The South African Human Language Technologies Audit

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
E-mail: asharma1@csir.co.za, gvhuyssteen@csir.co.za, tinus.pretorius@up.ac.za

Abstract

Human language technologies (HLT) can play a vital role in bridging the digital divide and thus the HLT field has been recognised as a priority area by the South African government. We present our work on conducting a technology audit on the South African HLT landscape across the country's eleven official languages. The process and the instruments employed in conducting the audit are described and an overview of the various complementary approaches used in the results' analysis is provided. We find that a number of HLT language resources (LRs) are available in SA but they are of a very basic and exploratory nature. Lessons learnt in conducting a technology audit in a young and multilingual context are also discussed.

1. Introduction

Besides the government as a major role-player/funder, the current human language technologies (HLT) landscape in South Africa consists mostly of a relatively young research and development (R&D) community (universities and science councils), and a handful of private companies. The R&D agenda for HLT in South Africa is by and large uncoordinated, and is either academic-driven (e.g. linguists, or computer scientists with an HLT interest), and/or of a highly pragmatic nature (e.g. make something that works well enough for its purpose, without extensive documentation or distribution plans).

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 create a thriving HLT industry. One of the key challenges is the perceived fragmentation of the R&D activities in this domain: there is insufficient codified knowledge about existing South African language resources (LRs) and applications. Hence, in 2009 the South African National HLT Network (NHN) – an informal online community of South African HLT role-players – undertook a large-scale technology audit for the HLT landscape in South Africa (henceforth SAHLTA).

In the next section we describe related work, in section 3 we present an overview of the SAHLTA process and instruments used. A concise overview of some of our results is presented in section 4, while section 5 gives directions for future work.

2. Related Work

In the international field of HLT, a number of 'technology audit'-like efforts have been undertaken. The earliest of such formal audits was the Dutch HLT survey (Binnenpoorte *et al.*, 2002), which applied Krauwer's (1998) concept of the 'basic language resource kit' (BLaRK) 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 advanced HLT development through the Extended Language Resource Kit (Mapelli *et al.*, 2003) and condensed to an entry-level BLaRK, termed the BLaRKette (Krauwer, 2006) for severely under-resourced languages.

3. Process and Instruments

3.1 Terminology, audit criteria and cursory inventory

We commenced by developing an HLT audit terminology list to establish the nomenclature, taxonomy and descriptions for the data, modules and applications to be used in the audit (and thereby creating a common frame of reference). Whilst the Dutch and Arabic efforts provided a useful point of departure, some adaptation was required for the South African context. For example, the application categories defined by Binnenpoorte *et al.* (2002) were the most relevant ones for Dutch at that moment in time. We refined these taking into account the differences in the South African market needs and the current level of technological advancement of South African languages, e.g. categories for audio search, and reference works were added.

Our second step involved defining an HLT inventory criteria framework that specified the criteria or dimensions on which the HLT components would be audited and documented. Here, we closely followed the Dutch and Arabic BLaRK efforts; however, we customised it to include the most important and relevant criteria for the South African context (see below). Concurrently, we built a cursory inventory to identify existing HLT components for each of the eleven South African languages, across the major HLT role- players in the country. As input, we used the preliminary results of a previous informal 2008 mini-BLaRK survey, and complemented this with web-searches and consultations

with a few local HLT experts.

3.2 Audit workshop

An audit workshop with eight South African HLT experts (both speech and text) was organised in July 2009. During this workshop, a session was dedicated to refine and obtain consensus on the terminology list within the South African context. Another session developed the first draft of priorities for applications and associated LRs. The most relevant factors considered were:

- International trends;
- Local market needs; and
- Technical feasibility.

The inventory criteria framework developed earlier was also refined and verified during the workshop. The final audit criteria/dimensions framework is summarised below:

- Technical description (e.g. description, size (tokens, hrs), stratum, etc.);
- Availability:
  o Accessibility (e.g. available for commercial purposes, etc.);
  o Maturity (under development; released, etc.);
  o Distribution (e.g. website, CD-ROM, etc.);
  o Licensing;
  o Cost;
- Documentation (details of publications, reports, website, user manuals, patents, etc.);
- Quality:
  o Verification and/or proof of quality (manual verification of data sets, accuracy, etc.);
  o Compatibility with standards (based on standards or guidelines); and
- Reusability/adaptability (compatibility with other data formats, standard tools/platforms, etc.).

3.3 Audit questionnaire

The inventory criteria framework formed the backbone for the audit questionnaire (spreadsheet) which consists of three major sections: data, module and applications and includes the most relevant audit criteria for that particular section. Another section, 'Tools/Platforms', was also added later to accommodate technologies that are typically language-independent, or aid the development of HLTs (e.g. annotation or corpus searching tools). Prior to roll-out the questionnaire was piloted with a few HLT experts; this proved to be most valuable in identifying information fields that could be potentially misinterpreted. To further aid participants (and stimulate response rates), example entries that illustrate how to record information were provided for each category.

The audit questionnaire was sent to all major HLT role-players in the country. Organisations approached were classified as primary (universities, science councils, companies-15) or secondary (national lexicography units, government departments-12) participants, based on their historical core HLT competence in R&D. Participants

were sent a list of the HLT components identified at their institution (as identified in the cursory inventory), and were requested to fill out all sections for the eleven languages. All primary participants were paid a minimal honorarium to compensate for the considerable effort that was required from them.

3.4 Data analysis

We experimented with various (subjective e.g. indexes) ways to quantify the data, to represent it in a "consumable", bird's eye-view format and provide an impressionistic view of the HLT landscape in South Africa. The final step in our audit process involved an inventory gap analysis, which identified the gaps between the current status of HLT components in South Africa, and the prioritised South African HLT components (as identified during the workshop). This inventory gap analysis could be highly informative for future road-mapping exercises, as well as to immediately identify areas or languages that should receive special attention (see below).

4. Results and Discussion

Due to the limited scope of this paper, we present here only a general overview of some of our results.

4.1 The South African HLT landscape

In order to compare the state of HLT development 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, as well as the stage of maturity and accessibility of their HLT LRs and applications.

Figure 1 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 LRs and applications, followed by the local vernacular of English (with a significant difference between the two). This picture is 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.
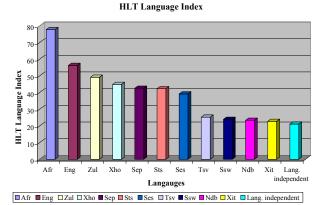

Figure 1: The South African 'HLT Language Index'[1]

[1] Afr – Afrikaans, Eng – South African English, Zul – isiZulu, Xho –

IsiZulu, isiXhosa, Sepedi, Setswana and Sesotho (the five African languages in South Africa with the most native speakers) follow behind English; the two Nguni languages (isiZulu and isiXhosa) have slightly more activity in the field of HLT, compared to the Sotho languages (Sepedi, Setswana and Sesotho). This can be attributed to the fact that isiZulu and isiXhosa are by far the two largest languages in South Africa, used in a variety of domains and in various provinces – and thus often of larger commercial and/or academic interest.

At the tail-end are the lesser-used languages such as Tshivenda, Siswati, isiNdebele, Xitsonga (language independent items are also included in the index, right at the end). 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.

Secondly, we also developed an 'HLT Component Index' that provides an alternative perspective of the quantity of activity taking place within each of the data, modules, and applications categories on a HLT component grouping level (e.g. pronunciation resources); this allows HLT practitioners to ascertain areas where further work is required across the different languages. Figure 2 (see end of paper) illustrates the HLT Component Index for modules.

For example, from this figure, a funding agency could deduce that money should rather be invested in, for example, syntactic analysis, than in grapheme-to-phoneme conversion; or that special attention should be given to morphological analysis of Sesotho, Xitsonga and Tshivenda. It should be noted that the relative size of a bubble does not necessarily indicate that technologies in that section or for that language are mature – it is merely an indication of technologies and the South African languages relative to each other (and also not relative to other world languages).

Similar to the above results and representations, we have also developed a maturity index, an accessibility index, a detailed HLT inventory analysis (per sub-category within data, modules, applications), and an inventory gap analysis (see Sharma Grover (2009) for the comprehensive results).

Based on these, our over-all impression of the South African HLT landscape is that very few basic LRs and applications exist across all eleven languages; it is especially the four smallest languages that lag far behind in terms of HLT development. It is also clear that there are a great many areas that lie fallow across all the South African languages in terms of the variety, number and maturity of items, especially compared to other world languages.

4.2 About the audit process

Besides the audit findings, we also learnt a number of lessons about how an HLT audit should be conducted, especially in a young and multilingual context.

Data collection through the audit questionnaire proved to be a major burden. Many participants, although active role-players in the field, provided only basic information and had to be prompted personally in the post-questionnaire phase. Although this is common in research, a number of specific reasons for this slow/low response could be identified. We recall that for the Dutch BLaRK, a checklist approach was followed and a number of field workers were used to gather information (Binnenpoorte, 2002). The financial scope of this audit did not allow the luxury of field workers, and we therefore had to use a questionnaire instead. We draw the conclusion that the audit questionnaire may have been too comprehensive in terms of the number of information fields (dimensions) required.

Maegaard *et al.* (2009) experienced similar challenges in their most recent Arabic BLaRK effort. Comparing our experiences, we conjecture that these challenges arise because measuring quality and other subjective dimensions is a time-consuming, (often) costly and effortful process, requiring dedicated human resources. However, we do believe that it is worth the investment, since the value of an LR should take into account several audit dimensions; for a participant to merely state that a certain LR exists for a certain language, 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 an HLT landscape.

The above-mentioned issues warrant further investigations in the optimisation of data collection approaches (interviews, field workers, web-based questionnaires), and the measurement of the value of an LR in order to build a useful HLT inventory.

In addition, we also experienced that despite monetary incentives, the response rate was in some cases rather slow/low. Only after explaining the value of an audit – its findings and the possibility of a national HLT database that captures this information (freely available for their perusal) – did participation become more active. This reluctance may be possibly ascribed to factors such as the lack of a shared, coherent vision on the potential of a strong HLT R&D industry in a young R&D community.

In hindsight, we have learnt that such audits should follow a bottom-up approach: if the community doesn't share an understanding of the value and need of the audit (e.g. a national database, or the potential to get funding), the process is hampered considerably.

5. Conclusions and Future Work

As part of the feedback and information dissemination process, the findings of the audit will be presented to the South African HLT community at appropriate fora. Refinement and verification of especially the final

priority list for South African LRs and applications will be high on the agenda. Additionally, it is imperative that the audit data should be captured in a national, online database that is freely accessible by the local and international HLT community, and that is kept up to date on a regular basis. In our opinion, one should strive to do a once-off extensive audit like this; subsequent to this, auditing should be organic, supported by good governance and buy-in from the community.

6. References

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

Elenius, K., Forsborm, E., and Megyesi, B. (2008). Language Resources and Tools for Swedish: A Survey. In: Proc. LREC 2008, Marrakesh, Morocco.

Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. (2006). The BLARK concept and BLARK for Arabic. In: Proc. LREC 2006, Genova: 773-778.

Maegaard, B., Krauwer, S., and 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)

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: Proc. of 5th Slovenian and 1st International Language Technologies Conference, October 9-10, 2006, Ljubljana, Slovenia.

Mapelli V., and 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. [Online]. Available: http://www.ilc.cnr.it/enabler-network/reports.htm (accessed June 2009)

Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., and Doikoff D. (2004). A Language Resources Infrastructure for Bulgarian. In: Proc LREC 2004, Lisbon, Portugal: 1685-1688.

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.
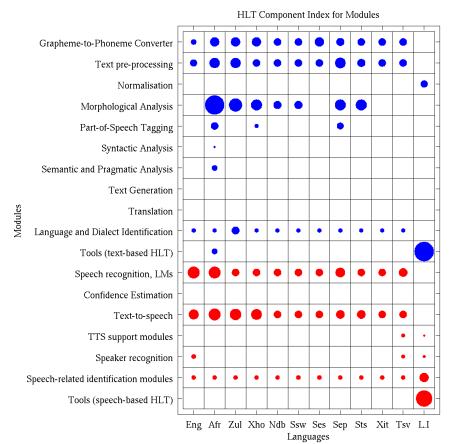
Acknowledgements

Figure 2: HLT Component Index for modules.