HLT DEVELOPMENT IN SUB-SAHARAN AFRICA

REPORT TO COCOSDA / WRITE WORKSHOP
LREC 2008
MARRAKECH

Justus C Roux
Centre for Language and Speech Technology
Stellenbosch University, South Africa

OVERVIEW

Activities in

- South Africa at
  - Academic institutions
  - Semi-governmental institutions
  - Governmental institutions
- West Africa
- East Africa

Websites

South Africa
Academic Institutions (7)

University of Cape Town

University of Limpopo

University of the North West (Potchefstroom)

University of Pretoria

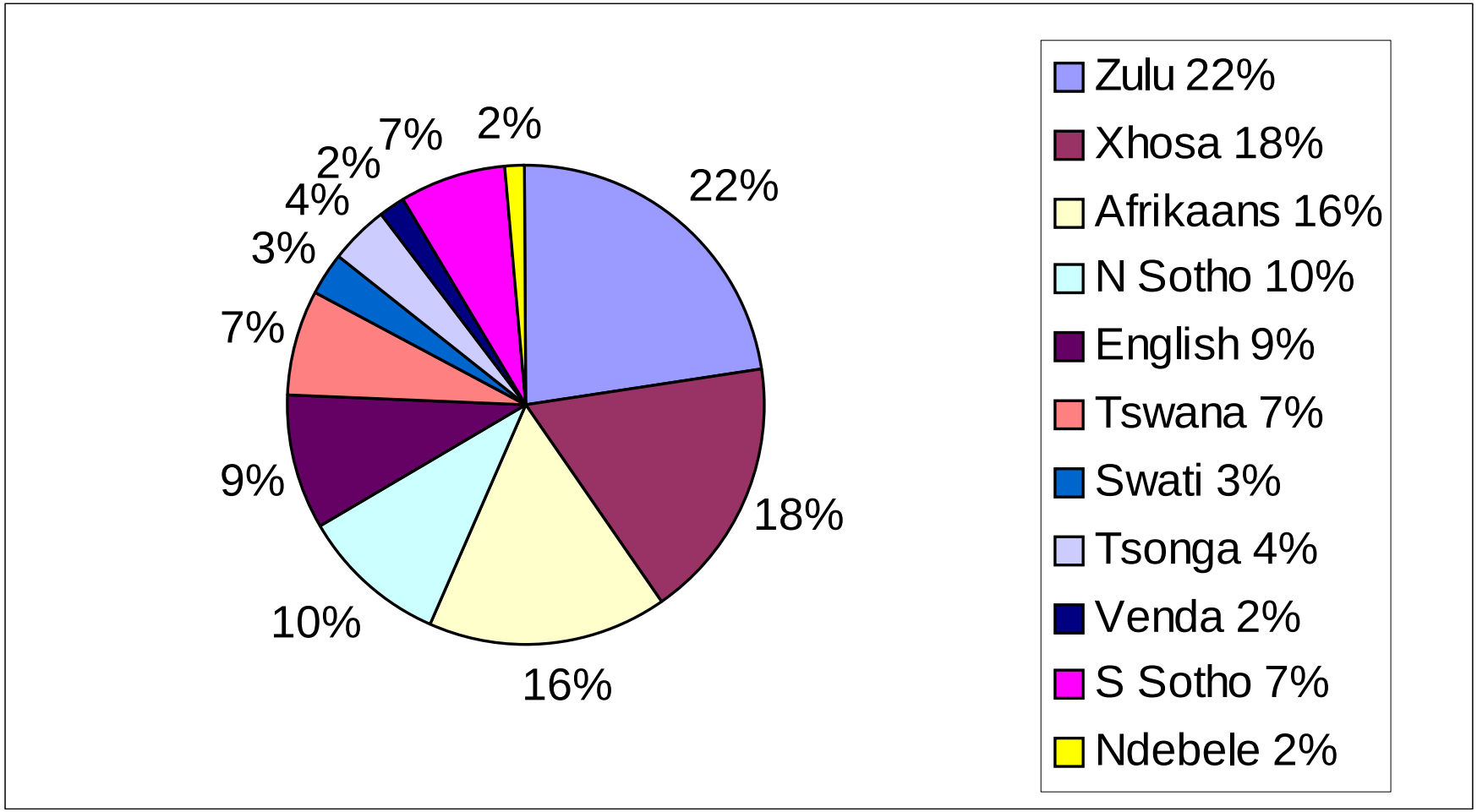
University of South Africa

University of Stellenbosch

University of the Witwatersrand (Johannesburg)

Language Situation

Mother tongue division (n=44,8 mil speakers - 2004)



University of Cape Town

Speech Technology and Research (STAR, S*)

http://www.star.za.net

Electronic Engineering **-** Focus on speaker recognition.

University of Limpopo

Telkom Centre of Excellence for Speech Technology
http://www.ul.ac.za [Mr Jonas Manamela]

- Synthesizers for African languages: N Sotho (Sepedi), Tswana, Venda and Tsonga;
- Tools for automatic recognition of continuous speech for N Sotho (Sepedi), Tswana, Venda and Tsonga;

North West University (Potchefstroom)

Centre for Text Technology (CTexT™)

http://www.ctext.co.za [Prof Gerhard van Huyssteen]

- OpenSource machine translation systems: English to Afrikaans, Zulu and Northern Sotho.
- One of three centres in the world to develop all future spelling checkers for Microsoft.
  - In 2007, CTexT developed lexical data for 5 African Languages (Igbo, Hausa, Kinyarwanda, Wolof, Yoruba),
- Spelling checkers for ten South African languages for use in the public administration domain (2007-2008).
- The first Afrikaans grammar and style checker (to be released by late 2008)

CTexT (2)

- **Tools**
  - Stemmer for Zulu, Xhosa /Lemmatiser for Tswana, Afrikaans
  - POS Taggers (Afrikaans)
  - WebCrawler – Engine for automatic extraction and organisation of specified texts from the WWW, using novel techniques in automatic language identification.
  - Corpus alignment tools
- **Corpora**
  - Wide variety of corpora for local languages, Dutch Hausa, Igbo, Yoruba, Wolof and Kinyarwanda.
- **Commercial products**
  - Variety spell checkers, language learning systems

University of Pretoria

Department of African Languages

http://web.up.ac.za [Prof Danie Prinsloo]

- Leader in the building of machine-readable corpora for all of the official languages of South Africa.
- Developing HLT-tools for the computational processing of Northern Sotho (Sepedi).
- Localisation of Microsoft Windows XP operating system into Zulu

University of South Africa (UNISA)

Pretoria

Department of African Languages [Prof Sonja Bosch]

http://www.unisa.ac.za

- **Computational morphological analysis project**
  - Computational morphological analysis of Zulu, Xhosa, Ndebele, Swati, N-Sotho, Tswana,
  - development of XML machine-readable lexicons for above languages
- **Morpho-syntactic annotation**
- **Collaboration with various universities providing expertise in morphological analyses**

Stellenbosch University Centre for Language and Speech Technology

http://www.sun.ac.za/su_clast [Justus Roux]

- **Speech processing**
  - Speech-to-Speech Translation
  - High quality speech synthesis in African tone languages (with ATR, Japan)
  - Large vocabulary speech recognition for SA English
  - Multilingual acoustic modelling.
  - Multilingual and multi-accent speech recognition in SA languages
- **Language learning systems**
- **Mobile learning systems**
  - Implementing multilingual TTS based learning systems on mobile platforms in collaboration with University of Copenhagen

SOUTH AFRICA
SEMI-GOVERNMENT INSTITUTIONS

MERAKA INSTITUTE (Pretoria)

http://www.meraka.org/nhn [Dr Marelie Davel]

Lwazi:

- Development of tools and technologies that support the provision of multilingual telephony-based information services in all South Africa's official languages.
  - the creation of an Open Content linguistic resource repository containing annotated speech data for ASR and TTS purposes in 11 languages
  - the development of ASR & TTS systems in 11 languages
  - all data and tools are distributed free of charge under an Open Source license.

MERAKA (2)

OpenPhone

- Development, piloting and evaluation of a telephony-based information service capable of assisting home caregivers in caring for HIV-positive children in Gabarone, Botswana.
  - an ASR and TTS system in the dialect of Setswana spoken in the Gabarone region.

PAST (Phonetics for Advanced Speech Technology)

- A collaborative project – Universities of Witwatersrand and North-West - focus on important linguistic questions for the development of advanced speech technology in South Africa.
  - Include questions related to phonetic and phonological systems, the prevalence and influence of dialects, and the modelling of tone.

South African Government

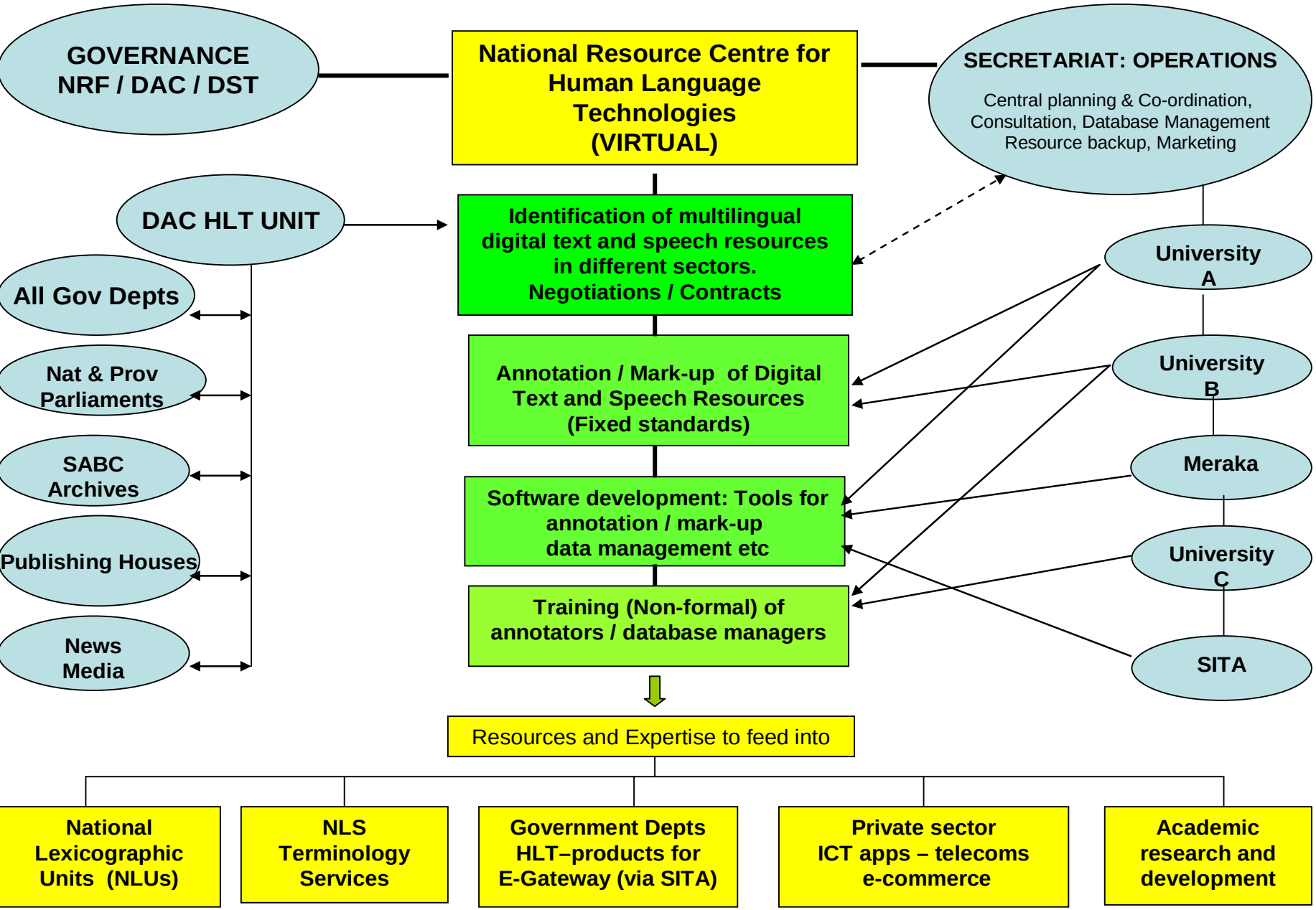
Human Language Technology Unit

Department of Arts and Culture (DAC)

National Language Service [Dr Mbulelo Jokweni]

http://www.dac.gov.za

Responsible for the establishment of a National HLT Resource Centre (in 2009 ?) in co-operation with a National Steering Committee (of role players)


GOVERNANCE
NRF / DAC / DST
National Resource Centre for Human Language Technologies (VIRTUAL)
SECRETARIAT: OPERATIONS
Central planning & Co-ordination, Consultation, Database Management Resource backup, Marketing
DAC HLT UNIT
Identification of multilingual digital text and speech resources in different sectors. Negotiations / Contracts
All Gov Depts
Nat & Prov Parliaments
SABC Archives
Publishing Houses
News Media
Annotation / Mark-up of Digital Text and Speech Resources (Fixed standards)
Software development: Tools for annotation / mark-up data management etc
Training (Non-formal) of annotators / database managers
University A
University B
Meraka
University C
SITA
Resources and Expertise to feed into
National Lexicographic Units (NLUs)
NLS Terminology Services
Government Depts HLT–products for E-Gateway (via SITA)
Private sector ICT apps – telecoms e-commerce
Academic research and development

West Africa

African Language Technology Inititiative (ALT-i)

Nigeria [Dr Tunde Adegbola]

http://www.alt-i.org

- **Corpus Development and Language Modelling** Yoruba
- **Yoruba Keyboard and Word processor**
- **Speech recognition and TTS for Yoruba**
- **Igbo-English Machine Translation**
- **Addressing other Nigerian Languages**

  Active working relationship with the National Institute for Nigerian Languages (NINLAN), in Aba, Nigeria

(East Africa)

Djibouti Centre of Speech Research

[Dr Nimaan Abdillahi] nimaan.abdillahi@gmail.com

- ASR on AFAR language (the second language of Djibouti)

- Automatic translation between Somali-French-Afar

East Africa

Kenya

Teknobyte Speech Technologies

[Dr Mucemi Gakuru] www.teknobyte.co.ke

- Speech synthesis in Kiswahili

  Lead consultant in the development of **NAFIS**, the first ever automated IVR for use by rural farmers in Kenya. NAFIS uses Teknobyte's Kiswahili and Kenyan English voices, and has been developed in collaboration with **Speechnet Ltd** for the Ministry of Agriculture and Ministry of Livestock development

  www.nafis.go.ke

Websites

AfLaT – African Language Technology

http://www.aflat.org/

- Aims to catalogue language resources for the benefit of researchers interested in African language technology
- AfLaT.org contains a steadily growing collection of bibliographic resources, web links and tools, provided by AfLaT members.

National HLT Network (South Africa)

- http://www.meraka.org/nhn
- Facilitates local and international collaboration
- Administered by Meraka

Conclusion

- Exceptional growth in Southern Africa over the last two years
- Steady growth in West and East Africa
- Need for co-operation within the African continent and wider.
  - Topic of Humboldt Alumni Conference on HLT in Africa - Rabat, 3-5 June 2008