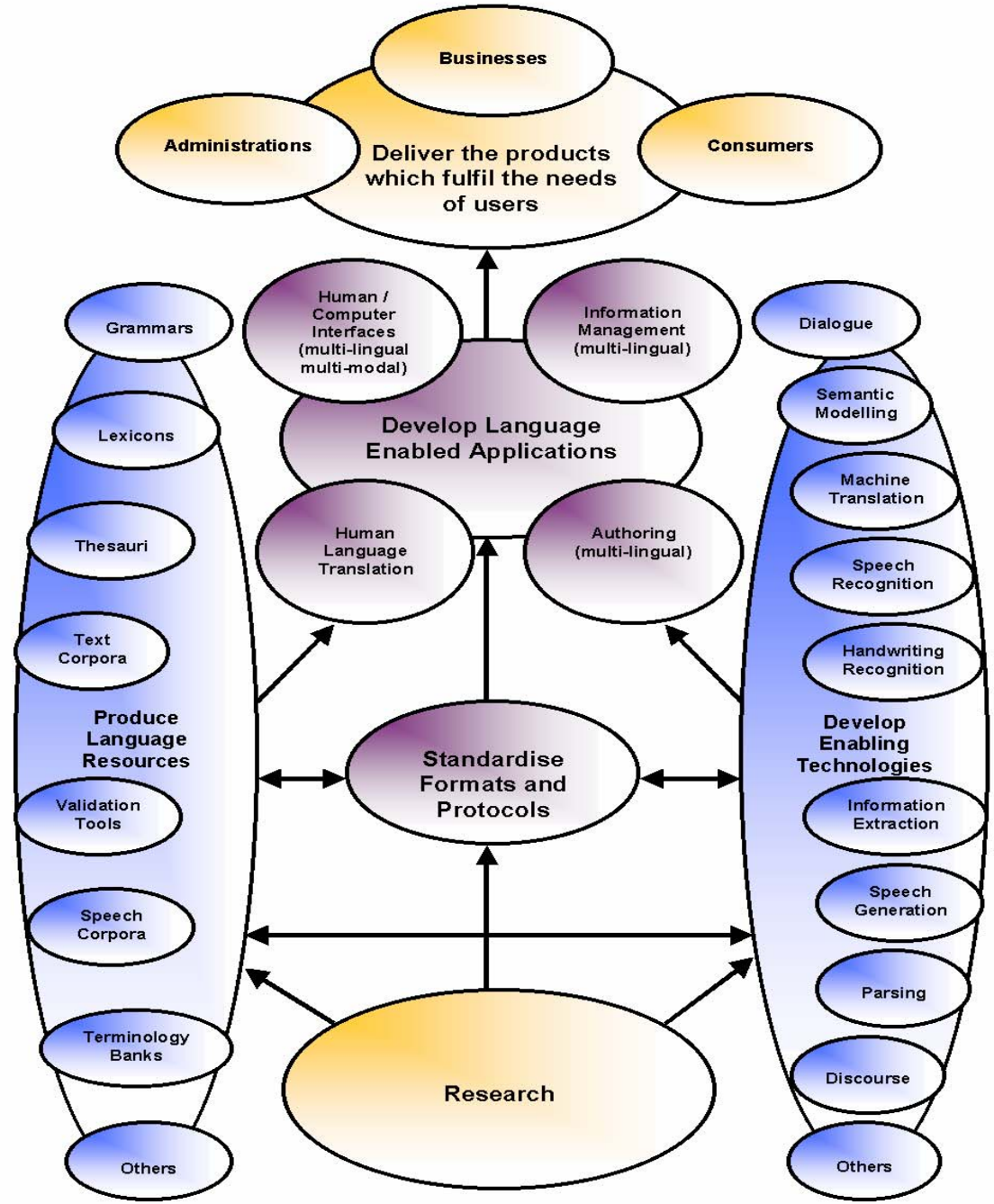
# Networking language resources in Africa: Future plans and proposals

Justus Roux
Stellenbosch University
Centre for Language and Speech Technology

**Aim**

- **The nature and use of language resources**
- **International resource centres**
- **National resource facilities**
- **DAC initiative:  HLT Resource Centre**
- **Resource networks in Africa**

# GENERAL DEVELOPMENT MODEL FOR HUMAN LANGUAGE TECHNOLOGIES




Model of Language Engineering Activities

## Language resources are imperative prerequisites for technology development and applications

- **Language resources:**
  –Text, Speech, Multimodal / Multimedia, Tools

- **Distinguish between**
  – ‘Traditional’ language resources >> support human users in creating and processing text

  – Digital language resources >> support development of technologies that will enable automated processing of text or that will facilitate, *inter alia*, human-machine interaction through speech

## Digital Language Resource (Defined)

"… a set of speech or language data and descriptions in machine readable form, used e.g.

- for building, improving or evaluating natural language and speech algorithms or systems, or,

- as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject area specialists and endusers."

(ELRA, www.elra.info )

## Digital Language Resources (Text)

- Lexica: representing lexical knowledge, machine-understandable dictionaries, word networks, etc.

- Corpora: representing examples of language usage including general language and sub-languages, cf. weather forecasts, medical reports, technical manuals etc,

- Terminologies: representing specialised vocabularies, standardised terminological databases, nomenclatures, ontologies, etc

# Digital Language Resources (Speech)

Speech corpora, pronunciation lexica: mono- & multi-lingual

- Speech varieties related to dialect, accent, age, gender, transmission lines (microphone, telephone), environment (quiet, noisy), pathological speech

needed for

automatic speech recognition, speaker identification, speaker verification, language identification, dialect identification, speaker adaptation, speech synthesis, etc

in applications such as

computer assisted language learning, access control, dialogue systems, information access, clinical intervention systems, etc

**Digital Language Resources (Multimodal / Multimedia)**

Spoken dialogue with gestures (audio & video) / static gesture images (graphic images)

used for

complete and/or robust speech recognition, pragmatic analysis, semantic analysis, speaker verification

applications in

applications in computer assisted language learning, sign language transformation, etc

# Digital Language Resources (Tools)

Tools: representing software modules that are used in conjunction with other resources for their acquisition, analysis, management, integration, employment etc

Typical tools: morphological and syntactic analysers, taggers, lemmatisers, chunkers, grapheme-phoneme converters, automatic phonetic transcribers, automatic phonetic segmentation, automatic phoneme aligners, etc

## Key priorities to consider in developing Digital Language Resources (DLRs)

- DLRs need to fit an open and standards based framework

- DLRs need to be reusable, of large scale, and multi-layered

- DLRs need to be dynamic and sustainable – it is a continuous process involving resources that need to be maintained and updated taking into account developments in data storage technologies.

# International Resource Facilities

- European Language Resource Association (ELRA), Paris (www.elra.org)

- Linguistic Data Consortium (LDC) in Pennsylvania in the USA

- European Network of Excellence in HLT (ELSNET), Utrecht, the Netherlands,

- Network for Euro-Mediterranean Language Resources (NEMLAR), Cairo, Egypt.

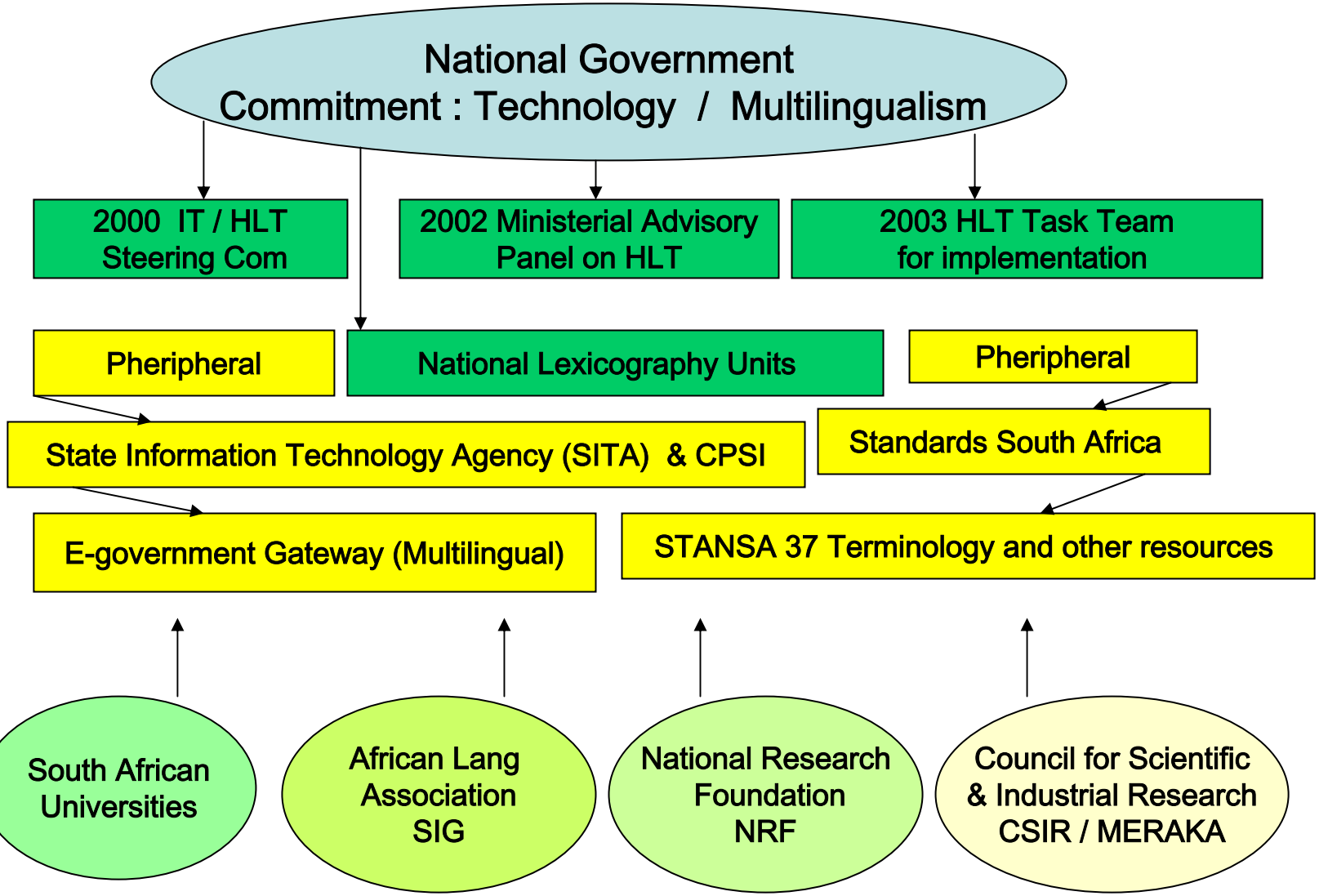
## International Resource Facilities (2)

- France: nine major projects dedicated to the creation of DLRs
- Japan & Hong Kong
- USA: fourteen centres for National Language Resources established
- National governments are highly involved in the establishment of DLR centres.

# National resource facilities

- National Lexicographic Units  (PanSALB)
- National Language Service (DAC)
  - –Terminology development
- Tertiary Institutions of Higher Learning
- Language Development Centres (?)

# Top – Down Initiatives



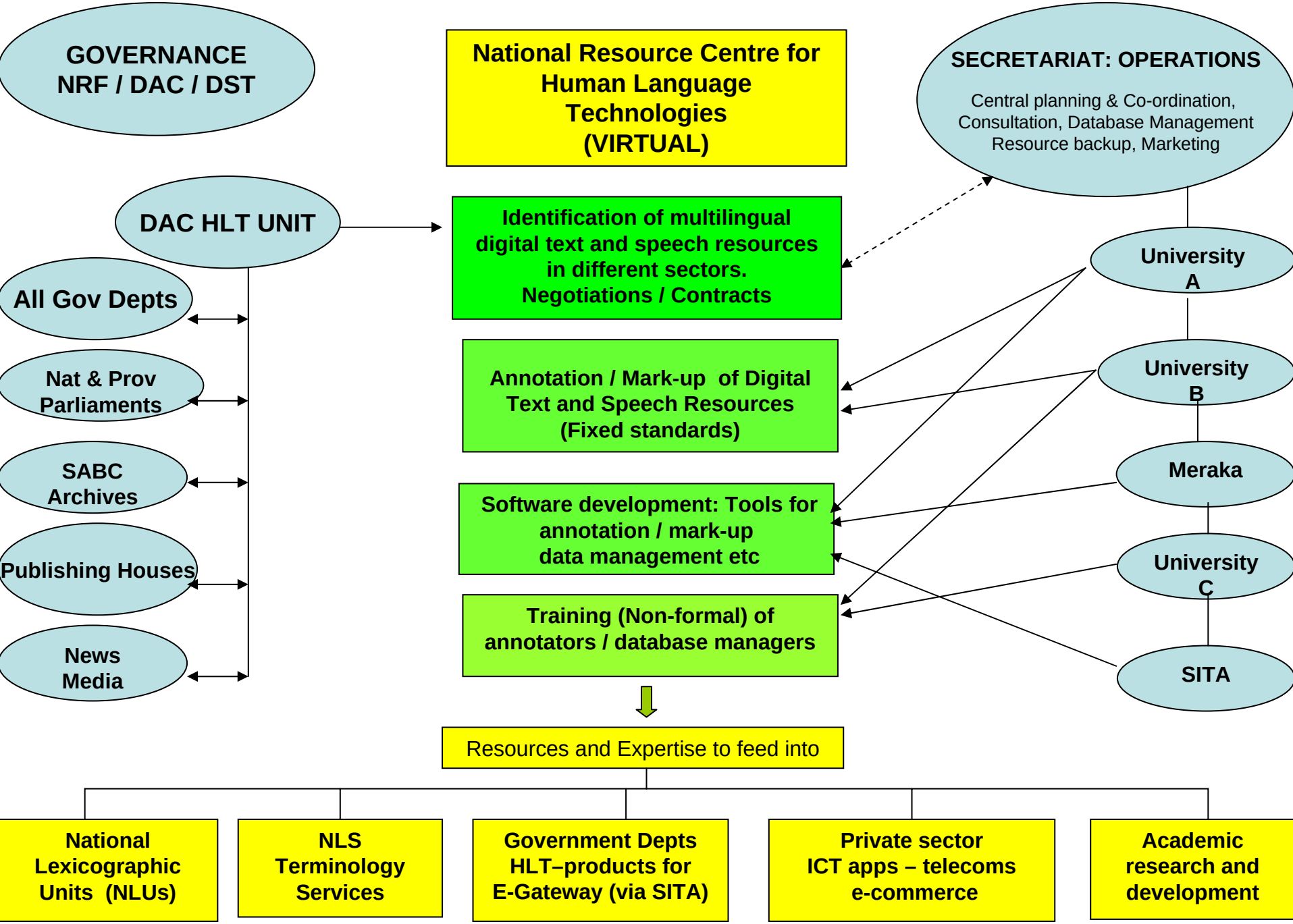

# Bottom – Up Initiatives

# POTENTIAL MODEL

# FOR

# RESOURCE CENTRE IN HUMAN LANGUAGE TECHNOLOGIES IN SOUTH AFRICA

- Central co-ordination hub
- Interlinking development nodes at expert centres
- Sharing model


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

# Networking Language Resources in Africa

African Language Association of Southern Africa – Special Interest Group for Language and Speech Technology (ALASA-SIG)

Pre-conference workshop at LREC 2006 in Genoa, Italy (May 2006)

**Why?**

- Africa is part of global village (ICT / HLT = global)
- Processing of African Languages with similar characteristics
- Determine the status of language resources in East, West, Central and Southern Africa
- Develop standards within Africa (ISO)
- Joint development of tools / applications
- ..? ?

- **Structure?**
  - Physical Network (web)
  - Academic Network
- **Activities?**
  - Workshops
  - Electronic bulletin board
- **Funding?**
  - Nepad?
  - World Bank?
  - ISO?
  - International Speech Communication Association (ISCA)?
- **Suggestions?**