An overview of HLTs for South African Bantu languages

Aditi Sharma Grover[1,2], Karen Calteaux[1], Gerhard van Huyssteen[1,3], Marthinus 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
Telephone: +27 12 841 3028
Email: asharma1@csir.co.za, kcalteaux@csir.co.za, gvhuyssteen@csir.co.za, tinus.pretorius@up.ac.za

ABSTRACT
South Africa (SA) is one of the few countries in the world that boasts a large number of official languages. Due to the efforts of government and the local research and development (R&D) community (comprising universities, science councils and a few private sector companies) all the official languages are – to varying degrees – enabled with regard to human language technology (HLT). We present in this paper the current status of HLTs for a few selected official South African languages, namely isiZulu, Sepedi, Tshivenda and, Xitsonga based on a national HLT audit covering all official languages of South Africa. We discuss the HLT position of the above languages in relation to other official South African languages, and also explore the types of data collections, technology modules and applications currently available in the R&D community for these four languages.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]

General Terms
Languages, Management.

Keywords
Human language technology, language resources, language audit, language resource infrastructure, technology audit, Bantu languages, isiZulu, Sepedi, Tshivenda, Xitsonga.

1. INTRODUCTION
Approximately 25 languages are spoken in SA, of which 11 have been afforded official status. The South African National Language Policy Framework (NLPF), which provides guidelines for managing this linguistic diversity, mentions that "human language technologies applications (e.g. machine-assisted translation, translation memories, spelling checkers) for the indigenous languages will play a major supporting role in language facilitation activities" (Department of Arts and Culture, 2002:15). The aim of this paper is to give an overview of HLT work that has been done for a selection of Bantu languages to date, based on a technology audit of the HLT landscape in South Africa that was completed in 2009.

In order to provide an overview of HLTs for South Africa's Bantu languages, we have selected four languages that represent the various language groupings in the country, namely isiZulu (from the Nguni language group), Sepedi (also known as Sesotho sa Leboa– from the Sotho language group), Tshivenda (a smaller language not related to either of the above groups) and, Xitsonga (also a smaller language unrelated to either group). These four languages, together with English and Afrikaans correlate with the six language groupings proposed in the NLPF.

SAICSIT '10, October 11–13, 2010, Bela Bela, South Africa.

2. OVERVIEW OF AFRICAN HLTs IN SA
Most of the work on HLTs for the African languages began with efforts relating to machine translation, as this was seen as a way of managing the linguistic diversity in our country. In 1986, the Human Sciences Research Council (HSRC) commissioned the LEXINET investigation 'to determine the extent to which computer processing of languages abroad might be relevant to South Africa, and to formulate proposals for possible local developments' (Bosch, 2007: 171). The 1988 report on this programme indicated that only two University departments of African languages were involved in computational linguistics at that stage (University of Stellenbosch and Rhodes University), with only one publication on computer-assisted research in the African languages being reported till date. The report also indicated that although several researchers had data corpora at their disposal, very few of these were computerised (Bosch, 2007: 171).

In 1990, the Department of Computer Science at the University of Pretoria developed a uniquely South African machine translation device, named *Lexica* which rendered comprehensible translations from Setswana to English. However, this project ended in 1997, possibly due to poor results in translation accuracy (Bosch, 2007: 172). Since the mid-90s, pockets of expertise in developing HLTs for the African languages have been nurtured at various universities and research institutions. The new millennium saw the launching of a project called African Speech Technology (AST), executed by a consortium of researchers lead by the University of Stellenbosch and completed in 2004. The project involved developing a prototype of a fully automated telephone-based multilingual query and booking system for the hotel industry which would function in five languages viz. – South African English, isiZulu, isiXhosa, Sesotho and Afrikaans.

Over the past decade, the Centre for Text Technology (CTexT™) at the North-West University has also become a leading role player in the development of text-based technologies for the official languages. A noteworthy highlight of their work is the Autshumato project which is aimed at developing a terminology management system, word translators for ten South African language pairs (covering eleven languages), machine-translation systems for three South African language pairs (covering four languages), and a basic document

management solution. CTexT has also developed spell checkers for all official languages; the first-ever Afrikaans grammar checker; multimedia language acquisition software packages for isiZulu, isiXhosa, Setswana and Afrikaans that enable a learner to acquire the basics of a new language in less than 48 hours; web search engines; and translation programs (CTexT).

Since 2003, the Human Language Technologies Research Group (HLTRG) of the Meraka Institute at the CSIR has emerged as a leading role player in the development of speech technologies. Its flagship project, known as Lwazi, was completed in 2009. Lwazi delivered a telephone-based information service which has been piloted in various locations in SA in the eleven official languages. Lwazi packages automatic speech recognition (ASR), pronunciation dictionaries and text-to-speech (TTS) technologies on a telephony platform to make information available to users in spoken form. The HLTRG has also developed various data resources as well as tools which enable the faster and more efficient development of speech technologies. These include the Lwazi ASR and TTS corpora for all the official languages, and tools such as ASR-Builder, DictionaryMaker, and Speect (Meraka Institute).

The late 1990s and early part of the 21st century were also marked by attempts by academics and researchers to encourage government to develop policies with regard to HLT research and resource management in South Africa. Most notably, this lead to the establishment of the National Centre for HLT (NCHLT) in 2009, which could play a significant role in the development and management of HLTs for all South African languages in future. Note, this section is intended as a brief illustrative overview of South African HLT efforts for the African languages and does not attempt to encompass all efforts, for example, smaller instances of isolated work on African languages occurring internationally.

3. HLT STATUS OF AFRICAN LANGUAGES

In 2009, the South African National HLT Network (NHN; an informal online community of HLT role-players in SA), undertook a large-scale technology audit for the HLT landscape in SA (henceforth SAHLTA), which was sponsored by Department of Science and Technology (DST). In this section we further expound on the audit findings specifically related to isiZulu, Sepedi, Tshivenda, and Xitsonga.

3.1 SAHLTA process

The concept of a BLaRK (Binnenpoorte *et al.*, 2002 and Maegaard *et al.,* 2009) was used to guide the audit, which provides a well-defined classification of the different HLT components that were to be audited, namely language resources (LRs, which include *data* and *modules*), and *applications*. Data refers to text or speech data sets in a machine readable form, and which are used to create, evaluate and improve HLT technology modules and includes items such as corpora, lexica and grammars. Modules refer to the basic software units or processes that are usually required to create HLT applications and products, and include items such as part-of-speech taggers, tokenisers, and language and acoustic models for speech recognition. Applications refer to the categories of different end-user or enterprise application areas where HLT is used, and may include domains such as computer-assisted language learning, document production, proofing/authoring tools and machine translation.

The primary means of the audit data collection was a questionnaire that 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. Twenty seven organisations were approached, and feedback from sixteen was received. See Sharma Grover *et al.* (2010) for a more comprehensive overview of the SAHLTA process.

3.2 Language Index

Based on the questionnaire, a number of indexes were designed to give a comparative overview of the position of the South African languages in the local HLT landscape see (Sharma Grover *et al.* 2010). The HLT language index provides an impressionistic comparison on the overall status of HLT development for the eleven official South African languages. This index allows the South African languages to be compared on the basis of the total quantity of HLT activity (data, modules, applications) within a language, whilst also taking into account the stage of maturity and accessibility of the outputs of the HLT activity in South Africa; see Figure 1. From this we see that Afrikaans scores the highest, followed by South African English, then isiZulu, isiXhosa, Sepedi, Setswana, and Sesotho (languages with more native speakers) and finally, the smaller languages such as Tshivenda, siSwati, isiNdebele, and Xitsonga.

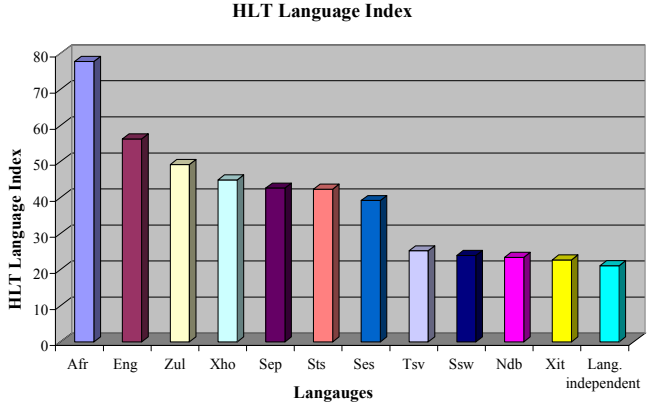

Figure 1. The South African HLT language index.

3.3 Availability of HLT Components

We present in Table 1 (at the end of this paper) a detailed overview of the HLT components (data, modules and applications) available in South Africa for isiZulu, Sepedi, Tshivenda, and Xitsonga for both text and speech-based HLT.

Under text applications we see that across the four languages some of the most significant work has been done in the development of proofing/authoring tools (i.e. spelling checkers and hyphenators). Currently, there are both open source and Microsoft-based spelling checkers available for all of these languages. South Africa also has text-based computer-assisted language learning (CALL) in all the languages, though isiZulu has three items in this area (an interactive PC-based software CD, an open-source CALL game for children, and a proprietary multilingual illustrated dictionary with interactive games). Sepedi, Xitsonga and Tshivenda, on the other hand, only have an open source CALL game for children to date. In terms of machine-aided human translation we find that only isiZulu and Sepedi have some activity, specifically in the context of the Autshumato project.

In the speech domain, all the languages have two telephone-based information services, with the exception of isiZulu, which also has another legacy IVR demo. In addition, under the accessibility area, isiZulu has two ongoing projects in beta

stages, one of which is a communication device for the visually impaired and the other an augmentative and alternate communication (AAC) device that allows speech to be generated through icons.

Focusing our attention on HLT modules, we observe that there are text-based modules such as grapheme-to-phoneme converters, tokenisers and language identification modules for all four languages. Again, we observe that isiZulu and Sepedi have a wider variety of modules such as sentencisers, hyphenators and morphological parser/decomposers. In the speech domain, all the languages have basic ASR and TTS due to the recently completed Lwazi project. However isiZulu has slightly more activity in TTS compared to the other languages, since some prior South African TTS efforts focused on isiZulu and currently multiple research institutes are active in it.

Finally, for data, isiZulu leads in terms of text corpora/data set size, closely followed by Sepedi and finally Xitsonga and Tshivenda (the size of the largest resource type is indicated in brackets for each entry where available). On the speech side, all the languages have some basic annotated ASR speech corpora, pronunciation dictionaries and models, phoneme sets and TTS corpora. However, Sepedi and Tshivenda have a slightly larger speech corpora since there is some developmental work (beta and alpha stages respectively) occurring in these languages at multiple research institutions. In general, these findings highlight that the majority of the HLT activities in isiZulu, Sepedi, Xitsonga and, Tshivenda in SA are currently focussed on building basic and core LRs and applications and not so much advanced LRs and applications.

3.4 HLT Component Index

The HLT component index provides an alternative perspective on the quantity of activity taking place within each of the data, modules, and applications on an HLT component grouping level (e.g. pronunciation resources can consist of phoneme sets, pronunciation dictionaries, pronunciation models, intonation models, etc). This index is calculated by summing the maturity index and the accessibility index (Sharma Grover *et al*, 2010). It is plotted in a grid using a bubble plot and highlights the gaps in the availability of HLT LRs which in turn provides HLT stakeholders with the necessary information to align their future R&D activities accordingly.

Figure 2 shows this index for data for isiZulu, Sepedi, Tshivenda, and Xitsonga. 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). Note, the size of the bubbles within a plot is proportional to the highest value of the HLT component index within the entire plot. From this figure it can be seen that in areas such as 'unannotated monolingual text corpora' and 'lexica' the four languages seem to have similar quantities of work (and maturity/accessibility thereof), whereas in areas such as 'semantic networks' and 'aligned multilingual text corpora', isiZulu and Sepedi have far greater activity taking place as opposed to Xitsonga and Tshivenda.

In general, we see 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 is not uniform across all the languages. We note here that, many of the data collections available currently are of a very basic quality (small sizes, developmental stages, and restricted use), for example, aligned multilingual corpora and lexicons for text domain and annotated speech corpora and intonation models on the speech side. Figure 3 shows the HLT component index for modules. Within the speech modules (in red) we see that basic ASR and TTS are available in these four languages; however, these have only recently come into existence, and thus the activities here are still very much in their formative years. Many other speech modules statistical language models, non-native SR, confidence measures, advanced pre-processing and prosody generation for TTS are unavailable.

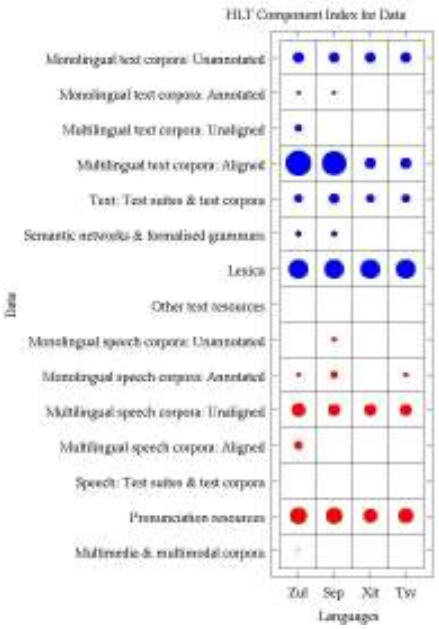

Figure 2. HLT component index for data.

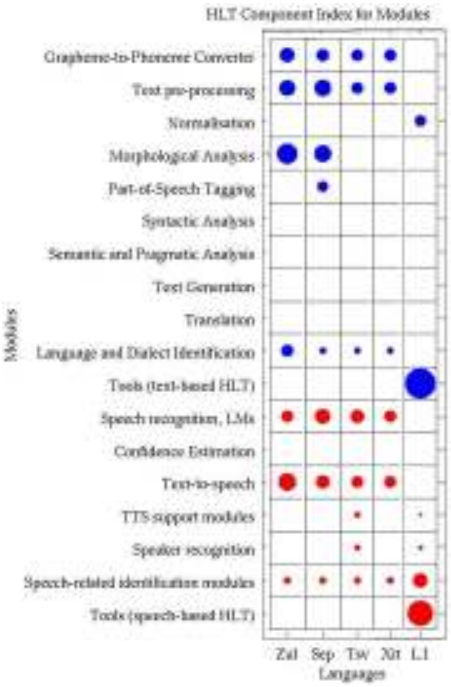

Figure 3. HLT component index for modules.

Similarly, in the text domain numerous important modules are unavailable such as part-of-speech-taggers, shallow parsers, compound analysers and word meaning disambiguators. Many of the other text modules cited as currently available (Table 1) are also of a basic nature and many a times still in

developmental stages. Note, L.I. in figures 3 and 4 refers to language independent HLT components. From Figure 4 which illustrates the HLT component index of applications, we observe that (document production) (proofing/authoring tools) and telephony-based services are the only applications that are currently available across all four languages. In general, we see that for majority of the other applications, there is barely any activity which spreads across the four languages.

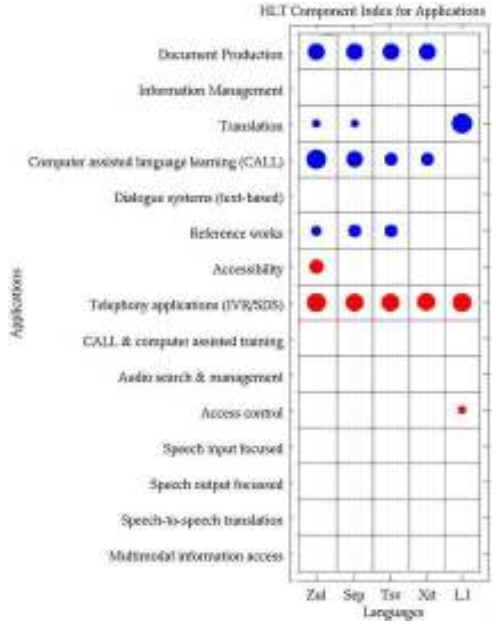

Figure 4. HLT component index for applications.

4. DISCUSSION

Based on the audit findings[1] for isiZulu, Sepedi, Xitsonga and, Tshivenda discussed, not surprisingly, we find that isiZulu has the greatest amount of HLT activity, very closely followed by Sepedi, and then the two smaller languages of Xitsonga and Tshivenda that lag behind. Overall, we see that very few basic HLT LRs and applications exist in these African languages. There are a great many areas (as identified in the gap analysis) that lie bare in terms of HLT development of these languages as compared to Afrikaans, South African English and other world languages. Some of the differences across the four languages in question can be ascribed to a combination of the following aspects:

Firstly, the availability of expert knowledge (both in linguistics and HLT) for a certain language has a significant influence on the HLT profile of that language; based on the proportionate sizes of the languages investigated, we see that there is usually a greater likelihood of finding experts for the larger languages as opposed to the smaller languages like Xitsonga, Tshivenda, isiNdebele, and siSwati.

Secondly, the availability of data collections, such as text sources (e.g. newspapers, books, periodicals, documents) and speech sources (e.g. audio recordings) for the larger languages is far greater than that for the smaller languages. Also, the sheer fact that the larger languages are spoken in the economic hubs of South Africa where the largest research institutions are located (e.g. Gauteng and Western Cape), makes the recruitment of native speakers and data collection tasks practically much easier for these languages in contrast to the smaller languages (which tend to be spoken more generally in remote areas such as the northern parts of Limpopo and eastern parts of Mpumalanga).

Thirdly, the market needs of HLT in a particular language can be viewed as i) a combination of supply and demand factors; and ii) the functional status of the language in the public domain. The former 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 latter 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 (even more so for the smaller languages) are less widely used in such commercial environments. This significantly lowers the economic feasibility of HLT endeavours for the smaller African languages.

The slow pace at which the African languages are responding to their 'new' official status has been a cause for concern for academics, researchers and government role players, such as PanSALB, for some time. A 2001 survey of language use and language interaction [PanSALB] in SA confirmed that English is seen as a language of upward social mobility and the language of business, while the African languages are mainly relegated to community and social use. The South African government will therefore have to play a significant role in order to enable all languages in the HLT domain.

Ironically, higher levels of illiteracy among the speakers of the smaller languages make these languages ideal candidates for HLT-enabled applications such as speech-driven telephone-based information systems aimed at improving access to government service. Ensuring uptake of HLT applications for the African languages in a developing world must take cognizance of this phenomenon, and this should receive particular attention in the short term.

Lastly, relatedness amongst language families could also play a role; isiZulu and Sepedi are members of the Nguni (others are isiXhosa, isiNdebele and siSwati) and Sotho (related to Setswana and Sesotho) language groups respectively, and thus to some extent have also benefited through any HLT development within their language groups, as opposed to Xitsonga and Tshivenda, which are unrelated to the above language groups. Often local researchers also tend to work with a particular larger African language and then naturally expand their work to the other languages within that language group (in contrast to solely focusing on one of the smaller languages). This implies that HLT efforts for the smaller languages commence from the bottom of the development lifecycle, and start by investing in basic LR and linguistic knowledge generation.

5. CONCLUSION

The socio-economic and political background of South Africa in combination with the above aspects has shaped HLT efforts and resulted in significant differences in the level of HLT activity across the eleven official languages of South Africa. Whilst there is noteworthy HLT development occurring in SA, much still needs to done to HLT-enable all of the South African languages and additional effort is needed for the smaller African languages. Foremost on the agenda, should be HLT resource collection initiatives. Here not only government but also industry needs to be tapped into, for example, sources such

[1] The audit data is currently being made available through an online database at www.meraka.org.za/nhnaudit.

as television and radio broadcasts, parliamentary speeches, and other media channels should be more intensely leveraged on, to gather data for the African languages. Furthermore, to encourage innovation, these LRs should preferably be made available in the open source domain. Also in these formative years of HLT for the African languages, it is essential that Government continues to play a crucial role in enabling the development of these languages and more so stimulate industry participation in these languages. This may require further in-depth studies into determining the market needs of the African languages and how best to tap into and service them. Lastly, a greater investment needs to be made in generating HLT practitioners who can feed into the emergent HLT industry's pipeline.

6. REFERENCES

[1] 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: Proceedings LREC 2002, (3rd International Conference on Language Resources and Evaluation), Spain 2002.

[2] Bosch, S. 2007 "African languages – is the writing on the screen?", Southern African Linguistics and Applied Language Studies, 25: 2, 169-181.

[3] CTexT, Centre for Text Technology. www.puk.ac.za/fakulteite/lettere/ctext/index_e.html. [Accessed 31 May 2010]

[4] Department of Arts and Culture. 2002. National Language Policy Framework. Final Draft. Unpublished policy document. Pretoria.

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

[6] Meraka Institute,CSIR. Human Language Technologies – Projects, http://www.meraka.org.za/hlt_projects.htm. [Accessed 31 May 2010]

[7] PanSALB, Pan South African Language Board. 2001. Language Use and Language Interaction in SA. Unpublished language survey report. Pretoria.

[8] Roux, J.C. and Bosch, S. 2006. Language resources and tools in Southern Africa. Paper presented at a workshop of the African Language Association of Southern Africa, Special Interest Group for Language and Speech Technology, 22 May 2006.

[9] Sharma Grover, A. 2009. Technology audit: The state of Human Language technologies R&D in South Africa. Unpublished Masters Thesis. Pretoria: University of Pretoria.

[10] Sharma Grover, A., van Huyssteen G.B, Pretorius, M.W. 2010. The South African Human Language Technologies Audit. In: Proc. LREC 2010, Malta, 2847-2850.

[11] Statistics South Africa. 2001. Census 2001. Key Results. http://www.statssa.gov.za/census01/html/Key%20results_files/Key%20results.pdf. [Accessed 20 May 2010]

Table 1. Overview of HLT components available for isiZulu, Sepedi, Tshivenda, and Xitsonga.

Language / Component		isiZulu	Sepedi	Tshivenda	Xitsonga
Applications	Text	- Proofing/authoring tools (2) (Microsoft & open source) - Computer assisted language learning (CALL) (3) - Machine-aided human translation - Reference works	- Proofing/authoring tools (2) (Microsoft & open source) - Computer assisted language learning (CALL) (1) - Machine-aided human translation - Reference works	- Proofing/authoring tools (2) (Microsoft & open source) - Computer assisted language learning (CALL) (1) - Reference works	- Proofing/authoring tools (2) (Microsoft & open source) - Computer assisted language learning (CALL) (1)
	Speech	- Accessibility (2) - Telephony services (IVR/SDS) (3)	- Telephony services (IVR/SDS) (2)	- Telephony services (IVR/SDS) (2)	- Telephony services (IVR/SDS) (2)
Modules	Text	- Grapheme-2-phoneme converter (2), sentenciser, tokeniser, hyphenator, morphological parser/decomposer (2), language ID.	- Grapheme-2-phoneme converter (2), sentenciser, tokeniser, hyphenator, morphological parser/decomposer (2), POS tagger, language ID.	- Grapheme-2-phoneme converter, tokeniser, language ID.	- Grapheme-2-phoneme converter, tokeniser, language ID.
	Speech	- Complete speech recognition, domain independent TTS (3), spoken language ID	- Complete speech recognition, domain independent TTS (2), spoken language ID	- Complete speech recognition, domain independent TTS (1), spoken language ID	- Complete speech recognition, domain independent TTS (1), spoken language ID
Data	Text	- Unannotated monolingual corpora: ~ 7.4 mil words (7 mil) - Annotated monolingual: 100k - Aligned multilingual corpora: 216k - Test suites & corpora: 30k - Wordnets: 10k synsets - Monolingual lexicons: 7.7 mil (7.4 mil) - Terminology lists: 35k	- Unannotated monolingual corpora: ~ 7.3 mil words (7 mil) - Annotated monolingual: 7 mil - Aligned multilingual corpora: 284k - Test suites & corpora: 30k - Wordnets: 10k synsets - Monolingual lexicons: 1.5 mil (1.4 mil) - Terminology lists: 42k	- Unannotated monolingual corpora: ~ 2.05 mil words (2 mil) - Aligned multilingual corpora: 38k - Test suites & corpora: 30k - Monolingual lexicons: 1.3 mil (1.3 mil) - Terminology lists: 28k	- Unannotated monolingual corpora: ~ 2.05 mil words (2 mil) - Aligned multilingual corpora: 38k - Test suites & corpora: 30k - Monolingual lexicons: ~1.6mil (1.6 mil) Terminology lists: 35k
	Speech	- Annotated speech corpora: ASR: ~15 hrs (12 hrs) ; TTS: 38 mins - Pronunciation dictionaries: ~9k (5k) - Pronunciation models (g2p) - Phoneme sets - Intonation models	- Annotated speech corpora: ASR: ~28 hrs (25hrs) ; TTS: 26mins - Pronunciation dictionaries: ~127k (122k) - Pronunciation models (g2p) - Phoneme sets - Intonation models	- Annotated speech corpora: ASR: ~2 hrs (2hrs) ; TTS: 26mins - Pronunciation dictionaries: ~5k (5k) - Pronunciation models (g2p) - Phoneme sets	- Annotated speech corpora: ASR: ~12 hrs (10hrs) ; TTS: 23mins - Pronunciation dictionaries: ~5k (5k) - Pronunciation models (g2p) - Phoneme sets