Human Language Technology Audit 2018: Design Considerations and Methodology

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

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

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

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

***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, contributing to its economy and bridging the digital divide. 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.**

***Index Terms*—human language technology, technology audit, language resources, text resources, speech resources, digital humanities**

I. 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 [1]. The South African Centre for Digital Language Resources (SADiLaR) [2] was recently established as part of the South African Research Infrastructure Roadmap (SARIR) [3]. 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 [1].

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 [4], 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 South Africa. The 2018 HLT Audit aimed to provide updated information on the maturity and availability of HLT resources in the country.

The first-ever HLT audit was conducted by the European Network of Excellence in Human Language Technologies (ELSNET) [5] in 1991, and was based on the idea of a roadmap where information on HLT resources would be collected on a continual basis [6]. 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 [7], took the ELSNET audit as point of departure.

The sections that follow discuss the methodology, the audit design and approach, a description of the audit instrument development, the execution of the audit and finally a brief analysis of the data and the conclusion.

II. 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) [8]

and the Basic Language Resource Kit (BLARK) [9] initiated by ELSNET [6] 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 [7] adapted the BLARK methodologies described above and undertook an audit of HLT resources in South Africa. Taking the Dutch BLARK [10] 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 [11] 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.
- 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).

III. Audit design and approach

A. 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.

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

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 2000 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.

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.

B. 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.

IV. Audit instrument development

A. 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).

B. 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 [12].

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;

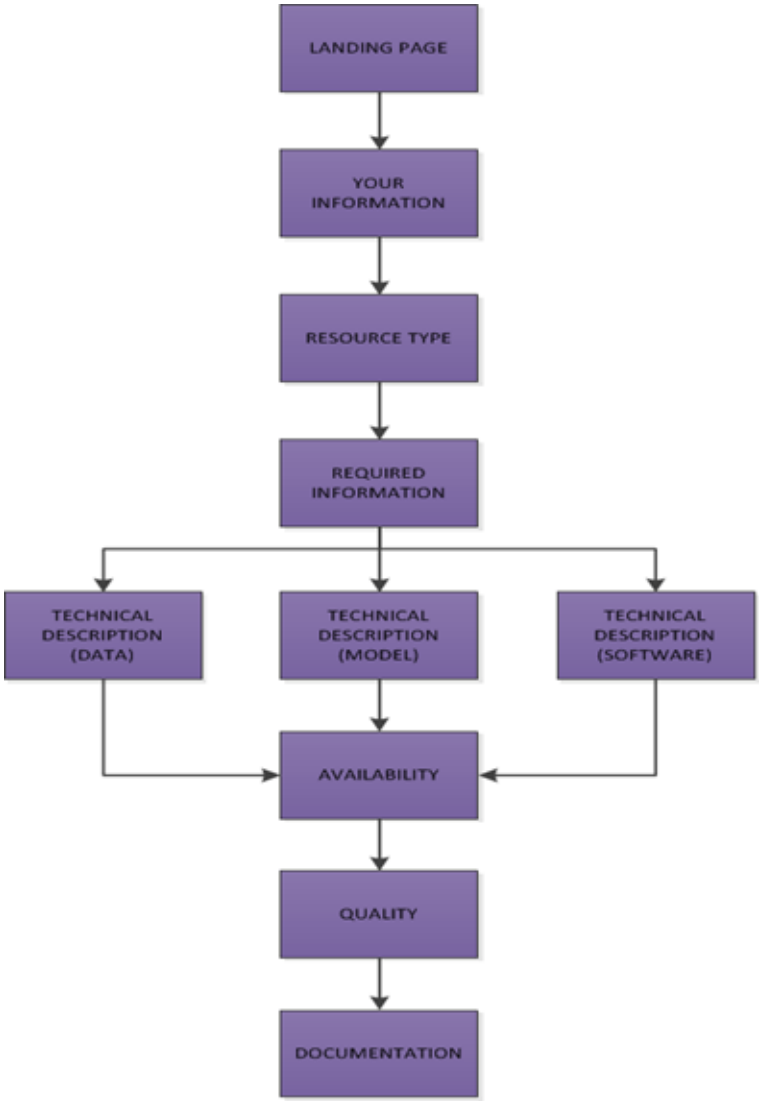

Fig. 1: High-level overview of the 2018 Audit survey

- 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.

C. 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.

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).

D. 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.

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").

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

A separate website [13] 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.

F. 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 (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.

V. Audit execution

A. 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.

B. 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 scheduling 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 are presented a paper by Moors, Wilken, Calteaux and Gumede [14]. In the section below, we provide an overview on the process we follow in analysing the Audit data as well as the actual results of the analysis.

VI. Data analysis

A. Introduction and process

The purpose of analysing the data is to determine what the language resource development trends are and to identify if any gaps exist in the availability of resources in specific South African languages. We obtained the actual 2009 HLT Audit data which was uploaded onto the RMA in 2013 and the data that was uploaded from 2014 until the 2018 HLT Audit from the RMA. We were therefore able to cluster the data into three datasets (2009, 2014 and 2018) to be able to compare resource types. We matched the resource types from the 2009 HLT Audit and 2014 RMA data with the resource types modified in the Audit design, as mentioned in section III. From our matching of resource types we were able to compare the availability of a specific resource in a specific

language. We used a graphical representation (stacked column chart) indicating the number of resources submitted per dataset per language (each data set is a different colour). We then tallied the number of text versus speech resources submitted from 2009 until 2018. This provided us with information on resource types which lack resources and languages which have minimal resources. In addition, by subtracting the final number of resources available in 2018 from the resources available in 2009 only, we were able to provide a graphical representation on resource development per language over a period of 10 years.

An example of how we matched and compared the resource types across datasets is as follows: The multilingual lexicon resource type exists in the 2009 HLT Audit, 2014 RMA data and the 2018 Audit. In the 2009 HLT Audit, for English six multilingual lexicons were submitted, in 2014 one more was submitted and in the 2018 Audit another one was submitted. Therefore, a total of eight multilingual resources exist for English from 2009 until 2018. To measure the increase in this resource, we deducted the original six resources from the final number (ten) and converted the difference into a percentage.

B. Results

Based on the comparisons between datasets and calculating the increase in resources, we were able to determine that there is an increase in resource availability for most South African languages. However, languages such as Xitsonga, Tshivenda, Sesotho, siSwati and isiNdebele still remain under-resourced. We were further able to deduce that more text than speech resources are currently available in South Africa.

In addition to the comparison between resource types, we also determined the maturity and accessibility of the resources in all official languages in South Africa. The maturity calculation is based on whether the resource is under development, in its alpha or beta version or released. In terms of maturity, we deduced that speech corpora is the most mature resource type. In terms of accessibility of resources, we used a calculation based on whether the resource is not available/proprietary or closed, if the availability of the resource is undecided, for research or commercial purposes or is openly or freely available. From these calculations, we deduced that text corpora is the most accessible resource type.

A third calculation was done, where the results of the maturity and accessibility calculations were summed for each resource type, in order to get an overview of HLT development in South Africa. Overall, text corpora is the most developed resource type in South Africa, followed closely by speech corpora. Fig. 2 provides an overview of the development of resources in South Africa.

The results obtained from the analysis of the data is able to provide an overview to academics and other interested parties on which resources still need to be developed and in which South African languages. This information is vital for decision-making on resource development investment.

VII. 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] Department of Science and Technology, "South African Research Infrastructure Roadmap. First Edition," Department of Science and Technology, Pretoria, 2016.

[2] South African Centre for Digital Language Resources (SADiLaR), 2018. [Online]. Available: https://www.sadilar.org/

[3] South African Centre for Digital Language Resources, "Research Infrastructure (RI) Proposal for the South African Research Infrastructure Roadmap (SARIR)," 2017.

[4] U. Bross, "Technology audit as a policy instrument to improve innovations and industrial competitiveness in countries in transition," *Innovation: The European Journal of Social Science Research*, vol. 12, no. 3, pp. 397–412, 1999.

[5] European Network of Excellence in Human Language Technologies (ELSNET), 2018. [Online]. Available: http://www.elsnet.org/index_nof.html/

[6] S. Krauwer, "The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap," in *Proceedings of SPECOM*, Moscow, Russia, 27-29 October 2003.

[7] A. Sharma Grover, G. B. van Hyssteen, and M. W. Pretorius, "An HLT profile of the official South African languages," in *Proceedings of the second workshop on African Language Technology (AFLaT 2010)*, Valetta, Malta, 18 May 2010.

[8] International Conference on Language Resources and Evaluation, "LREC Conferences," 2018. [Online]. Available: http://www.lrec-conf.org/

[9] Evaluations and Language resources Distribution Agency, "BLARK: Basic LAnguage Resource Kit," 2018. [Online]. Available: http://www.blark.org/

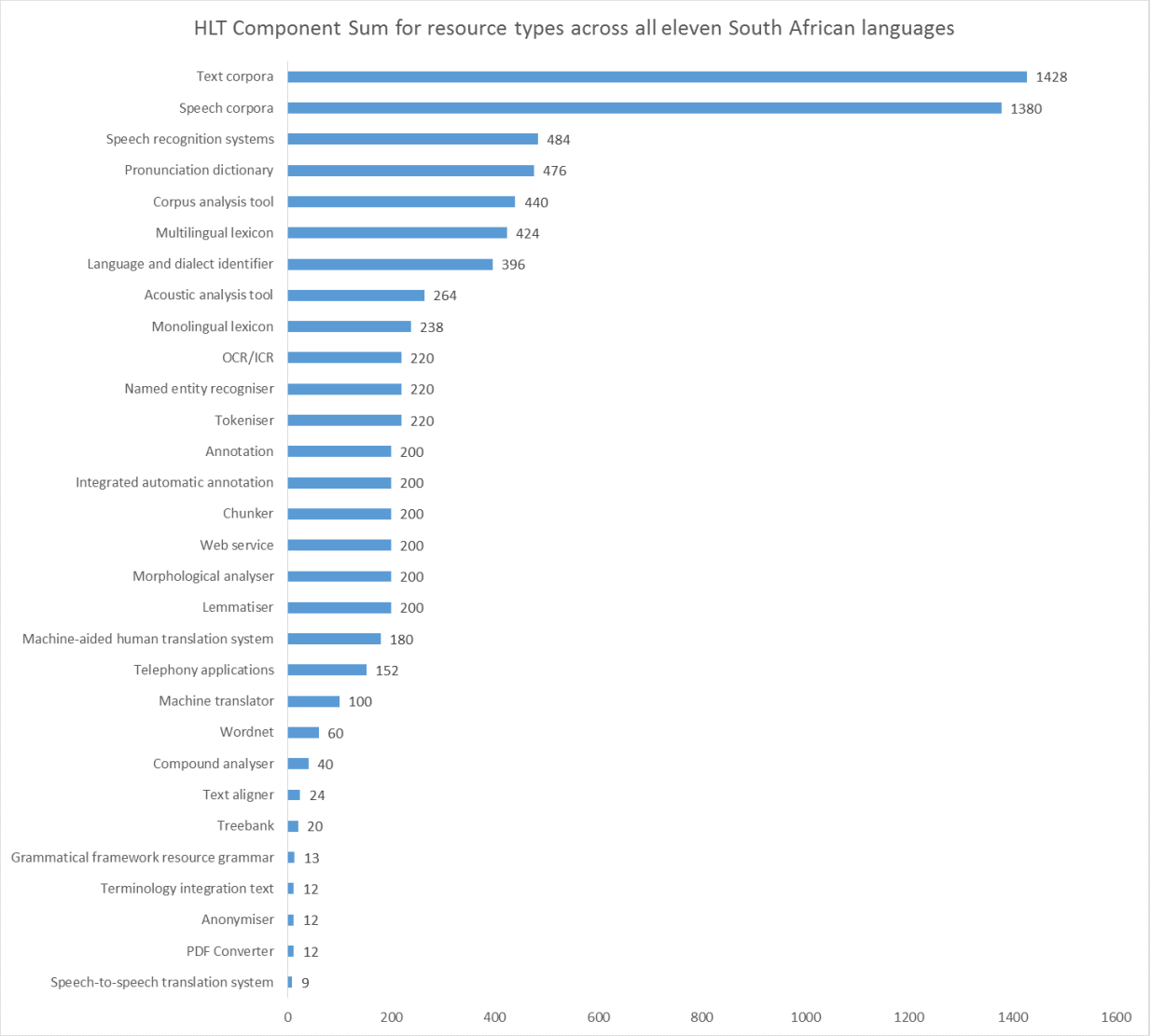

Fig. 2: HLT Component Sums for existing resource types

[10] Helmer Strik, "BLaRK: Basic Language Resource Kit," 2018. [Online]. Available: http://hstrik.ruhosting.nl/blark/

[11] A. Sharma Grover, "Technology Audit: The State of Human Language Technologies R&D in South Africa," Master's thesis, University of Pretoria, South Africa, 2009.

[12] LimeSurvey, "Professional online surveys with limesurvey," 2018. [Online]. Available: https://www.limesurvey.org/

[13] Human Language Technology and Language Resources Audit 2017/8, "List of definitions," 2017. [Online]. Available: https://sites.google.com/view/hlt-audit-defi nitions/home

[14] C. Moors, I. Wilken, K. Calteaux, and T. Gumede, "Human language technology audit 2018: analysing the development trends in resource availability in all South African languages," in *Proceedings of the South African Institute of Computer Scientists and Information Technologists (SAICSIT)*, Port Elizabeth, South Africa, 26-28 September 2018.