The South African Human Language Technology Audit

Aditi Sharma Grover[1,2], Gerhard B. van Huyssteen[3], Marthinus W. Pretorius[2]

Human Language Technology Research Group, Meraka Institute, CSIR[1],
Graduate School of Technology Management, University of Pretoria[2],
Centre for Text Technology (CTexT), North-West University[3]

Abstract – Human language technology (HLT) has been identified as a priority area by the South African government. However, despite efforts by government and the research and development (R&D) community, South Africa has not yet been able to maximise the opportunities of HLT and create a thriving HLT industry. One of the key challenges is the fact that there is insufficient codified knowledge about the current South African HLT components, their attributes and existing relationships. Hence a technology audit was conducted for the South African HLT landscape, to create a systematic and detailed inventory of the status of the HLT components across the eleven official languages. Based on the Basic Language Resource Kit (BLaRK) framework (Krauwer, 1998), we used various data collection methods (such as focus groups, questionnaires and personal consultations with HLT experts) to gather detailed information. The South African HLT landscape is analysed using a number of complementary approaches and based on the interpretations of the results, recommendations are made on how to accelerate HLT development in South Africa, as well as on how to conduct similar audits in other countries and contexts.

Keywords: Technology audit, human language technology, language resources, BLaRK, language audit, language resource infrastructure, resource-scarce languages

1. Introduction

Over the past few years, the South African government has come to realise the important role that human language technology (HLT) could play in bridging the digital divide and challenges of a multilingual society in South Africa[1]. Various research and development (R&D) projects[2] and initiatives have been funded by

[1] Roughly twenty five languages are spoken in South Africa; eleven of these have been declared official languages based on the grounds that their usage includes about 98% of the total population (DAC, 2002). These official languages are Afrikaans (Afr), English (Eng), isiNdebele (Ndb), isiXhosa (Xho), isiZulu (Zul), Sepedi (Sesotho sa Leboa) (Sep), Sesotho (Ses), Setswana (Sts), Siswati (Ssw), Tshivenda (Tsv) and Xitsonga (Xit). The South African government has launched various initiatives and mechanisms to ensure a truly multilingual society (e.g. establishment of the Pan South African Language Board; www.pansalb.org.za).

[2] See for example the contribution by Badenhorst, *et al*. (2011) in this volume.

government, notably through its Department of Arts and Culture (DAC), Department of Science and Technology (DST), and National Research Foundation (NRF).

The National HLT Network (NHN), sponsored by DST, is an informal, online community that aims to strengthen synergies between HLT research practitioners in South Africa. It currently consists mainly of members from the major South African tertiary institutions and science councils who are actively involved in HLT R&D activities. In 2009, NHN undertook a large-scale technology audit for the HLT landscape in South Africa, called the South African HLT Audit (SAHLTA).

Despite a number of efforts by government and the R&D community, South Africa has not yet been able to maximise on the opportunities of HLT, and to create a thriving HLT industry (in comparison to, for example, The Netherlands or the United States of America). One of the key challenges in addressing this problem is the perceived fragmentation of R&D activities in this domain: there is insufficient codified knowledge about the currently available South African HLT resources and applications. These challenges, which motivated the SAHLTA, are very similar to those faced by other resource-scarce languages, especially in the developing world. The lack of language resources (LRs), limited availability of and access to existing LRs, quality of LRs, small-scale and uncoordinated HLT development, and the lack of infrastructure for LR management, are all common issues faced by the development of LRs in resource-scarce languages. Thus, our experiences in the SAHLTA process and representation of the results would be of value to other countries that face similar challenges in HLT development.

Thus, the objective of our study was to codify and present a profile of HLT components in the South African R&D environment. This article aims to provide an overview of the SAHLTA process, as well as to present a selection of the results (cf. Sharma Grover (2009) for an extensive report). In the next section we refer to related work, before discussing in section 3 an overview of the SAHLTA process and instruments we used. In section 4 we present a concise overview of some of our results, in section 5 the discussion and recommendations, and in section 6 the conclusion and directions for future work.

2. Related work

A technology audit is valuable in any technology field, since it allows stakeholders to identify and measure problems and/or performance gaps, while providing information to set the background for the definition of future action plans (Khalil, 2000). The outcome of a technology audit is analogous to a "balance sheet": a snapshot of the organisation's current technological status, as well as an indication of the future directions for maintaining and improving the organisation's status (Martino, 1994; Probert *et al*., 1999). Bross (1999: 397) highlights that technology audits can also be used to identify technological strengths and weaknesses of a sector or industry. As an instrument it supports policy makers in designing appropriate strategies for shaping science and technology (S&T) policies. Thus, within a national landscape, a technology audit can be conducted at a sector level (e.g. HLT, biotechnology, nanotechnology,

etc.), allowing an assessment of the technological competitiveness and innovation potential of the country in that particular sector.

In the international field of HLT, a number of such technology audits have been undertaken. The earliest example is the Dutch HLT survey (Binnenpoorte *et al.*, 2002), which applied Krauwer's (1998) concept of the 'basic language resource kit' (BLaRK) – a set of basic language resources available to conduct preliminary research in HLT – to conduct a field survey for Dutch LRs. Over the past few years, the BLaRK concept and the Dutch survey have inspired HLT surveys for a few other languages, including Arabic (carried out by NEMLAR and MEDAR; Maegaard *et al.*: 2006, 2009), Swedish (Elenius *et al.*, 2008), and Bulgarian (Simov *et al.*, 2004). The BLaRK concept has also been broadened to cater for not just pre-competitive research, but also for advanced HLT development. The Extended Language Resource Kit (ELARK; Mapelli *et al.*, 2003) serves as a definition of HLT components for advanced research or commercial development (i.e. more sophisticated modules, tools and a larger variety of data). On the other extreme, many world languages have very little or no HLTs, and for their purposes an entry-level BLaRK, termed the BLaRKette (Krauwer, 2006), is defined. The BLaRKette caters for very basic research, as well as training and education in academia.

Another well-known HLT survey was the EUROMAP project that benchmarked and measured European countries' progress in the HLT field. Countries were compared on two broad measures, *viz*. an 'Opportunity index' and a 'HLT Benchmark index', where the former represents "the robustness of the opportunity to exploit HLT", and the latter measures the "prospects for and success of HLT research and technology transfer" (Joscelyne & Lockwood, 2003).

Recently, initiatives like FlaReNet[3] and META-NET[4] are also concerned with aspects related to fostering the development of language resources, including auditing of available resources for various languages, setting priorities, managing a sustainable LR environment, etc.

3. SAHLTA process

Based on the above, the BLaRK concept was chosen to guide the audit, since it provides a well-defined structure to capture the different HLT components. The subsequent sections describe the process we followed in conducting the South African HLT technology audit, which could serve as basis for other subsequent audits in other countries or for other languages.

3.1 Terminology, inventory criteria framework and cursory inventory

We commenced by developing an HLT terminology list to establish the nomenclature, taxonomy and descriptions for the HLT components to be used in the audit, according to the following BLaRK classification:

[3] www.flarenet.eu

[4] www.meta-net.eu

- Data (either speech or text; e.g. corpora, lexica, grammars, etc.);
- Modules (software units or processes required to create HLT applications and products; e.g. part-of-speech taggers, language models, etc.); and
- Applications (used by end-users with a dedicated user interface; e.g. proofing/authoring tools, dictation systems, etc.).

Whilst the Dutch and Arabic efforts provided a useful point of departure, some adaptation was required for the South African context. For example, we note that the environments of these languages are significantly different from that of South African languages due to various reasons:

- South Africa has different market needs, where the target users are not only multilingual, but also of diverse socio-economic and cultural backgrounds; and
- Significant technological advancement has been made in the field of HLT for Dutch and Arabic; their applications category thus encompasses more advanced applications that may not currently be feasible in the South African context, both from a technical and market-related viewpoint.

Departing from these application categories, we refined them (for example, categories for audio search and reference works were added), taking into account the above-mentioned issues. (For further details on the ontology of these HLT components please refer to the supplementary online resources available on the WWW.[5])

Our second step involved the establishment of an HLT inventory criteria framework that defined the criteria or dimensions on which the HLT components would be audited and documented. It is not sufficient to know whether a component exists or not, but rather a detailed assessment (e.g. quality, availability, adaptability, etc.) is required on the component to determine its usability.

Concurrently, we built a cursory inventory to identify existing HLT components (using the data, modules, and applications categories defined above) for each of the eleven South African official languages, across the major HLT role players in the country. This was based on research groups' websites, published project reports, academic publications and consultations with a few local HLT experts.

3.2 Audit workshop

The above-mentioned outputs (section 3.1) served as inputs for an audit workshop with eight South African HLT experts (representing speech and text technologies). The workshop was aimed on the one hand at establishing and verifying the audit process and instruments (including terminology and inventory criteria framework), as well as identifying priorities for applications and related LRs (data and modules) for South Africa.

[5] tinyurl.com/6lb8z6x

3.2.1 Inventory criteria framework

Each audit dimension of the inventory criteria framework was described in terms of either the possible states that it could be in (e.g. maturity could be 'under development', 'alpha version', 'beta version' or 'released'), or in terms of subjective descriptions on the details of the item (e.g. the documentation dimension would be described in terms of the availability of publications, reports, etc.). The final inventory criteria framework is summarised below (see Sharma Grover *et al.* (2010a) for more details):

- Technical description (e.g. description, size, stratum, programming language, I/O specifications, operating environment, file and encoding format, etc.);
- Availability (i.e. accessibility, maturity, distribution, licensing, and cost);
- Documentation (e.g. details of publications, reports, websites, user manuals, etc.);
- Quality (i.e. verification and/or proof of quality, and compatibility with standards); and
- Reusability/adaptability (e.g. compatibility with other data formats, standard tools/platforms, relevance to other LRs and applications, open source, etc.). (This criterion is the only one that was subsequently not included as a compulsory field in the audit questionnaire – see section 3.3.)

3.2.2 Prioritisation of HLT components

During the workshop, a first draft of priorities for applications and associated LRs was also developed. The most relevant factors that should be taken into consideration were identified, and include:

- International trends: best practices and current developments in the field of HLT in international R&D efforts and industries;
- Local market needs: factors such as socio-economic constraints, culture, multilingualism, and the readiness of local markets to introduce HLT products and services; and
- Feasibility: technical feasibility and practical implications such as cost and time-lines for R&D.

A basic descending 3-point priority scale was defined for assigning priorities to the HLT components as follows:

- 1 = Requires definite attention for furthering development;
- 2 = Attention should be given based on specific needs and trends; or
- 3 = Nice-to-have for further research or enhancement.

3.3 Audit questionnaire

The inventory criteria framework formed the backbone for the audit questionnaire, which was aimed at gathering data on available HLT components in South Africa. The audit questionnaire consists of four major sections: one for each HLT component category (i.e. 'Data', 'Module', 'Application'), as well as a section, 'Tools/Platforms', which was added to accommodate technologies that are

typically language-independent and aid the development of HLTs (e.g. annotation tools, or corpus searching tools, or platforms that provide easy interfaces to other tools). Each section included the most relevant criteria for that particular category.

The audit questionnaire was sent to all major HLT role-players in South Africa. Based on their historical core HLT competence in R&D, organisations were classified as primary (e.g. universities) or secondary (e.g. national lexicographic units) participants. The response rate was 80% for primary participants, and 33% for secondary participants (as expected, since these participants are in general not very active in the field of HLT in South Africa). All primary participants were paid a minimal honorarium to compensate for the considerable effort that was required from them.

4. Results and discussion

4.1 HLT priorities

We present first the prioritisation of the HLT applications for South Africa, as determined during the audit workshop (section 3.2.2). It can be observed from the priority list in Table 1 that many advanced HLT applications were given only priority 2 or 3, despite the fact that much attention is being paid to these internationally (e.g. question answering or dictation); however, local market needs and feasibility factors played a strong role in determining this priority list, as indicated in section 3.2.2 above. As the HLT industry in South Africa develops further, this priority list will need to be updated.

Subsequently, the prioritisation of LRs (data and modules) was also done during the workshop, using the same priority scale as defined above (see Sharma Grover, 2009, or supplementary online resources[6]).

Table 1 Prioritisation of HLT applications for South Africa

Priority	Applications	
	Text	**Speech**
1	Proofing/authoring tools	Accessibility
	Information retrieval	Telephony applications (IVR/SDS; Transactional/Information)
	Information extraction	Computer assisted language learning (CALL)
	Human-aided machine translation	Audio Search
	Machine-aided human translation	Audio Management
2	Optical character recognition (OCR)/ Intelligent character recognition (ICR)	Access control
	Multilingual comprehension assistants	Embedded speech recognition
	Computer assisted language learning (CALL)	Speaking devices
	Authorship Identification	Computer-assisted training
3	Text generation	Transcription and dictation

[6] tinyurl.com/6lb8z6x

	Document classification	Multimodal information access
	Automatic summarisation	Command & Control
	Question Answering (QA)	Announcement systems
	Dialogue systems (text-based)	Audio books
	Reference works	Speech-to-speech translation

4.2 HLT components

In order to compare the state of HLT components for all eleven languages, we created the 'HLT Language Index', an impressionistic index that relatively ranks languages based on the total quantity of HLT activity within a language, together with the stage of maturity and accessibility of their LRs and applications. This index was created by summation of the 'Maturity Index' and the 'Accessibility Index'. Finally, we also did a gap analysis to determine the current status quo vs. the identified priorities.

4.2.1 Maturity Index

The Maturity Index provides a measure of the maturity of HLT components in a language by taking into account the maturity stage of an item against the relative importance or contribution of each maturity stage. The 'maturity sum' per item grouping (e.g. 'pronunciation resources') for each language is calculated as:

$$\text{Maturity Sum} = (UD * 1) + (AV * 2) + (BV * 4) + (RV * 8) \quad (1)$$

where UD is the number of components in the 'under development' phase, AV is the number of 'alpha version' components, BV is the number of 'beta version' components, and RV is the number of 'released version' components. The weights for the different versions are relative weights, in order to give greater importance to the final, released versions of components. The weight assigned to a maturity stage is double that of the preceding maturity stage, i.e. an 'alpha version' item counts twice as much as an 'under development' item, or a 'beta version' item counts twice as much as an 'alpha version' item, etc. (For an example of these calculations, see Sharma Grover *et al.*, 2010b).

Maturity sums were calculated across component groupings for all data, modules and applications per language. To obtain a comparative approximation of the maturity across the different languages, the 'Maturity Index' (per language) was calculated by normalising the total of all the maturity sums (i.e. all item groupings across data, modules and applications for a language) by the sum of weights for the maturity stages (1+2+4+8 =15) , as shown below:

$$\text{Maturity Index} = \frac{\sum \text{Maturity sums (data, modules, applications)}}{\sum \text{Weights of maturity stages}} \quad (2)$$

Figure 1 illustrates this 'Maturity Index' that was calculated for each language; note this index is a relative index since it is based on the total number of HLT

components that exist in a language[7], and indicates how mature the language is in terms of the development stages of its HLT components.

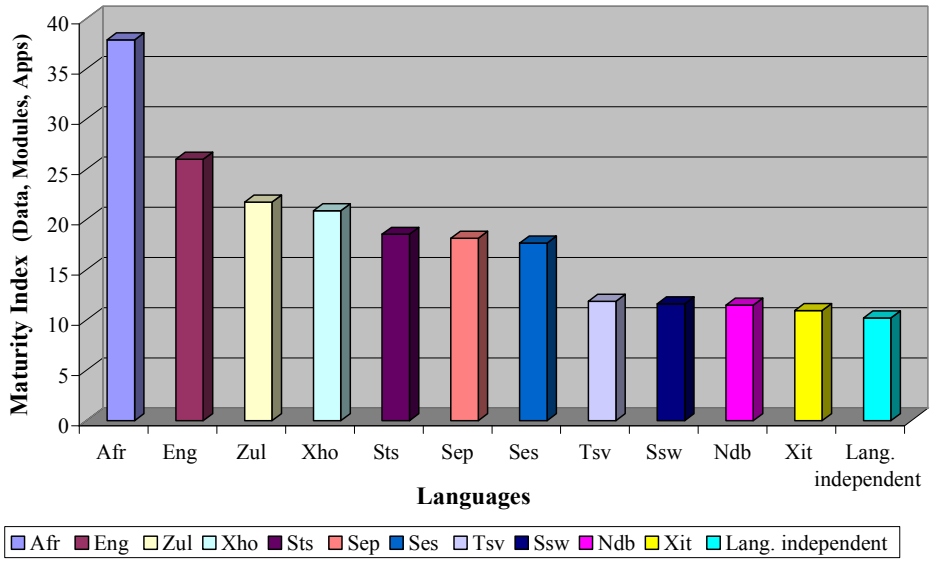

Fig. 1 Maturity index per language[8]

4.2.2 Accessibility Index

The Accessibility Index provides a measure of the accessibility of HLT components in a language by considering the accessibility stage of an item as well as the relative importance of each accessibility stage. The 'accessibility sum' is calculated per HLT component grouping for each language as follows;

$$\text{Accessibility Sum} = (UN * 1) + (NA * 2) + (RE * 4) + (CO * 8) + (CRE * 12) \quad (3)$$

where UN is the number of components that are classified as 'Unspecified' in terms of the accessibility stage, NA is the number of components that are listed as 'Not available (proprietary or contract R&D[9])', RE is the number of components 'available for research and education (R&E)', CO is the number of components 'available for commercial purposes' and CRE is the number of components 'available for commercial purposes and R&E' in terms of accessibility. Similar to

---

[7] In our work 'English (Eng)' refers to 'South African English (SAE)' which has significant linguistic differences (e.g. pronunciation of words) from other accents of English such as 'British' or 'American' English.

[8] Lang. independent/L.I. = language independent.

[9] Not available (NA) items refer to proprietary resources or contract R&D resources which may not be fully available (e.g. resources from the defence environment). In a resource-scarce environment we found it significant that a resource exists even if NA, so that the HLT community is aware of it. Since UN items have more uncertainty with regard to their accessibility status (which may take significant time to get resolved), NA items are given a higher score than UN.

the maturity sum, relative weights were assigned to the different accessibility stages, with higher weights for stages that make a component more accessible (e.g. available for commercial). Since the 'available for commercial purposes and R&E' stage is a combination of the previous 'commercial only' and 'R&E only' categories, it was assigned only 1.5 times the weight of the preceding score, i.e. 1.5*8=12).

Similar to the Maturity Index, the Accessibility Index was determined by calculating accessibility sums across component groupings for all data, modules and applications per language (for more details, see Sharma Grover, 2009).

4.2.3 HLT Component Indexes

The HLT Component Index provides an alternative perspective on the quantity of activity taking place within each of the data, modules, and applications categories on an HLT component grouping level (e.g. pronunciation resources), and is calculated as follows:

HLT Component Index = Maturity Index (per item grouping) + Accessibility Index (per item grouping) [10]

(4)

The HLT Component Indexes for all languages are plotted in a grid using a bubble plot. The value of the HLT Component Index for a particular component grouping determines the size of the bubble; i.e. the higher the index, the larger the bubble. However, it is important to note that the size of the bubbles plotted within a plot is proportional to the highest value of the HLT Component Index within the entire plot. Thus, this index provides a relative comparison of the HLT activity within the various groupings of data, modules or applications within a single plot, as opposed to an absolute comparison of languages.

For the sake of brevity, we present here only the plot for the HLT Component Indexes for applications (see Sharma Grover *et al.* (2010b), or the supplementary online resources[11] for plots for data and modules). Figure 2 illustrates the HLT Component Indexes for text (in blue) and speech (in red) applications. The greatest quantity of activity (more mature and accessible) is in the reference works (text) for Afrikaans and English, as well as in telephony-based services in English. There is some medium-scale activity in the document production, translation, CALL, and accessibility areas for the text and speech domains respectively. A number of speech-based applications have very few activities in English, Afrikaans and IsiXhosa, since these applications are in their early development phases. The remainder of the South African languages do not have any activity in terms of speech-based applications.

[10] The Maturity Index and Accessibility Index used here are calculated for each grouping of HLT components within data, modules and applications.

[11] tinyurl.com/6lb8z6x

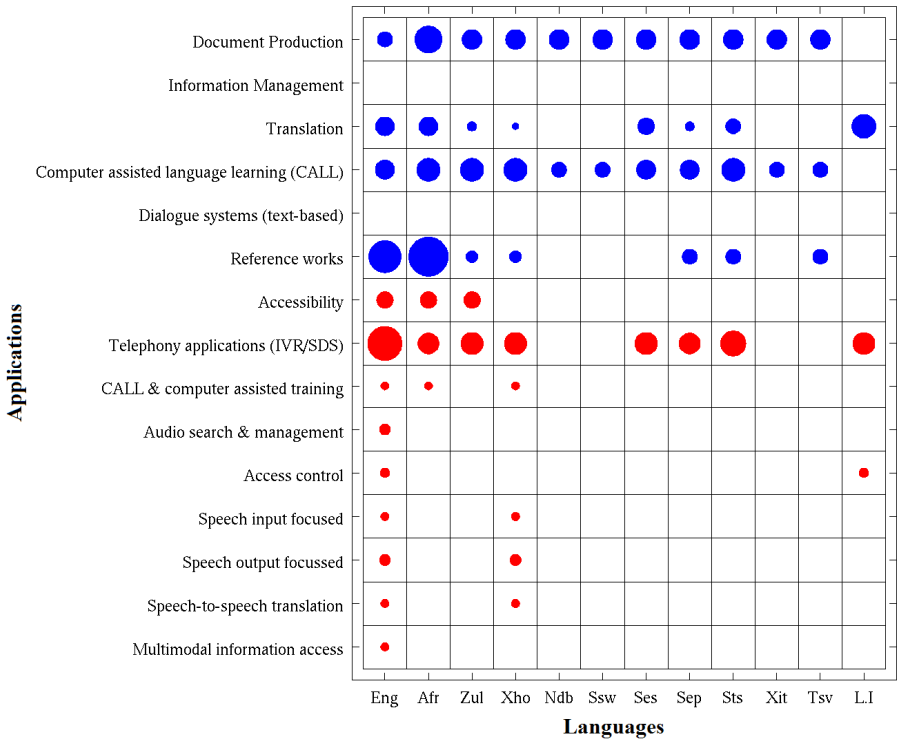

Fig. 2 HLT Component Index for applications

Based on these plots, one could also perform a gap analysis, which could serve to identify gaps between the current status and the prioritised components (section 4.1). This information could be highly informative for future road-mapping exercises, as well as to immediately identify areas or languages that should receive particular attention. (For examples of such gap analyses, refer to the supplementary online resources.[12])

4.2.4 HLT Language Index

The 'HLT Language Index' provides a comparison on the overall status of HLT development for the eleven South African languages, and was calculated by summation of the Maturity Index and the Accessibility Index for each language (across all HLT components):

$$\text{HLT Language Index} = \text{Maturity Index} + \text{Accessibility Index}^{13} \text{ (per language)} \quad (5)$$

This index allows languages to be compared against each other based on the total quantity of HLT activity within a language whilst also taking into account the stage of maturity and accessibility of the outputs of the HLT activity. In interpreting the results of the 'HLT Language Index' it is important to note that statistical data analysis is unsuitable, since there are a relatively small number of items available per language. Thus rather an impressionistic representation of the

[12] tinyurl.com/6lb8z6x

[13] The Maturity Index and the Accessibility Index here is on a per language basis, taken across all data, modules, applications as discussed in sections 4.2.1 and 4.2.2 respectively.

HLT landscape is being portrayed through this index. Also, the limited number of items available per sub-category makes the data analysis subject to significant changes if additional items are identified for a language.

Figure 3 depicts the HLT Language Index for the South African languages. It shows that Afrikaans is by far the most developed language in South Africa with regard to HLT LRs and applications, followed by the local vernacular of South African English (with a significant difference between the two). This picture is slightly skewed by the fact that very little work on South African English is required within the text domain, which means that South African English will almost always only be measured in terms of activity related to speech technologies.

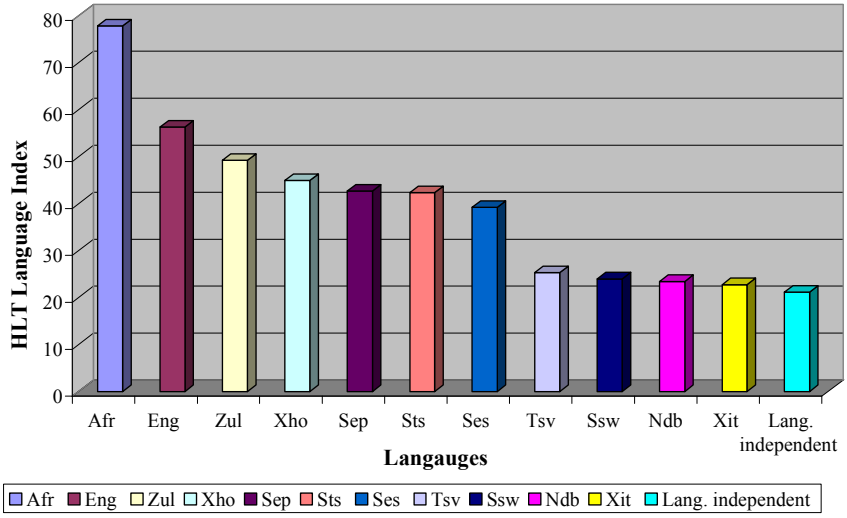

Fig. 3 The South African 'HLT Language Index'

Based on the above, our overall impression of the South African HLT landscape is that very few basic LRs and applications exist across all eleven languages; unsurprisingly, it is especially the four smallest languages that lag far behind in terms of HLT development (see section 5 for an explanation). It is also clear that there are a great many areas that lie fallow in terms of the variety, number and maturity of items, especially compared to other world languages.

5. Discussion and recommendations

Although this audit was conducted in South Africa, lessons learnt could be applied to other similar contexts, especially to developing countries and resource-scarce environments. The discussions and recommendations that follow focus on the South African audit and context, but could be extrapolated easily to other settings.

SAHLTA's findings reveal that whilst there is a considerable level of HLT activity in South Africa, there are significant differences in the amount of activity across the eleven languages, and in general the language resources and

applications currently available are of a very basic nature. In reflecting on these findings, several factors need to be considered holistically in order to understand the current HLT landscape in South Africa.

- *HLT expert knowledge:* HLT is a highly specialised field in science, engineering and technology (SET), which requires technical skills ranging from linguistics and computer science to mathematics and computer engineering. Linguistics plays a crucial role, since work on HLT in a specific language requires basic linguistic knowledge of that language. In general, the availability of SET experts is limited, and even more so the case for HLT. This, coupled with historical imbalances, leads to more linguistic expertise and foundational work being available for Afrikaans and South African English.

- *Availability of data resources:* The cornerstone for building and enhancing HLT modules and applications is the availability of data collections, such as text sources (e.g. newspapers, books, periodicals, and documents) and speech sources (e.g. audio recordings) for a language. The availability of such sources is far greater for languages such as Afrikaans and South African English, as opposed to the African languages (and even more so for the smaller languages).
  Another facet of this factor is the geographic location of the available linguistic and HLT experts. Since most of the HLT role-players are geographically located in the northern/north-western (i.e. Gauteng and North-West Province), and south-western (i.e. Western Cape) regions of the country, it is often more practical and feasible to work on the languages more commonly spoken in these areas (i.e. Afrikaans, South African English, isiZulu, isiXhosa, Setswana and Sepedi), since data can be collected much more easily, and native speakers of these languages can be recruited more easily.

- *Market needs of a language:* The market needs of HLT in a particular language can be viewed as a combination of supply and demand factors, and the functional status of the language in the public domain. By supply and demand, one mostly refers to the size and nature of the target population for the language (e.g. number of people who use the language, demographic and socio-economic profile of users, needs of end-users, etc.), whilst the functional status refers to the usage of a language in various public domains (e.g. by government, in business sectors, in education, in the media, and for various cultural activities). In South Africa, English (and to a somewhat lesser extent Afrikaans) is by and large the lingua franca in the business domain, while the African languages are less widely used in such commercial environments. This significantly lowers the economic feasibility of HLT endeavours for these languages, and the South African government will therefore have to play a vital role in order to enable all languages in the HLT domain.

- *Relatedness to other world languages:* Since HLT development for a new language involves considerable investment in resources, one could employ cross-language information and bootstrapping approaches (Davel and Barnard, 2003) to initiate HLT development of new languages, based on other linguistically similar languages. Linguistically, Afrikaans is very similar to

Dutch, and thus has benefitted from and leveraged on the HLT developments for Dutch. Similarly, South African English has also taken advantage of the international HLT activity for American and British English in general.

In contrast to the above, African languages are linguistically not similar to any of the European languages where there has been HLT development, and thus cannot leverage on an existing pool of HLT knowledge. This fact, coupled with the complexity of African languages (e.g. tone, clicks, morphological complexity, etc.), leads to these languages having to commence their HLT efforts from the bottom of the development lifecycle, and start by investing in basic LR and linguistic knowledge generation.

South African HLT role-players and stakeholders are faced with the challenging task of balancing political needs (i.e. to pay attention to all official languages equally) with economic viability (i.e. to create a thriving HLT industry, where there is return on investment on the HLT outputs produced for a certain language). A number of recommendations can be made for accelerating HLT development in South Africa:

- *Resource development and distribution:* From the gap analysis, it is easily discernible that basic core LRs need to be built for all languages. However, it is also important to note that whilst building basic LRs should be prioritised, the South African HLT community needs to start building experience in developing more advanced LRs (priority 2 and 3) for future fast-tracking of HLT applications. In addition, market needs and trends should be a prime consideration in the development of such LRs.

  It was also observed in the results that licensing agreements were often not defined for numerous LRs (often for government funded research projects). Thus, although some of these LRs may be declared as accessible (available for commercial and R&E usage) the ambiguity around the licensing leads to delays and obstacles in using them. Therefore, in order to encourage innovation, such government funded LRs should preferably be made freely available in the open source domain. Alternatively, where LRs are subject to intellectual property rights for commercial use, they should be available at a price that does not prohibit their usage.

- *Funding:* The principal sponsor of HLT development in the country thus far has been the South African government, with some commercial work funded by international companies. In contrast, the South African HLT industry only comprises a handful of companies that focus on a few languages, since the initial investment required does not cover the potential income from the projected market needs for most languages. Thus, in these formative years the government needs to continue to invest in HLT efforts to build a strong foundation of HLT outputs, which would enable the creation of a thriving HLT industry in South Africa, and ensure that HLT in all eleven languages continues to progress.

- *Industry stimulation programmes:* Besides funding, government needs to ensure that there are more initiatives across various government departments to encourage the existing industry's participation in national HLT activities,

and to enable the establishment of new HLT-based start-up companies. In addition, industry participation in resource-scarce languages may need to be motivated proactively by the South African government.

- *Collaborations:* Closely related to the above-mentioned stimulation programmes is the need for greater collaborations within the local HLT community and the larger international community. One of the challenges that lie ahead is to harness the knowledge and skills developed in local pockets of excellence into a collaborative South African endeavour. Thus, a more coordinated effort across the HLT community is required (like the Dutch Language Union's STEVIN[14] programme; D'Halleweyn, 2006), in order to ensure that there is a well-mapped trajectory for LR creation and HLT market development in South Africa.

- *Human capital development (HCD):* The shortage of linguistic and HLT expertise (and general scientific capacity) is a prohibitive factor in the progress of HLT; thus, HCD efforts within the field of HLT should be accelerated. Currently, there is only one South African undergraduate HLT degree programme of its kind (Pilon, *et al.,* 2005), while most other training courses are at the postgraduate level. The general sentiment amongst HLT academia is that more awareness of the field needs to be created among undergraduate and secondary school students, and greater investment needs to be made in generating HLT practitioners who can feed into the emerging HLT industry's pipeline.

  A final noteworthy point is that collaboration across disciplines (e.g. linguistics, engineering, and computer science) should be encouraged, since HLT involves crossing silos of academic disciplines.

- *Cultivation of niche expertise:* It was observed from the audit results that a number of language independent methods have been adopted in creating HLT components for South African languages. This approach (depending on the LR in question) has the potential to fast-track the development of HLTs across South Africa's languages, and already a number of achievements have been made in South Africa in producing HLT items with limited LRs. The local HLT community should continue to enhance this capability of producing portable language independent HLTs, to create a niche expertise area for itself in producing HLTs for resource-scarce languages. This capability could result in knowledge transfer to other countries with resource-scarce languages (e.g. other African languages and the smaller European languages). Thus, the South African HLT R&D community should focus on nurturing niche expertise relevant to a larger, international audience.

6. Conclusion

Besides the audit findings, we also learnt a number of lessons on how an HLT audit should be conducted, especially in a developing, multilingual context. For example, this audit confirmed once again that measuring quality and other

[14] taalunieversum.org/taal/technologie/stevin/

subjective dimensions is a time-consuming, costly and effortful process, requiring dedicated human resources. These constraints might dictate a more abbreviated and superfluous process; however, we are of the opinion that an audit should gather as much detailed information as possible on several audit dimensions. For a participant to merely state that a certain LR exists, does not give an impression of how mature this LR is, and could therefore skew the results if one wants to get an impression of the depth and breadth of an HLT landscape, especially in a resource-scarce context.

Data collection through the audit questionnaire proved to be a major burden. We recall that for the Dutch BLaRK, a checklist approach was followed, while a number of field workers were used to gather information (Binnenpoorte *et al*., 2002). However, the financial scope of the SAHLTA did not allow for the luxury of field workers, and we therefore had to use a comprehensive questionnaire instead. In hindsight, we have to draw the conclusion that the audit questionnaire might have been too cumbersome in terms of the number of information fields required from the participants. In many instances the compulsory information required for an HLT item (e.g. technical description) was not easily accessible or available on-hand for the participant, or if a LR was out-dated or not well-documented, then it was even more challenging to gather information about it. Thus, if one cannot use field workers, one should rather opt for a simpler check-list approach; the qualitative data might then be less informational (e.g. not have information about performance, or file sizes, or related publications), but one might get data more easily.

In addition, we also experienced that, despite monetary incentives for primary participants, the response rate was in some cases rather slow/low. In personal communications with such participants, it became apparent that they either did not value the true contribution of such a detailed audit, or that they were reluctant to share information because they were uncertain what the implications of it might be. Only after explaining the value of an audit, its findings and the possibility of a national HLT database that captures this information (which would be freely available for their perusal), did they become more active in their participation.

In hindsight, we have learned that an audit like this should follow a bottom-up approach: if the community does not share an understanding of the real value of such an audit, or if they do not have a real need to get access to the results of such an audit (e.g. a national database, or the potential to get funding), the process is hampered considerably.

We therefore also conclude that it is imperative that data should be captured in a national, online database that is freely accessible by the local and international HLT community, and that it is kept up to date on a regular basis. Such a database can be used for a number of purposes, such as road-mapping exercises, networking and identifying collaborations, determining availability of LRs, gauging the gaps in HLT development for a language, and setting funding priorities. In this regard, the South African National Centre for HLT could play a pivotal role in keeping the information for South Africa updated. In addition, it could play a role in helping to create awareness about such audit efforts, and liaise

with other local and international LR infrastructures (such as the ISCA[15] special interest group on speech and language technology for minority languages and AfLaT[16]).

Technology audits are primarily used as tools for further planning around a technology domain. In this paper, we have presented an audit methodology for codifying knowledge about the HLT domain. We have explicated a process that could be repeated in other contexts/countries, and described some of the instruments (e.g. questionnaires, indexes, etc.) through which data could be captured and presented. In our opinion, one should strive to do a once-off extensive audit like this; after that, auditing should be organic, supported by good governance and buy-in from the community.

Acknowledgments

The Department of Science and Technology of the South African Government is acknowledged hereby for financial support of the SAHLTA. We would also like to express our gratitude to anonymous reviewers for their detailed feedback. All fallacies remain ours.

References

Badenhorst, J., Van Heerden, C., Davel, M., & Barnard, E. (2011). Collecting and evaluating speech recognition corpora for 11 South African languages. *Language Resources and Evaluation*. Special Issue: African Language Technology.

Bross, U. (1999). Technology Audit as a Policy Instrument to Improve Innovations and Industrial Competitiveness in Countries in Transition. Innovation, 12(3), 397-412.

Binnenpoorte, D., De Vriend, F., Sturm, J., Daelemans, W., Strik, H., & Cucchiarini, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, In: Proceedings LREC 2002, (3rd International Conference on Language Resources and Evaluation), Las Palmas, Spain 2002, 1862-1866.

Davel M., & Barnard, E., (2003). Bootstrapping in Language Resource Generation. In: Proceedings of the Symposium of Pattern Recognition Society of South Africa, Langebaan, South Africa, November 2003, 97-100.

Department of Arts and Culture (DAC). (2002). National Language Policy Framework. Department of Arts and Culture, Pretoria, South Africa. www.info.gov.za/otherdocs/2002/langpolicyfinal.pdf. Accessed February 2008.

[15] ixa2.si.ehu.es/saltmil

[16] www.aflat.org

D'Halleweyn, E., Odijk, J., Teunissen, L.M., & Cucchiarini, C. (2006). Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 761-766.

Elenius, K., Forsbom, E., & Megyesi B. (2008). Language Resources and Tools for Swedish: A Survey. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesh, Morocco, 600-604.

Joscelyne, A., & Lockwood, R. (2003). Benchmarking HLT progress in Europe. EUROMAP Language Technologies. Center for Sprogteknologi, Copenhagen. www.cervantes.es/seg_nivel/lect_ens/oesi/EUROMAP-Final-Report-Full-May-2003.pdf. Accessed June 2009.

Khalil, T.M. (2000). Management of technology – the key to competitiveness and wealth creation. McGraw-Hill: New York.

Krauwer, S., (1998). ELSNET and ELRA: A common past and a common future. In: The ELRA Newsletter, 3(2).

Krauwer, S. (2006). Strengthening the smaller languages in Europe. In: Proceedings of the 5th Slovenian and 1st International Language Technologies Conference, October 9-10, 2006, Ljubljana, Slovenia.

Maegaard, B., Krauwer, S., Choukri, K., & Jørgensen, L. (2006). The BLARK concept and BLARK for Arabic. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, 773-778.

Maegaard, B., Krauwer, S., & Choukri, K. (2009). BLaRK for Arabic. MEDAR – Mediterranean Arabic Language and Speech Technology. www.medar.info/MEDAR_BLARK_I.pdf. Accessed June 2009.

Mapelli V., & Choukri K. (2003). Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps. European National Activities for Basic Language Resources (ENABLER) Thematic Network. www.ilc.cnr.it/enabler-network/reports.htm. Accessed June 2009.

Martino J.P. (1994). A technology audit: Key to technology planning. In: Proceedings of the IEEE National Aerospace and Electronics Conference NAECON 1994, Dayton, Ohio, USA, 1241-1247.

Pilon, S., Van Huyssteen, G.B., & Van Rooy, B. (2005). Teaching Language Technology at the North-West University, In: Proceedings of the Second ACL-TNLP Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Michigan, Ann Arbor, United States of America, 57-61.

Probert, D., Farrukh, C., Gregory, M.., & Robinson, N. (1999). Linking technology to business planning: theory and practice. *International Journal of Technology Management*, 18(1-2), 11-30.

Sharma Grover, A. (2009). A Technology Audit: The State of Human Language Technologies R&D in South Africa (Masters research report). Graduate School of Technology Management, University of Pretoria.

Sharma Grover, A., van Huyssteen G.B, Pretorius, M.W. (2010a). The South African Human Language Technologies Audit. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Malta, 2847-2850.

Sharma Grover, A, Van Huyssteen, GB & Pretorius, MW. (2010b). A technological profile of the official South African languages. 2nd Workshop on African Language Technology: AfLaT 2010 at the 7th International Conference on Language Resources and Evaluation (LREC 2010), Malta, 3-7.

Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., & Doikoff D. (2004). A Language Resources Infrastructure for Bulgarian. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 1685-1688.