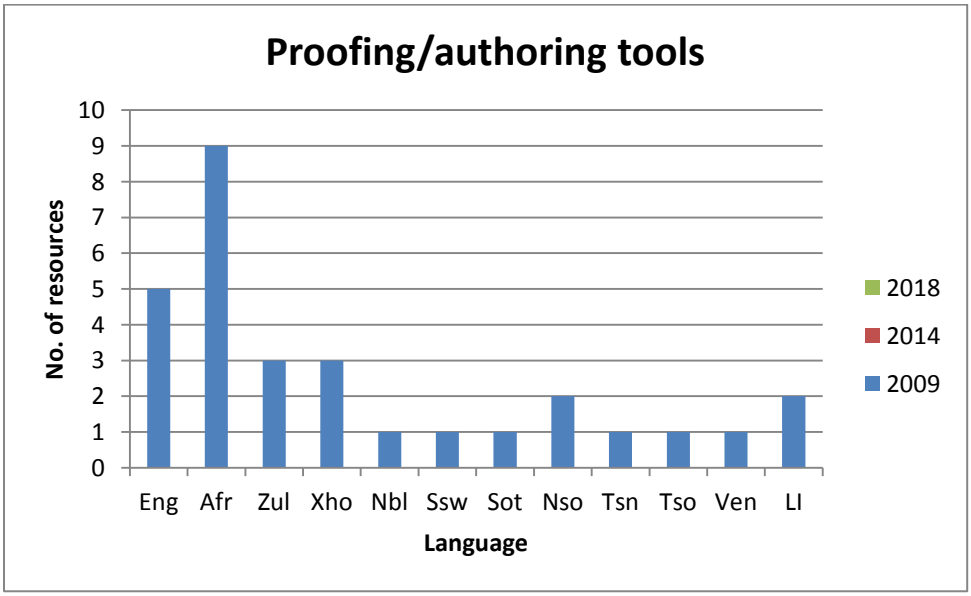
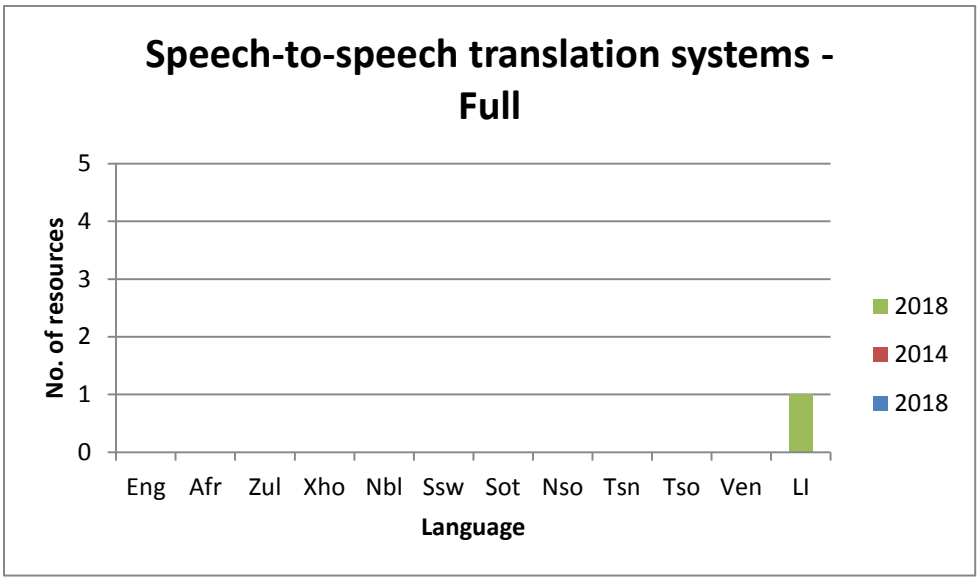


FIGURE 53: REPRESENTATION OF PROOFING/AUTHORING TOOLS

8.4.14.19 Speech-to-speech translation systems

One full language independent speech-to-speech translation system was added in 2018. Speech-to-speech translation systems were already available for English and isiXhosa since 2009.



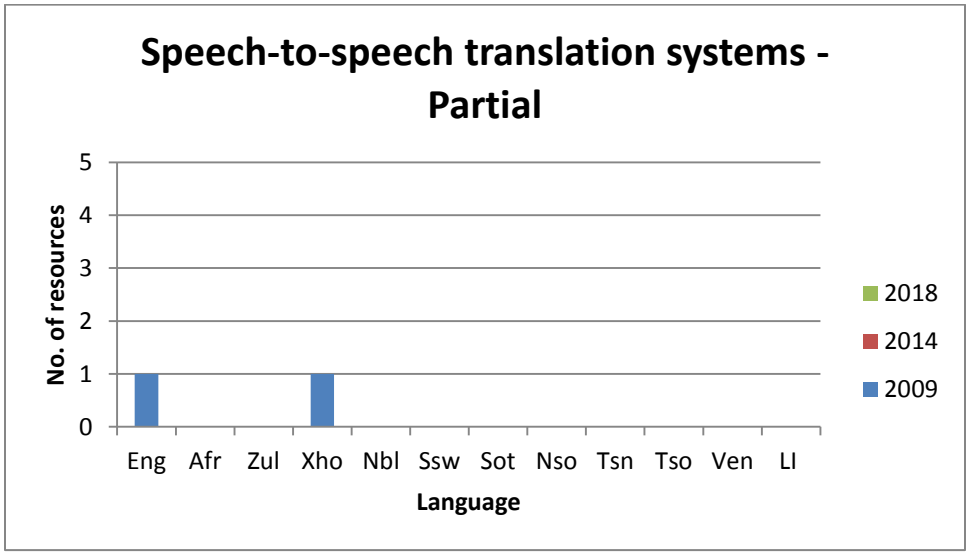


FIGURE 54: REPRESENTATION OF SPEECH-TO-SPEECH TRANSLATION SYSTEMS

8.4.14.20 Named-entity recognisers

Full named-entity recognisers were made available for all the 11 official South African languages in 2009 and 2014. No additional resources were added in 2018.

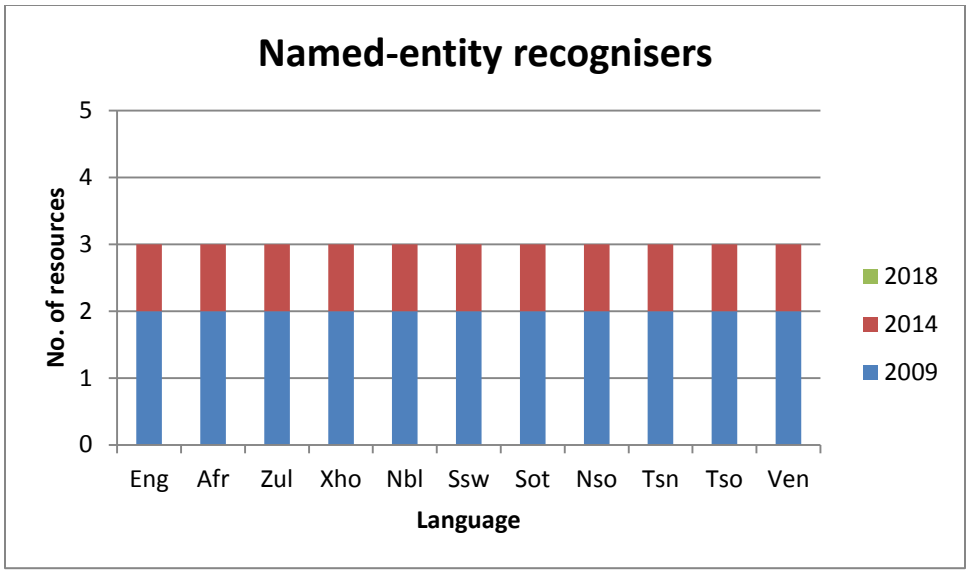


FIGURE 55: REPRESENTATION OF NAMED-ENTITY RECOGNISERS

8.4.14.21 Corpus analysis tools

Full corpus analysis tools were made available in 2014 and in 2018 for all of the 11 official South African languages.

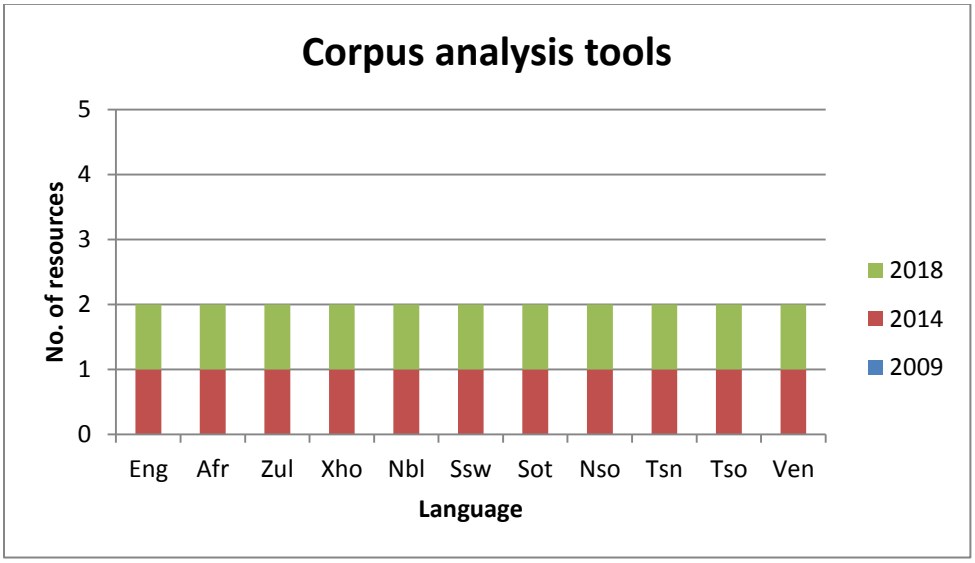


FIGURE 56: REPRESENTATION OF CORPUS ANALYSIS TOOLS

8.4.14.22 Acoustic analysis tools

Full acoustic analysis tools were made available in 2014 for all of the 11 official South African languages. No additional resources were made available in 2018.

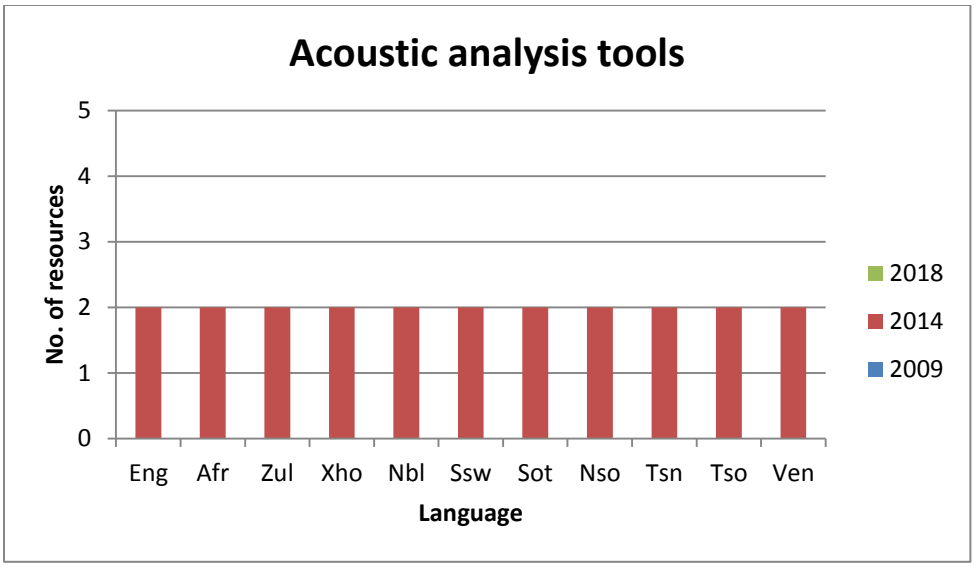


FIGURE 57: REPRESENTATION OF ACOUSTIC ANALYSIS TOOLS

8.4.14.23 OCR/ICR

Full OCR/ICR resources were made available in 2014 all of the 11 official South African languages. No additional resources were made available in 2018.

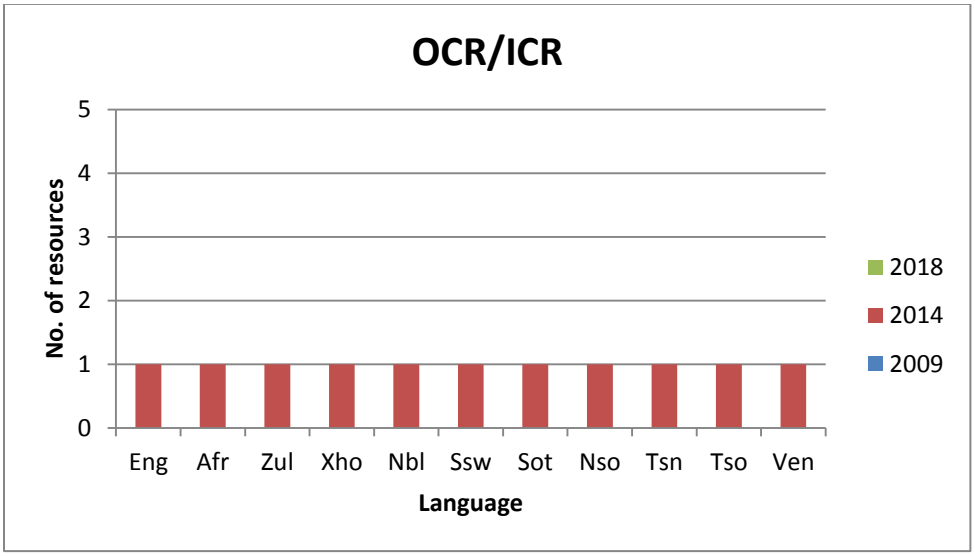


FIGURE 58: REPRESENTATION OF OCR/ICR RESOURCE TYPE

8.4.14.24 Integrated automatic annotation

Full integrated automatic annotation resources were made available in 2014 for all of the 11 official South African languages, except English. No additional resources were made available in 2018.

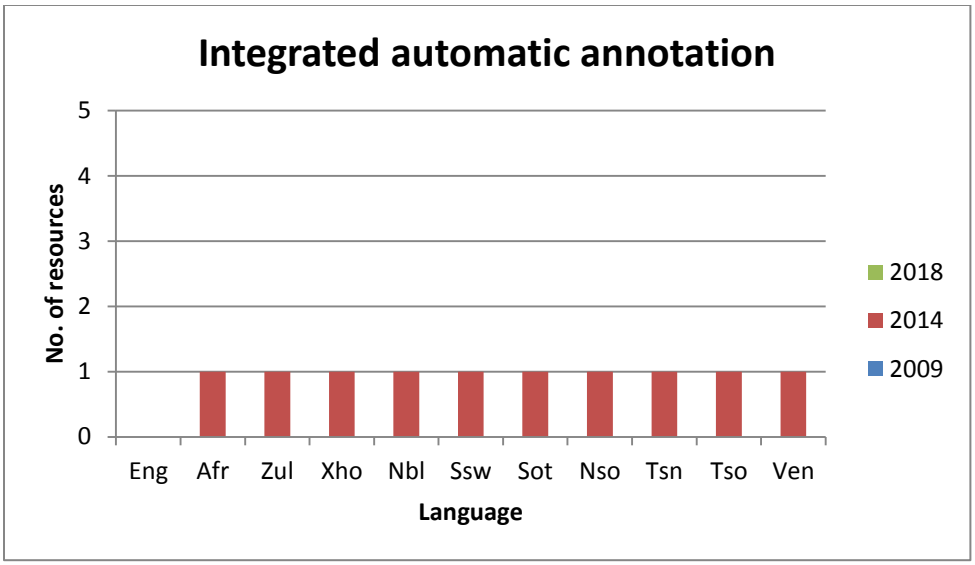


FIGURE 59: REPRESENTATION OF INTEGRATED AUTOMATIC ANNOTATION

COMPARISON OVER TWO DATASETS

8.4.14.25 Grapheme-to-phoneme converters

A partial G2P converter exists for Sepedi.

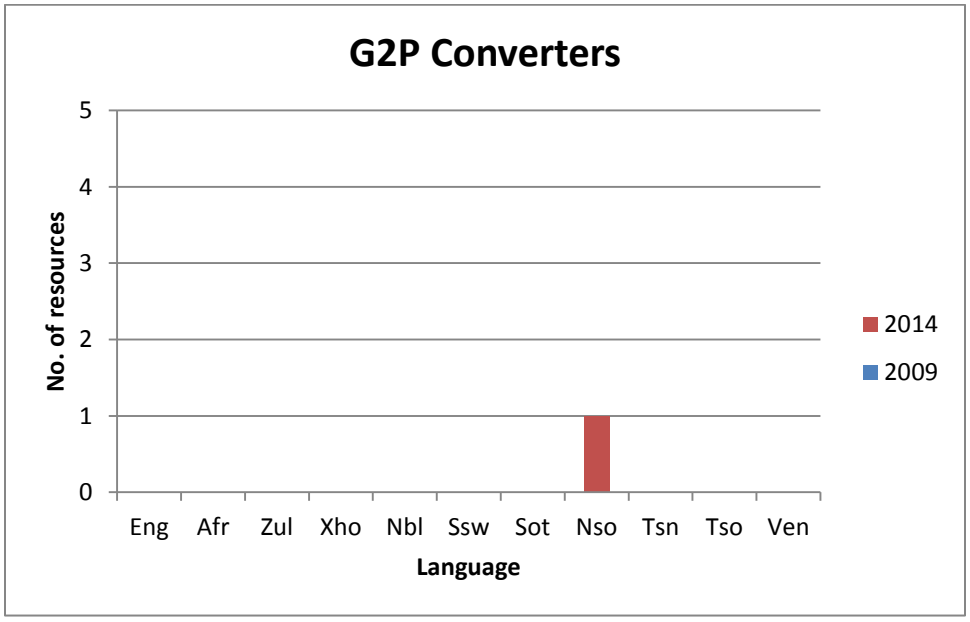
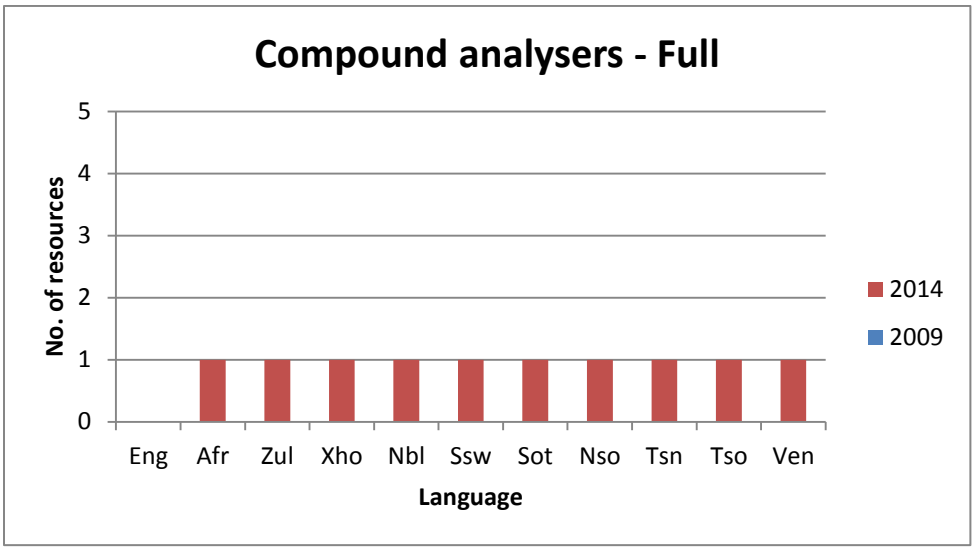


FIGURE 60: REPRESENTATION OF G2P CONVERTERS

8.4.14.26 Compound analysers

Full compound analysers were made available in 2014 for all languages except English. Additional partial compound analysers were made available for Afrikaans in 2014.



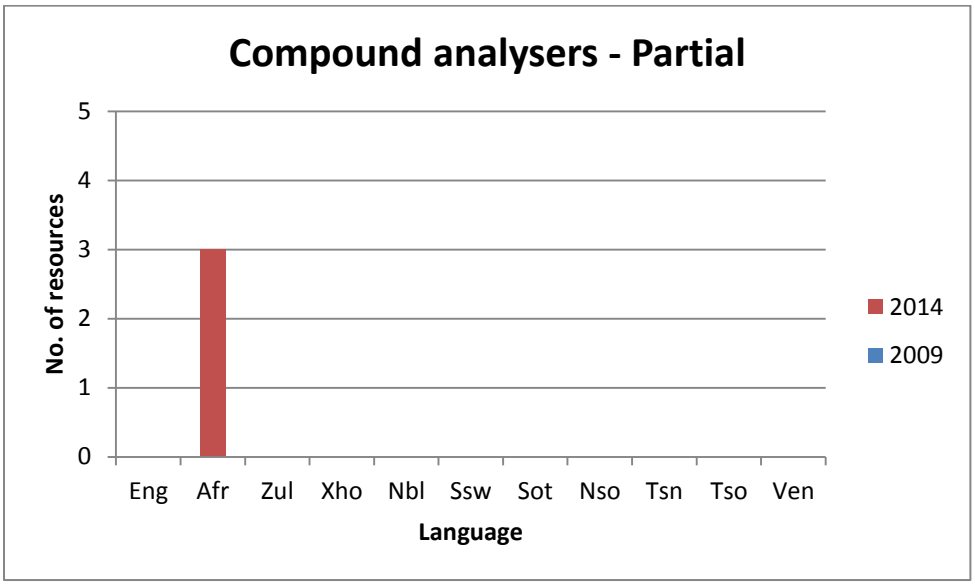


FIGURE 61: REPRESENTATION OF COMPOUND ANALYSERS

8.4.14.27 Computer-assisted Language Learning (CALL)

Partial CALL resources (CALL application) are available in English, Afrikaans, isiZulu, isiXhosa, isiNdebele, Sepedi and Setswana only from 2009. One language independent resource was also made available in 2009. No additional resources were added in 2014 and 2018.

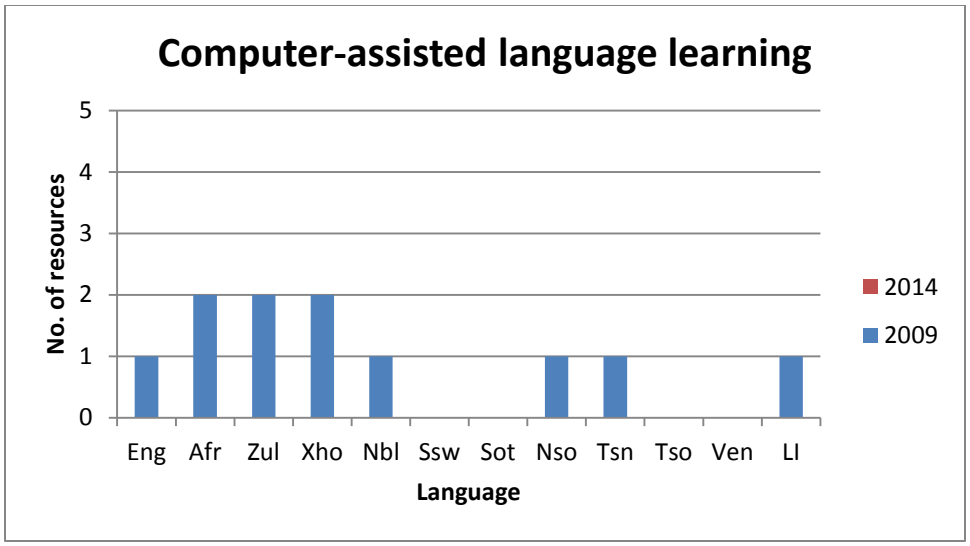


FIGURE 62: REPRESENTATION OF CALL RESOURCES

8.4.14.28 Audio search

Partial audio search software is available in English only.

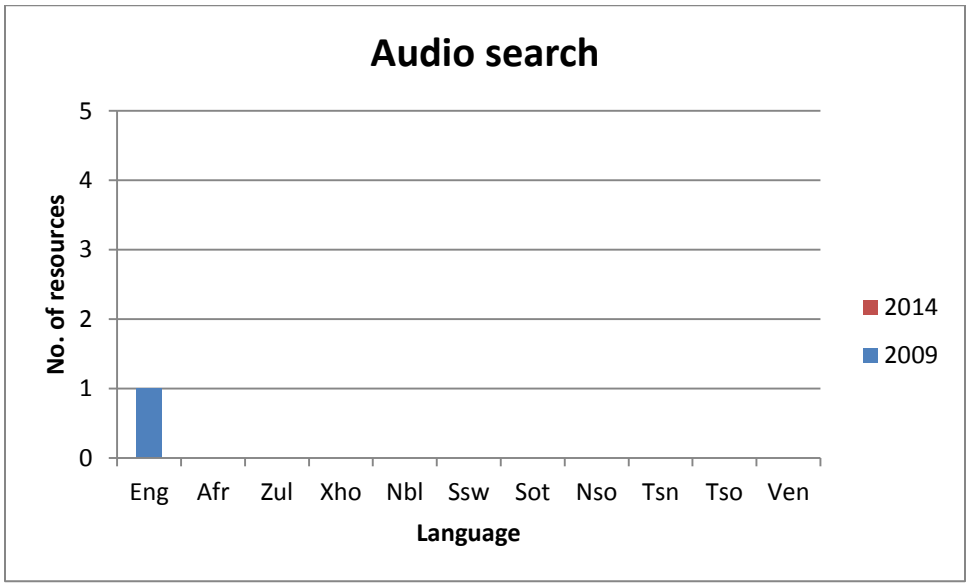


FIGURE 63: REPRESENTATION OF AUDIO SEARCH

8.4.14.29 Access control

A partial access control resource (application) is available for English only.

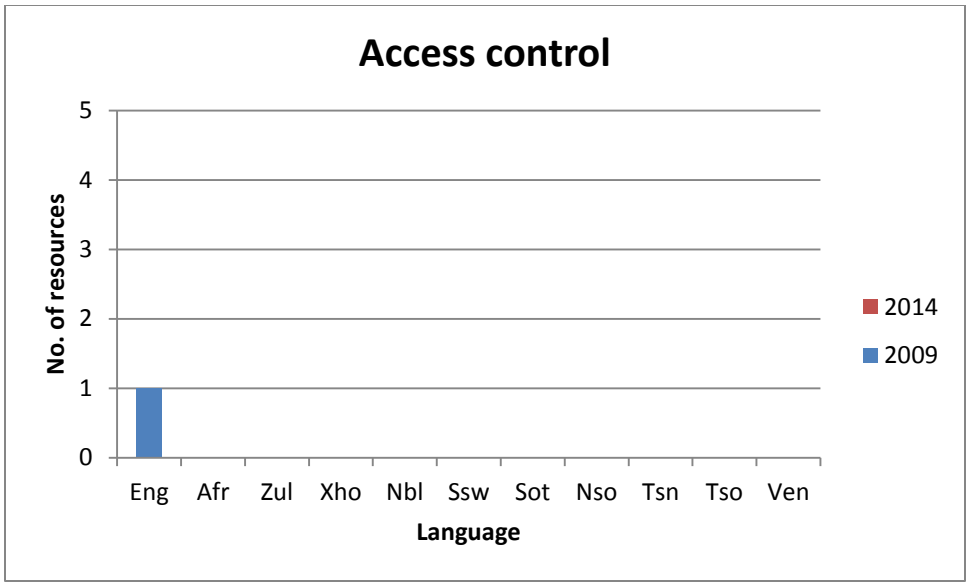


FIGURE 64: REPRESENTATION OF ACCESS CONTROL RESOURCES

8.4.14.30 Speaking devices

Partial speaking devices are available for English and isiXhosa only.

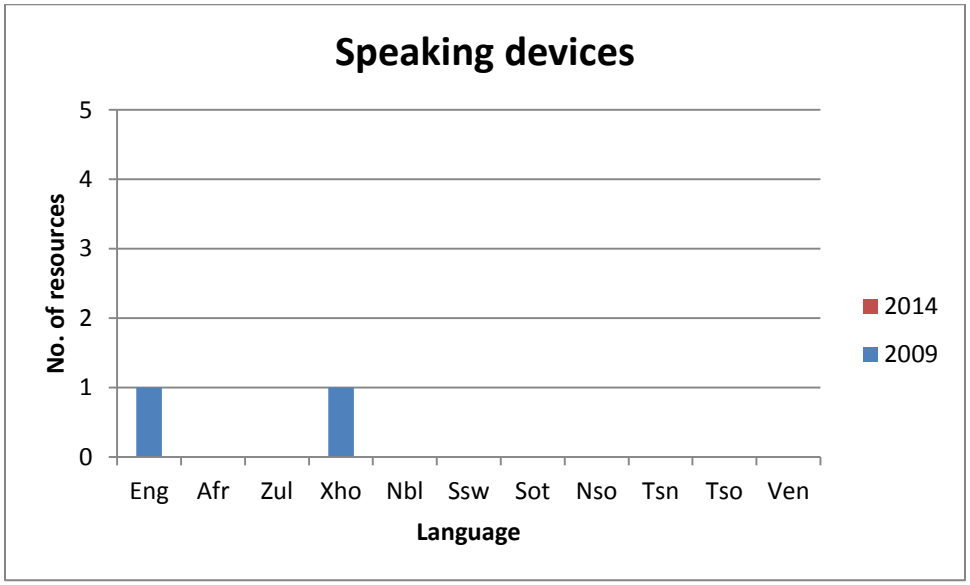


FIGURE 65: REPRESENTATION OF SPEAKING DEVICES

8.4.14.31 Telephony applications

Full and partial telephony applications are available for all 11 South African official languages in 2009. An additional full resource was added in 2014.

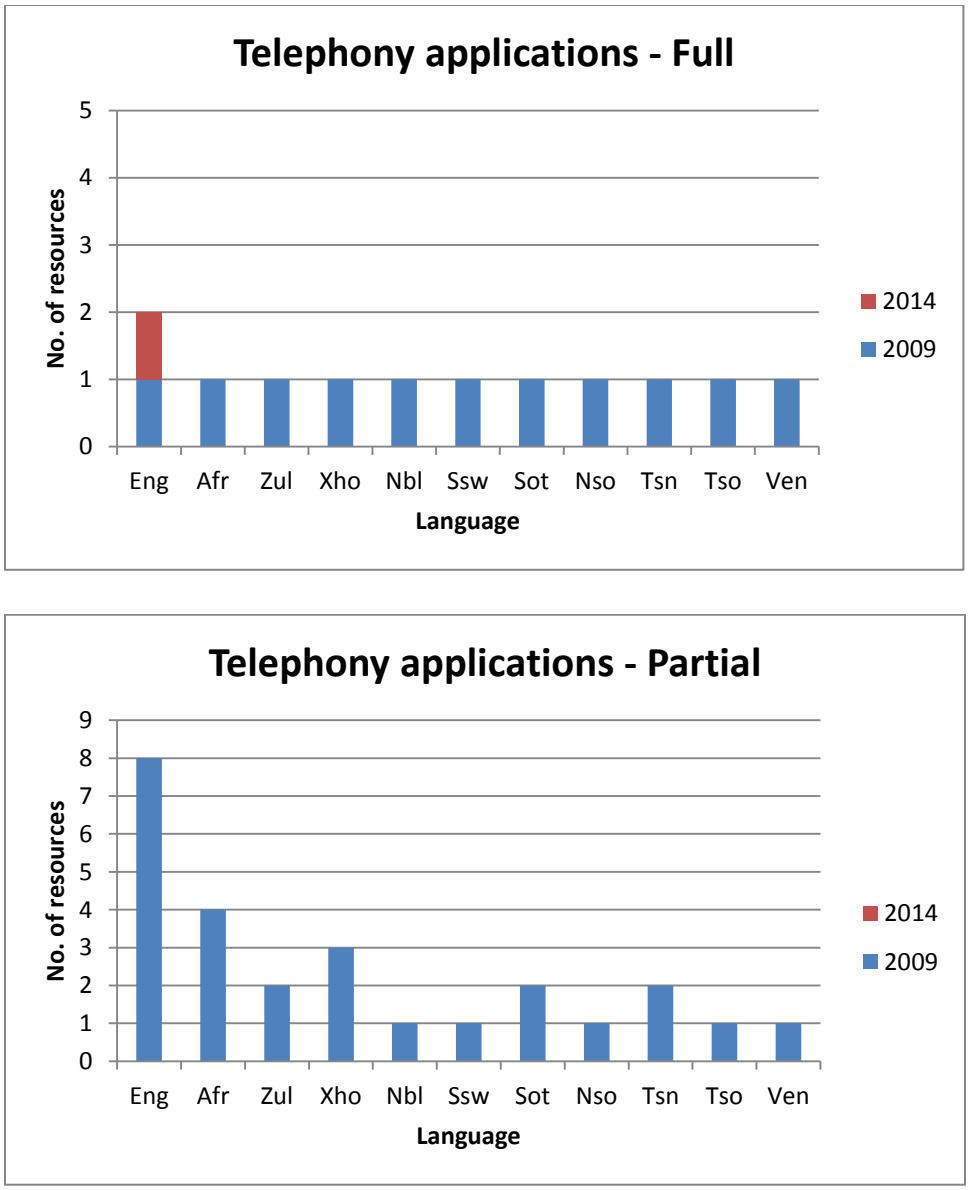


FIGURE 66: REPRESENTATION OF TELEPHONY APPLICATIONS

8.4.14.32 Text selection tools

One language independent partial text selection tool was made available in 2009. No additional resources were made available in 2014.

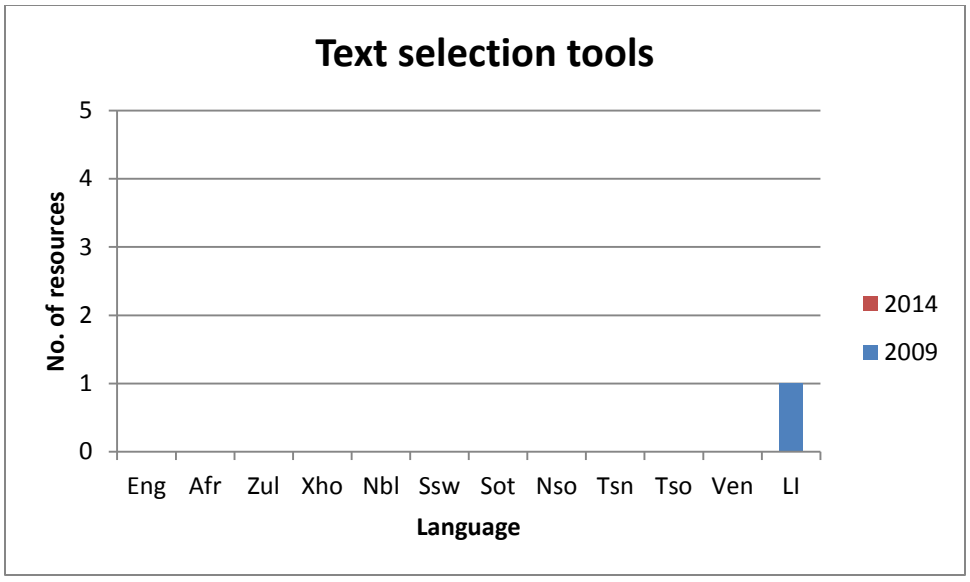


FIGURE 67: REPRESENTATION OF TEXT SELECTION TOOLS

8.4.14.33 Parameter search

One language parameter search resource was made available in 2009. No additional resources were made available in 2014.

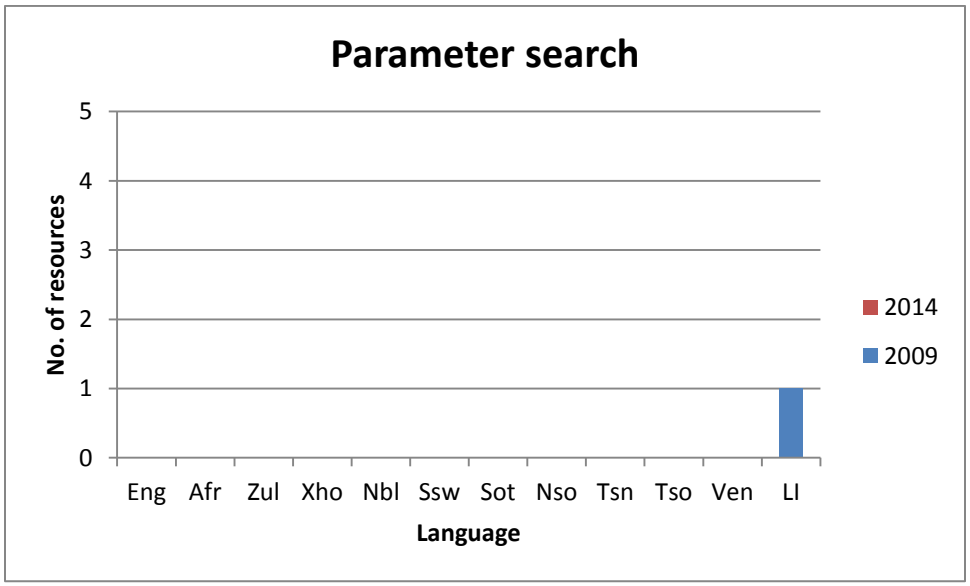
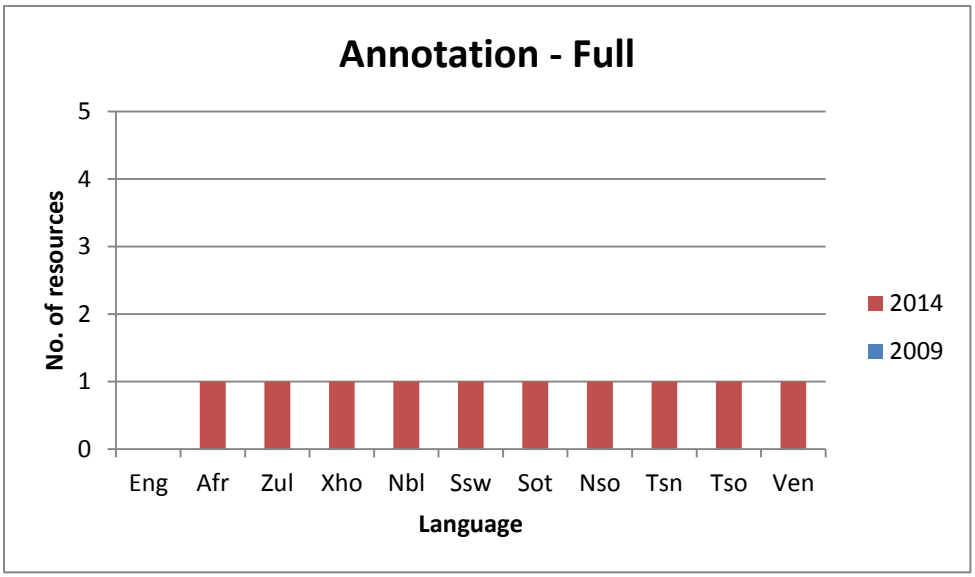


FIGURE 68: REPRESENTATION OF PARAMETER SEARCH RESOURCES

8.4.14.34 Annotation

One full annotation resource was made available for each official South African language (except English) in 2014. One partial language independent resource was available in 2009.



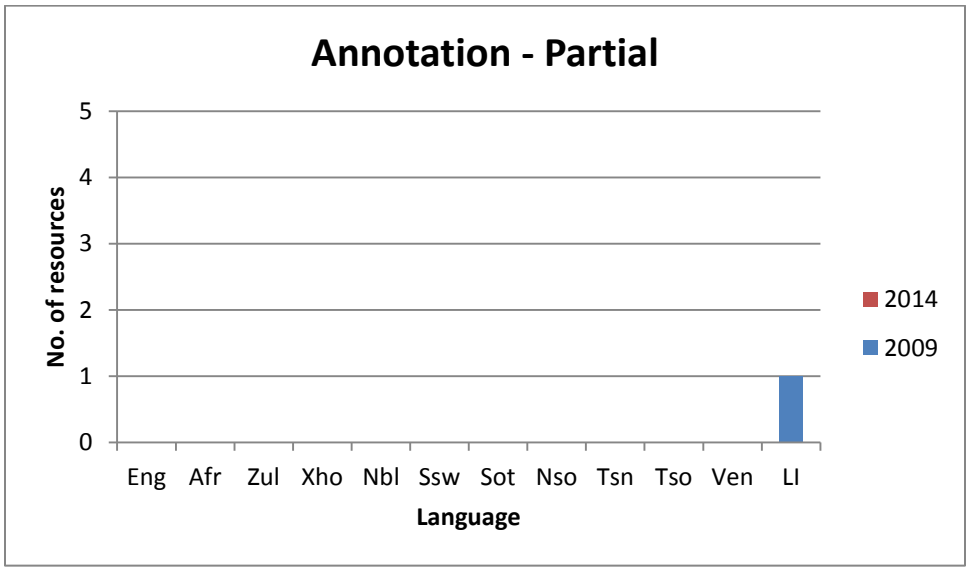


FIGURE 69: REPRESENTATION OF ANNOTATION RESOURCES

8.4.14.35 Web crawler

One partial web crawler language independent resource was made available in 2009.

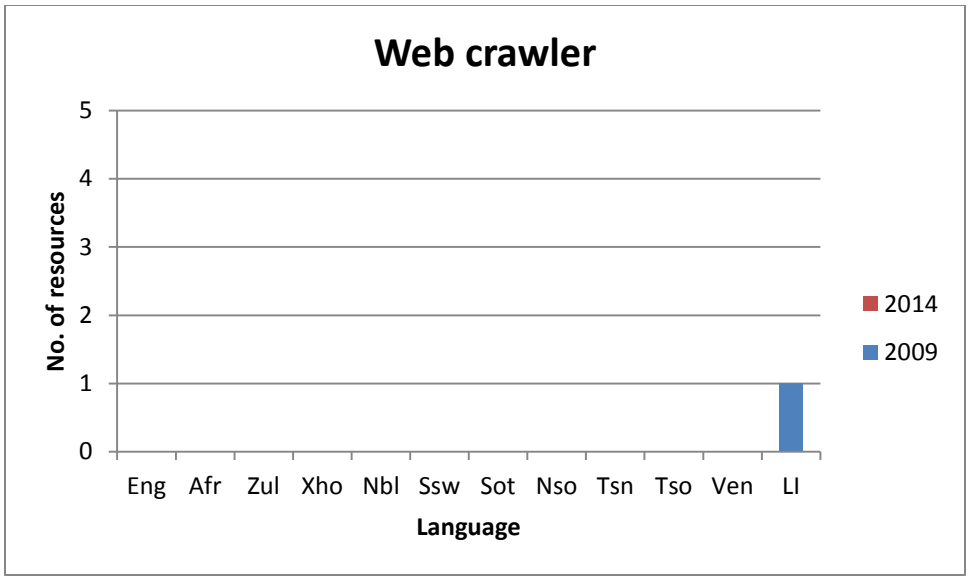


FIGURE 70: REPRESENTATION OF WEB CRAWLER RESOURCES

8.4.14.36 Accessibility

Two partial accessibility resources were made available in 2009 for English, Afrikaans and isiZulu. No resources were submitted in 2014.

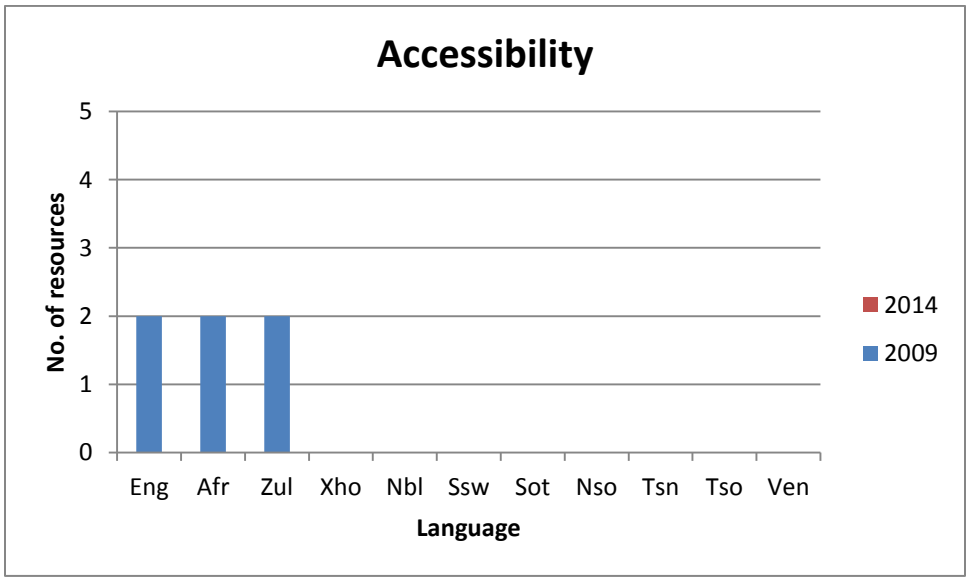


FIGURE 71: REPRESENTATION OF ACCESSIBILITY RESOURCES

8.4.14.37 Multimodal information access

One partial multimodal information access resource was made available for English only in 2009. No resources were submitted in 2014.

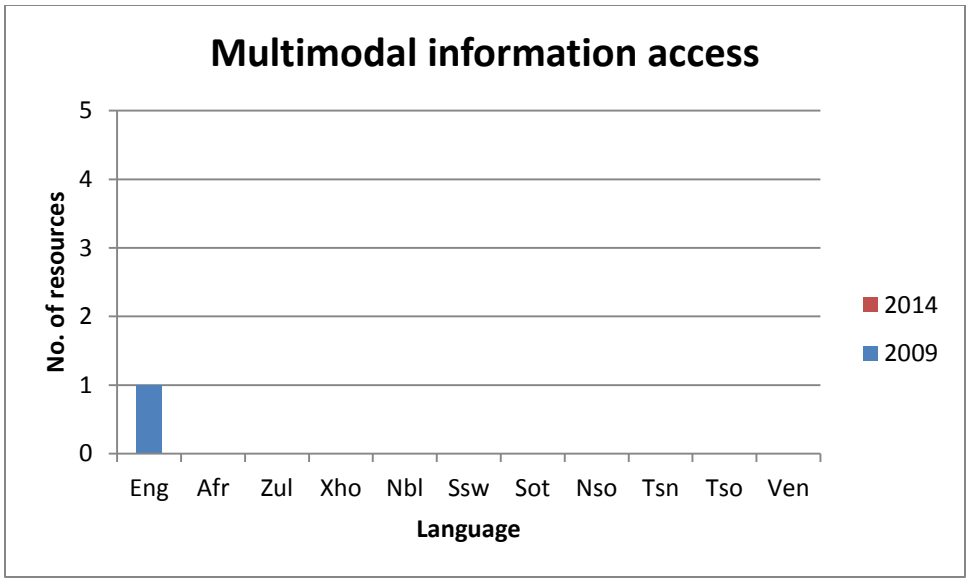


FIGURE 72: REPRESENTATION OF MULTIMODAL INFORMATION ACCESS

8.4.14.38 PDF converters

Full PDF converter resources were made available in 2009, and these resources are language independent.

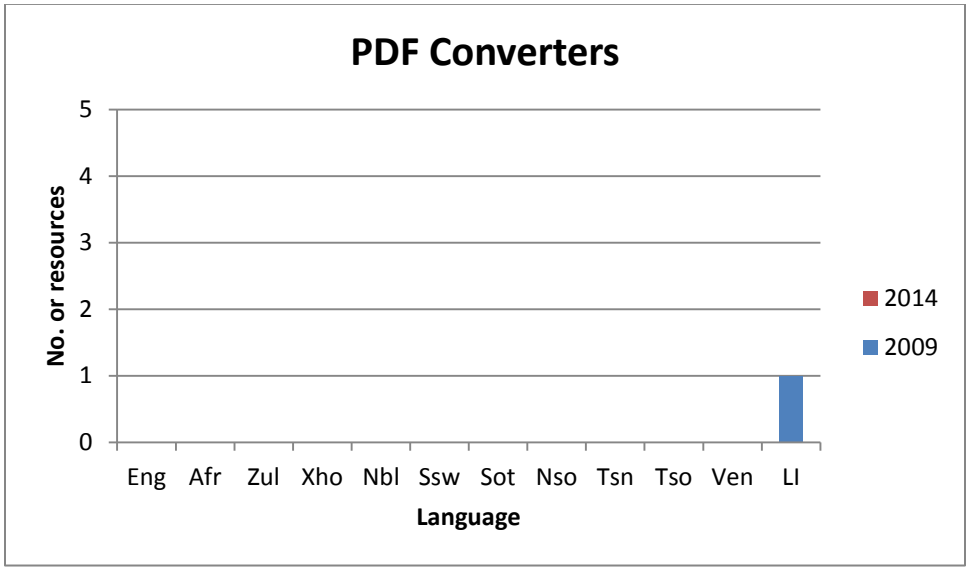


FIGURE 73: REPRESENTATION OF PDF CONVERTERS

8.4.14.39 Anonymisers

Full anonymiser resources were made available in 2009, and these resources are language independent.

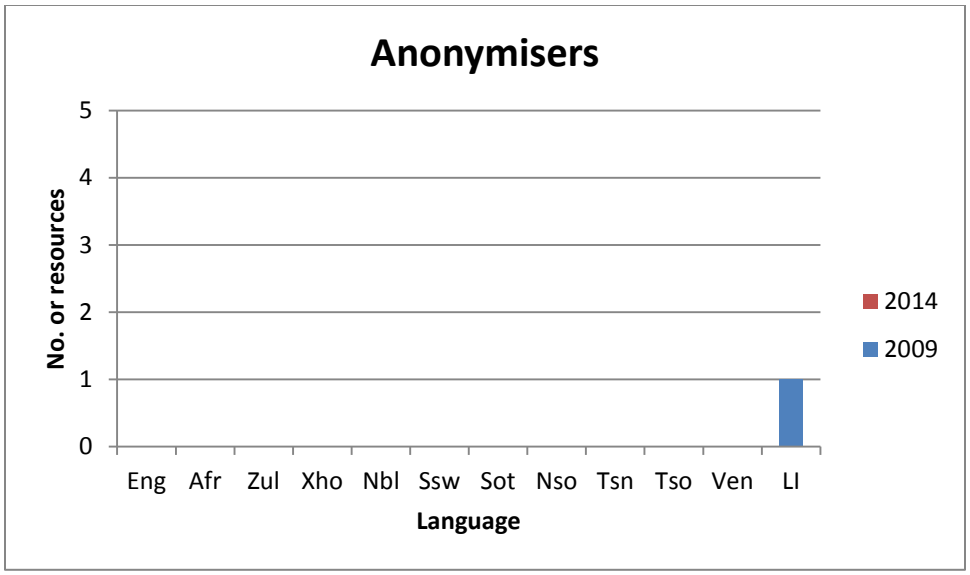


FIGURE 74: REPRESENTATION OF ANONYMISERS

8.4.14.40 Terminology integration texts

Full terminology integration text resources were made available in 2009, and these resources are language independent.

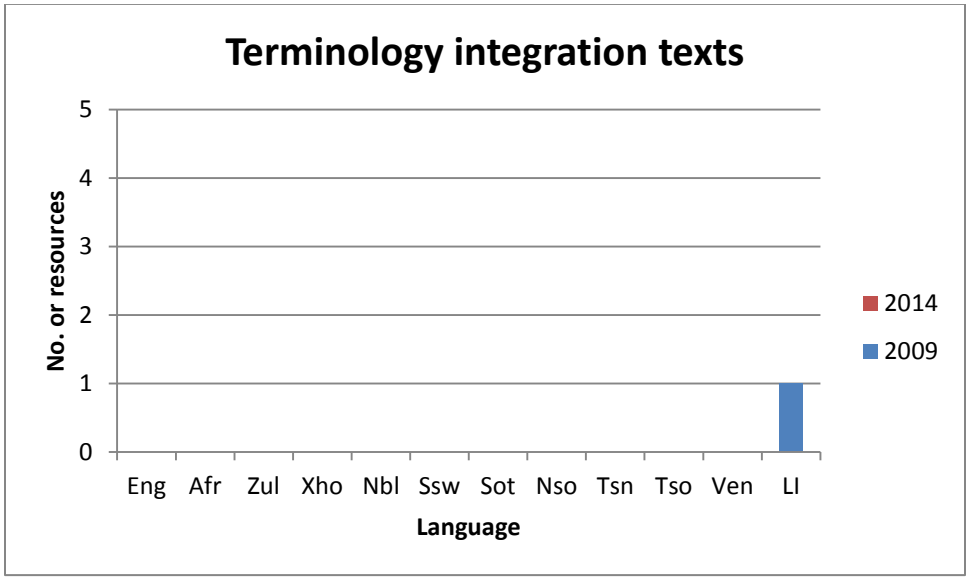


FIGURE 75: REPRESENTATION OF TERMINOLOGY INTEGRATION TEXTS

8.4.14.41 Text aligners

Full text aligner resources were made available in 2009, and these resources are language independent.

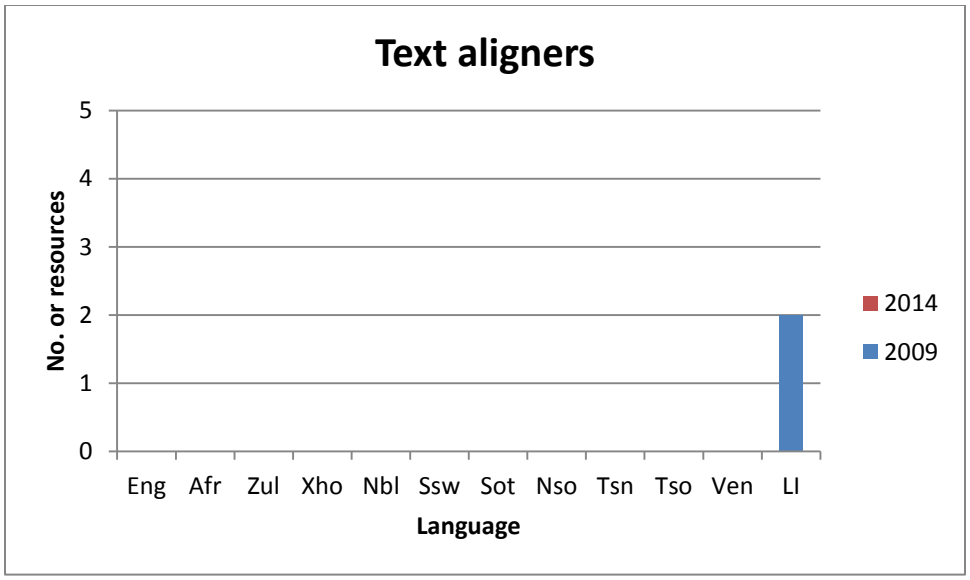


FIGURE 76: REPRESENTATION OF TEXT ALIGNERS

8.4.15 SUMMARY OF THE RESULTS COMPARISON

The above analysis indicates the resource development trends and development progress made from the 2009 Audit to the 2018 Audit. It is clear from the analysis, that while significant progress has been made since 2009 to develop additional resources across more languages, and to develop cutting-edge resources (as we see from the types of resources added in 2018), as well as language independent resources, the more marginalised indigenous languages (particularly Xitsonga, Tshivenda, and isiNdebele), remain severely under-resourced.

In the next section, we discuss the trends and gaps in resource development which still need to be addressed.

8.4.16 DATA ANALYSIS

8.4.16.1 Data analysis process

Detail is provided under each of the resource types represented and compared in section 8.4.9. From the comparison and analysis, it is evident that there are still many resource types that are not available for a number of South African languages. Text resources are better represented than speech resource types and South African English, Afrikaans, isiZulu, isiXhosa, Sepedi and Setswana are the languages in which most resource types are available. In this section, a Maturity Sum, an Accessibility Sum and an HLT Component Sum for all existing resource types will be provided.

As discussed in section 8.4.6 above, the resources that have been submitted over the years are classified according to their level of maturity (under development, alpha version, beta version and released), as well as according to their level of accessibility (not available/proprietary/closed, undecided, research, commercial, open/freely available). As a means to compare the available resources across resource types and across languages, a maturity sum, an accessibility sum and an HLT component sum (Annexure G) were calculated for each resource type across all languages. These calculations were made based on the number of full resources per resource type. The partial resources have not been added to these calculations, as there is no way of knowing whether or not a partial resource has since become a full resource[8]. In Table 13 below, we provide an overview of the number of partial resources that exist across all languages, according to the current available data in all three datasets. Table 14 contains an overview of all the full resources that exist.

8 For the 2018 Audit, we requested that participants only submit resources that are not yet listed on the RMA website. We then requested the 2009 audit data and data on the resources that were submitted between the two audits, which was provided by the RMA.

TABLE 13: EXISTING PARTIAL RESOURCES

Resource type	Language											
	Eng	**Afr**	**Zul**	**Xho**	**Nbl**	**Ssw**	**Sot**	**Nso**	**Tsn**	**Tso**	**Ven**	**LI**
Data												
Text corpora	4	4	5	4	1	1	1	3	3	1	2	0
Wordnets	0	1	0	0	0	0	0	0	0	0	0	0
Monolingual lexicons	1	0	0	0	0	0	0	0	0	0	0	0
Multilingual lexicons	0	2	0	0	0	0	0	0	0	0	0	0
Treebanks	0	0	0	0	0	0	0	1	0	0	0	0
Speech corpora	4	5	0	0	0	0	0	2	1	0	1	0
Pronunciation dictionaries	1	2	0	0	0	0	0	1	0	0	1	0
Multilingual terminology lists	9	9	9	9	9	9	9	9	9	9	9	0
Intonation models	0	0	1	0	0	0	1	1	1	0	0	0
Lexical databases	1	1	0	0	0	0	0	0	0	0	0	0
Test suites and test corpora	0	0	0	0	0	0	0	0	1	0	0	0
Multimedia corpora	0	0	1	1	0	0	0	0	0	0	0	0
Phone mappings	1	1	1	1	1	1	1	1	1	1	1	0
Statistical language models	0	0	0	0	0	0	0	1	0	0	0	0
Other text resources	1	1	1	1	1	1	1	1	1	1	1	1
Software												
Lemmatiser	0	1	0	0	0	0	0	0	0	0	0	0
Morphological analyser	0	0	2	0	0	0	0	0	1	0	0	0
POS tagger/disambiguator	0	2	0	1	0	0	0	1	0	0	0	0
Multilingual comprehension assistants	1	1	0	0	0	0	1	0	0	0	0	0
Format normaliser	1	1	1	1	1	1	1	1	1	1	1	0
Tokeniser	0	0	0	0	0	0	1	1	0	0	0	0
Speech-based tools	0	0	0	0	0	0	0	0	0	0	0	1
Hyphenator	0	1	0	0	0	0	0	0	0	0	0	0
Speech recognition	4	1	0	1	0	0	0	0	0	0	0	0
Speech-to-speech translation system	1	0	0	1	0	0	0	0	0	0	0	0
Domain independent TTS	1	0	1	1	0	0	0	0	0	0	0	0
Chunkers	0	1	0	0	0	0	0	0	0	0	0	0
Automatic phonetic transcription	0	0	0	0	0	0	0	0	0	0	0	1
Proofing/authoring tools	5	9	3	3	1	1	1	2	1	1	1	0
Limited domain TTS	2	2	1	2	1	1	1	1	1	1	1	0
Human-aided machine translation	1	1	0	1	0	0	0	0	1	0	0	0
G2P Converter	0	0	0	0	0	0	0	1	0	0	0	0
Compound analysers	0	3	0	0	0	0	0	0	0	0	0	0

Resource type	Language											
	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Telephony applications	8	4	2	3	1	1	2	1	2	1	1	0
Computer assisted language learning	1	2	2	2	1	0	0	1	1	0	0	1
Audio search	1	0	0	0	0	0	0	0	0	0	0	0
Access control	1	0	0	0	0	0	0	0	0	0	0	0
Speaking devices	1	0	0	1	0	0	0	0	0	0	0	0
Text selection tool	0	0	0	0	0	0	0	0	0	0	0	1
Parameter search	0	0	0	0	0	0	0	0	0	0	0	1
Annotation	0	0	0	0	0	0	0	0	0	0	0	1
Web crawler	0	0	0	0	0	0	0	0	0	0	0	1
Accessibility	2	2	2	0	0	0	0	0	0	0	0	0
Multimodal information access	1	0	0	0	0	0	0	0	0	0	0	0

TABLE 14: EXISTING FULL RESOURCES

Resource type	Language											
	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Data												
Text corpora	9	8	8	7	5	5	6	6	7	8	5	0
Wordnets	0	0	1	1	0	0	0	1	1	0	1	0
Monolingual lexicons	1	1	1	1	1	1	1	1	1	2	1	0
Multilingual lexicons	8	4	3	2	1	1	1	3	1	1	1	0
Treebanks	0	0	0	0	0	0	0	0	1	0	0	0
Speech corpora	16	13	10	7	4	4	10	4	5	4	4	0
Pronunciation dictionaries	2	2	2	2	2	2	3	3	3	2	2	0
Software												
Lemmatisers	0	1	1	1	1	1	1	1	1	1	1	0
Morphological analysers	0	1	1	1	1	1	1	1	1	1	1	0
Machine translators	0	1	1	0	0	0	0	1	1	1	0	0
Language and dialect identifiers	2	2	2	2	2	2	2	2	2	2	2	0
Tokenisers	1	1	1	1	1	1	1	1	1	1	1	0
Speech-based tools	2	2	2	2	2	2	2	2	2	2	2	0
Machine-aided human translation	2	2	2	1	1	1	1	2	1	1	1	0
Speech recognition systems	3	3	3	3	3	3	3	3	3	3	3	0
Speech-to-speech translation systems	0	0	0	0	0	0	0	0	0	0	0	1
Named entity recognisers	1	1	1	1	1	1	1	1	1	1	1	0
Chunkers	0	1	1	1	1	1	1	1	1	1	1	0

Resource type	Language											
	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Corpus analysis tools	2	2	2	2	2	2	2	2	2	2	2	0
Acoustic analysis tools	2	2	2	2	2	2	2	2	2	2	2	0
Compound analysers												
Annotations	0	1	1	1	1	1	1	1	1	1	1	0
OCR/ICRs	1	1	1	1	1	1	1	1	1	1	1	0
Integrated automatic annotation	0	1	1	1	1	1	1	1	1	1	1	0
Telephony applications (IVR/SDS)	2	1	1	1	1	1	1	1	1	1	1	0
Web services	0	1	1	1	1	1	1	1	1	1	1	0
Grammatical framework resource grammars	0	0	0	0	0	0	0	1	0	0	0	0
PDF converters	0	0	0	0	0	0	0	0	0	0	0	1
Anonymisers	0	0	0	0	0	0	0	0	0	0	0	1
Terminology integration texts	0	0	0	0	0	0	0	0	0	0	0	1
Text aligners	0	0	0	0	0	0	0	0	0	0	0	2

8.4.17 MATURITY SUM

The maturity sum provides a measure of the maturity of resources in a language. There are four maturity stages, each with an associated weight. The maturity sum is calculated per resource type by multiplying the number of resources at each maturity stage with the weight associated with the maturity stage, and then summing the products. The weight assigned to a maturity stage is double that of the preceding maturity stage and can thus be either 1 (under development), or 2 (alpha version), or 4 (beta version), or 8 (released version). Table 15 shows how the maturity weight for Afrikaans resource types is calculated. **It is important to note here that for this purpose, the term maturity refers to whether or not a resource has been released, and not to the size, scope or coverage of the resources.**

TABLE 15: MATURITY SUM (MS) FOR AFRIKAANS

Resource type	Number of resources	MS per resource type	Total
Monolingual lexicon	1	**(1 x 1)** + (2 x 0) + (4 x 0) + (8 x 0)	1
Multilingual lexicon	4	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 4)**	32
Text corpora	8	(1 x 0) + (2 x 0) + **(4 x 1)** + **(8 x 7)**	60
Pronunciation dictionary	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Speech corpora	13	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 13)**	104
Lemmatiser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Morphological analyser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Machine translator	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Language and dialect identifier	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16

Resource type	Number of resources	MS per resource type	Total
Machine aided human translation system	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Web service	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Corpus analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Tokeniser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Named entity recogniser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Compound analyser	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Chunker	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
OCR/ICR	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Integrated automatic annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Telephony applications	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Speech recognition systems	3	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 3)**	24
Acoustic analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
	Total		405

Figure 77 shows an overview of the maturity sums for each language. South African English and Afrikaans are the most mature languages, followed closely by isiZulu, Sesotho, Sepedi, isiXhosa, Setswana and Xitsonga, and Tshivenda, isiNdebele and siSwati are the least mature languages. Language independent resources are also shown, but have a very low maturity in comparison to the language specific resources.

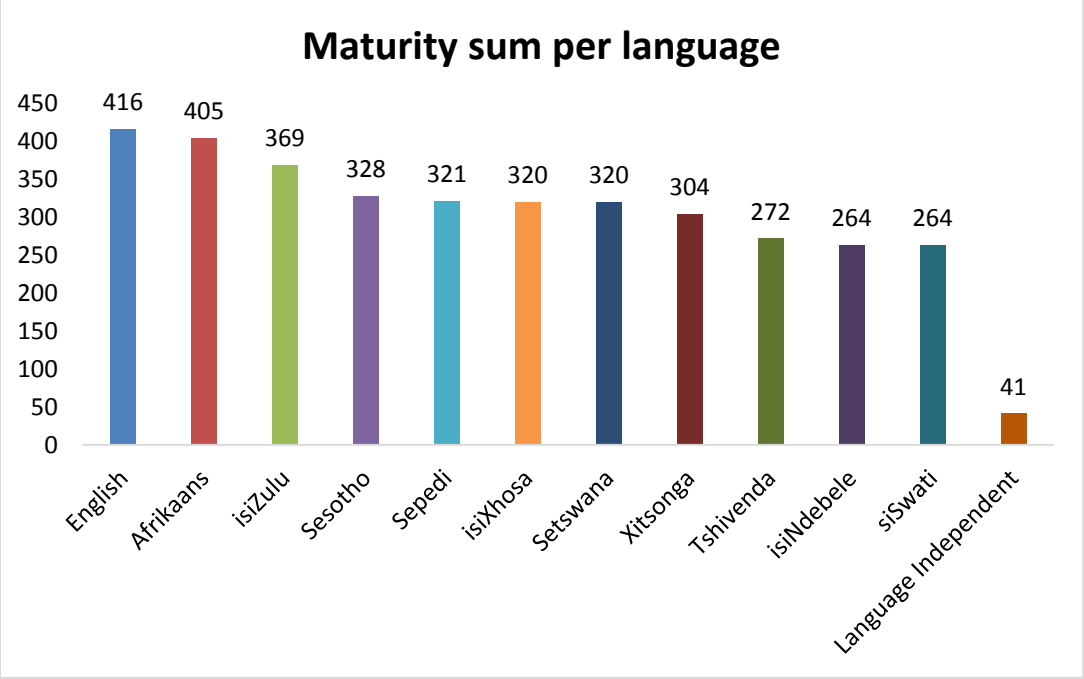


FIGURE 77: MATURITY SUM PER LANGUAGE

Figure 78 illustrates an overview of the maturity sums for all the resource types for which full resources exist in the South African languages. From this chart, it is evident that speech corpora resources are the most mature resource type across all languages. Text corpora and speech recognition systems are the second and third most mature resource types. Speech-to-speech translation systems and grammatical framework resource grammars are the least mature of the languages.

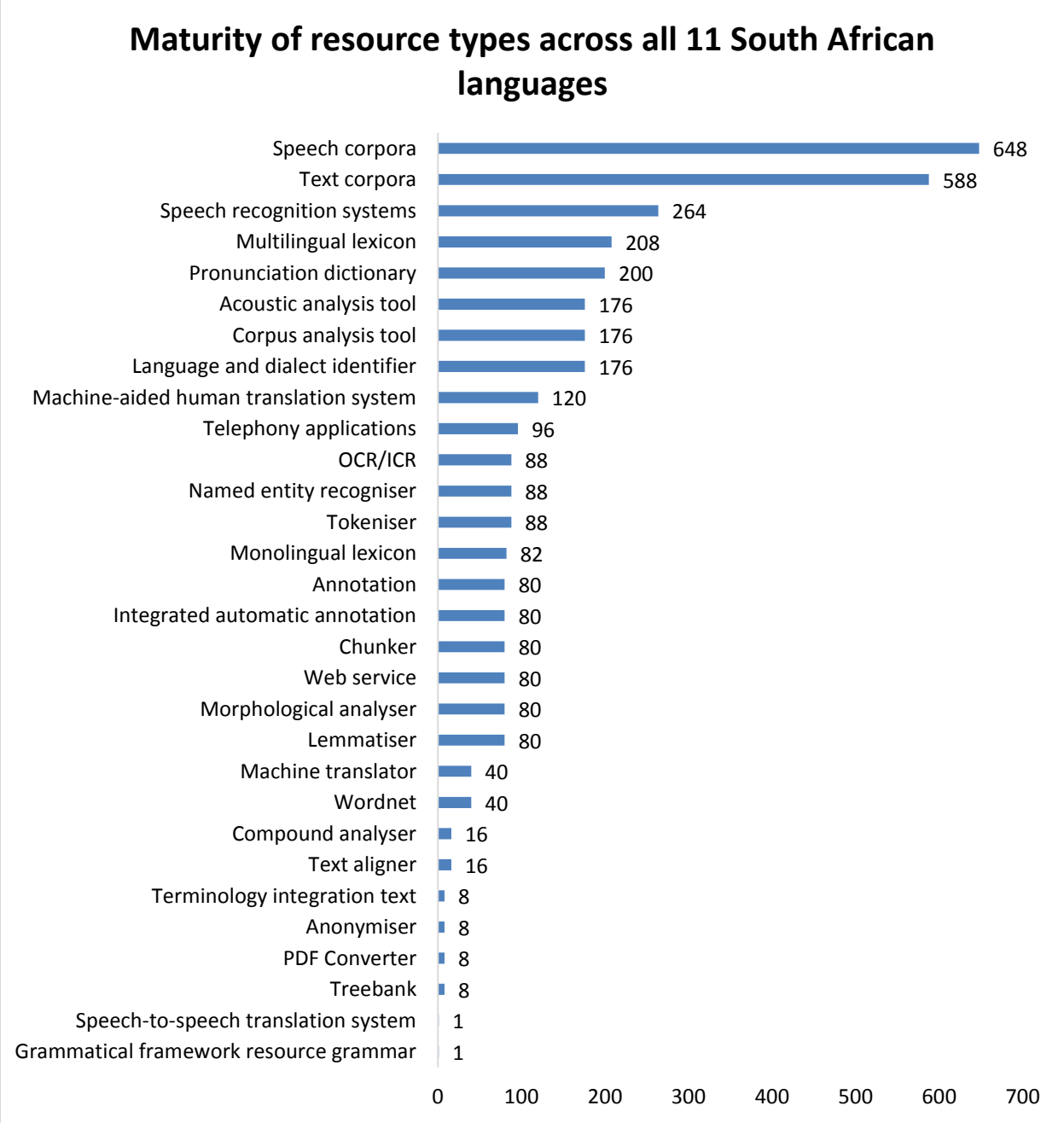


FIGURE 78: MATURITY OF RESOURCE TYPES ACROSS ALL 11 OFFICIAL SOUTH AFRICAN LANGUAGES

8.4.18 ACCESSIBILITY SUM

The accessibility sum provides a measure of the accessibility of resources in a language. The accessibility sum is calculated similar to the maturity sum. For each resource type, the number of resources at each accessibility stage is multiplied with the weight associated with the accessibility stage, before summing the products. The weight assigned to an accessibility stage is double that of the preceding accessibility stage, however, the last stage is a combination of the third and fourth stages, and thus those weights have been added together to create the weight for the last stage. In other words, the weights can be either 1 (not available/proprietary/closed), or 2 (undecided), or 4 (research), or 8 (commercial), or 12 (open/freely available). Table 16 shows how the accessibility weight for Afrikaans resources is calculated.

TABLE 16: ACCESSIBILITY SUM (AS) FOR AFRIKAANS

Resource type	Number of resources	AS per resource type	Total
Monolingual lexicon	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Multilingual lexicon	4	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + **(12 x 2)**	32
Text corpora	8	(1 x 0) + (2 x 0) + **(4 x 1)** + (8 x 0) + **(12 x 7)**	88
Pronunciation dictionary	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Speech corpora	13	(1 x 0) + (2 x 0) + **(4 x 5)** + (8 x 0) + **(12 x 8)**	116
Lemmatiser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Morphological analyser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Machine translator	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Language and dialect identifier	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)** + **(12 x 1)**	20
Machine aided human translation system	2	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	8
Web service	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Corpus analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Tokeniser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Named entity recogniser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Compound analyser	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Chunker	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
OCR/ICR	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Integrated automatic annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Telephony applications	1	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	4
Speech recognition systems	3	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + **(12 x 1)**	20
Acoustic analysis tool	2	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	8
Annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
	Total		500

Figure 79 shows an overview of the accessibility sums for each language. Afrikaans resources are the most accessible, followed by South African English and isiZulu, and Sesotho, isiXhosa, Setswana, Xitsonga and Sepedi follow within close proximity. Tshivenda, isiNdebele and siSwati are the least accessible of the languages. Language independent resources are also shown, but have a very low maturity in comparison to the official languages.

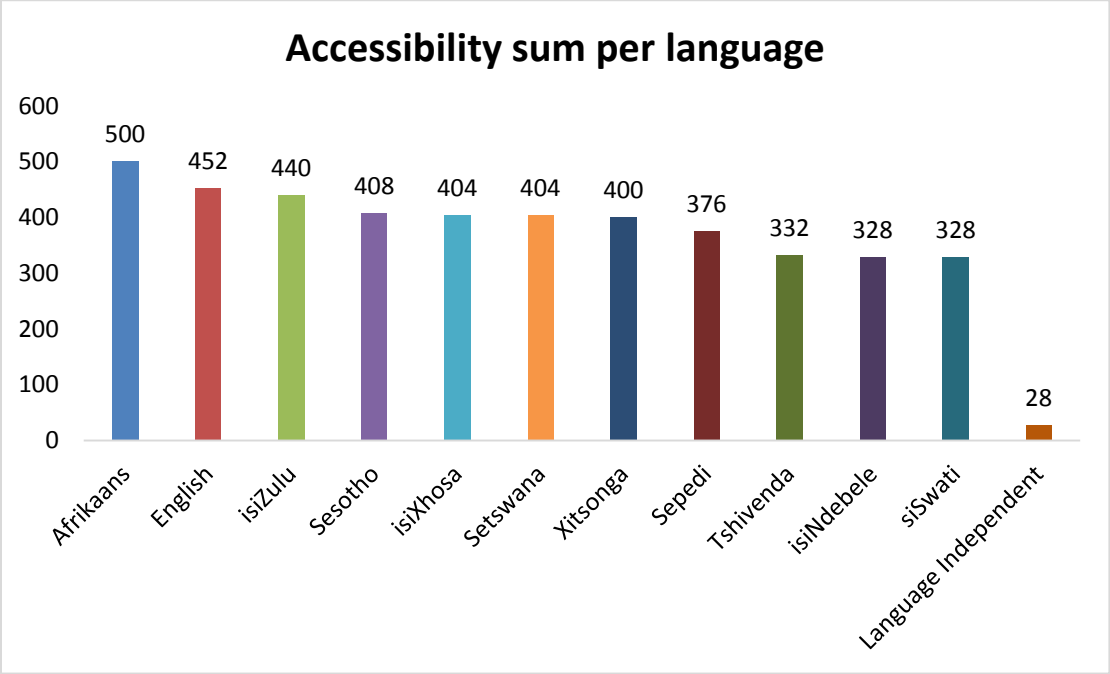


FIGURE 79: ACCESSIBILITY SUM PER LANGUAGE

Figure 80 illustrates an overview of the accessibility for all the resource types for which resources exist in the South African languages. From this chart, it is evident that text corpora resources the most accessible resource type. Terminology integration texts, anonymisers and PDF converters are the least accessible resource types currently available.

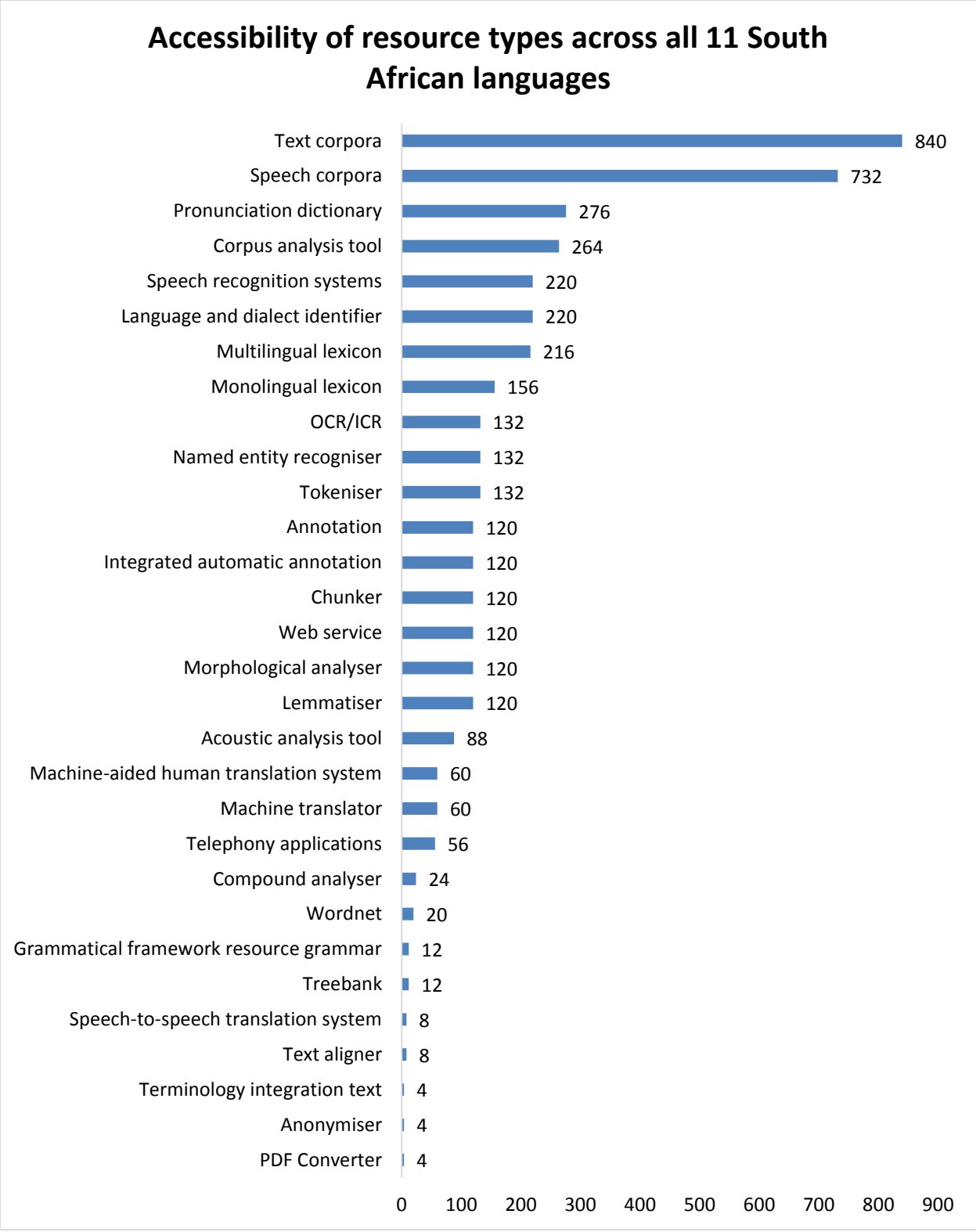


FIGURE 80: ACCESSIBILITY OF RESOURCE TYPES ACROSS ALL 11 SOUTH AFRICAN OFFICIAL LANGUAGES

8.4.19 HLT COMPONENT SUM FOR RESOURCE TYPES

The HLT component sum for resource types provides an overview of the status of HLT development for the eleven South African languages. The HLT component sum for resource types is calculated by adding the maturity sum and the accessibility sum for each language:

$$HLT\ Component\ Sum\ =\ Maturity\ Sum\ +\ Accessibility\ Sum$$

Figure 81 below shows the HLT component sum for resources types across the official languages. The chart indicates that text corpora and speech corpora are the most developed resource types. The least developed resource type is speech to speech translation systems.

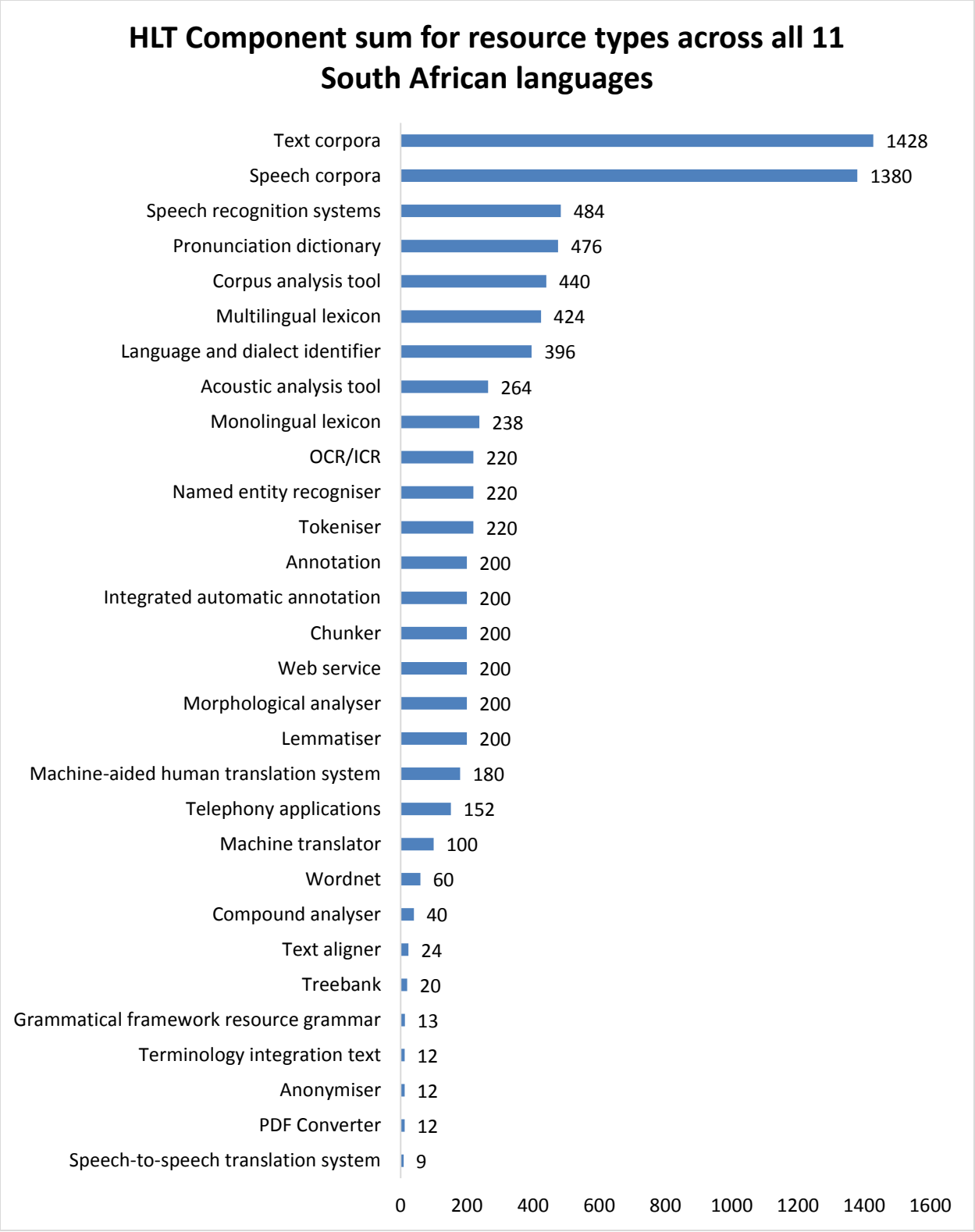


FIGURE 81: HLT COMPONENT SUM FOR RESOURCE TYPES ACROSS ALL 11 OFFICIAL SOUTH AFRICAN LANGUAGES

8.4.20 Overview of existent and non-existent resource types

Table 17 lists the resource types from the Data, Models and Software Categories for which full, partial or no resources are available in any of the datasets (2009, 2014 or 2018). The table classifies the resource types according to the 2018 Audit classification. We have not captured the resource types of 2009 and 2014 for which no resources were submitted as the 2018 Audit resource types will be used in future[9]. This table indicates that there are a number of resource types for which resources still need to be developed.

TABLE 17: RESOURCE TYPES FOR WHICH RESOURCES NEED TO BE DEVELOPED

Resource		Fully available	Partially available	Non existent
Data: Text	Monolingual lexicon	X	X	
	Multilingual lexicon	X	X	
	Terminology list		X	
	Controlled vocabulary			X
	Named entities list			X
	Thesaurus			X
	Wordnets	X	X	
	Ontologies			X
	Text corpora	X	X	
	Treebanks	X	X	
	Statistical language model		X	
	Formal grammar			X
	Tagset			X
	Lexical database		X	
	Test suites and test corpora*		X	
	Other text resources*		X	
Data: Speech	TTS sentence splitting rule sets (manually created)			X
	TTS tokenisation rule sets (manually created)			X
	TTS normalisation rule sets (manually created)			X
	TTS language ID rule sets (manually created)			X
	Language grammar			X
	Phoneme sets			X
	Phone mappings		X	
	Pronunciation dictionaries	X	X	

[9] Data on non-existent resource types from 2009 can be accessed in Sharma-Grover's thesis.

Resource		Fully available	Partially available	Non existent
	Pronunciation rule sets (manually created)			X
	Intonation rule sets (manually created)			X
	Phrasing rule sets			X
	Tone rules sets			X
	Stress rule sets			X
	Syllabification rule sets			X
	Speech corpora	X	X	
	Text corpora for speech			X
Corpora	Multimodal corpora		X	
Models: Speech	TTS Sentence Splitting models			X
	TTS Tokenisation models			X
	TTS Normalisation models			X
	TTS LID models			X
	Language models			X
	Acoustic models			X
	G2P models			X
	Intonation models		X	
	Phrasing models			X
	Tone models			X
	Stress models			X
	Syllabification models			X
Software: Text	G2P Converter		X	
	Tokeniser	X	X	
	Sentenciser			X
	Spelling corrector			X
	Full-form normaliser			X
	Format normaliser		X	
	Number normaliser			X
	Diacritics normaliser			X
	Anonymiser	X		
	Lemmatiser	X	X	
	Stemmer			X
	Morphological analyser	X	X	
	Morphological synthesiser			X
	Part-of-speech tagger/disambiguator		X	
	Syllabifier			X
	Hyphenator		X	
	Dependency parser			X

Resource		Fully available	Partially available	Non existent
	Constituent recogniser			X
	Chunker	X	X	
	Event extractor			X
	Named entity recogniser	X		
	Terminology extractor			X
	Topic modelling			X
	Sentiment analysis/affect/emotion analyser			X
	Referent resolver			X
	Word meaning disabmiguator			X
	Pragmatic analyser			X
	Text generator			X
	Summariser			X
	Machine translator	X		
	Semantic analyser: Frame extractor			X
	Language and dialect identifier	X		
	Shallow parser: Relation finder			X
	Proofing/authoring tool		X	
	Information retrieval system			X
	Information extractor			X
	Human-aided machine translation system		X	
	Machine-aided human translation system	X		
	OCR/ICR	X		
	Computer-aided language learning (CALL) system		X	
	Document classifier			X
	Authorship identifier			X
	Question answering (QA) system			X
	Dialogue system (text-based)			X
	Comprehension assistant		X	
	Web service	X		
	Grammatical framework resource grammar	X		
	Corpus analysis tool	X		
	Compound analyser	X	X	
	PDF Converter	X		
	Anonymiser	X		
	Terminology integration text	X		
	Text aligner	X		
	Text selection tool		X	
	Web crawler		X	

Resource		Fully available	Partially available	Non existent
	Annotation	X	X	
	Parameter search		X	
	Integrated automatic annotation	X		
Software: Speech	Language modelling tool			X
	Pronunciation dictionary creation tool			X
	G2P tool			X
	Acoustic modelling tool			X
	Intonation tool			X
	Phrasing tool			X
	Tone tool			X
	Stress tool			X
	Syllabification tools			X
	Vocoder			X
	TTS Sentence Splitting tool			X
	TTS Tokenisation tool			X
	TTS Normalisation tool			X
	TTS LID tool			X
	Large vocabulary speech recognition system	X	X	
	Command and control system			X
	Non-native speech recognition system			X
	Code-switched speech recognition system			X
	Multilingual speech recognition system	X	X	
	Noise robust speech recognition system		X	
	Embedded speech recognition system	X	X	
	Alignment system			X
	Automatic phonetic transcription system			X
	Confidence measures			X
	Acoustic Language ID			X
	Acoustic Age ID			X
	Acoustic Gender ID			X
	Acoustic Dialect ID			X
	Acoustic Emotion ID			X
	Key-word spotting system			X
	Voice activity detection			X
	Speaker tracking			X
	Acoustic speaker ID			X
	Speaker verification system			X
	Diarisation			X

Resource		Fully available	Partially available	Non existent
	Complete TTS System		X	
	Speech-to-speech translation system	X	X	
	Audio search		X	
	Access control		X	
	Speaking devices		X	
	Accessibility		X	
	Telephony applications	X	X	
	Acoustic analysis tool	X		
	Multimodal information access		X	
	Speech-based tools*		X	
*These resource types existed for the 2009 and 2014 data sets. The types were not included in the 2018 list of resource types, but we still decided to list these items for the purposes of indicating existent resources.				

8.4.21 SUMMARY OF DATA ANALYSIS PER CATEGORY AND LANGUAGE

In Figures 82 – 85 below, we have summarised the total full text and speech resources in the data and software categories for the period 2009-2018 across the languages.

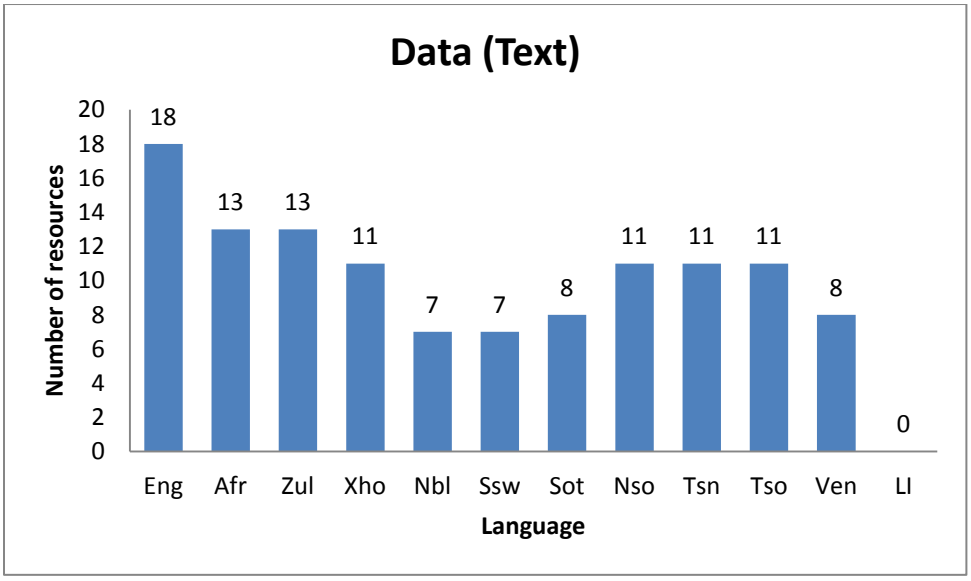


FIGURE 82: SUMMARY OF DATA (TEXT) RESOURCE TYPES FROM 2009 - 2018

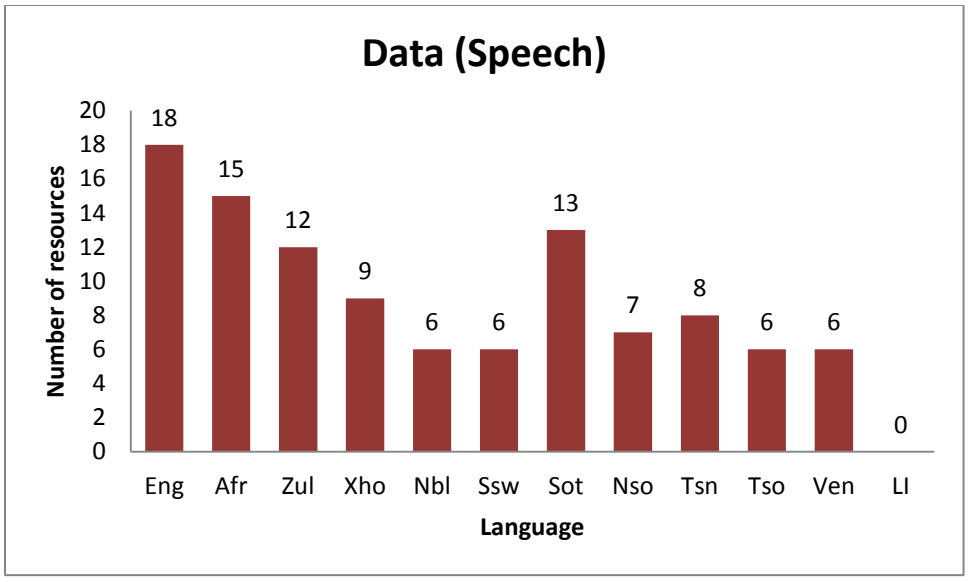


FIGURE 83: SUMMARY OF DATA (SPEECH) RESOURCE TYPES FROM 2009 - 2018

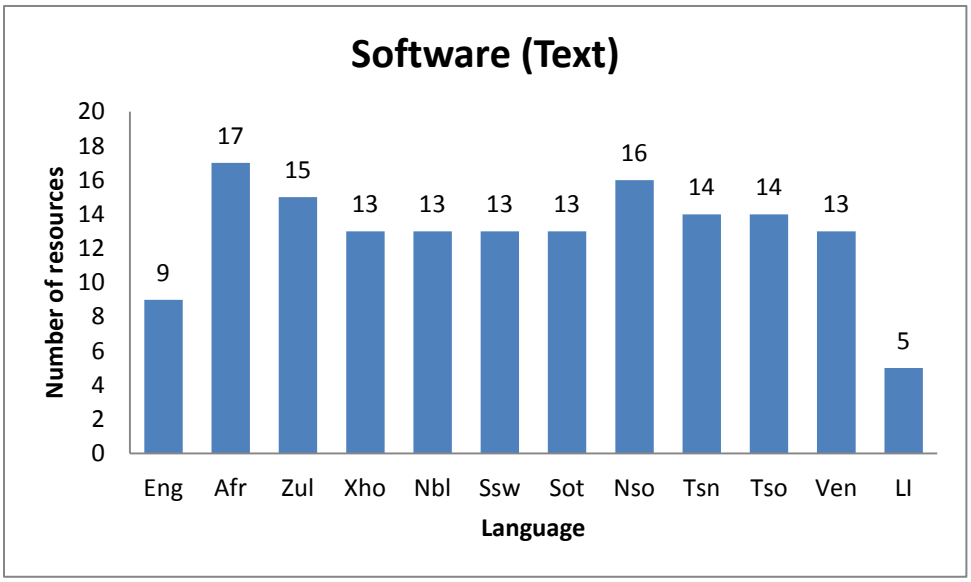


FIGURE 84: SUMMARY OF SOFTWARE (TEXT) RESOURCE TYPES FROM 2009 - 2018

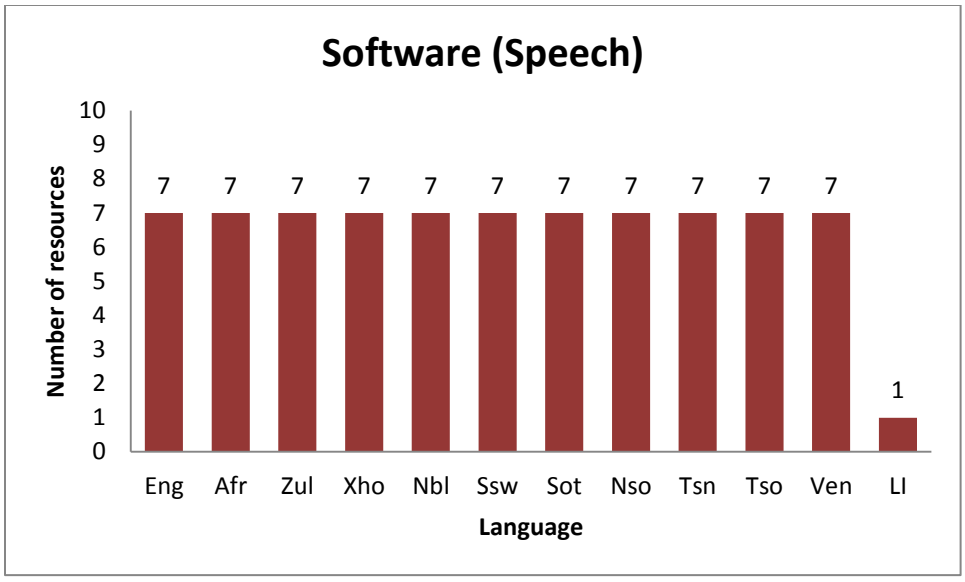


FIGURE 85: SUMMARY OF SOFTWARE (SPEECH) RESOURCE TYPES FROM 2009 - 2018

As shown in Figures 86 and 87 below, there has been an increase of full data resource types in a number of languages from 2009 to 2018.

In the data category, there was a significant increase in text resources for Sesotho, Setswana, Xitsonga and Tshivenda. However, there was a significant increase in the availability of English and Afrikaans speech resources.

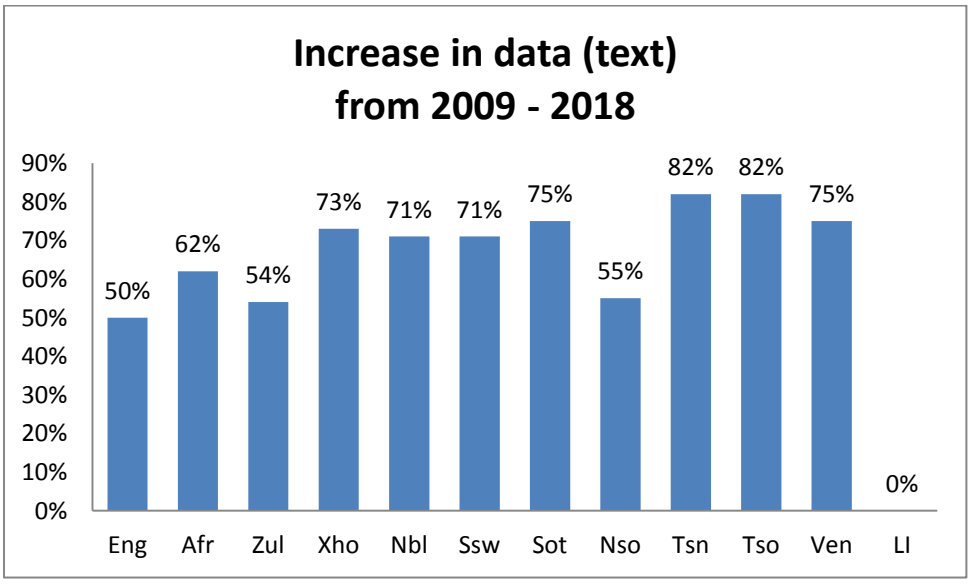


FIGURE 86: INCREASE IN DATA (TEXT) FROM 2009 - 2018

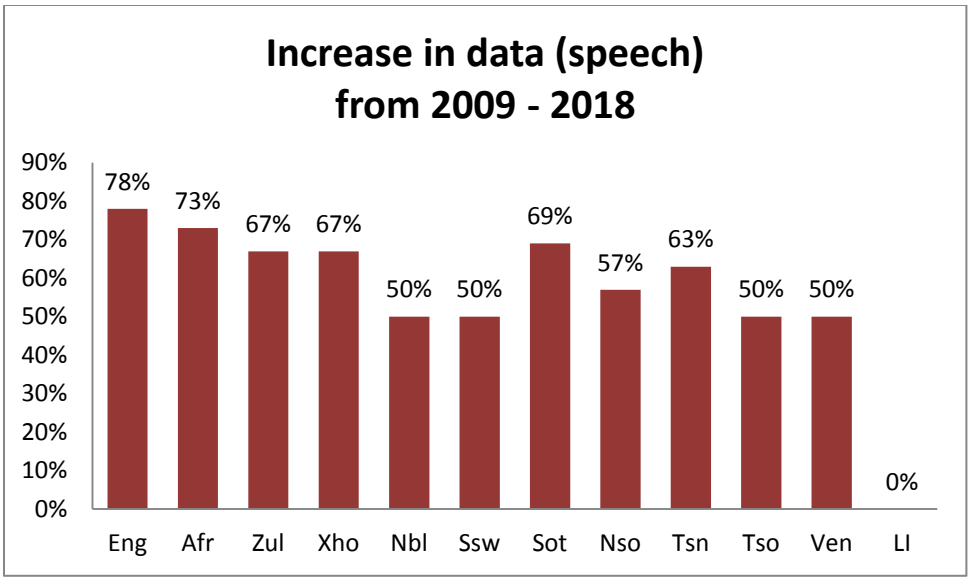


FIGURE 87: INCREASE IN DATA (SPEECH) FROM 2009 - 2018

As shown in Figures 88 and 89 below, there has been an increase of full software resource types in a number of languages from 2009 to 2018.

In the software category, there was a significant increase in text resources for 10 of the 11 South African official languages. However, there was only a significant increase in the availability of language independent speech resources. This specifically refers to the telephony applications in 2009 and 2014.

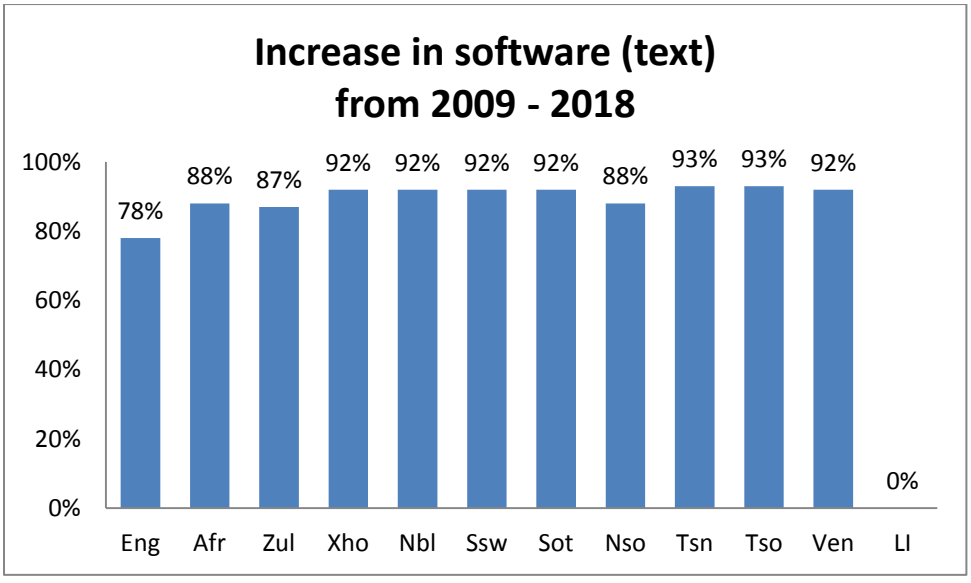


FIGURE 88: INCREASE IN SOFTWARE (TEXT) FROM 2009 – 2018

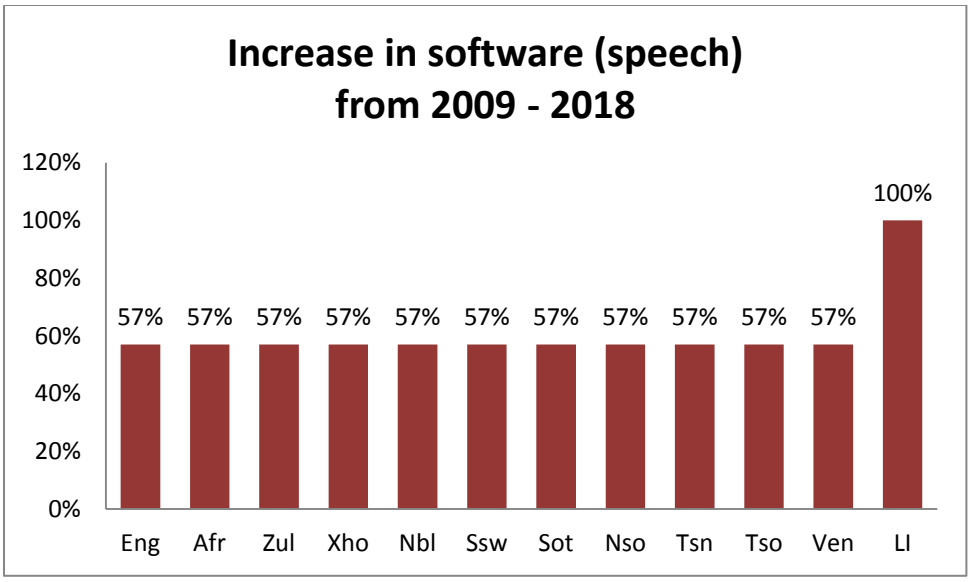


FIGURE 89: INCREASE IN SOFTWARE (SPEECH) FROM 2009 – 2018

8.5 WP 5: RECOMMENDATIONS FOR A DYNAMIC AUDIT UPDATE SYSTEM

8.5.1 INTRODUCTION

In the modern era of big data (collection, fusion and visualisation), distributed networking, and large infrastructure programmes, automating the updating of information on HLT R&D outputs using a distributed platform is imperative. Undertaking technology audits every few years should no longer be the only way to obtain an overview of HLT R&D in the country, as the information should be continuously updated by the developers thereof.

This work package originally aimed to design solution architecture for a system that would enable entities involved in HLT R&D to upload the outputs of their work as these become available. The envisaged outcome was a secure distributed platform which provides for quality control over the uploaded content.

During the execution of the project, it became evident that a separate automated update system would not be required, but that the 2018 Audit online survey could fulfil the requirements. This was confirmed with SADiLaR management and documented in the minutes of that meeting. All that remained to be done in this WP was therefore to provide input on the process to ensure that the outputs of HLT R&D get submitted via the online tool, as these become available. To this end, we familiarised ourselves with the information management architecture of CLARIN and the LRE, and their approaches to ensuring continual submission of language resources to their databases. This report contains an overview of their processes and provides recommendations on how SADiLaR can implement a similar process locally.

In work packages 1 and 2 of this project, we designed and developed an online survey tool to conduct the 2018 Audit, as detailed in the deliverable report entitled “HLT Audit design, instrument development and execution” submitted to SADiLaR on 12 March 2018.

Although we used the same methodology (BLaRK) followed by Sharma Grover in the 2009 Audit, we customised the Audit design and instruments based on the inputs from HLT Experts in the Audit Design Workshop which took place in October 2017. At this workshop, a number of changes to the 2009 Audit tool and process were proposed. One such change was in the categories of the Audit, namely consolidating the previous modules, tools and applications categories into a software category. Another significant change was to the Audit instrument. We agreed that an online Audit tool was required to ensure that resource submission becomes a more dynamic process - feedback had been received that the previous Audit tool (Microsoft Excel spreadsheet) was administratively cumbersome and would not function well to integrate the audit results with the current SADiLaR database. A significant amount of work went into the 2018 Audit design and subsequently the development of the Audit instrument/online survey.

Upon implementing the online survey it became evident that a separate automated update mechanism was no longer required. We discussed this observation with SADiLaR on 7 February 2018 and agreed that the online survey tool would become the automated update mechanism. The Minutes of the meeting on 7 February 2018 are attached as Annexure H.

The following actions remained to be addressed:

1. Transferring the online survey tool, including all administrative rights, to SADiLaR.
2. Integrating the output of the online survey with the SADiLaR resources database.
3. Ensuring that HLT resources are continually submitted to the database via the online survey, as they become available.

8.5.2 TRANSFERRING THE ONLINE SURVEY TOOL TO SADILAR

The 2018 Audit data received via the survey which was hosted on a CSIR server and the actual data analysis graphical representations were transferred to SADiLaR on 29 June 2018. The 2018 Audit data was exported from LimeSurvey into a number of file types. All of the various file types were transferred to SADiLaR.

The process to access the survey will require the following:

- ***Step 1:*** SADiLaR to download the OpenSource LimeSurvey [4] Community Edition tool
- ***Step 2***: The CSIR to export the survey files (questions and structure) and send them to SADiLaR
- ***Step 3:*** SADiLaR to import the 2018 Audit data files to LimeSurvey
- ***Step 4:*** No admin rights need to be transferred as SADiLaR will take ownership of the LimeSurvey tool once it has been downloaded.

8.5.3 *Integrating the output of the 2018 Audit with the SADiLaR resources database*

All data from the 2018 Audit was packaged and uploaded onto SADiLaR's NextCloud repository on 29 June 2018. This included both the raw (unanalysed) data and the three datasets emanating from the analysis of the 2009, 2014 and 2018 data. The data also included all manual categorisation of resource types and all graphs developed as part of the data analysis process. Much of the data is at present captured in Microsoft Excel spreadsheets.

SADiLaR will take responsibility for integrating the 2018 Audit data with its existing database and for representing the recently-added resources in the Catalogue and Index.

8.5.4 *Ensuring continual submission of resources to SADiLaR as these become available*

In order to recommend a process to ensure that HLT resources are continually submitted to SADiLaR as they become available, we perused the approaches followed by ELRA and CLARIN. We determined that ELRA updates the resources on their database by linking registration to relevant conferences with updating the available resources on their system first.

8.5.4.1 Language Resources and Evaluation (LRE) Map

The LRE Map is a database of NLP resources managed by ELRA. The LRE Map collects NLP resources when papers are submitted to NLP conferences. They then clean up the submissions and record them into the LRE database. The purpose and uses of the LRE Map [10], and the process of accessing and uploading resources are discussed below.

The purpose of the LRE Map is to -

- collect information on language resources;
- become a community of users;
- share resources;
- provide feedback on resources; and
- search for resources.

The Map can provide information on -

- most frequent type of resource;
- most represented language;
- the applications for which resources are developed;
- new versus existing resources;
- the distribution of resources; and
- extending resources beyond the current community.

The process to upload resources onto the LRE map requires users to -

- register on the database;
- log-on as users;
- provide basic information about all the resources described in their conference papers, such as -
    - resource name, language type (mono, bilingual etc.), availability, license;
- provide conference/journal information such as -
    - conference/journal name, year and URL
    - conference description; and
- provide more detailed information on resources such as -
    - resource size, status usage and documentation.

All the above-mentioned information is gathered in a global matrix called the LREC Map.

8.5.4.2 Common Language Resources and Technology Infrastructure (CLARIN)

CLARIN implements a VLO (Virtual Language Observatory) as a means of archiving and processing language-related resources in the field of humanities and social sciences. CLARIN also uses the VLO database as a means of sharing resources. The resources included in the CLARIN database are publicly available for scholars and research purposes and not for commercial use. The VLO functions as a browser to explore linguistic resources, tools and services available through CLARIN and to provide an easy interface for access to many resources from a number of domains.

The purpose of CLARIN [11] and the VLO [12], and the processes for accessing and uploading resources are described below.

The purpose of CLARIN is to provide -;

- an accessible online environment for researchers;
- infrastructure for sharing, use and sustaining language data and tools for research purposes;
- access to digital language data (written or/and spoken);
- language data repositories, service and knowledge centres; and
- information and tools on language resources.

Resources may be accessed through the VLO and resources may be uploaded by sending an email with the data to vlo@clarin.eu.

8.5.5 *Recommendations for continually updating the SADiLaR Catalogue and Index*

It is our recommendation that SADiLaR implements a similar process to that followed by ELRA, but extends the approach to include all relevant conferences and journals as SADiLaR does not organise its own conference. This would entail that SADiLaR puts in place agreements with the organisers of the relevant conferences and the editors-in-chief of the relevant journals, and requires authors who submit papers to these conferences and articles to these journals to indicate whether any resources were developed during the execution of the work being reported on in the paper/article. Where this is the case, the authors will be required to submit those resources to the SADiLaR database and to provide proof of same, before their paper/article is approved for publication.

Such an approach would require the following:

- Determining the relevant conferences and journals to which resource-related papers and articles can be submitted.
- Contacting the conference organisers and journal editors-in-chief to obtain buy-in for implementing such a process.
- Developing an interface linking the conference paper/journal article submissions to the SADiLaR online resource capturing tool (online survey).
- Developing a feedback mechanism to confirm successful submission of resources to the SADiLaR database.

While much of the above work lies outside the scope of this project, we have developed a list of relevant international, regional and local conferences, and journals, and below provide details on these to enable SADiLaR to put in place the proposed system.

8.5.5.1 Relevant language resource-related conferences

The table below contains details on conferences which list language and/or technology resource development as one of their areas of interest.

TABLE 18: LIST OF LANGUAGE RESOURCE-RELATED CONFERENCES

Full name of conference	Acronym	Next conference	Web address/contact
International			
Workshop on Spoken Language Technologies for Under-resourced languages	SLTU	August 2018 (every 2nd year)	http://www.mica.edu.vn/sltu/
Workshop on Speech and Language Technology in Education	SLaTE	June 2018 (Annual)	http://slate-conf.org/2018/home
International Conference on Language Resources and Evaluation	LREC	May 2018 (every 2nd year)	http://www.lrec-conf.org/

Full name of conference	Acronym	Next conference	Web address/contact
Digital Humanities Conference	DH	June 2018 (Annual)	https://dh2018.adho.org/en/
Annual Conference of the International Speech Communication Association	Interspeech	September 2018 (Annual)	http://interspeech2018.org/
Workshop on Speech and Language Processing for Assistive Technologies	SLPAT	TBC	http://www.slpat.org/
Annual Meeting of the Association for Computational Linguistics	ACL	July 2018 (Annual)	https://acl2018.org/
International Conference on Acoustics, Speech and Signal Processing	ICASSP	April 2018 (Annual)	https://2018.ieeeicassp.org/
International Conference on Machine Learning	ICML	July 2018 (Annual)	https://icml.cc/
International Conference on Pattern Recognition	ICPR	August 2018 (every 2nd year)	http://www.icpr2018.org/
Spoken Language Technology Conference	SLT	December 2018 (Annual)	https://signalprocessingsociety.org
International Conference on Human-Computer Interaction	Interact	September 2019 (every 2nd year)	https://interact2019.org/
Pattern Recognition Association of South Africa - Robmech International Conference	PRASA-Robmech	2018 (Annual)	Depending on the organiser
Regional			
IST Africa	IST Africa	May 2018 (Annual)	http://www.ist-africa.org/home/
Local			
South African Institute of Computer Scientists and Information Technologists	SAICSIT	2018 (Annual)	http://www.saicsit.org/

8.5.5.2 Relevant language resource-related conferences

The table below contains details on journals which list language and/or technology resource development as one of their areas of interest.

TABLE 19: LIST OF LANGUAGE RESOURCE-RELATED JOURNALS

Full name of journal	Acronym	Topic	Editor-in-chief	Web address/contact
International Journal of Advanced Computer Technology	IJACT	Computer Science	Prof. Harish Bairathiya, Maulana Azad National Institute of Technology	http://www.ijact.org/index.htm
IEEE Transactions on	NA	Speech	Mari Ostendorf,	https://ieeexplore.ieee.org/xpl/R

Full name of journal	Acronym	Topic	Editor-in-chief	Web address/contact
Speech and Audio Processing			University of Washington	ecentIssue.jsp?punumber=89
IEEE/ACM Transactions on Audio, Speech and Language Processing	NA	Speech	Haizhou Li, University of Singapore	https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570655
Speech Communication	NA	Speech	Dr. F. Bimbo, Centre National de la Recherche Scientifique (CNRS)	https://www.journals.elsevier.com/speech-communication/
EURASIP Journal on Audio, Speech, and Music Processing	ASMP	Speech	Emanuël Habets, University of Erlangen-Nuremberg	https://asmp-eurasipjournals.springeropen.com/
Computer Speech and Language	NA	Speech and NLP	Prof. Roger K. Moore, University of Sheffield	https://www.journals.elsevier.com/computer-speech-and-language/
Literator	NA	Speech and NLP	Phil van Schalkwyk Ian Bekker, NWU	https://literator.org.za/index.php/literator/index
Computational Linguistics	NA	NLP	Paola Merlo, University of Geneva	https://www.mitpressjournals.org/loi/coli
South African Journal of African Languages	SALAL	African Languages	Prof. Inge Kosch, UNISA	http://www.nisc.co.za/products/21/journals/south-african-journal-of-african-languages
Southern African Linguistics and Applied Language Studies	SALALS	SA Languages	Prof. Leketi Makalela, UNISA	http://www.nisc.co.za/products/16/journals/southern-african-linguistics-and-applied-language-studies
South African Computer Journal	SACJ	ICT	Philip Machanick	http://sacj.cs.uct.ac.za/
SAIEE Africa Research Journal	ARJ	General	Prof. Bea Lacquet, WITS	https://www.saiee.org.za/DirectoryDisplay/DirectoryCMSPages.aspx?name=Publications&name=Publications#id=8339&dirname=Africa%20Research%20Journal&dirid=337
South African Journal of Science	SAJS	General	John Butler-Adam	https://www.sajs.co.za/
The Journal of the Pattern Recognition Society	NA	ML/PR	Edwin Hancock	https://www.journals.elsevier.com/pattern-recognition/
IEEE Transactions on Pattern Analysis and Machine Intelligence	NA	ML/PR	Sven Dickinson, University of Toronto	https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=34

Full name of journal	Acronym	Topic	Editor-in-chief	Web address/contact
Journal of Machine Learning Research	JMLR	ML/PR	Francis Bach, INRIA David Blei, Columbia University Bernhard Schölkopf, MPI for Intelligent Systems	http://jmlr.org/
IEEE Transactions on Neural Networks	NA	ML/PR	Derong Liu, Chinese Academy of Sciences	https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=72
Data Mining and Knowledge Discovery	NA	ML/PR	Johannes Fürnkranz	https://www.springer.com/computer/database+management+&+information+retrieval/journal/10618

Transferring the 2018 Audit online survey tool, all data collected, and all analyses of the data to SADiLaR concludes the work to be completed in WP5 of this project. The online survey tool will be repurposed by SADiLaR to become a method for continually updating HLT resources on the SADiLaR database as these become available.

A process for ensuring that developers of HLT resources provide information on these as they become available has been proposed. SADiLaR will consider whether it is feasible to implement this proposal, which entails that submission of South African language resources referenced in conference papers and journal articles becomes a prerequisite for publication. A list of relevant conferences and journals has been provided for ease of reference.

9. SUSTAINABILITY

One of the main aims of the project was to ensure that a sustainable environment for continually updating HLT resources as these become available is created.

The initial thinking was to design and develop a distributed platform which could be used to automate the updating of information on HLT R&D outputs. This would enable us to move away from undertaking technology audits every few years as the only way to obtain an overview of HLT R&D in the country. We therefore planned to design solution architecture for a system that would enable entities involved in HLT R&D to upload the outputs of their work as this becomes available. The envisaged outcome was a secure distributed platform which would provide for quality control over the uploaded content. The information management architecture of CLARIN and the RMA would serve as a point of departure for such a system. The actual development of the system was outside the scope of the project, but the project would deliver solution architecture design diagrams, functional specifications, and a recommendation report on how to implement the proposed solution.

In selecting the online tool used for the 2018 Audit (instead of the previously used Excel spreadsheets), the need for an automated update system was simultaneously addressed. The design of the survey tool was planned carefully and in consultation with SADiLaR, so that the output can be incorporated into the existing resources database. The mappings between the previous resource categories and the 2018 resource categories will be provided to SADiLaR to be incorporated into the database and possibly to guide a revamp of the representation of the data on the SADiLaR website. The survey software/system will be carried over to SADiLaR for further administration and will be hosted on the SADiLaR website in future. As such, the online survey tool represents a continuously available system which researchers can use to capture resources as these become available. Many researchers are now familiar with the system, and the system is completely customisable, should changes be required.

The savings effected by reusing the survey tool as the automated update system, were utilised towards covering the additional costs incurred in designing and developing the survey tool (instead of re-using the previous Excel spreadsheets, as had originally been planned).

10. DISSEMINATION OF RESULTS

CONFERENCE PAPERS

The South African Institute of Computer Scientists and Information Technology (SAICSIT) 2018 conference has initially been identified as an appropriate platform to present two proposed papers on the 2018 Audit. The first paper is on the 2018 Audit design, instrument development and execution, and the second paper is based on the 2018 Audit results and analysis. If one or both of the papers are accepted, they will be presented at the SAICSIT 2018 conference in September 2018. The conference papers are attached as Annexures I and J.

If possible, an international conference will be targeted for an additional presentation or as a substitute for one or two of the local conference papers should they not be accepted.

DISSEMINATION WORKSHOP

The CSIR hosted a workshop on **26 July 2018** at which the outcome of the 2018 Audit and the analysis of the results were presented to interested stakeholders. The aim was to ensure that the stakeholder community is aware of the resources which are available for research and other purposes, and to highlight gaps in resource development which still need to be addressed.

The dissemination workshop programme and CSIR presentation is attached as Annexures K and L.

11. LESSONS LEARNT

In the table below, we have captured lessons learnt in the project. These lessons captured are both positive and negative. The recommended solutions can be to either avoid negative lessons learnt or to continue with positive lessons learnt.

TABLE 20: LESSONS LEARNT

Lesson	Lesson detail	Recommended solution	Impact
Duration of the project	We determined after taking into account the timelines and actual work that the duration of the project was not sufficient for a project of this nature. The design and instrument development work packages took a significant amount of time, and there was therefore less time for the executable Audit to take place which also contributed towards a limited number of resources received.	A project of this nature should be implemented over a 2 year period which will allow sufficient time for the technical work to be completed and sufficient time to participate in the Audit.	Limited Audit responses were received.
Consulting experts in the processes	The CSIR researchers who worked on this project were not HLT Experts. We found it useful to consult with HLT Experts in the text and speech areas which greatly contributed to the Audit design and gap analysis.	To involve technical experts in each stage of a project of this nature.	We successfully designed and executed the Audit.
Participation in the Audit	We received limited responses to the Audit. We continuously reminded invited institutions to submit resources; however, this did not make a large impact on obtaining updating resources.	To follow the recommended approach of linking conference paper submissions to resource collection.	A limited number of resources were submitted.
Open Source online survey tool	We managed to identify an Open Source online survey tool to conduct the Audit. The tool was at no cost and very simple to configure and use. We were able to configure the tool successfully and Audit responses were easy to export and analyse.	To proceed with an Open Source option to updating and managing resource.	We were able to successfully develop and execute the Audit at no cost.
Data analytics tool	We used a manual process to analyse the Audit data and conduct the gap analysis. By using this process, we are open to human error when analysis the data.	To obtain a data analytics tool to analyse the data	There may be errors

12. SCIENTIFIC IMPACT

The below-mentioned publications were submitted to SAICSIT 2018 on 29 June 2018. Only the paper entitled "A Human Language Technology Audit: Analysing the development trends and gap in resource availability in all South African languages" was accepted to present at the SAICSIT 2018 conference in September 2018.

TABLE 21: LOCAL CONFERENCE PAPERS

Paper title	Authors	Abstract
Human Language Technology Audit 2018: Design considerations and Methodology	I.Wilken, C.Moors, K.Calteaux, T.Gumede	Technology audits can play a significant role in surfacing information which can be used by researchers, policy-makers and funders alike to build a country's research and development system of innovation towards increasing its competitiveness and contributing to its economy. In 2016, South Africa established a Centre for Digital Language Resources (SADiLaR) with the aim of supporting a large research infrastructure programme tasked with bringing South African language resources into the digital age. This paper discusses the design considerations and methodology employed to undertake one of the first projects funded by SADiLaR: an updated audit of human language technology resources in South Africa. The paper aims to provide sufficient information to replicate such a technology audit in other environments. The design considerations aim to ensure a pleasant user experience, in order to facilitate as much input as possible. The approach aims to ensure that a sustainable audit tool is developed which can be hosted by SADiLaR in future.
A Human Language Technology Audit: Analysing the development trends in resource availability in all South African languages	C.Moors, I.Wilken, K.Calteaux, T.Gumede	Almost a decade has passed since the first audit on the state of HLT in South Africa was published in 2009. An increase in HLT R&D in South Africa since then, as well as developments in language resource management in the country surfaced the need for an updated audit of HLT resources. Consolidating information on the availability and maturity of HLT resources provides valuable information for both researchers and decision-makers. On the one hand, information on available HLT resources enables researchers to identify new opportunities for multidisciplinary research. On the other, decision-makers can use the information to determine

Paper title	Authors	Abstract
		priorities for resource development and where to focus their investments. The paper presents an overview of the main findings of an audit of HLT resources, undertaken in 2017/8, and makes some suggestions for ensuring that the current information is continually updated as new resources are developed.

13. FINANCIAL REPORT

The project expenses (exclusive of VAT) are summarised in the following table:

TABLE 22: FINANCIAL REPORT

	HR	Running	Total
Budget allocated	R1,273,921.00	R117,391.00	R1,391,314.00
Actual spent	R1,353,965.00	R37,584.48	R1,391,549.00
Over/under expenditure			-R235.00

The audited financial report is attached to this report as Annexure M. Please note that the contract was audited and not per project, therefore, this statement includes both the HLT Audit and the Meraka Node Management income and expenditures.

14. CONCLUSION

In summary, the 2017/2018 followed a similar approach to that of Aditi Sharma in her Master's thesis of 2008/2009. Whilst going through her thesis in detail and reading up related research papers, we determined that the approach would still be relevant to the current Audit. However, in the design phase of the Audit, we modified the approach slightly based on various inputs in the field. We also opted for an online tool on which to conduct the Audit which was less manual than the 2008/2009 Audit. However, the analysis and consolidation of the 2017/2018 Audit was a manual task. We analysed and represented the 2017/2018 Audit results in the form of graphs and tables. We also used this approach to analyse and compare the various datasets (2009, 2014 and 2018). We used the above data which was compared to determine the current gaps and development trends in language resources in South Africa. We used the same approach used by Aditi Sharma to conduct the gap analysis.

The data available can also be used for further analysis by interested parties and for decision-making on future resource development trends. We also propose that more awareness should be raised on the availability of HLT resources for R&D, specifically those less represented languages resources (as indicated in the gap analysis). Although significant progress was made in the increase in resources within the 10 years from the previous HLT Audit, there is still a large gap in available resources.

ANNEXURES

Annexure	WP	Description
A	1	HLT Audit design workshop programme
B	1	HLT Component categories
C	1	Questionnaire used in 2009 HLT Audit
D	2	Front-end screen shots of the online Audit tool
E	2	Email to participants to notify them of the Audit
F	2	Automated email to participate in the Audit
G	4	Maturity sum, accessibility sum, HLT component sum
H	5	Minutes of meeting to discuss WP5
I	4	Conference paper: HLT Audit 2018: Design considerations and Methodology
J	4	Conference paper: HLT Audit 2018: Analysing the development trends in resource availability in all South African languages
K	5	Dissemination workshop programme
L	5	CSIR Presentation at dissemination of findings workshop
M	0	Audited financial report

REFERENCES

[1] Cetindamar, D; Phaal, R and Probert, D. (eds) 2010. Technology Management: Activities and Tools. Palgrave Macmillan, Hampshire, United Kingdom.

[2] Aditi Sharma Grover. 2009. Technology Audit: The State of Human Language Technologies R&D in South Africa. Masters' Thesis. University of Pretoria.

[3] Sharma Grover, A; Calteaux, K; Van Huyssteen, G and Pretorius, M. 2010. An overview of HLTs for South African Bantu languages, in Proceedings of the South African Institute for Computer Scientists and Information Technologists (SAICSIT '10), October 11-13, Bela-Bela. South Africa, pp 370-375.

[4] Sharma Grover, A.; Van Huyssteen, G and Pretorius, M. 2011. The South African Human Language Technology Audit, in Language Resources and Evaluation 45(3), September 2011, pp 271-288.

[5] Krauwer, S. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap, in Proceedings of the International Workshop Speech and Computer (SPECOM 2003), October 27-29, Moscow State Linguistic University, Russia, pp 8-15.

[6] Strik, H.; Daelemans, W.; Binnenpoorte, D.; Sturm, J.; de Vriend, F. and Cucchiarini, C. 2002. Dutch HLT resources: from BLARK to priority lists, in Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002), September 16-20, Denver, USA, pp. 1549-1552.

[7] Aditi Sharma Grover, Gerhard B. van Huyssteen, and Marthinus W. Pretorius. 2010. The South African Human Language Technologies Audit. In Proceedings of the 7th Language Resource and Evaluation Conference, Valletta, Malta. 17-23 May 2010, pp 2847-2850.

[8] Steven Krauwer. 2003. "The Basic Language Resource Kit (BLARK) as the first Milestone for the Language Resources Roadmap." In proceedings of ELNET' s workshop on Speech and Computer (SPECOM 2003).

[9] A. Sharma-Grover, G.B. van Huyssteen and MW. Pretorius, An HLT profile of the official South African Languages. In proceedings of the second workshop on African Language Technology (AFLAT 2010)

[10] Language Resources and Evaluation (LRE) Map http://lremap.elra.info/

[11] Common Language Resources and Technology Infrastructure (CLARIN) https://www.clarin.eu/

[12] CLARIN Virtual Language Observatory https://www.clarin.eu/content/virtual-language-observatory-vlo

[13] LimeSurvey https://www.limesurvey.org/

AGENDA

ANNEXURE A – AUDIT DESIGN WORKSHOP PROGRAMME

HLT Audit Design Workshop

Meraka Auditorium, Building 43B

Wednesday, 30 August 2017 and Thursday, 31 August 2017

DAY 1: Wednesday, 30 August 2017

Topic	Time
1. Registration *Arrival coffee, tea and muffins are served*	08:30 – 08:45
2. Welcome and introduction • SADiLaR • HLT Audit 2017/2018	08:45 – 09:15
3. Updating the component categories • Introduction and background • Group discussions (Data) • Feedback discussion and consolidation (Data)	09:15 – 10:30
TEA BREAK	**10:30 – 10:45**
4. Updating the components categories • Group discussions (Modules) • Feedback discussion and consolidation (Modules)	10:45 – 12:00
5. Updating the components categories • Group discussions (Applications) • Feedback discussion and consolidation (Applications)	12:00 – 13:00

LUNCH	**13:00 – 13:45**
6. Identification of institutions • Update draft list and match components and institutions	13:45 – 14:30
7. Prioritising text and speech components • Introduction and background • Group discussion (Applications) • Feedback (Applications)	14:30 – 15:30
TEA BREAK	**15:30 – 15:45**
8. Prioritising text and speech components • Group discussion (Modules and Data) • Feedback (Modules and Data)	15:45 – 16:30
9. Campus ramble	17:00 – 18:00
10. Dinner at Newton's, on the deck	18:00 – 20:00

DAY 2: Thursday, 31 August 2017

Topic	Time
1. Arrival coffee, tea and muffins	08:30 – 09:00
2. Recap of Day 1	09:00 – 09:30
3. Introduction to the questionnaire session	09:30 – 10:00
4. Formulating the audit questionnaire (Session 1)	10:00 – 11:00
TEA BREAK	**11:00 – 11:15**
5. Formulating and finalising the audit questionnaire (Session 2)	11:15 – 13:15
LUNCH	**13:15 – 14:00**

6. Discussion on audit execution	14:00 – 15:00
7. Conclusion and way forward	15:00 – 15:30

ANNEXURE B- UPDATED COMPONENT CATEGORIES

TEXT

DATA		Metadata questions
Monolingual lexicon		Indicate the level of detail added in the annotation?
Multilingual lexicon		Indicate the level of detail added in the annotation?
Terminology list		Indicate the domain the terminology belongs to.
Controlled vocabulary		- To which domain does the terminology belong?- Indicate the level of detail added in the annotation?
Named entities list		Indicate the type(s) of named entities that are annotated. Provide details.
Thesaurus		-
Wordnet		Indicate the type(s) of wordnet. Provide details.
Ontology		Indicate the type(s) of ontology. Provide details.
Monolingual text		- Indicate the type of annotation added to the text, if any. Provide details.- If the text is aligned, what type of alignment is applied?
Multilingual text		- Indicate the type of annotation added to the text, if any. Provide details.- If the text is aligned, what type of alignment is applied?
Treebank		- Indicate the type(s) of treebank. Provide details.- If the treebank is aligned, what type of alignment is applied?
Statistical language model		- Indicate the language modelled.- Indicate the n-gram size that was applied to the model.
Formal grammar		Indicate the formalism that was applied.
Tagset		Indicate the type of tagging this model describes.

SOFTWARE		Metadata questions
Tokeniser		-
Sentenciser		-
Spelling corrector		-
Full-form normaliser		Indicate the type(s) of full-form that are normalised.
Format normaliser		Indicate the type(s) of format that is normalised.
Number normaliser		Indicate the type(s) of number that is normalised.
Diacritics normaliser		Indicate the type(s) of diacritics that are normalised.
Anonymiser		Indicate the type(s) of terms that are anonymised.
Lemmatiser		-
Stemmer		-
Morphological analyser		Indicate the depth of analysis that was applied.
Morphological synthesiser		Indicate the type(s) of synthesis that was applied.
Part-of-speech tagger/disambiguator		-
Syllabifier		-
Hyphenator		-
Deep parser	Dependency parser	-
Deep parser	Constituent recogniser	-
Shallow parser	Chunker	-
Semantic analyser	Event extractor	Indicate the type(s) of event(s) that are extracted.
Semantic analyser	Frame extractor	Indicate the type(s) of frame(s) that are extracted.
Semantic analyser	Named entity recogniser	Indicate the type(s) of entities that are extracted.
Semantic analyser	Terminology extractor	-
Semantic analyser	Topic modelling	-
Sentiment analysis/affect/emotion analyser		Indicate the level of detail added to the emotion/analysis.
Referent resolver		-
Word meaning disambiguator		-
Pragmatic analyser		-
Text generator		-
Summariser		-
Machine translator		- Indicate the target language.- Indicate the source language.
Language and dialect identifier		Indicate the level of detail added to the annotation/analysis.
Proofing/authoring tool		-
Information retrieval system		-
Information extractor		-
Human-aided machine translation system		-
Machine-aided human translation system		-
OCR/ICR		-
Computer aided language learning (CALL) system		-
Document classifier		-
Authorship identifier		-
Question answering (QA) system		-
Dialogue system (text-based)		-
Comprehension assistant		-

SPEECH

DATA	Metadata questions
TTS sentence splitting rule sets (manually created)	- Specify the notation/rule format. - Specify any other markup.
TTS tokenisation rule sets (manually created)	- Specify the notation/rule format. - Specify any other markup.
TTS normalisation rule sets (manually created)	- Specify the notation/rule format. - Specify any other markup.
TTS LID rule sets (manually created)	- Specify the notation/rule format. - Specify any other markup. - List any additional languages applicable to the resource that were not selected in the "Required Information" section.
Language grammar	- Specify the notation/rule format. - Specify any other markup.
Phoneme sets	- Indicate the phonetic transcription used. - List any phone features present.
Phone mappings	-Indicate the source and destination phone sets.
Pronunciation dictionaries	- Indicate the phone set used. - Indicate any lexical features present.
Pronunciation rule sets (manually created)	- Indicate the phone set used. - Specify the notation/rule format. - Specify any other markup.
Intonation rule sets (manually created)	- Indicate the phone set used, if applicable. - Specify the notation/rule format. - Specify any other markup.
Phrasing rule sets	- Specify the notation/rule format. - Specify any other markup.
Tone rule sets	- Indicate the phone set used, if applicable. - Specify the notation/rule format. - Specify any other markup.
Stress rule sets	- Indicate the phone set used, if applicable. - Specify the notation/rule format. - Specify any other markup.
Syllabification rule sets	- Indicate the phone set used, if applicable. - Specify the notation/rule format. - Specify any other markup.

Speech Corpora	- Is the corpus monolingual or multilingual? - Is the corpus parallel? If yes, indicate the language pairs. - Is the corpus unannotated (raw) or annotated? If the corpus is annotated, provide: a) the URL/link to the protocol (or upload it at the end of this section) and b) the definition of the tag set - Is the corpus unaligned or aligned? If the corpus is aligned, provide: a) the URL/link to the protocol (or upload it at the end of this section) and b) the granularity of alignment - Does the corpus contain human or synthetic speech? - Does the corpus contain code-switched speech? If the corpus contains code-switched speech, list any applicable languages that were not selected in the "Required Information" section. - Provide more information on the production style. - Provide more information on the production environment. - Was a professional voice artist used? - Indicate the number of speakers. - Indicate the degree of accentedness. - Does the corpus contain normal or pathological speech? If the corpus contains pathological speech, provide more detail.
Text corpora for speech	- Is the corpus monolingual or multilingual? - Is the corpus unannotated (raw) or annotated? If the corpus is annotated, provide: a) the URL/link to the protocol (or upload it at the end of this section) and b) the definition of the tag set - If transcriptions are available, name and/or provide a link to the text corpus. - Indicate the domain(s) of the text. - Does the corpus contain code-switched speech? If the corpus contains code-switched speech, list any applicable languages that were not selected in the "Required Information" section.
Intonation models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Indicate the domain(s) of the text. - Indicate the phone set used, if applicable.
Phrasing models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Describe the domain.
Tone models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Indicate the domain(s) of the text. - Indicate the phone set used, if applicable.
Stress models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Indicate the domain(s) of the text. - Indicate the phone set used, if applicable.
Syllabification models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Indicate the domain(s) of the text. - Indicate the phone set used, if applicable.

MODEL	Metadata questions
TTS Sentence Splitting models	- Indicate the machine learning method used. - Provide a link to the training corpus.
TTS Tokenisation models	- Indicate the machine learning method used. - Provide a link to the training corpus.
TTS Normalisation models	- Indicate the machine learning method used. - Provide a link to the training corpus.
TTS LID models	- Indicate the machine learning method used. - Provide a link to the training corpus. - List any additional languages applicable to the resource that were not selected in the "Required Information" section.
Language models	- Indicate the machine learning method used. - Provide a link to the training corpus.
Acoustic models	- Indicate the machine learning method used. - Provide a link to the training corpus. - Provide the link to the pronunciation dictionary.
G2P models	- Indicate the machine learning method used. - Provide a link to the training corpus.

SOFTWARE	Metadata questions
Language modelling tool	- Indicate the machine learning method used.
Pronunciation dictionary creation tool	- Indicate the machine learning method used.
G2P tool	- Indicate the machine learning method used.
Acoustic modelling tool	- Indicate the machine learning method used.
Intonation tool	- Indicate the machine learning method used.
Phrasing tool	- Indicate the machine learning method used.
Tone tool	- Indicate the machine learning method used.
Stress tool	- Indicate the machine learning method used.
Syllabification tools	- Indicate the machine learning method used.
Vocoder	- Indicate the method used.
TTS Sentence Splitting tool	- Indicate the machine learning method used.
TTS Tokenisation tool	- Indicate the machine learning method used.
TTS Normalisation tool	- Indicate the machine learning method used.
TTS LID tool	- Indicate the machine learning method used.
Large vocabulary speech recognition system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Command and control system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Non-native speech recognition system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - Provide the L1 language of L2 users, if known.
Code switched speech recognition system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - List any additional languages applicable to the resource that were not selected in the "Required Information" section.
Multilingual speech recognition system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - List any additional languages applicable to the resource that were not selected in the "Required Information" section.
Noise robust speech recognition system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Alignment system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Automatic phonetic transcription system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Confidence measures	- Provide a link to the training corpus, if any.
Acoustic Language ID	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - List any additional languages applicable to the resource that were not selected in the "Required Information" section.
Acoustic Age ID	- Provide a link to the corpus. - Provide links to the ASR system(s) and/or model(s) used. - Indicate the method used.

Acoustic Gender ID	- Provide a link to the corpus. - Provide links to the ASR system(s) and/or model(s) used. - Indicate the method used.
Acoustic Dialect ID	- Provide a link to the corpus. - Provide links to the ASR system(s) and/or model(s) used. - Indicate the method used.
Acoustic Emotion ID	- Provide a link to the corpus. - Provide links to the ASR system(s) and/or model(s) used. - Indicate the method used.
Key-word spotting system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Voice activity detection	- Provide a link to the corpus. - Provide links to the ASR system(s) and/or model(s) used. - Indicate the method used.
Speaker tracking	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - Indicate the method used.
Acoustic speaker ID	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - Indicate the method used.
Speaker verification system	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - Indicate the method used.
Diarisation	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system. - Indicate the method used.
Complete TTS System	- Provide a link to the training corpus. - Provide links to other basic components that constitute the system.
Speech to speech translation system	- Indicate the source language. - Indicate the target language.
Other	-

MULTIMODAL	Metadata questions
Multimodal corpora	- List the modalities in the corpus. - Is the corpus unannotated (raw) or annotated? If the corpus is annotated, provide: a) the URL/link to the protocol (or upload it at the end of this section) and b) the definition of the tag set - Is the corpus aligned or unaligned? If the corpus is aligned, provide: a) the URL/link to the protocol (or upload it at the end of this section) and b) the granularity of alignment

ANNEXURE C - QUESTIONNAIRES

DATA

AUDIT DIMENSIONS	DATA	English
Technical Description	**Please choose an option from the drop-down menu:**	If "ANNOTATED", give details on type of annotation (e.g. TEXT: Words + POS, Words + Concept, Sentence + Constituent Structure, Sentence + Semantic Annotation SPEECH: Orthographic transcription, phonetic annotation, prosodic annotation, other annotations (speaker tags, emotion tags))
	Name and Version number of Data set:	
	Name(s) of principal developer(s)	
	Contact person(s) and Email(s):	
	Affiliation(s):	
	Description/Background/Purpose (1-2 sentences):	
	Source: (e.g. Books, studio recordings, radio, media agency, newspaper, web, institution, etc)	
	Stratum (structure of data): (e.g. 50% fiction, 25% non-fiction, or 20 male, 20 female speakers, or 30 children, 10 adults)	
	Data specification: (For TEXT: state a) Format (tab de-limited, comma separated) b) Encoding (UTF8, ASCII, etc) c) File format (txt, xml, xml,doc) (For SPEECH: state a) Sampling frequency (24 KHz) b) Bit rate (8 samples/second c) Recording channel (broadband, mono/stereo ,microphone, cell phone d) Format (wav, mp3))	
	Size: TEXT: Number of tokens SPEECH: Duration	
	Size: File	
	Specialised software required?	
Availability	Accessibility	Please tick all that apply: If any further details are available provide here:
	Maturity stage	Please tick ONE option ONLY:
	Distribution (e.g. CD-ROM, a website (where it is available for download), etc)	
	Licensing (e.g. License info, any intellectual property rights or legal info)	
	Cost: Price (for each type of use): - Academic - Research - Commercial If the pricing varies for 'Single vs. Multiple users', please indicate so.	Price (for each type of use): - Academic: - Research: - Commercial:
Quality	Verification and Proof of quality	Has the data set been through any verification or quality measures/procedures ? IF 'Yes' please provide details on the verification/quality measure/procedures used:
	Compatibility with standards	Is the item built based on… ? : (Please tick one option only) Please provide further details on the common standard/guidelines used:
Documentation	Details of documentation available relating to the data set e.g. publications, reports, user manuals, a brief description on usage of the data set, patents, website, API, etc. If available, please provide any Technical report (history, specifications, background, compatibility)	
Reusability/ adaptability/ extendibility	Compatibility with standard format types, standard packages or tools/platforms ?	

TOOLS/PLATFORMS

AUDIT DIMENSIONS	TOOLS/PLATFORMS	INFO
Technical Description	**Is the item a tool/platform ?** **Please tick ONE option ONLY:**	Tools and platforms are usually language independent. If the tool/platform is linked to any languages please indicate so.
	Name and Version number of Tool/Platform:	
	Name(s) of principal developer(s):	
	Contact person(s):	
	Affiliation(s):	
	Description/Background/Purpose (1-2 sentences):	
	Programming language (e.g. C++)/Format: (e.g. xfst, HTK, C5.0 generated tree)	
	Operating system: (e.g. Linux, Windows 98/2000/NT/XP, MAC)	
	Execution location: (Local or server based)	
	Input specification (e.g. language, format of input)	
	Output specification (e.g. language, format of output)	
	Size: File	
	Specialised software required?	
Availability	Accessibility	Please tick all that apply: If any further details are available provide here:
	Maturity stage	Please tick ONE option ONLY:
	Distribution (e.g. CD-ROM, a website (where it is available for download), etc)	
	Licensing (License info, any intellectual property rights and legal info)	
	Cost: Price (for each type of use): - Academic - Research - Commercial If the pricing varies for 'Single vs. Multiple users', please indicate so.	Price (for each type of use): - Academic: - Research: - Commercial:
Quality	Verification and Proof of quality (e.g.. Quality of the module can be measured by a) Accuracy: Could be measured in terms of number of correct answers, recall (how many of the desired outputs the module correctly produced), and precision (how many of the outputs of the module were correct), or b) Efficiency: The time and memory required for normal use of the module must be acceptable.	Please provide details on the accuracy and/or efficiency of the module as well as a brief description of testing suites or conditions or any other quality checks used:
	Compatibility with standards	Is the item built based on… ? : (Please tick one option only) Please provide further details on the common standard/guidelines used:
Documentation	Details of documentation available relating to the tool/platform e.g. publications, reports, user manuals, a brief description on usage of the data set, patents, website, API, etc. If available, please provide any Technical report (history, specifications, background, compatibility)	
Reusability/ adaptability/ extendibility	Compatibility with standard packages or platforms e.g. MATLAB, Praat, HTK, etc ?	
	Relevance to other tasks and applications ? (e.g. can the tool/platform be easily used/extended to accomplish other tasks ?)	
	Open-source (source code and binaries are available)	Is the module open-source ? Please tick ONE option ONLY:

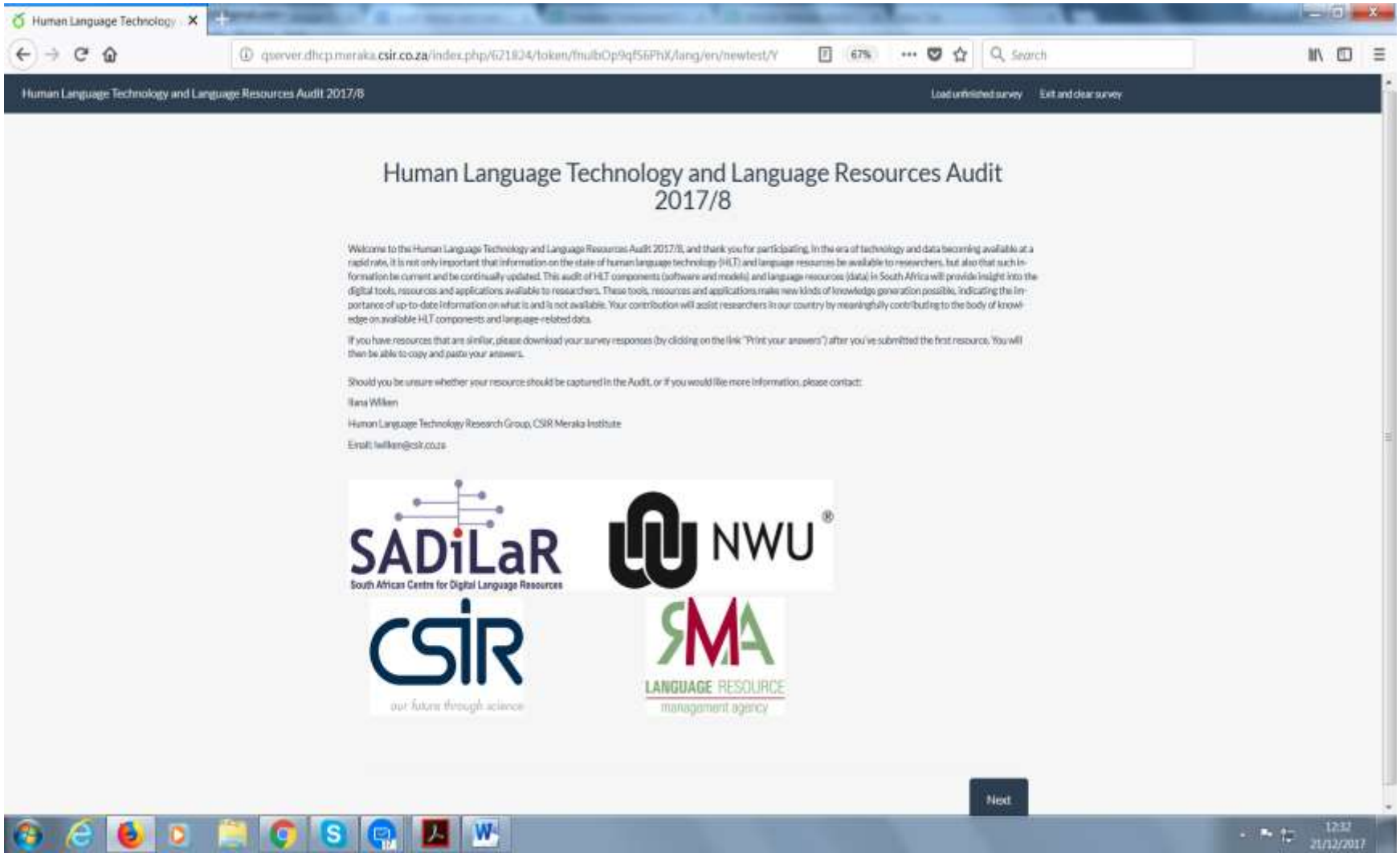
MODULES

AUDIT DIMENSIONS	MODULES	English
Technical Description	**Please choose an option from ONE of the drop-down menus:** **Text-based modules:** **Speech-based modules:**	If a module is language-independent then please state so.
	Name and Version number of Module:	
	Name(s) of principal developer(s)	
	Contact person(s) and Email(s):	
	Affiliation(s):	
	Description/Background/Purpose (1-2 sentences):	
	Programming language (e.g. C++)/Format: (e.g. xfst, HTK, C5.0 generated tree)	
	Operating system: (e.g. Linux, Windows 98/2000/NT/XP, MAC)	
	Execution location: (Local or server based)	
	Input specification (e.g. language, format of input)	
	Output specification (e.g. language, format of output)	
	Size: File	
	Specialised software required?	
Availability	Accessibility	Please tick all that apply: If any further details are available provide here:
	Maturity stage	Please tick ONE option ONLY:
	Distribution (e.g. CD-ROM, a website (where it is available for download), etc)	
	Licensing (License info, any intellectual property rights and legal info)	
	Cost: Price (for each type of use): - Academic - Research - Commercial If the pricing varies for 'Single vs. Multiple users', please indicate so.	Price (for each type of use): - Academic: - Research: - Commercial:
Quality	Verification and Proof of quality (e.g.. Quality of the module can be measured by a) Accuracy: Could be measured in terms of number of correct answers, recall (how many of the desired outputs the module correctly produced), and precision (how many of the outputs of the module were correct), or b) Efficiency: The time and memory required for normal use of the module must be acceptable.	Please provide details on the accuracy and/or efficiency of the module as well as a brief description of testing suites or conditions or any other quality checks used:
	Compatibility with standards	Is the item built based on… ? : (Please tick one option only) Please provide further details on the common standard/guidelines used:
Documentation	Details of documentation available relating to the module e.g. publications, reports, user manuals, a brief description on usage of the data set, patents, website, API, etc. If available, please provide any Technical report (history, specifications, background, compatibility)	
	Compatibility with standard packages or platforms e.g. MATLAB, Praat, HTK, etc ?	

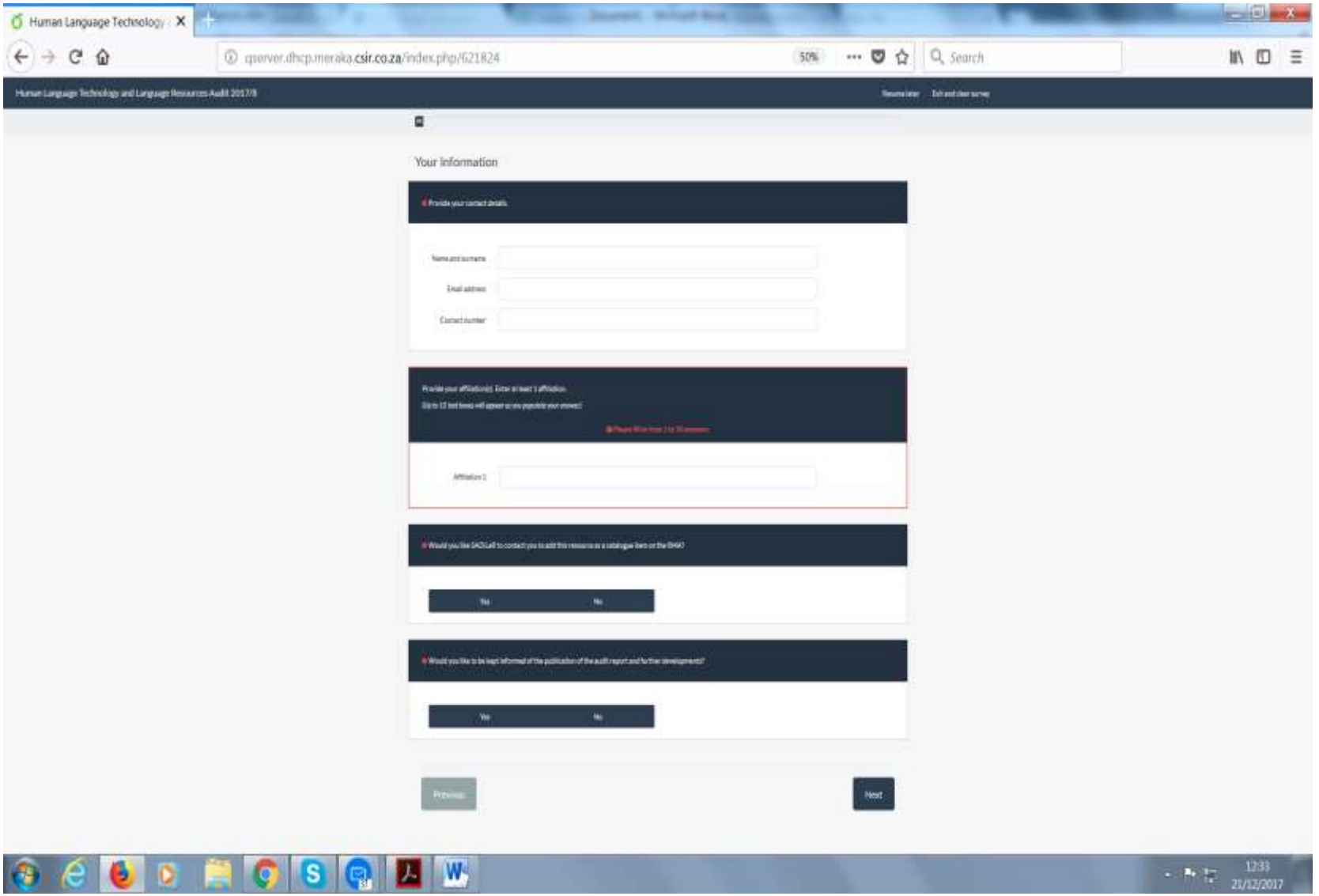
Reusability/ adaptability/ extendibility	Relevance to other tasks and applications ? (e.g. can the module be easily extended to accomplish other tasks ?)	
	Open-source (source code and binaries are available)	Is the module open-source ? Please tick ONE option ONLY:

ANNEXURE D – FRONT END SURVEY SCREENSHOTS

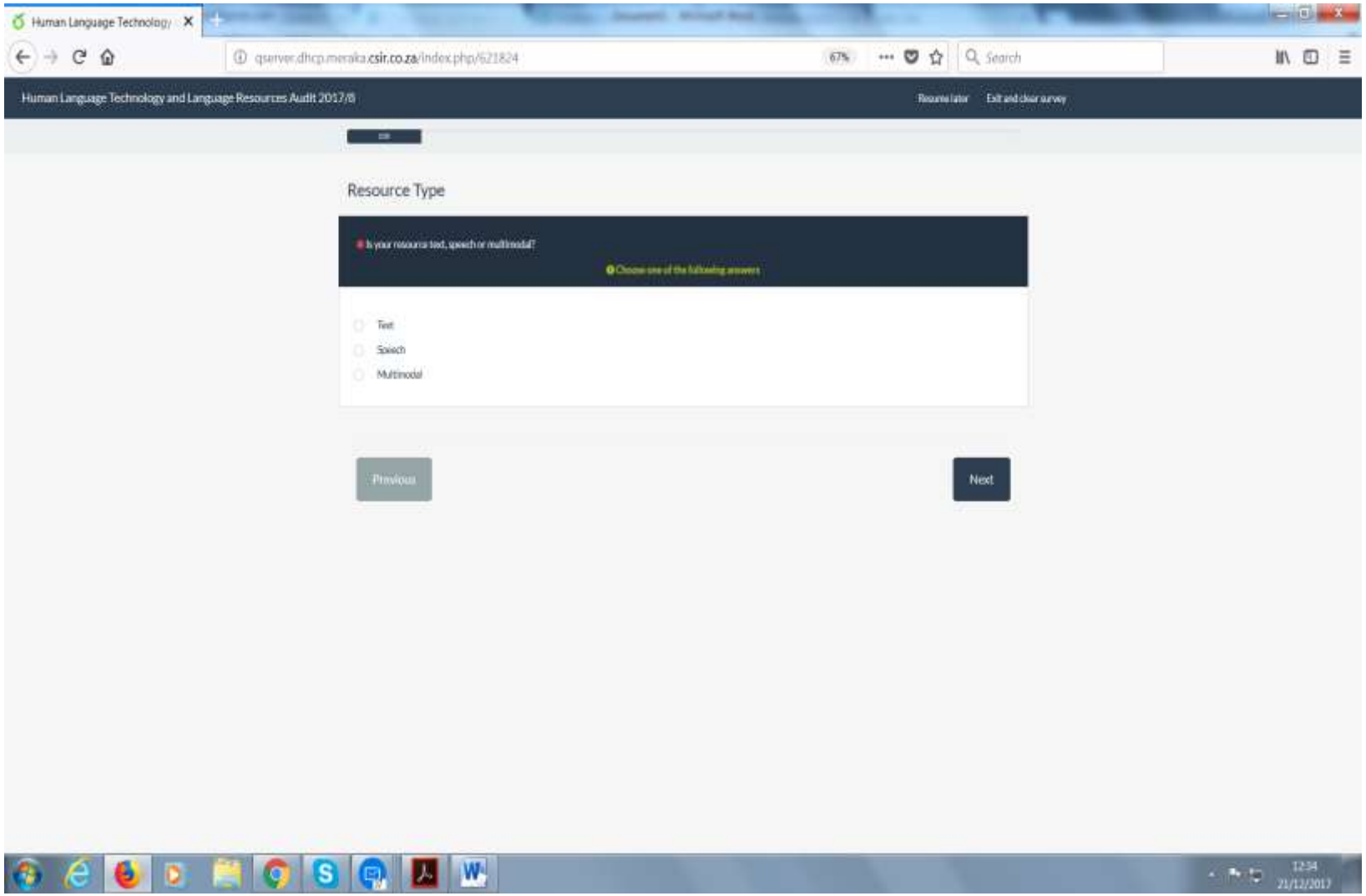
Landing page



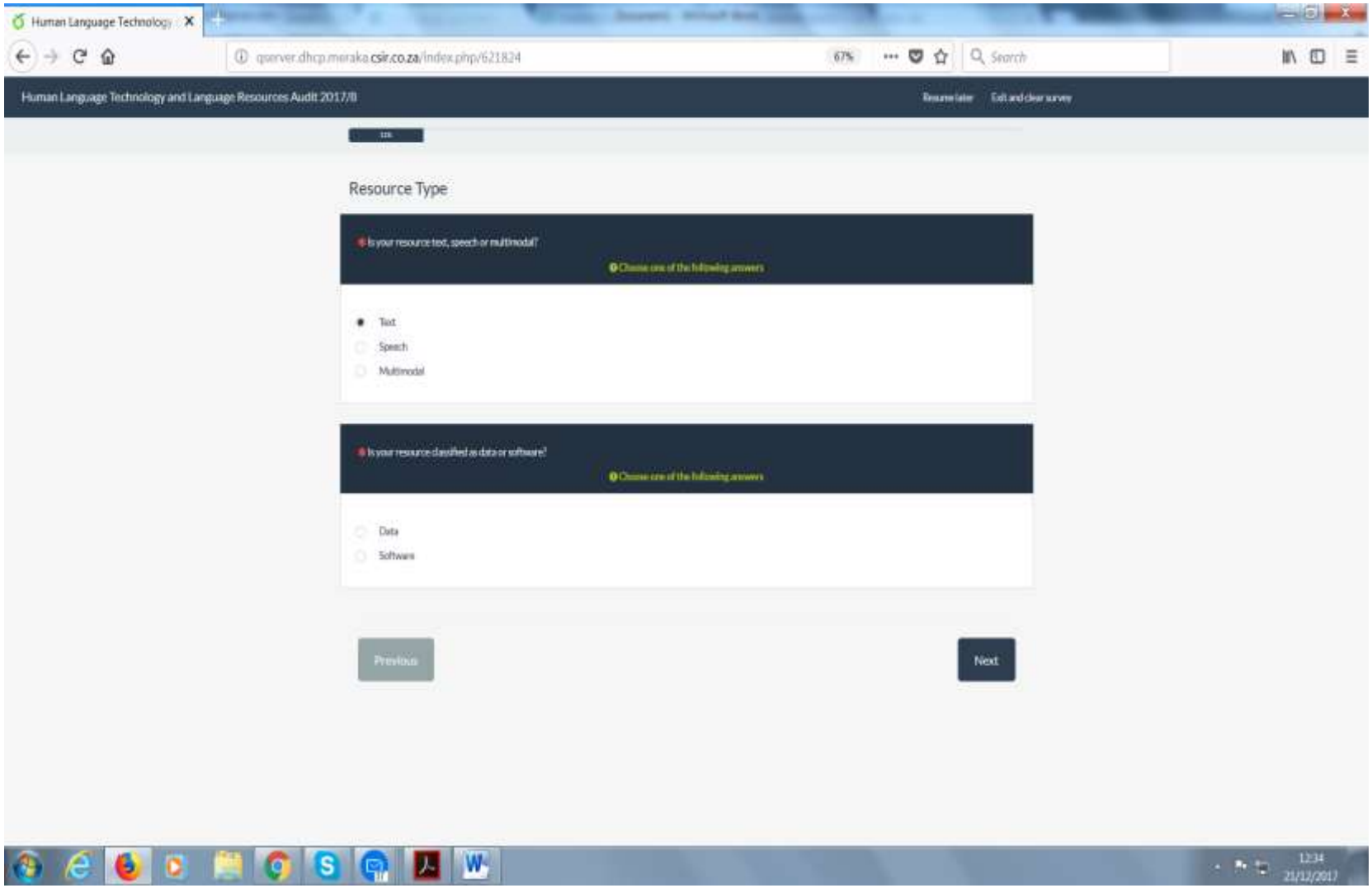
Your information page



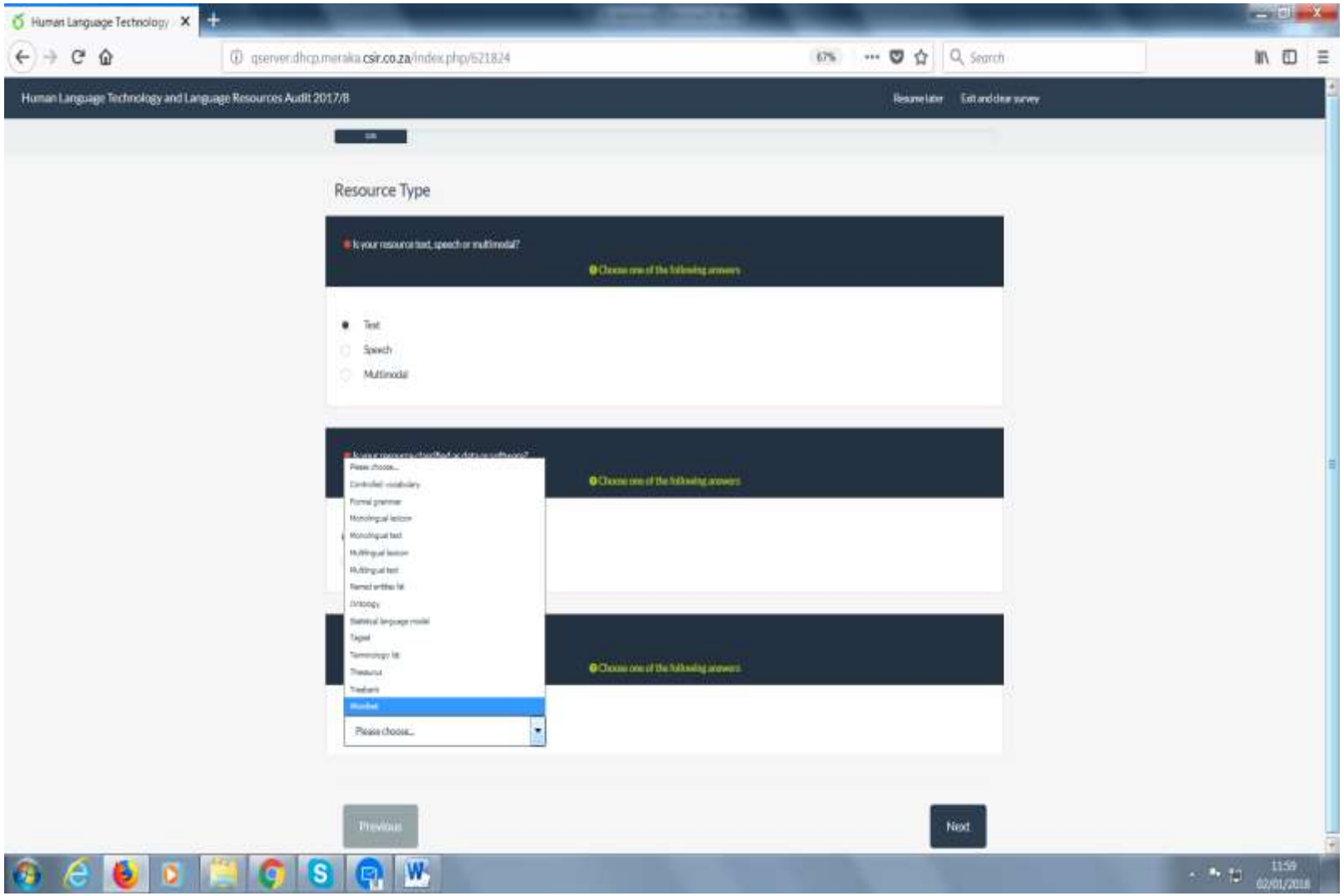
Resource type page



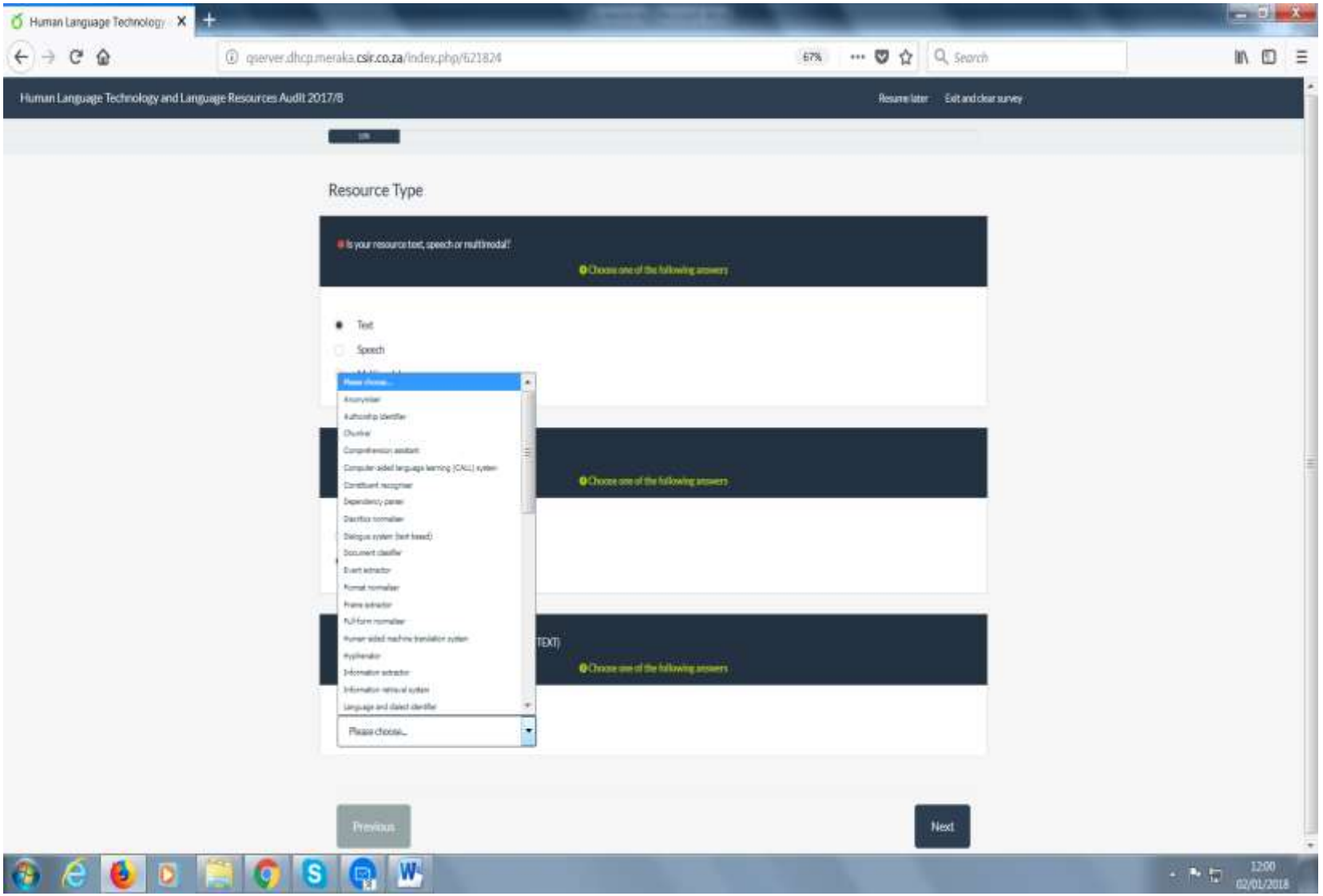
Resource type – text page



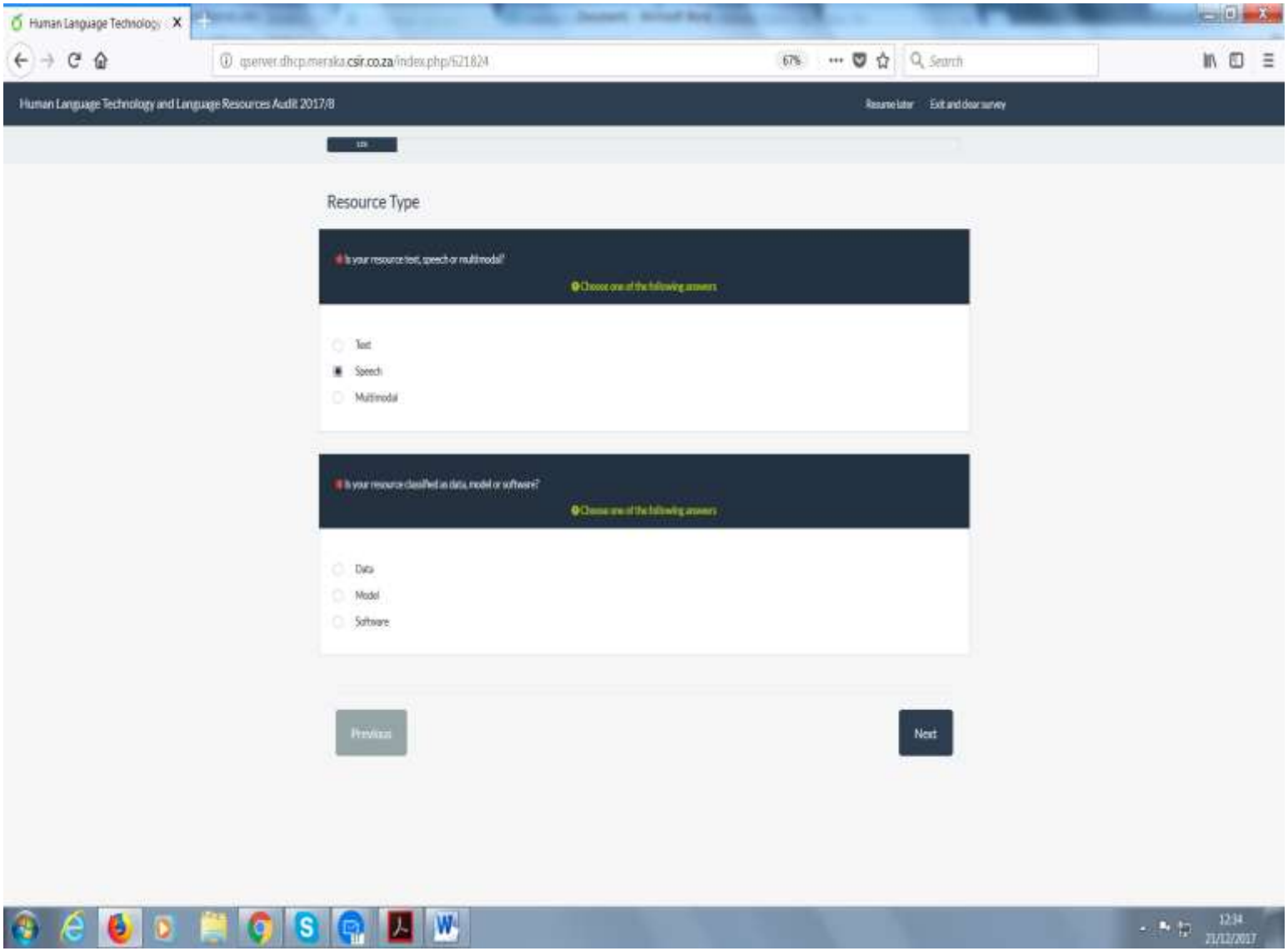
Resource type – text – data – selection page



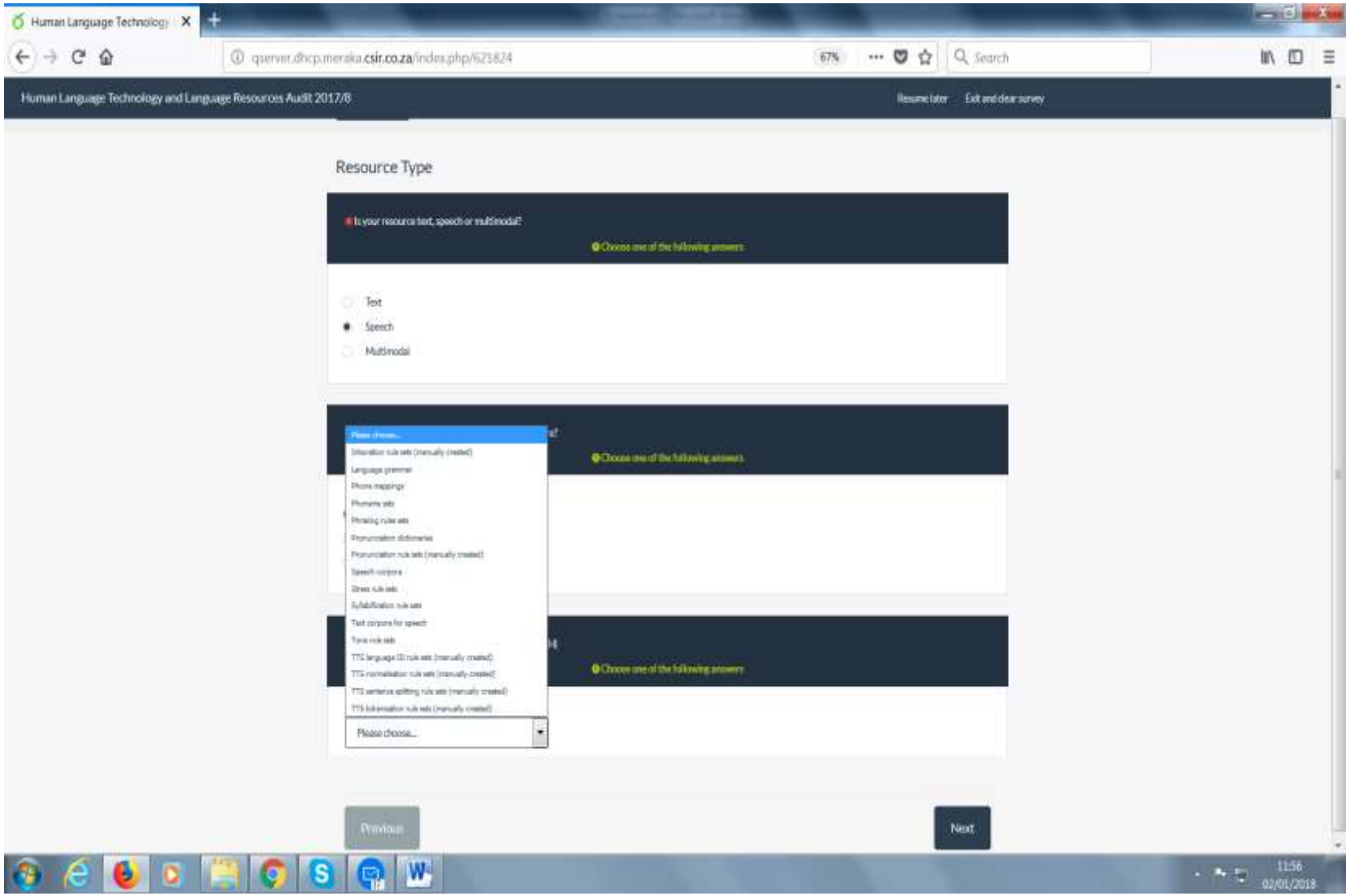
Resource type – text – software – selection page



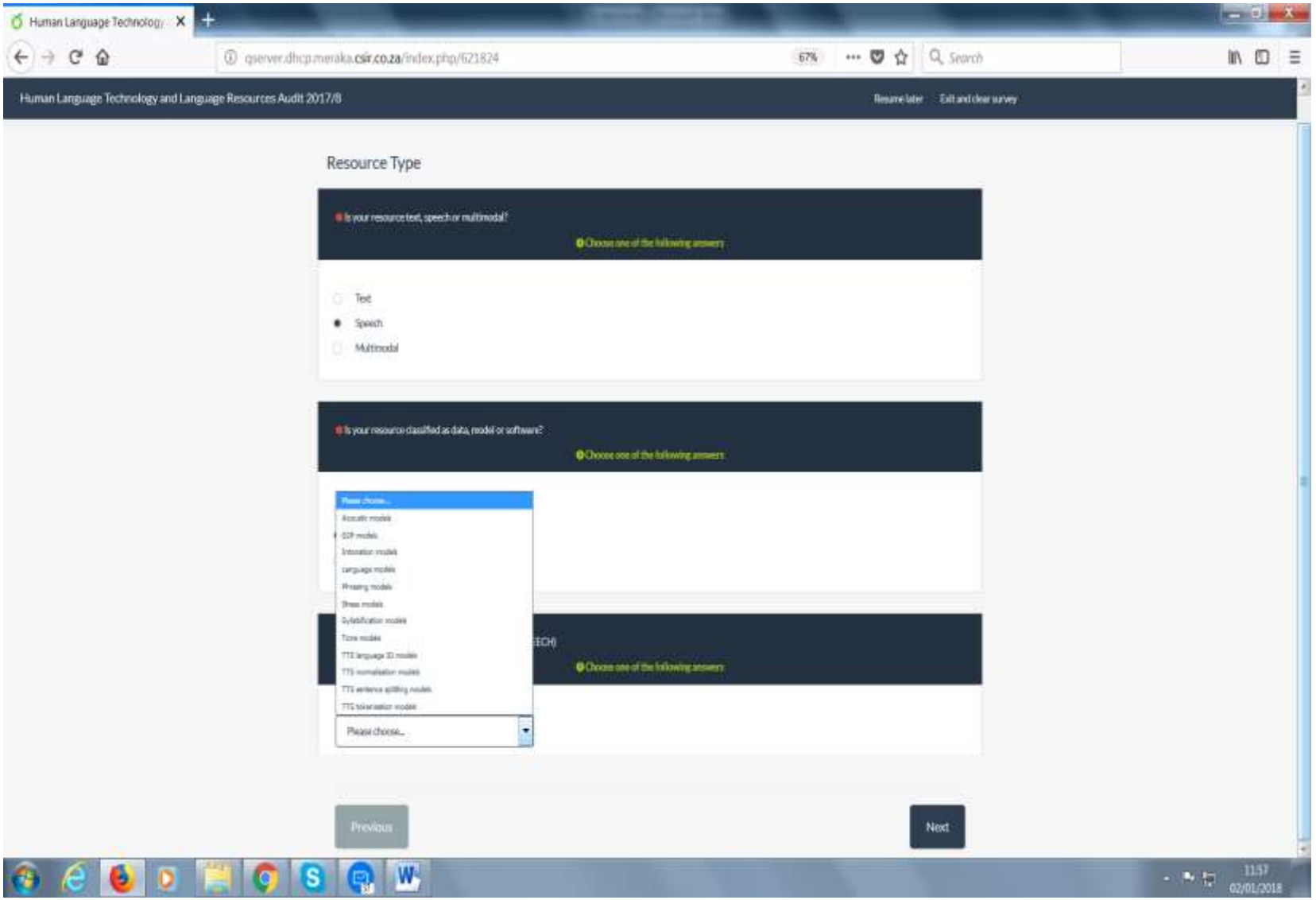
Resource type – speech page



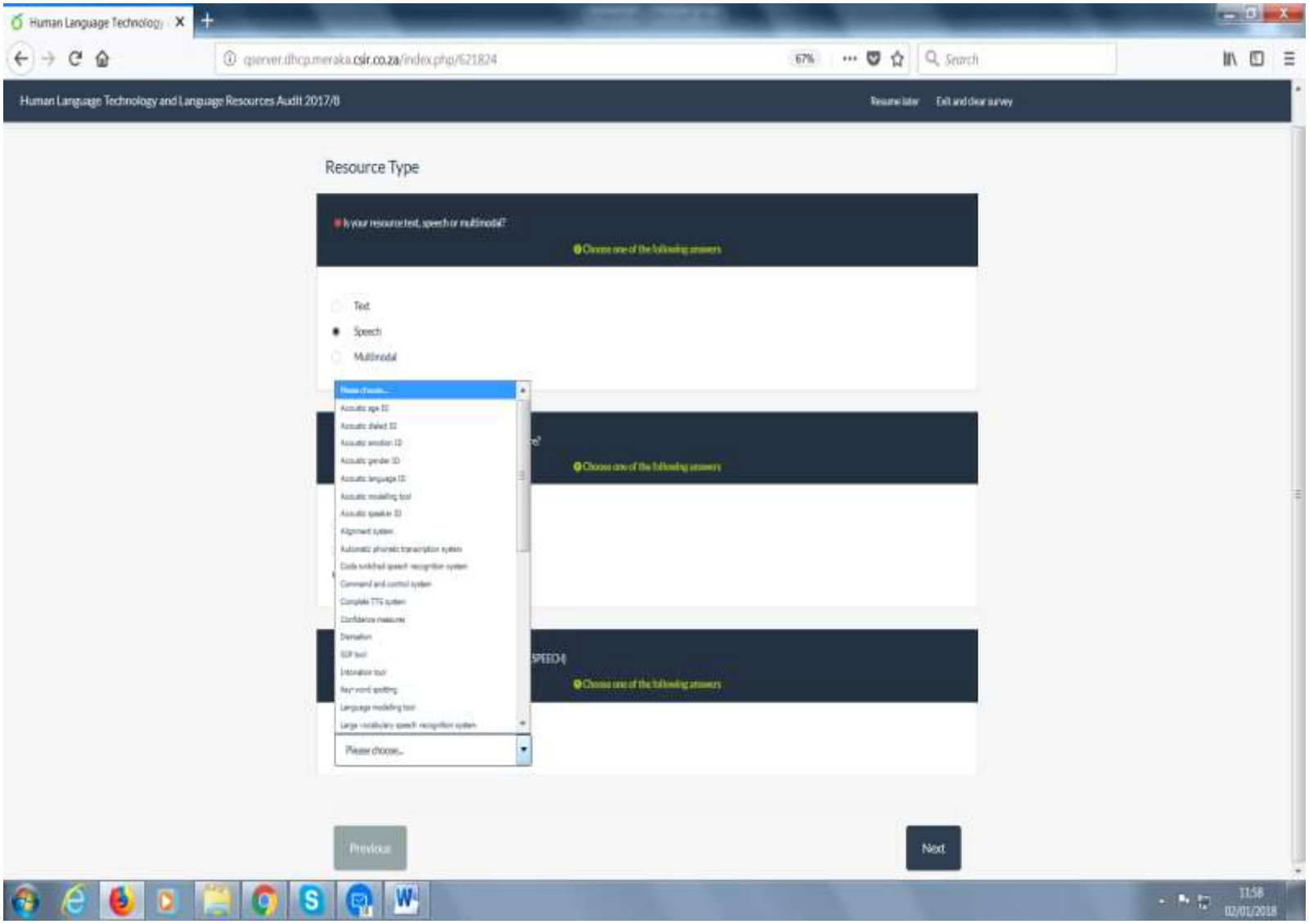
Resource type – speech – data – selection page



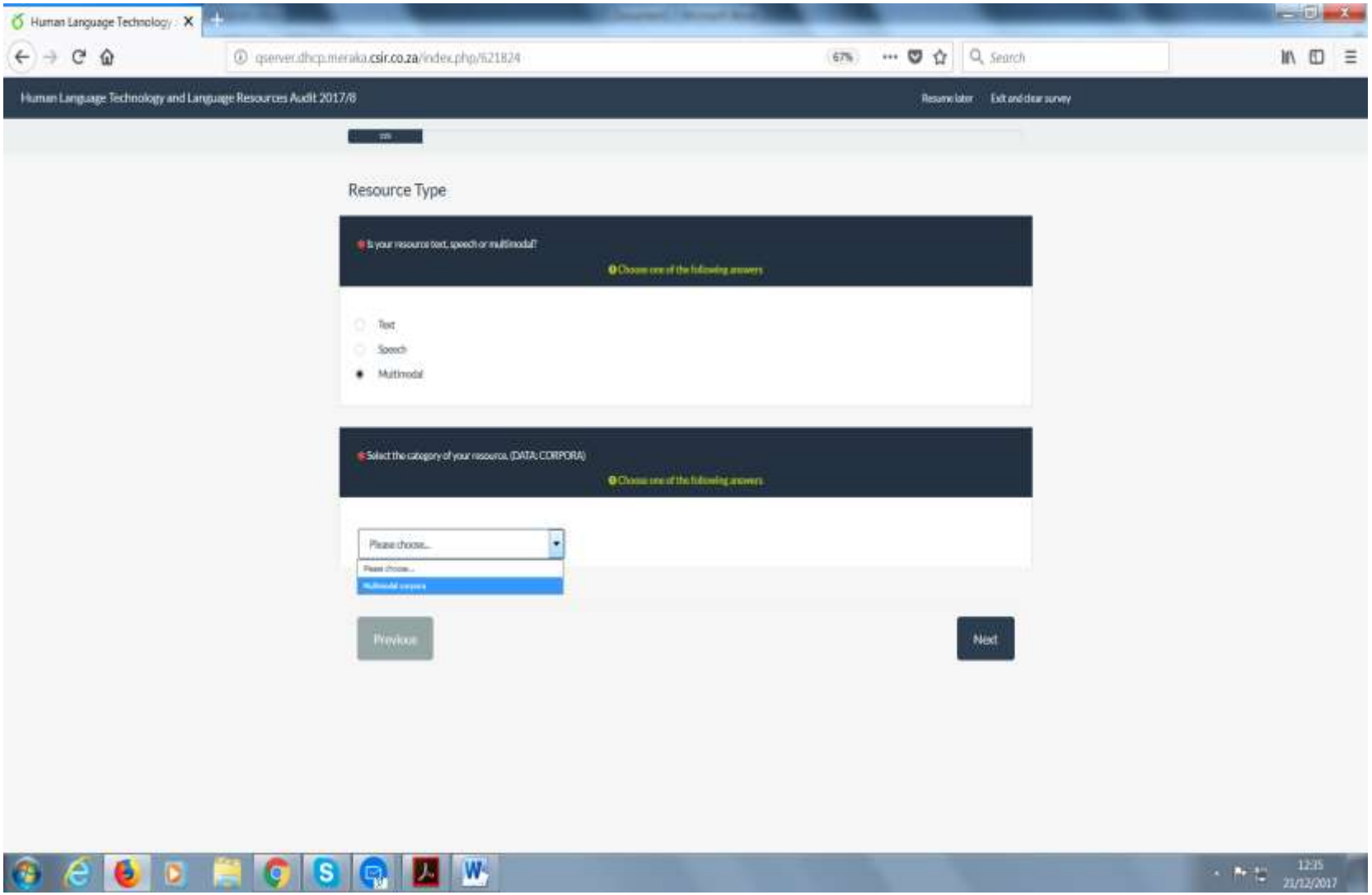
Resource type – speech – model – selection page



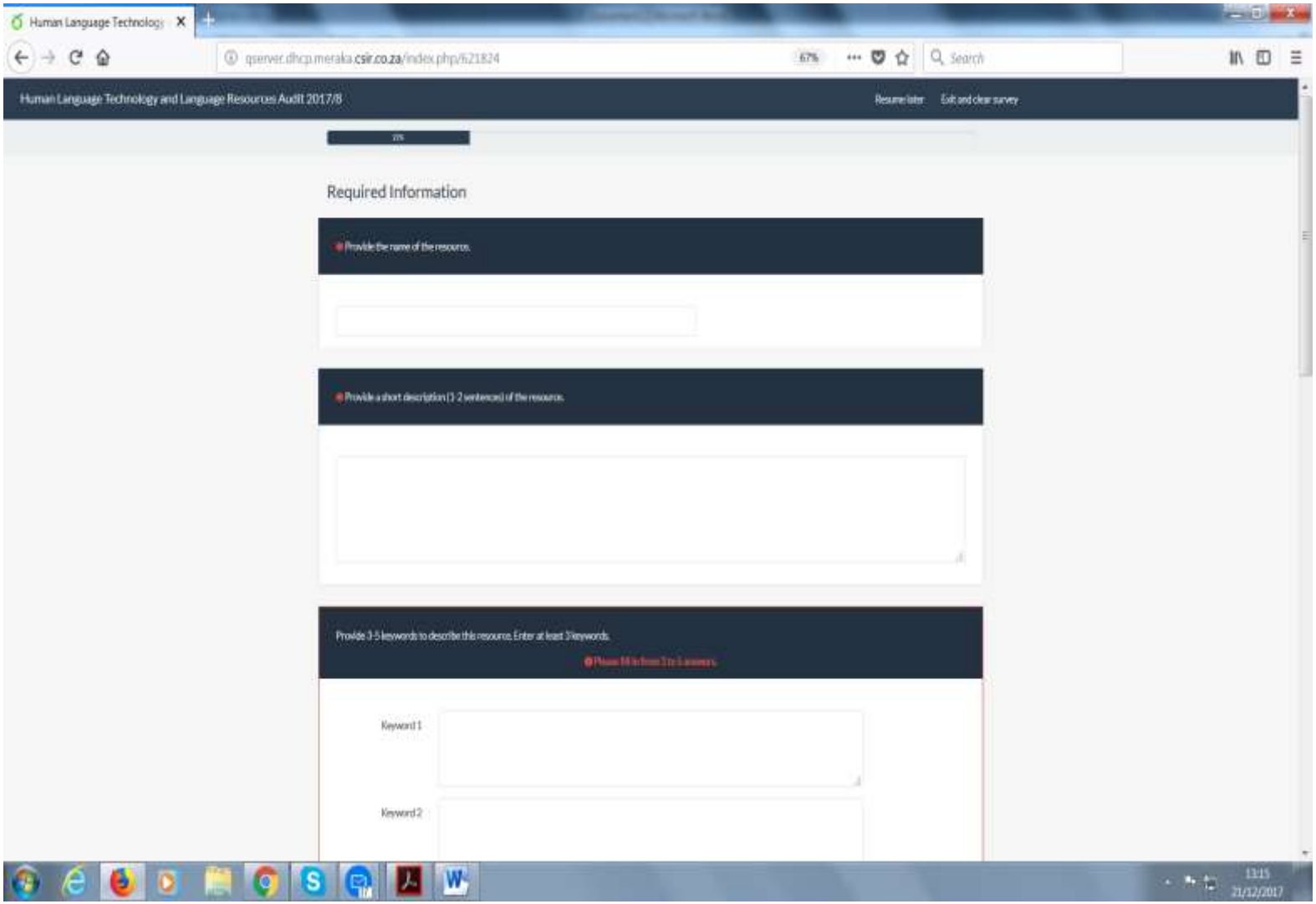
Resource type – speech – software – selection page



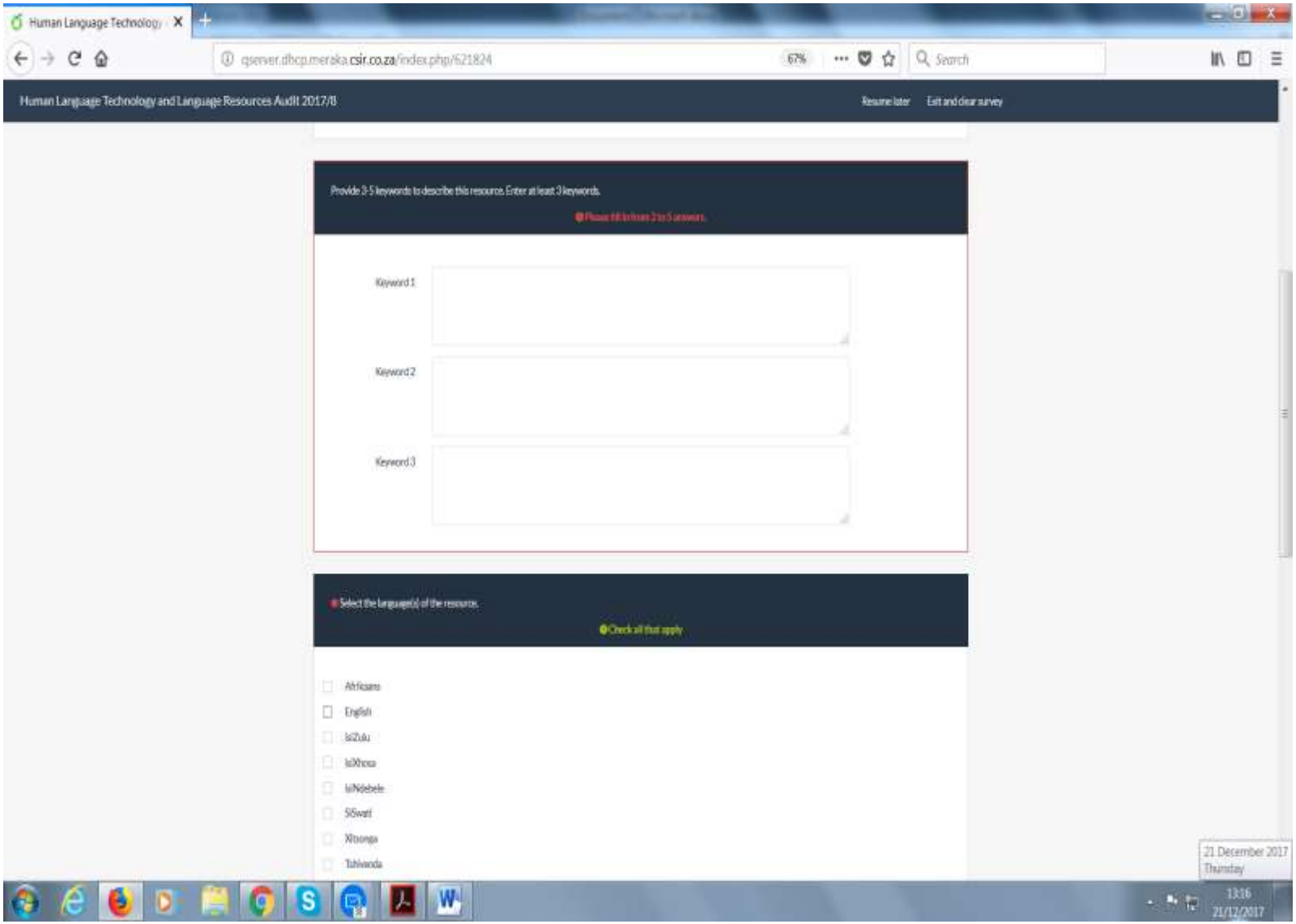
Resource type – multimodal page



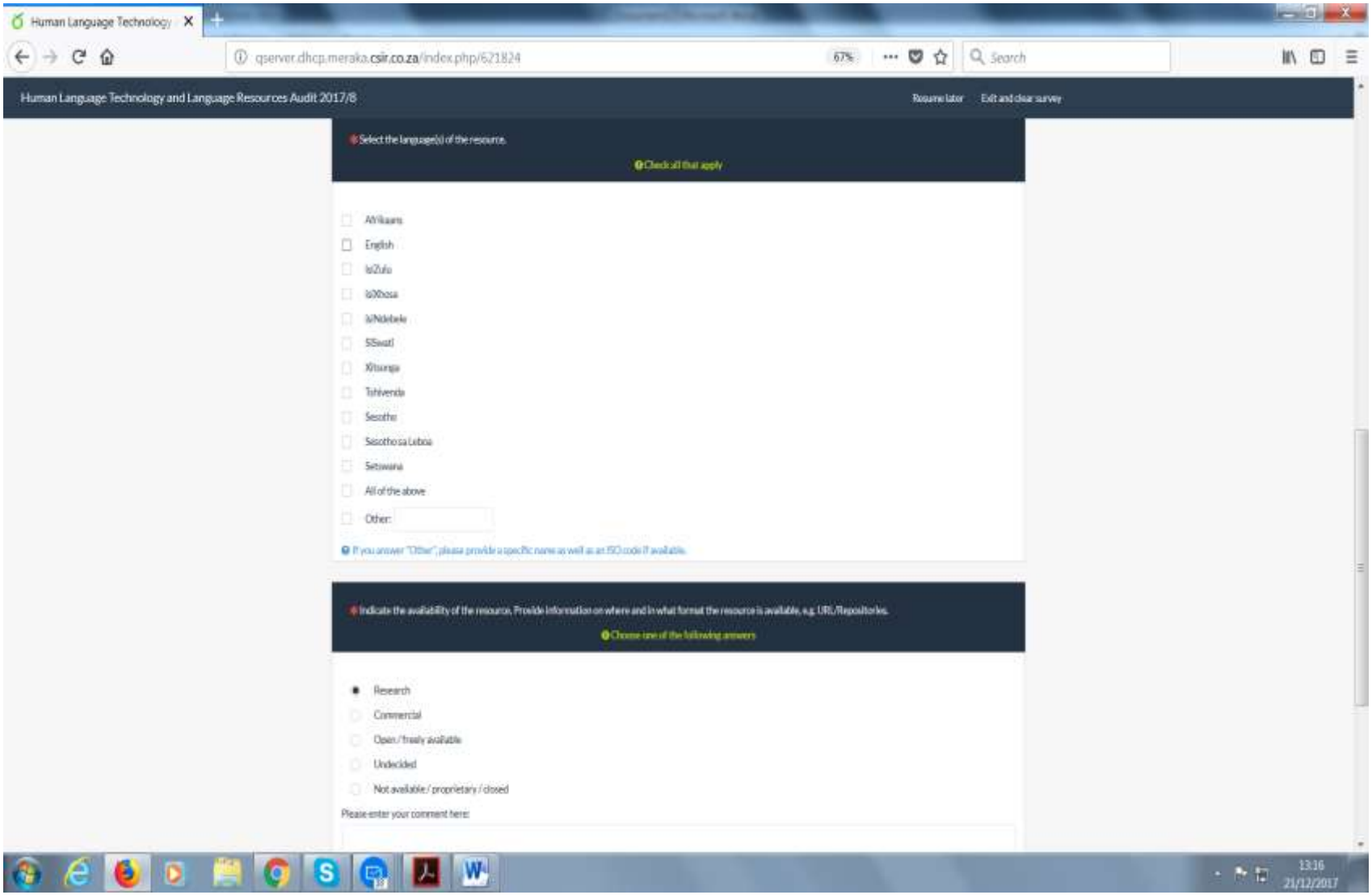
Required information page (1)



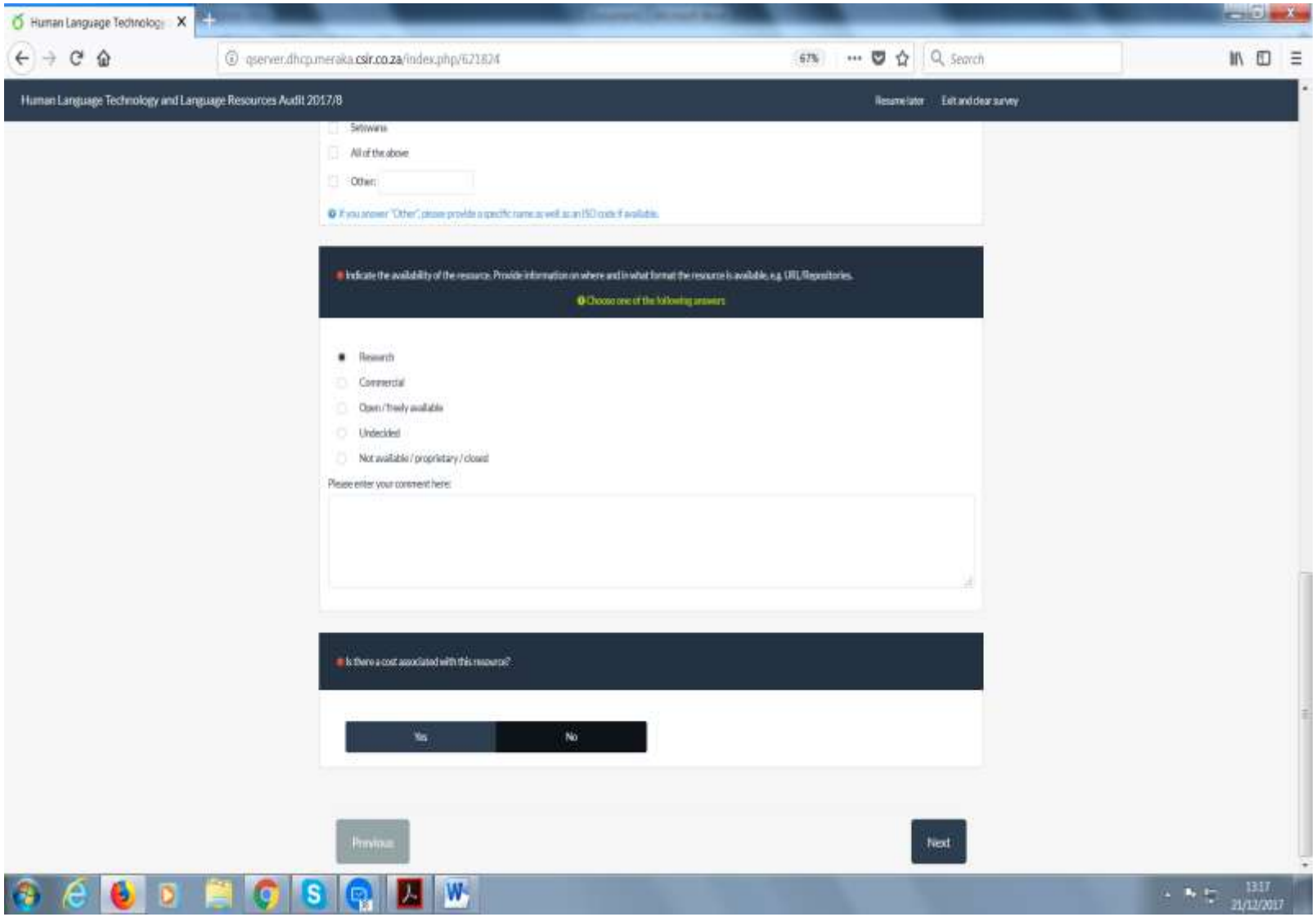
Required information page (2)



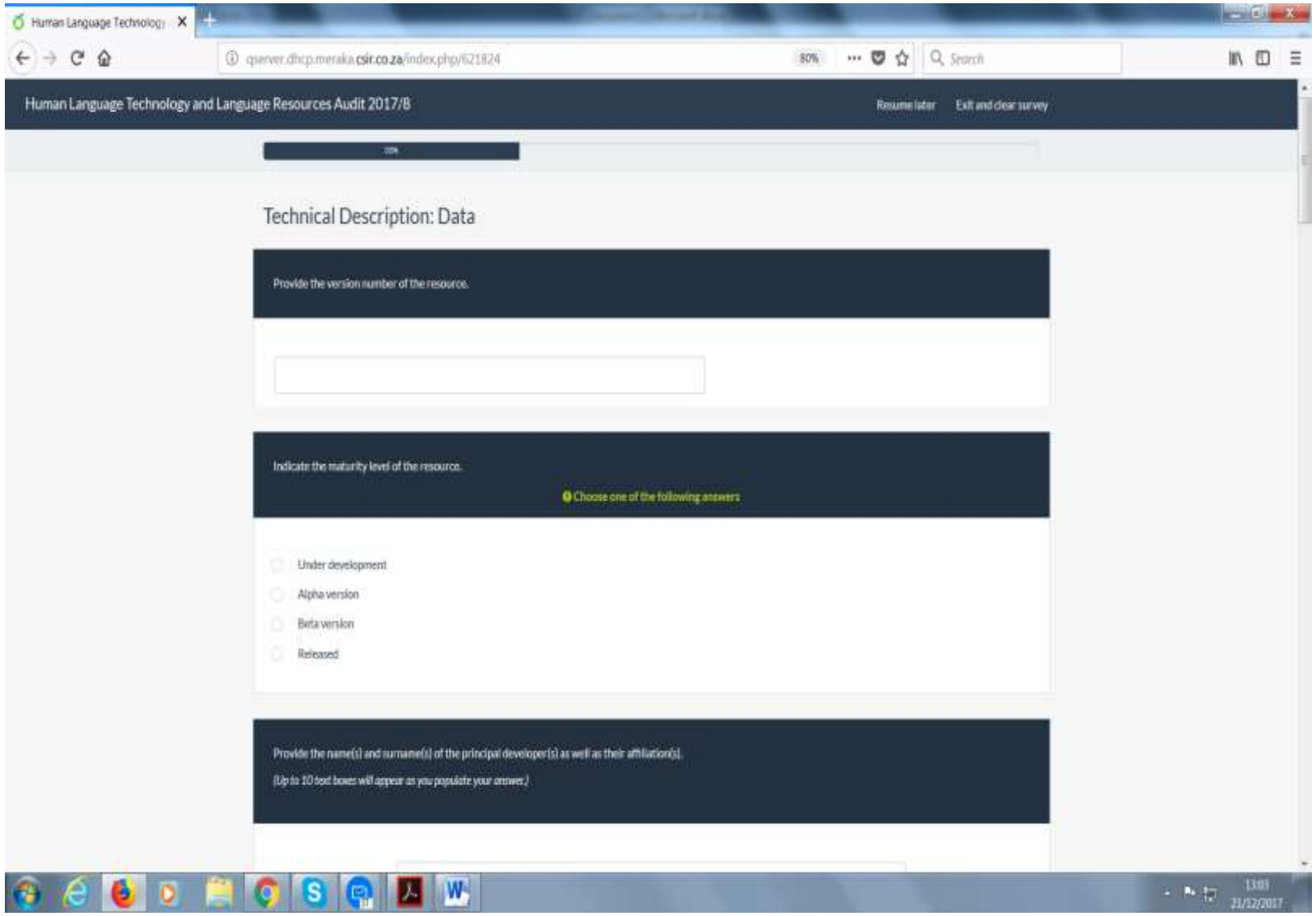
Required information page (3)



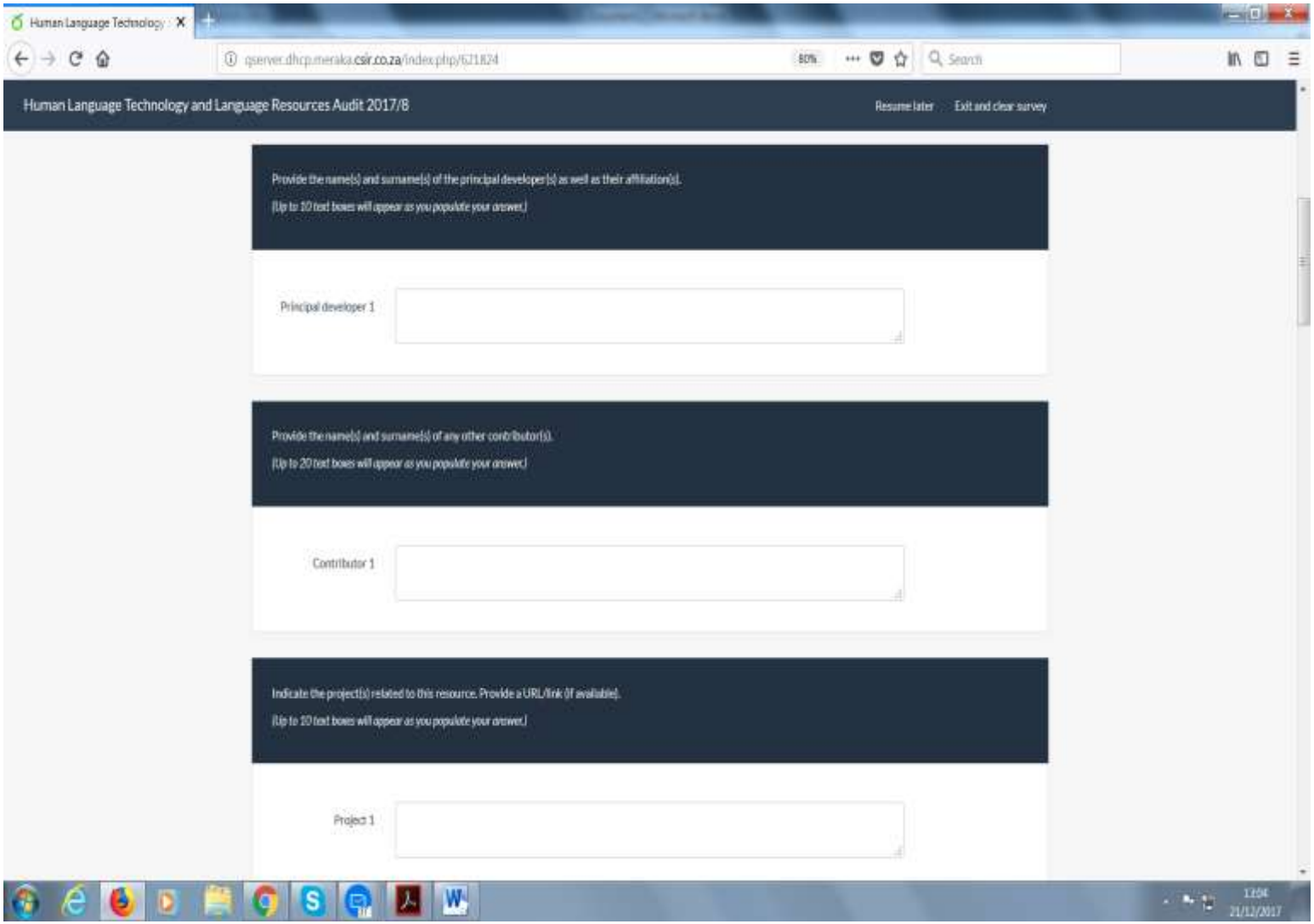
Required information page (4)



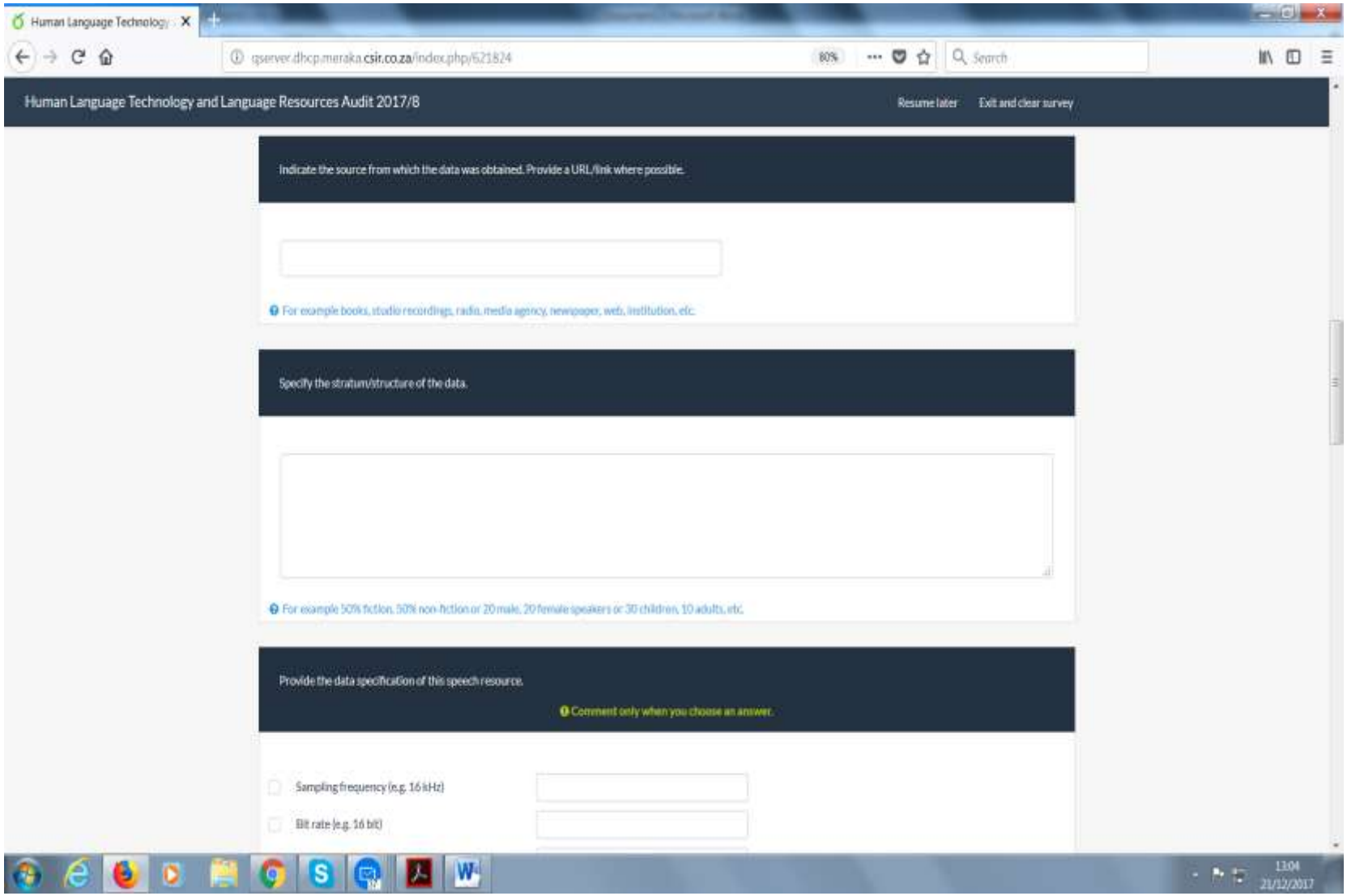
Technical Description: Data page (1)



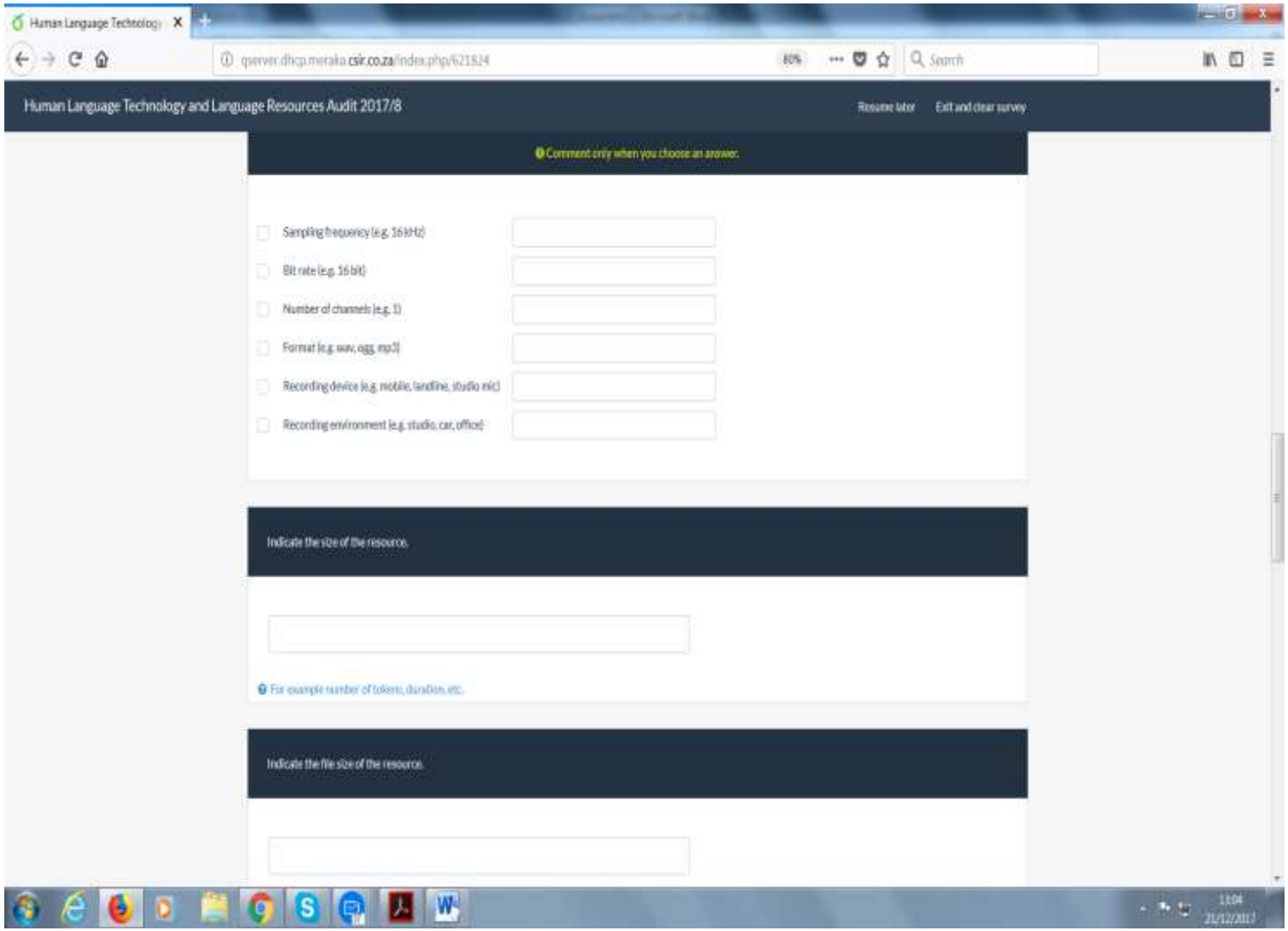
Technical description – data page (2)



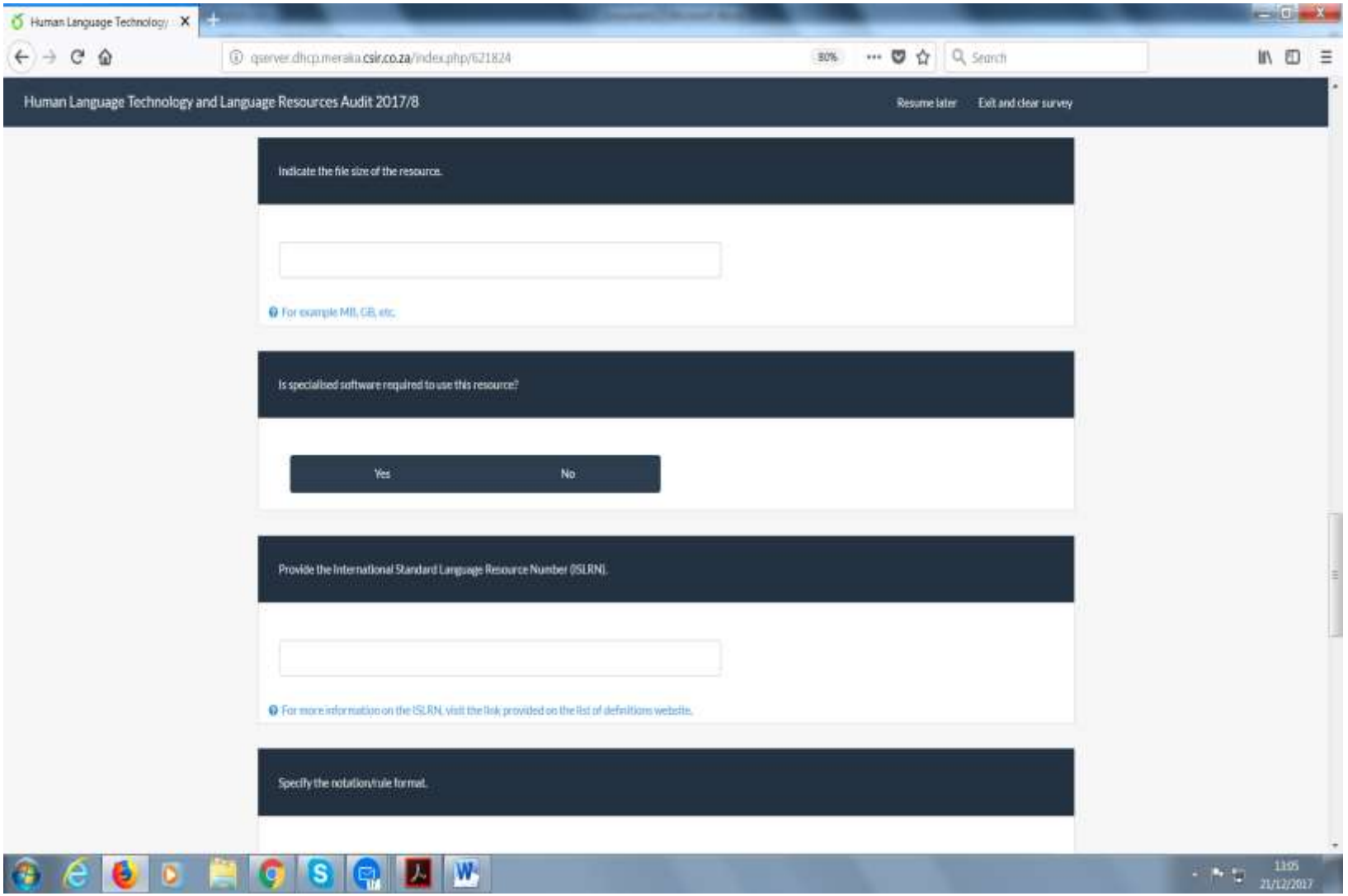
Technical description – data page (3)



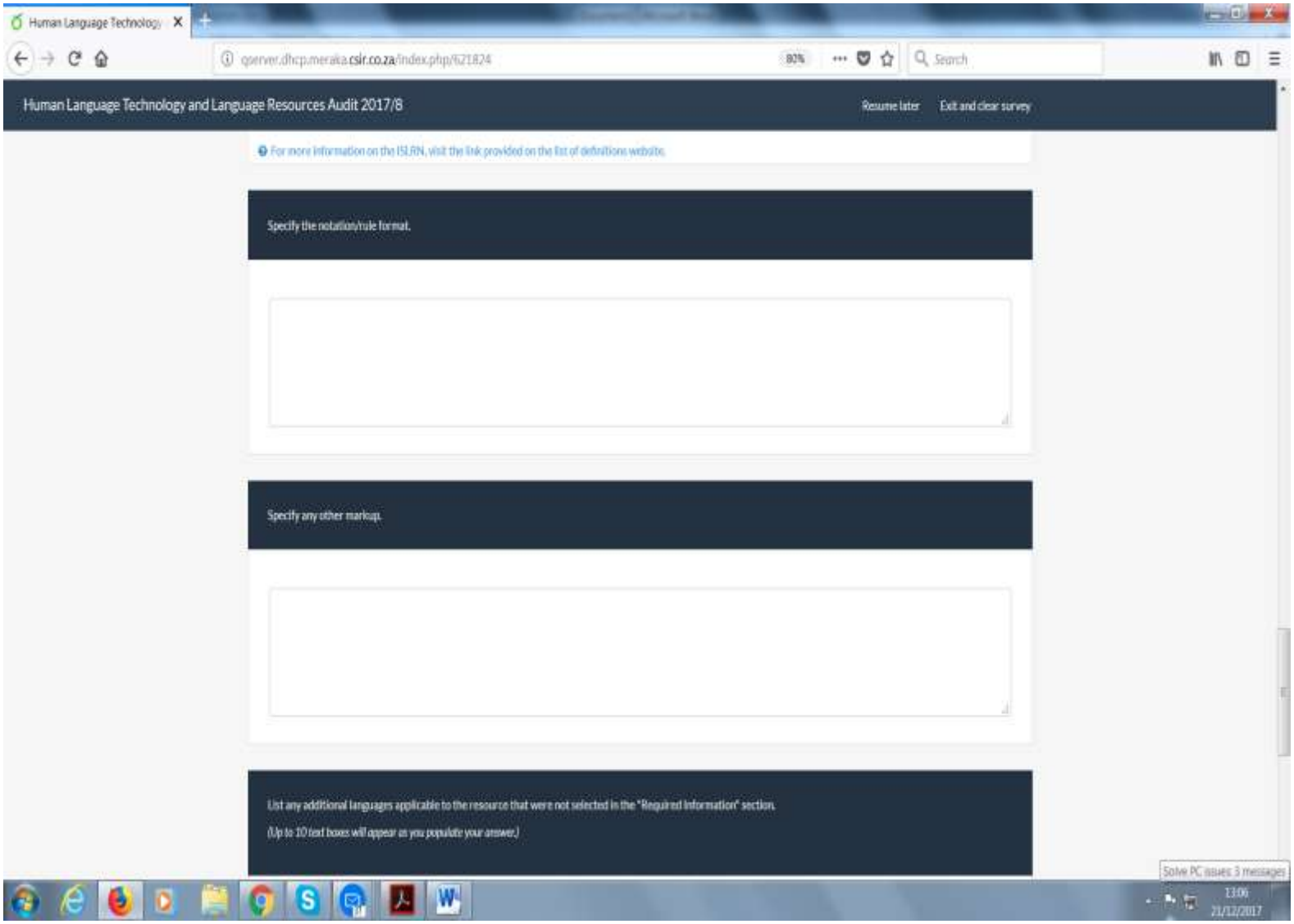
Technical description – data page (4)



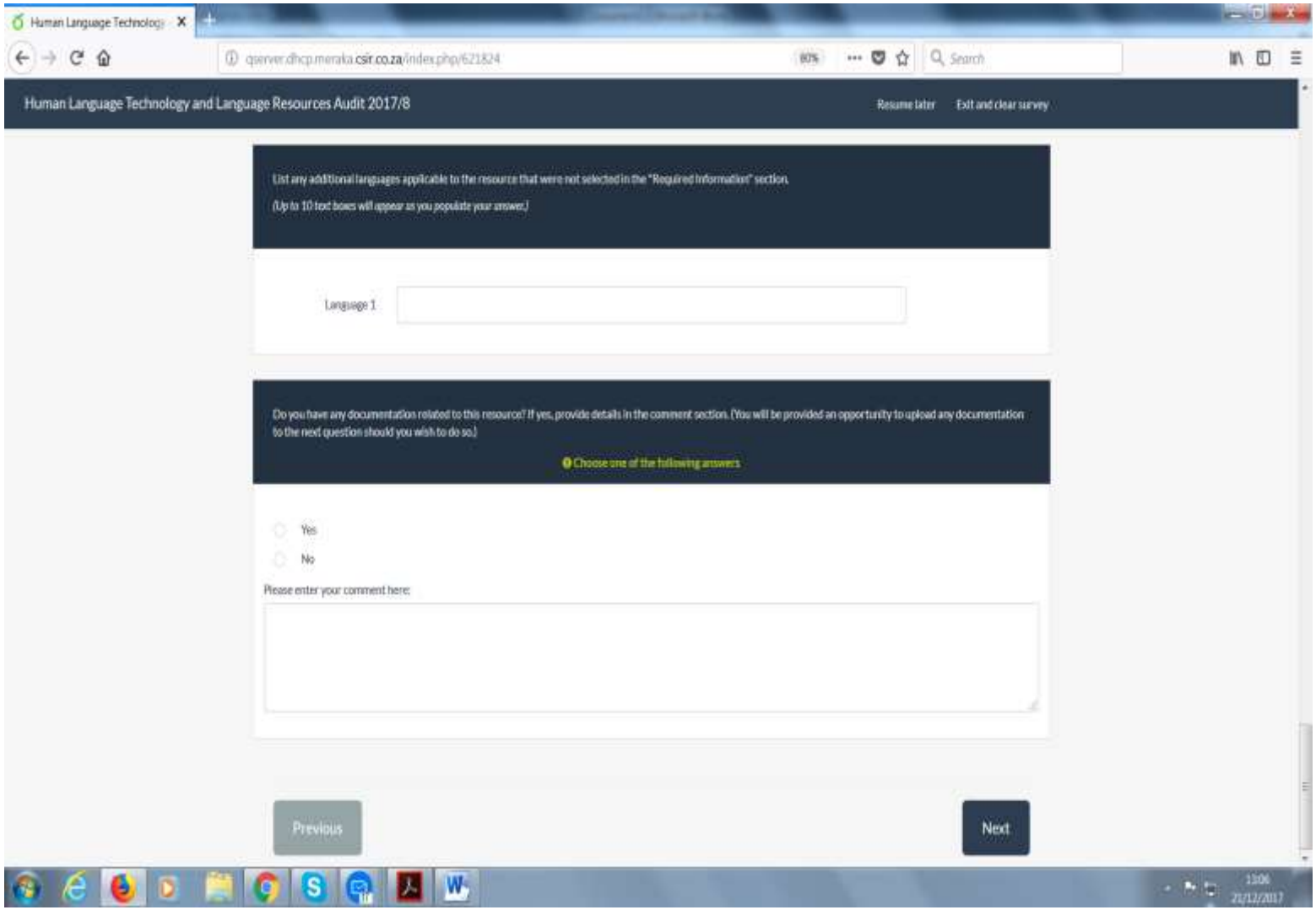
Technical description – data page (5)



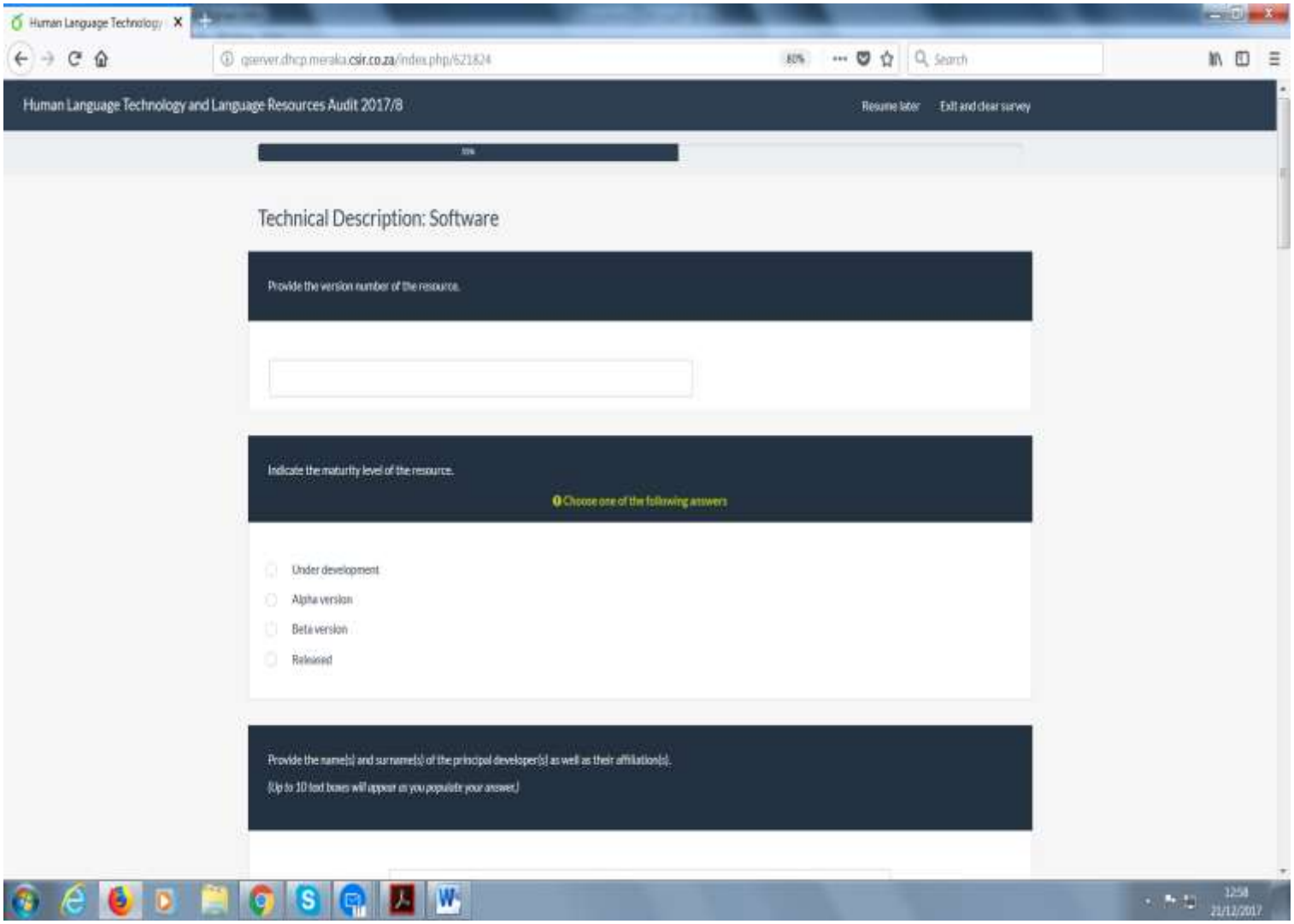
Technical description – data page (6)



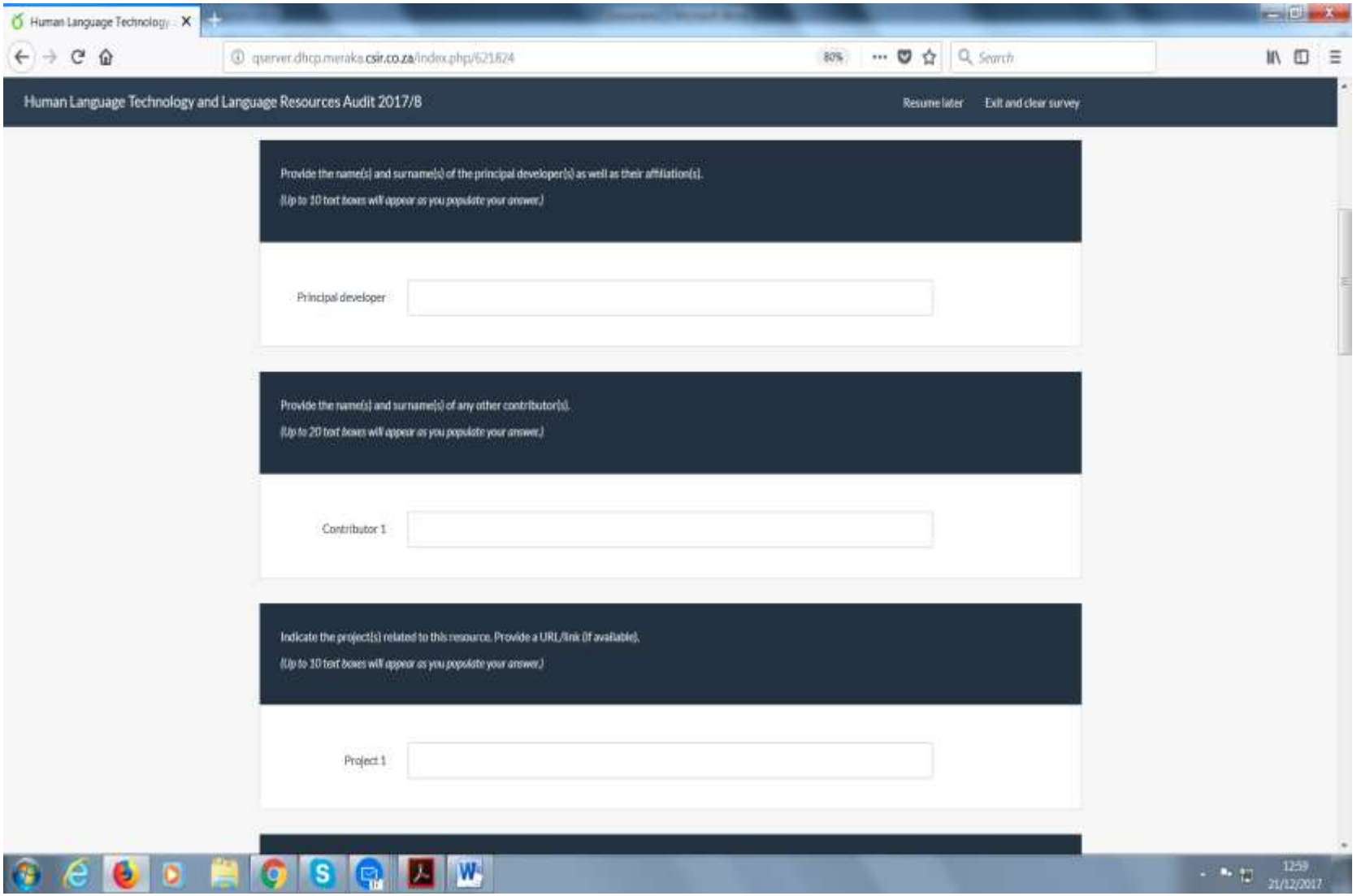
Technical description – data page (7)



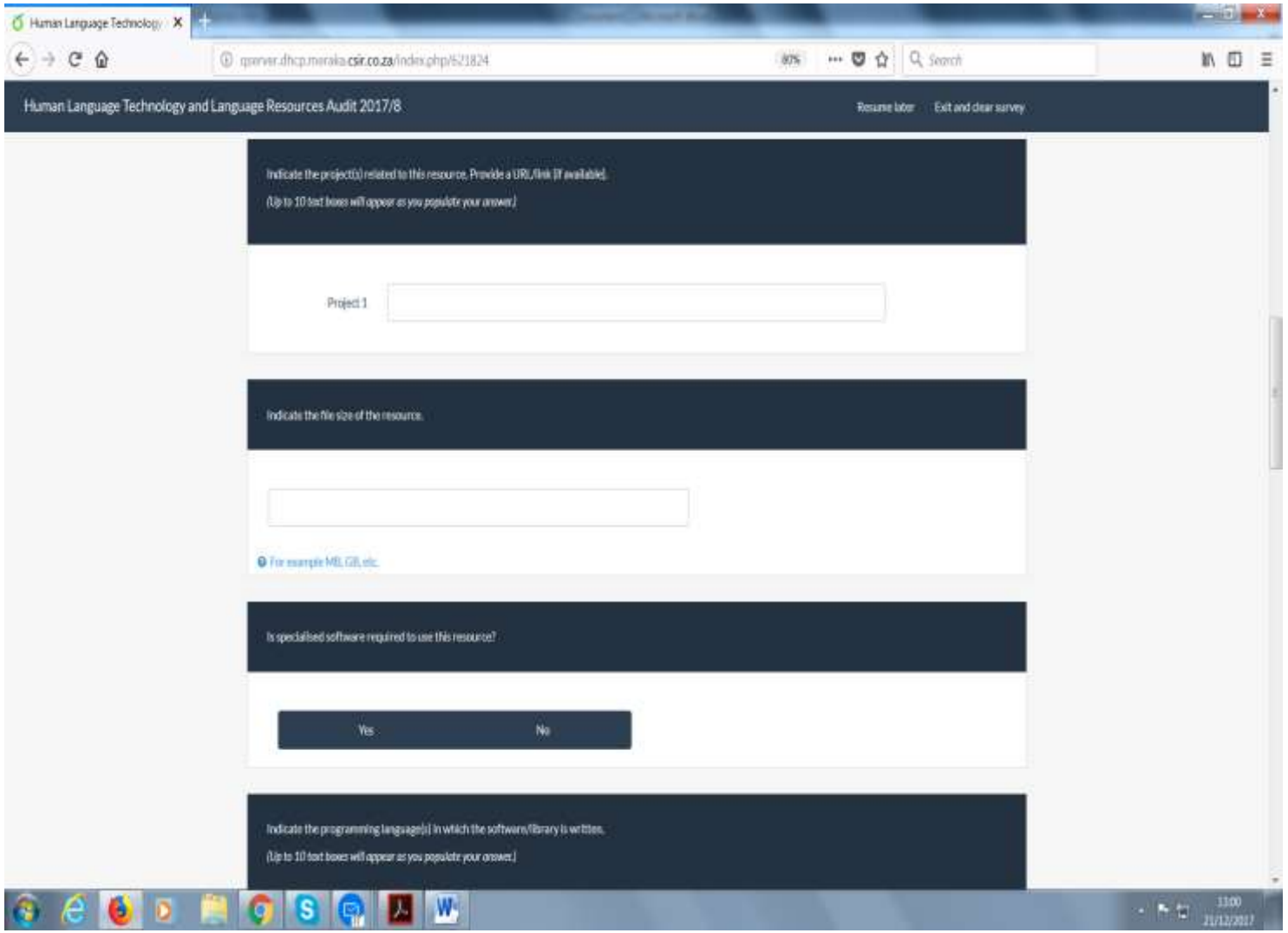
Technical Description - Software page (1)



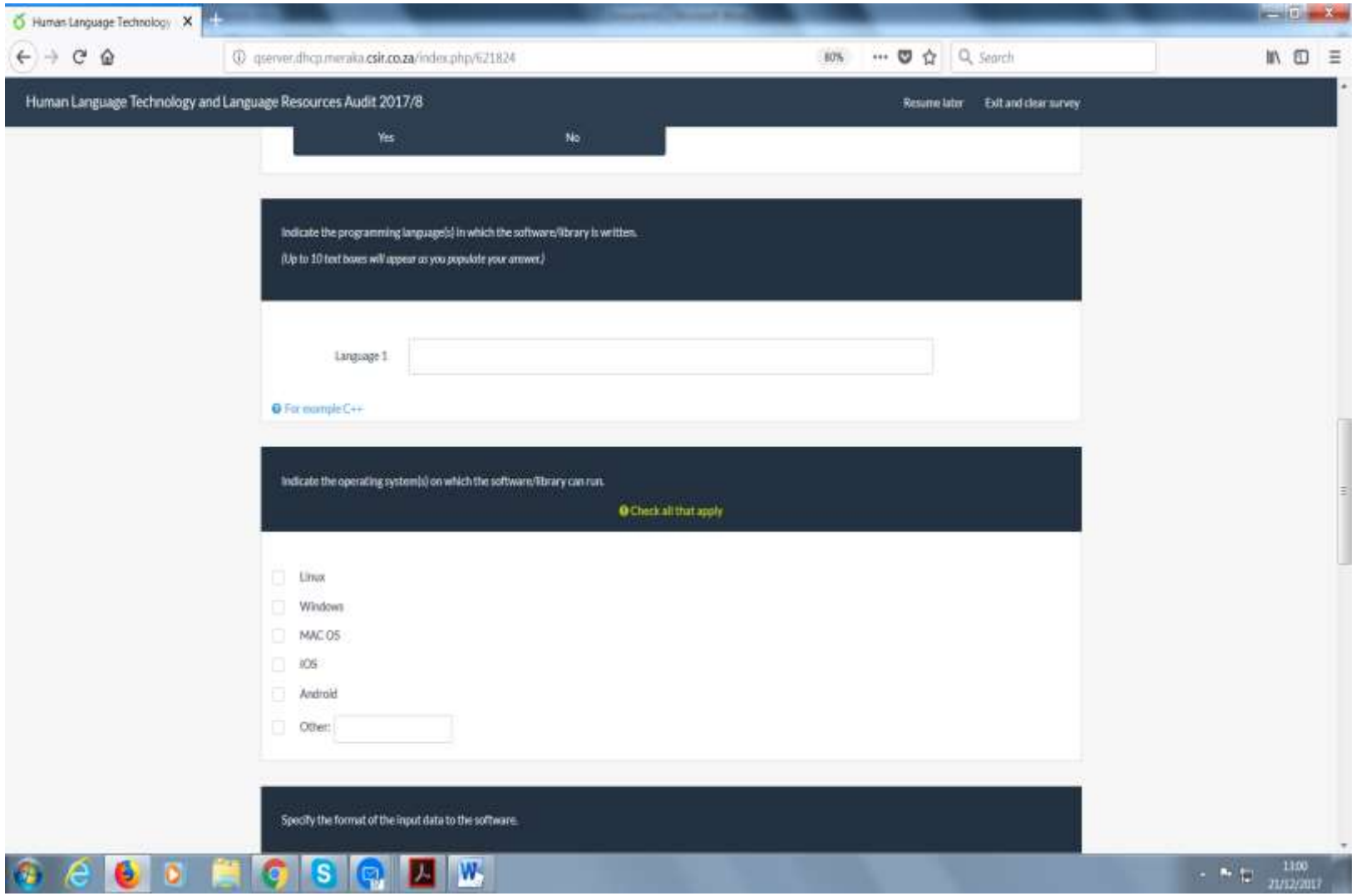
Technical description – software page (2)



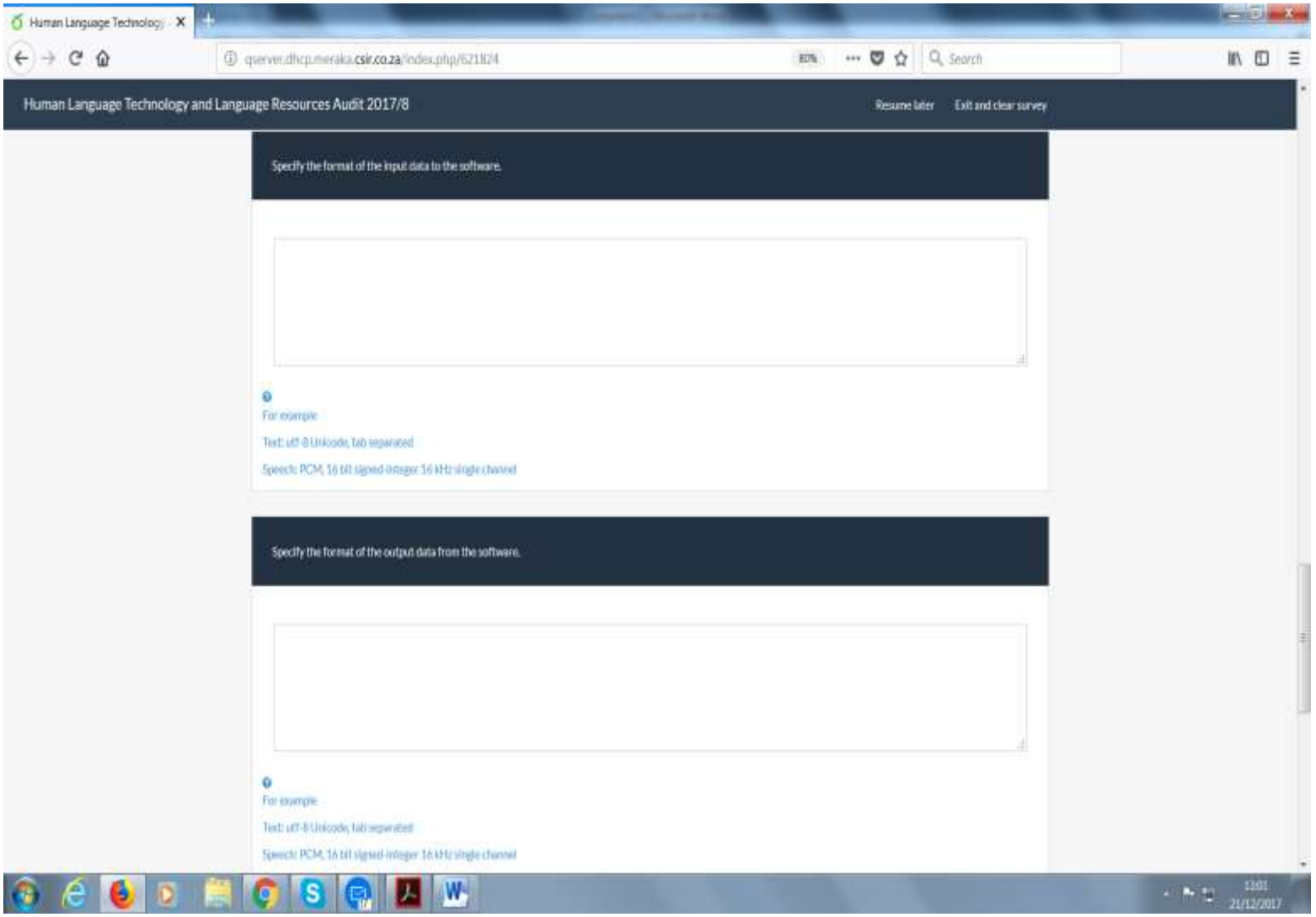
Technical description – software page (3)



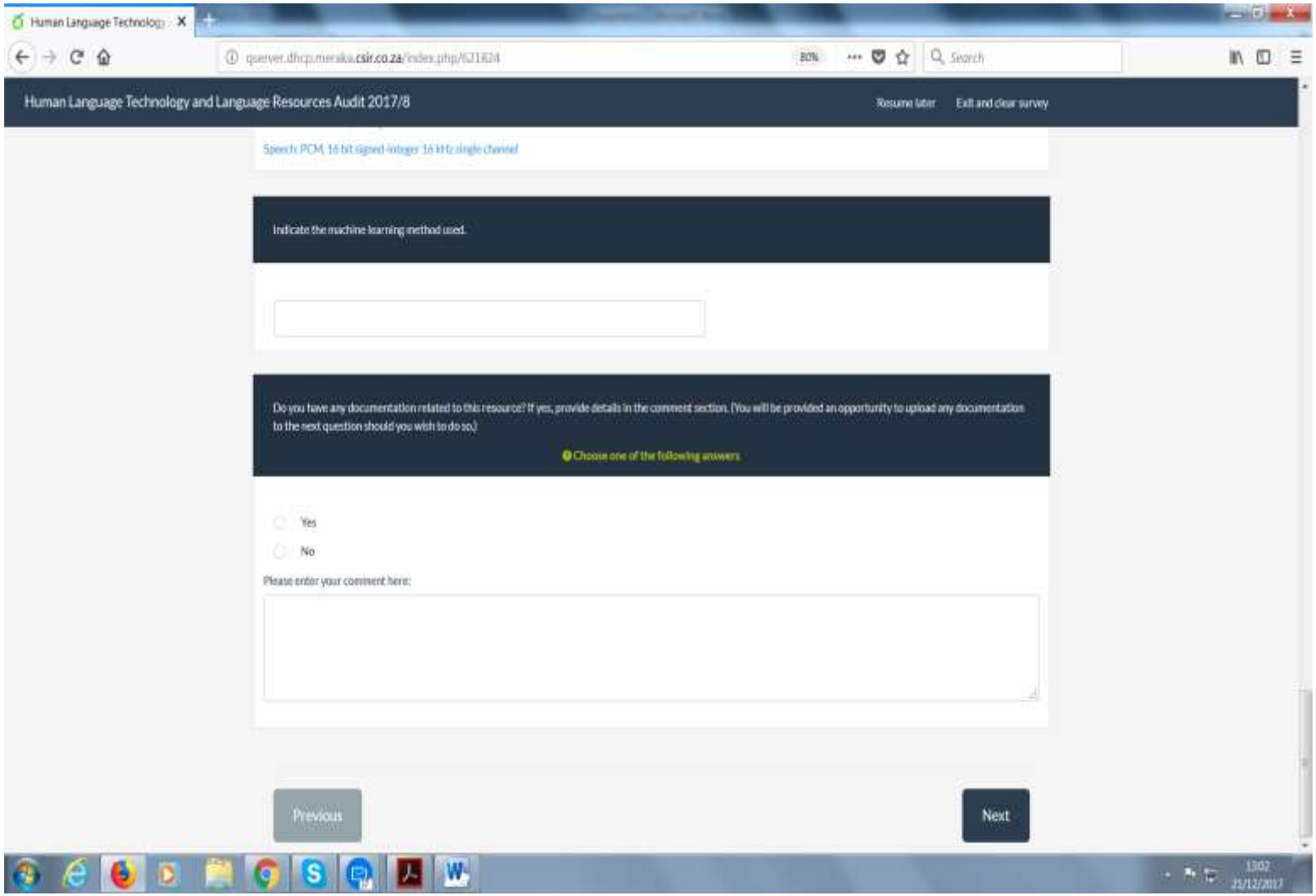
Technical description – software page (4)



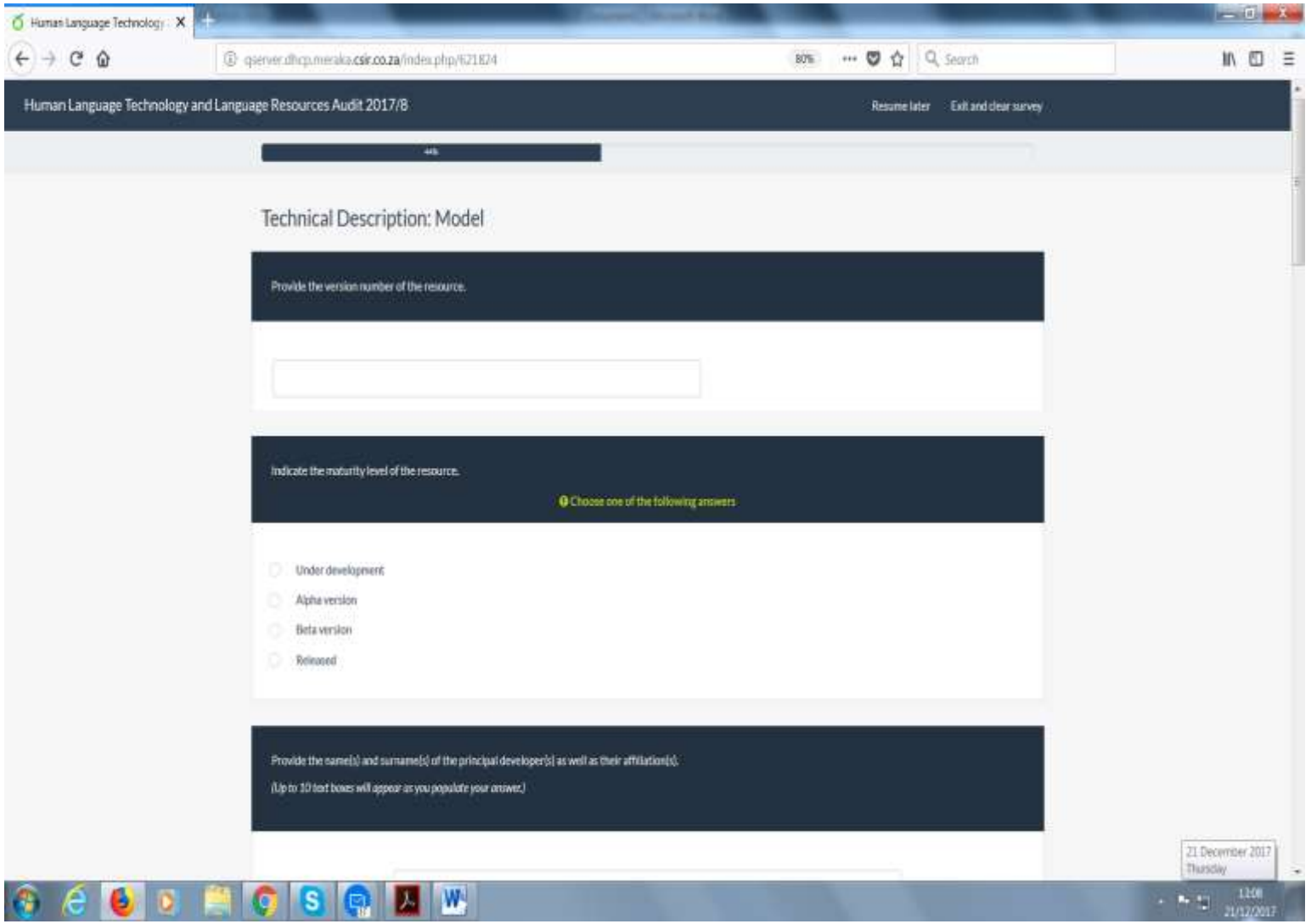
Technical description – software page (5)



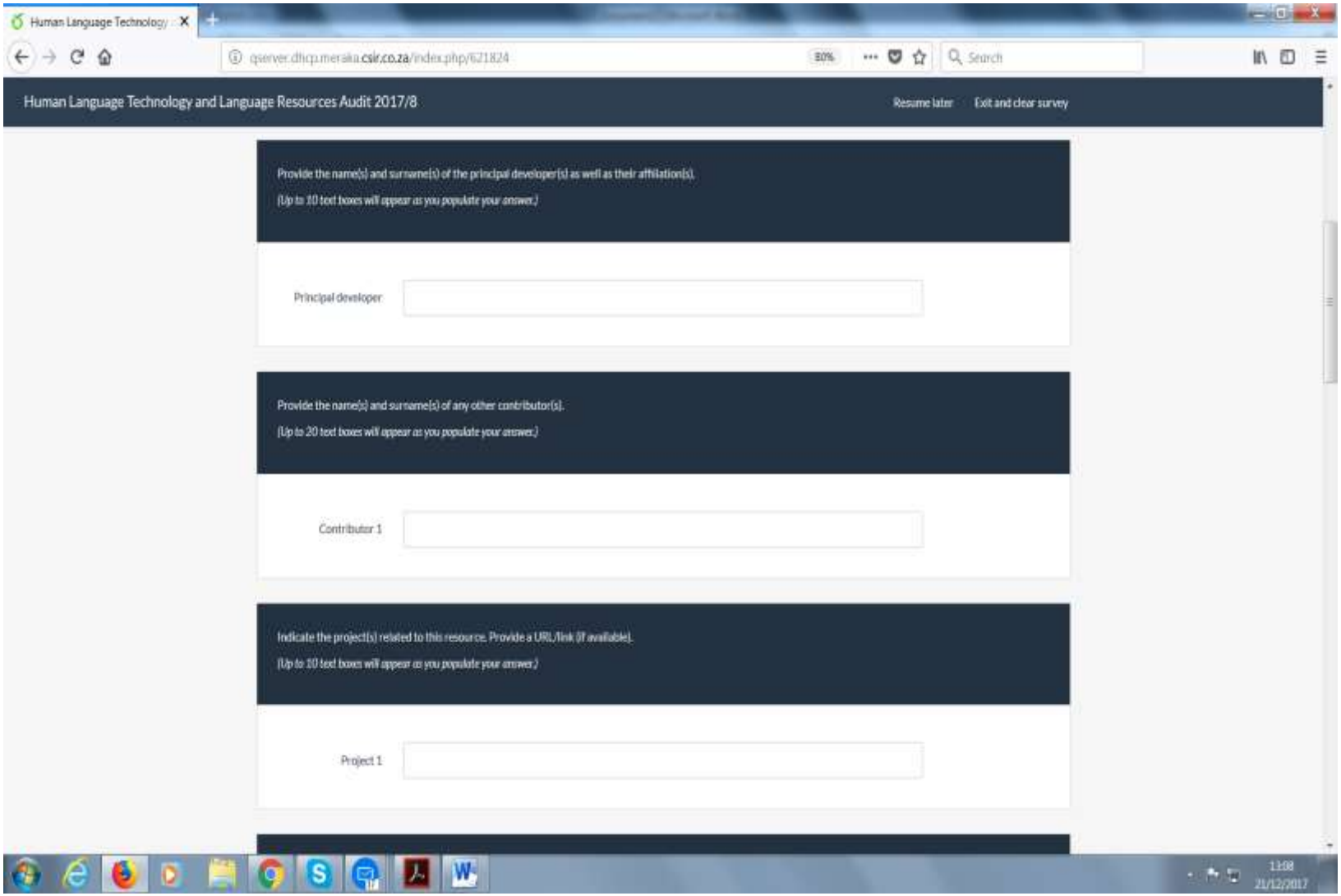
Technical description – software page (6)



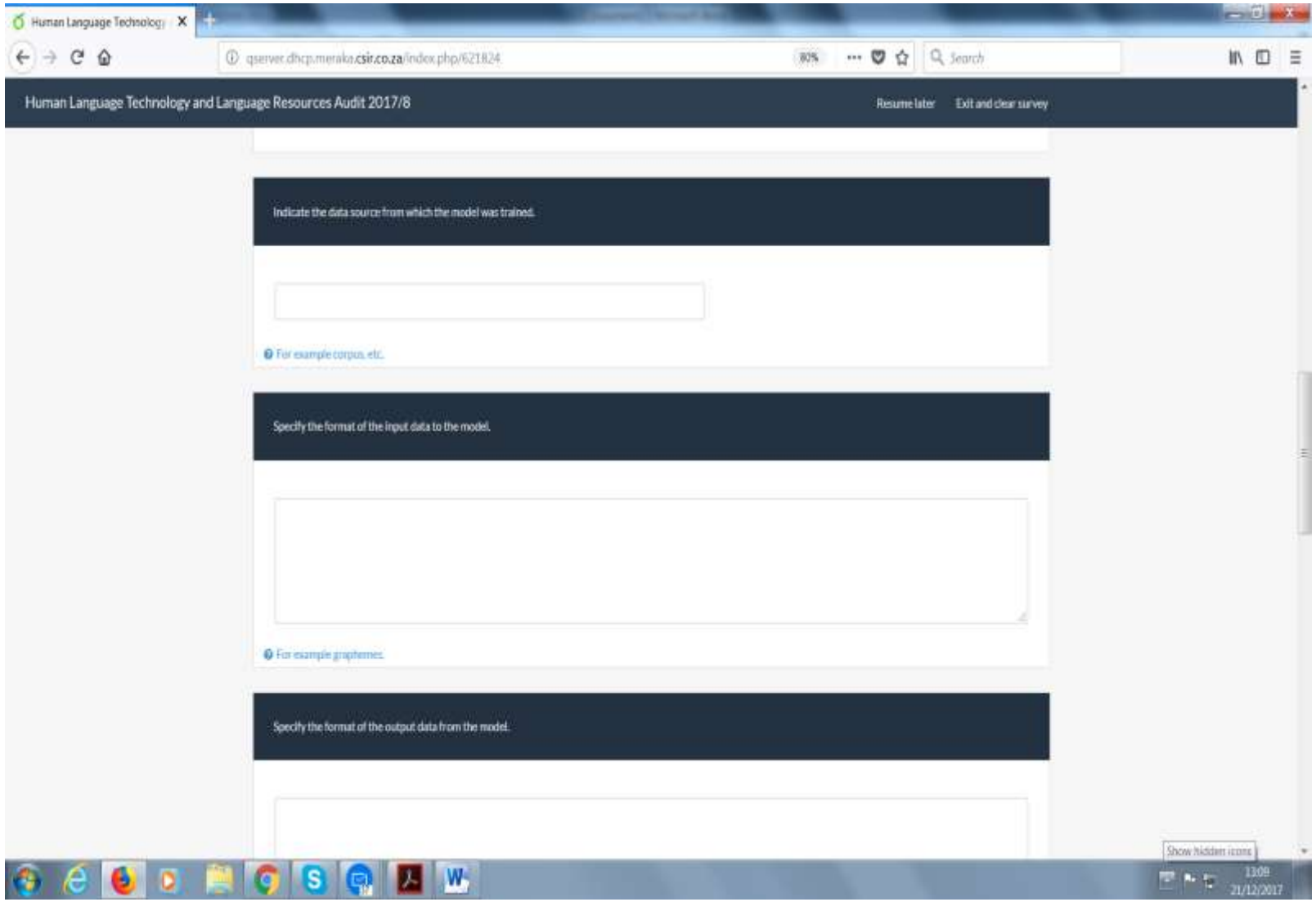
Technical Description - Model page (1)



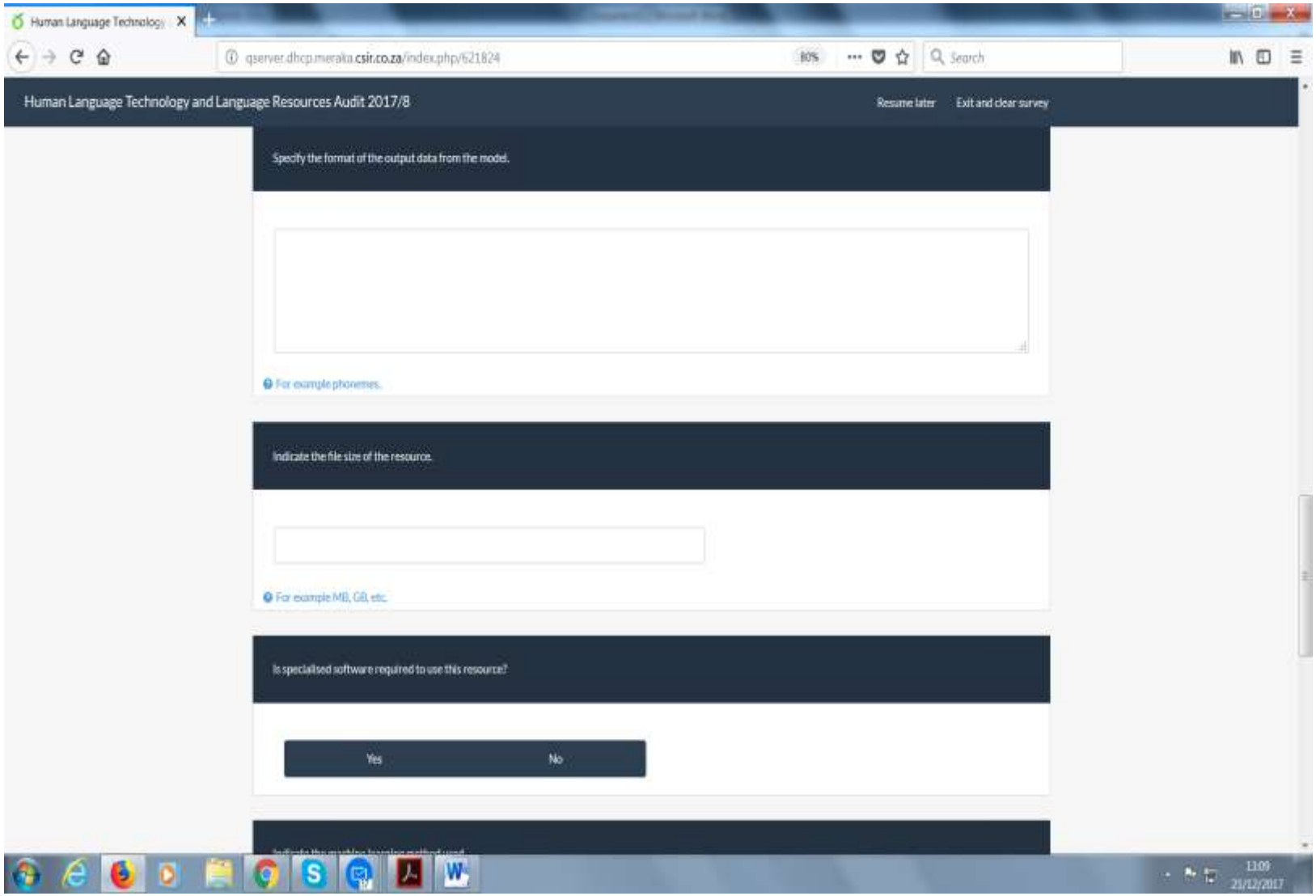
Technical Description - Model page (2)



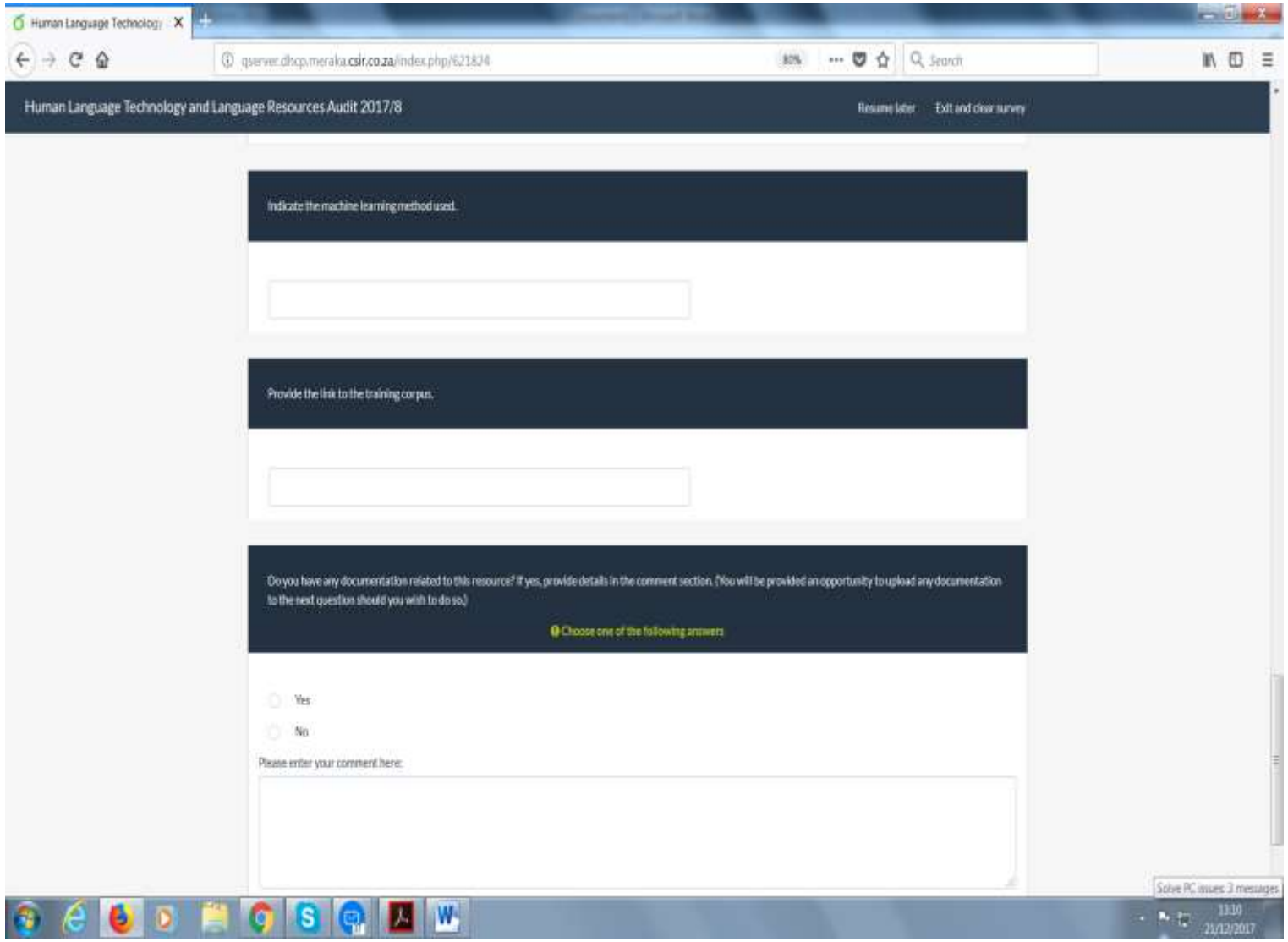
Technical Description - Model page (3)



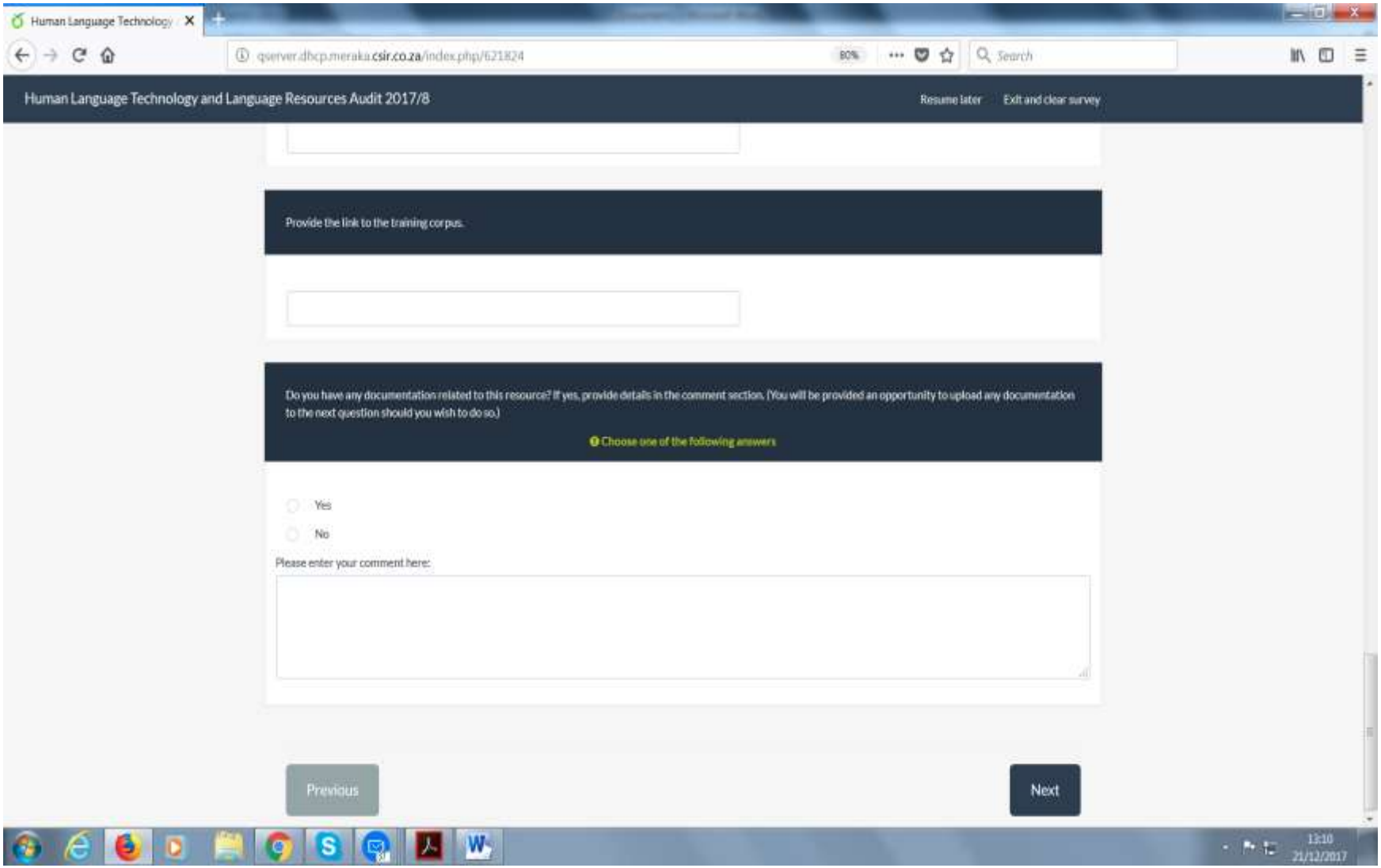
Technical Description - Model page (4)



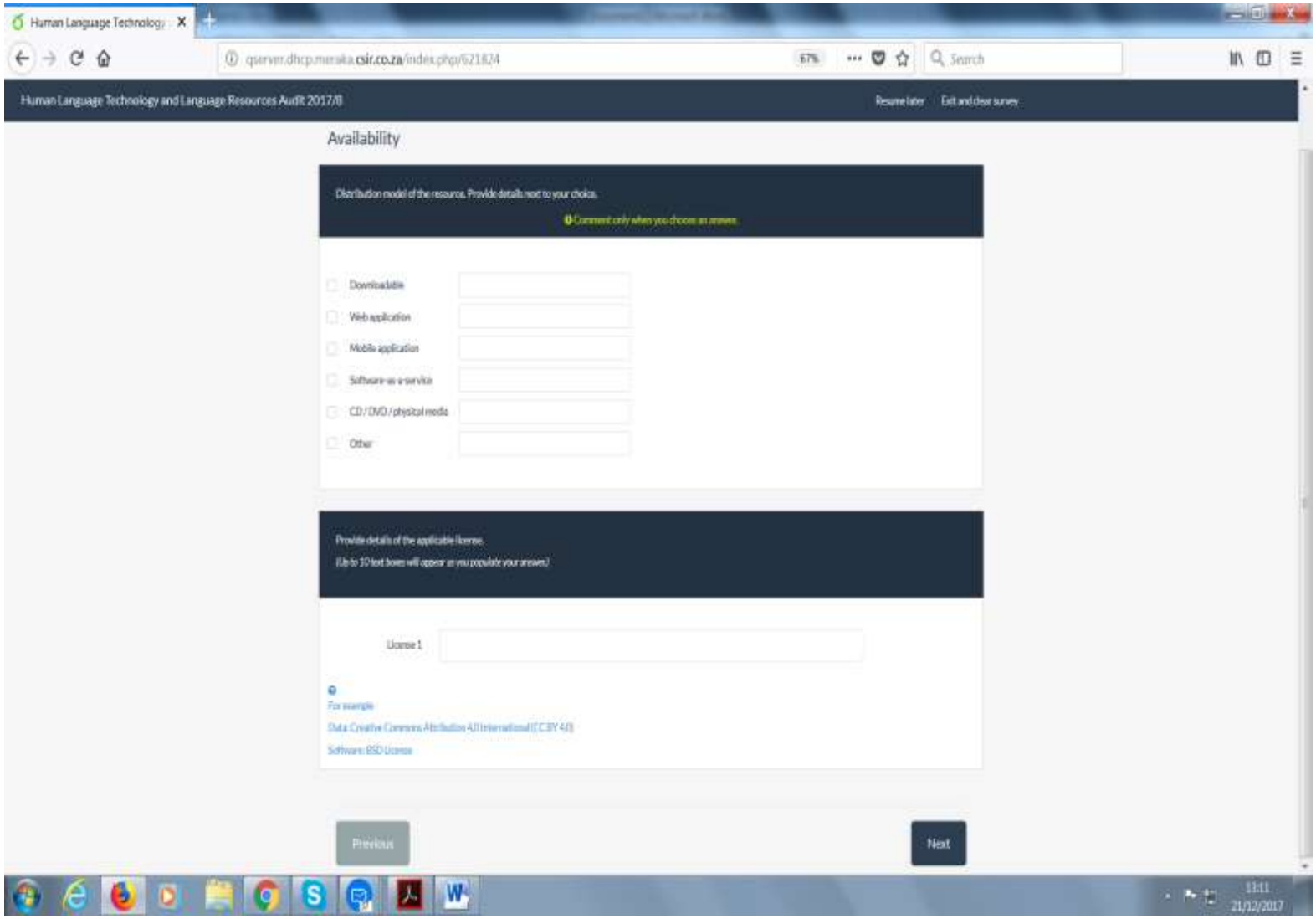
Technical Description - Model page (5)



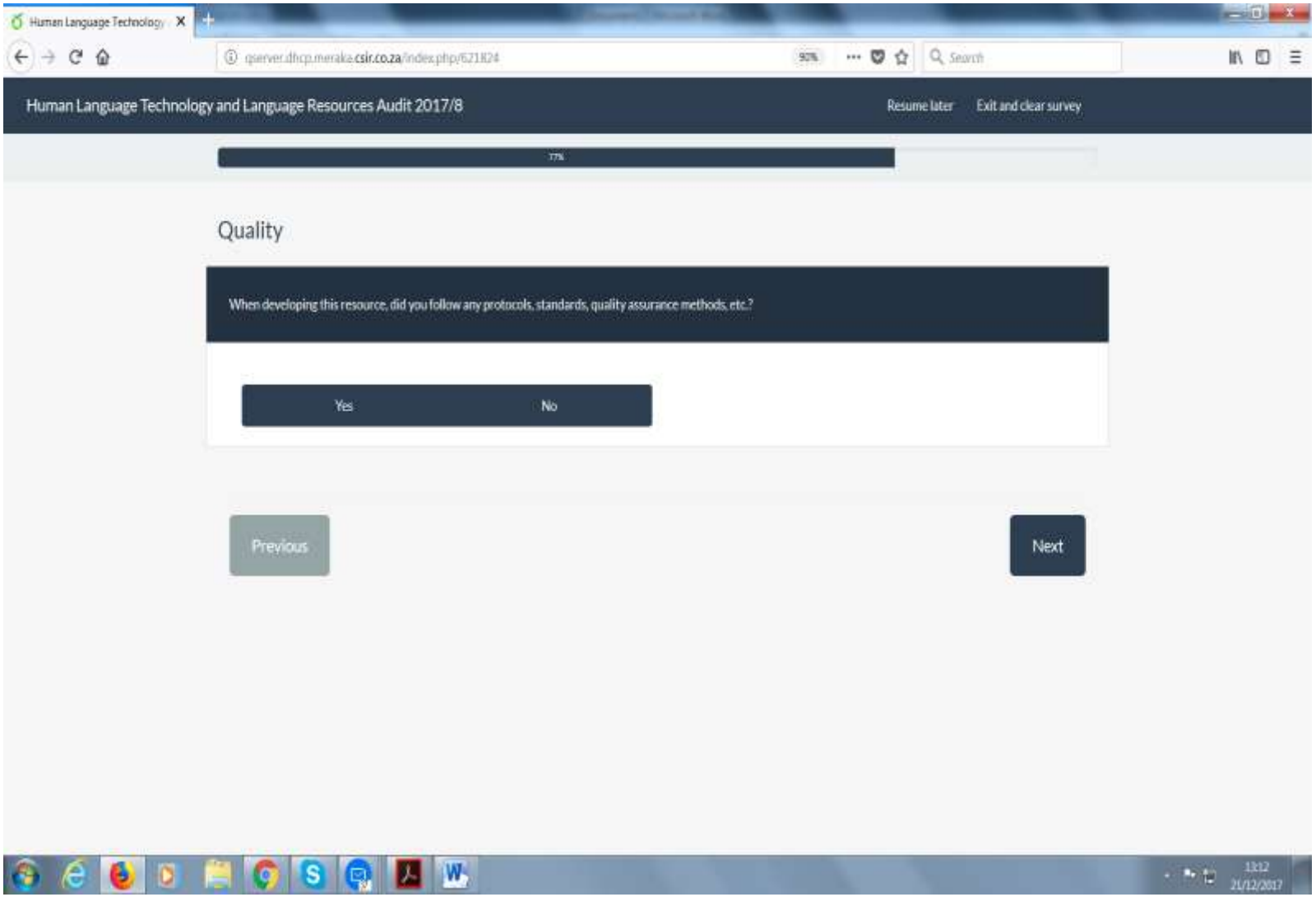
Technical Description - Model page (6)



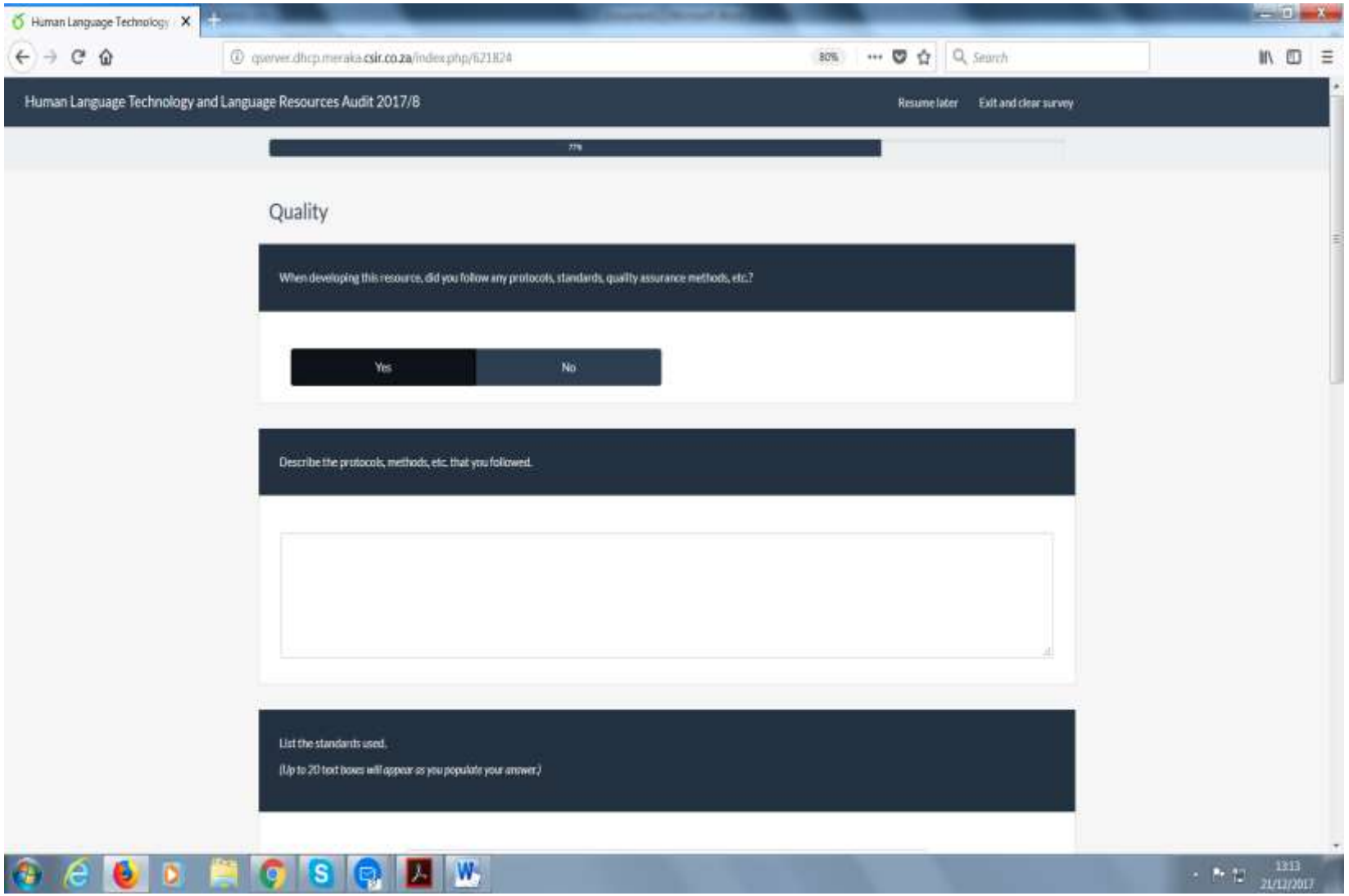
Availability page



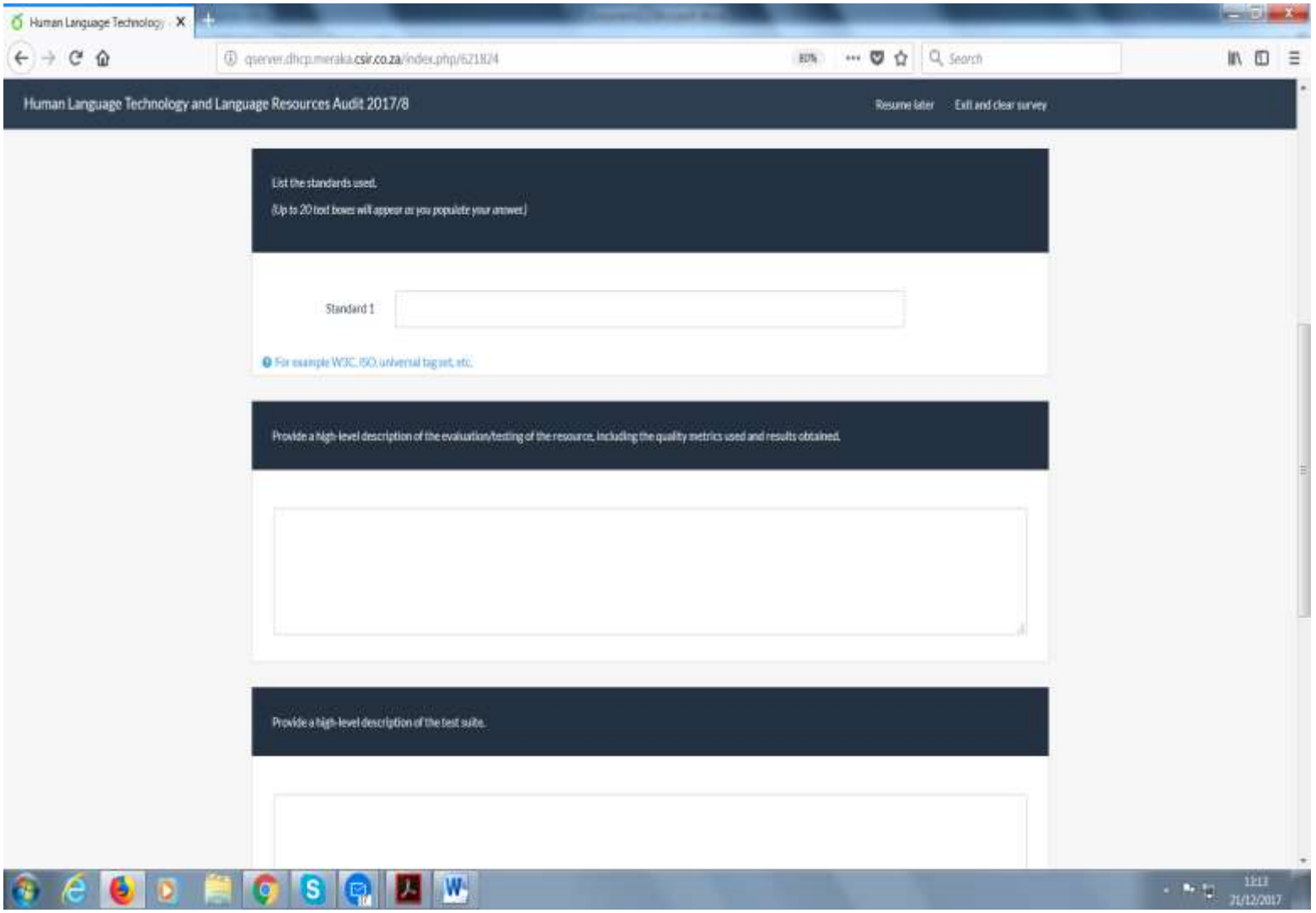
Quality page



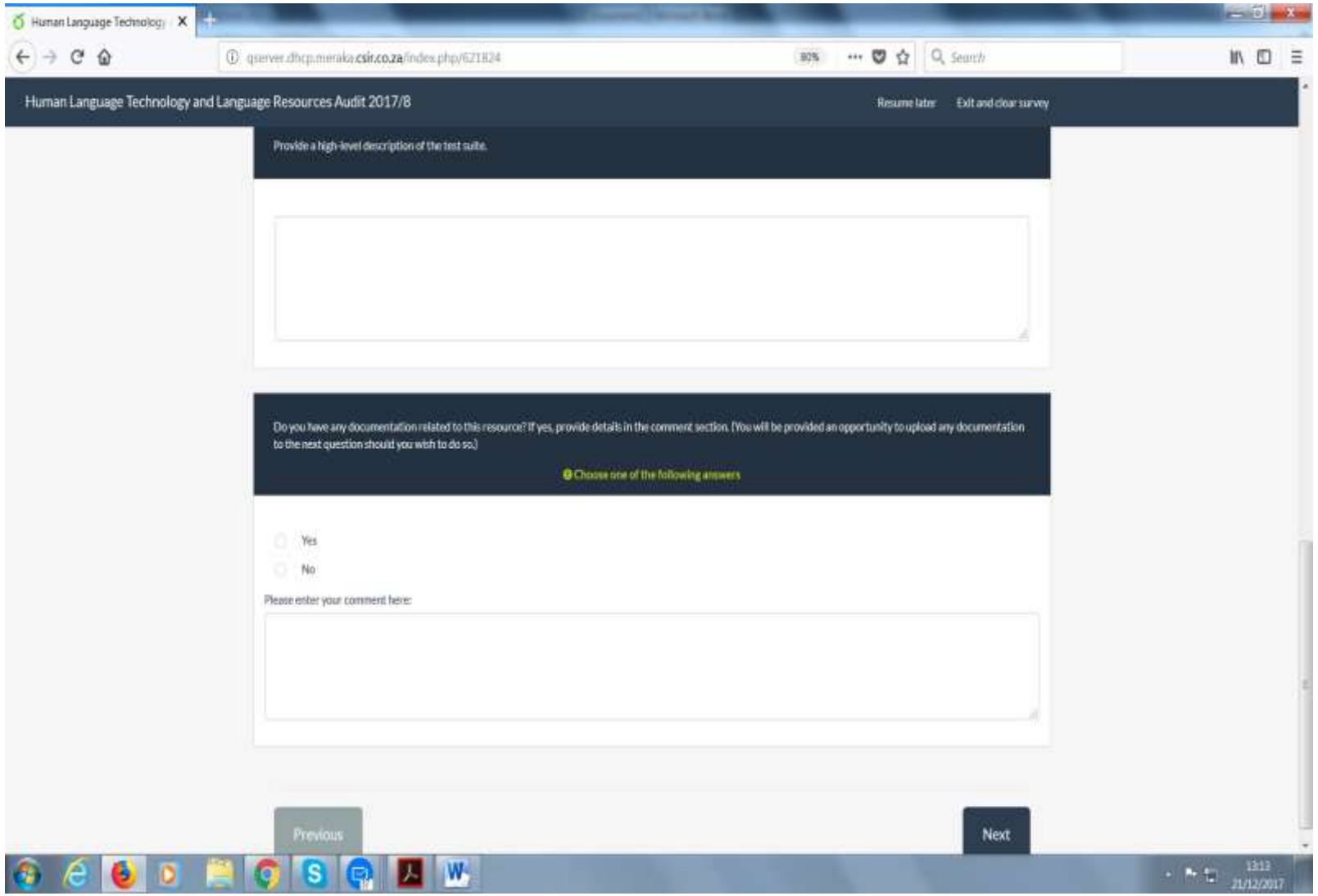
Quality page -Yes (1)



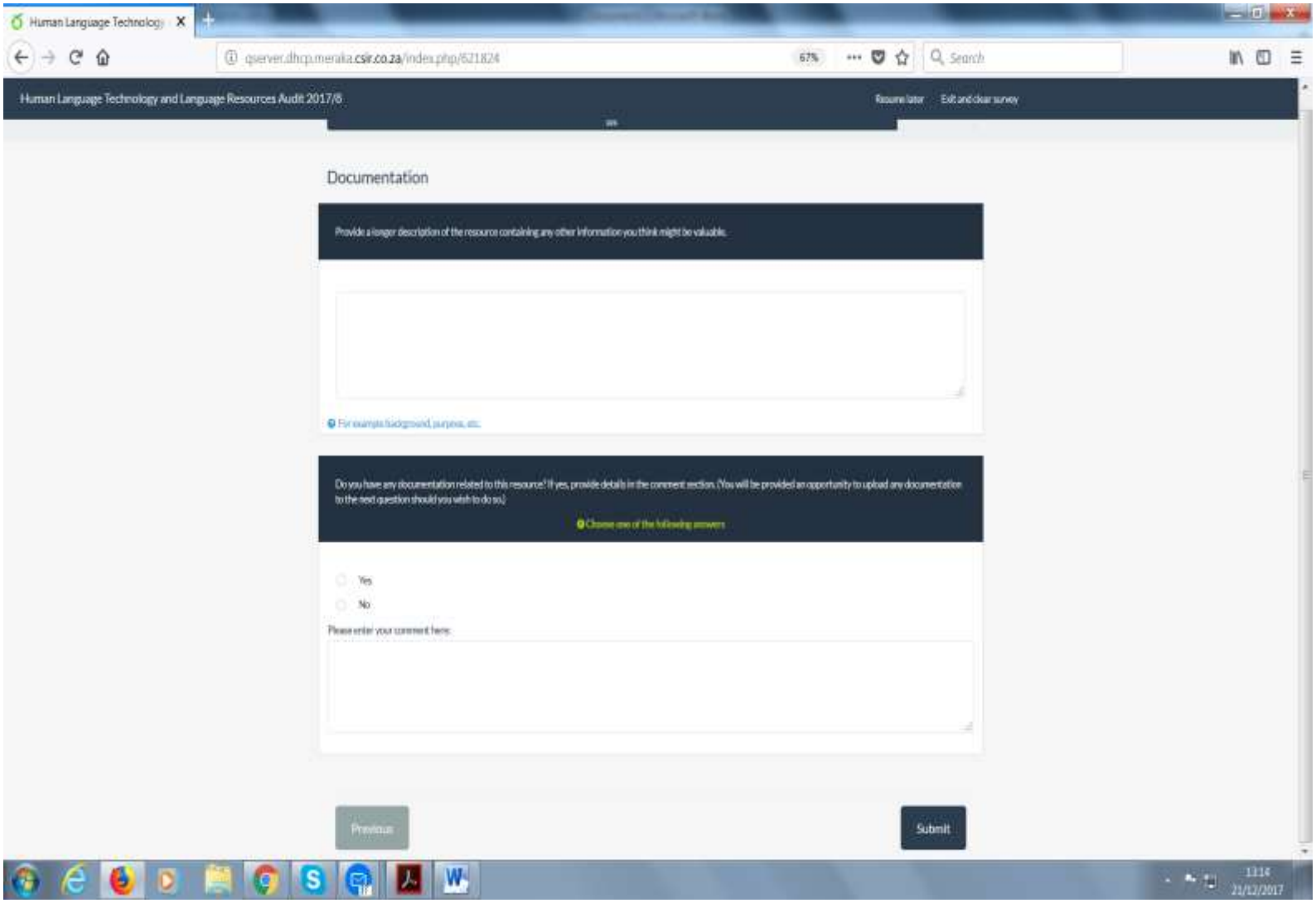
Quality page -Yes (2)



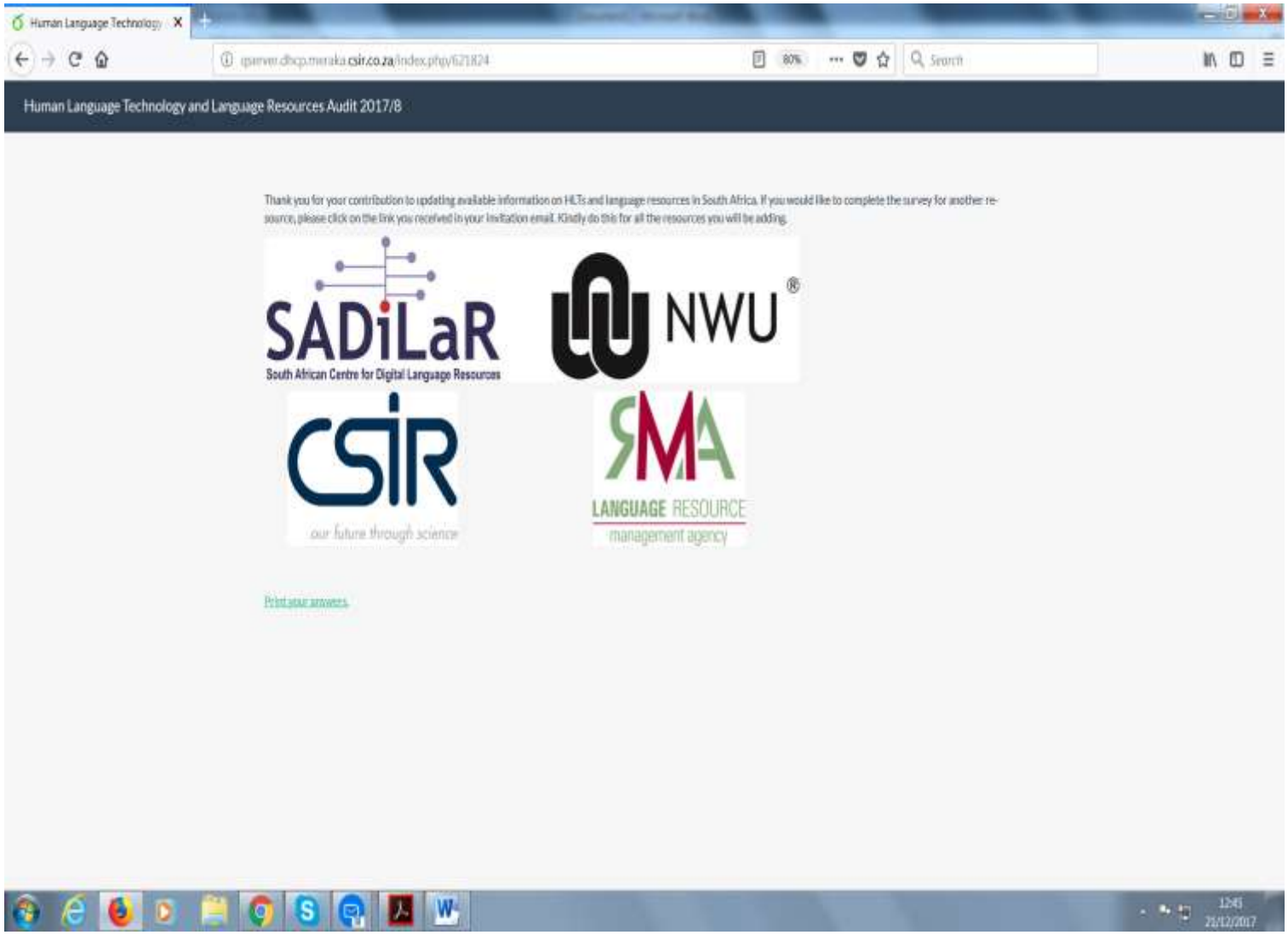
Quality page - Yes (3)



Documentation page



End page


Human Language Technology
Human Language Technology and Language Resources Audit 2017/8
Thank you for your contribution to updating available information on HLTs and language resources in South Africa. If you would like to complete the survey for another resource, please click on the link you received in your invitation email. Kindly do this for all the resources you will be adding.
SADiLaR
South African Centre for Digital Language Resources
NWU
CSIR
our future through science
RMA
LANGUAGE RESOURCE
management agency
Print your answers

ANNEXURE E - QUESTIONNAIRES: EMAIL TO INFORM POTENTIAL PARTICIPANTS OF THE SURVEY

SUBJECT: Invitation to participate in the Human Language Technology and Language Resources Audit 2017/8

MESSAGE:

Dear colleagues

APOLOGIES FOR MULTIPLE POSTINGS

The first Human Language Technology (HLT) Audit in South Africa was conducted in 2009. Eight years later, noting the increased activity in the HLT field, an increase in the number of institutions conducting HLT research and the introduction of digital humanities in South Africa, the Human Language Technology Research Group (HLTRG) at the CSIR Meraka Institute is conducting a follow-up Audit. This Audit is being conducted under the auspices of the South African Centre for Digital Language Resources (SADiLaR) (https://rma.nwu.ac.za/index.php/about-sadilar/)

The objectives of this Audit are:

1. To provide a systematic and detailed inventory of the current HLT components* (software, models) and resources (data) for the official South African languages;

2. To set out the most important dimensions/criteria for the documentation of the HLT components and language resources;

3. To describe the status of the HLT components and language resources in the SA R&D environment for the 11 official languages:

- Which components and resources exist and are freely available?

- Which components and resources are most important for the SA context?

- What differences exist in components and resources across the 11 official languages? and

4. To indicate the gaps between available components/resources and the most important components/resources for the South African context.

The information emanating from the Audit will be collated and displayed on the SADiLaR website, in the form of a catalogue of resources available for download, and a list of resources known to exist but not available for download (to see what is currently available, click here: https://rma.nwu.ac.za/). This information will make it possible for South African language-related research endeavors to be enhanced through collaboration between researchers interested in similar topics, through reusing/repurposing available resources, by extending available resources, and by reducing duplication.

We have tried to follow as inclusive an approach as possible, to identify the institutions and individuals who are involved in developing or managing language technology components and language resources. Your participation is hereby kindly requested and would be much appreciated.

The invitation to participate in the Audit will be sent from “Ilana Wilken <meraka.hlt.audit@gmail.com>”, so please check your spam folder and let us know if you have not received the email. Should you know of someone at your institution or someone you collaborate with who has HLT components/language resources, and who you think should also participate in this Audit, kindly forward this email to them or send an email to iwilken@csir.co.za with their name, surname, email address, telephone number and affiliation.

Members of the RMA and NHN mailing lists:

Should you wish to participate in the Audit, please send your name, surname, email address, telephone number and affiliation to Ilana Wilken (iwilken@csir.co.za), who will send you an invitation to the survey containing a link and a participation token.

If you have any questions do not hesitate to send an email to iwilken@csir.co.za.

Sincerely,

Ilana Wilken

Human Language Technology Research Group

Meraka Institute, CSIR

ANNEXURE F: AUTOMATED FORMAL INVITATION EMAIL

SUBJECT: Invitation to participate in the Human Language Technology and Language Resources Audit 2017/8

MESSAGE:

Dear {FIRSTNAME} {LASTNAME}

You are hereby invited to participate in the Human Language Technology and Language Resources Audit 2017/8.

The link you receive in this email is unique to your participation. We ask that you record each component or resource you have, or manage, separately. (Teams should designate the completion of the survey to an individual please.) You will be able to record up to 100 components/resources*, by clicking on the link each time you want to upload a component/resource. Alternatively, you can enter the token provided below when prompted. Please note that you do not need to recapture components or resources which already appear on the RMA website. Any components or resources you have access to, which do not yet appear on the RMA website, should be captured.

Should you have any questions or experience any challenges, do not hesitate to send an email to iwilken@csir.co.za.

The due date for completion of the survey is 28 February 2018.

Your unique token for this survey is: {TOKEN}

Click here to complete the survey: {SURVEYURL}

* The definitions for the components mentioned in the survey can be viewed here: https://sites.google.com/view/hlt-audit-definitions/home

Background to the Audit

The first Human Language Technology (HLT) Audit in South Africa was conducted in 2009. Eight years later, noting the increased activity in the HLT field, an increase in the number of institutions conducting HLT research and the introduction of digital humanities in South Africa, the Human Language Technology Research Group (HLTRG) at the CSIR Meraka Institute is conducting a follow-up Audit. This Audit is being conducted under the auspices of the South African Centre for Digital Language Resources (SADiLaR) (https://rma.nwu.ac.za/index.php/about-sadilar/)

The objectives of this Audit are:

1. To provide a systematic and detailed inventory of the current HLT components (software, models) and resources (data) for the official South African languages;

2. To set out the most important dimensions/criteria for the documentation of the HLT components and language resources;

3. To describe the status of the HLT components and language resources in the SA R&D environment for the 11 official languages:

- Which components and resources exist and are freely available?

- Which components and resources are most important for the SA context?

- What differences exist in components and resources across the 11 official languages? and

4. To indicate the gaps between available components/resources and the most important components/resources for the South African context.

The information emanating from the Audit will be collated and displayed on the SADiLaR website, in the form of a catalogue of resources available for download, and a list of resources known to exist but not available for download (to see what is currently available, click here: https://rma.nwu.ac.za/). This information will make it possible for South African language-related research endeavors to be enhanced through collaboration between researchers interested in similar topics, through reusing/repurposing available resources, by extending available resources, and by reducing duplication.

Sincerely,

Ilana Wilken

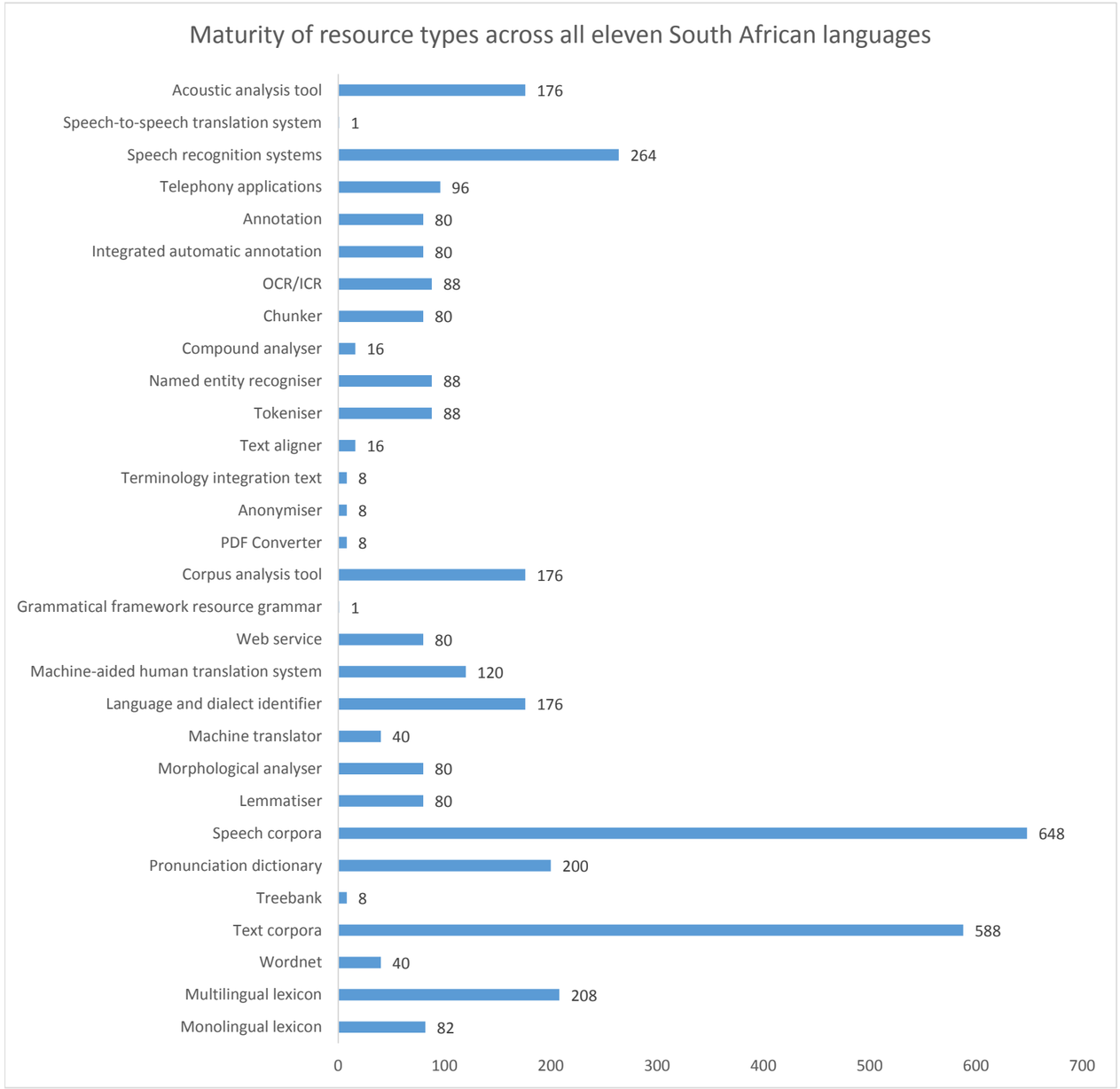
Human Language Technology Research Group

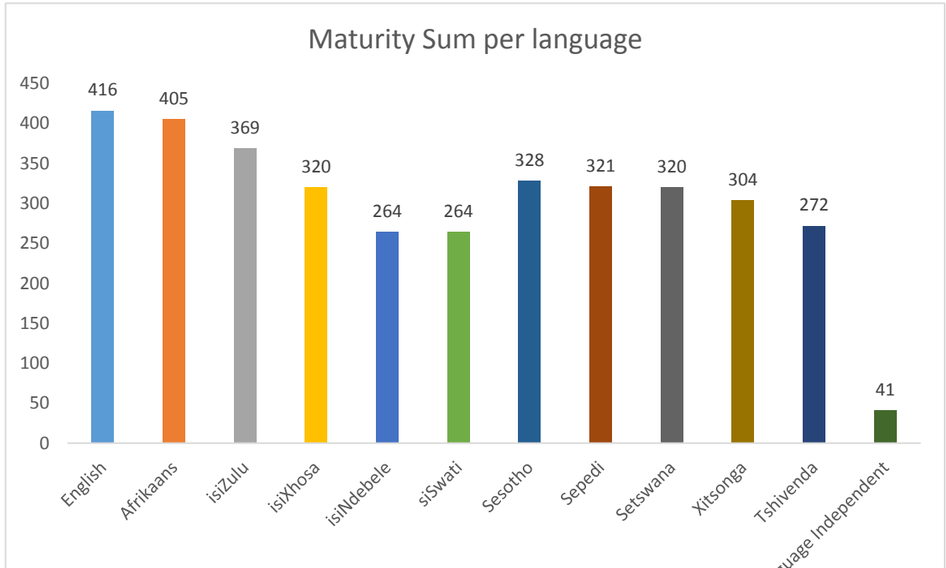
Meraka Institute, CSIR

MATURITY SUM

		English		Afrikaans		isiZulu		isiXhosa		isiNdebele		siSwati		Sesotho		Sepedi		Setswana		Xitsonga		Tshivenda		Language Independent		SUB TOTAL	TOTAL
DATA: TEXT	Monolingual lexicon	8	1.9%	1	0.2%	1	0.3%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	16	5.3%	8	2.9%	0	0.0%	**82**	**2.3%**
	Multilingual lexicon	64	15.4%	32	7.9%	24	6.5%	16	5.0%	8	3.0%	8	3.0%	8	2.4%	24	7.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**208**	**5.7%**
	Wordnet	0	0.0%	0	0.0%	8	2.2%	8	2.5%	0	0.0%	0	0.0%	0	0.0%	8	2.5%	8	2.5%	0	0.0%	8	2.9%	0	0.0%	**40**	**1.1%**
	Text corpora	72	17.3%	60	14.8%	64	17.3%	56	17.5%	40	15.2%	40	15.2%	48	14.6%	48	15.0%	56	17.5%	64	21.1%	40	14.7%	0	0.0%	**588**	**16.2%**
	Treebank	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	2.5%	0	0.0%	0	0.0%	0	0.0%	**8**	**0.2%**
DATA: SPEECH	Pronunciation dictionary	16	3.8%	16	4.0%	16	4.3%	16	5.0%	16	6.1%	16	6.1%	24	7.3%	24	7.5%	24	7.5%	16	5.3%	16	5.9%	0	0.0%	**200**	**5.5%**
	Speech corpora	128	30.8%	104	25.7%	80	21.7%	56	17.5%	32	12.1%	32	12.1%	80	24.4%	32	10.0%	40	12.5%	32	10.5%	32	11.8%	0	0.0%	**648**	**17.9%**
SOFTWARE: TEXT	Lemmatiser	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
	Morphological analyser	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
	Machine translator	0	0.0%	8	2.0%	8	2.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	2.5%	8	2.5%	8	2.6%	0	0.0%	0	0.0%	**40**	**1.1%**
	Language and dialect identifier	16	3.8%	16	4.0%	16	4.3%	16	5.0%	16	6.1%	16	6.1%	16	4.9%	16	5.0%	16	5.0%	16	5.3%	16	5.9%	0	0.0%	**176**	**4.9%**
	Machine-aided human translation system	16	3.8%	16	4.0%	16	4.3%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	16	5.0%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**120**	**3.3%**
	Web service	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
	Grammatical framework resource gramma	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	0.3%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	**1**	**0.0%**
	Corpus analysis tool	16	3.8%	16	4.0%	16	4.3%	16	5.0%	16	6.1%	16	6.1%	16	4.9%	16	5.0%	16	5.0%	16	5.3%	16	5.9%	0	0.0%	**176**	**4.9%**
	PDF Converter	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	19.5%	**8**	**0.2%**
	Anonymiser	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	19.5%	**8**	**0.2%**
	Terminology integration text	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	19.5%	**8**	**0.2%**
	Text aligner	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	16	39.0%	**16**	**0.4%**
	Tokeniser	8	1.9%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**88**	**2.4%**
	Named entity recogniser	8	1.9%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**88**	**2.4%**
	Compound analyser	0	0.0%	16	4.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	**16**	**0.4%**
	Chunker	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
	OCR/ICR	8	1.9%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**88**	**2.4%**
	Integrated automatic annotation	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
	Annotation	0	0.0%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**80**	**2.2%**
SOFTWARE: SPEECH	Telephony applications	16	3.8%	8	2.0%	8	2.2%	8	2.5%	8	3.0%	8	3.0%	8	2.4%	8	2.5%	8	2.5%	8	2.6%	8	2.9%	0	0.0%	**96**	**2.6%**
	Speech recognition systems	24	5.8%	24	5.9%	24	6.5%	24	7.5%	24	9.1%	24	9.1%	24	7.3%	24	7.5%	24	7.5%	24	7.9%	24	8.8%	0	0.0%	**264**	**7.3%**
	Speech-to-speech translation system	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	2.4%	**1**	**0.0%**
	Acoustic analysis tool	16	3.8%	16	4.0%	16	4.3%	16	5.0%	16	6.1%	16	6.1%	16	4.9%	16	5.0%	16	5.0%	16	5.3%	16	5.9%	0	0.0%	**176**	**4.9%**
Total		**416**	**100.0%**	**405**	**100.0%**	**369**	**100.0%**	**320**	**100.0%**	**264**	**100.0%**	**264**	**100.0%**	**328**	**100.0%**	**321**	**100.0%**	**320**	**100.0%**	**304**	**100.0%**	**272**	**100.0%**	**41**	**100.0%**	**3624**	**100.0%**

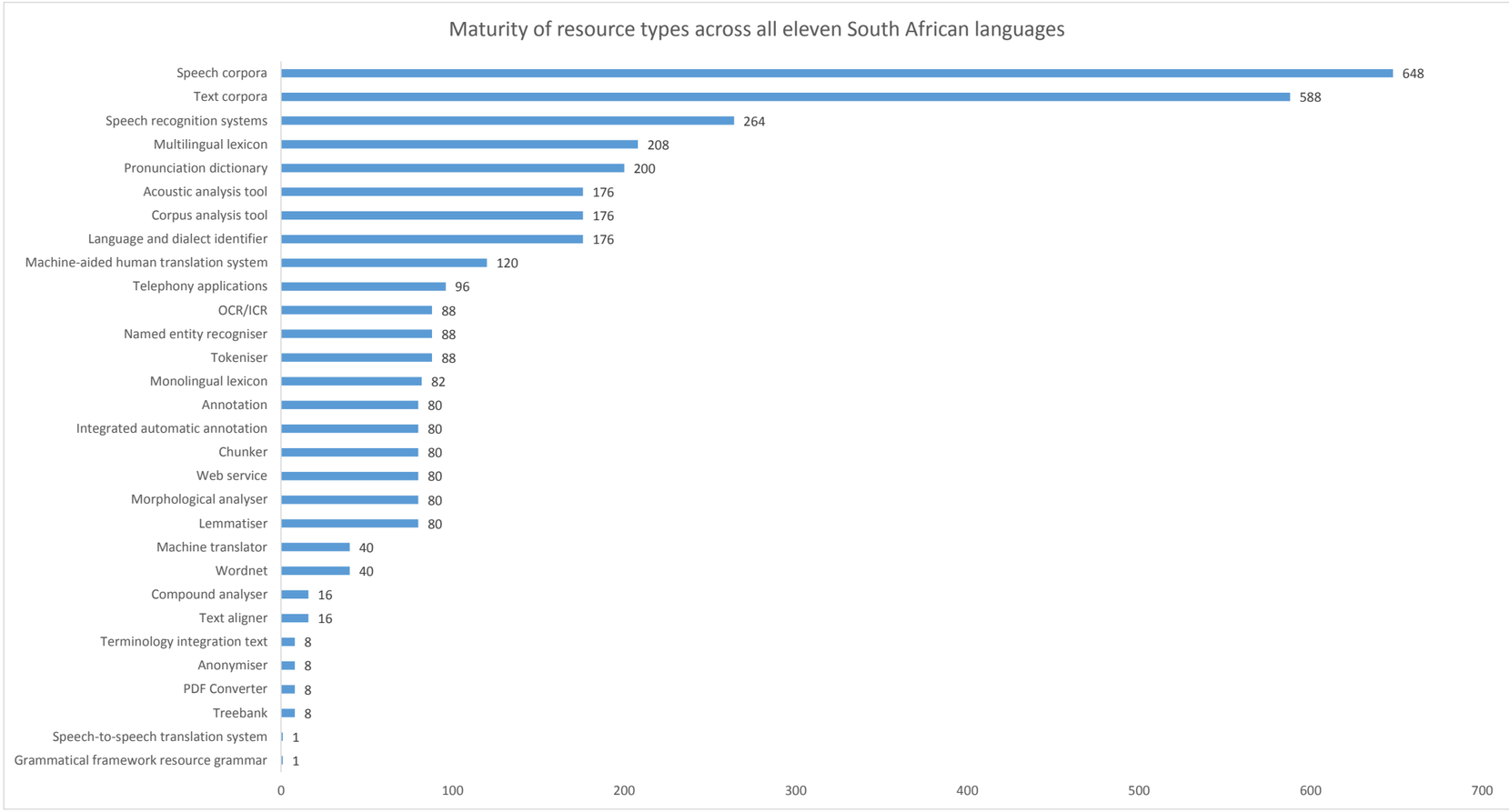
GRAPHS ILLUSTRATING THE MATURITY OF RESOURCES IN SOUTH AFRICA



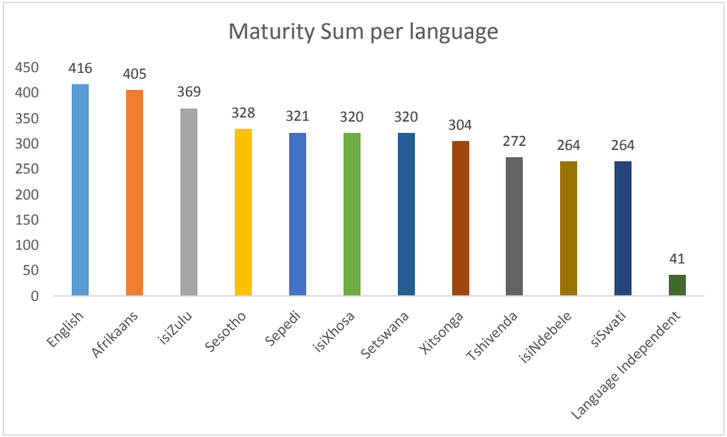


GRAPHS ILLUSTRATING THE MATURITY OF RESOURCES IN SOUTH AFRICA (ORDERED)

Resource Type	Maturity Sum
Grammatical framework resource grammar	1
Speech-to-speech translation system	1
Treebank	8
PDF Converter	8
Anonymiser	8
Terminology integration text	8
Text aligner	16
Compound analyser	16
Wordnet	40
Machine translator	40
Lemmatiser	80
Morphological analyser	80
Web service	80
Chunker	80
Integrated automatic annotation	80
Annotation	80
Monolingual lexicon	82
Tokeniser	88
Named entity recogniser	88
OCR/ICR	88
Telephony applications	96
Machine-aided human translation system	120
Language and dialect identifier	176
Corpus analysis tool	176
Acoustic analysis tool	176
Pronunciation dictionary	200
Multilingual lexicon	208
Speech recognition systems	264
Text corpora	588
Speech corpora	648

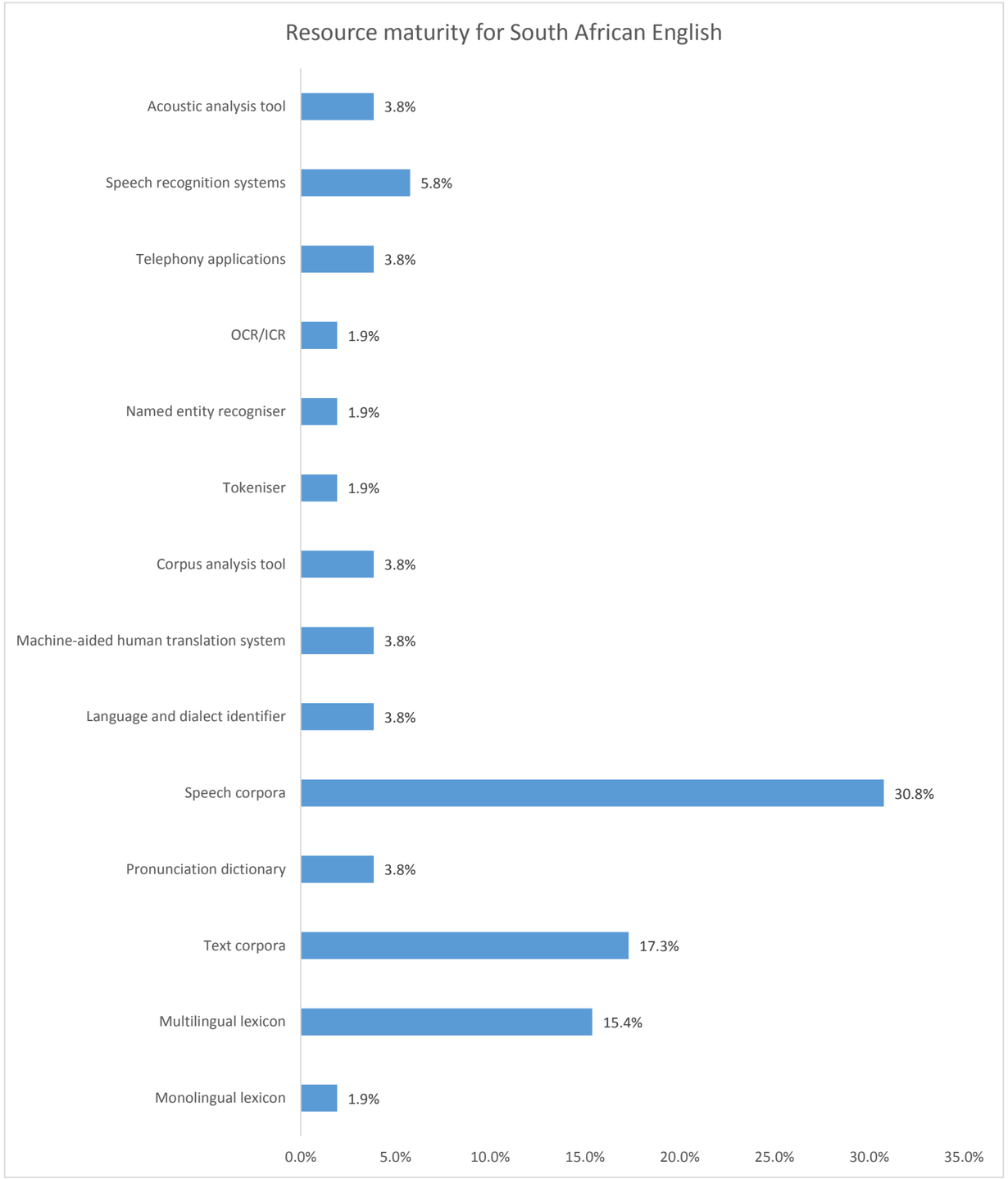


Language	Maturity Sum
English	416
Afrikaans	405
isiZulu	369
Sesotho	328
Sepedi	321
isiXhosa	320
Setswana	320
Xitsonga	304
Tshivenda	272
isiNdebele	264
siSwati	264
Language Independent	41



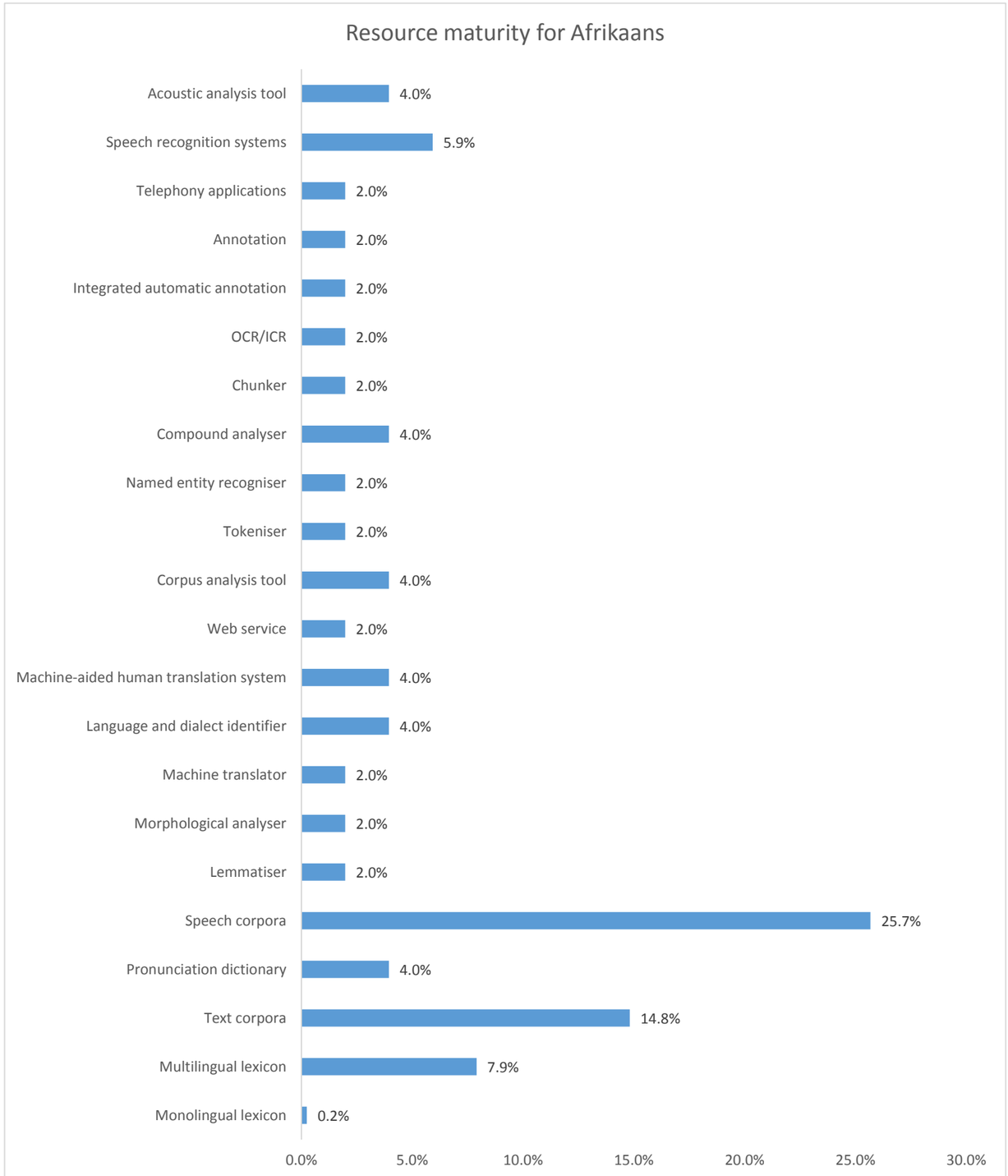
MATURITY SUM PER LANGUAGE

		English	
DATA: TEXT	Monolingual lexicon	8	1.9%
	Multilingual lexicon	64	15.4%
	Text corpora	72	17.3%
DATA: SPEECH	Pronunciation dictionary	16	3.8%
	Speech corpora	128	30.8%
SOFTWARE: TEXT	Language and dialect identifier	16	3.8%
	Machine-aided human translation system	16	3.8%
	Corpus analysis tool	16	3.8%
	Tokeniser	8	1.9%
	Named entity recogniser	8	1.9%
	OCR/ICR	8	1.9%
SOFTWARE: SPEECH	Telephony applications	16	3.8%
	Speech recognition systems	24	5.8%
	Acoustic analysis tool	16	3.8%
	Total	**416**	**100.0%**



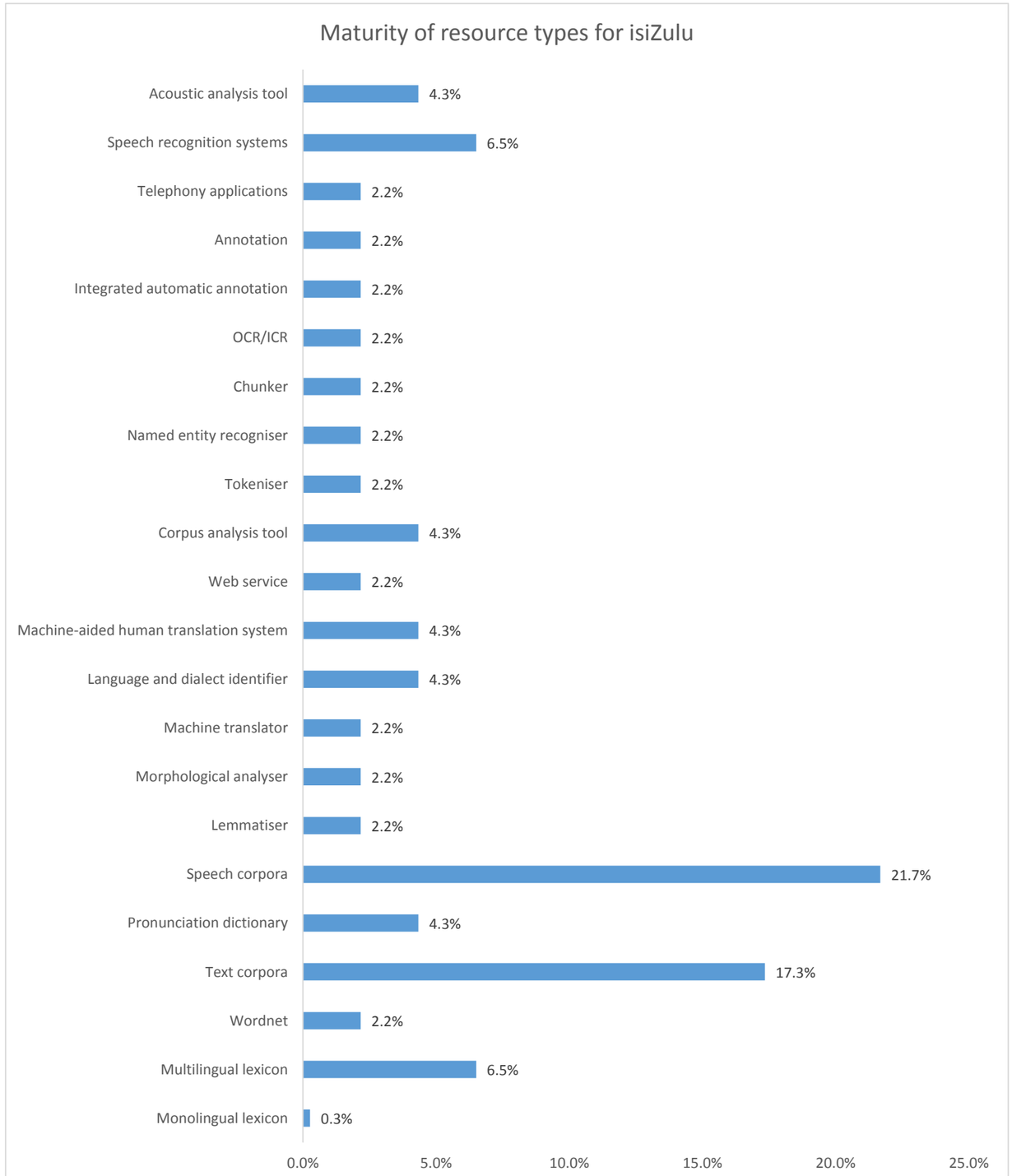
MATURITY SUM PER LANGUAGE

		Afrikaans	
DATA: TEXT	Monolingual lexicon	1	0.2%
	Multilingual lexicon	32	7.9%
	Text corpora	60	14.8%
DATA: SPEECH	Pronunciation dictionary	16	4.0%
	Speech corpora	104	25.7%
SOFTWARE: TEXT	Lemmatiser	8	2.0%
	Morphological analyser	8	2.0%
	Machine translator	8	2.0%
	Language and dialect identifier	16	4.0%
	Machine-aided human translation system	16	4.0%
	Web service	8	2.0%
	Corpus analysis tool	16	4.0%
	Tokeniser	8	2.0%
	Named entity recogniser	8	2.0%
	Compound analyser	16	4.0%
	Chunker	8	2.0%
	OCR/ICR	8	2.0%
	Integrated automatic annotation	8	2.0%
	Annotation	8	2.0%
SOFTWARE: SPEECH	Telephony applications	8	2.0%
	Speech recognition systems	24	5.9%
	Acoustic analysis tool	16	4.0%
	Total	**405**	**100.0%**



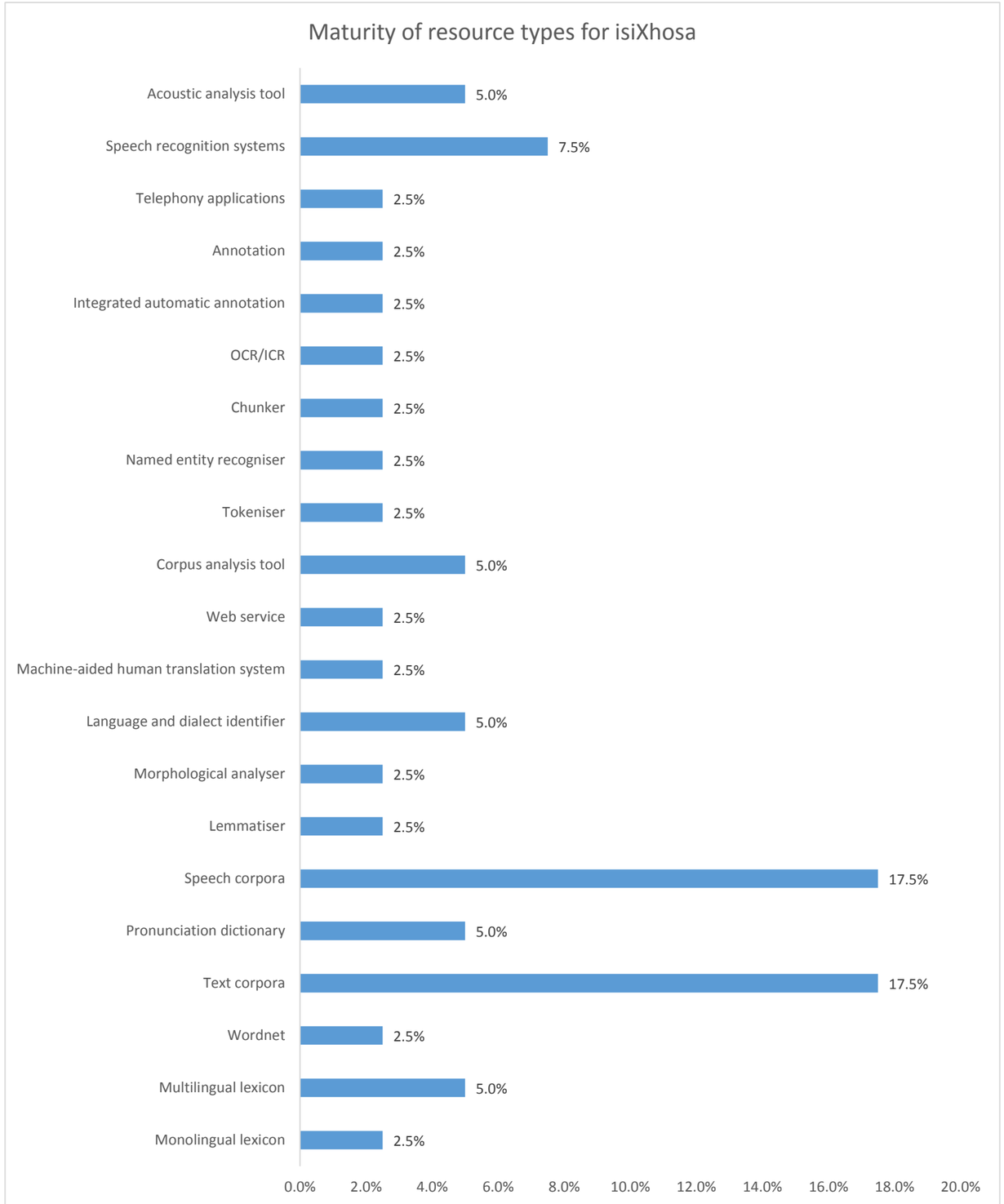
MATURITY SUM PER LANGUAGE

		isiZulu	
DATA: TEXT	Monolingual lexicon	1	0.3%
	Multilingual lexicon	24	6.5%
	Wordnet	8	2.2%
	Text corpora	64	17.3%
DATA: SPEECH	Pronunciation dictionary	16	4.3%
	Speech corpora	80	21.7%
SOFTWARE: TEXT	Lemmatiser	8	2.2%
	Morphological analyser	8	2.2%
	Machine translator	8	2.2%
	Language and dialect identifier	16	4.3%
	Machine-aided human translation system	16	4.3%
	Web service	8	2.2%
	Corpus analysis tool	16	4.3%
	Tokeniser	8	2.2%
	Named entity recogniser	8	2.2%
	Chunker	8	2.2%
	OCR/ICR	8	2.2%
	Integrated automatic annotation	8	2.2%
	Annotation	8	2.2%
SOFTWARE: SPEECH	Telephony applications	8	2.2%
	Speech recognition systems	24	6.5%
	Acoustic analysis tool	16	4.3%
	Total	**369**	**100.0%**



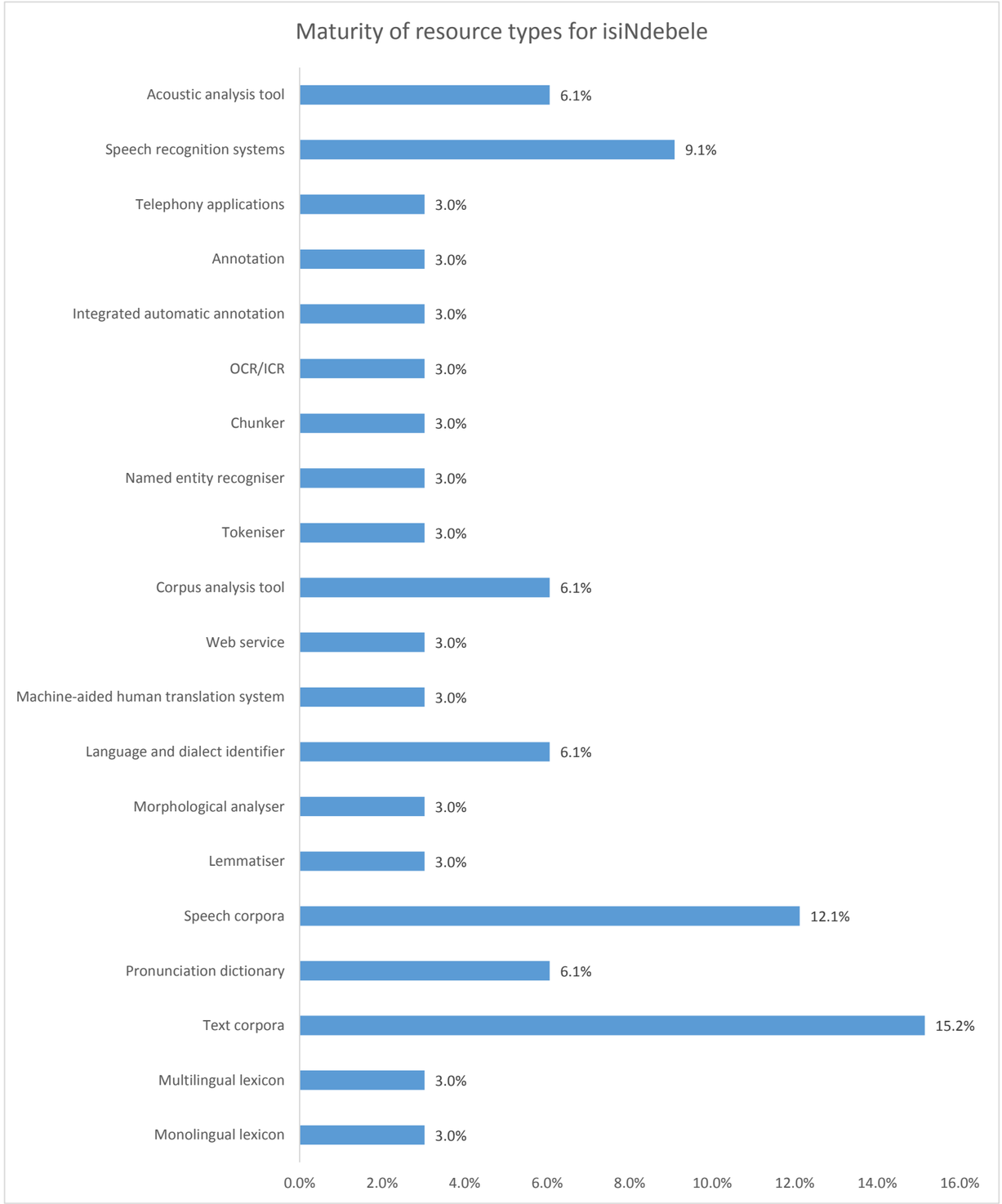
MATURITY SUM PER LANGUAGE

		isiXhosa	
DATA: TEXT	Monolingual lexicon	8	2.5%
	Multilingual lexicon	16	5.0%
	Wordnet	8	2.5%
	Text corpora	56	17.5%
DATA: SPEECH	Pronunciation dictionary	16	5.0%
	Speech corpora	56	17.5%
SOFTWARE: TEXT	Lemmatiser	8	2.5%
	Morphological analyser	8	2.5%
	Language and dialect identifier	16	5.0%
	Machine-aided human translation system	8	2.5%
	Web service	8	2.5%
	Corpus analysis tool	16	5.0%
	Tokeniser	8	2.5%
	Named entity recogniser	8	2.5%
	Chunker	8	2.5%
	OCR/ICR	8	2.5%
	Integrated automatic annotation	8	2.5%
	Annotation	8	2.5%
SOFTWARE: SPEECH	Telephony applications	8	2.5%
	Speech recognition systems	24	7.5%
	Acoustic analysis tool	16	5.0%
	Total	**320**	**100.0%**



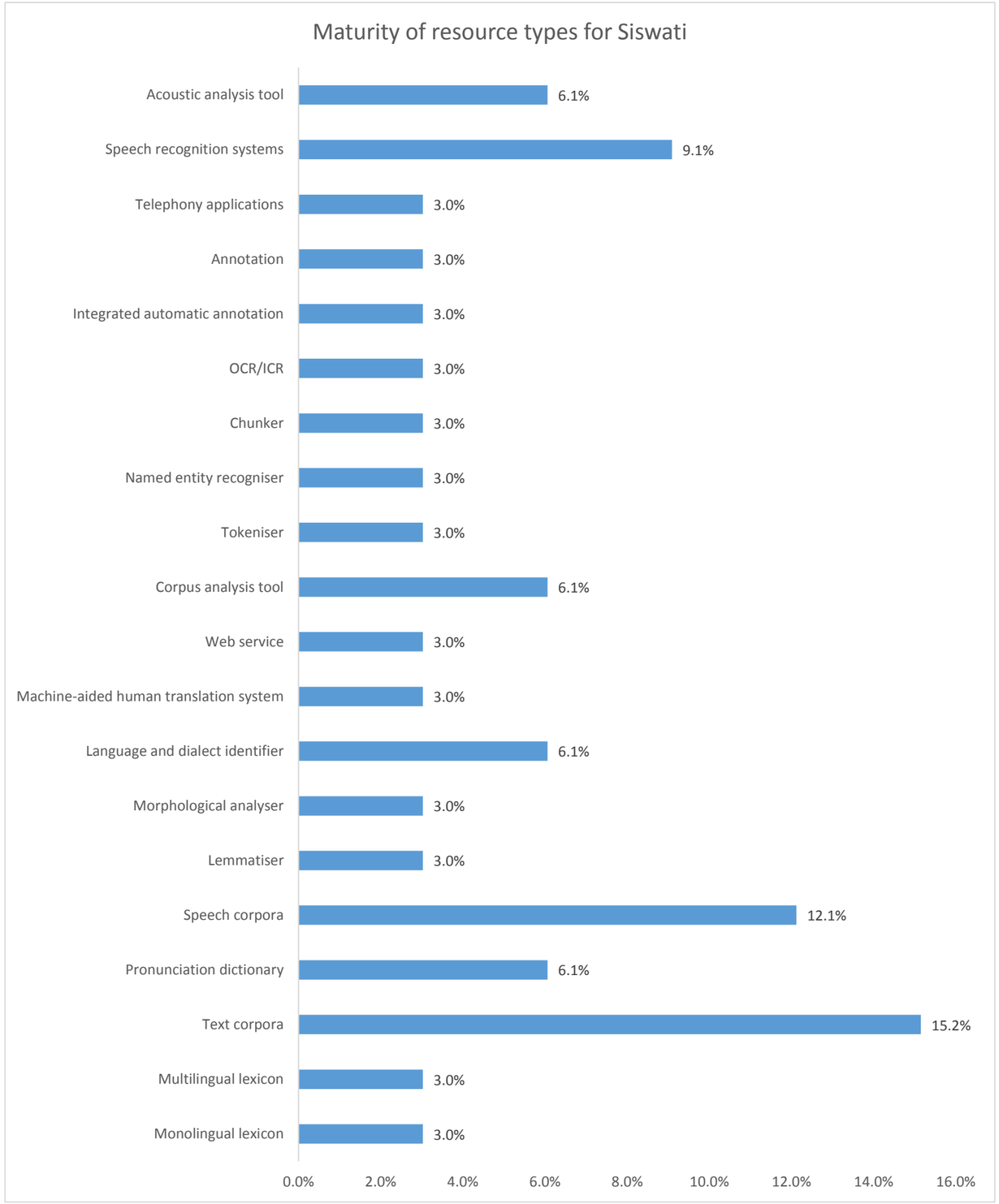
MATURITY SUM PER LANGUAGE

		isiNdebele	
DATA: TEXT	Monolingual lexicon	8	3.0%
	Multilingual lexicon	8	3.0%
	Text corpora	40	15.2%
DATA: SPEECH	Pronunciation dictionary	16	6.1%
	Speech corpora	32	12.1%
SOFTWARE: TEXT	Lemmatiser	8	3.0%
	Morphological analyser	8	3.0%
	Language and dialect identifier	16	6.1%
	Machine-aided human translation system	8	3.0%
	Web service	8	3.0%
	Corpus analysis tool	16	6.1%
	Tokeniser	8	3.0%
	Named entity recogniser	8	3.0%
	Chunker	8	3.0%
	OCR/ICR	8	3.0%
	Integrated automatic annotation	8	3.0%
	Annotation	8	3.0%
SOFTWARE: SPEECH	Telephony applications	8	3.0%
	Speech recognition systems	24	9.1%
	Acoustic analysis tool	16	6.1%
	Total	**264**	**100.0%**



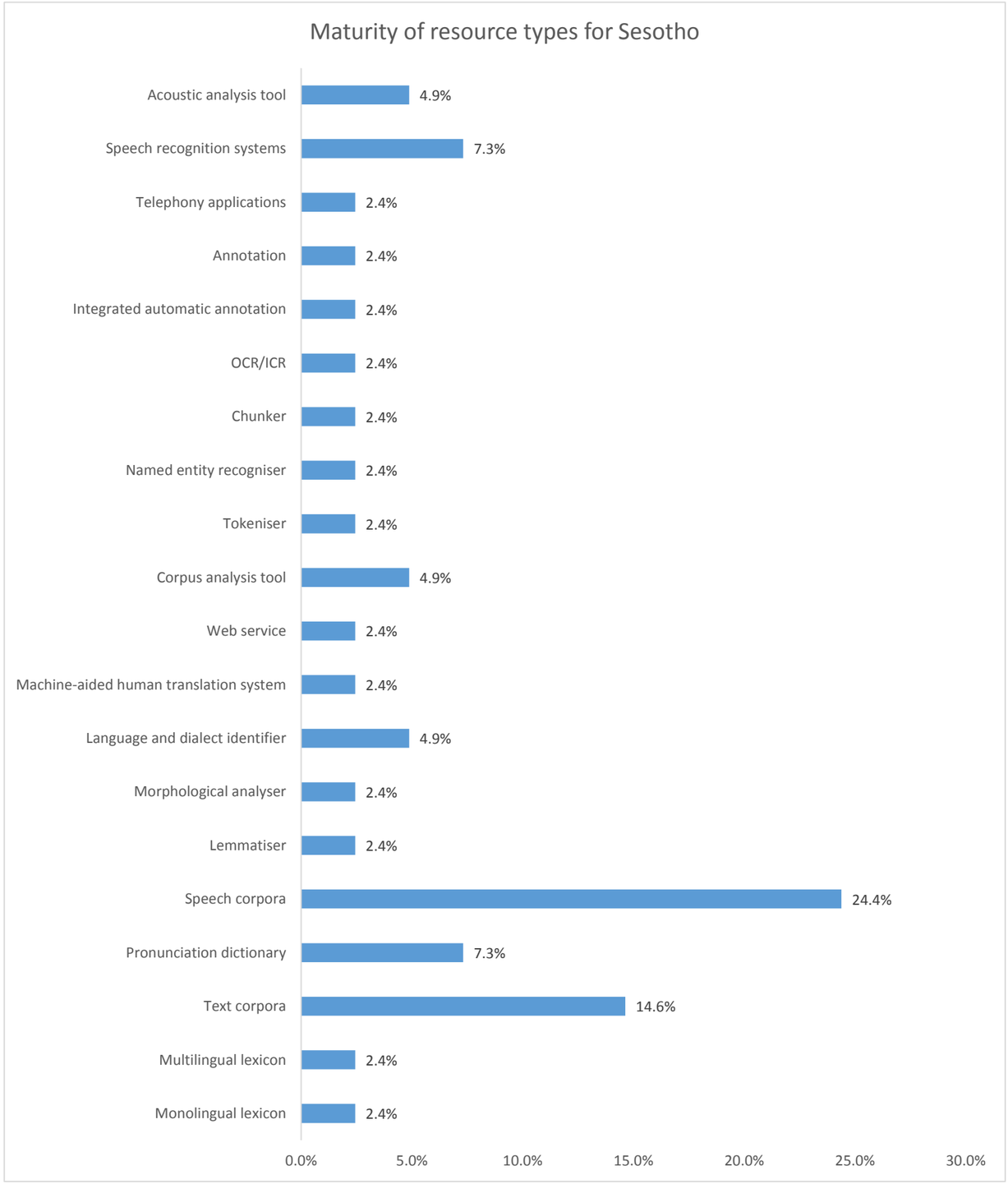
MATURITY SUM PER LANGUAGE

		siSwati	
DATA: TEXT	Monolingual lexicon	8	3.0%
	Multilingual lexicon	8	3.0%
	Text corpora	40	15.2%
DATA: SPEECH	Pronunciation dictionary	16	6.1%
	Speech corpora	32	12.1%
SOFTWARE: TEXT	Lemmatiser	8	3.0%
	Morphological analyser	8	3.0%
	Language and dialect identifier	16	6.1%
	Machine-aided human translation system	8	3.0%
	Web service	8	3.0%
	Corpus analysis tool	16	6.1%
	Tokeniser	8	3.0%
	Named entity recogniser	8	3.0%
	Chunker	8	3.0%
	OCR/ICR	8	3.0%
	Integrated automatic annotation	8	3.0%
	Annotation	8	3.0%
SOFTWARE: SPEECH	Telephony applications	8	3.0%
	Speech recognition systems	24	9.1%
	Acoustic analysis tool	16	6.1%
	Total	**264**	**100.0%**



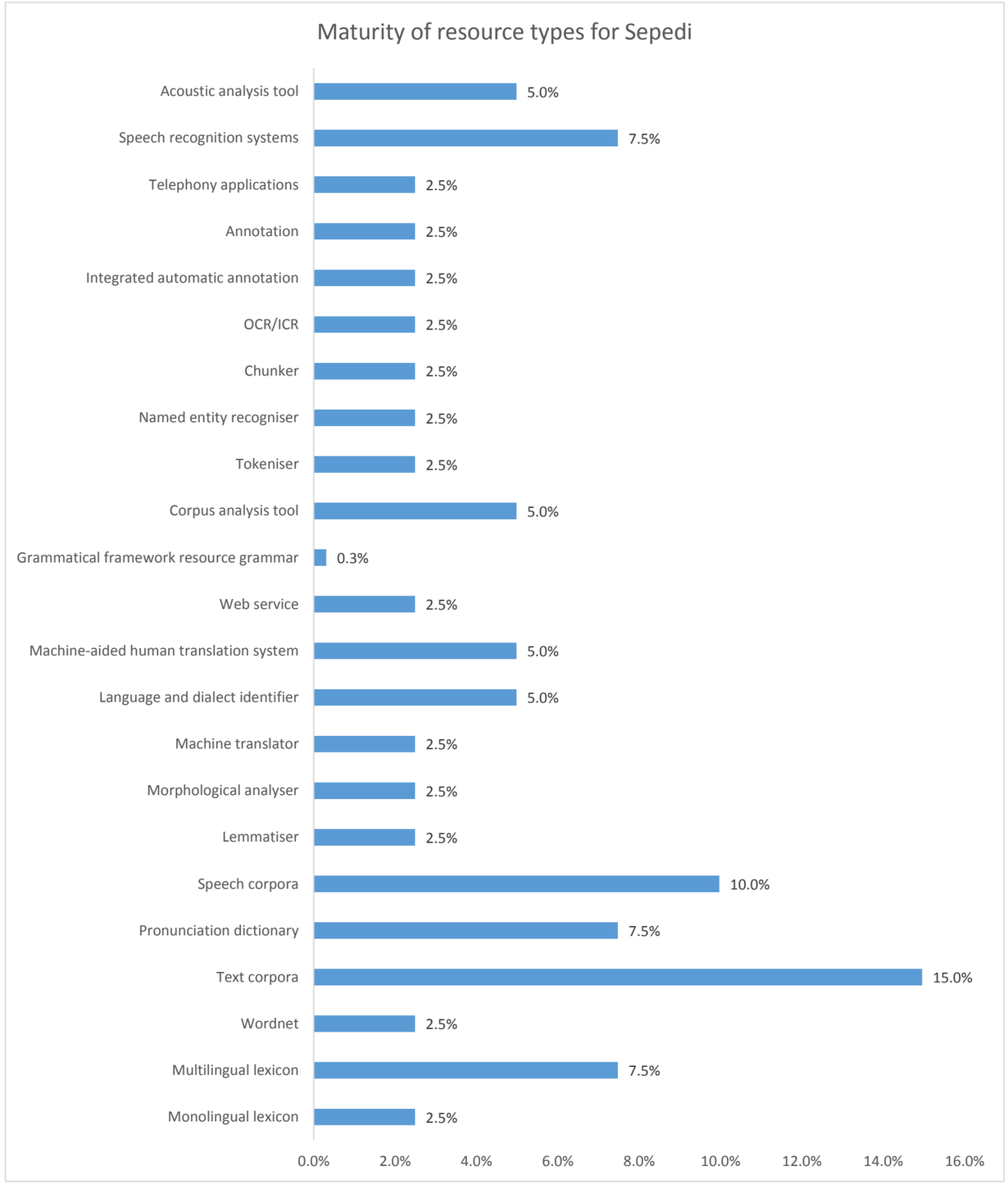
MATURITY SUM PER LANGUAGE

		Sesotho	
DATA: TEXT	Monolingual lexicon	8	2.4%
	Multilingual lexicon	8	2.4%
	Text corpora	48	14.6%
DATA: SPEECH	Pronunciation dictionary	24	7.3%
	Speech corpora	80	24.4%
SOFTWARE: TEXT	Lemmatiser	8	2.4%
	Morphological analyser	8	2.4%
	Language and dialect identifier	16	4.9%
	Machine-aided human translation system	8	2.4%
	Web service	8	2.4%
	Corpus analysis tool	16	4.9%
	Tokeniser	8	2.4%
	Named entity recogniser	8	2.4%
	Chunker	8	2.4%
	OCR/ICR	8	2.4%
	Integrated automatic annotation	8	2.4%
	Annotation	8	2.4%
SOFTWARE: SPEECH	Telephony applications	8	2.4%
	Speech recognition systems	24	7.3%
	Acoustic analysis tool	16	4.9%
	Total	**328**	**100.0%**



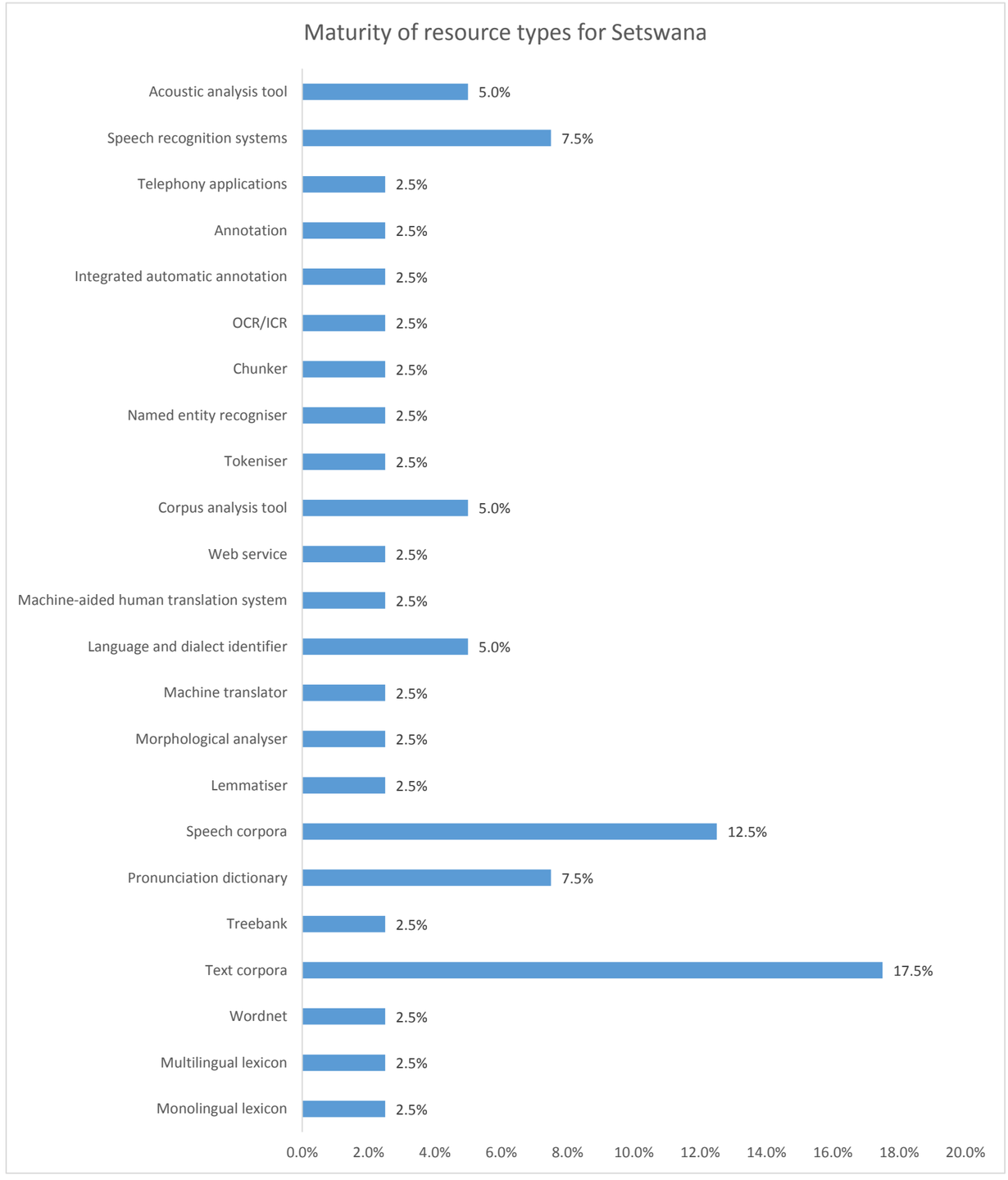
MATURITY SUM PER LANGUAGE

		Sepedi	
DATA: TEXT	Monolingual lexicon	8	2.5%
	Multilingual lexicon	24	7.5%
	Wordnet	8	2.5%
	Text corpora	48	15.0%
DATA: SPEECH	Pronunciation dictionary	24	7.5%
	Speech corpora	32	10.0%
SOFTWARE: TEXT	Lemmatiser	8	2.5%
	Morphological analyser	8	2.5%
	Machine translator	8	2.5%
	Language and dialect identifier	16	5.0%
	Machine-aided human translation system	16	5.0%
	Web service	8	2.5%
	Grammatical framework resource grammar	1	0.3%
	Corpus analysis tool	16	5.0%
	Tokeniser	8	2.5%
	Named entity recogniser	8	2.5%
	Chunker	8	2.5%
	OCR/ICR	8	2.5%
	Integrated automatic annotation	8	2.5%
	Annotation	8	2.5%
SOFTWARE: SPEECH	Telephony applications	8	2.5%
	Speech recognition systems	24	7.5%
	Acoustic analysis tool	16	5.0%
	Total	**321**	**100.0%**



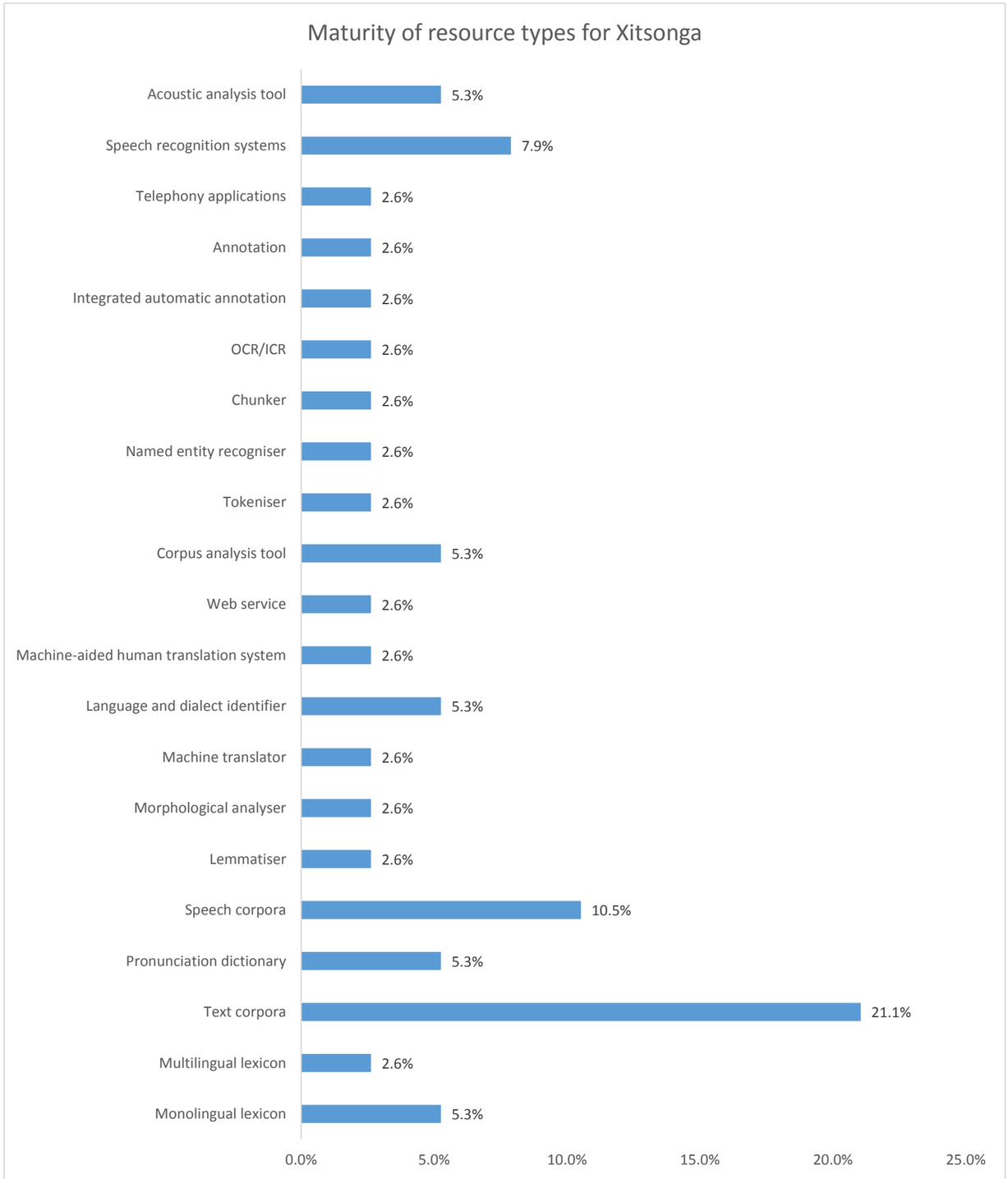
MATURITY SUM PER LANGUAGE

		Setswana	
DATA: TEXT	Monolingual lexicon	8	2.5%
	Multilingual lexicon	8	2.5%
	Wordnet	8	2.5%
	Text corpora	56	17.5%
	Treebank	8	2.5%
DATA: SPEECH	Pronunciation dictionary	24	7.5%
	Speech corpora	40	12.5%
SOFTWARE: TEXT	Lemmatiser	8	2.5%
	Morphological analyser	8	2.5%
	Machine translator	8	2.5%
	Language and dialect identifier	16	5.0%
	Machine-aided human translation system	8	2.5%
	Web service	8	2.5%
	Corpus analysis tool	16	5.0%
	Tokeniser	8	2.5%
	Named entity recogniser	8	2.5%
	Chunker	8	2.5%
	OCR/ICR	8	2.5%
	Integrated automatic annotation	8	2.5%
	Annotation	8	2.5%
SOFTWARE: SPEECH	Telephony applications	8	2.5%
	Speech recognition systems	24	7.5%
	Acoustic analysis tool	16	5.0%
	Total	**320**	**100.0%**



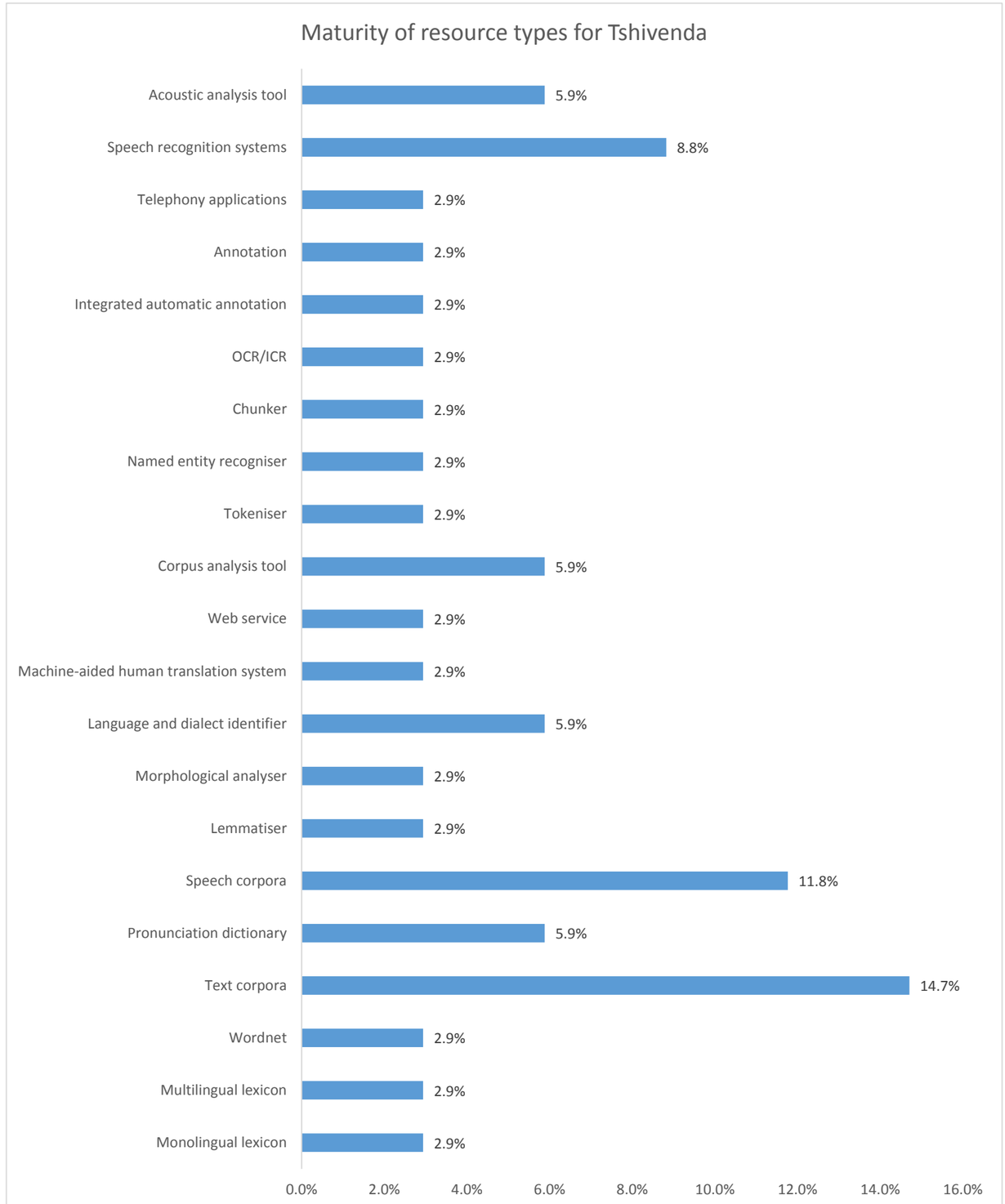
MATURITY SUM PER LANGUAGE

		Xitsonga	
DATA: TEXT	Monolingual lexicon	16	5.3%
	Multilingual lexicon	8	2.6%
	Text corpora	64	21.1%
DATA: SPEECH	Pronunciation dictionary	16	5.3%
	Speech corpora	32	10.5%
SOFTWARE: TEXT	Lemmatiser	8	2.6%
	Morphological analyser	8	2.6%
	Machine translator	8	2.6%
	Language and dialect identifier	16	5.3%
	Machine-aided human translation system	8	2.6%
	Web service	8	2.6%
	Corpus analysis tool	16	5.3%
	Tokeniser	8	2.6%
	Named entity recogniser	8	2.6%
	Chunker	8	2.6%
	OCR/ICR	8	2.6%
	Integrated automatic annotation	8	2.6%
	Annotation	8	2.6%
SOFTWARE: SPEECH	Telephony applications	8	2.6%
	Speech recognition systems	24	7.9%
	Acoustic analysis tool	16	5.3%
	Total	**304**	**100.0%**



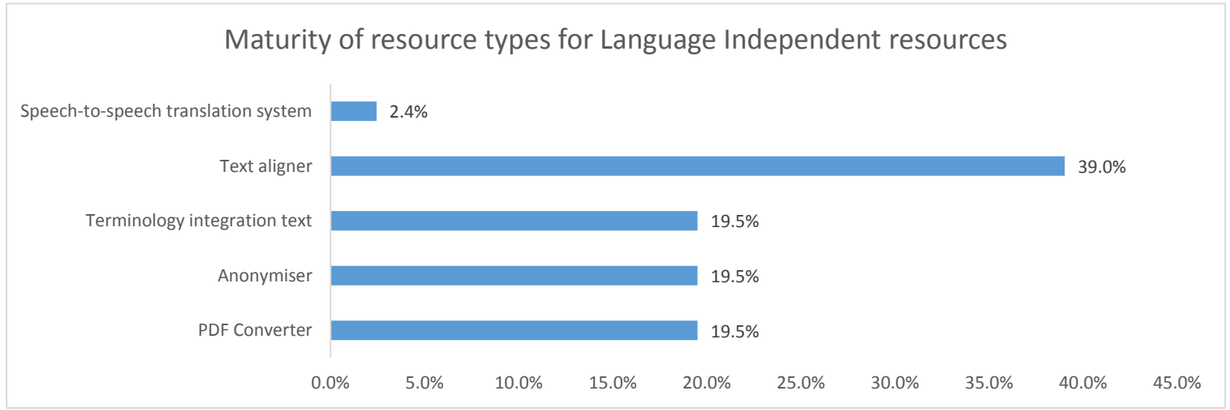
MATURITY SUM PER LANGUAGE

		Tshivenda	
DATA: TEXT	Monolingual lexicon	8	2.9%
	Multilingual lexicon	8	2.9%
	Wordnet	8	2.9%
	Text corpora	40	14.7%
DATA: SPEECH	Pronunciation dictionary	16	5.9%
	Speech corpora	32	11.8%
SOFTWARE: TEXT	Lemmatiser	8	2.9%
	Morphological analyser	8	2.9%
	Language and dialect identifier	16	5.9%
	Machine-aided human translation system	8	2.9%
	Web service	8	2.9%
	Corpus analysis tool	16	5.9%
	Tokeniser	8	2.9%
	Named entity recogniser	8	2.9%
	Chunker	8	2.9%
	OCR/ICR	8	2.9%
	Integrated automatic annotation	8	2.9%
	Annotation	8	2.9%
SOFTWARE: SPEECH	Telephony applications	8	2.9%
	Speech recognition systems	24	8.8%
	Acoustic analysis tool	16	5.9%
	Total	**272**	**100.0%**



MATURITY SUM PER LANGUAGE

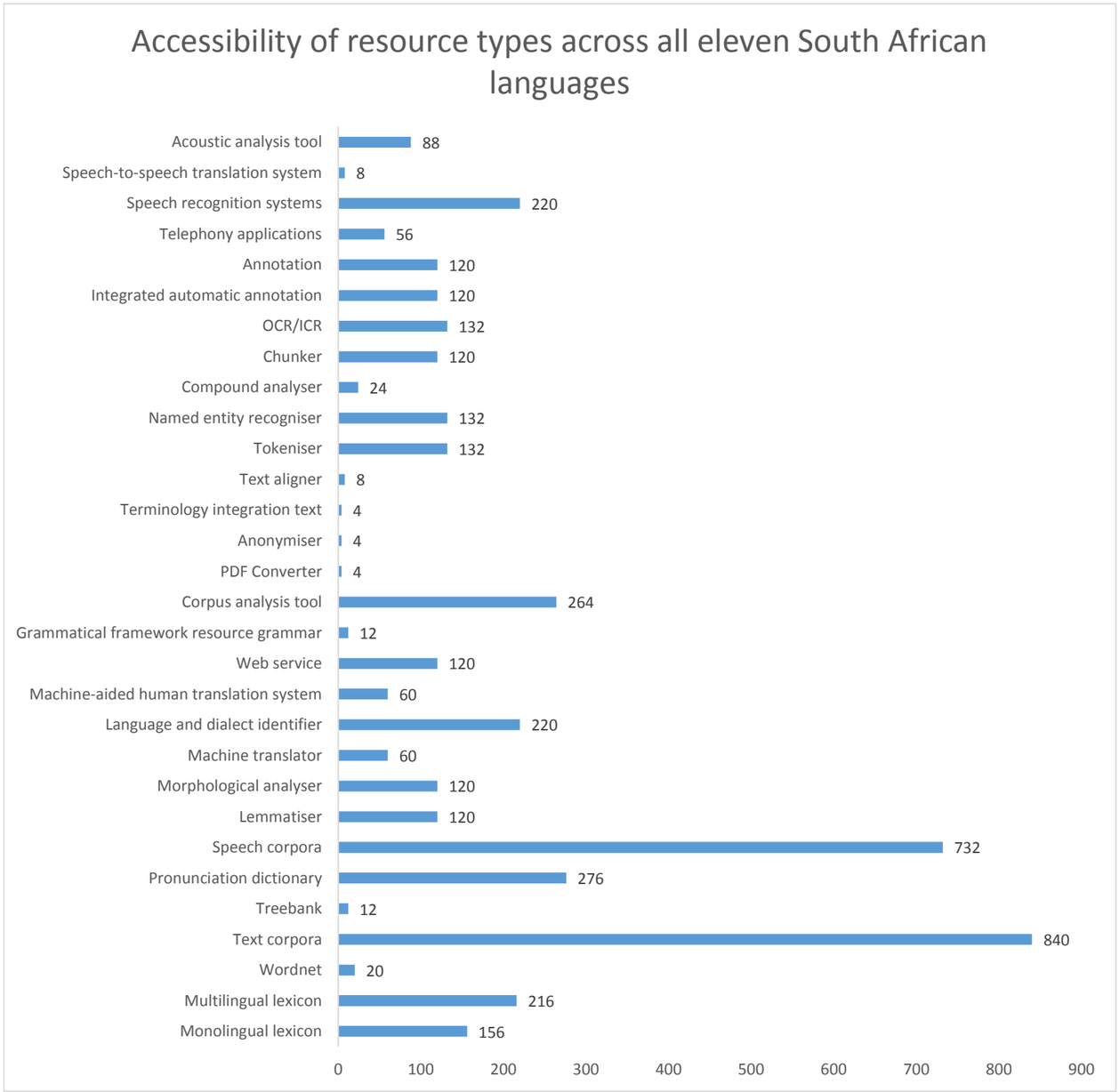
		Lang. Indep.	
SOFTWARE: TEXT	PDF Converter	8	19.5%
	Anonymiser	8	19.5%
	Terminology integration text	8	19.5%
	Text aligner	16	39.0%
SOFTWARE: SPEECH	Speech-to-speech translation system	1	2.4%
	Total	**41**	**100.0%**

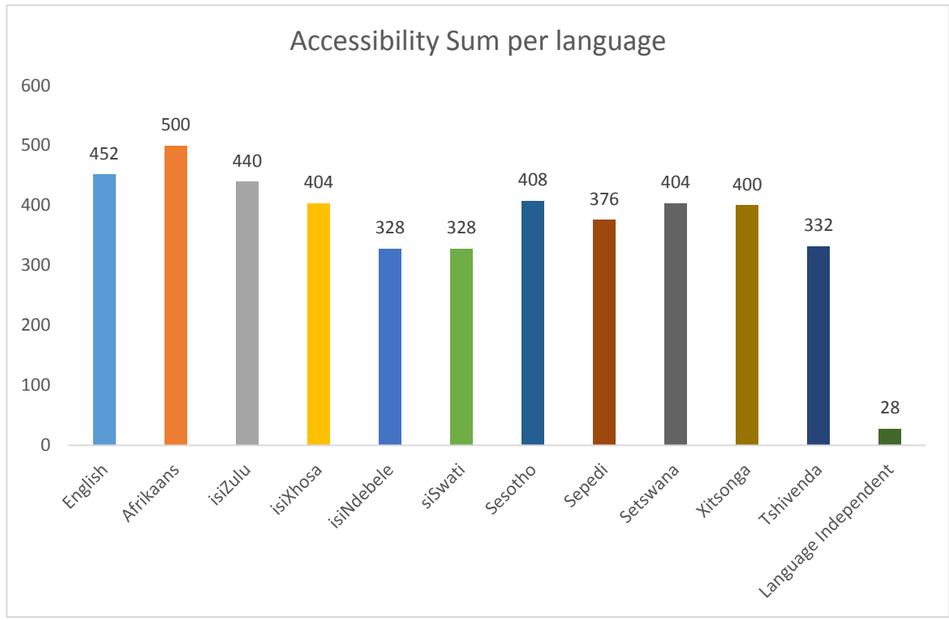


ACCESSIBILITY SUM

		English		Afrikaans		isiZulu		isiXhosa		isiNdebele		siSwati		Sesotho		Sepedi		Setswana		Xitsonga		Tshivenda		Language Independent		SUB TOTAL	TOTAL
DATA: TEXT	Monolingual lexicon	12	2.7%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	36	9.0%	12	3.6%	0	0.0%	**156**	**3.5%**
	Multilingual lexicon	48	10.6%	32	6.4%	20	4.5%	24	5.9%	12	3.7%	12	3.7%	12	2.9%	20	5.3%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**216**	**4.9%**
	Wordnet	0	0.0%	0	0.0%	4	0.9%	4	1.0%	0	0.0%	0	0.0%	0	0.0%	4	1.1%	4	1.0%	0	0.0%	4	1.2%	0	0.0%	**20**	**0.5%**
	Text corpora	84	18.6%	88	17.6%	88	20.0%	84	20.8%	60	18.3%	60	18.3%	72	17.6%	64	17.0%	84	20.8%	96	24.0%	60	18.1%	0	0.0%	**840**	**19.1%**
	Treebank	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	12	3.0%	0	0.0%	0	0.0%	0	0.0%	**12**	**0.3%**
DATA: SPEECH	Pronunciation dictionary	24	5.3%	24	4.8%	24	5.5%	24	5.9%	24	7.3%	24	7.3%	28	6.9%	28	7.4%	28	6.9%	24	6.0%	24	7.2%	0	0.0%	**276**	**6.3%**
	Speech corpora	152	33.6%	116	23.2%	88	20.0%	68	16.8%	32	9.8%	32	9.8%	96	23.5%	32	8.5%	52	12.9%	32	8.0%	32	9.6%	0	0.0%	**732**	**16.6%**
SOFTWARE: TEXT	Lemmatiser	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
	Morphological analyser	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
	Machine translator	0	0.0%	12	2.4%	12	2.7%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	12	3.2%	12	3.0%	12	3.0%	0	0.0%	0	0.0%	**60**	**1.4%**
	Language and dialect identifier	20	4.4%	20	4.0%	20	4.5%	20	5.0%	20	6.1%	20	6.1%	20	4.9%	20	5.3%	20	5.0%	20	5.0%	20	6.0%	0	0.0%	**220**	**5.0%**
	Machine-aided human translation system	8	1.8%	8	1.6%	8	1.8%	4	1.0%	4	1.2%	4	1.2%	4	1.0%	8	2.1%	4	1.0%	4	1.0%	4	1.2%	0	0.0%	**60**	**1.4%**
	Web service	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
	Grammatical framework resource grammar	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	12	3.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	**12**	**0.3%**
	Corpus analysis tool	24	5.3%	24	4.8%	24	5.5%	24	5.9%	24	7.3%	24	7.3%	24	5.9%	24	6.4%	24	5.9%	24	6.0%	24	7.2%	0	0.0%	**264**	**6.0%**
	PDF Converter	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	4	14.3%	**4**	**0.1%**
	Anonymiser	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	4	14.3%	**4**	**0.1%**
	Terminology integration text	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	4	14.3%	**4**	**0.1%**
	Text aligner	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	28.6%	**8**	**0.2%**
	Tokeniser	12	2.7%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**132**	**3.0%**
	Named entity recogniser	12	2.7%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**132**	**3.0%**
	Compound analyser	0	0.0%	24	4.8%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	**24**	**0.5%**
	Chunker	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
	OCR/ICR	12	2.7%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**132**	**3.0%**
	Integrated automatic annotation	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
	Annotation	0	0.0%	12	2.4%	12	2.7%	12	3.0%	12	3.7%	12	3.7%	12	2.9%	12	3.2%	12	3.0%	12	3.0%	12	3.6%	0	0.0%	**120**	**2.7%**
SOFTWARE: SPEECH	Telephony applications	16	3.5%	4	0.8%	4	0.9%	4	1.0%	4	1.2%	4	1.2%	4	1.0%	4	1.1%	4	1.0%	4	1.0%	4	1.2%	0	0.0%	**56**	**1.3%**
	Speech recognition systems	20	4.4%	20	4.0%	20	4.5%	20	5.0%	20	6.1%	20	6.1%	20	4.9%	20	5.3%	20	5.0%	20	5.0%	20	6.0%	0	0.0%	**220**	**5.0%**
	Speech-to-speech translation system	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	28.6%	**8**	**0.2%**
	Acoustic analysis tool	8	1.8%	8	1.6%	8	1.8%	8	2.0%	8	2.4%	8	2.4%	8	2.0%	8	2.1%	8	2.0%	8	2.0%	8	2.4%	0	0.0%	**88**	**2.0%**
Total		**452**	**100.0%**	**500**	**100.0%**	**440**	**100.0%**	**404**	**100.0%**	**328**	**100.0%**	**328**	**100.0%**	**408**	**100.0%**	**376**	**100.0%**	**404**	**100.0%**	**400**	**100.0%**	**332**	**100.0%**	**28**	**100.0%**	**4400**	**100.0%**

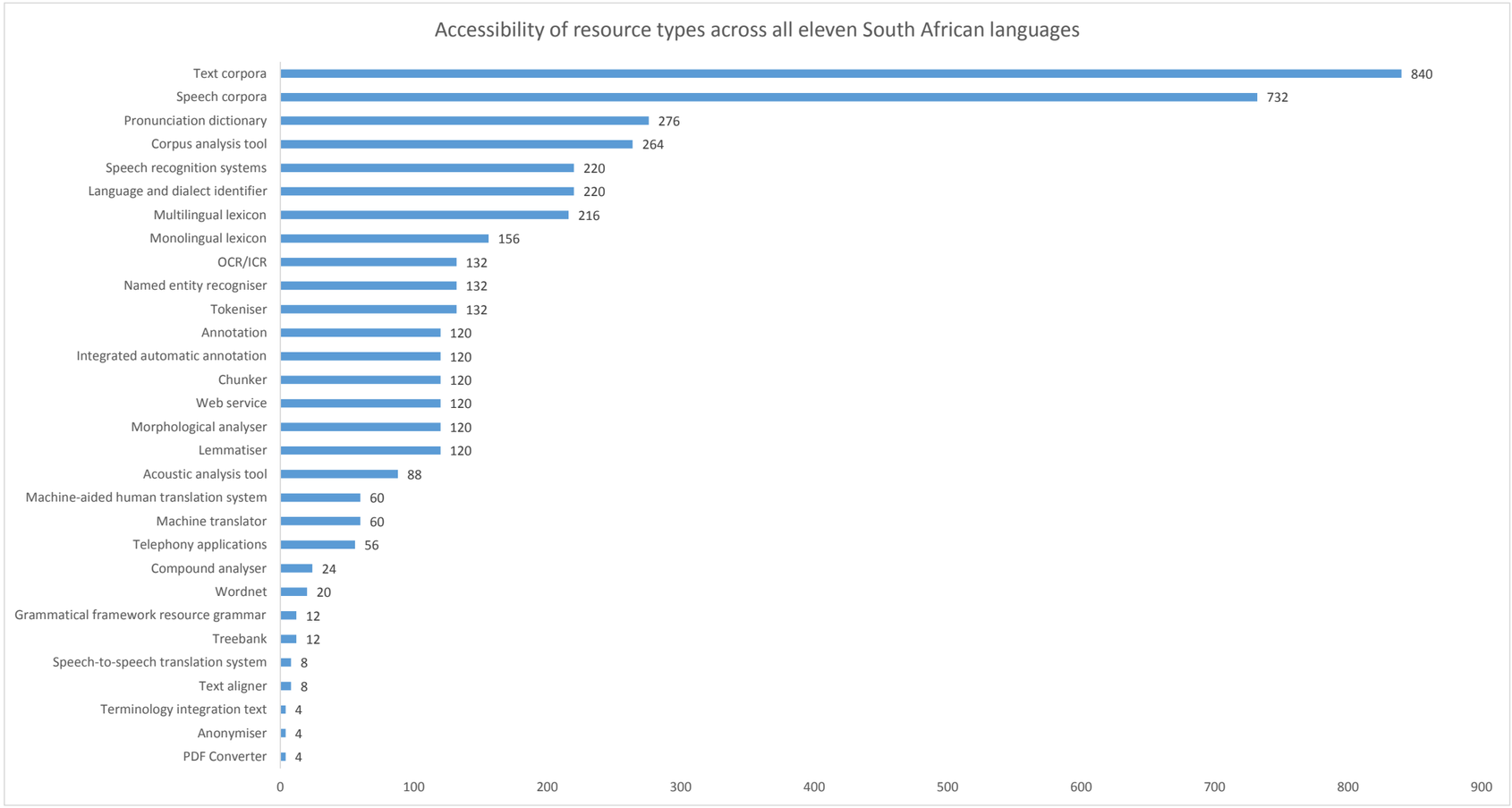
GRAPHS ILLUSTRATING THE LEVEL OF ACCESSIBILITY OF RESOURCES IN SOUTH AFRICA



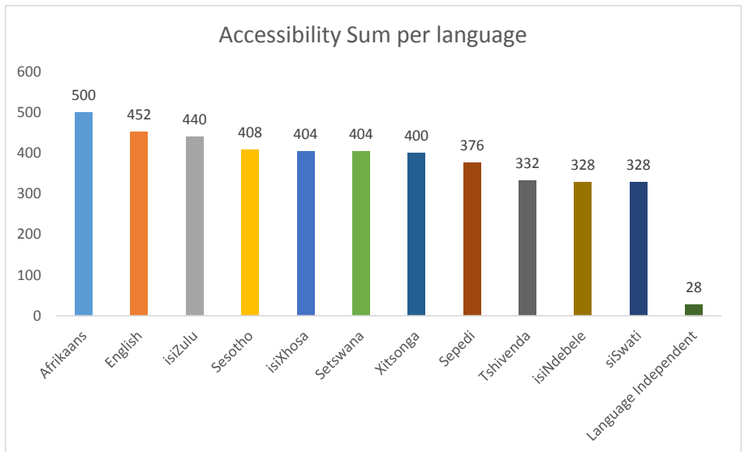


GRAPHS ILLUSTRATING THE LEVEL OF ACCESSIBILITY OF RESOURCES IN SOUTH AFRICA (ORDERED)

Resource Type	Accessibility Sum
PDF Converter	4
Anonymiser	4
Terminology integration text	4
Text aligner	8
Speech-to-speech translation system	8
Treebank	12
Grammatical framework resource grammar	12
Wordnet	20
Compound analyser	24
Telephony applications	56
Machine translator	60
Machine-aided human translation system	60
Acoustic analysis tool	88
Lemmatiser	120
Morphological analyser	120
Web service	120
Chunker	120
Integrated automatic annotation	120
Annotation	120
Tokeniser	132
Named entity recogniser	132
OCR/ICR	132
Monolingual lexicon	156
Multilingual lexicon	216
Language and dialect identifier	220
Speech recognition systems	220
Corpus analysis tool	264
Pronunciation dictionary	276
Speech corpora	732
Text corpora	840

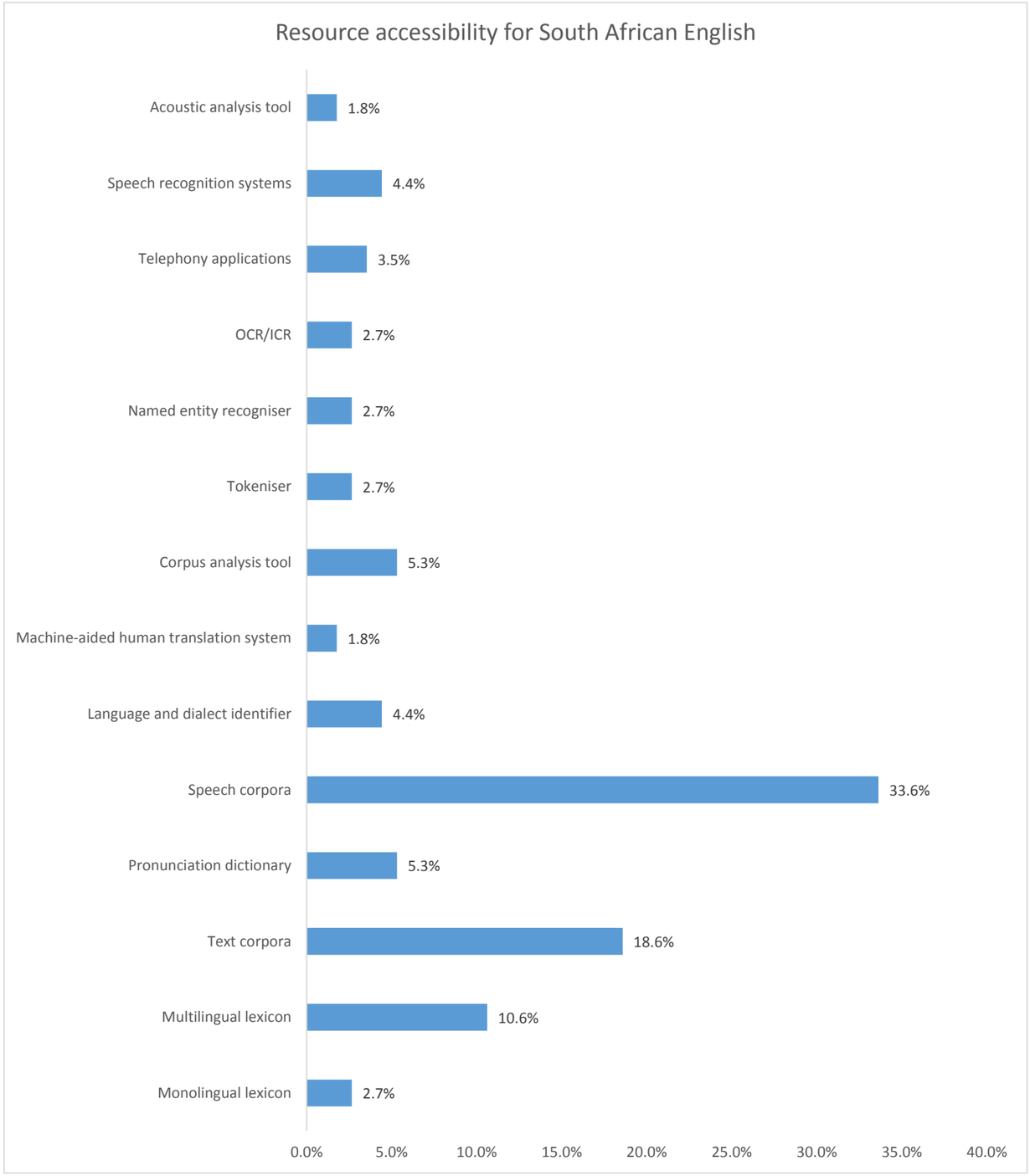


Language	Accessibility Sum
Afrikaans	500
English	452
isiZulu	440
Sesotho	408
isiXhosa	404
Setswana	404
Xitsonga	400
Sepedi	376
Tshivenda	332
isiNdebele	328
siSwati	328
Language Independent	28



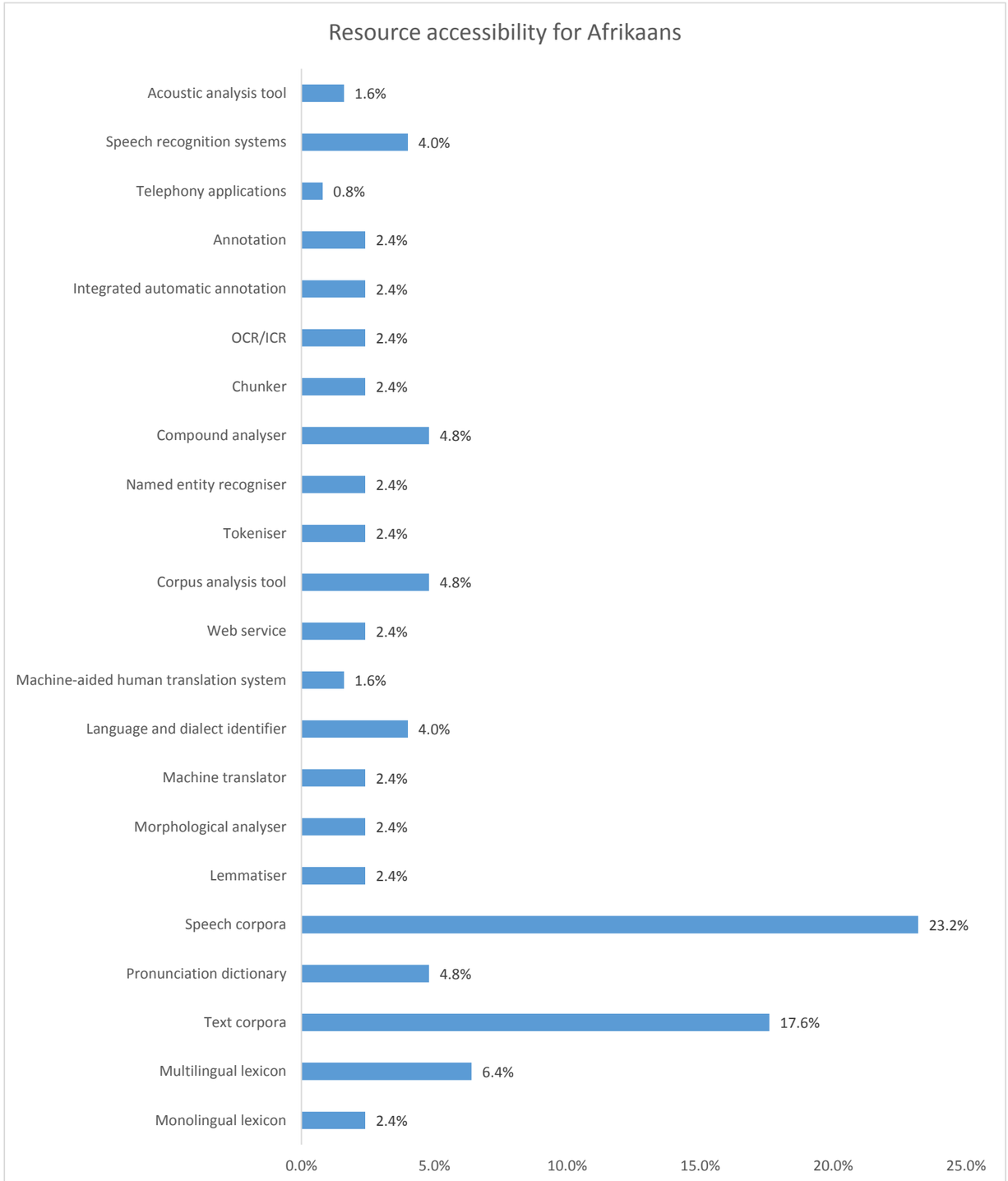
ACCESSIBILITY SUM PER LANGUAGE

		English	
DATA: TEXT	Monolingual lexicon	12	2.7%
	Multilingual lexicon	48	10.6%
	Text corpora	84	18.6%
DATA: SPEECH	Pronunciation dictionary	24	5.3%
	Speech corpora	152	33.6%
SOFTWARE: TEXT	Language and dialect identifier	20	4.4%
	Machine-aided human translation system	8	1.8%
	Corpus analysis tool	24	5.3%
	Tokeniser	12	2.7%
	Named entity recogniser	12	2.7%
	OCR/ICR	12	2.7%
SOFTWARE: SPEECH	Telephony applications	16	3.5%
	Speech recognition systems	20	4.4%
	Acoustic analysis tool	8	1.8%
	Total	**452**	**100.0%**



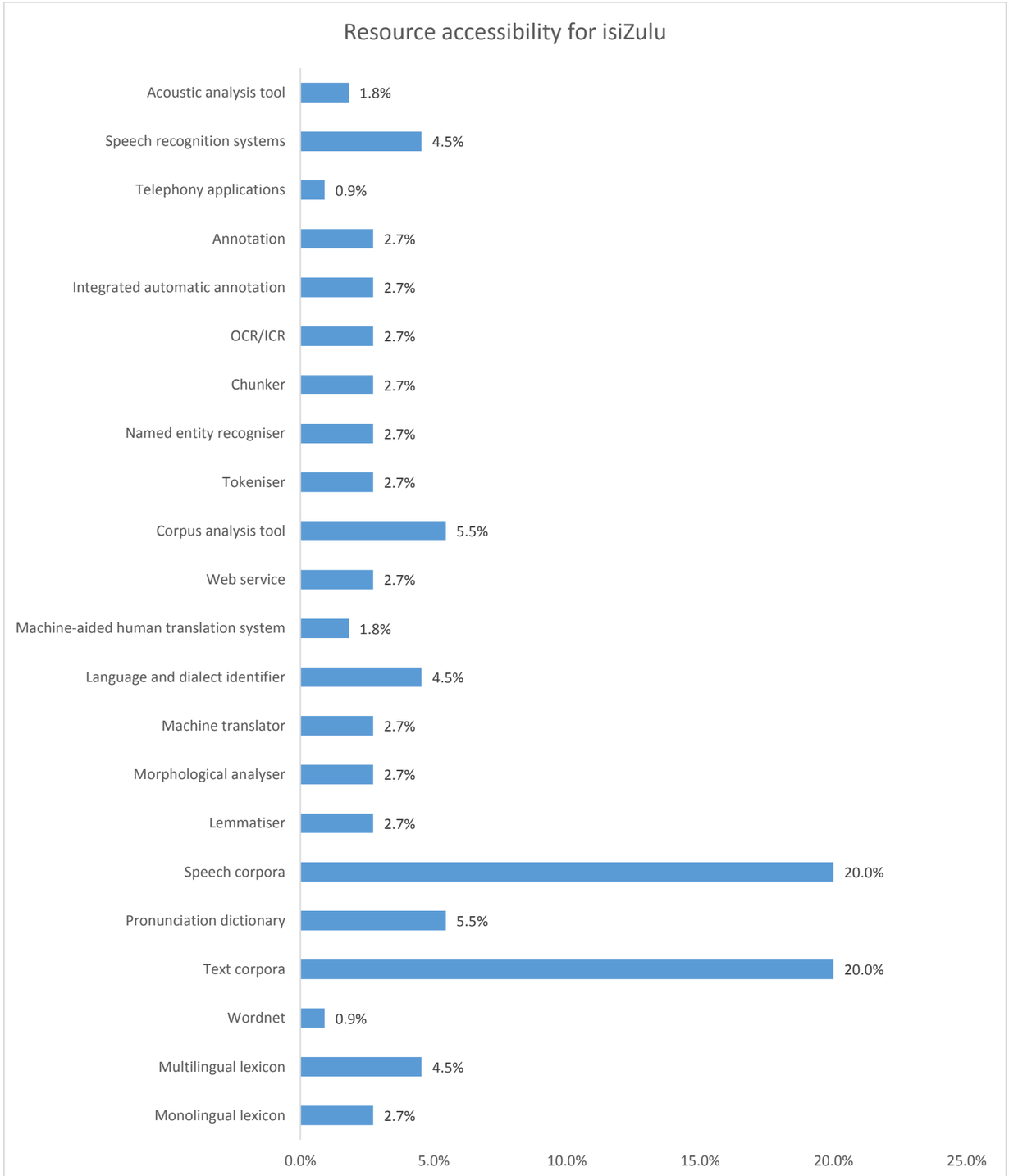
ACCESSIBILITY SUM PER LANGUAGE

		Afrikaans	
DATA: TEXT	Monolingual lexicon	12	2.4%
	Multilingual lexicon	32	6.4%
	Text corpora	88	17.6%
DATA: SPEECH	Pronunciation dictionary	24	4.8%
	Speech corpora	116	23.2%
SOFTWARE: TEXT	Lemmatiser	12	2.4%
	Morphological analyser	12	2.4%
	Machine translator	12	2.4%
	Language and dialect identifier	20	4.0%
	Machine-aided human translation system	8	1.6%
	Web service	12	2.4%
	Corpus analysis tool	24	4.8%
	Tokeniser	12	2.4%
	Named entity recogniser	12	2.4%
	Compound analyser	24	4.8%
	Chunker	12	2.4%
	OCR/ICR	12	2.4%
	Integrated automatic annotation	12	2.4%
	Annotation	12	2.4%
SOFTWARE: SPEECH	Telephony applications	4	0.8%
	Speech recognition systems	20	4.0%
	Acoustic analysis tool	8	1.6%
	Total	**500**	**100.0%**



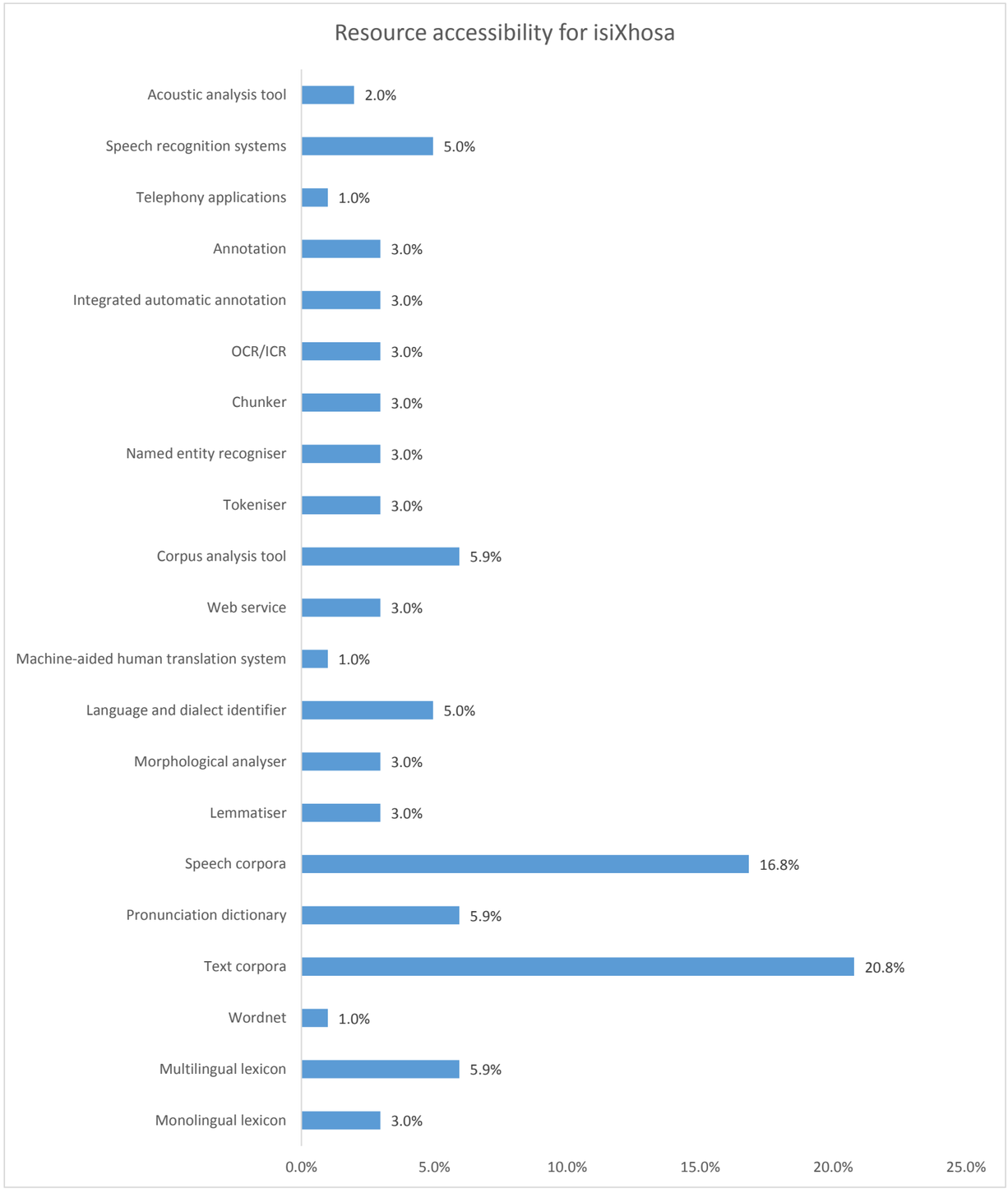
ACCESSIBILITY SUM PER LANGUAGE

		isiZulu	
DATA: TEXT	Monolingual lexicon	12	2.7%
	Multilingual lexicon	20	4.5%
	Wordnet	4	0.9%
	Text corpora	88	20.0%
DATA: SPEECH	Pronunciation dictionary	24	5.5%
	Speech corpora	88	20.0%
SOFTWARE: TEXT	Lemmatiser	12	2.7%
	Morphological analyser	12	2.7%
	Machine translator	12	2.7%
	Language and dialect identifier	20	4.5%
	Machine-aided human translation system	8	1.8%
	Web service	12	2.7%
	Corpus analysis tool	24	5.5%
	Tokeniser	12	2.7%
	Named entity recogniser	12	2.7%
	Chunker	12	2.7%
	OCR/ICR	12	2.7%
	Integrated automatic annotation	12	2.7%
	Annotation	12	2.7%
SOFTWARE: SPEECH	Telephony applications	4	0.9%
	Speech recognition systems	20	4.5%
	Acoustic analysis tool	8	1.8%
	Total	**440**	**100.0%**



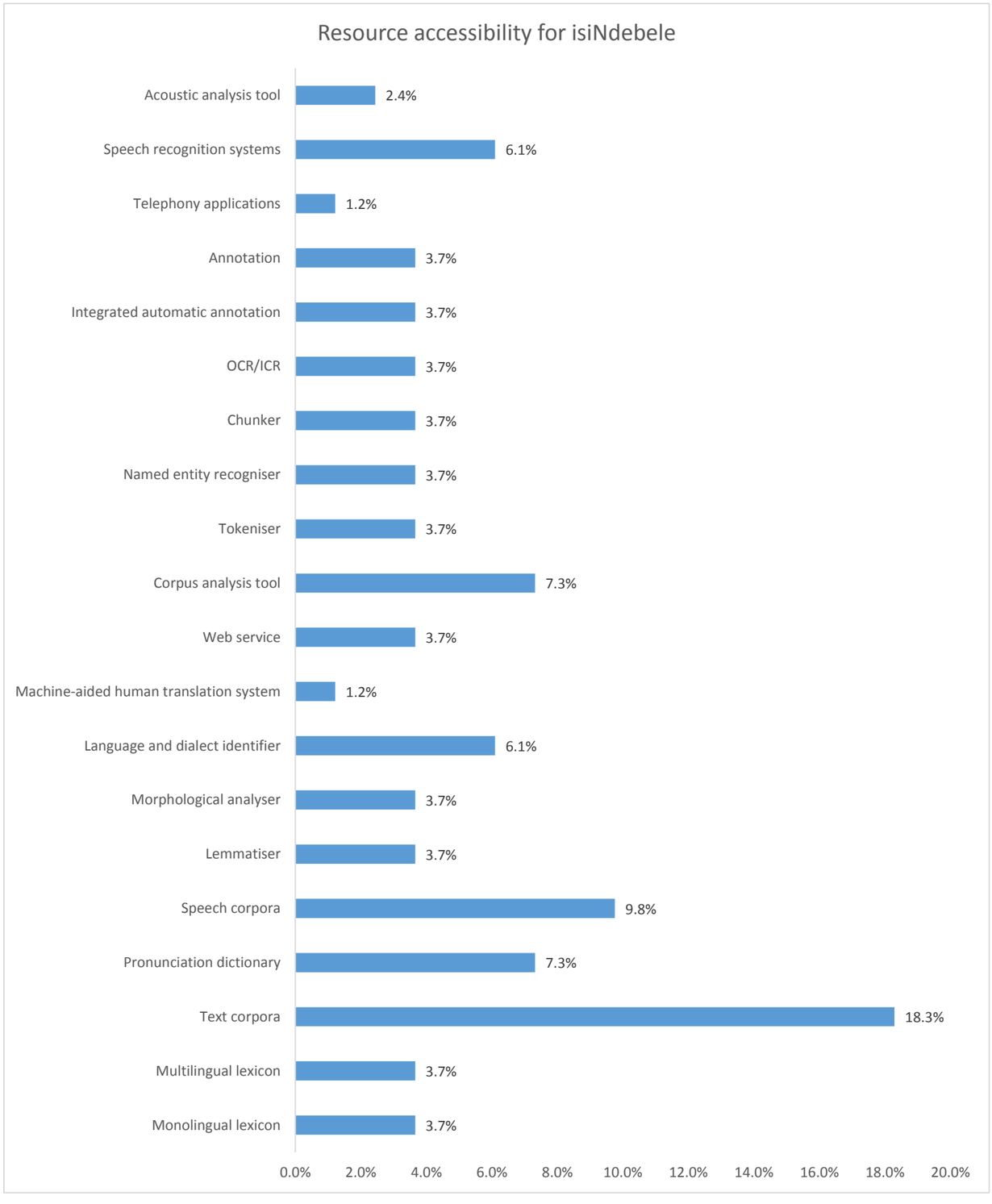
ACCESSIBILITY SUM PER LANGUAGE

		isiXhosa	
DATA: TEXT	Monolingual lexicon	12	3.0%
	Multilingual lexicon	24	5.9%
	Wordnet	4	1.0%
	Text corpora	84	20.8%
DATA: SPEECH	Pronunciation dictionary	24	5.9%
	Speech corpora	68	16.8%
SOFTWARE: TEXT	Lemmatiser	12	3.0%
	Morphological analyser	12	3.0%
	Language and dialect identifier	20	5.0%
	Machine-aided human translation system	4	1.0%
	Web service	12	3.0%
	Corpus analysis tool	24	5.9%
	Tokeniser	12	3.0%
	Named entity recogniser	12	3.0%
	Chunker	12	3.0%
	OCR/ICR	12	3.0%
	Integrated automatic annotation	12	3.0%
	Annotation	12	3.0%
SOFTWARE: SPEECH	Telephony applications	4	1.0%
	Speech recognition systems	20	5.0%
	Acoustic analysis tool	8	2.0%
	Total	**404**	**100.0%**



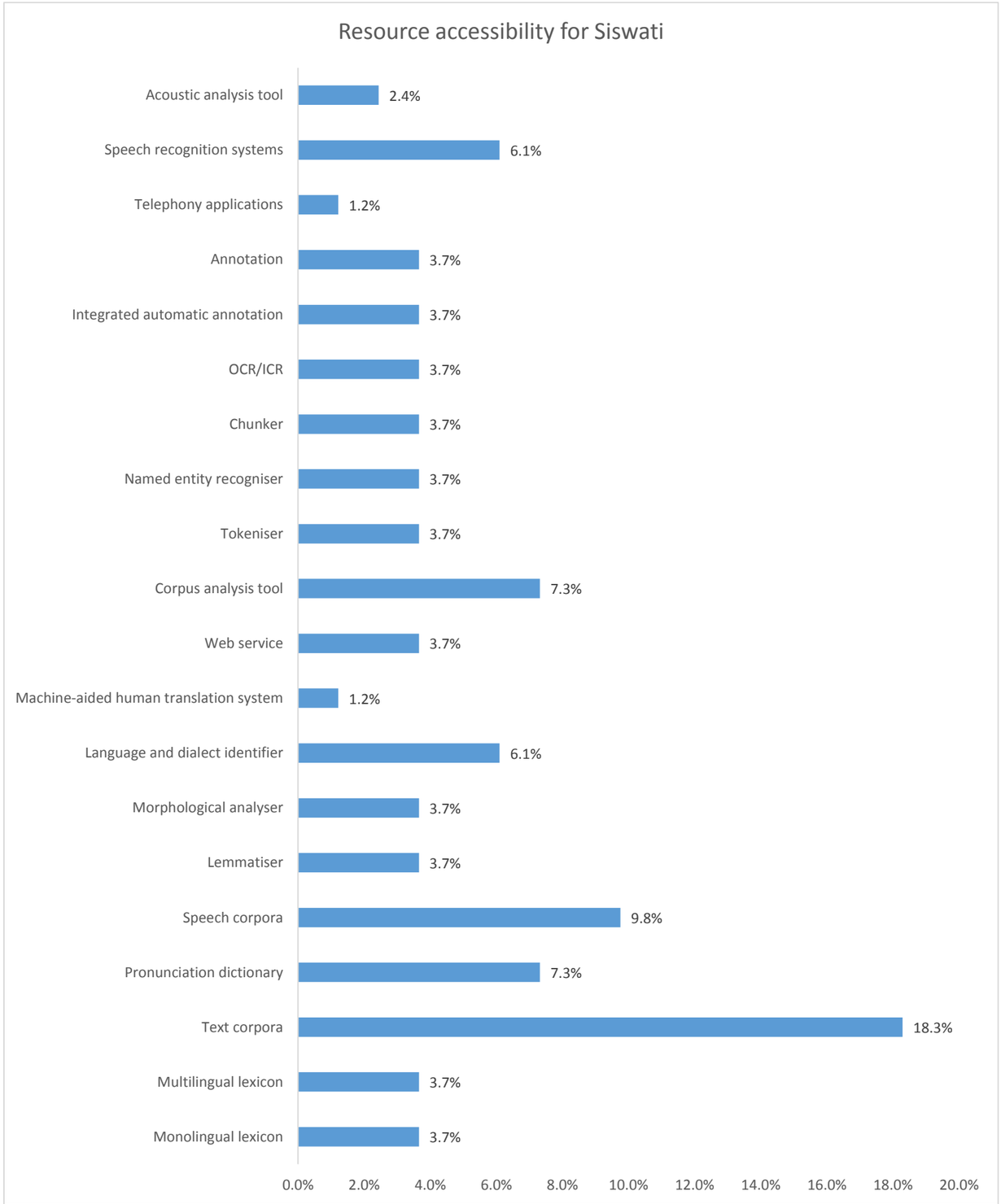
ACCESSIBILITY SUM PER LANGUAGE

		isiNdebele	
DATA: TEXT	Monolingual lexicon	12	3.7%
	Multilingual lexicon	12	3.7%
	Text corpora	60	18.3%
DATA: SPEECH	Pronunciation dictionary	24	7.3%
	Speech corpora	32	9.8%
SOFTWARE: TEXT	Lemmatiser	12	3.7%
	Morphological analyser	12	3.7%
	Language and dialect identifier	20	6.1%
	Machine-aided human translation system	4	1.2%
	Web service	12	3.7%
	Corpus analysis tool	24	7.3%
	Tokeniser	12	3.7%
	Named entity recogniser	12	3.7%
	Chunker	12	3.7%
	OCR/ICR	12	3.7%
	Integrated automatic annotation	12	3.7%
	Annotation	12	3.7%
SOFTWARE: SPEECH	Telephony applications	4	1.2%
	Speech recognition systems	20	6.1%
	Acoustic analysis tool	8	2.4%
	Total	**328**	**100.0%**



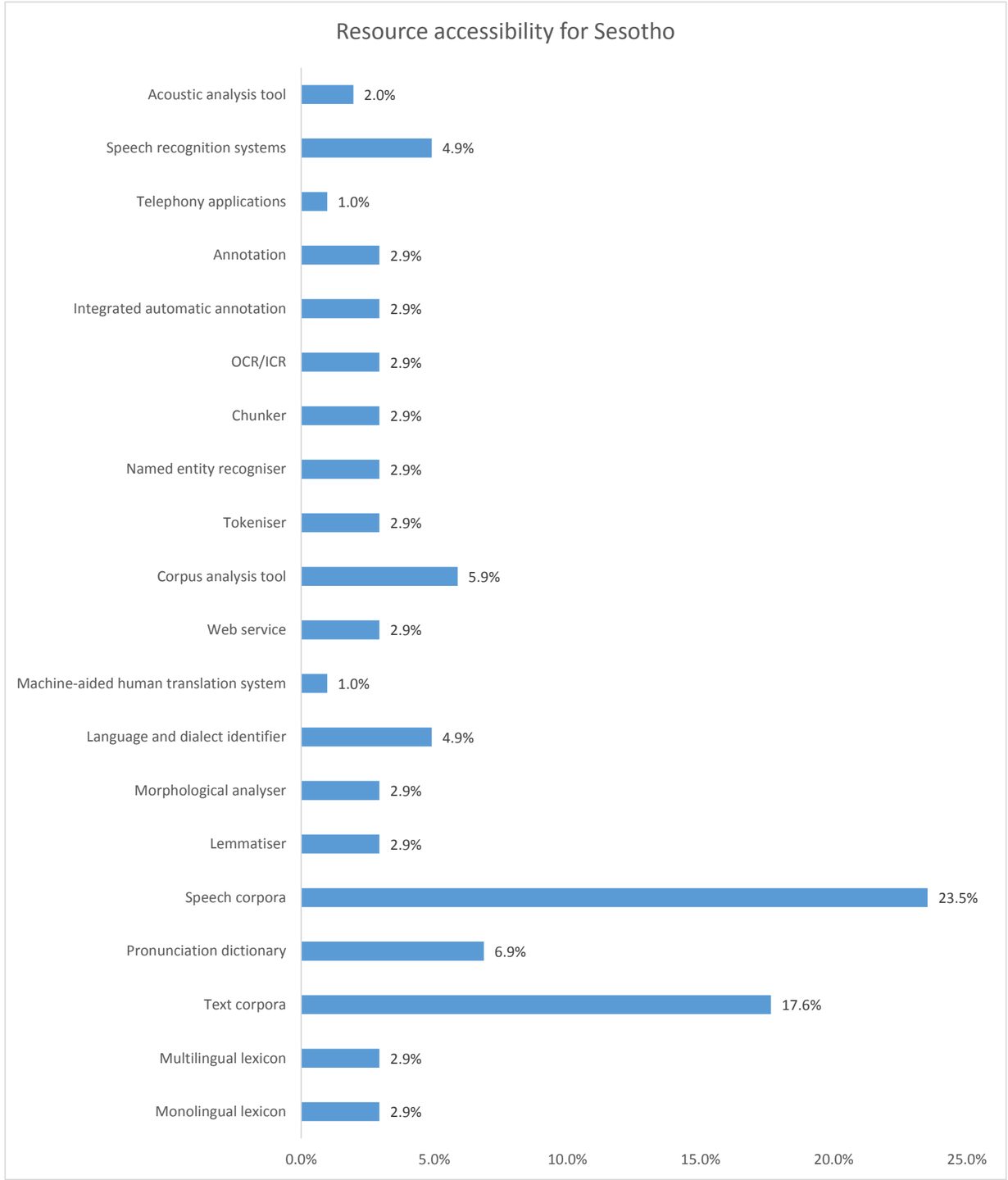
ACCESSIBILITY SUM PER LANGUAGE

		siSwati	
DATA: TEXT	Monolingual lexicon	12	3.7%
	Multilingual lexicon	12	3.7%
	Text corpora	60	18.3%
DATA: SPEECH	Pronunciation dictionary	24	7.3%
	Speech corpora	32	9.8%
SOFTWARE: TEXT	Lemmatiser	12	3.7%
	Morphological analyser	12	3.7%
	Language and dialect identifier	20	6.1%
	Machine-aided human translation system	4	1.2%
	Web service	12	3.7%
	Corpus analysis tool	24	7.3%
	Tokeniser	12	3.7%
	Named entity recogniser	12	3.7%
	Chunker	12	3.7%
	OCR/ICR	12	3.7%
	Integrated automatic annotation	12	3.7%
	Annotation	12	3.7%
SOFTWARE: SPEECH	Telephony applications	4	1.2%
	Speech recognition systems	20	6.1%
	Acoustic analysis tool	8	2.4%
	Total	**328**	**100.0%**



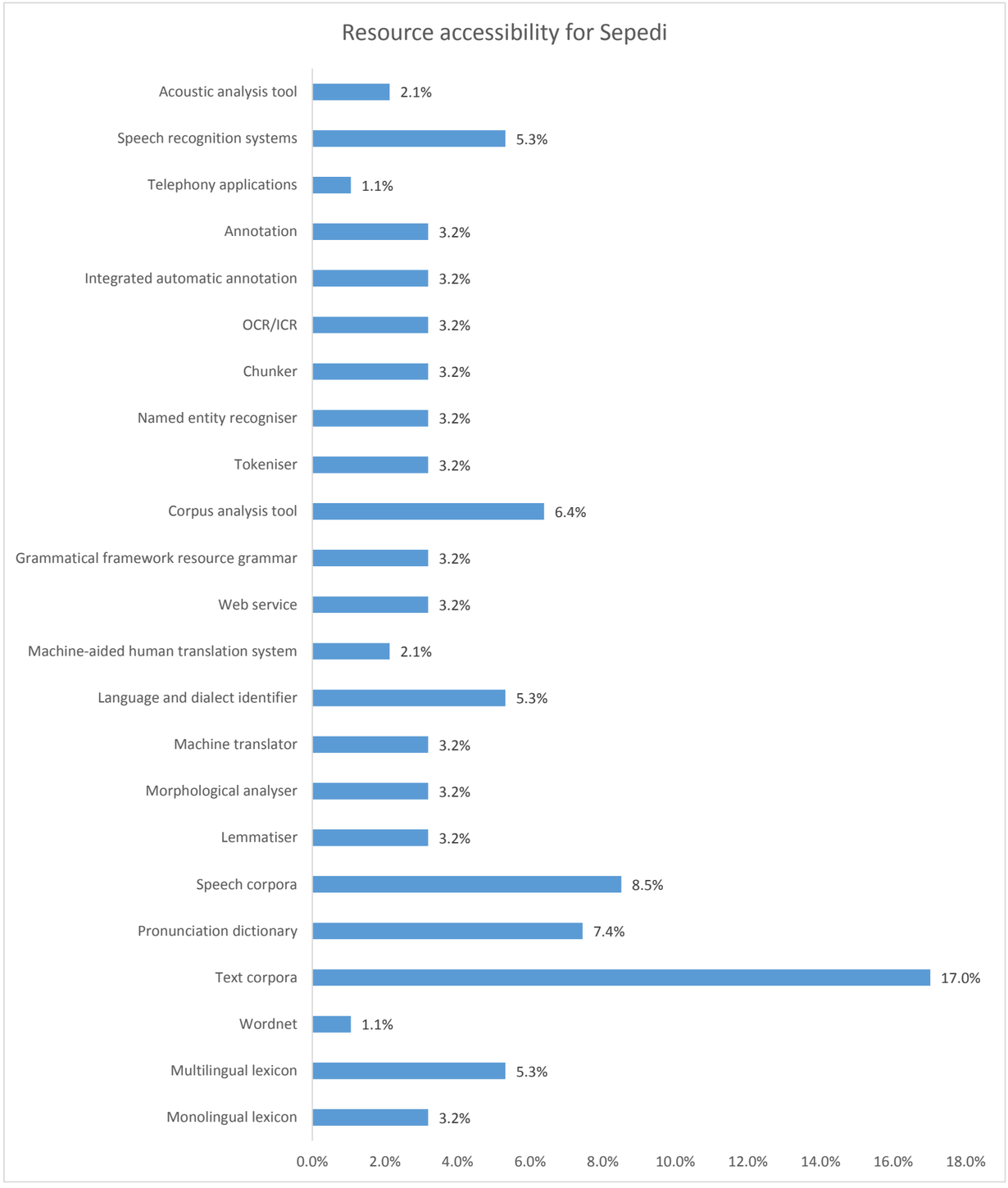
ACCESSIBILITY SUM PER LANGUAGE

		Sesotho	
DATA: TEXT	Monolingual lexicon	12	2.9%
	Multilingual lexicon	12	2.9%
	Text corpora	72	17.6%
DATA: SPEECH	Pronunciation dictionary	28	6.9%
	Speech corpora	96	23.5%
SOFTWARE: TEXT	Lemmatiser	12	2.9%
	Morphological analyser	12	2.9%
	Language and dialect identifier	20	4.9%
	Machine-aided human translation system	4	1.0%
	Web service	12	2.9%
	Corpus analysis tool	24	5.9%
	Tokeniser	12	2.9%
	Named entity recogniser	12	2.9%
	Chunker	12	2.9%
	OCR/ICR	12	2.9%
	Integrated automatic annotation	12	2.9%
	Annotation	12	2.9%
SOFTWARE: SPEECH	Telephony applications	4	1.0%
	Speech recognition systems	20	4.9%
	Acoustic analysis tool	8	2.0%
	Total	**408**	**100.0%**



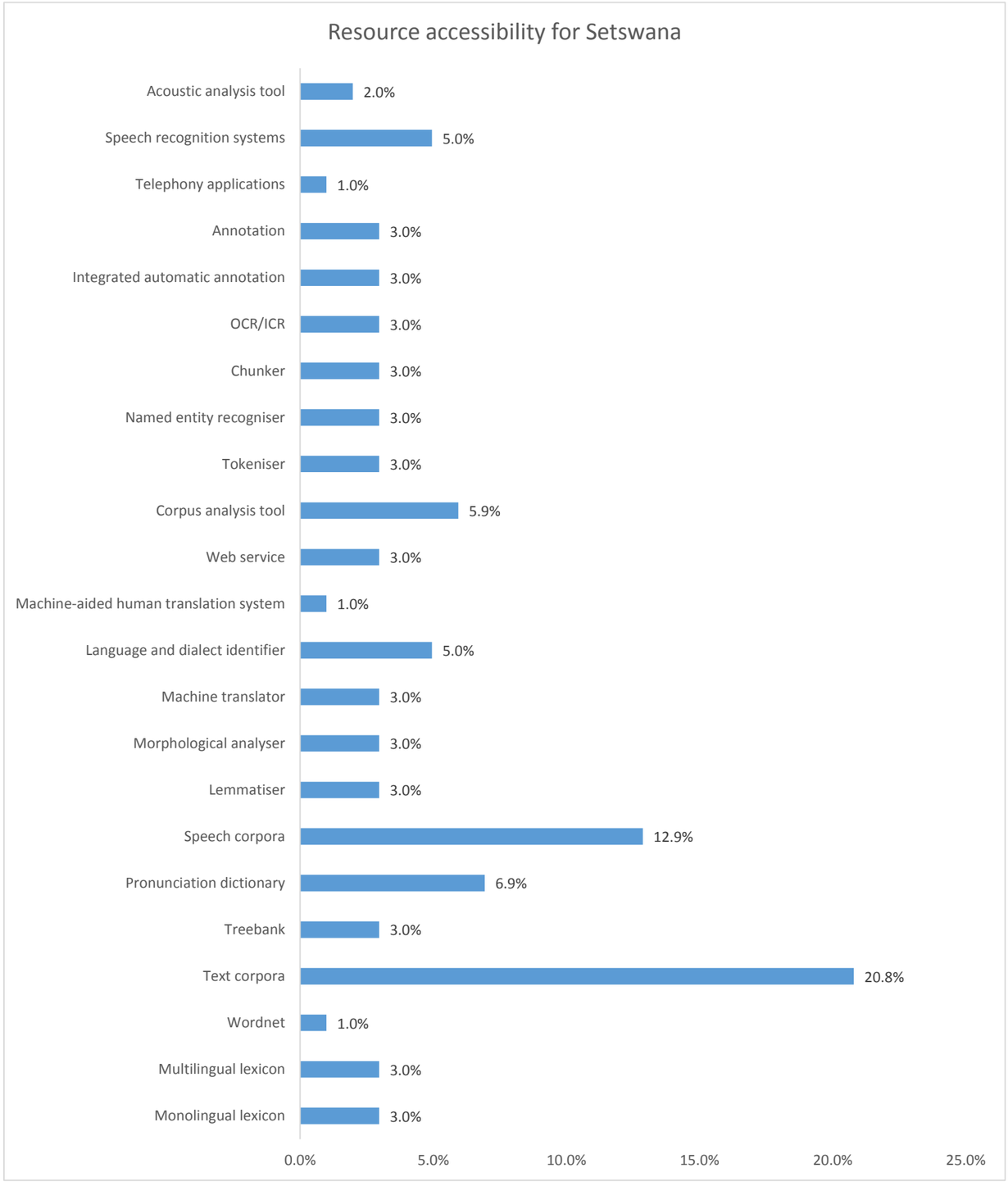
ACCESSIBILITY SUM PER LANGUAGE

		Sepedi	
DATA: TEXT	Monolingual lexicon	12	3.2%
	Multilingual lexicon	20	5.3%
	Wordnet	4	1.1%
	Text corpora	64	17.0%
DATA: SPEECH	Pronunciation dictionary	28	7.4%
	Speech corpora	32	8.5%
SOFTWARE: TEXT	Lemmatiser	12	3.2%
	Morphological analyser	12	3.2%
	Machine translator	12	3.2%
	Language and dialect identifier	20	5.3%
	Machine-aided human translation system	8	2.1%
	Web service	12	3.2%
	Grammatical framework resource grammar	12	3.2%
	Corpus analysis tool	24	6.4%
	Tokeniser	12	3.2%
	Named entity recogniser	12	3.2%
	Chunker	12	3.2%
	OCR/ICR	12	3.2%
	Integrated automatic annotation	12	3.2%
	Annotation	12	3.2%
SOFTWARE: SPEECH	Telephony applications	4	1.1%
	Speech recognition systems	20	5.3%
	Acoustic analysis tool	8	2.1%
	Total	**376**	**100.0%**



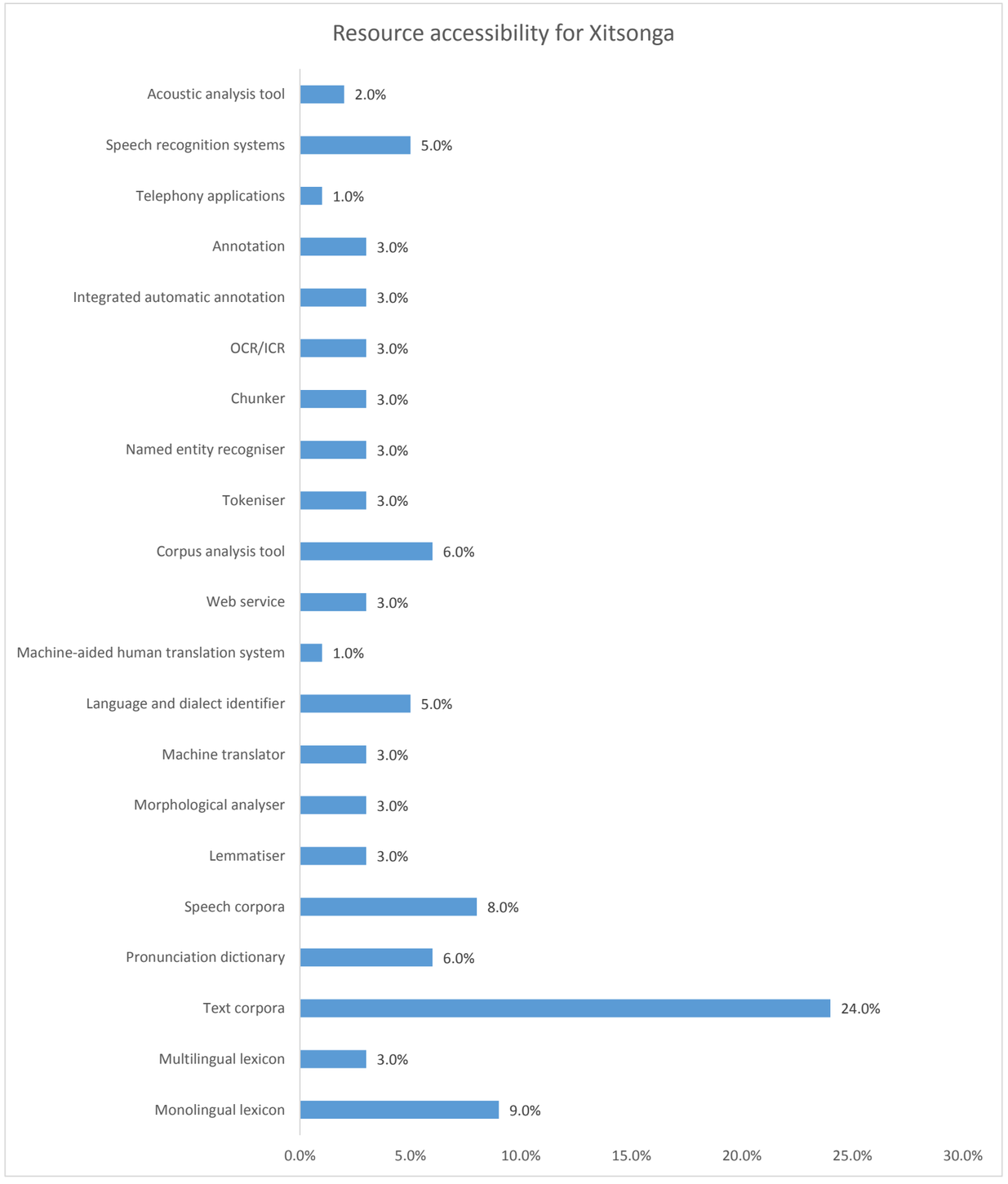
ACCESSIBILITY SUM PER LANGUAGE

		Setswana	
DATA: TEXT	Monolingual lexicon	12	3.0%
	Multilingual lexicon	12	3.0%
	Wordnet	4	1.0%
	Text corpora	84	20.8%
	Treebank	12	3.0%
DATA: SPEECH	Pronunciation dictionary	28	6.9%
	Speech corpora	52	12.9%
SOFTWARE: TEXT	Lemmatiser	12	3.0%
	Morphological analyser	12	3.0%
	Machine translator	12	3.0%
	Language and dialect identifier	20	5.0%
	Machine-aided human translation system	4	1.0%
	Web service	12	3.0%
	Corpus analysis tool	24	5.9%
	Tokeniser	12	3.0%
	Named entity recogniser	12	3.0%
	Chunker	12	3.0%
	OCR/ICR	12	3.0%
	Integrated automatic annotation	12	3.0%
	Annotation	12	3.0%
SOFTWARE: SPEECH	Telephony applications	4	1.0%
	Speech recognition systems	20	5.0%
	Acoustic analysis tool	8	2.0%
	Total	**404**	**100.0%**



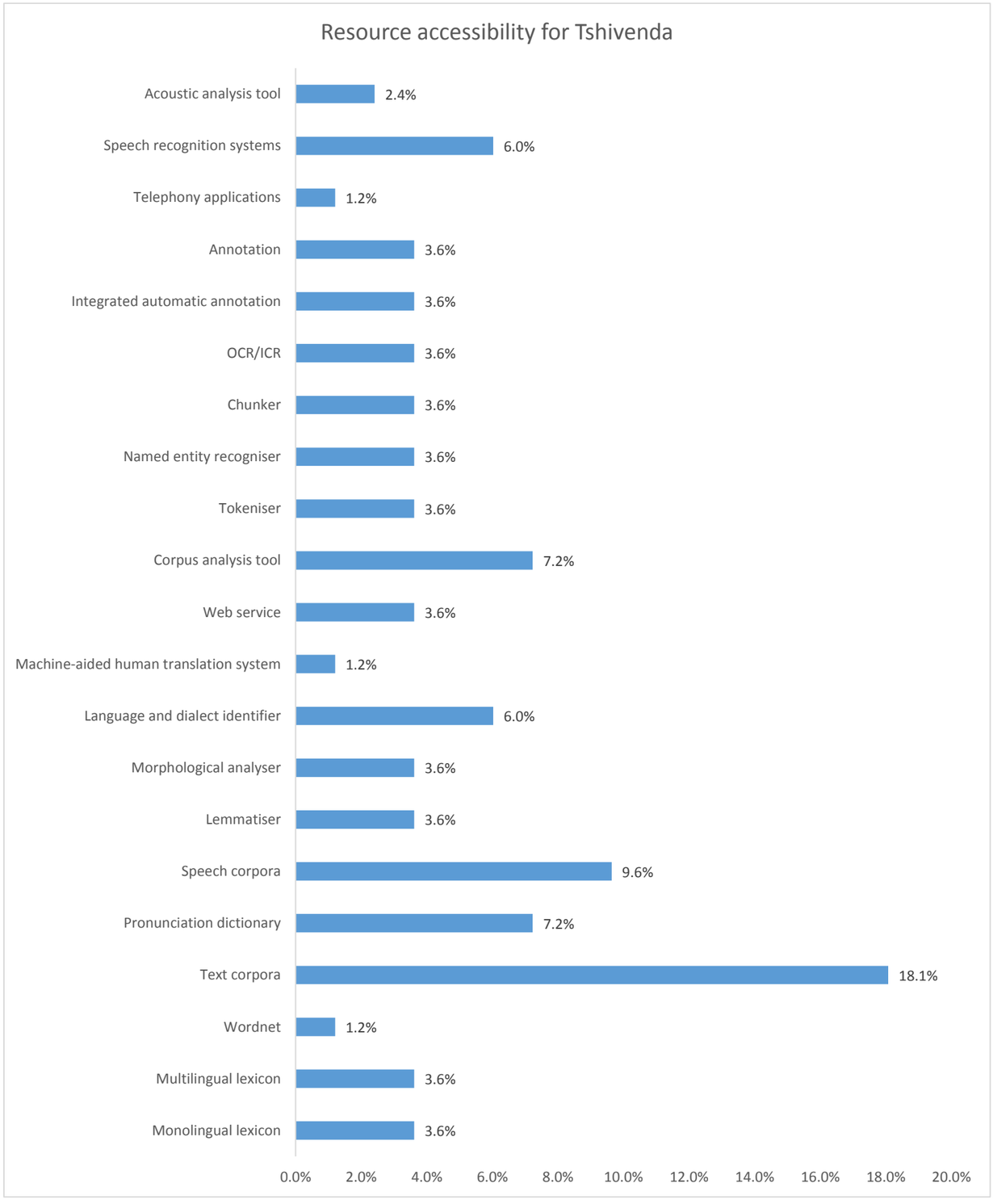
ACCESSIBILITY SUM PER LANGUAGE

		Xitsonga	
DATA: TEXT	Monolingual lexicon	36	9.0%
	Multilingual lexicon	12	3.0%
	Text corpora	96	24.0%
DATA: SPEECH	Pronunciation dictionary	24	6.0%
	Speech corpora	32	8.0%
SOFTWARE: TEXT	Lemmatiser	12	3.0%
	Morphological analyser	12	3.0%
	Machine translator	12	3.0%
	Language and dialect identifier	20	5.0%
	Machine-aided human translation system	4	1.0%
	Web service	12	3.0%
	Corpus analysis tool	24	6.0%
	Tokeniser	12	3.0%
	Named entity recogniser	12	3.0%
	Chunker	12	3.0%
	OCR/ICR	12	3.0%
	Integrated automatic annotation	12	3.0%
	Annotation	12	3.0%
SOFTWARE: SPEECH	Telephony applications	4	1.0%
	Speech recognition systems	20	5.0%
	Acoustic analysis tool	8	2.0%
	Total	**400**	**100.0%**



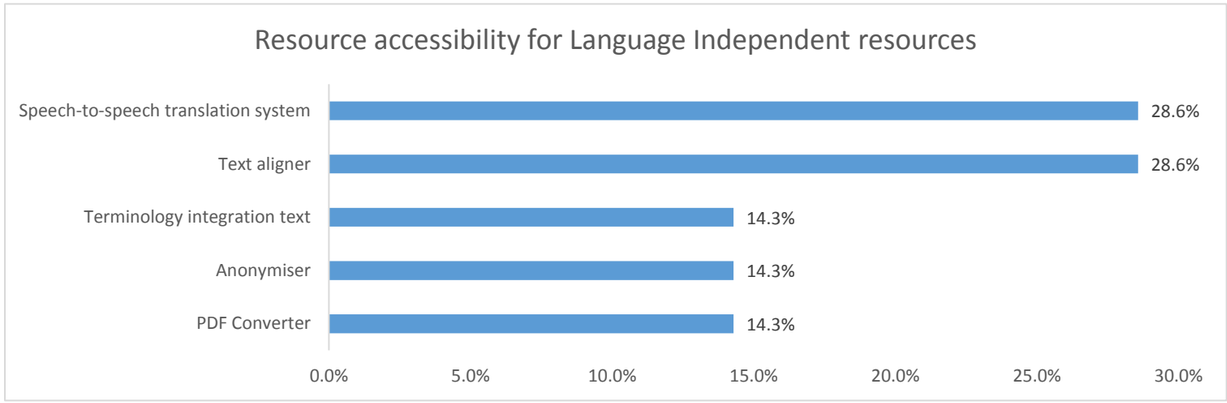
ACCESSIBILITY SUM PER LANGUAGE

		Tshivenda	
DATA: TEXT	Monolingual lexicon	12	3.6%
	Multilingual lexicon	12	3.6%
	Wordnet	4	1.2%
	Text corpora	60	18.1%
DATA: SPEECH	Pronunciation dictionary	24	7.2%
	Speech corpora	32	9.6%
SOFTWARE: TEXT	Lemmatiser	12	3.6%
	Morphological analyser	12	3.6%
	Language and dialect identifier	20	6.0%
	Machine-aided human translation system	4	1.2%
	Web service	12	3.6%
	Corpus analysis tool	24	7.2%
	Tokeniser	12	3.6%
	Named entity recogniser	12	3.6%
	Chunker	12	3.6%
	OCR/ICR	12	3.6%
	Integrated automatic annotation	12	3.6%
	Annotation	12	3.6%
SOFTWARE: SPEECH	Telephony applications	4	1.2%
	Speech recognition systems	20	6.0%
	Acoustic analysis tool	8	2.4%
	Total	**332**	**100.0%**



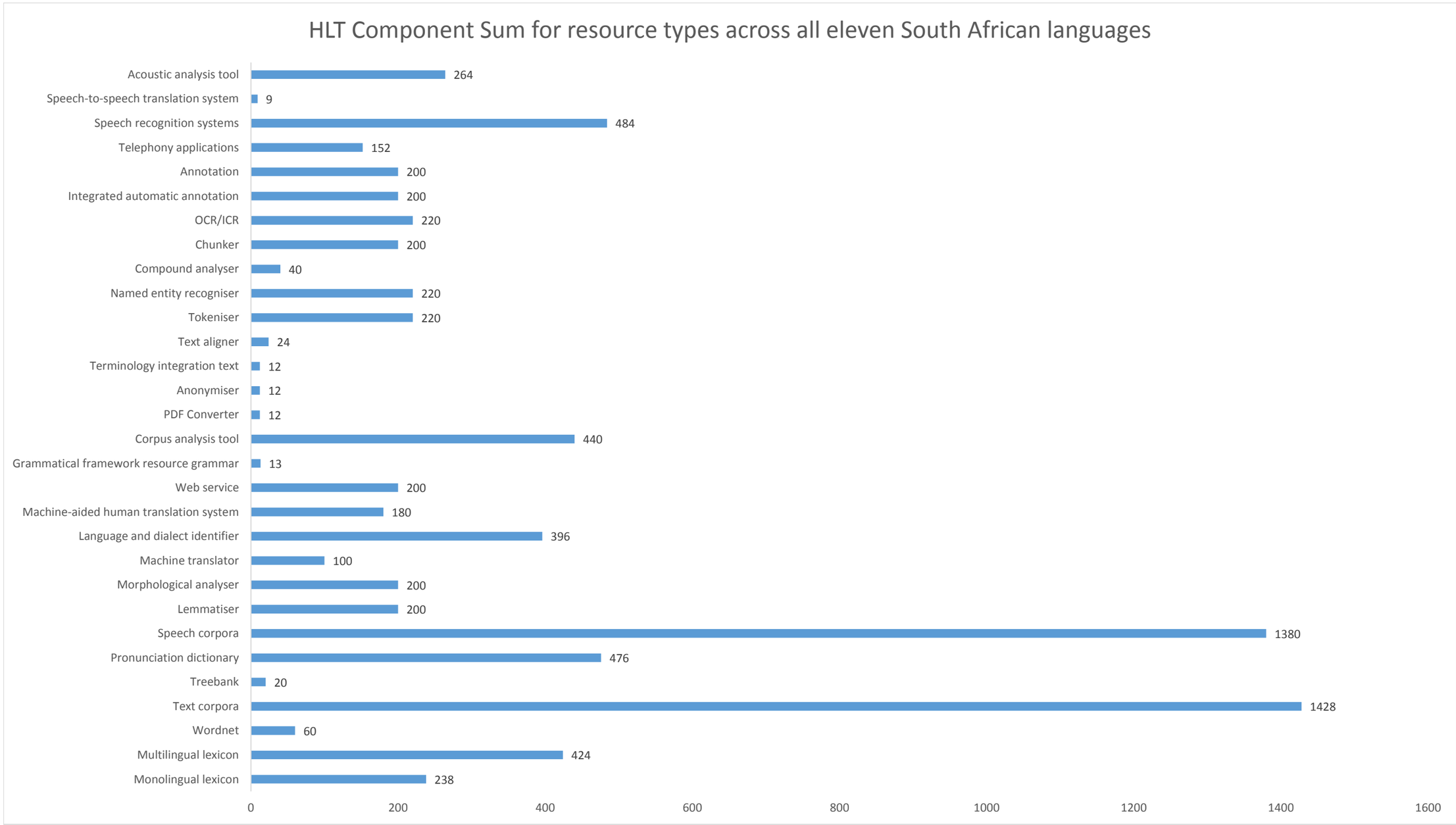
ACCESSIBILITY SUM PER LANGUAGE

		Lang. Indep.	
SOFTWARE: TEXT	PDF Converter	4	14.3%
	Anonymiser	4	14.3%
	Terminology integration text	4	14.3%
	Text aligner	8	28.6%
SOFTWARE: SPEECH	Speech-to-speech translation system	8	28.6%
	Total	**28**	**100.0%**

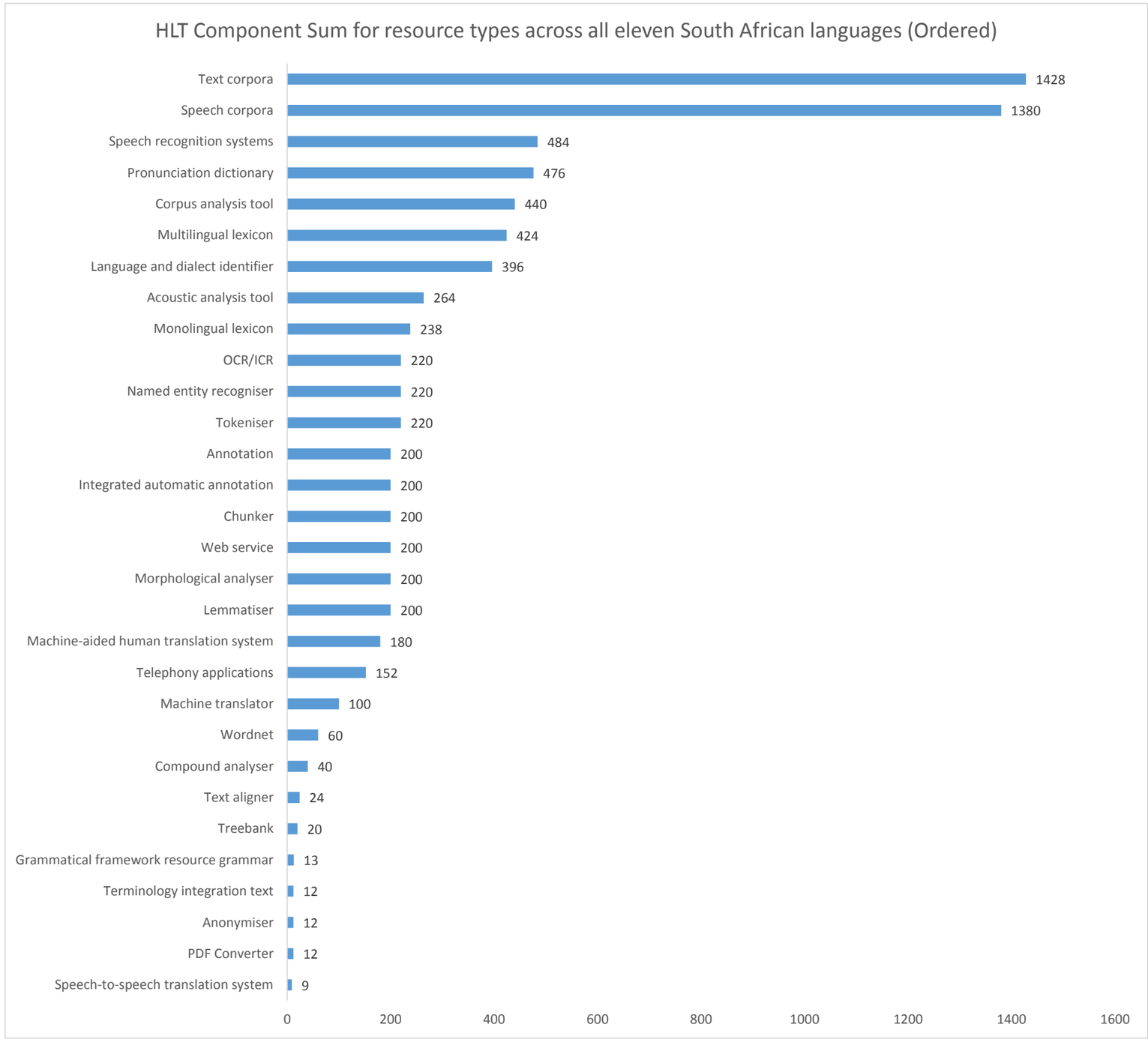


HLT COMPONENT SUM

		English			Afrikaans			isiZulu			isiXhosa			isiNdebele			siSwati			Sesotho			Sepedi			Setswana			Xitsonga			Tshivenda			Language Independent			TOTAL
		MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	MS	AS	HLT CS	HLT CS
DATA: TEXT	Monolingual lexicon	8	12	20	1	12	13	1	12	13	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	16	36	52	8	12	20	0	0	0	**238**
	Multilingual lexicon	64	48	112	32	32	64	24	20	44	16	24	40	8	12	20	8	12	20	8	12	20	24	20	44	8	12	20	8	12	20	8	12	20	0	0	0	**424**
	Wordnet	0	0	0	0	0	0	8	4	12	8	4	12	0	0	0	0	0	0	0	0	0	8	4	12	8	4	12	0	0	0	8	4	12	0	0	0	**60**
	Text corpora	72	84	156	60	88	148	64	88	152	56	84	140	40	60	100	40	60	100	48	72	120	48	64	112	56	84	140	64	96	160	40	60	100	0	0	0	**1428**
	Treebank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	12	20	0	0	0	0	0	0	0	0	0	**20**
DATA: SPEECH	Pronunciation dictionary	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	24	28	52	24	28	52	24	28	52	16	24	40	16	24	40	0	0	0	**476**
	Speech corpora	128	152	280	104	116	220	80	88	168	56	68	124	32	32	64	32	32	64	80	96	176	32	32	64	40	52	92	32	32	64	32	32	64	0	0	0	**1380**
SOFTWARE: TEXT	Lemmatiser	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
	Morphological analyser	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
	Machine translator	0	0	0	8	12	20	8	12	20	0	0	0	0	0	0	0	0	0	0	0	0	8	12	20	8	12	20	8	12	20	0	0	0	0	0	0	**100**
	Language and dialect identifier	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	16	20	36	0	0	0	**396**
	Machine-aided human translation system	16	8	24	16	8	24	16	8	24	8	4	12	8	4	12	8	4	12	8	4	12	16	8	24	8	4	12	8	4	12	8	4	12	0	0	0	**180**
	Web service	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
	Grammatical framework resource grammar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12	13	0	0	0	0	0	0	0	0	0	0	0	0	**13**
	Corpus analysis tool	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	16	24	40	0	0	0	**440**
	PDF Converter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	4	12	**12**
	Anonymiser	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	4	12	**12**
	Terminology integration text	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	4	12	**12**
	Text aligner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	8	24	**24**
	Tokeniser	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**220**
	Named entity recogniser	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**220**
	Compound analyser	0	0	0	16	24	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**40**
	Chunker	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
	OCR/ICR	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**220**
	Integrated automatic annotation	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
	Annotation	0	0	0	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	8	12	20	0	0	0	**200**
SOFTWARE: SPEECH	Telephony applications	16	16	32	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	8	4	12	0	0	0	**152**
	Speech recognition systems	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	24	20	44	0	0	0	**484**
	Speech-to-speech translation system	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	9	**9**
	Acoustic analysis tool	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	16	8	24	0	0	0	**264**
Total		**416**	**452**	**868**	**405**	**500**	**905**	**369**	**440**	**809**	**320**	**404**	**724**	**264**	**328**	**592**	**264**	**328**	**592**	**328**	**408**	**736**	**321**	**376**	**697**	**320**	**404**	**724**	**304**	**400**	**704**	**272**	**332**	**604**	**41**	**28**	**69**	**8024**


HLT Component Sum for resource types across all eleven South African languages
Acoustic analysis tool 264
Speech-to-speech translation system 9
Speech recognition systems 484
Telephony applications 152
Annotation 200
Integrated automatic annotation 200
OCR/ICR 220
Chunker 200
Compound analyser 40
Named entity recogniser 220
Tokeniser 220
Text aligner 24
Terminology integration text 12
Anonymiser 12
PDF Converter 12
Corpus analysis tool 440
Grammatical framework resource grammar 13
Web service 200
Machine-aided human translation system 180
Language and dialect identifier 396
Machine translator 100
Morphological analyser 200
Lemmatiser 200
Speech corpora 1380
Pronunciation dictionary 476
Treebank 20
Text corpora 1428
Wordnet 60
Multilingual lexicon 424
Monolingual lexicon 238
0
200
400
600
800
1000
1200
1400
1600

Resource Type	HLT CS
Speech-to-speech translation system	9
PDF Converter	12
Anonymiser	12
Terminology integration text	12
Grammatical framework resource grammar	13
Treebank	20
Text aligner	24
Compound analyser	40
Wordnet	60
Machine translator	100
Telephony applications	152
Machine-aided human translation system	180
Lemmatiser	200
Morphological analyser	200
Web service	200
Chunker	200
Integrated automatic annotation	200
Annotation	200
Tokeniser	220
Named entity recogniser	220
OCR/ICR	220
Monolingual lexicon	238
Acoustic analysis tool	264
Language and dialect identifier	396
Multilingual lexicon	424
Corpus analysis tool	440
Pronunciation dictionary	476
Speech recognition systems	484
Speech corpora	1380
Text corpora	1428



SOUTH AFRICAN CENTRE FOR DIGITAL LANGUAGE RESOURCES

CSIR node meeting: HLT Audit discussion

07/02/2018

CSIR, Meraka institute

MEMBERS: R Eiselen (ERE), K Calteaux (KC), C Moors (CM), A. Louw (AL)

ITEM | DESCRIPTION	DISCUSSION	RESPONSIBLE PERSON	DEADLINE
1. Purpose	To discuss WP5 (dynamic audit update system in the HLT Audit project) • This WP is to provide a system to be able to easily update HLT resources onto the SADiLaR database	KC	NA

	• Agreed previously that this WP5 will first be discussed between CSIR and NWU before commencing to determine the relevance based on the completion of WPs 1- 4.
	• Deliverables for this WP5 included in the proposal were:
	  ○ Architecture design diagrams.
	  ○ Functional specifications.
	  ○ Recommendation report to implement the solution.		

2. Discussion on need and relevance of WP5	• Is there still a need to develop a separate dynamic updating system as envisaged when the proposal was written or does the survey tool developed (not planned in the proposal) function as a way to dynamically update the resources database? • During the interaction with experts in the instrument design phase, the tool for conducting the 2018 Audit changed from an excel spreadsheet (as used in 2009) to an online survey. • The work to design, develop and test the online survey tool required much more work than originally planned. • An output of WP2 (Instrument design) is a design for the online audit tool. • The online survey tool will be transferred to SADiLaR. As such, it is proposed that the online survey tool replaces the need for a separate dynamic update system. An additional system is no longer required. • The design of the new tool was undertaken with the	CSIR and NWU	NA

	guidance of HLT experts – and can be seen to be their recommendations for the structuring of the Audit and by implication the information required in a dynamic update system. • The design of the new tool has many benefits including that it better aligns the information required for text resources to that required for speech resources and that the categories of resources defined in the tool have been selected and defined based on expert analysis and input. The additional hours spent on developing the online survey tool are therefore justified. • The proposal did not include implementation of the dynamic update system by the CSIR (only designs, specs and a recommendations report).		
3. Outcome of discussion on WP5	• The online survey tool will become the new system through which HLT resources will be added to the catalogue and/or list of available resources • NWU is in agreement with the online survey design documentation replacing the dynamic system architecture design deliverable. • The transfer of the online survey (code) to SADiLaR sufficiently covers the deliverables of WP5 (designs, specs). • The project report will detail recommendations for continuously updating the information on available HLT resources.	CSIR will explain the reasons behind using the WP2 design as the WP5 architecture design. CSIR will provide information on the specifications of the survey tool in the final project report. CSIR is not responsible to implement the solution, only to provide recommendations on how to implement the solution CSIR will look into the processes followed with the LRE Map, among others, and make recommendations for how best to get researchers to update	As per the deliverable schedule.

	Question arose on how the mapping between the audit results and resources on the RMA will take place once the audit data has been transferred? • The additional time spent on WP2 and the time planned for WP5 will cancel one another out, so that the project will still deliver on time and on budget.	the information on the database. NWU will map the results to the RMA database once the 2018 Audit data has been received.	

Meeting adjourned. Date of next meeting to be determined

Signature_____ **Date:**_____

Prof Attie de Lange
Director
South African Centre for Digital Language Resources (SADiLaR)
NWU

Human Language Technology Audit 2018: Design considerations and Methodology

ABSTRACT

Technology audits can play a significant role in surfacing information which can be used by researchers, policy-makers and funders alike to build a country's research and development system of innovation towards increasing its competitiveness and contributing to its economy. In 2016, South Africa established a Centre for Digital Language Resources (SADiLaR) with the aim of supporting a large research infrastructure programme tasked with bringing South African language resources into the digital age. This paper discusses the design considerations and methodology employed to undertake one of the first projects funded by SADiLaR: an updated audit of human language technology resources in South Africa. The paper aims to provide sufficient information to replicate such a technology audit in other environments. The design considerations aim to ensure a pleasant user experience, in order to facilitate as much input as possible. The approach aims to ensure that a sustainable audit tool is developed which can be hosted by SADiLaR in future.

CCS CONCEPTS

• **Human-centered computing** → *User interface design*;

KEYWORDS

Human language technology, Technology audit, Language resources, Text resources, Speech resources, Digital Humanities

ACM Reference Format:

. 2018. Human Language Technology Audit 2018: Design considerations and Methodology. In *Proceedings of SAICSIT 2018.* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION AND NEEDS

The establishment of research infrastructure can play a significant role in South Africa's social and economic development, if such infrastructure programmes create opportunities for innovative national research and development. The National Development Plan by the National Planning Commission acknowledges the need for more investment in research and development [6]. The South African Centre for Digital Language Resources (SADiLaR) was recently established as part of the South African Research Infrastructure Roadmap (SARIR). SADiLaR aims to address the need for access to large corpora of authentic digital data and applicable software tools to enable South African researchers to advance localised research endeavours in the Humanities, Social Sciences, and Information and Communication Technologies in order to address the challenges of unemployment, poverty and inequality [3].

However, researchers, educators, developers, service providers and funders need a roadmap to enable them to decide where to concentrate their efforts in order to give a maximum push to the development of a particular field [2], and to know what is available to enable further technology development and research. Technology audits are an important instrument to provide such a roadmap, with the result that in 2017/8, SADiLaR funded a project to undertake an audit of human language technology (HLT) resources in the country. The 2018 HLT Audit aimed to provide updated information on the availability and maturity of HLT resources in the country.

The first-ever HLT audit was conducted by the European Network of Excellence in Human Language Technologies (ELSNET) in 1991, and was based on the idea of a roadmap where information on HLT resources would be collected on a continual basis [4]. The dynamic nature of the ELSNET audit made it suitable for the fast changing nature of the HLT field and therefore suitable to be adapted to similarly gather information on the HLT resources available in South Africa. As a result, the audit of South African HLTs, undertaken by Sharma Grover in 2009 [1], took the ELSNET audit as point of departure.

2 METHODOLOGY

The 2018 Audit commenced with a process aimed at identifying and understanding the frameworks available to conduct HLT audits. This investigation uncovered few such frameworks, although a substantial number of references to the Language Resources and Evaluation Conference (LREC) and the Basic Language Resource Kit (BLaRK) initiated by ELSNET [4] were found. The ELSNET approach to their audit was to first conduct a workshop with experts in the field. This was followed by sharing the results via a website and inviting the HLT community to provide inputs. The inputs were then workshopped again, with the concept of a BLaRK emerging. This process continued in a cyclical fashion with researchers adding information about their work to the website and the BLaRK team updating the information on the website.

In 2008/9, Sharma Grover [1] adapted the BLaRK methodologies described above and undertook an audit of HLT resources in South Africa. Taking the Dutch BLaRK as point of departure, Sharma Grover redefined all the HLT components in detail and then produced the first detailed audit on South African HLT resources. The 2009 Audit classified the HLT resources into three categories, namely:

- Data
  - Linguistic data sets or collections (speech or text), in a machine-readable form, used to create, evaluate and improve HLT modules.
  - Includes corpora, lexicas and grammars.

SAICSIT 2018, September 26–28, 2018, Port Elizabeth, South Africa

- Modules
  - Basic software units or processes usually required to create HLT applications and products.
  - Includes part-of-speech taggers, sentence tokenisers, language models, acoustic models.
- Applications
  - Categories of different application areas where HLT is used.
  - Includes application domains such as speech input, document production, proofing/authoring tools, and translation.

The data gathered in the 2009 Audit was transferred to the National Department of Arts and Culture's Resource Management Agency (RMA), hosted by the Centre for Text Technology (CTexT) at the North-West University. The RMA subsequently obtained access to many of the resources identified in the 2009 Audit, and made these available via a catalogue (containing downloadable resources), and an index (listing non-downloadable resources).

3 AUDIT DESIGN AND APPROACH

3.1 2018 Audit design process

The 2018 HLT Audit initially aimed to replicate the 2009 HLT Audit, in order to provide comparable data. A similar process to that followed for the 2009 HLT Audit was thus embarked on. This process entailed the following:

- Familiarising ourselves with the 2009 Audit design process, including:
  - The HLT audit terminology development process
  - The HLT inventory criteria selection process
  - The process for defining the HLT components (and selecting priorities)
  - The HLT audit execution process
  - The HLT inventory gap analysis
- Deciding on the resource categories to be included in the design
- Compiling a list of respondents to be approached to participate in the Audit
- Reviewing the 2009 Audit tool (questionnaire) and determining fit-for-purpose for the 2018 Audit
- Obtaining a thorough understanding of the data analysis techniques used in the 2009 Audit.

3.1.1 Defining the structure of the 2018 Audit. In designing the 2018 Audit, we consulted HLT experts in order to assist us to modernise the design. In a workshop with these experts, we updated the component categories which form the basis of the audit; obtained inputs into the audit questionnaire; and compiled a list of institutions which would be approached to participate in the Audit.

The workshop attendants were divided into two working groups: one for speech resources and one for text resources. The working groups were tasked with the following:

- Reviewing the 2009 components
  - Determining which components are still relevant
  - Determining which components need to be changed, added or deleted
- Ensuring that components pertaining to all languages are covered.

The working groups agreed that the Modules and Applications categories are no longer applicable. We therefore only included a Data category and combined the Modules and Applications categories into a Software category. A Model category was added for speech components only. The Data, Model and Software categories were then updated with the resource types which fall into each category, and relevant metadata was added to each component.

Once we had updated the data categories and resource types, we needed to develop definitions for each of the resource types and provide technical descriptions to enable respondents to submit their resources under the correct headings. We nominated a sub-group of experts to assist with this task: three experts for text resources and three for speech resources.

3.1.2 Identifying the respondents. Parallel to the process of consulting with the HLT experts on the design of the Audit, we compiled a list of all individuals and institutions involved in HLT research and development in South Africa. This list comprises individuals (contacts) at universities, private companies and research institutions.

3.2 Audit workflow design

Participating in a technology audit can be a very cumbersome process. If the instrument used to collect the data has not been designed carefully, or is not completely fit-for-purpose, it can lead to a poor user experience and create a barrier to participation. The 2009 Audit employed a Microsoft Excel spreadsheet as the tool with which to collect the data. Navigating through the spreadsheet became cumbersome when large amounts of information needed to be entered. Negative feedback on the usability and user experience of the 2009 Audit instrument, led us to consider alternatives. We elected to use an online survey tool, instead of a spreadsheet.

In designing the workflow for the 2018 Audit, we studied the 2009 Audit questionnaire and discussed it with the HLT experts at the above-mentioned workshop. Based on these discussions, we designed a new workflow for the 2018 Audit. We defined a number of distinct pages, each containing/requesting information on a specific topic:

- The **Landing page** provides a brief introduction on the 2018 HLT audit, including an overview of how the 2018 Audit will work.
- The **Your Information page** allows users to complete their general information such as name, contact information and affiliation. Users can also choose to be contacted by SADiLaR to have their resource uploaded to the resource catalogue or index.
- The **Resource type page** allows users to select the type of resource that they are uploading, such as text, speech or multimodal.
  - The **Resource type - text selection page**. The user then selects whether their resource is Data or Software. Finally, under either Data or Software, the user may then select the resource type which their resource will be classified as.

  - The **Resource type - speech selection page**. The user then selects whether their resource falls under the Data, Model or Software category. Finally, under Data, Model or Software, the user may then select the resource type which their resource will be classified as.
  - The **Resource type - multimodal selection page**. The user then selects Multimodal corpora.
- The **Required information page** allows the user to complete information on the resource they are uploading. This information includes the name, description and keywords associated with the resource, the language(s) (should the resource be multilingual), the availability, and the cost of the resource.
- The **Technical description page** allows the user to complete further technical information on the resource under the Data, Model and Software pages - this is dependent on the resource type selected earlier in the questionnaire.
- The **Availability page** allows the user to indicate the model of distribution and the license associated with the resource.
- The **Quality page** allows the user to select to complete any protocols, standards and quality assurance methods followed in compiling the resource. Should a user select YES to this question, he/she will be prompted to answer follow-up questions that require detailed information.
- The **Documentation page** allows the user to include a more detailed description of the resource which may not have been covered elsewhere, as well as to upload any other documentation related to the resource.
- The **End page** thanks the user for his/her participation in the Audit and acknowledges the partners in the Audit.

Fig. 1 provides a high-level overview of the flow of the survey.

4 AUDIT INSTRUMENT DEVELOPMENT

4.1 Methodology and tool requirements

In selecting an appropriate instrument (tool) for conducting a technology audit, various factors need to be considered. These include cost, functionality and hosting, among other things. We defined the following requirements as a basis for selecting an audit tool:

- Client/user requirements:
  - Online tool (cloud-based or hosted in-house)
  - Attractive to the user (modern look and feel)
  - Clear and easy to use
  - Logical flow
- Functionality:
  - Drop down menus
  - Multiple choice options
  - Yes/No questions
  - Short narrative descriptions possible
  - Document/file upload available
- Technical requirements:
  - Accessible free-of-charge (open platform)
  - Accessible to invited participants (managed participation)
  - Multiple simultaneous inputs possible
  - Ability to store (large) documents (in specific format(s))
  - Ability to export to a database
  - Ability to convert raw data into Microsoft Excel format
- Success criteria:
  - Completeness of information received
  - Scalability
- Outputs:
  - Export raw data to Microsoft Excel format (required)
  - Dashboard with a consolidated view of the audit outcome (optional)
  - Transfer to client website/database (required).

4.2 Selection of an audit tool

We undertook an Internet search for online questionnaire/survey tools which would suit the needs of the 2018 Audit. We compared different tools, and selected an online tool called LimeSurvey [5].

LimeSurvey is leading open source survey software which is available as Software-as-a-Service (SaaS) or as a self-hosted Community Edition. LimeSurvey is a powerful survey tool which is highly customisable. We opted for the Community Edition, as the solution -

- can be self-hosted and is free of charge;
- is easy to set up and customise to the users' needs;
- meets the functionality requirements described above; and
- is accessible using a screen reader.

The user manual and the community forum were then utilised to self-learn the functionalities offered by LimeSurvey.

4.3 Configuring the audit tool

Limesurvey offers the functionality of creating a questionnaire using an existing template, or completely from scratch. Since none of the existing templates met the needs of the 2018 Audit, we created a questionnaire from scratch.

The properties for every questionnaire created can be changed to suit specific needs. To create a new questionnaire, the following is required:

- Questionnaire title
- Description
- Welcome message
- End message.

There are general settings for each created questionnaire which can be changed as needed. Some of these include:

- Administrator contact details
- How the questions are displayed (question by question vs question group by question group vs all questions on one page)
- Navigation settings (will the user be allowed to navigate backwards or not)
- Displaying the number of questions
- Displaying the progress a user is making
- Access to the questionnaire (open to everyone vs open to anyone who has an access token).

4.4 Developing the questionnaire

Careful consideration was given to the types of questions to be used for each piece of information required. Usability and user-experience further guided decisions on layout and wording.

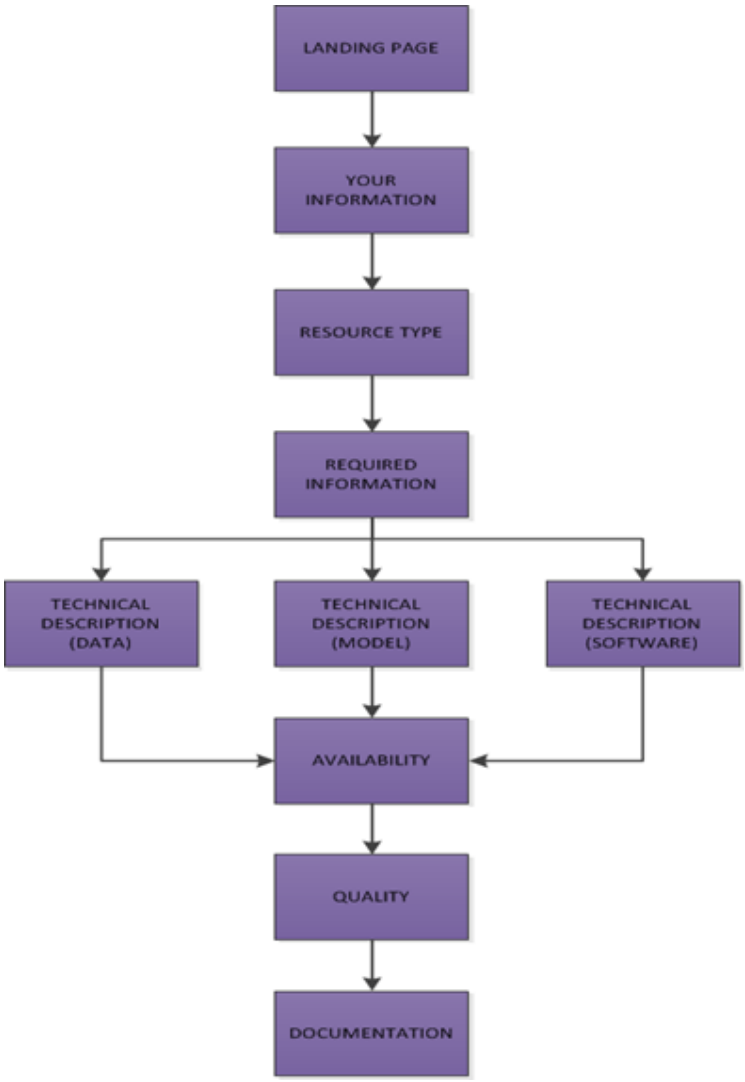


Figure 1: High-level overview of the 2018 Audit survey

The development of the online questionnaire consists of two sections, namely the back-end and the front-end (user interface). The questionnaire workflow was used as the basis for populating the online questionnaire in the back-end of LimeSurvey. Each question was manually created. This entailed:

- Typing the question
- Defining the question type
  - Short text, long text, multiple choice, multiple choice with comments, radio list, radio list with comments, drop down lists, yes/no questions, file upload questions, etc.
- Adding the predetermined answer options (for the multiple choice and radio list type questions)
- Creating conditions for certain questions (for example, "Ask Question 3 if the answer to Question 2 is blue").

LimeSurvey provides the user with a predefined "look" and "feel" (front-end) which met our needs and did not require modification. Fig. 2 and Fig. 3 below show the user interface for two of the pages of the questionnaire, namely the Resource Type page and the Technical Description page for Data.

4.5 Beta testing of the audit tool

A beta version of the Audit tool was tested with a small group of beta testers and the feedback was incorporated to the extent possible given the constraints of the online tool. Some of the changes made, based on the feedback received, included:

- Refining/rewording questions
- Changing conditions on certain questions
- Adding an "other" option to some multiple choice questions
- Adding a list of definitions for the components.

One of the current constraints of the Audit tool, is that it does not allow a user to copy the data from one submitted resource to

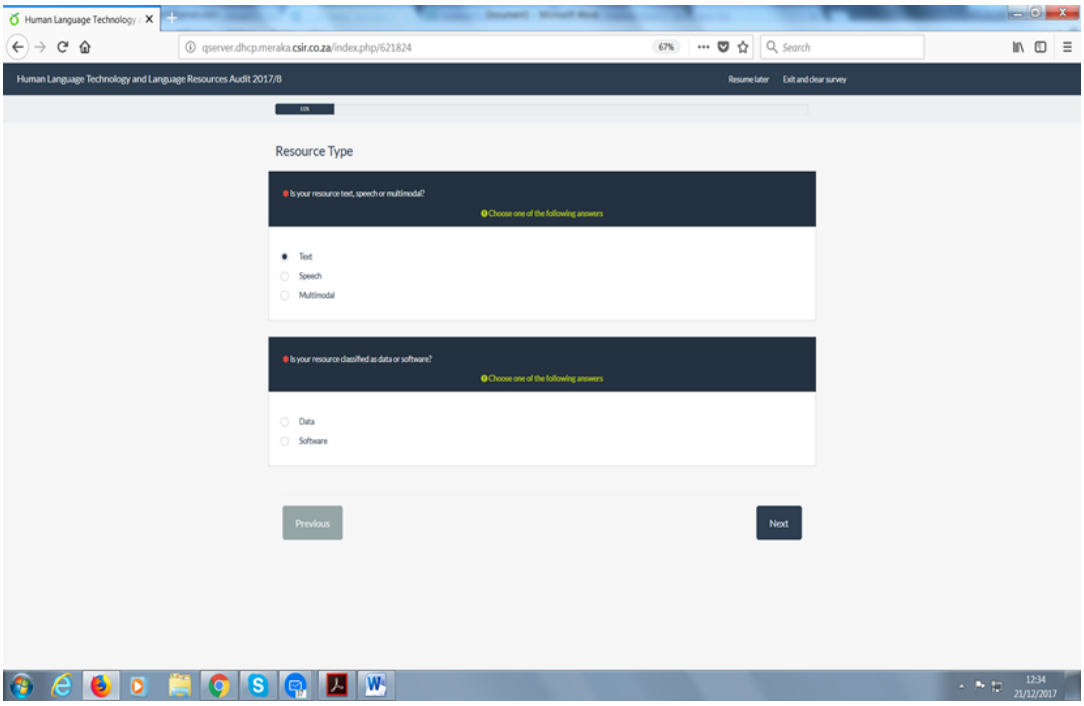


Figure 2: Resource Type page

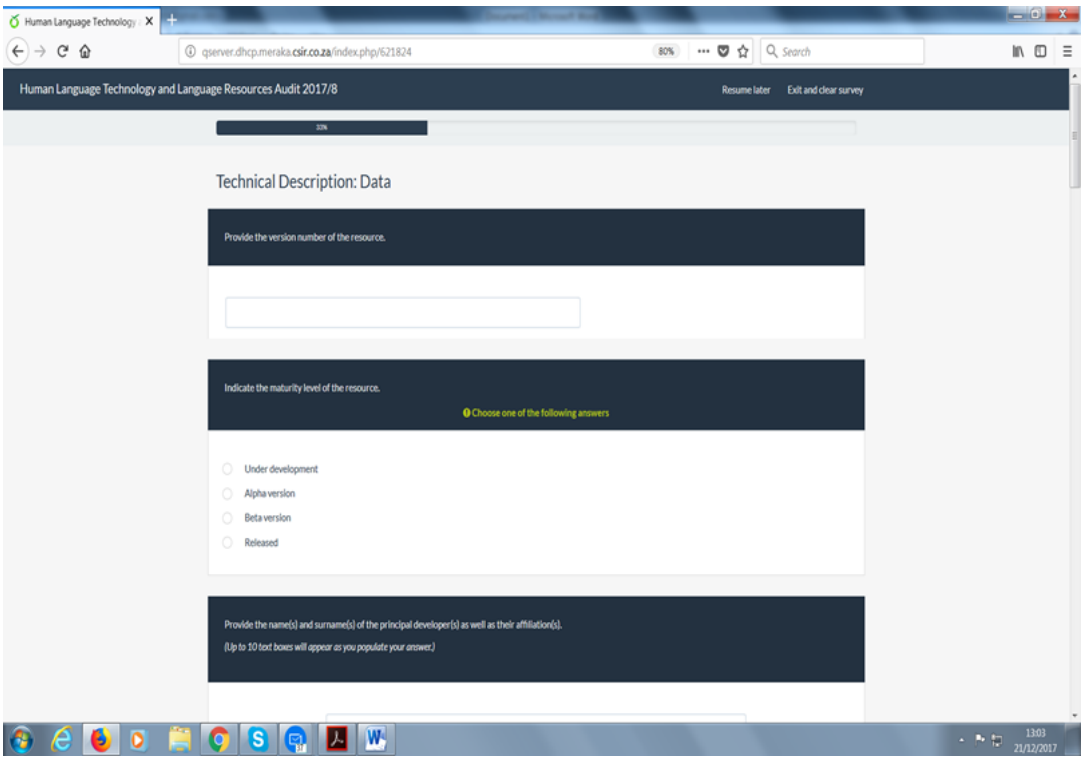


Figure 3: Technical Description: Data

enable multiple submissions of a similar resource, e.g. where only one or two fields differ across multiple similar resources.

A separate website [7] was also created to provide an easy reference to the list of definitions for the components, as adding all the definitions to the questionnaire would have cluttered the layout and overwhelmed the participants.

4.6 Providing access to the audit tool

For security purposes, we granted access to the questionnaire by issuing tokens to participants. Each token is valid for a certain number of uses - we set the limit at 100 uses as this was deemed to be sufficient (i.e. it is unlikely that one participant would upload more than 100 resources). A unique token was generated per participant and each participant was sent a personalised email containing a link to the questionnaire as well as their unique token.

5 AUDIT EXECUTION

5.1 Invitation to participate

During the audit design workshop, a decision was made to extend the 2018 Audit to include generic language resources in addition to HLT resources. This was communicated in the email notifying potential participants of the Audit. This email was sent to known members of the HLT community, as well as government departments, the National Lexicography Units of the Pan South African Language Board, publishers, private companies, professional associations, tertiary institutions (we targeted the language, computer science and engineering departments, as well as the language units and requested they disseminate the email to relevant colleagues at the institutions), and the mailing lists of the National HLT Network (NHN) and the Resource Management Agency. The Audit notification email was distributed on 5 and 6 December 2017. The notification email provided background information on the Audit, and requested potential participants to provide us with their contact details should they wish to participate. In addition, the recipients were requested to forward the email to other potential participants within their own networks.

Responses to the notification email generated an automated formal invitation to participate in the Audit. This invitation email contained a link to the online questionnaire (titled "Human Language Technology and Language Resources Audit 2017/8"), the participant's unique token (valid for up to 100 entries), as well as a link to the list of the definitions of the resource components.

5.2 Responses

The Audit spanned four months, from December 2017 to March 2018. Participants were initially given three months to complete the questionnaire. At the end of month two, follow-up reminder emails were sent out. These were followed by calendar schedulings and phone calls at the end of month three. The latter communication resulted in the extension of the deadline to accommodate additional responses.

A total of 26 completed responses were received. These responses included resources from eight different institutions across South Africa, as well as an institution situated in Germany. Of the 26 responses, 10 were speech-related and 16 were text-related. A total of 76 resources were submitted. An in-depth representation and analysis of the results will be presented in another paper.

6 CONCLUSION

The design and development of the 2018 Audit tool involved extensive research into past and current related audits and methodologies. The experts who participated in this process assisted in creating a simplified and modernised design for collecting information on existing HLT and language resources. The design was implemented in an online tool as method to collect the data. Both the design and the resultant tool can be re-used (with minimal effort) to design future audits (if required) and continually capture HLT resources as these become available.

Future work includes addressing the current challenges with the online tool, particularly the functionality to capture several similar resources with minimal effort. Further work includes implementing a system(s) to ensure that HLT resources (and other language resources) are continually submitted to SADiLaR as these become available. Raising awareness on the benefits of contributing to the body of knowledge and making resources available to others for further research and development, will require focused attention.

ACKNOWLEDGMENTS

The authors would like to thank the HLT experts who assisted with formulating and updating the component categories, and providing feedback on the flow of the questionnaire. The same gratitude is extended to all who participated in the 2018 Audit. The 2018 HLT Audit was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

REFERENCES

[1] Gerhard B. van Hyssteen Aditi Sharma Grover and Marthinus W. Pretorius. 2010. An HLT profile of the official South African Languages. In *Proceedings of the second workshop on African Language Technology (AFLaT 2010)*.
[2] Ulrike Bross. 1999. Technology audit as a policy instrument to improve innovations and industrial competitiveness in countries in transition. *Innovation: The European Journal of Social Science Research* 12, 3 (1999).
[3] South African Centre for Digital Language Resources. 2017. Research Infrastructure (RI) Proposal for the South African Research Infrastructure Roadmap (SARIR). Personal communication with the Authors.
[4] Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first Milestone for the Language Resources Roadmap. In *Proceedings of ELSNET's workshop on Speech and Computer (SPECOM 2003)*.
[5] LimeSurvey. 2018. Professional online surveys with LimeSurvey. https://www.limesurvey.org/
[6] Department of Science and Technology. 2016. South African Research Infrastructure Roadmap. First Edition.
[7] Human Language Technology and Language Resources Audit 2017/8. 2017. List of definitions. https://sites.google.com/view/hlt-audit-definitions/home

Human Language Technology Audit 2018: Analysing the development trends in resource availability in all South African languages

Carmen Moors
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
cmoors@csir.co.za

Ilana Wilken
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
iwilken@csir.co.za

Karen Calteaux
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
kcalteaux@csir.co.za

Tebogo Gumede
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
tgumede@csir.co.za

ABSTRACT

Almost a decade has passed since the first audit on the state of HLT in South Africa was published in 2009. An increase in HLT R&D in South Africa since then, as well as developments in language resource management in the country surfaced the need for an updated audit of HLT resources. Consolidating information on the availability and maturity of HLT resources provides valuable information for both researchers and decision-makers. On the one hand, information on available HLT resources enables researchers to identify new opportunities for multidisciplinary research. On the other, decision-makers can use the information to determine priorities for resource development and where to focus their investments. The paper presents an overview of the main findings of an audit of HLT resources, undertaken in 2017/8, and makes some suggestions for ensuring that the current information is continually updated as new resources are developed.

CCS CONCEPTS

• **Applied computing** → Language translation; Digital libraries and archives;

KEYWORDS

Human language technology, Technology audit, Language resources, Text resources, Speech resources, Digital Humanities

ACM Reference Format:
Carmen Moors, Ilana Wilken, Karen Calteaux, and Tebogo Gumede. 2018. Human Language Technology Audit 2018: Analysing the development trends in resource availability in all South African languages. In *Proceedings of 2018 Annual Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2018).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

SAICSIT 2018, 26-28 September 2018, Port Elizabeth, South Africa

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

South Africa is a highly multilingual country in which communication barriers and the digital divide are still largely prevalent. Research and development (R&D) activities on language form a fertile field of study which can contribute to nation-building, as well as regional growth and economic development.

The South African government has realised the role which technology, and in particular human language technology (HLT), can play in bridging communication barriers and addressing the digital divide. In addition, there has been an increase in HLT R&D and digital humanities as fields of study since the first HLT Audit was undertaken in 2009. Making HLT resources available to stimulate multi-disciplinary R&D is therefore an important activity within the South African R&D landscape.

HLTs fall roughly into two main categories - text technologies and speech technologies. Through its custodian of language matters, the Department of Arts and Culture, the South African government has contributed significantly to promoting HLT R&D and the development of HLT tools, data and applications for both text and speech technology. In addition, there have been several industry-funded initiatives to develop and make available HLT data, tools and applications in the South African languages over the last 15 years.

In order to make decisions about future investment in ICT R&D, HLT R&D, digital humanities-related R&D, and language resource management and infrastructure development, a unified picture of the current South African language technology domain (among others) is required. One way of achieving this is through a technology audit.

Technology audits aim to identify, assess and catalogue technologies according to different criteria, ranging from categories of technologies, maturity of technologies, competitive position, location in the supply chain, and levels of competencies; through to the impact of technologies [2]. By employing a mapping technique known as a technology matrix, technology audits can provide an

overview of the related technology landscape in a company, market or country. This information can then be used in various decision-making scenarios such as, among others, investment decisions to be taken by the newly-established South African Centre for Digital Language Resources (SADiLaR).

This paper emanates from an HLT audit project aimed at updating our knowledge of the available HLT resources in South Africa, undertaken in 2018. To enable comparison with previous data, the 2018 HLT Audit replicated the 2009 HLT Audit [3].

2 METHODOLOGY

The 2018 HLT Audit commenced with a process aimed at identifying and understanding the frameworks available to conduct HLT audits. This investigation uncovered few such frameworks, although a substantial number of references to the Language Resources and Evaluation Conference (LREC) and the Basic Language Resource Kit (BLaRK) initiated by ELSNET [7] were found. The ELSNET approach to their audit was to first conduct a workshop with experts in the field. This was followed by sharing the results via a website and inviting the HLT community to provide inputs. The inputs were then workshopped again, with the concept of a BLaRK emerging. This process continued in a cyclical fashion with researchers adding information about their work to the website and the BLaRK team updating the information on the website.

In 2009, Sharma Grover [4] [5] adapted the BLaRK methodologies described above and undertook an audit of HLT resources in South Africa. Taking the Dutch BLaRK as point of departure, Sharma Grover redefined all the HLT components in detail and then produced the first detailed audit on South African HLT resources. The 2018 HLT Audit built on these approaches.

3 THE DATA

3.1 Data categorisation

The 2018 HLT Audit was designed under the guidance of local experts in the field of HLT. While the data categories from the 2009 HLT Audit were taken as point of departure, these were modernised in the 2018 HLT Audit.

In a workshop with these experts, we updated the component categories which form the basis of the audit; obtained inputs into the audit questionnaire; and compiled a list of institutions which would be approached to participate in the Audit.

The 2009 HLT Audit divided the resources into four resource categories:

- DATA - split between text and speech resources
- MODULES - split between text and speech resources
- APPLICATIONS - split between text and speech resources
- TOOLS - split between text and speech resources.

In the 2018 HLT Audit, Modules, Applications and Tools are collapsed into one category (Software), and a Models category is added (only applicable to speech resources), resulting in the following three categories being defined:

- DATA - split between text and speech resources
- MODELS - only includes speech resources
- SOFTWARE - split between text and speech resources and consolidates the previous MODULES, APPLICATIONS AND TOOLS categories.

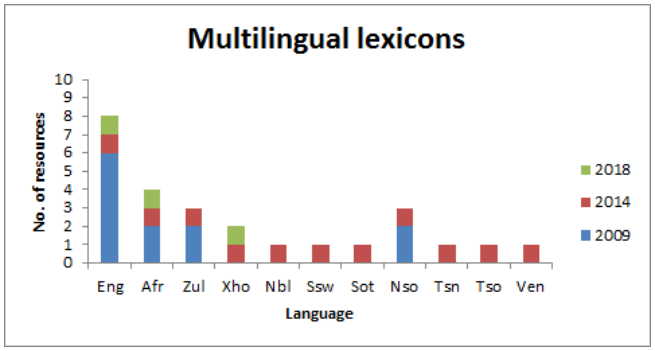


Figure 1: Multilingual lexicons

In addition to the datasets collected during the 2009 and 2018 HLT Audits, we obtained a dataset from the Resource Management Agency (the government-appointed custodians of HLT resources in the country and the precursor to the newly-established SADiLaR) containing the resources which had been added subsequent to the 2009 Audit. We therefore present an analysis of three datasets of HLT resources in this paper. We term these the 2009 dataset (emanating from the 2009 HLT Audit), the 2014 dataset (resources added between the two HLT Audits), and the 2018 dataset (emanating from the 2018 HLT Audit).

3.2 Data analysis process

In order to compare the three datasets, we match the resource types across all three datasets to one another, although we use the 2018 resource categories to structure the presentation of the results. We only match resource types for which there is at least one entry in one dataset.

Because the resource types differ across the three datasets, the matching of resource types results in three sets of comparisons being made:

- A comparison of resource types which match across all datasets (2009, 2014 and 2018)
- A comparison of resource types which match across two datasets (2009 and 2014), but do not match those in the 2018 data
- A representation of the resource types from the 2018 data that cannot be matched to data in the 2009 and 2014 datasets.

The comparison of the resources available for the Multilingual Lexicon resource type presented in Figure 1, is an example of a comparison which could be made across all three datasets, and clearly indicates the growth in this resource type over the last 10 years.

4 RESOURCE DEVELOPMENT PROGRESS AND TRENDS

In Figure 2 the comparison of the three datasets indicates that while significant progress has been made since 2009 to develop

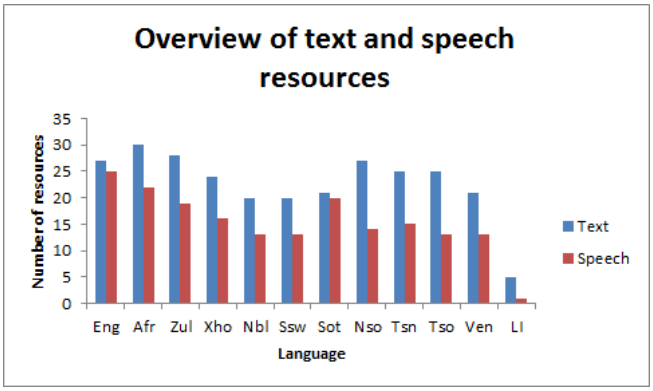


Figure 2: Overview of text and speech resources

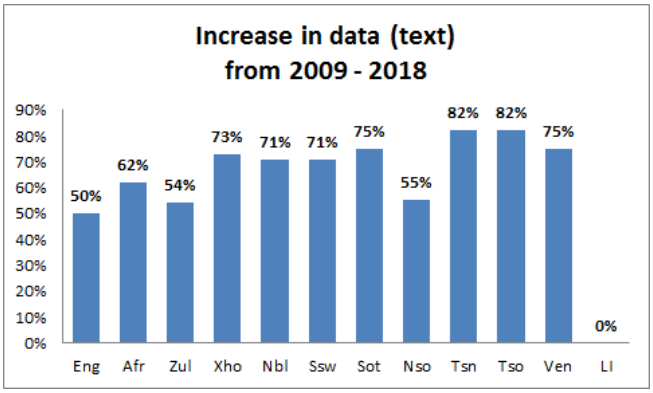


Figure 3: Increase in data (text)

additional resources across more languages, to develop cutting-edge resources, and to develop language independent resources, the more marginalised indigenous languages (particularly Xitsonga, Tshivenda, Sesotho, siSwati, and isiNdebele), remain severely under-resourced. In addition, there are more text resource types than speech resource types across all the languages.

In the Data category, there is a notable increase from 2009 to 2018, in text resources for Sesotho, Setswana, Xitsonga and Tshivenda, and in speech resources for English and Afrikaans as presented in Figures 3 and 4.

In the Software category, there is a notable increase in text resources for 10 of the 11 South African official languages, but this is not the case in speech resources. The 100 percent increase in language independent speech resources relates to one application and is not significant. This is presented in Figures 5 and 6.

5 DATA ANALYSIS

5.1 Maturity and accessibility indicators

In this section, we compare resource types in terms of their levels of maturity and accessibility. We use the indicators set out in Table 1 (adapted from [3]) to evaluate the maturity and accessibility of existing resources.

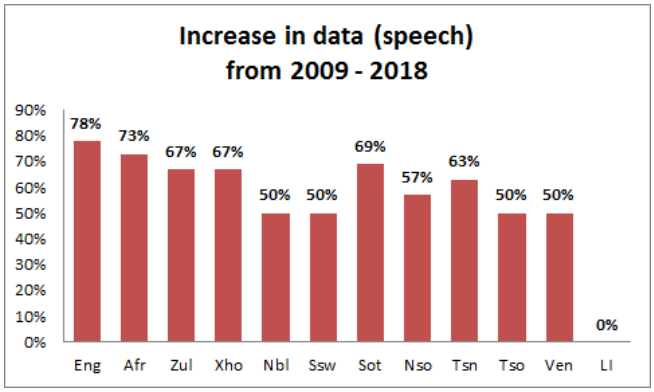


Figure 4: Increase in data (speech)

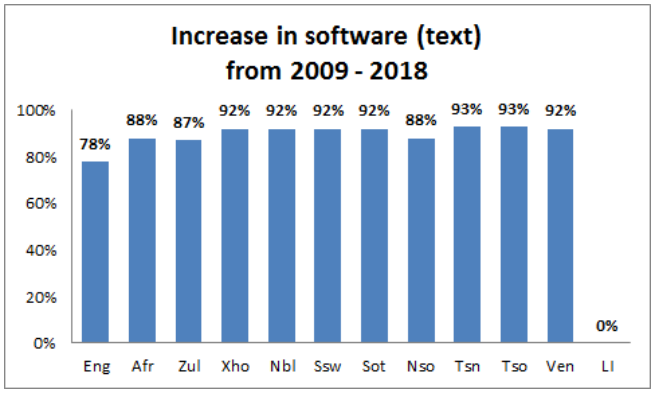


Figure 5: Increase in software (text)

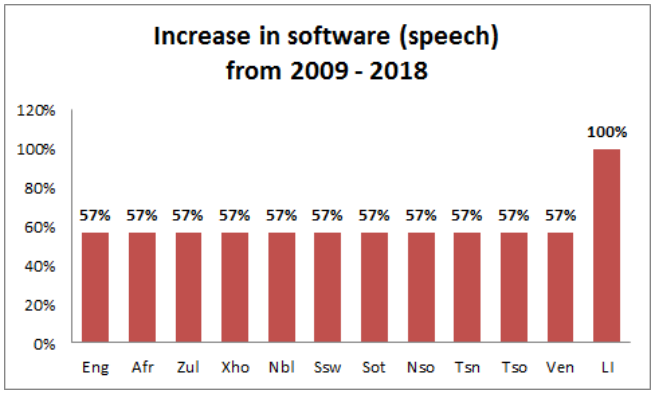


Figure 6: Increase in software (speech)

Based on a combination of its maturity and accessibility, each resource is categorised as either a full resource (F), or a partial resource (P): A full resource is a resource which is fully mature (released) and is fully available (commercially, openly/freely), i.e. has full circles in terms of maturity and accessibility. A partial resource is a resource which is only partially mature (under development, alpha version, beta version) and only partially available

Maturity		Accessibility	
Level	Representation	Level	Representation
Under development	◑	Not available/ proprietary/ closed	◑
Alpha version	◑	Undecided	◑
Beta version	◑	Research	◑
Released	●	Commercial	●
		Open/ freely available	●

Table 1: Representation of maturity and accessibility

(not available/proprietary/closed, undecided, for research purposes), i.e. features half circles in terms of maturity and accessibility.

5.2 Maturity and accessibility sums

As a means to compare the available resources across resource types and across languages, a maturity sum and accessibility sum was calculated for each language. These calculations were made based on the number of full resources per resource type. The partial resources have not been added to these calculations, as there is no way of knowing whether or not a partial resource has since become a full resource.

5.2.1 Maturity Sum. The maturity sum provides a measure of the maturity of resources in a language. There are four maturity stages, each with an associated weight. The maturity sum is calculated per resource type by multiplying the number of resources at each maturity stage with the weight associated with the maturity stage, and then summing the products. The weight assigned to a maturity stage is double that of the preceding maturity stage and can thus be either 1 (under development), or 2 (alpha version), or 4 (beta version), or 8 (released version). Table 2 shows how the maturity sum for Afrikaans resource types is calculated. It is important to note here that for this purpose, the term maturity refers to whether or not a resource has been released, and not to the size, scope or coverage of the resources.

Fig. 7 shows an overview of the maturity sums for each language. South African English and Afrikaans are the most mature languages, followed closely by isiZulu, Sesotho, Sepedi, Setswana and Xitsonga. IsiNdebele, siSwati and Tshivenda are the least mature languages. Language independent resources are also shown, but have a very low maturity compared to the language-specific resources.

Fig. 8 illustrates an overview of the maturity sums for all the resource types for which full resources exist in the South African languages. From this chart, it is evident that speech corpora is the most mature resource type across all languages. Text corpora and speech recognition systems are the second and third most mature resource types. Speech-to-speech translation systems and grammatical framework resource grammars are the least mature of the South African languages.

5.2.2 Accessibility Sum. The accessibility sum provides a measure of the accessibility of resources in a language. The accessibility sum is calculated in a similar manner to the maturity sum. For each resource type, the number of resources at each accessibility stage is multiplied with the weight associated with the accessibility stage, before summing the products. The weight assigned to an accessibility stage is double that of the preceding accessibility stage, however, the last stage is a combination of the third and fourth stages, and thus those weights are added together to create the weight for the last stage. In other words, the weights can be either 1 (not available/proprietary/closed), or 2 (undecided), or 4 (research), or 8 (commercial), or 12 (open/freely available). Table 3 shows how the accessibility sum for Afrikaans resources is calculated.

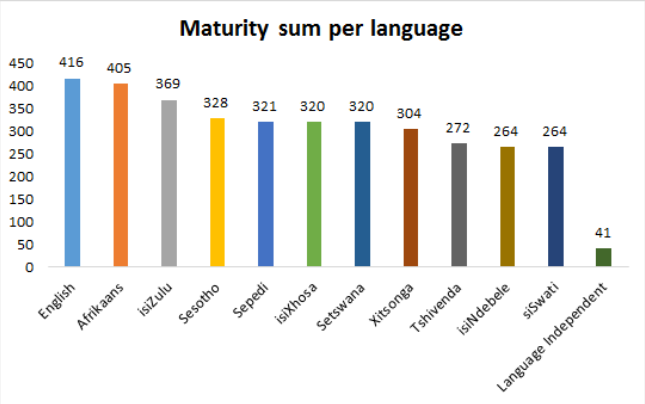


Figure 7: Maturity sum per language

Figure 9 shows an overview of the accessibility sums for each language. Afrikaans resources are the most accessible, followed by South African English and isiZulu. Sesotho, isiXhosa, Setswana, Xitsonga and Sepedi follow within close proximity. Tshivenda, isiNdebele and siSwati are the least accessible of the languages. Language independent resources are also shown, but have a very low maturity compared to the individual languages.

Figure 10 illustrates an overview of the accessibility sums for all the resource types for which resources exist in the South African languages. From this chart, it is evident that text corpora is the most accessible resource type across all languages. Terminology integration texts, anonymisers and PDF converters are the least accessible resource types currently available.

5.2.3 HLT Component Sum. The HLT Component Sum provides an overview of the status of HLT development for the eleven official South African languages. The HLT Component Sum is calculated by adding the Maturity Sum and the Accessibility Sum for each language:

$$HLT\ Component\ Sum = Maturity\ Sum + Accessibility\ Sum$$

Figure 11 shows the HLT Component Sum for resource types across the official languages. The chart indicates that text corpora and speech corpora are the most developed resource types. The least developed resource types are speech-to-speech translation systems.

Table 2: Maturity sum (MS) for Afrikaans

Resource type	Number of resources	MS per resource type	Total
Monolingual lexicon	1	**(1 x 1)** + (2 x 0) + (4 x 0) + (8 x 0)	1
Multilingual lexicon	4	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 4)**	32
Text corpora	8	(1 x 0) + (2 x 0) + **(4 x 1)** + **(8 x 7)**	60
Pronunciation dictionary	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Speech corpora	13	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 13)**	104
Lemmatiser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Morphological analyser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Machine translator	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Language and dialect identifier	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Machine aided human translation system	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Web service	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Corpus analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Tokeniser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Named entity recogniser	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Compound analyser	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Chunker	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
OCR/ICR	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Integrated automatic annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Telephony applications	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
Speech recognition systems	3	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 3)**	24
Acoustic analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 2)**	16
Annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)**	8
		Total	405

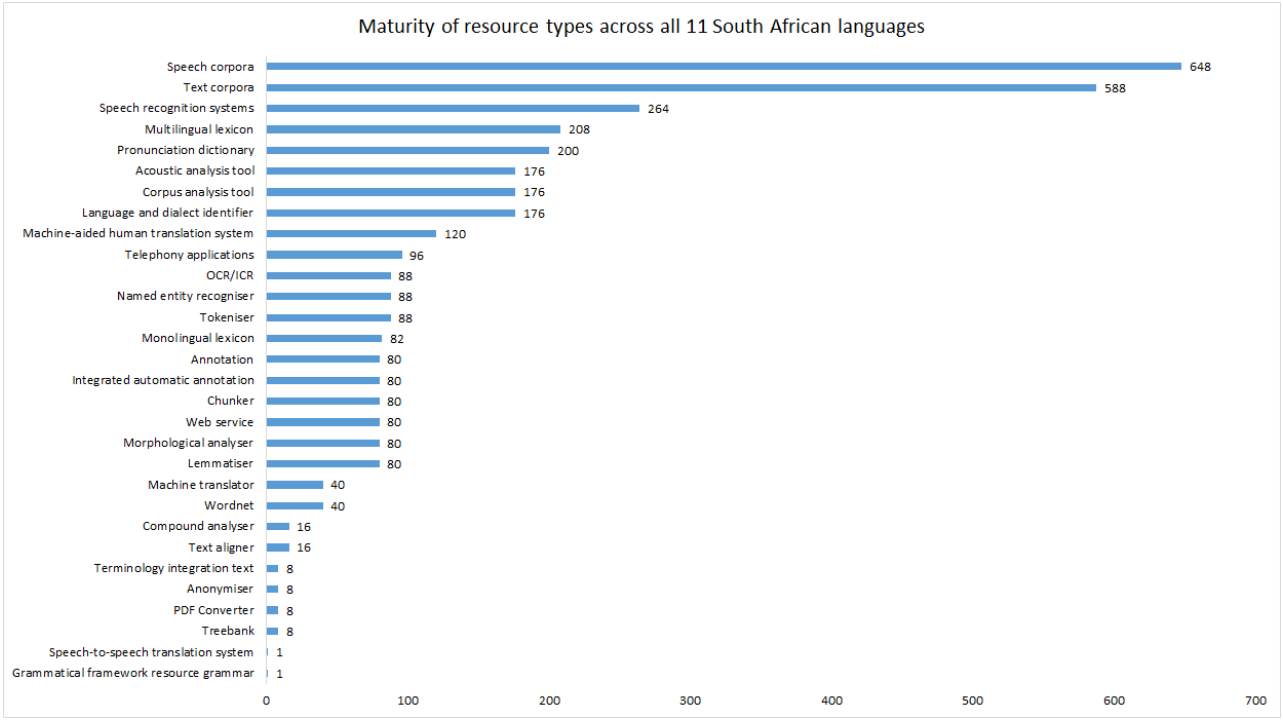


Figure 8: Maturity sum per resource type

Table 3: Accessibility sum (AS) for Afrikaans

Resource type	Number of resources	MS per resource type	Total
Monolingual lexicon	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Multilingual lexicon	4	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + **(12 x 2)**	32
Text corpora	9	(1 x 0) + (2 x 0) + **(4 x 1)** + (8 x 0) + **(12 x 7)**	92
Pronunciation dictionary	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Speech corpora	13	(1 x 0) + (2 x 0) + **(4 x 5)** + (8 x 0) + **(12 x 8)**	116
Lemmatiser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Morphological analyser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Machine translator	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Language and dialect identifier	2	(1 x 0) + (2 x 0) + (4 x 0) + **(8 x 1)** + **(12 x 1)**	20
Machine aided human translation system	2	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	8
Web service	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Corpus analysis tool	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Tokeniser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Named entity recogniser	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Compound analyser	2	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 2)**	24
Chunker	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
OCR/ICR	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Integrated automatic annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
Telephony applications	1	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	4
Speech recognition systems	3	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + **(12 x 1)**	20
Acoustic analysis tool	2	(1 x 0) + (2 x 0) + **(4 x 2)** + (8 x 0) + (12 x 0)	8
Annotation	1	(1 x 0) + (2 x 0) + (4 x 0) + (8 x 0) + **(12 x 1)**	12
		Total	500

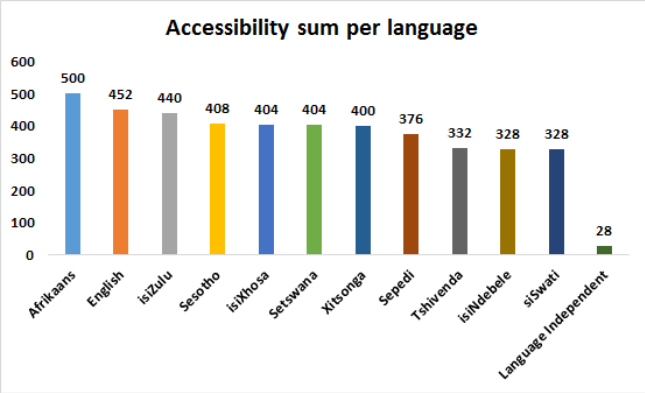


Figure 9: Accessibility sum per language

5.3 Overview of existent and non-existent resource types

Table 4 lists the resource types from the Data, Models and Software Categories for which full, partial or no resources exist in any of the datasets (2009, 2014 or 2018). The table classifies the resource types according to the 2018 Audit classification. We have not captured the resource types of 2009 and 2014 for which no resources were submitted as the 2018 Audit resource types will be used in future. This table indicates that there are a number of resource types for which resources still need to be developed.

6 RECOMMENDATIONS

The following section outlines some recommendations for increasing the accessibility of South African language resources.

6.1 Dynamically capture new resources as these become available

An online tool, such as that developed for the 2018 HLT Audit, greatly facilitates the capturing of language resources by the developers thereof, as these become available. The benefits of an online tool include that it minimises the administrative burden to submit and verify resources, is cost effective, and can be modified as new data management requirements emerge. In addition, the online tool selected for the 2018 Audit can be navigated with a screen reader (verified with five commonly-used screen readers), thereby making it accessible to all.

6.2 Manage the process to ensure resources are submitted to the SADiLaR database

We propose that a similar process to that followed by ELRA, be implemented in South Africa. This involves requiring authors of papers and articles which report on South African language resources, to provide proof of submission of such resources to SADiLaR before publications are finally accepted. This would require agreements to be put in place with the organisers of relevant conferences and the editorial boards of relevant journals to facilitate this process. Such

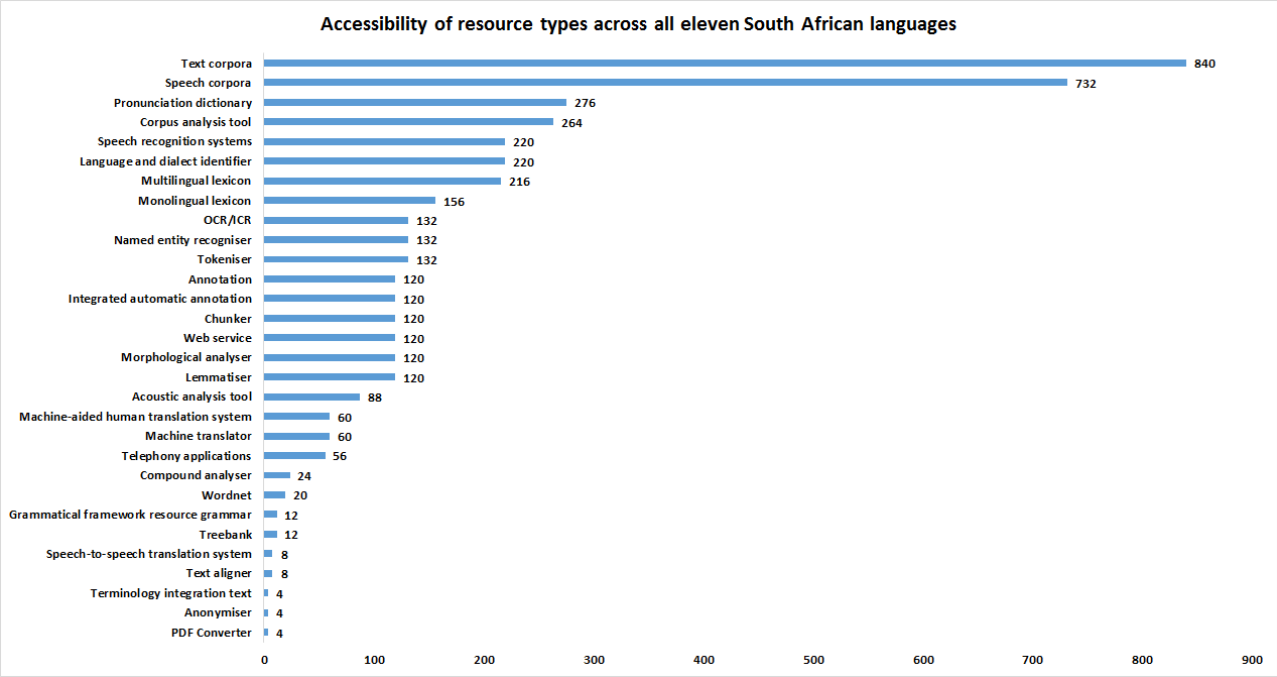


Figure 10: Accessibility sum per resource type

an endeavour would need to be supported by extensive awareness-raising on the importance and benefits of purposeful language resource management.

7 CONCLUSION

It was beyond the scope of this paper to compare our results directly to results of audits for other languages and countries, due to the divergent nature of HLT resources and their development. Some perspective might be gleaned, however, by comparing the most mature resource type for the South African languages, namely speech corpora, with speech corpora available for other languages. For example, the largest transcribed speech corpus for ASR for the South African languages is the NCHLT Speech corpus, which consists of approximately 56 hours of speech for all 11 languages. By comparison, the first Latvian speech recognition corpus consists of 100 hours of speech [1], an Icelandic speech corpus based on parliamentary data consists of 542 hours of speech [6], and the Librispeech corpus for English consists of roughly 1000 hours [8]. It is therefore evident that even South Africa's most mature speech corpus is still far smaller than those available in other countries.

In summary, the 2018 HLT Audit follows a similar approach to that of the 2009 HLT Audit. In the design phase of the 2018 HLT Audit, the 2009 approach is modified slightly based on various inputs from experts in the field. An online tool is selected to collect the data, improving on the manual process followed in 2009. The analysis and consolidation of the 2018 HLT Audit data is done manually and a selected set of analyses is presented in this paper.

By comparing the data from the two HLT Audits undertaken in South Africa to date and that available from submissions directly to the Resource Management Agency (now SADiLaR), we are able to obtain detailed insight into the available HLT resources per resource category, resource type and language. We are also able to identify resource types and languages which are lagging behind in terms of their development. Although text resource development has out-performed speech resource development, there are still many resources that need to be developed in both text and speech.

The data collected through the 2018 Audit is available for further analysis by researchers and for decision-making on future resource development investment. We propose that more awareness should be raised on the accessibility of HLT resources for R&D and the impact of dedicated language resource development activities.

The establishment of the South African Centre for Digital Language Resources (SADiLaR) marks a new era in language resource management in South Africa. SADiLaR presents a government-funded, formal, longer term (at least 10 years) infrastructure for curating and managing language resources. A formal process for submitting HLT resources and continually updating the database of HLT resources, is now possible. The current thinking is to follow a similar procedure to that implemented for the LRE Map in Europe, by requiring researchers to submit resources mentioned in research papers and articles to SADiLaR before publication of such articles and papers. To this end, the online survey developed in this project (as reported on in a related paper) will be hosted by SADiLaR in future. In this manner, we aim to create a sustainable system for capturing, curating and distributing HLT resources in South Africa.

ACKNOWLEDGMENTS

The authors would like to thank the HLT experts who assisted in designing the 2018 Audit and everyone who participated in the Audit.

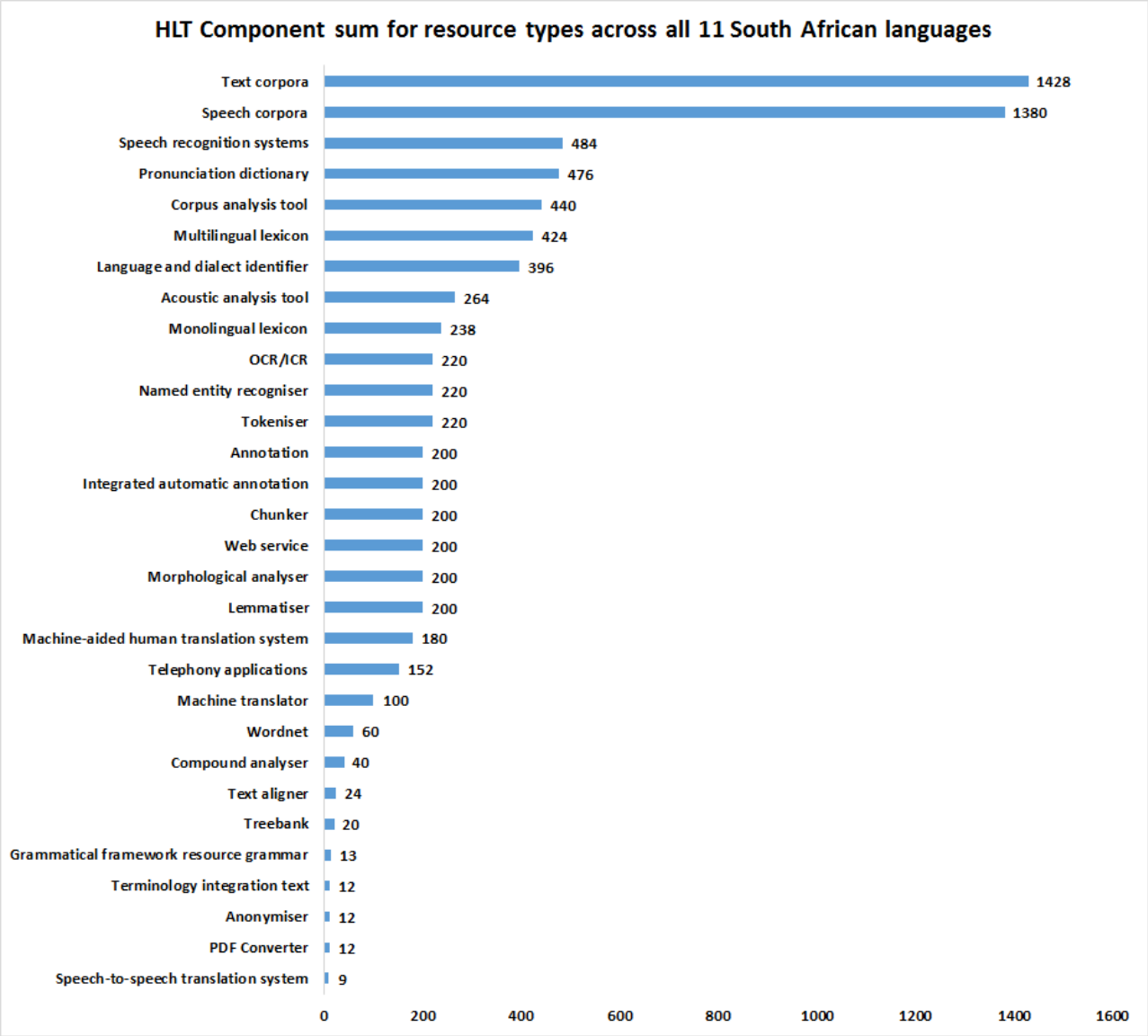


Figure 11: Component sum per resource type

The 2018 HLT Audit was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

REFERENCES

[1] M arcis Pinnas, Ilze Auzina, and K arlis Goba. 2014. Designing the Latvian Speech Recognition Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

[2] Dilek Cetindamar, Robert Phaal, and David Probert. 2010. *Technology Management Activities and Tools*. Palgrave Macmillan, Hampshire, England.

[3] Aditi Sharma Grover. 2009. *Technology Audit: The State of Human Language Technologies R&D in South Africa*. Master's thesis. University of Pretoria.

[4] Aditi Sharma Grover, Gerhard B. van Huyssteen, and Marthinus W. Pretorius. 2010. An HLT profile of the official South African Languages. In *Proceedings of the second workshop on African Language Technology (AFLaT 2010)*.

[5] Aditi Sharma Grover, Gerhard B. van Huyssteen, and Marthinus W. Pretorius. 2011. The South African Human Language Technology Audit. *Language Resources and Evaluation* 45, 3 (Sept. 2011), 271–288.

[6] Inga Rún Helgadottir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guonðson. 2017. Building an ASR corpus using Althingi's Parliamentary Speeches. In *Proceedings of Interspeech 2017*.

[7] Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first Milestone for the Language Resources Roadmap. In *Proceedings of ELSNET's workshop on Speech and Computer (SPECOM 2003)*.

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Table 4: Existent and non-existent resources in South African languages

		Fully available	Partially available	Non existent			Fully available	Partially available	Non existent
Data: Text	Monolingual lexicon	■	■		Software: Text (cont.)	Machine translator	■		
	Multilingual lexicon	■	■			Semantic analyser: Frame extractor			■
	Terminology list		■			Language and dialect identifier	■		
	Controlled vocabulary			■		Shallow parser: Relation finder			■
	Named entities list			■		Proofing/authoring tool		■	
	Thesaurus			■		Information retrieval system			■
	Wordnets	■	■			Information extractor			■
	Ontologies			■		Human-aided machine translation system		■	
	Text corpora	■	■			Machine-aided human translation system	■		
	Treebanks	■	■			OCR/ICR	■		
	Statistical language model		■			Computer-aided language learning (CALL) system		■	
	Formal grammar			■		Document classifier			■
	Tagset			■		Authorship identifier			■
	Lexical database		■			Question answering (QA) system			■
	Test suites and test corpora*		■			Dialogue system (text-based)			■
	Other text resources*		■			Comprehension assistant		■	
Data: Speech	TTS sentence splitting rule sets (manually created)			■		Web service	■		
	TTS tokenisation rule sets (manually created)			■		Grammatical framework resource grammar	■		
	TTS normalisation rule sets (manually created)			■		Corpus analysis tool	■		
	TTS language ID rule sets (manually created)			■		Compound analyser	■	■	
	Language grammar			■		PDF Converter	■		
	Phoneme sets			■		Anonymiser	■		
	Phone mappings		■			Terminology integration text	■		
	Pronunciation dictionaries	■	■			Text aligner	■		
	Pronunciation rule sets (manually created)			■		Text selection tool		■	
	Intonation rule sets (manually created)			■		Web crawler		■	
	Phrasing rule sets			■		Annotation	■	■	
	Tone rules sets			■		Parameter search		■	
	Stress rule sets			■		Integrated automatic annotation	■		
	Syllabification rule sets			■	Software: Speech	Language modelling tool			■
	Speech corpora	■	■			Pronunciation dictionary creation tool			■
	Text corpora for speech			■		G2P tool			■
Corpora	Multimodal corpora		■			Acoustic modelling tool			■
Models: Speech	TTS Sentence Splitting models			■		Intonation tool			■
	TTS Tokenisation models			■		Phrasing tool			■
	TTS Normalisation models			■		Tone tool			■
	TTS LID models			■		Stress tool			■
	Language models			■		Syllabification tools			■
	Acoustic models			■		Vocoder			■
	G2P models			■		TTS Sentence Splitting tool			■
	Intonation models		■			TTS Tokenisation tool			■
	Phrasing models			■		TTS Normalisation tool			■
	Tone models			■		TTS LID tool			■
	Stress models			■		Large vocabulary speech recognition system	■	■	
	Syllabification models			■		Command and control system			■
Software: Text	G2P Converter		■			Non-native speech recognition system			■
	Tokeniser	■	■			Code-switched speech recognition system			■
	Sentenciser			■		Multilingual speech recognition system	■	■	
	Spelling corrector			■		Noise robust speech recognition system		■	
	Full-form normaliser			■		Embedded speech recogntion system	■	■	
	Format normaliser		■			Alignment system			■
	Number normaliser			■		Automatic phonetic transcription system			■
	Diacritics normaliser			■		Confidence measures			■
	Anonymiser			■		Acoustic Language ID			■
	Lemmatiser	■	■			Acoustic Age ID			■
	Stemmer			■		Acoustic Gender ID			■
	Morphological analyser	■	■			Acoustic Dialect ID			■
	Morphological synthesiser			■		Acoustic Emotion ID			■
	Part-of-speech tagger/disambiguator		■			Key-word spotting system			■
	Syllabifier			■		Voice activity detection			■
	Hyphenator		■			Speaker tracking			■
	Dependency parser			■		Acoustic speaker ID			■
	Constituent recogniser			■		Speaker verification system			■
	Chunker	■	■			Diarisation			■
	Event extractor			■		Complete TTS System		■	
	Named entity recogniser	■				Speech-to-speech translation system	■	■	
	Terminology extractor			■		Audio search		■	
	Topic modelling			■		Access control		■	
	Sentiment analysis/affect/emotion analyser			■		Speaking devices		■	
	Referent resolver			■		Accessibility		■	
	Word meaning disabmiguator			■		Telephony applications	■	■	
	Pragmatic analyser			■		Acoustic analysis tool	■		
	Text generator			■		Multimodal information access		■	
	Summariser			■		Speech-based tools*		■	

*These resource types existed for the 2009 and 2014 data sets. The types were not included in the 2018 list of resource types, but we still decided to list these items for the purposes of indicating existent resources.

HLT Audit 2017/2018: Results dissemination workshop

CSIR Meraka Institute, Building 43 - Auditorium,

Thursday, 26 July 2018

09:00 – 12:15

Purpose of workshop **Disseminate the results of the HLT Audit 2017/2018**

Topic	Responsibility	Time
Arrival and tea/coffee	All	09:00 – 09:30
Welcome and introduction	CSIR, KC	09:30 – 09:45
Project overview	CSIR, IW	09:45 – 10:00
Results of the 2017/2018 Audit	CSIR, CM	10:00 – 10:15
Comparative analysis	CSIR, CM	10:15 – 10:45
Break	All	10:45 – 11:00
Gap analysis	CSIR, IW	11:00 – 11:30
Way forward and discussion	CSIR, KC	11:30 – 12:00
Workshop closure	SADiLaR. AdL and CSIR, KC	12:00 – 12:15

HLT Audit 2017/2018

Audit results dissemination workshop

26 July 2018

Carmen Moors and Ilana Wilken

Contents

- Project overview
- Results of 2017/2018 Audit
- Comparative analysis of 2009 – 2018 Audit
- Gap analysis
- Way forward

Project overview

Background

- Timeline = 1 July 2017 – 30 June 2018
- **Work packages**

1. Audit design
2. Audit instrument development
3. Audit execution and results
4. Analysis of results and dissemination of findings
5. Dynamic Audit update solution

Audit design, instrument development and execution

Audit design

- Audit design workshop with experts
- Revised categories
  - Old: **Data**, **Modules**, **Applications**, **Tools**
  - New: **Data**, **Models**, **Software**
- Alignment with RMA
- MS Visio design

Audit instrument development

- LimeSurvey selected as Audit tool
- Configured LimeSurvey

Audit execution and results

- Audit invitations and responses

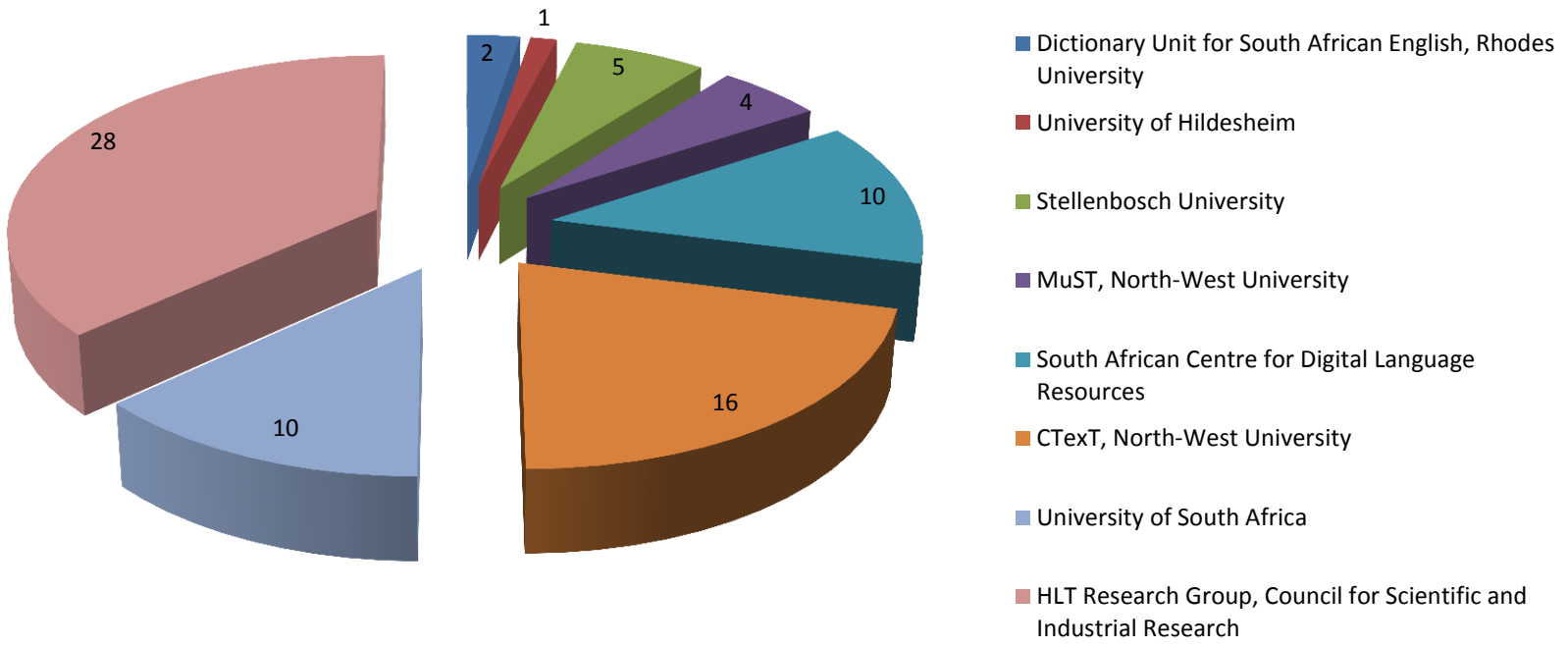
Analysis of results and dissemination of findings

Results analysis and findings dissemination

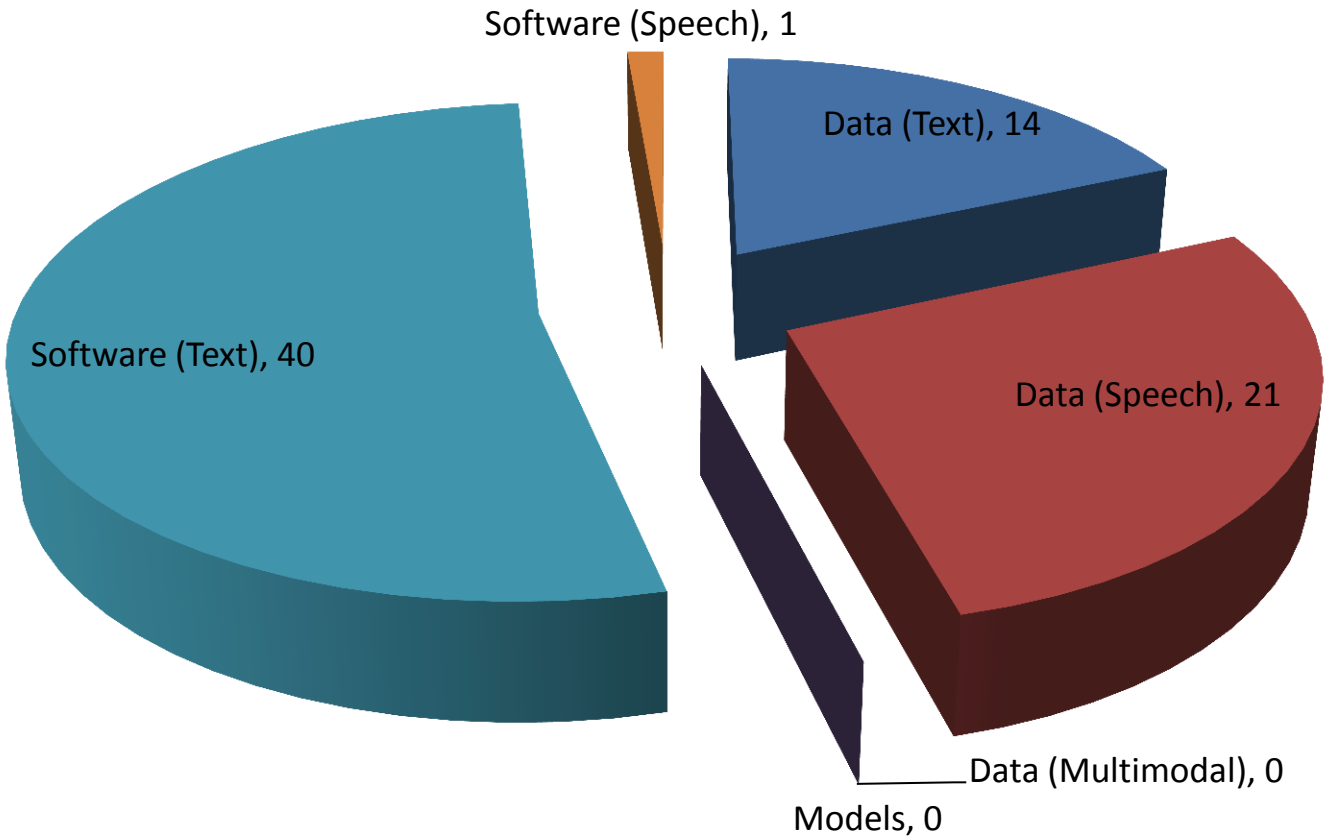
- Representation of 2017/2018 Audit results
- Comparative analysis of 3 datasets
- Gap analysis
- 2 x conference papers (pending acceptance)
  - **“Human Language Technology Audit 2018: Design considerations and Methodology”**
  - **“Human Language Technology Audit 2018: Analysing the development trends and gaps in resource availability in all South African languages”**
- Possible international conference in 2019/2020
- Dissemination of Audit results workshop – 26 July 2018

2017/2018 Audit results

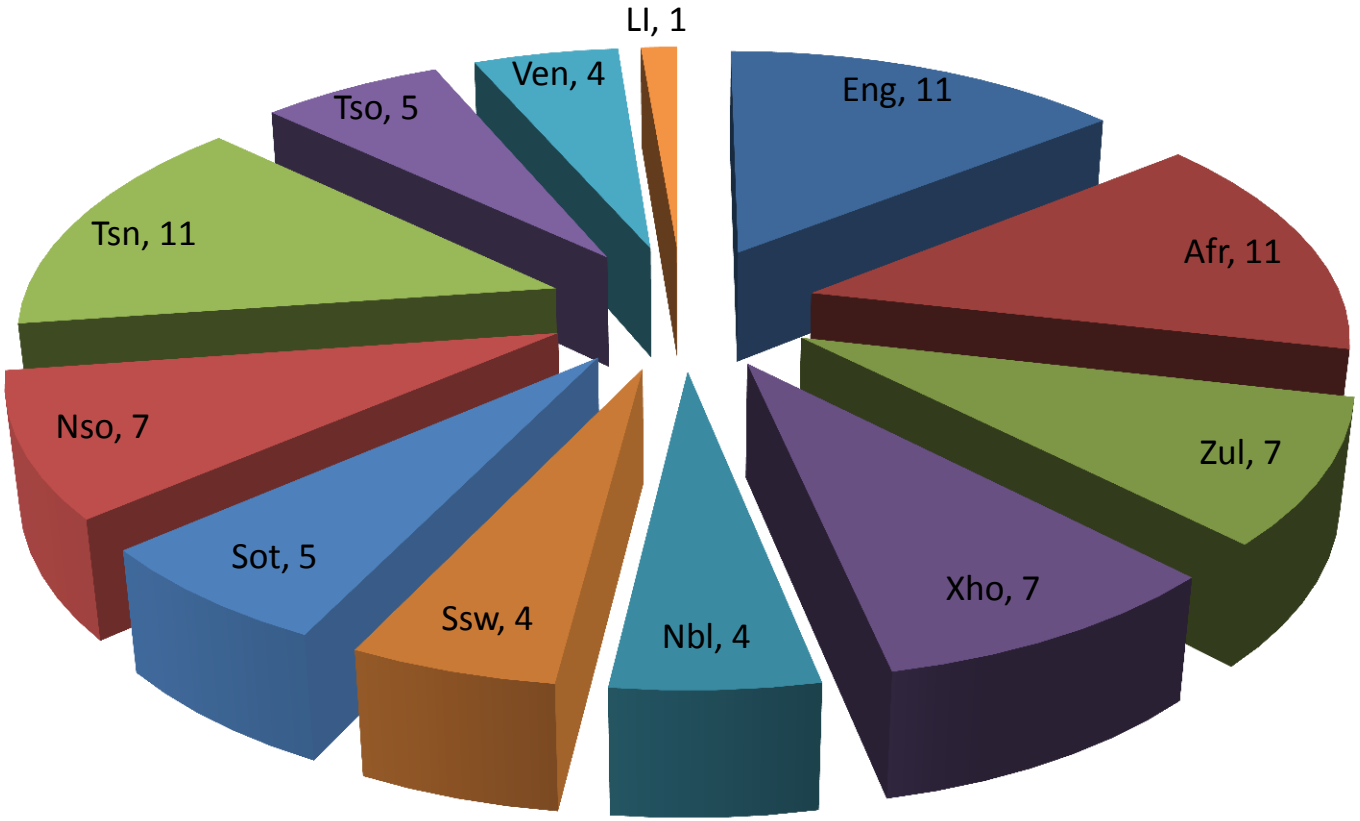
Resources per institution



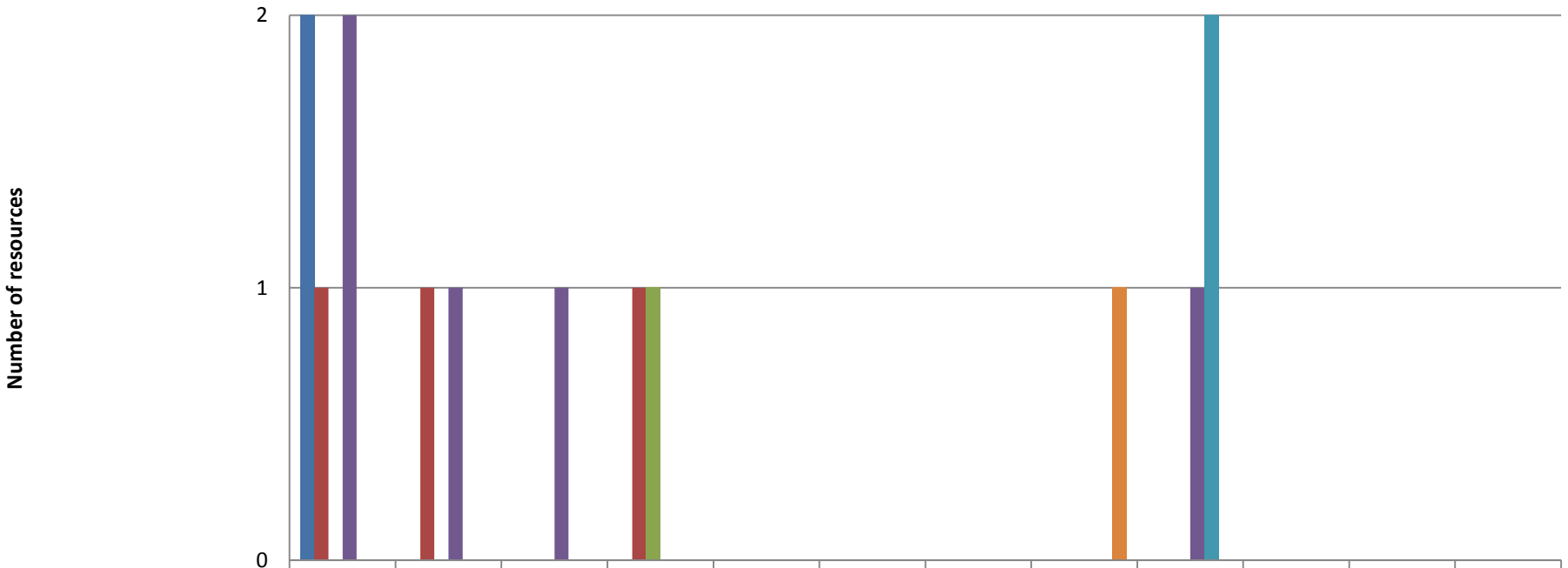
Resources per category



Resources per language

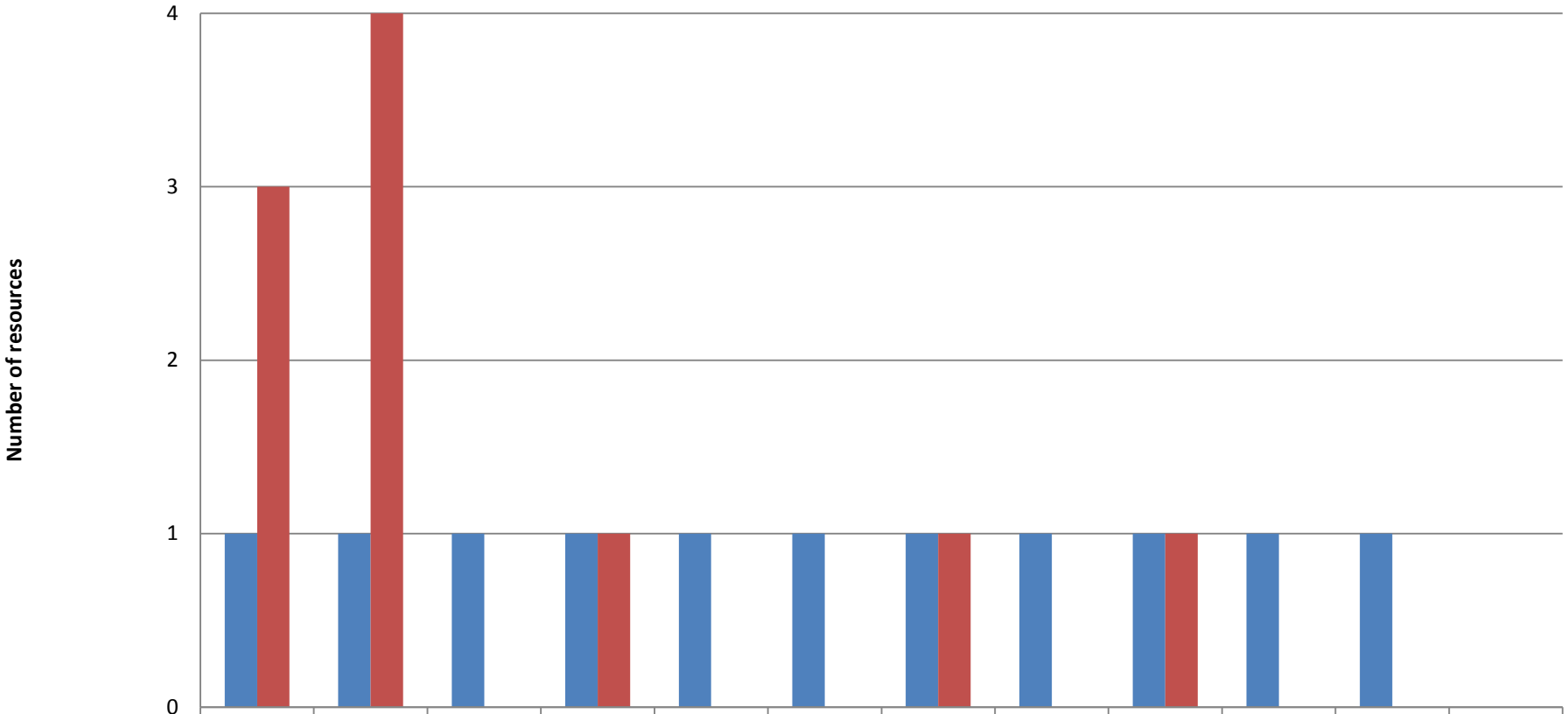


Resource per language for Data (text)



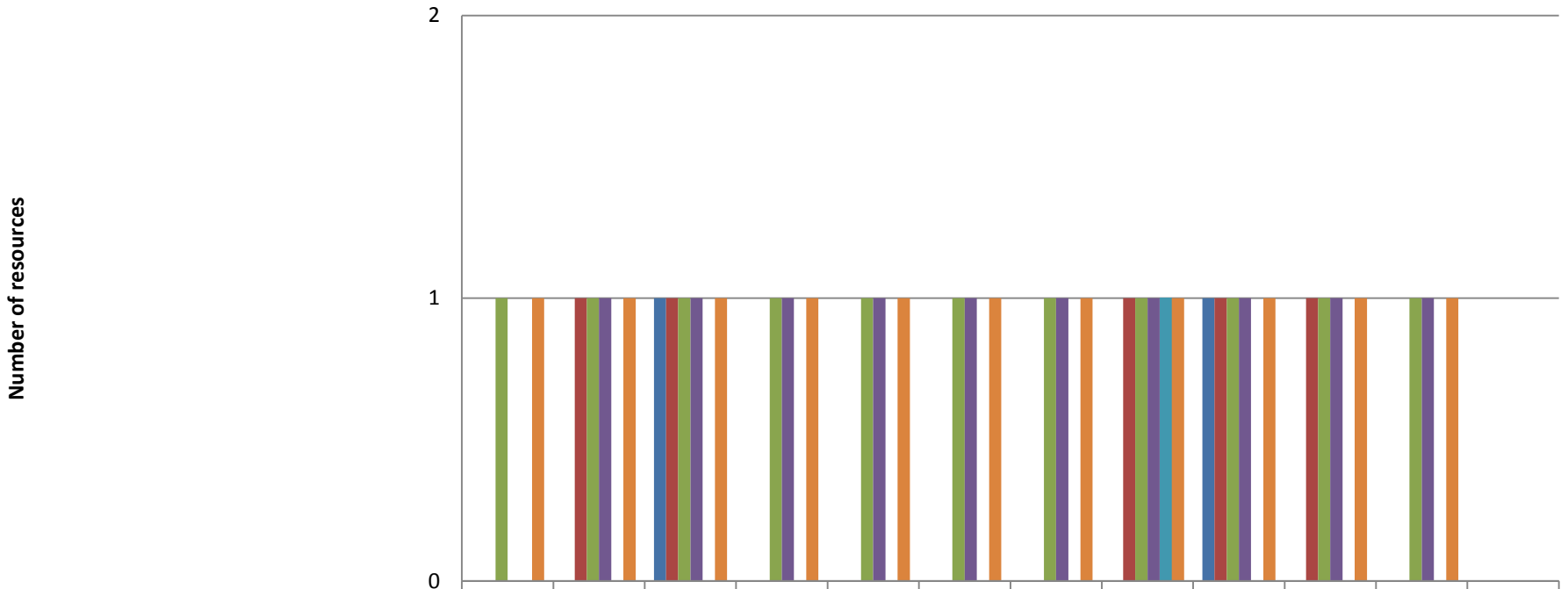
	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	Ll
Monolingual lexicon	2											
Multilingual lexicon	1	1		1								
Monolingual text				1								
Multilingual text	2	1	1						1			
Treebank									2			
Statistical language model								1				

Resource per language for Data (speech)



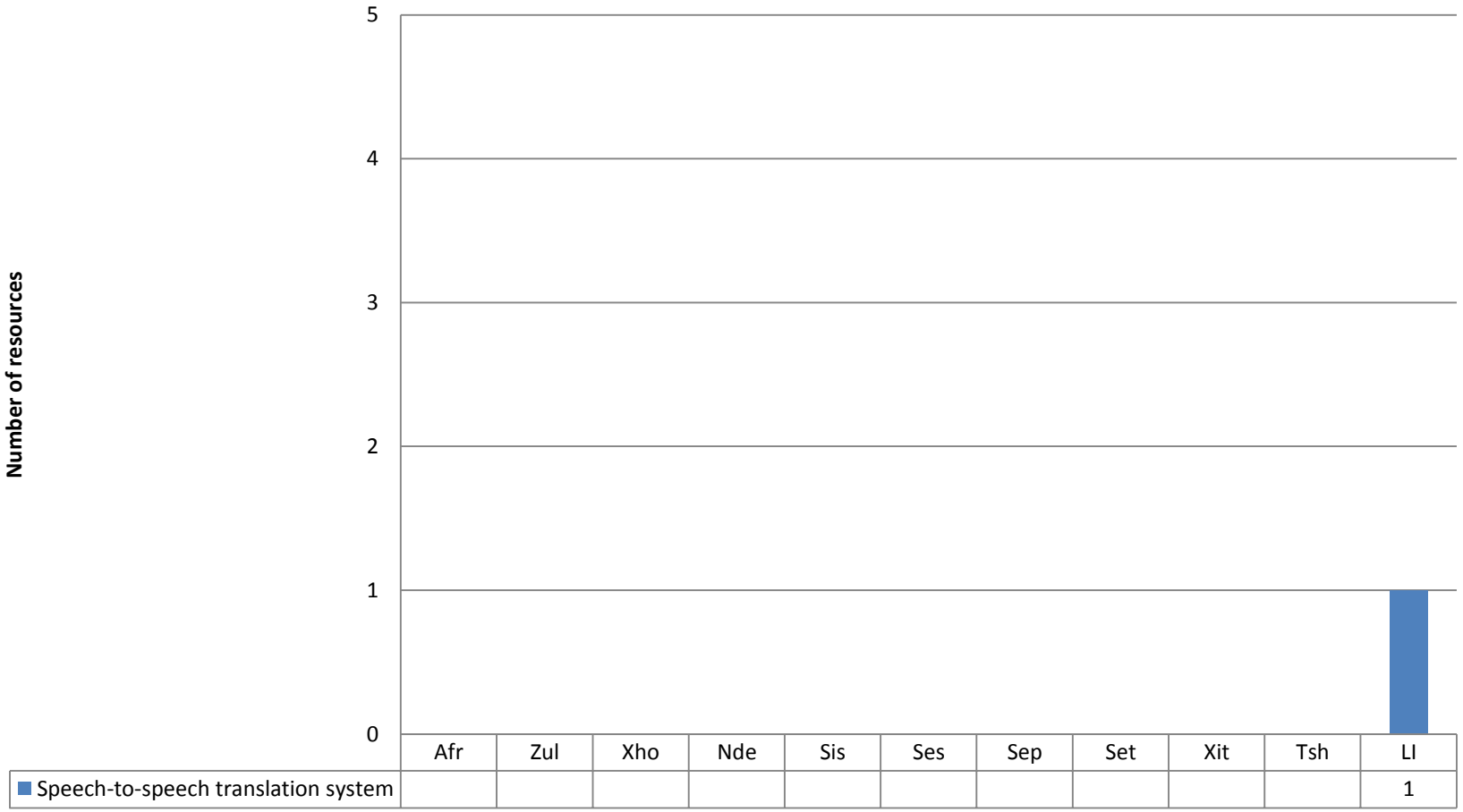
	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Phone mappings	1	1	1	1	1	1	1	1	1	1	1	
Speech corpora	3	4		1			1		1			

Resource per language for Software (text)



	Eng	Afr	Zul	Xho	Nbl	Ssw	Sot	Nso	Tsn	Tso	Ven	LI
Morphological analyser			1						1			
Machine translator		1	1					1	1	1		
Language and dialect identifier	1	1	1	1	1	1	1	1	1	1	1	
Web service		1	1	1	1	1	1	1	1	1	1	
Grammatical Framework resource grammar								1				
Corpus analysis tools	1	1	1	1	1	1	1	1	1	1	1	

Resource per language for Software (speech)



Maturity of resources

Maturity level	Category	Resource type	Language/s	Number
Under development ◑	No resources submitted			0
Alpha version ◑	Data (Text)	Multilingual text	Eng, Zul, Afri, Tsn	5
		Treebanks	Tsn	2
Beta version ◑	Data (Speech)	Speech corpora	Eng, Afr	2
		Phone mappings	Eng, Afr, Zul, Xho, Ven, Nbl, Ssw, Sot, Nso, Tsn, Tso	11
Released ●	Data (Text)	Monolingual lexicon	Eng	1
		Statistical language model	Sot	1
		Monolingual text	Xho	1
	Data (Speech)	Speech corpora	Xho, Afr, Tsn, Sot, Eng	8

Availability of resources

Availability	Category	Resource type	Language/s	Number
Commercial ●	Software (Speech)	Speech-to-speech translation system	All	11
	Software (Text)	Language and dialect ID	All	11
Open/ freely available ●	Data (Text)	Monolingual lexicon	Eng	1
		Multilingual lexicon	Eng, Afr, Xho	3
		Treebank	Tsn	1
	Data (Speech)	Speech corpora	Afr, Xho, Sot, Tsn	4
	Software (Text)	Other (web service)	Afr, Xho, Zul, Ven, Nbl, Ssw, Sot, Nso, Tsn, Tso	10
		Other (corpus analysis tools)	All	11
		Machine translation	Afr, Zul, Nso, Tsn, Tso	5
		Other (grammatical framework resource grammer)	Tsn	1
Research ◑	Data (Text)	Monolingual lexicon	Eng	1
		Statistical language model	Sot	1
	Data (Speech)	Phone mappings	All	11
		Speech corpora	Afr, Eng	4
Not available/ proprietary/ closed ◑	Software (Text)	Morphological analyser	Zul, Tsn	2
Undecided ◑	Data (Speech)	Speech corpora	Eng, Afr	2
	Data (Text)	Multilingual text	Eng, Zul, Afr, Tsn	5
		Monolingual text	Xho	1
		Treebank	Txn	1

Summary of 2017/2018 Audit results

- Most resourced languages: English, Afrikaans, isiZulu, isiXhosa, Sepedi and Setswana
- Most submissions received: Software (text) category

Comparative analysis

Comparative analysis process

- To determine if resource types have increased from the 2009 Audit to the 2018 Audit
- Matched resources types over all datasets to compare

Conducted three sets of comparisons:

- A comparison of resource types which matched across all datasets
- A comparison of resource types which matched across two datasets (2009 and 2014), but did not match those in the 2018 data
- A representation of the resource types from the 2018 data that could not be matched to 2009 and 2014 data

Measured maturity and availability of resources

- Capture full or partial resources based on measures in table below

Maturity and availability measures

Maturity		Availability	
Level	Representation	Level	Representation
Under development	◑	Research	◑
Alpha version	◑	Commercial	●
Beta version	◑	Open/ freely available	●
Released	●	Undecided	◑
		Not available/ proprietary/ closed	◑

Comparative analysis (3 datasets)

Data (Speech and text)			Software (Speech and text)		
Resource type	**Full**	**Partial**	**Resource type**	**Full**	**Partial**
Monolingual text	X	X	*Lemmatisers*	X	X
Multilingual text	X	X	*Morphological analysers*	X	X
Monolingual lexicons	X	X	*POS Taggers/ disambiguators*	X	X
Multilingual lexicons	X	X	*Text-based tools*	X	X
Speech corpora	X	X	*Speech-based tools*	X	X
Pronunciation dictionaries	X	X	*Speech recognition systems*	X	X
Intonation models		X	*(Language pair dependent) machine translators*	X	
Wordnets	X	X	*Language and dialect identifiers*	X	
Terminology lists		X	*Machine-aided human translation*	X	
Treebanks	X	X	*Human-aided machine translation*		X
			Format normalisers		X
			Compound analysers		X
			Shallow parsers/constituent recognisers		X
			Automatic phonetic transcriptions		X
			Tokenisers		X
			Limited domain TTS		X
			Domain independent TTS		X
			Hyphenaters		X
			Proofing/authoring tools		X

Comparative analysis (2 datasets)

Data (Speech and text)			Sotware (Speech and text)		
Resource type	**Full**	**Partial**	**Resource type**	**Full**	**Partial**
Lexical databases		X	*G2P convertors*		X
Other text resources		X	*Computer-assisted Language Learning*		X
Test suites and test corpora		X	*Audio search*		X
			Access control		X
			Speaking devices		X
			Telephony applications	X	X

Unmatched resource types (1 dataset)

Data (Speech and text)			Sotware (Speech and text)		
Resource type	**Full**	**Partial**	**Resource type**	**Full**	**Partial**
Statistical language models		X	*Web services*	X	
Phone mappings		X	*Grammatical framework resource grammars*	X	
			Corpus analysis tools	X	
			Speech-to-speech translation systems	X	X

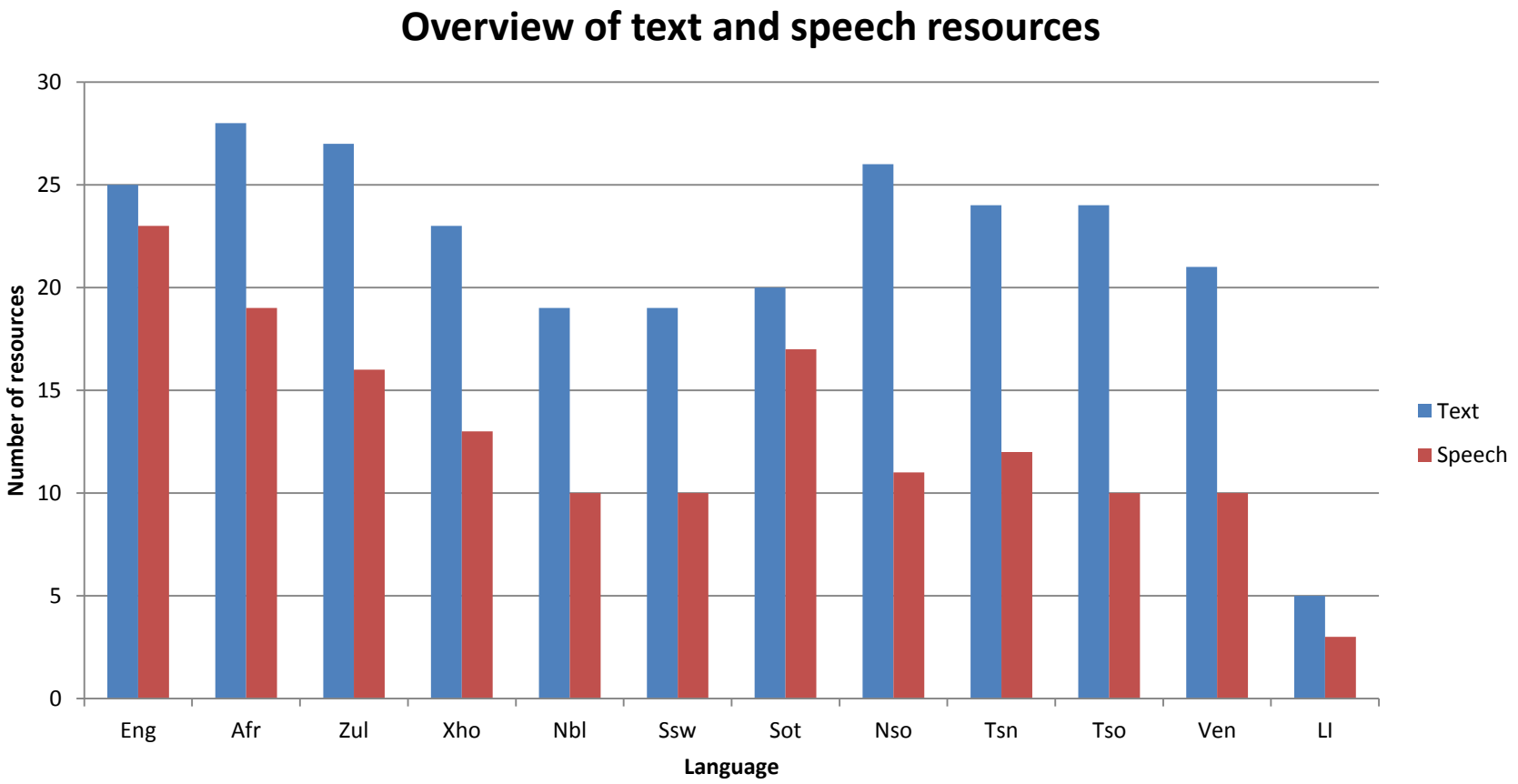
Summary of comparative analysis results

- The comparative analysis represents the resource development trends and development progress made from the 2009 Audit to the 2018 Audit

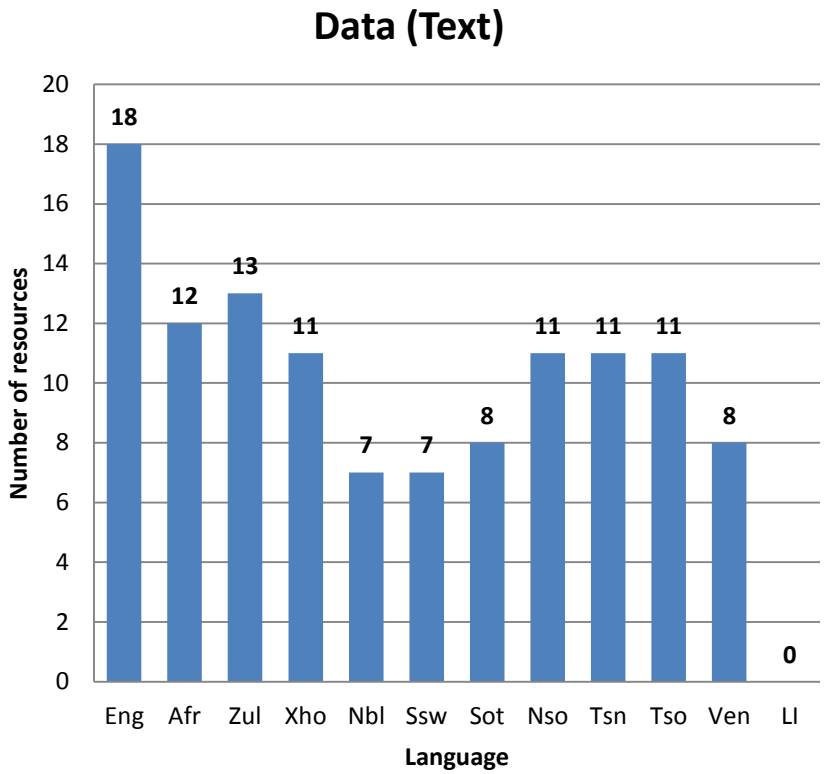
Refer to graphs below

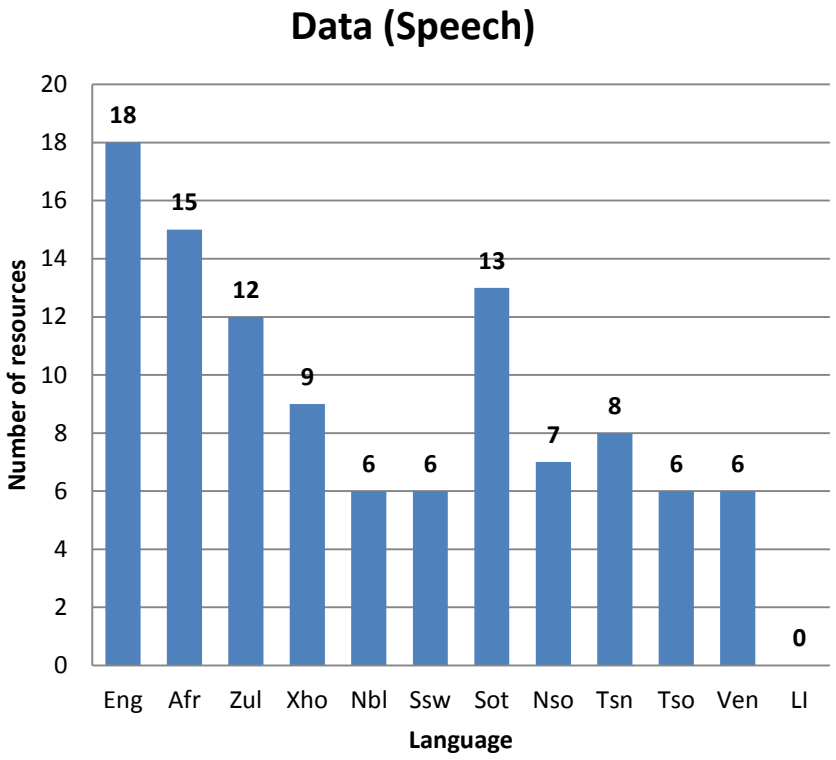
- Significant progress was made to develop additional resources across more languages
- However, more marginalised languages such as Xitsonga, Tshivenda and isiNdebele remain under-resourced

Available text and speech resources

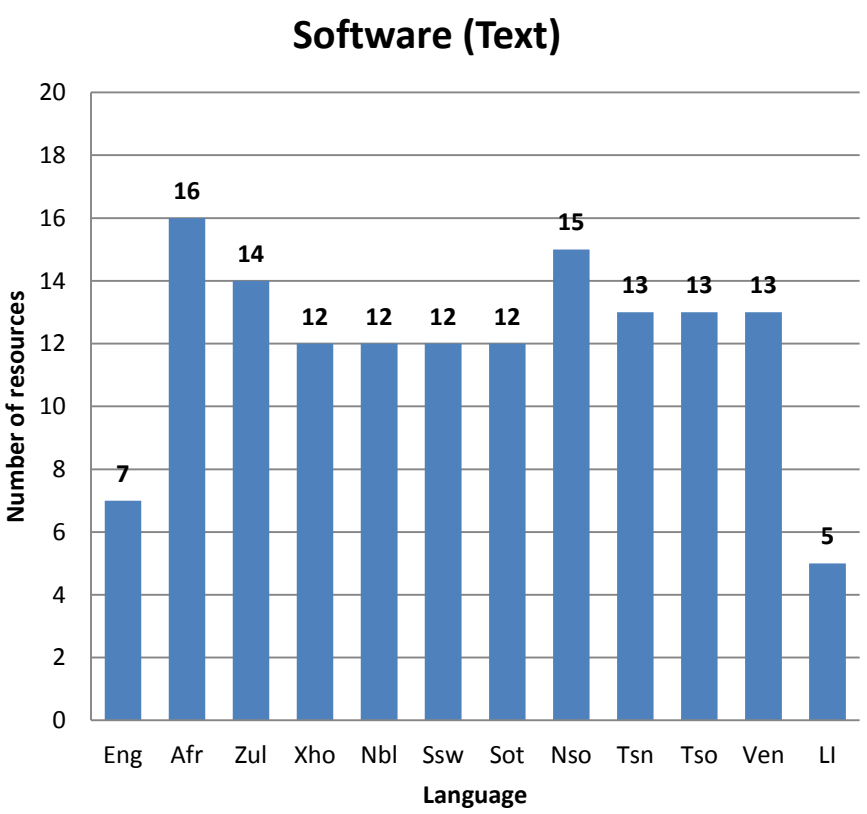


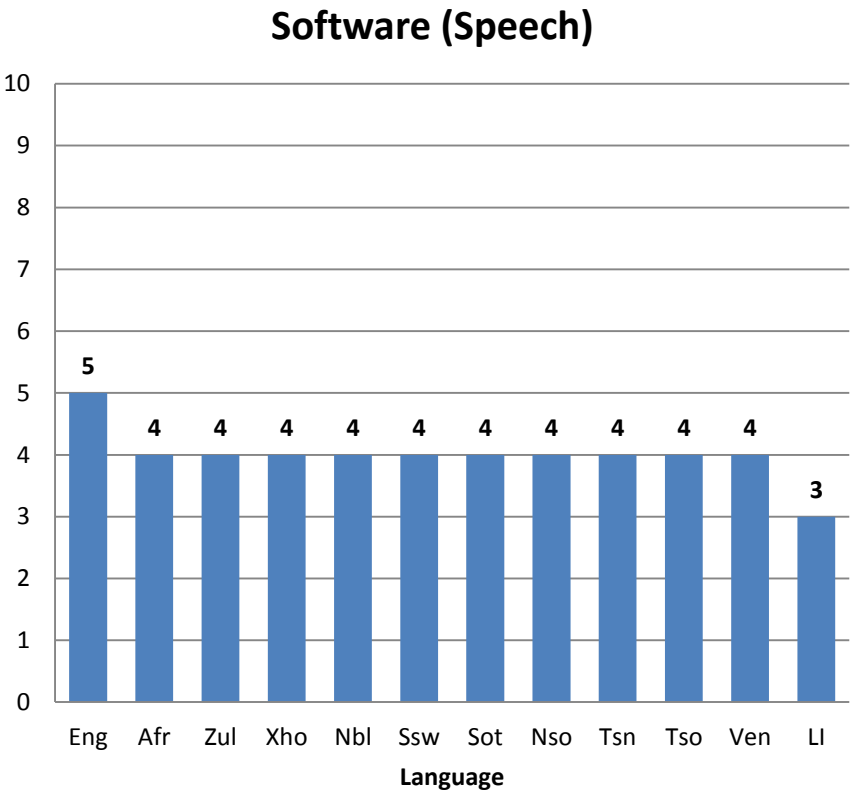
Available data resources 2009 - 2018





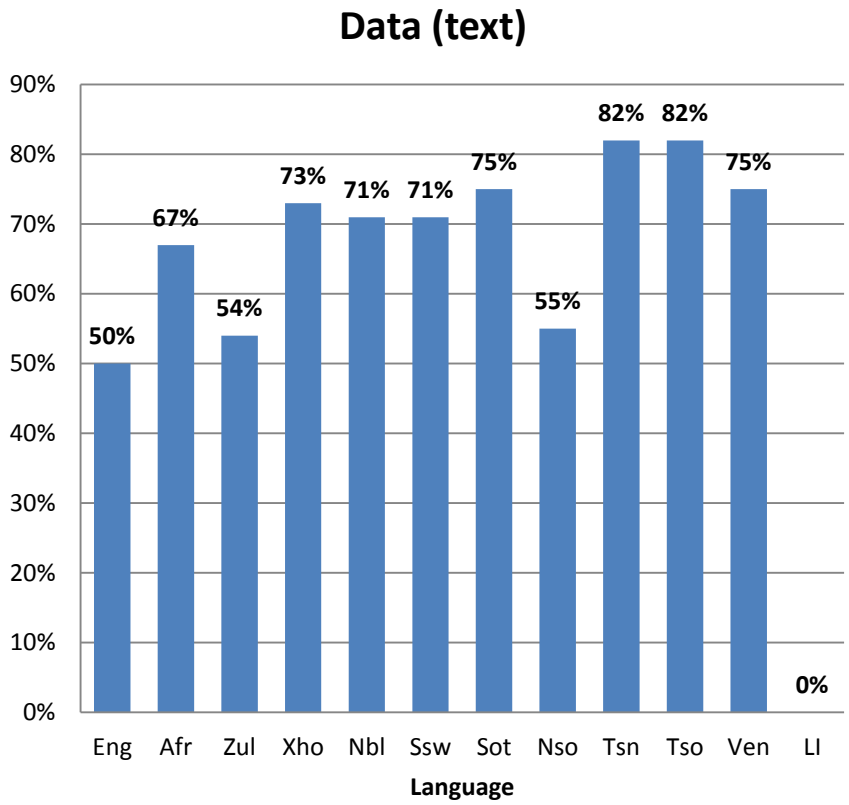
Available software resources 2009 - 2018

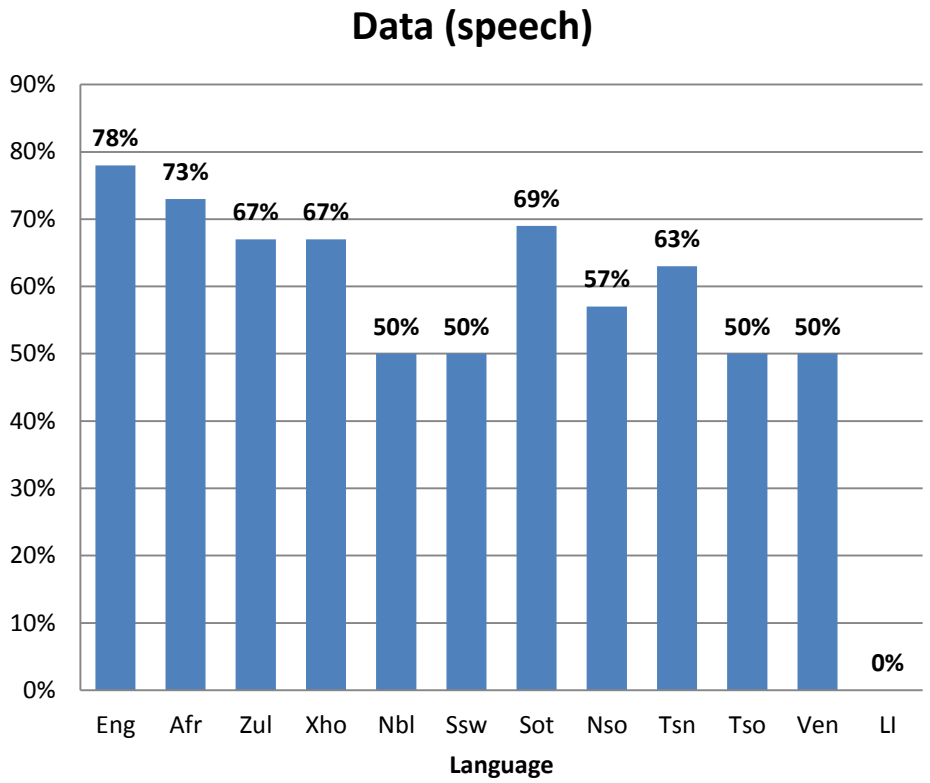




CSIR
our future through science

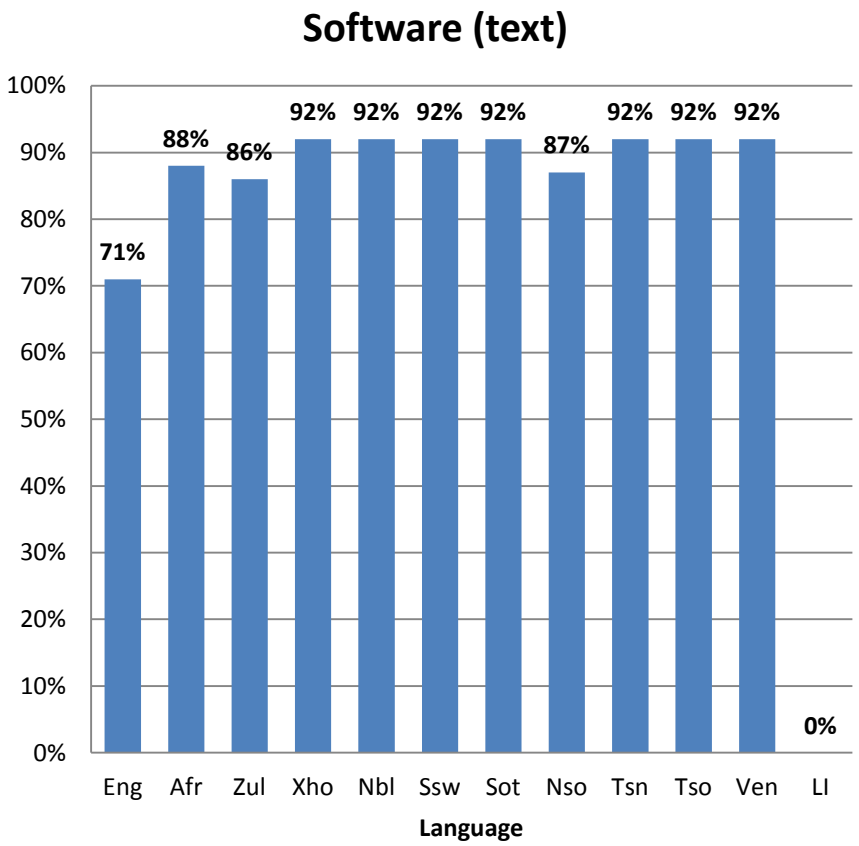
Increase in data resources from 2009 to 2018

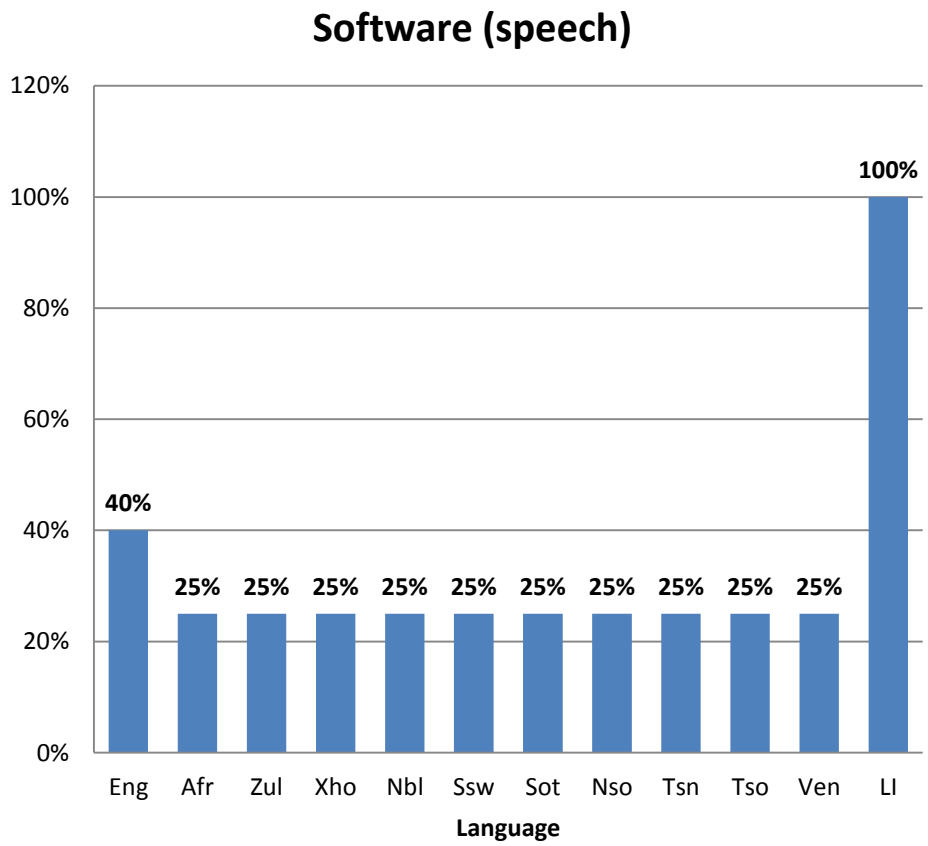




CSIR
our future through science

Increase in software resources from 2009 to 2018





CSIR
our future through science

Gap analysis

Gap analysis

- Many resource types are not available for a number of SA languages
- Text resources better represented than speech resources
- Eng, Afr, Zul, Xho, Nso, Tsn – most resource types
- Resources submitted: classified by level of maturity and level of accessibility
- To compare available resources: similar approach used as by A Sharma Grover in her Masters' thesis[1]
- Maturity Index
- Accessibility Index

[1] Aditi Sharma Grover. 2009. Technology Audit: The State of Human Language Technologies R&D in South Africa. Masters' Thesis. University of Pretoria.

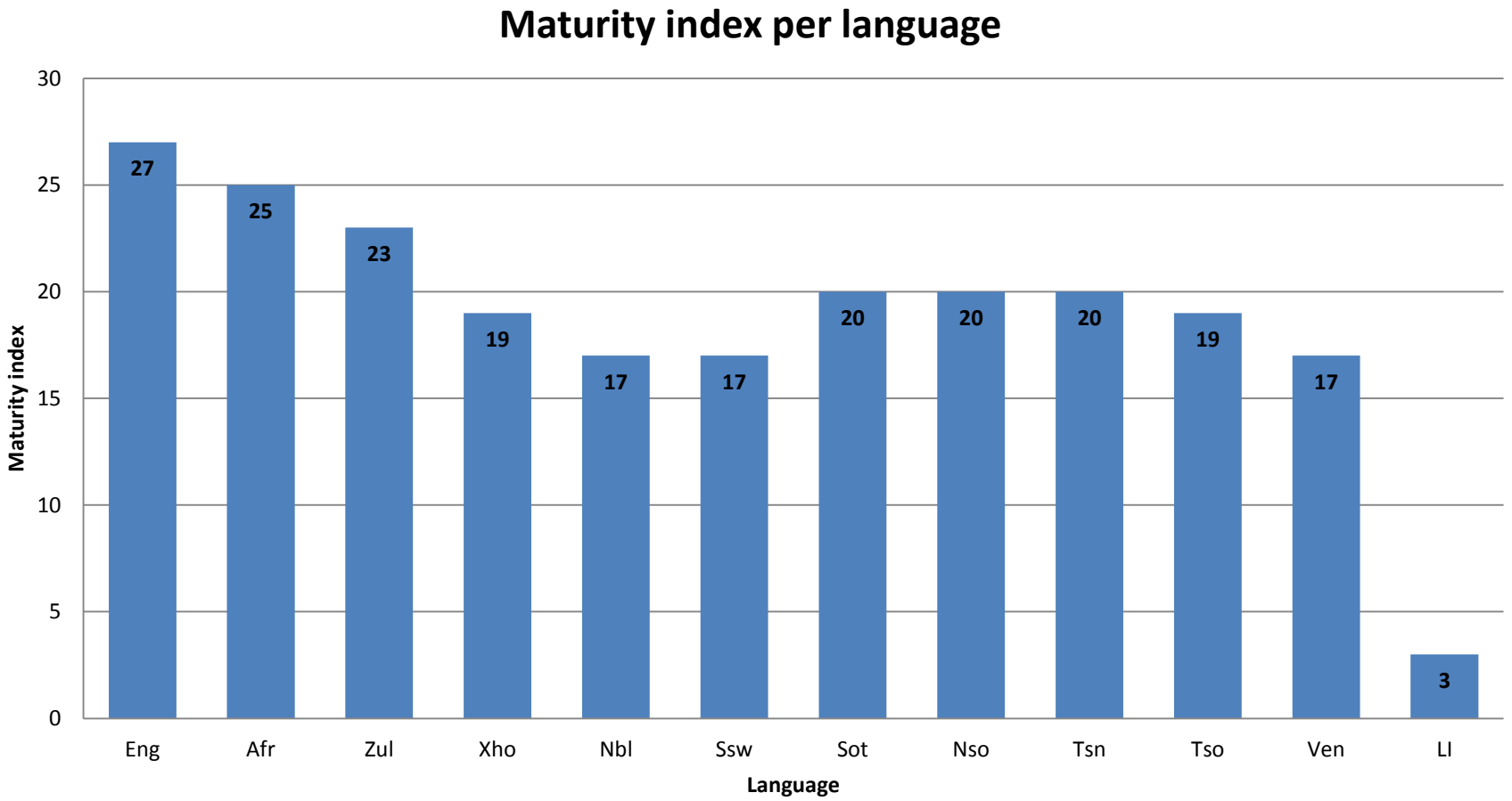
Maturity index

- The **maturity index** provides a measure of the maturity of resources in a language by taking into consideration the maturity weight of each resource

Weights

1 = under development
2 = alpha version
4 = beta version
8 = released version

Maturity index (cont.)



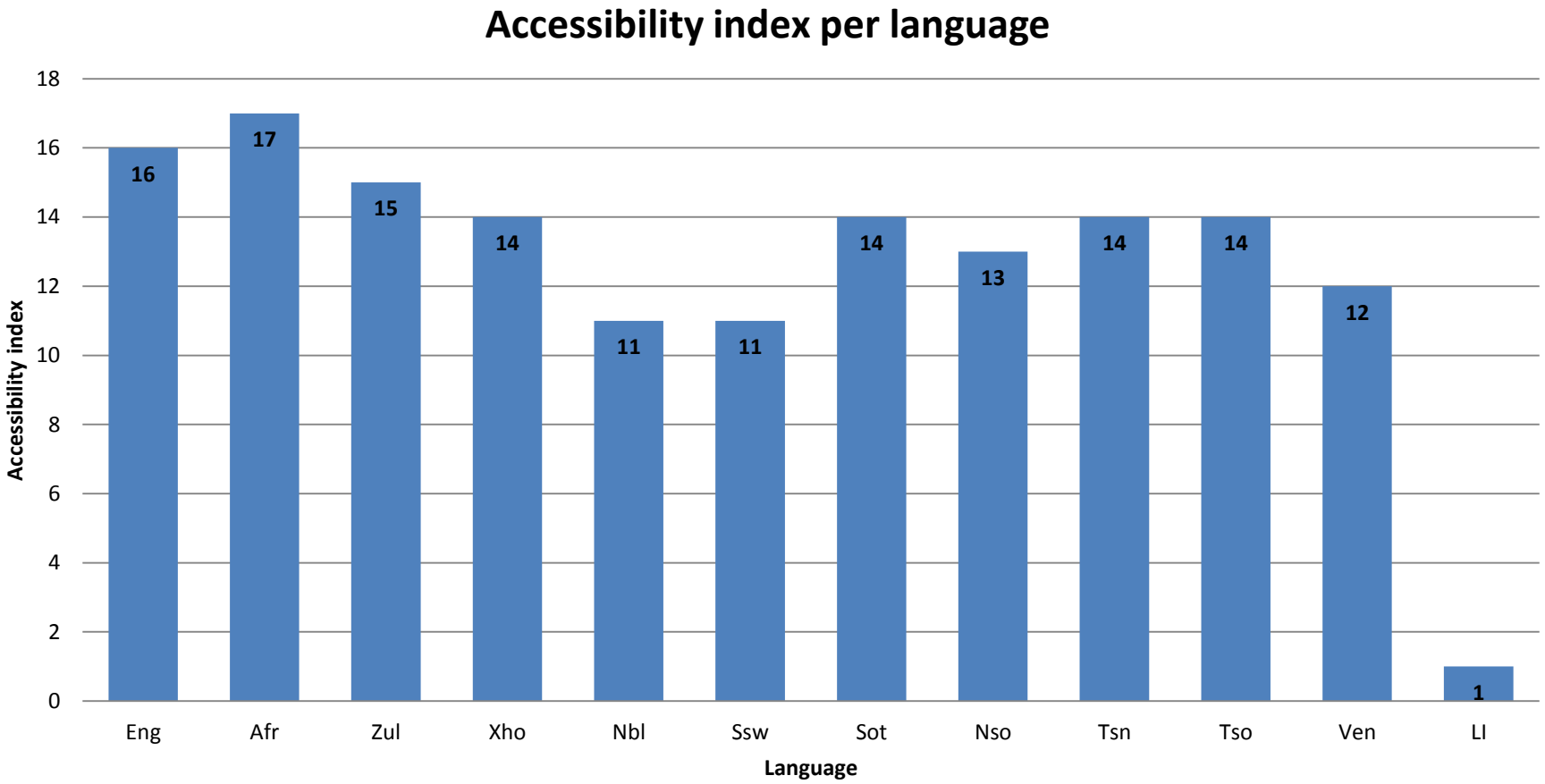
Accessibility index

- The **accessibility index** provides a measure of the accessibility of resources in a language by taking into consideration the accessibility stage of an item

Weights

1 = not available/proprietary/closed
2 = undecided
4 = research
8 = commercially available
12 = open/freely available

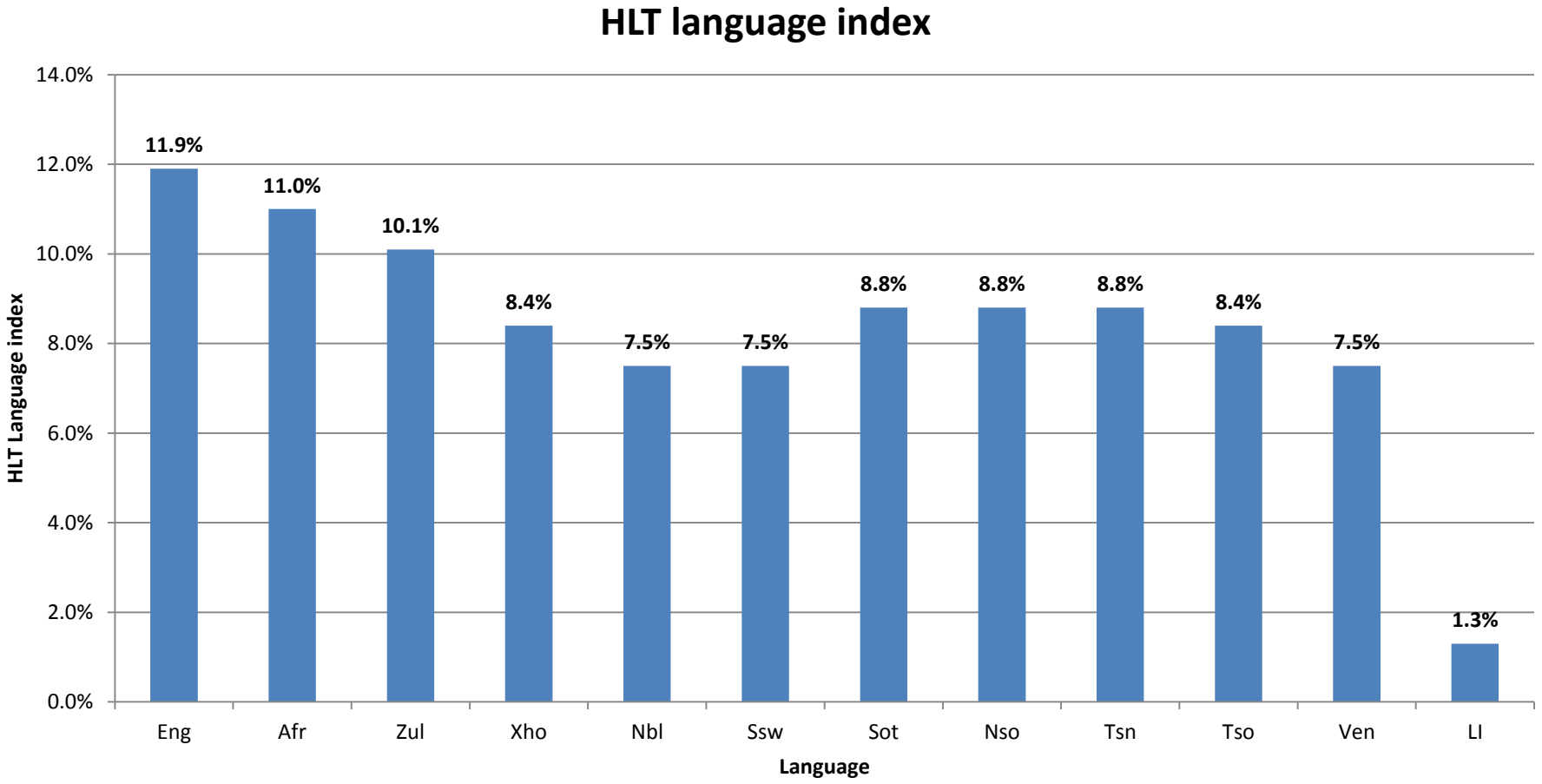
Accessibility index (cont.)



HLT language index

- The **HLT language index** provides an overview of the status of HLT development for all 11 SA languages
- Calculated by adding the maturity index and accessibility index of each language
- Represented as a percentage

HLT language index (cont.)



Data resource types for which no resources are available

Text	Speech		Multimodal
Controlled vocabularies	TTS sentence splitting rule sets	Phoneme sets	Multimodal corpora
Named entity lists	TTS tokenisation rule sets	Pronunciation rule sets	
Thesauri	TTS normalisation rule sets	Intonation rule sets	
Ontologies	TTS language ID rule sets	Phrasing rule sets	
Tagsets	Language grammars	Stress rule sets	
	Syllabification rule sets		

Model resource types for which no resources are available

Models	
TTS sentence splitting models	G2P models
TTS tokenisation models	Intonation models
TTS normalisation models	Phrasing models
TTS LID models	Tone models
Language models	Stress models
Acoustic models	Syllabification models

Software resource types for which no resources are available

Text		Speech	
Sentenciser	Word meaning disambiguator	Language modelling tool	Code-switched speech recognition system
Spelling corrector	Pragmatic analyser	Pronunciation dictionary creation tool	Multilingual speech recognition system
Full-form normaliser	Text generator	G2P tool	Alignment system
Number normaliser	Summariser	Acoustic modelling tool	Automatic phonetic transcription system
Diacritics normaliser	Information retrieval system	Intonation tool	Confidence measures
Anonymiser	Information extractor	Intonation tool	Acoustic language ID
Stemmer	OCR/ICR	Phrasing tool	Acoustic age ID
Morphological synthesizer	Document classifier	Tone tool	Acoustic gender ID
Syllabifier	Authorship identifier	Stress tool	Acoustic dialect ID
Dependency parser	Question answering (QA) system	Syllabification tool	Acoustic emotion ID
Event extractor	Dialogue system (text-based)	Vocoder	Keyword spotting system
Named entity recogniser	Comprehension assistant	TTS sentence splitting tool	Voice activity detection
Terminology extractor	Sentiment analysis/affect/emotion analyser	TTS tokenisation tool	Speaker tracking
Topic modelling	Referent resolver	TTS normalisation tool	Acoustic speaker ID
		TTS LID tool	Speaker verification system
		Large vocabulary speech recognition system	Diarisation
		Command and control system	Non-native speech recognition system

Way forward

Sustainable updates

- Follow same approach as ELRA and CLARIN
- Proposed solution:
  - Update resources when referenced in papers submitted to specific conferences
    - Identify relevant conferences
    - Negotiate with conference organisers
    - Develop interface to link the conference/journal article submissions to SADiLaR online
    - Develop feedback mechanism for successful submission of resources to SADiLaR
- LimeSurvey tool hosted by SADiLaR from now on
- All data will be visible on SADiLaR Catalogue and Index

In summary

- Use updated comparison for decision-making on future resource development needs
- More awareness raising required on availability of HLT resources for R&D
- More awareness raising needed on making other language resources available through SADiLaR

Thank you

INTERNAL AUDIT SERVICES

PO Box 395 Pretoria 0001 South Africa
Tel: +27 12 841 3524
Fax: +27 12 841 2113
Email:pngwato@csir.co.za

TO WHOM IT MAY CONCERN

Internal Audit Services (IAS) has been requested to provide a certificate to confirm the attached statement of funds received and expenditure incurred for the project **K7HLSAD-NWU** for the period 01 July 2017 to 30 June 2018.

IAS confirms that the statement agrees with the balances for the project in the financial records of the CSIR. In addition, IAS has performed the following audit procedures:

- Agreed income received to supporting documentation.
- Verified on a sample basis, the labour hours per the accounting records to the approved timesheets and the labour rates to the approved charge out rates per the accounting records.
- Selected a random sample of running expenses and agreed to supporting documentation.

The results of the above procedures are satisfactory and no exceptions were noted.

In addition IAS performs an annual review of the key financial controls to cover aspects such as authorisation, validity/authenticity of transactions and proper recording. The results of the audit recently completed did not yield any issues of concern.

The statement of funds received and expenditure incurred reviewed is attached and signed for identification by ourselves.

Regards

PONI NGWATO
ACTING GROUP MANAGER
INTERNAL AUDIT SERVICES
17 August 2018

CSIR MERAKA

K7HLSAD-NWU

PERIOD : 01 JULY 2017 TO 30 JUNE 2018

STATEMENT OF FUNDS RECEIVED AND EXPENDITURE INCURRED

	Notes	R
Funds received (Excl Vat)		**1 358 841**
Funds received from South African Centre for Digital Language Resources (Incl Vat)		1 549 079
Less : 14 % VAT	①	190 238
Interest received		-
Total funds available		**1 358 841**
Total HR costs		**1 760 855**
Manpower		1 760 855
Total running expenses		**42 512**
Travel		18 172
Conference Cost		21605
Entertainment		2635
Courier		100
Total expenses		**1 803 367**
Funds outstanding (shortfall)		**(444 526)**

Notes

① Vat has been paid over to SARS.

CSIR
Internal Audit Services
Scientia Campus, Building 3A
Meiring Naude Road, Brummeria, PTA

Sign:
Date: 17/02/2012

"We are here to assist you!"