An HLT profile of the official South African languages

Aditi Sharma Grover[1,2], Gerhard B. van Huyssteen[1,3], Marthinus W. Pretorius[2]
Human Language Technology Research Group, CSIR[1],
Graduate School of Technology Management, University of Pretoria[2],
Centre for Text Technology (CTexT), North-West University[3]
HLT RG, Meraka Institute, CSIR, P.O. Box 395, Pretoria 0001, South Africa
asharma1@csir.co.za, gvhuyssteen@csir.co.za, tinus.pretorius@up.ac.za

Abstract

Human language technologies (HLT) have been identified as a priority area by the South African government to enable its eleven official languages technologically. We present the results of a technology audit for the South African HLT landscape, and reveal that a number of HLT components are available in SA but are of a very basic and exploratory nature and much investment is needed in the development of HLT language resources (LRs) in SA. 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 SA.

1. Introduction

Over the past few years, the South African government has realised the role that human language technology (HLT) could play in bridging the digital divide in South Africa. Various research and development (R&D) projects and initiatives have been funded by government, notably through its Department of Arts and Culture (DAC), Department of Science and Technology (DST), and National Research Foundation (NRF). For a historical perspective on HLT policy and non-R&D initiatives in South Africa, see Roux & Du Plessis (2005) and Sharma Grover *et al.* (submitted) for recent initiatives.

In 2009 the National HLT Network (NHN), funded by the DST, conducted the South African HLT audit (SAHLTA). The need for a technology audit is evident in the HLT community where discourse with respect to R&D is vibrant, but with a lack of a unified picture that presents the technological profile of the South African HLT landscape. We present in this paper the results of SAHLTA, focussing on a technological profile of the official South African languages.

2. SAHLTA Process: A Brief Overview

The BLaRK concept (Binnenpoorte *et al.*, 2002 and Maegaard *et al.*, 2009) was chosen to guide the audit, since it provides a well-defined structure to capture the different HLT components as data, modules, and applications.

A questionnaire was used as the primary means to gather data, capturing relevant information according to set criteria. This questionnaire was sent to all major HLT role-players in the country, with the request to supply detailed information regarding LRs and applications developed at their institutions. This audit questionnaire consisted 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, or that aid the development of HLTs (e.g. annotation tools, or corpus searching tools); each section includes the most relevant audit criteria (e.g. maturity, accessibility, quality) for that particular category.

The audit questionnaire was sent to all major HLT role-players in the country. Organisations approached were classified as primary (universities, science councils, and companies-15) or secondary (national lexicography units, government departments-12) participants, based on their historical core HLT competence in R&D. All primary participants were paid a minimal honorarium to compensate for the considerable effort that was required from them.

For further details on the SAHLTA process and instruments used, see Sharma Grover *et al.* (2010).
In order to compare data (e.g. languages with each other), we experimented with various (subjective) ways to quantify the data. We developed a number or indexes in order to represent the technological profiles of the South African languages comparatively; these indexes are discussed and presented below.

3. Maturity Index

The Maturity Index measures the maturity of components by taking into account the maturity stage (i.e. development stage) of an item against the relative importance of each maturity stage. The 'maturity sum' per item grouping (e.g. 'pronunciation resources') for each language is calculated as:

$$MatureSum = 1 \times UD + 2 \times AV + 4 \times BV + 8 \times RV \quad (1)$$

where UD is the number of components in the 'under development' phase, AV is the number of 'alpha version' components, BV the number of 'beta version' components, and RV 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. Table 1 illustrates the maturity sum calculation for 'pronunciation resources'; the maturity sum for English will be 17, since English has one 'under development' item, no 'alpha' or 'beta version' items and two 'released' items; thus its maturity sum is

calculated to be (1×1+0×2+0×4+2×8)=17.

Data item grouping	Maturity Stage	Maturity Weight	Eng	Afr	Zul	Xho	Ndb	Ssw	Ses	Sep	Sts	Xit	Tsv	Lang. independent
Pronunciation resources	Under development	1	1	1	1	0	0	0	1	1	1	0	0	-
(Phoneme sets, Pronunciation dictionaries, Multilingual pronunciation lexicons, Pronunciation models, Intonation models)	Alpha version	2	0	0	0	0	0	0	0	0	0	0	1	-
	Beta version	4	0	0	0	0	0	0	0	1	0	0	0	-
	Released	8	2	5	4	4	3	3	4	3	3	3	3	-
	No. of Items		3	6	5	4	3	3	5	5	4	3	4	-
	Maturity sum		17	41	33	32	24	24	33	29	25	24	26	-

Table 1: Example of a 'maturity sum' calculation for pronunciation resources.

Maturity sums were calculated across component groupings for all data, modules and applications per language. To obtain a comparative approximation of the maturity across 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). Table 2 presents this Maturity Index per language; note that this index is a relative index, based on the number of components that exist in a language.

Lang	MatureInd	AccessInd	LangInd
SAE	26.0	28.2	54.2
Afr	37.9	36.7	74.6
Zul	21.7	25.0	46.7
Xho	20.9	22.3	43.2
Ndb	11.5	11.0	22.5
Ssw	11.6	11.4	23.0
Ses	17.7	20.4	38.1
Sep	18.1	22.3	40.4
Sts	18.5	21.9	40.4
Xit	10.9	11.0	21.9
Tsv	11.9	12.1	24.0
L.I	10.2	8.6	18.8

Table 2: Maturity index, Accessibility index and Language index per language[1].

4. 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:

$$AccessSum = 1 \times UN + 2 \times NA + 4 \times RE + 8 \times CO + 12 \times CRE \quad (2)$$

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

Accessibility sums were calculated across component groupings for all data, modules and applications per language. The Accessibility Index (per language) provides a comparative approximation of the accessibility of HLT components across all the languages. It was calculated by normalising the grand total of the accessibility sums from all the data, modules and applications component groupings per language, by dividing it with the sum of the weights of the accessibility stages (1+2+4+8+12=27). Results for the Accessibility Index are also presented in Table 2.

5. Language Index

The Language Index provides an impressionistic 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):

$$LangInd = Maturity\ Index + Accessibility\ Index^2 \quad (3)$$

From the Language Index presented in Table 2, it emerges that Afrikaans has the most prominent technological profile of all the languages, followed by the local vernacular of South African English. The fact that Afrikaans scores higher than English on this index, can be attributed to the fact that very relatively little work on South African English is required within the text domain; South African English will therefore almost always only be measured in terms of activity related to speech technologies.

The two languages with the most native speakers, isiZulu and isiXhosa (both Nguni languages) follow behind English, and have both slightly more prominent profiles compared to the Sotho languages (Sepedi, Setswana and Sesotho). This can be attributed to the fact that isiZulu and isiXhosa are often of larger commercial and/or academic interest, because they are used more widely throughout South Africa. At the tail-end are the lesser-used languages, viz. Tshivenda, Siswati, isiNdebele and Xitsonga. These four languages significantly lag behind in terms of HLT activity; the majority of items available for these languages were developed quite recently, and are mainly due to the South African government's investment in these languages.

[1] SAE – South African English, Afr – Afrikaans, Zul – isiZulu, Xho – isiXhosa, Ndb – isiNdebele, Ssw – SiSwati, Ses – Southern Sotho (Sesotho), Sep – Northern Sotho (Sesotho sa Leboa/Sepedi), Sts – Setswana, Xit – Xitsonga, Tsv – Tshivenda, L.I – language independent.

[2] The Maturity Index and the Accessibility Index here is on a per language basis, taken across all data, modules, applications as discussed in section 3 and 4 respectively.

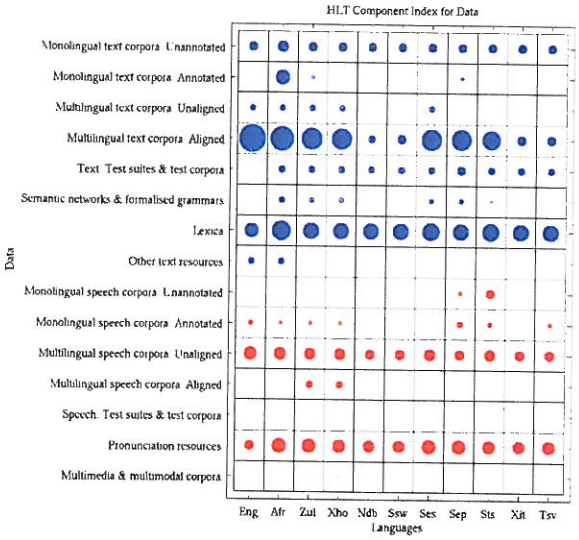

Figure 1: HLT Component Index for Data

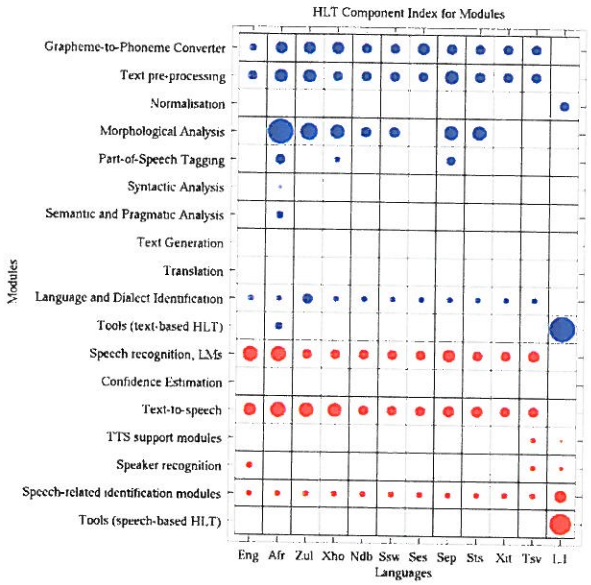

Figure 2: HLT Component Index for Modules

6. Component Indexes

The Component Indexes provide an alternative perspective on the quantity of activity taking place within the data, modules and application categories on a component grouping level (e.g. pronunciation resources), and is calculated as follows:

Component Index= Maturity Index (per item grouping) + Accessibility Index (per item grouping)[3] (4)

The Component Indexes for all languages are plotted in a grid using a bubble plot (see Figures 1 and 2; for the index on applications, see Sharma Grover *et al.*, submitted). The value of the Component Index for a particular component grouping determines the size of the bubble (i.e. the higher the index the larger the bubble). It is important to note that the size of the bubbles plotted within a plot is proportional to the highest value of the Component Index within that specific 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.

Figure 1 depicts the plot for the Component Index for data, where aligned multilingual text corpora have the highest score (which implies the greatest quantity of mature and accessible activities), followed by 'lexica' and so forth. In general it can be seen that speech data resources (in red) have less activity compared to text resources (in blue).

The figure also reveals that although there may be activity in many of the data sub-categories (e.g. 'semantic networks and formalised grammars'), it is very small (implying less maturity and accessibility), or does not exist across all the languages.

Figure 2 reveals that the largest amount of activity is in the field of morphological analysis in Afrikaans, as well as in text-based and speech-based tools that are language-independent (indicated by 'LI'). In general, there is some medium-scale activity in the basic HLT modules for both text and speech domains, such as grapheme-to-phoneme conversion, morphological analysis, speech recognition, and text-to-speech, while the more advanced modules are mostly non-existent or barely available for a few languages.

7. Discussion

The audit's findings reveal that while there is a significant level of HLT activity in South Africa, there are considerable differences in the amount of activity across the languages, and in general the LRs and applications currently available are of a very basic nature. In order to understand this, we need to take a holistic view of the current HLT landscape in South Africa; below we reflect on several factors that might have an influence on this.

HLT expert knowledge: Linguistic knowledge plays a crucial role in HLT enabling a language. In general, the availability of linguistic experts in South Africa is, compared to Europe or the USA, limited. Due to historical imbalances of the past, more linguistic expertise and foundational work are available for Afrikaans and South African English. This is followed by languages such as isiZulu, isiXhosa, Sepedi, Setswana and Sesotho, which have a larger pool of native speakers in South Africa, and thus a greater likelihood of linguistic experts' availability (as opposed to the smaller languages like Tshivenda, isiNdebele, SiSwati and Xitsonga).

Availability of data resources: The frequency and availability of text (e.g. newspapers, books, periodicals, documents) and speech sources (e.g. audio recordings) 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). This is a challenge since the R&D community is constantly faced with limited data collections when working with the

[3] The Maturity Index and Accessibility Index used here are calculated for each grouping of HLT components within data, modules and applications (for example, lexica, corpora, morphological analysis, translation, etc.).

African languages, which in turn severely inhibits the HLT development of the latter group.

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, while the functional status refers to the usage of a language in various public domains. 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.

Relatedness to other world languages: Cross-language information reuse and bootstrapping approaches (Davel & Barnard, 2003; De Pauw *et al.*, 2006) based on other linguistically similar languages can be used to initiate HLT development of new languages. Linguistically, Afrikaans is very similar to Dutch, and thus has benefitted and leveraged on the HLT approaches and developments made in Dutch. Conversely, South African English has not received significant attention in HLT R&D, since researchers leverage on adapting and reusing other international English LRs rather than investing in South African English LR generation from scratch. This leads to a lesser amount of home-grown South African English LRs, and explains the lesser position South African English takes to Afrikaans.
In contrast to the above, African languages are linguistically very dissimilar to any of the European or other world languages, and thus cannot leverage on an existing pool of closely related language resources. This fact, coupled with the complexity of African languages (e.g. tone, clicks, linguistic structure, 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.

Interplay of the above-mentioned factors with the socio-economic and political background of South Africa has shaped the HLT efforts across the South African languages, which has resulted in significant differences in the level of HLT activity across the eleven official languages of South Africa.

8. Conclusion

The South African HLT community is faced with the challenging task of balancing political needs (i.e. attention to all official languages equally) with economic viability (i.e. create a thriving HLT industry, with return on investment on HLT for a certain language). South Africa is far from being unique in this sense, and a number of recommendations can be made for accelerating HLT development in SA and other (African) countries.

Resource development and distribution: From the results it is discernible that basic core LRs need to be built for all languages. However, it is also important to note that while building basic LRs should be prioritised, one needs to start building experience in developing more advanced LRs 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 audit 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, LRs should preferably be made freely available in the open source domain; alternatively, where these are subject to intellectual property rights for commercial use, LRs should be available at a price that does not prohibit their usage.

Funding: The principal sponsors of HLT development for resource-scarce languages are often the governments of those countries. In contrast, the HLT industries in such countries often only comprise 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 the formative years governments needs to continue to invest in HLT efforts (especially in the development of LRs) to build a strong foundation of HLT outputs, which could create thriving HLT industries in such countries.

Industry stimulation programmes: Besides funding, governments need to ensure that there are more initiatives to encourage the existing industry's participation in national HLT activities, and to enable the establishment of new HLT-based start-up companies. For example, it has been noted that there is little awareness in the South African commercial sector about the opportunities and positive impact of HLT (e.g. in the financial or ICT sectors). HLT-focussed initiatives could be launched to stimulate R&D partnerships between academia and industry. In addition, industry participation in lesser resourced languages may need to be motivated proactively by such governments.

Collaborations: Closely related to the above-mentioned stimulation programmes is the need for greater collaborations within local HLT communities and the larger international community. One of the challenges is to harness the knowledge and skills developed in local pockets of excellence into a collaborative endeavour. Thus, a more coordinated effort across an HLT community is required to ensure that there is a well-mapped trajectory for LR creation and HLT market development. Also, collaboration across disciplines (e.g. linguistics, engineering, mathematics) should be encouraged, since HLT involves crossing silos of academic disciplines and national borders.

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. For example, there is currently 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. A greater investment needs to be made in generating HLT practitioners who can feed into the emergent HLT industry's pipeline.

Cultivation of niche expertise: It was observed in the audit that a number of language-independent methods have been adopted in creating HLT components for SA. This approach (depending on the LR in question) has the potential to fast-track the development of HLTs across other resource-scarce languages.

9. References

Binnenpoorte, D., De Friend, F., Sturm, J., Daelemans, W., Strik, H. & Cucchinari, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In *Proc. LREC 2002*, Spain.

Davel, M. & Barnard, E., (2003). Bootstrapping in Language Resource Generation. In *Proc. Symposium of Pattern Recognition Society of South Africa*, November 2003.

De Pauw G., de Schryver, G-M, & Wagacha, P.W., (2006) Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček, and K. Pala *(ed), Proc. Text, Speech and Dialogue, 9th International Conference*, LNCS, Berlin: Springer Verlag, vol. 4188, pp. 197–204.

Maegaard, B., Krauwer, S. & Choukri, K. (2009). *BLaRK for Arabic. MEDAR – Mediterranean Arabic Language and Speech Technology*. [Online]. Available: http://www.medar.info/MEDAR_BLARK_I.pdf (accessed June 2009)

Roux J. & Du Plessis, T., (2005). *The Development of Human Language Technology Policy in South-Africa.* In Daelemans, W., Du Plessis, T., Snyman, C. &Teck, L., Multilingualism and Electronic Language Management, Pretoria: Van Schaik, pp. 24–36.

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

Sharma Grover, A., van Huyssteen G.B. & Pretorius, M.W. (submitted). *The South African Human Language Technology Audit*. [submitted for publication in Journal for Language Resources and Evaluation]

Sharma Grover, A., van Huyssteen G.B & Pretorius, M.W. (2010). The South African Human Language Technologies Audit. In *Proc. LREC 2010*, Malta [accepted].

Acknowledgements

We would like to thank the Department of Science and Technology (DST) for funding this audit. We would also like to acknowledge Professor S. Bosch and Professor L. Pretorius from UNISA whose 2008 BLaRK questionnaire results (for the 2008 NHN-NTU workshop) and preliminary language-specific inventories were used to build the first draft of the cursory inventory of HLT items available in South Africa.