The Development of Language and Speech Resources: The South African Experience

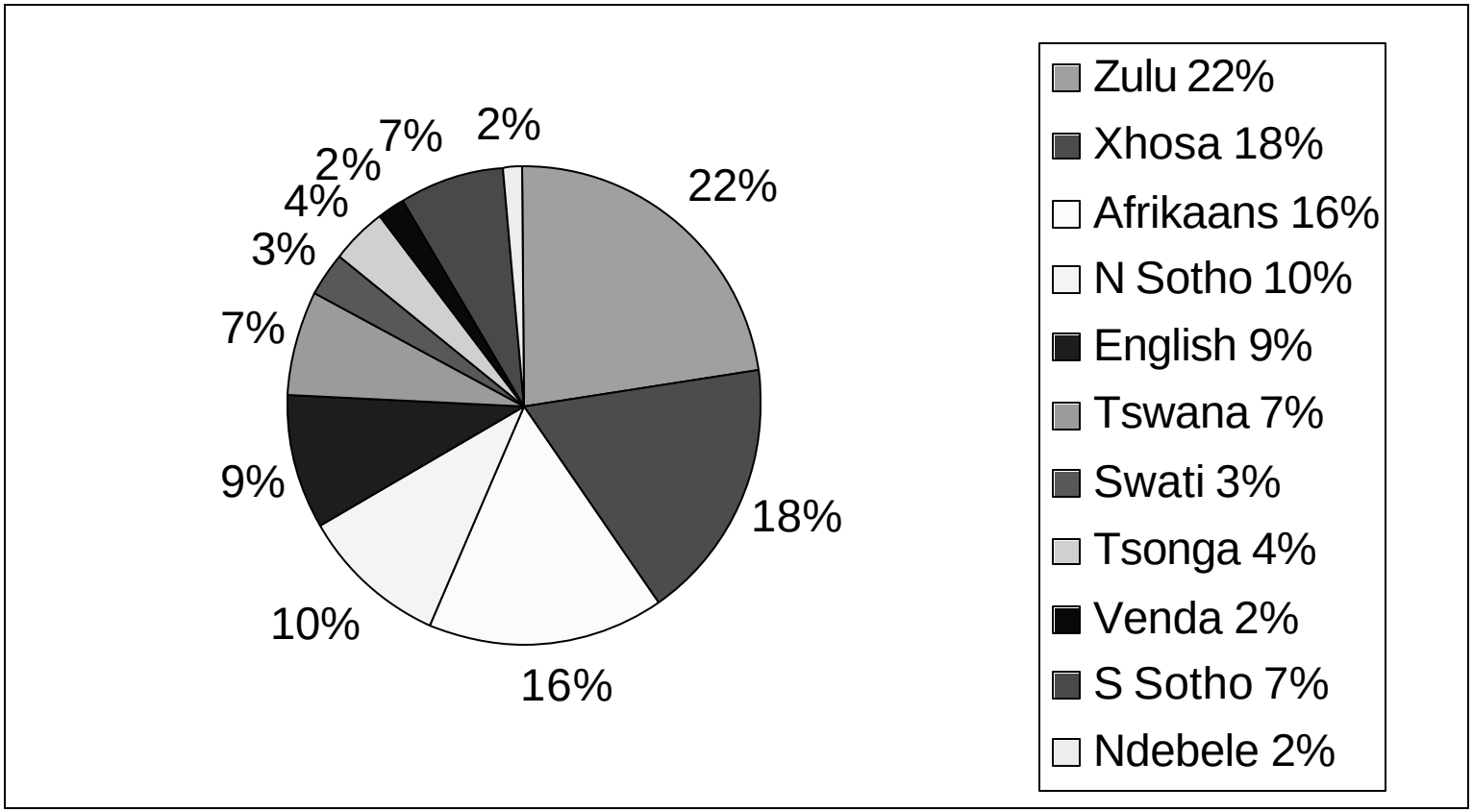
Justus Roux
Research Unit for Experimental Phonology
University of Stellenbosch
South Africa
jcr@sun.ac.za

Aims

- Description of the language scenario in South Africa and potential for HLT development
- ‘Top-down’ activities in developing language resources
- ‘Peripheral’ activities in developing language resources
- ‘Bottom-up’ activities in developing language resources
- Concluding remarks

Language Situation

Mother tongue division (n=40,5 mil speakers)



The Language Situation (2)

- Eleven official languages
  - Politically correct
  - Administrative ‘nightmare’
  - Extremely costly

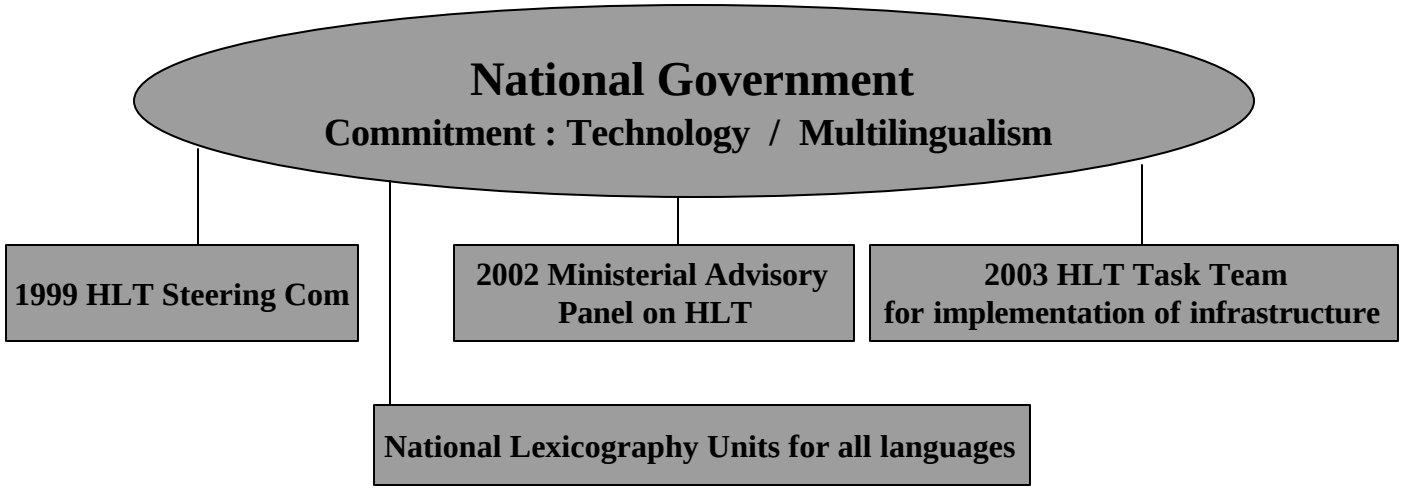
  BUT

  - Enormous challenges for HLT development

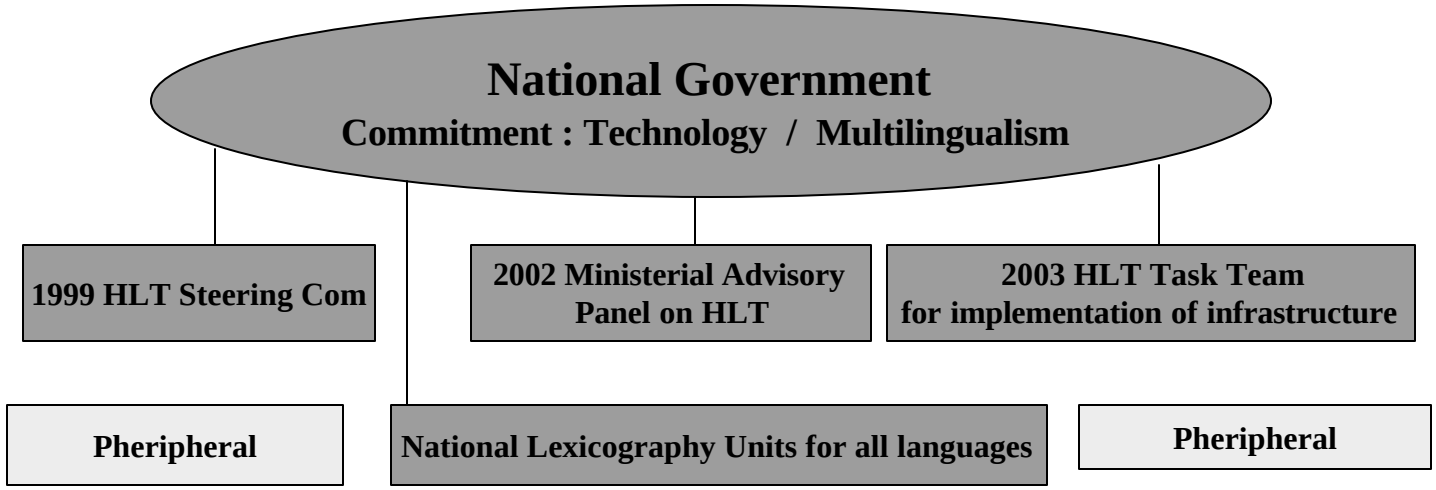
Language situation (3)

- Growth potential in delivery platforms (1995 - 2000)
  - PC's /1,000 people:
    - SA 27.9 > 61.8 (Sub-Saharan Africa 9.2)
  - Fixed line telephones / 1,000 people
    - SA 114 (Sub Saharan Africa 14)
  - Mobile telephone subscribers (1998 – 2002)
    - SA 2.55 million > 12.5 million
- Challenge for voice based systems to reach illiterate population – need for text and speech resources

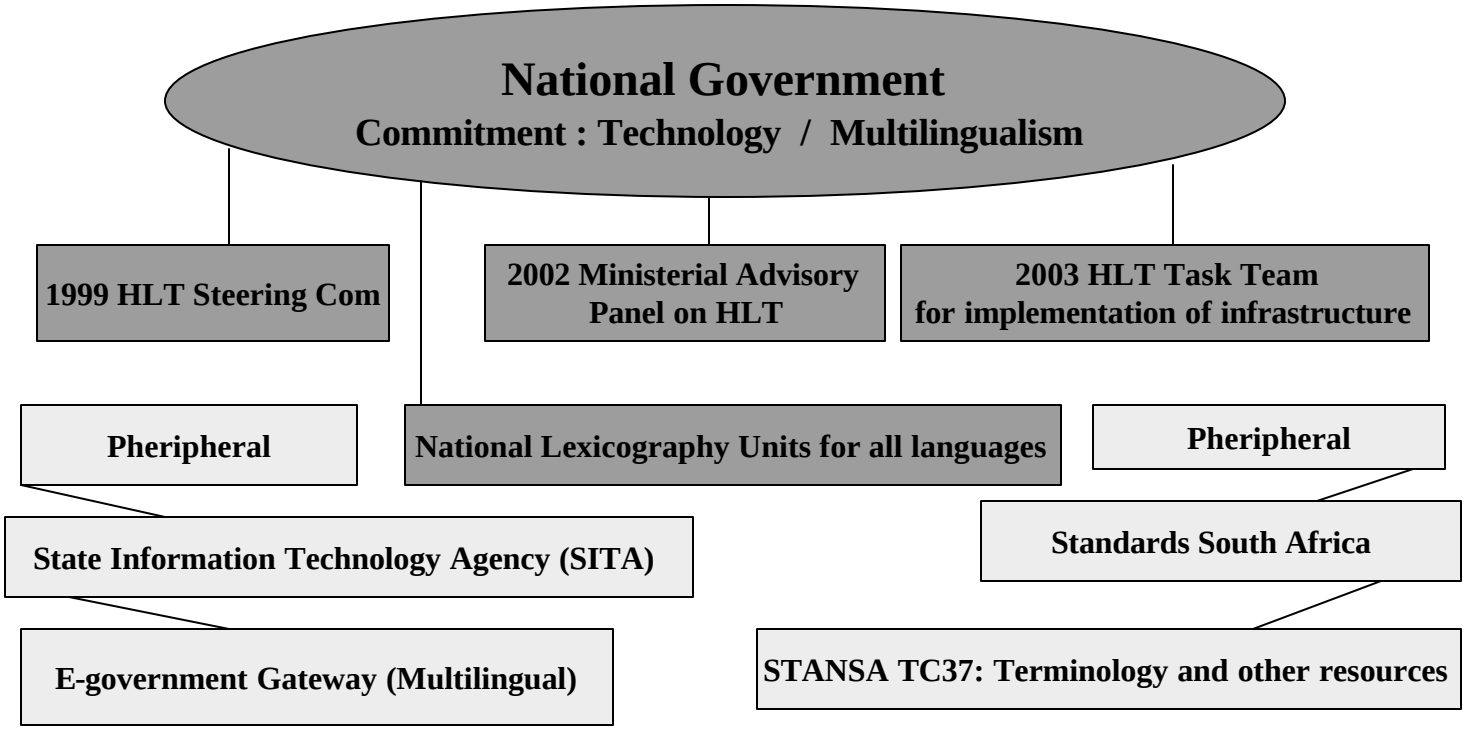
Top – Down Initiatives



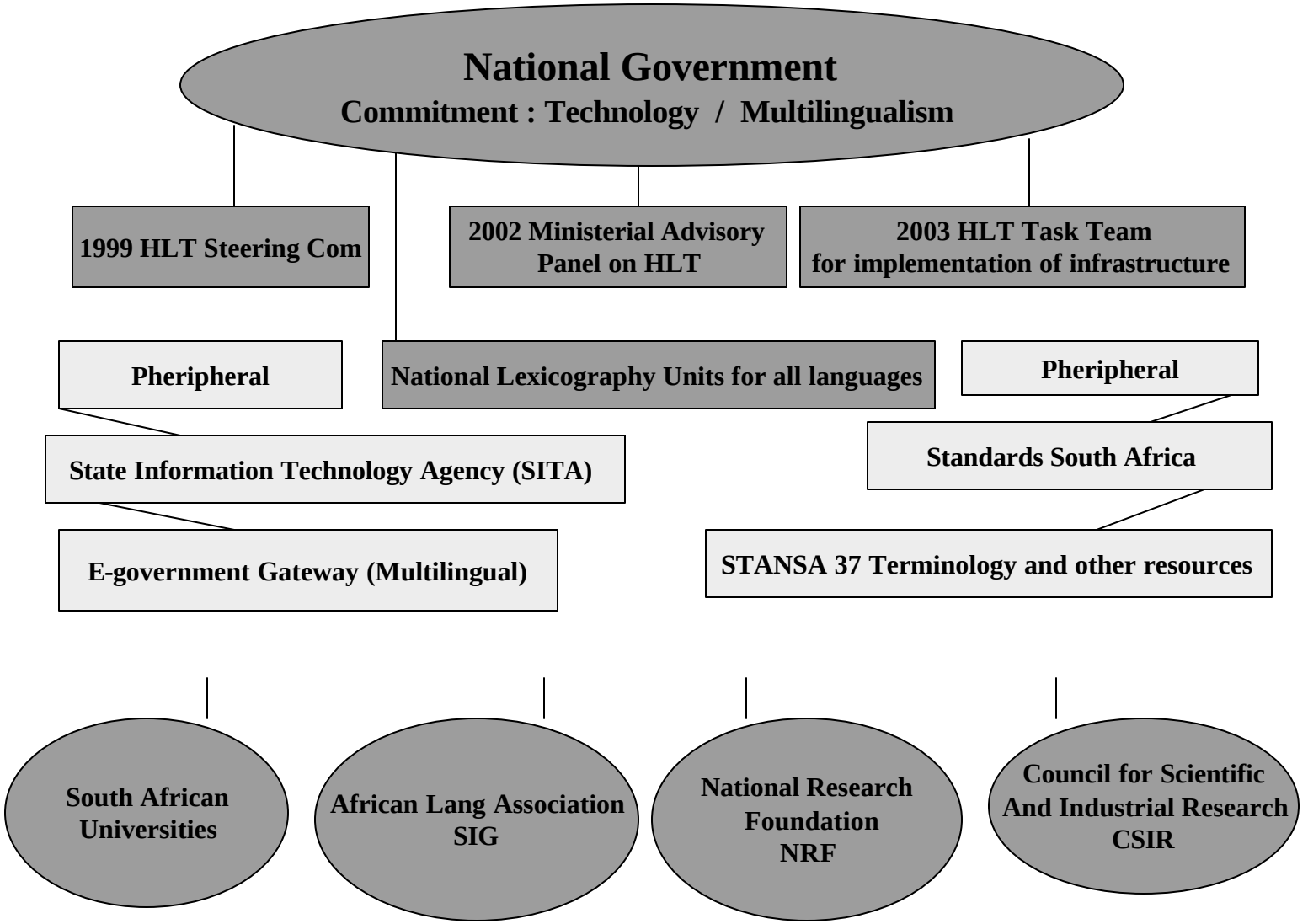
Top – Down Initiatives



Top – Down Initiatives



Top – Down Initiatives



Bottom – Up Initiatives

Main initiatives in South Africa on HLT

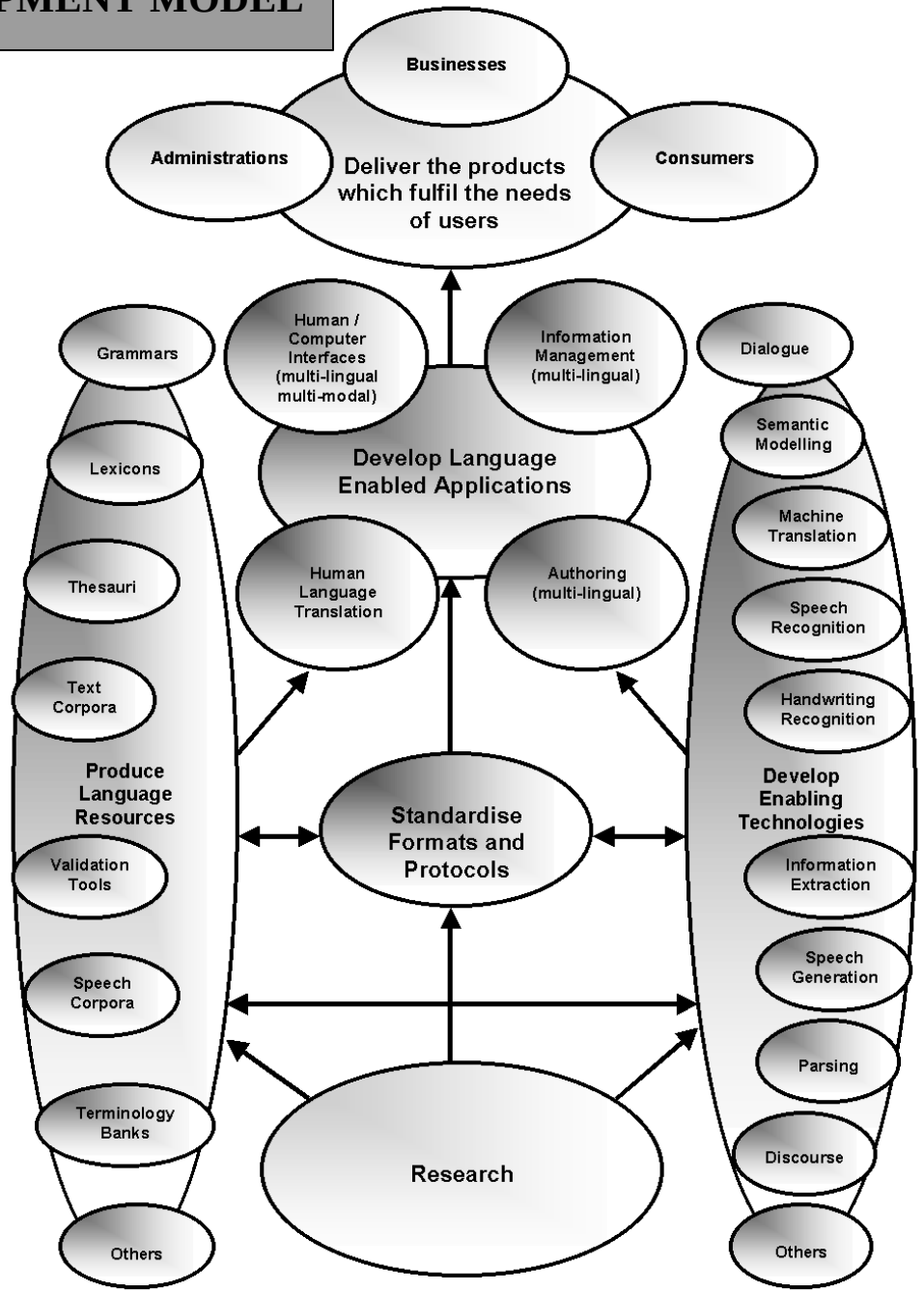
- **University of Stellenbosch**
  - African Speech Technology project (2000 – 2003)
  - Interactive telephone based information retrieval and booking system:
    - Annotated speech databases for SA English (five varieties), Afrikaans (two varieties), Zulu, Xhosa and Sesotho in SpeechDat format - approx 3 300 speakers in total
    - Lexicons with phonetic variations derived from above
    - Mixture of fixed line and mobile telephone data
    - Validation tools
    - Software toolkit for (new) application design
    - Awareness campaigns / Interim capacity building (ALASA-SIG)

- University of Pretoria
  - African language spellchecker project
    - Extensive text corpora for nine African languages

- University of South Africa (UNISA)
  - Morphological parsing: African Languages project
    - Finite state tools (Xerox) for Xhosa, Zulu, Ndebele and Northern Sotho

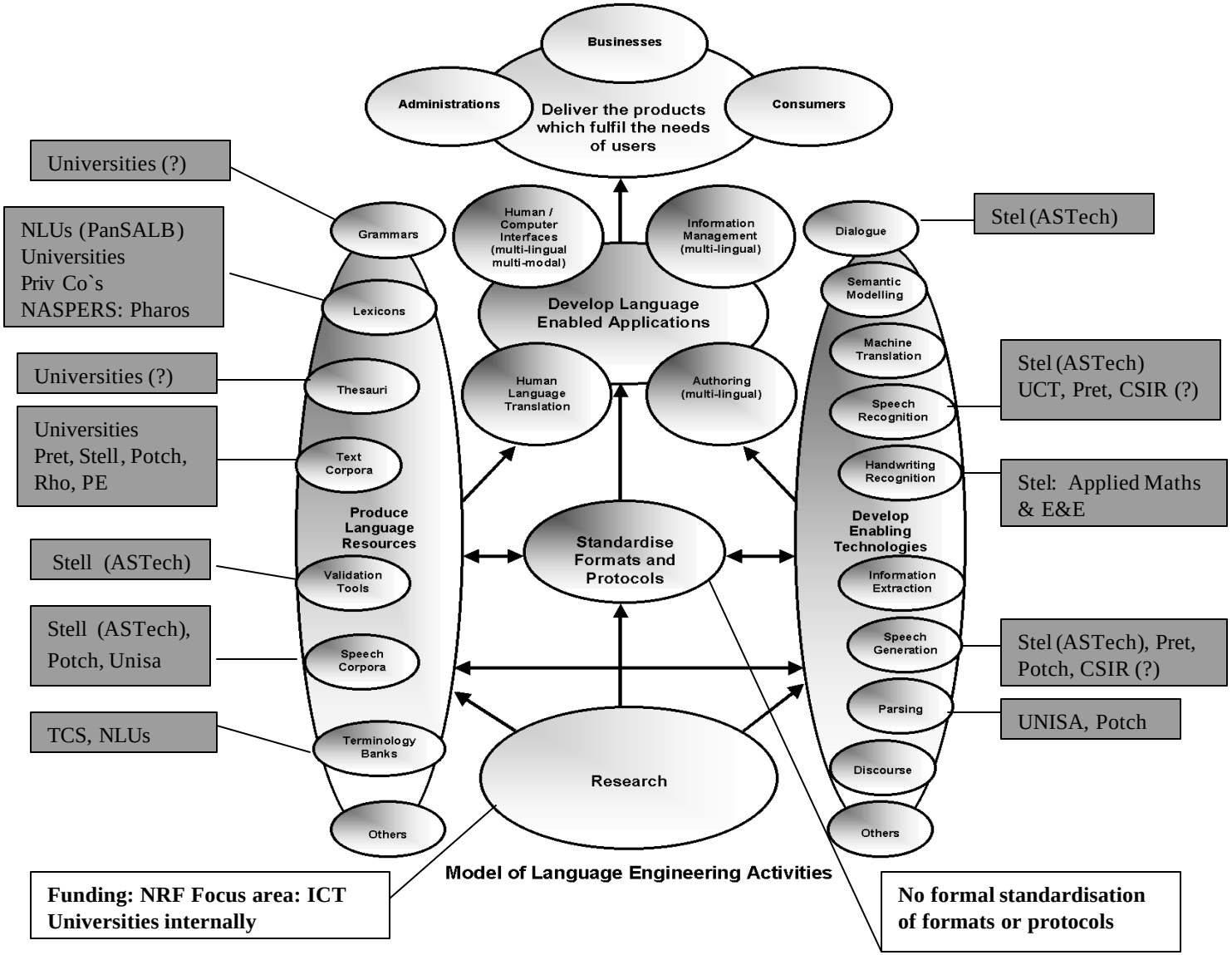
- University of Potchefstroom
  - Afrikaans text corpora for spellchecker project
  - Spoken Afrikaans corpus
    (+ University of Gent, Belgium)
  - SA English Corpus (text)
  - Afrikaans diphone database
  - Tools: POS taggers, lemmatisers, speech analysis tools

GENERAL HLT DEVELOPMENT MODEL



Model of Language Engineering Activities

HLT R&D: SOUTH AFRICAN SCENARIO



HLT Research Agenda

- Special Workshop to determine a research roadmap (18 August 2003 Pretoria)
- ELSNET co-operation (Steven Krauwer)
- Outcomes:
  - Too early due to
    - Limited text and speech resources
    - Limited human capacity
  - Need for a political champion to drive process
  - Need for sustained formal and non formal training

The Road Ahead

- Continued drive for development of language resources (text and speech) in South African context
- Embark on innovative training programmes
- Stay in close contact with international organisations and role players