NHN The National HLT Network

An HLT profile of the official South African languages

Aditi Sharma Grover[1,2], Gerhard B van Huyssteen[1,3] & Marthinus W. Pretorius[2]

[1]HLT Research Group, CSIR, South Africa

[2]Graduate School of Technology Management, University of Pretoria, South Africa

[3]Centre for Text Technology (CTexT), North-West University, South Africa

Overview

- Background
- Process
- Results
- Conclusion

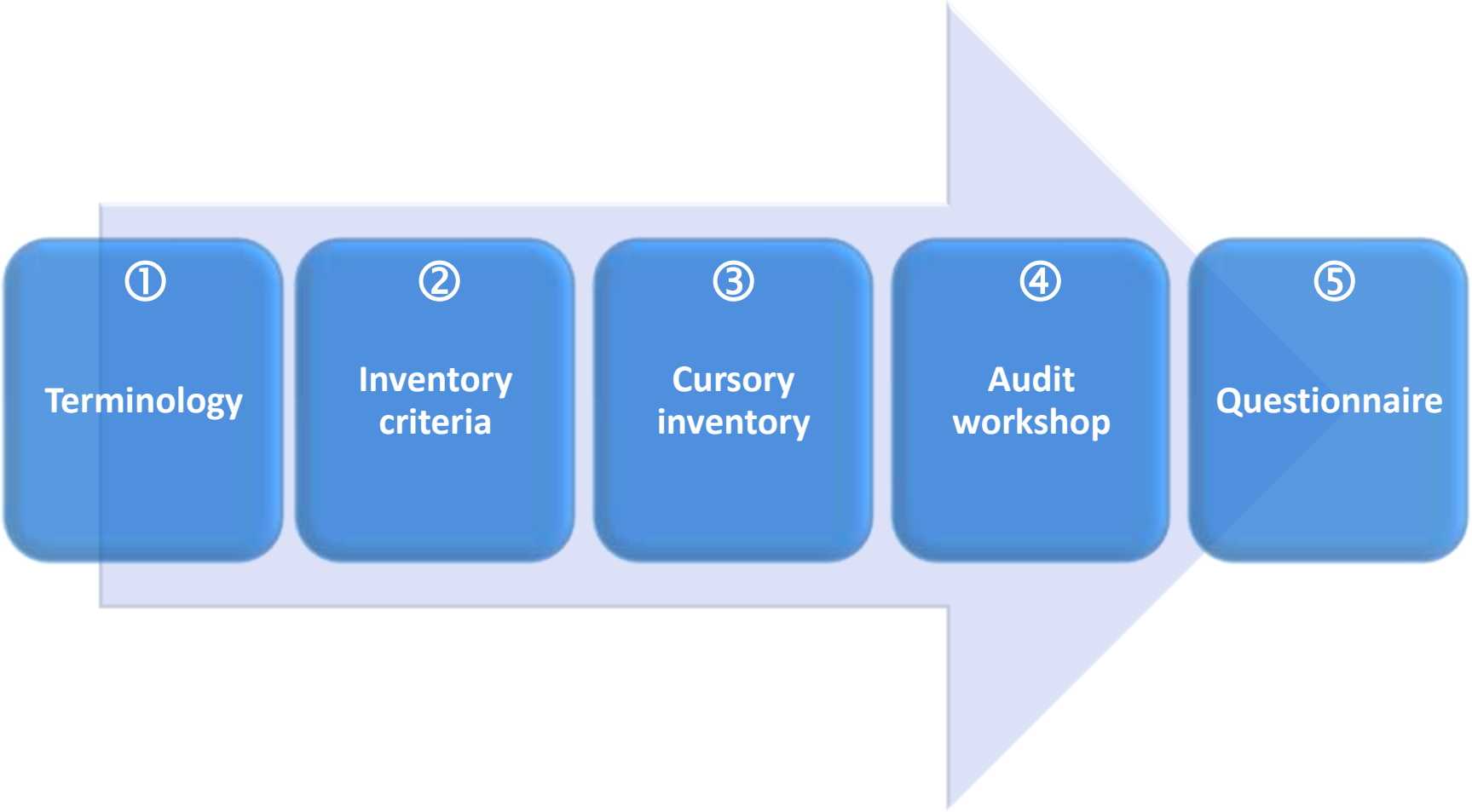
South African HLT landscape

- 11 official languages
- HLT community
  - R&D community (universities & science councils)
  - Very few private sector companies
- Various government initiatives
  - DST: HLT road-mapping process, NHN
  - DAC: HLT strategy, National Centre for HLT
  - NRF: research funding

Challenge

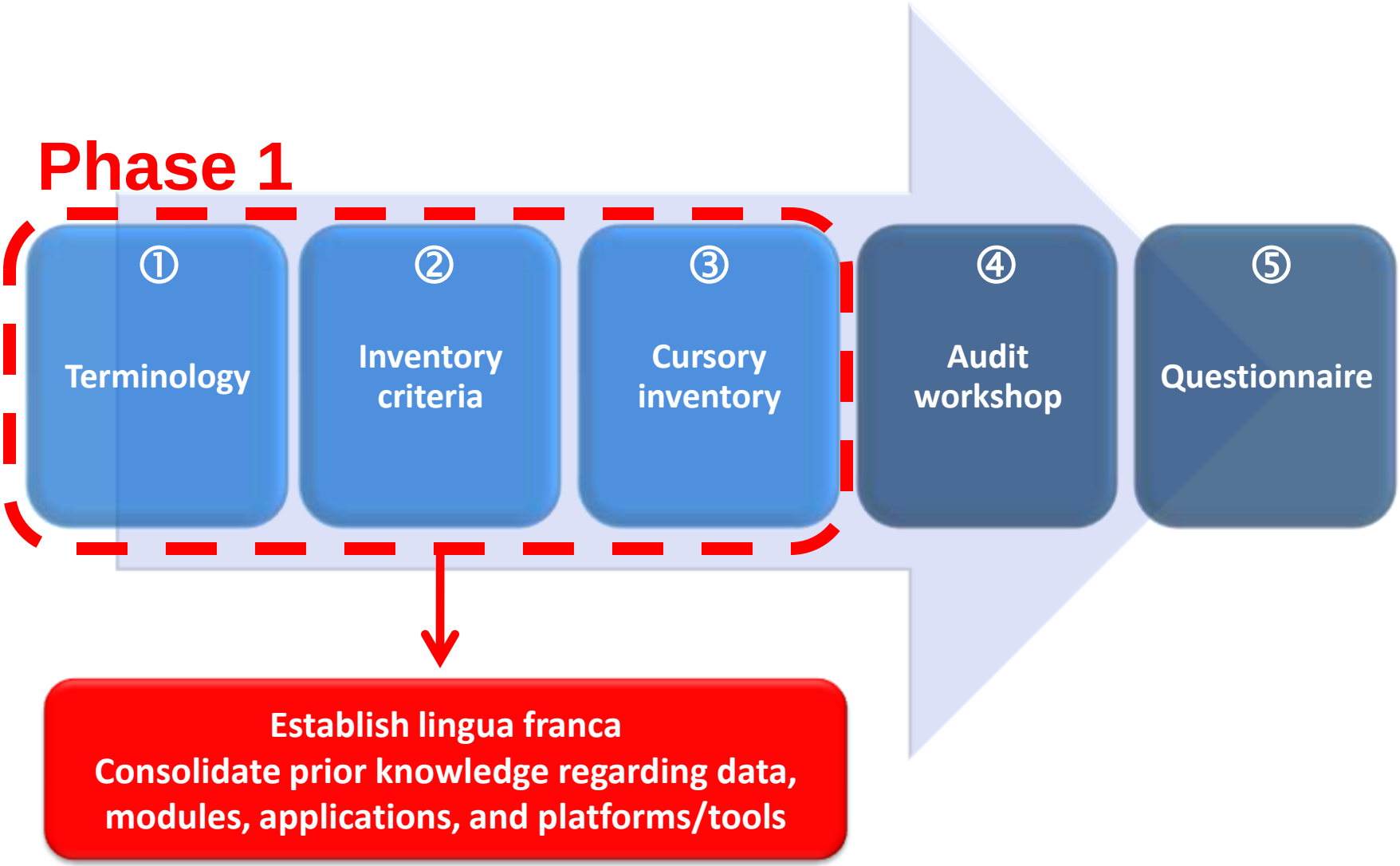
- SA has not yet capitalised on opportunities to create a thriving HLT industry
  - Lack of awareness within the local HLT community
- Perpetuated by perceived fragmentation of South African R&D activities
  - Lack of a unified technological profile of HLT activities across the 11 languages
- 2009: a technology audit for the South African HLT landscape (SAHLTA)
  - Align R&D activities and stimulate cooperation
  - Similar to Dutch (BLaRK), EuroMap

SAHLTA Process



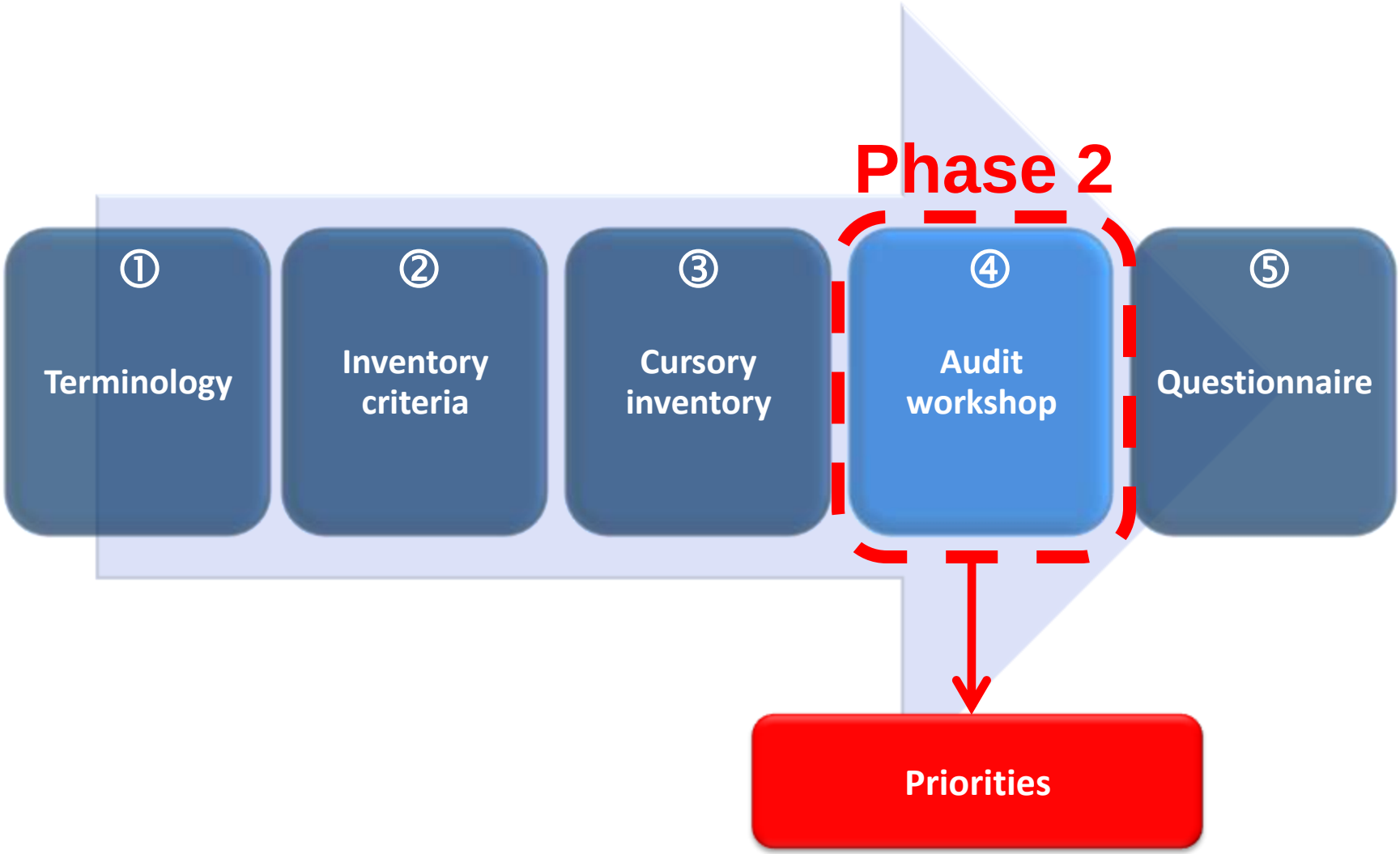
NHN The National HLT Network

SAHLTA Process



NHN The National HLT Network

SAHLTA Process



Prioritisation

- Based on international trends, local needs, and feasibility
- **Priority 1:** Basic & robust core HLT technology applications, modules and data
- **Priority 2, 3:** LRs that further enhance and complement core LRs (priority 1), and base their development on a strong foundation of core HLT LRs
  - Many advanced HLT applications are priority 2, 3
- Verification by larger SA HLT community
  - Need to be updated regularly

NHN The National HLT Network

Priority 1: Applications

Text

- Proofing tools
- Information Extraction
- Information Retrieval
- Human-aided machine translation
- Machine-aided human translation

Speech

- Accessibility
- Telephony applications
- Computer-assisted language learning
- Voice search
- Audio management

Priority 2: Applications

Text

- OCR/ICR
- Multilingual comprehension assistants
- CALL
- Authorship identification

Speech

- Access control
- Embedded speech recognition
- Speaking devices
- Computer-assisted training

Priority 3: Applications

Text

- Text generation
- Document classification
- Summarisation
- QA
- Dialogue systems
- Reference works

Speech

- Transcription/dictation
- Multimodal information access
- Command&Control
- Announcement systems
- Audio books
- S2S translation

NHN The National HLT Network

Priority 1: Modules

Text

- G2P
- Text pre-processing
- Normalisation
- Morphological analysis
- POS tagging
- Chunking
- WSD
- Language/dialect ID

Speech

- Complete ASR
- Non-native ASR
- Complete TTS
- Confidence measures
- Speaker ID
- Diarisation
- Language ID

Priority 1: Data

Text

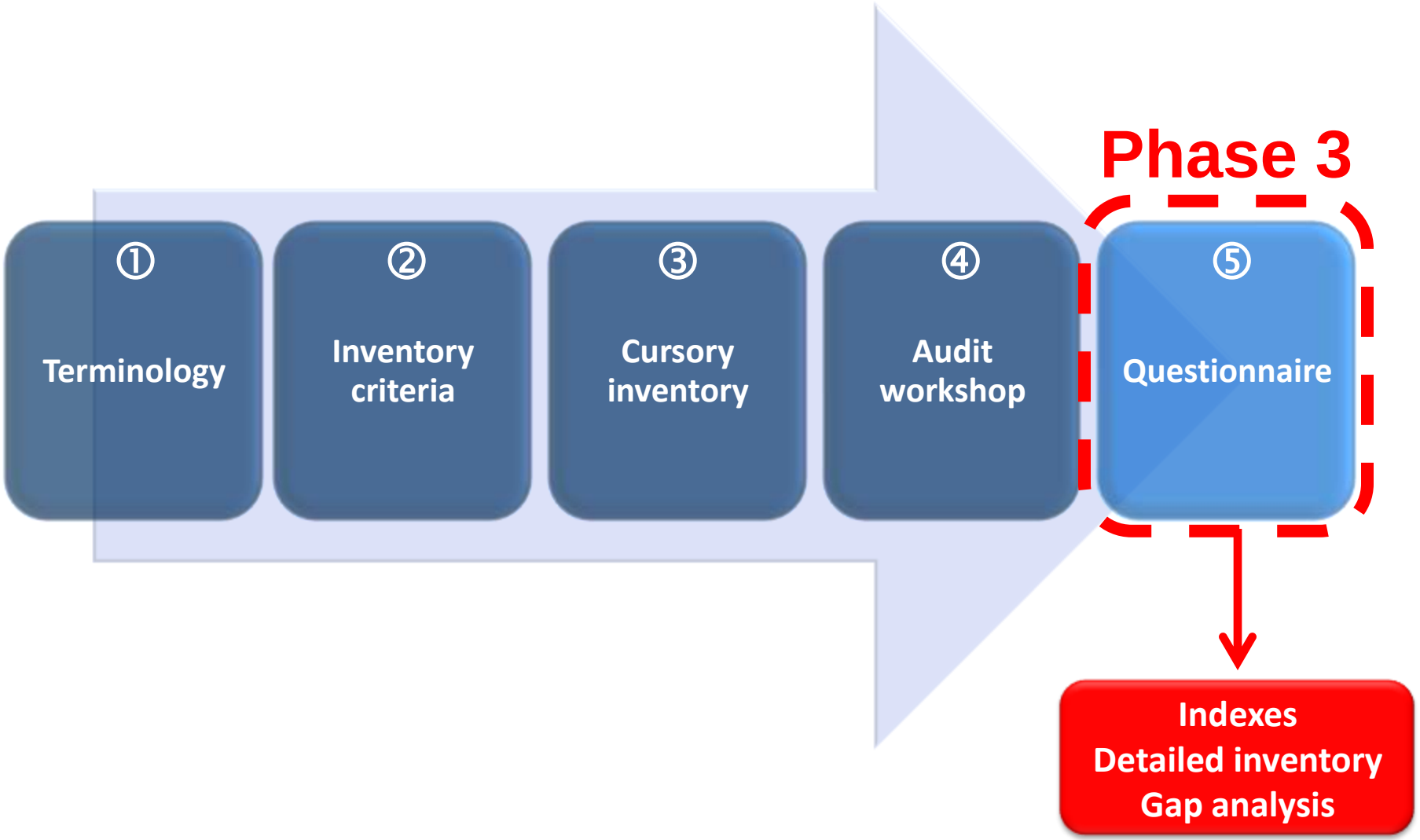
- Monolingual corpora
- Multilingual corpora
- Test suites and corpora
- Lexica (incl. named-entity lists)
- Domain-/Application-specific corpora

Speech

- Annotated monolingual corpora
- Domain-/Application-specific corpora
- Test suites and corpora
- Pronunciation resources (e.g. Phone sets, dictionaries, etc.)

NHN The National HLT Network

SAHLTA Process



Response rate

Participant	Contacted	Response received
Primary		
Universities	7	6
National science councils & independent research centres	2	2
Private companies	6	4
Total	**15**	**12**
Secondary		
National lexicography units	11	3
Government departments	1	1
Total	**12**	**4**

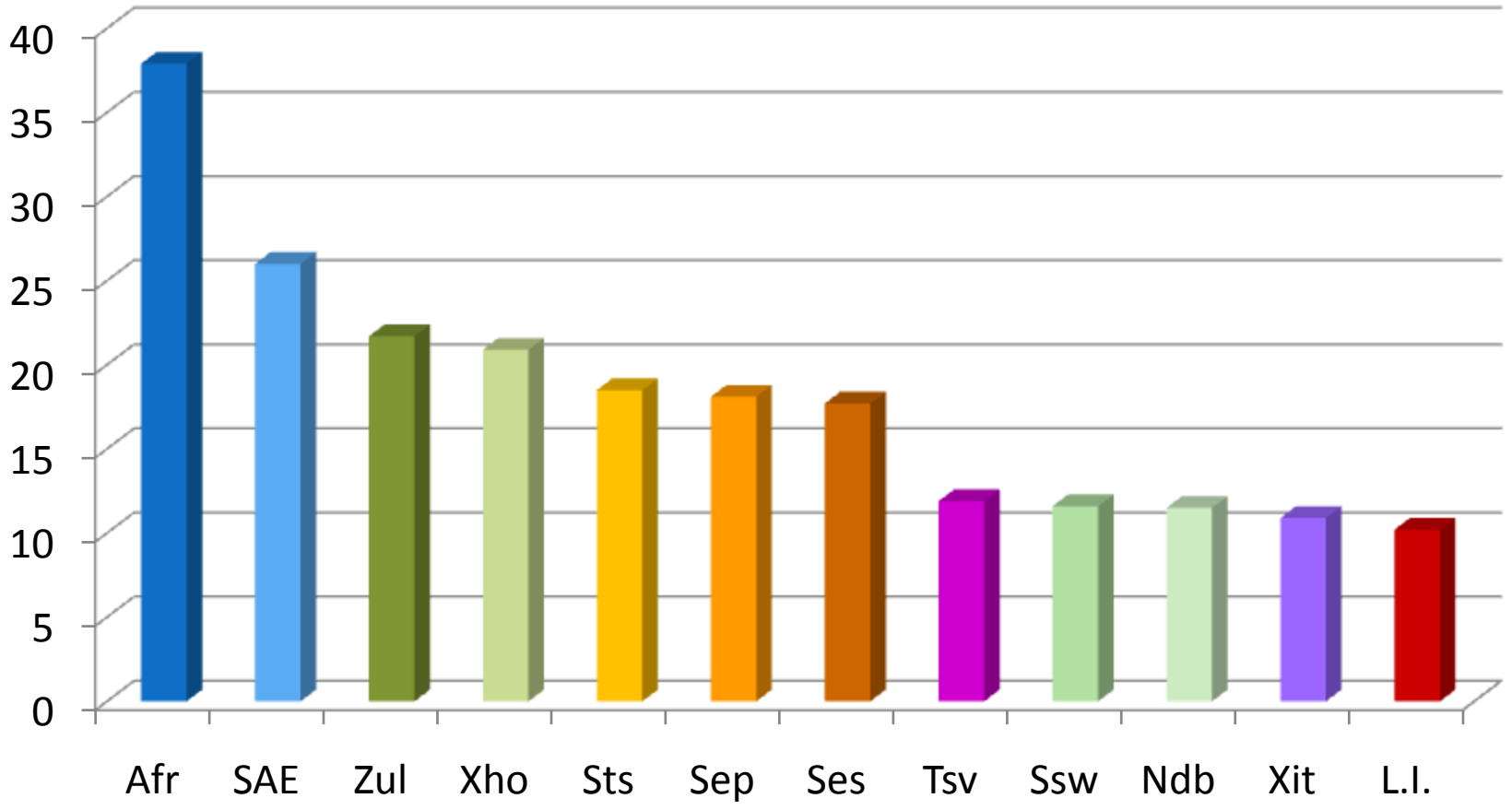
Maturity Index

- Maturity stages:
  - Under development (**UD**), Alpha version (**AV**), Beta version (**BV**) , Released (**RV**)
- Maturity Index
  - Measure of the maturity of HLT components in a language.
  - Considers the maturity stage of item against the relative importance of each maturity stage
  - MaturityInd = Σ (*1.UD+2.AV+4.BV+8.RV)/*

    Σ *Weights of maturity stages*

NHN The National HLT Network

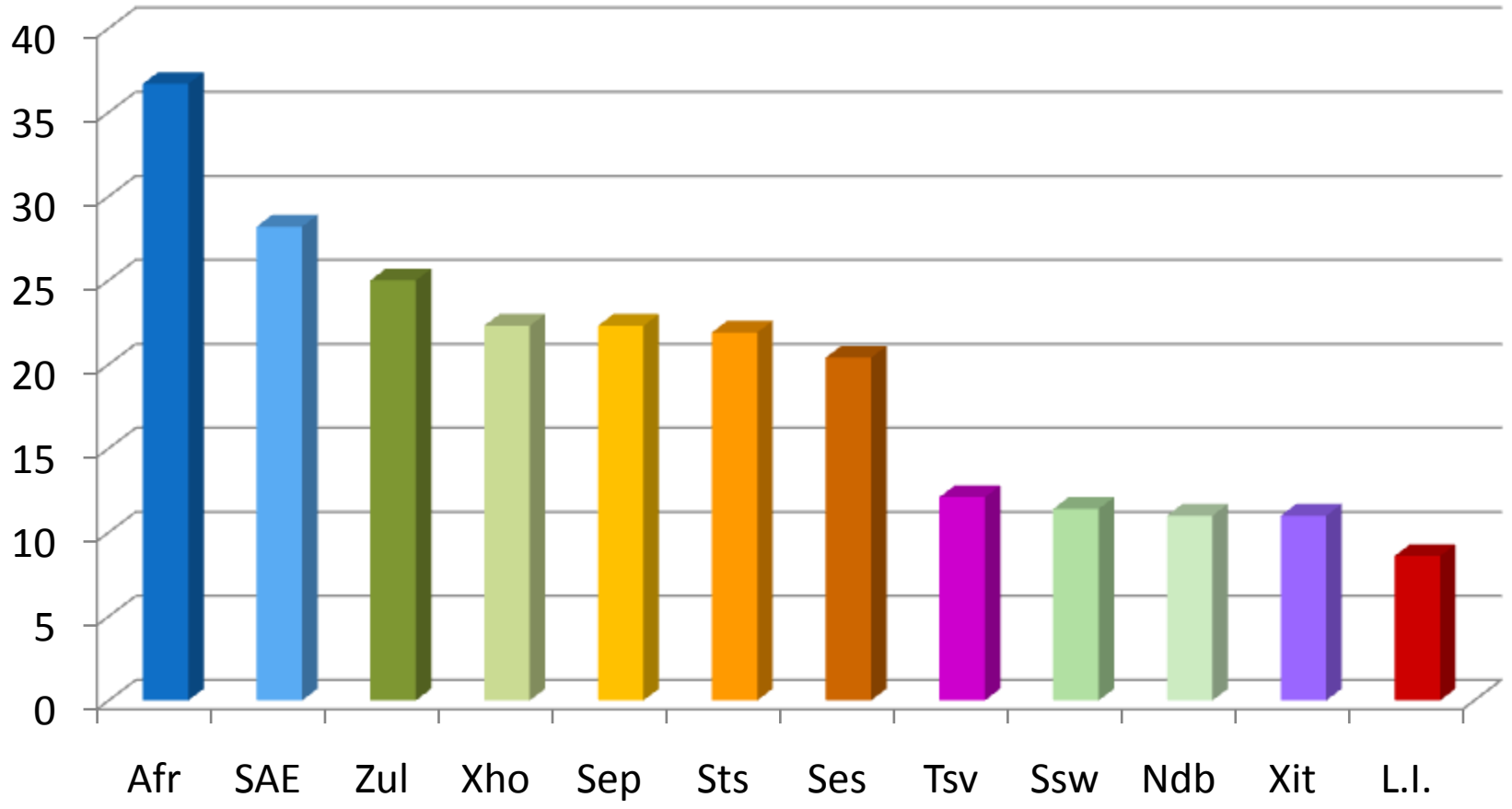
Maturity Index



Accessibility Index

- Accessibility stages:
  - Unspecified (**UN**), Not available (**NA**) (proprietary or contract R&D), Research and education (**RE**), Available for commercial purposes (**CO**), Available for commercial purposes and R&E (**CRE**)
- Accessibility Index
  - Measure of the accessibility of HLT components in a language
  - Considers the accessibility stage of an item against the relative importance of each accessibility stage
  - AccessInd = Σ (*1.UN+2.NA+4.RE+8.CO+12.CRE)*/ Σ *Weights of accessibility stages*

Accessibility Index



HLT Language Index

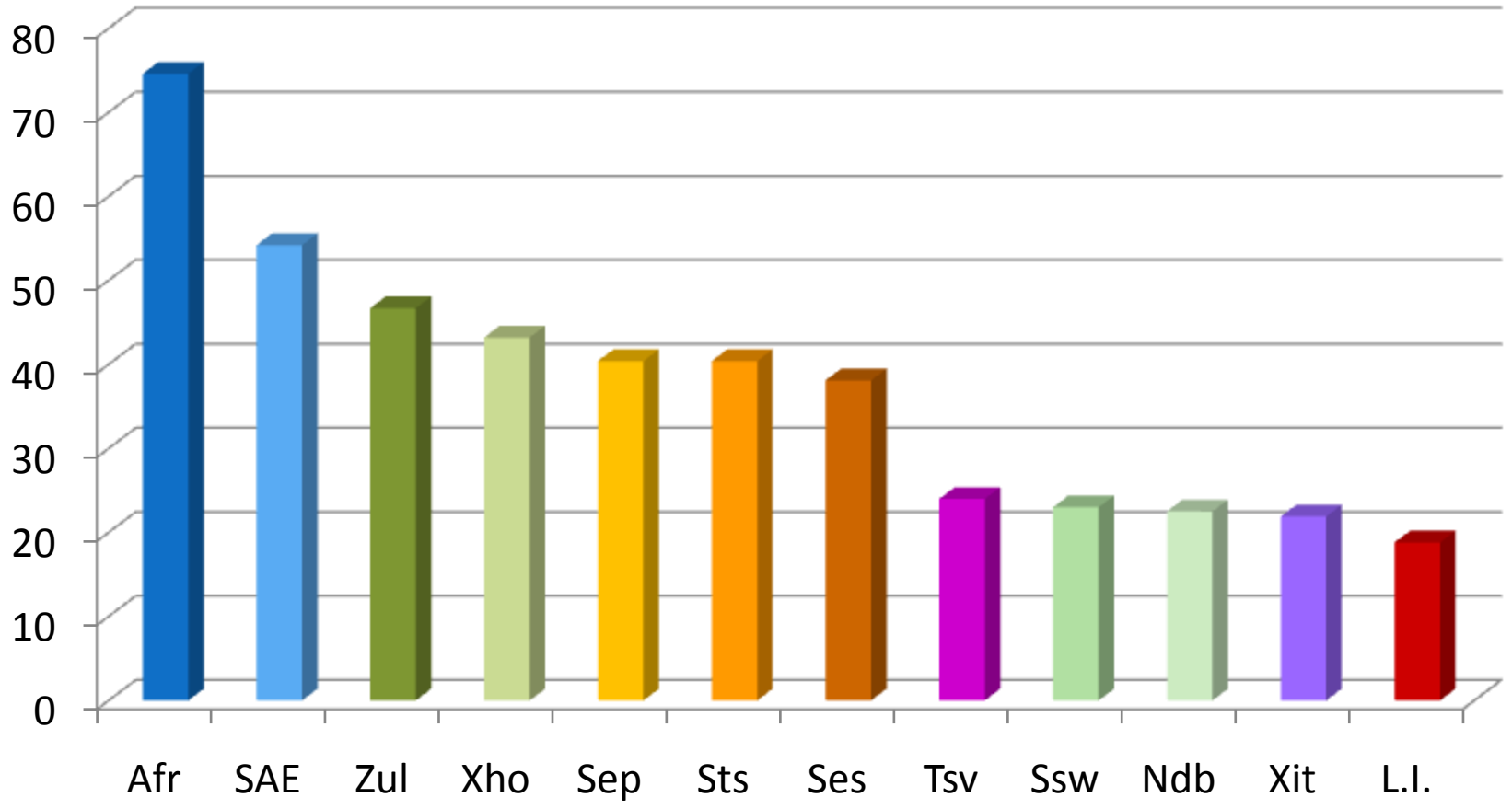
- Impressionistic index that relatively ranks languages based on the total quantity of HLT activity per language
- Considers the stage of maturity and accessibility of all the HLT components
- HLT Language Index = Maturity Index
  (per language, all components)

  \+

  Accessibility Index

NHN The National HLT Network

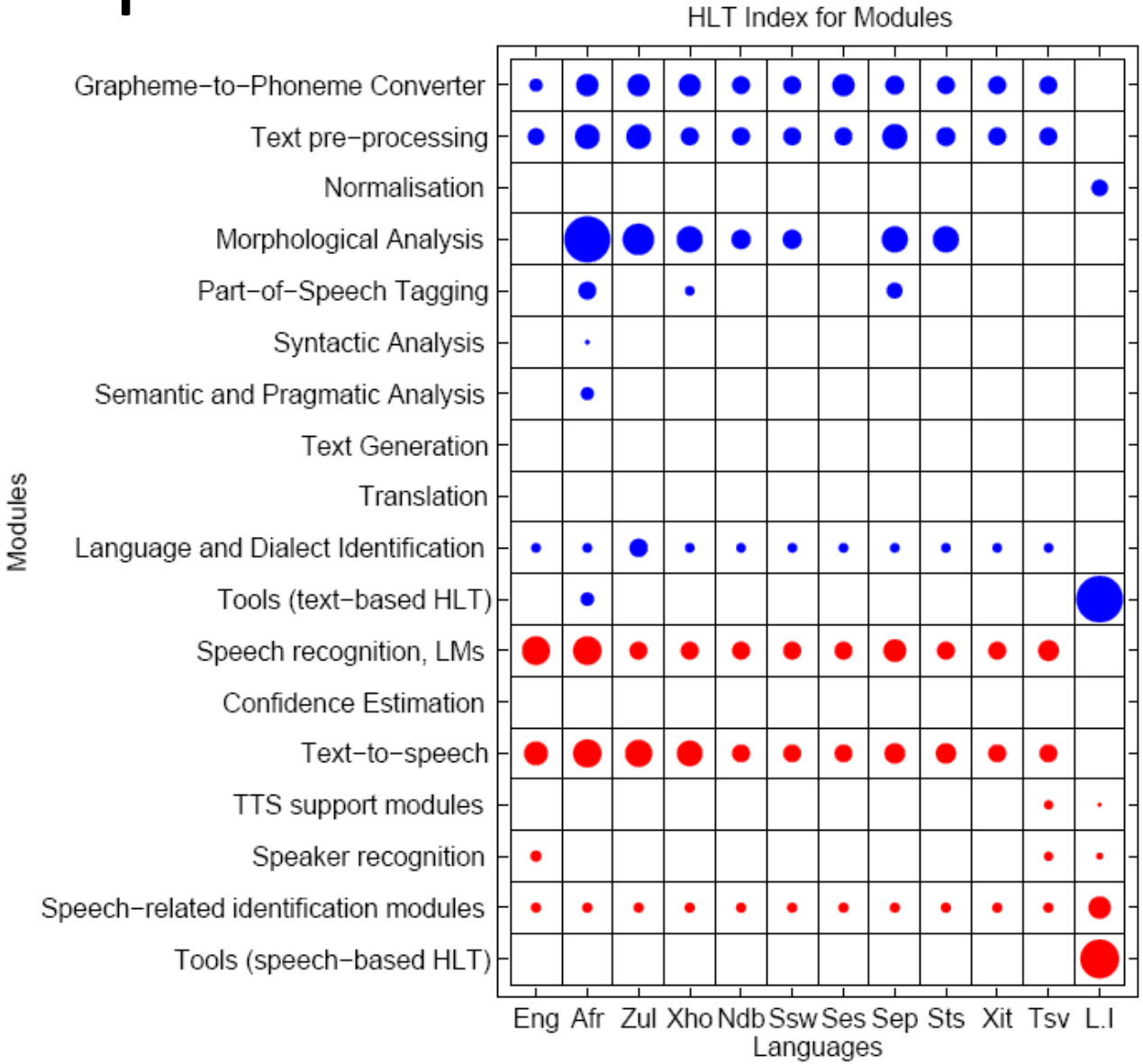
HLT Language Index



HLT Component Indexes

- Alternative perspective:
  - Quantity of activity taking place within each of the data, modules, and applications on a HLT component grouping level (e.g. pronunciation resources)

HLT Component Indexes: Modules



Results

HLT Detailed Inventory

● : Item exists, is accessible, released & of fairly adequate quality

◒ : Item may exist but available for restricted use not released/limited quality

◯ : Items do not exist

'–': Category is not applicable to the language

	APPLICATIONS	Eng	Afr	Zul	Xho	Ndb	Swz	Sso	Sep	Tsn	Xit	Tsv	Lang. Index
TEXT	Proofing/authoring tools	●	●	●	●	●	●	●	●	●	●	●	-
	Authorship Identification	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Text generation	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	OCR/ICR	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Information retrieval	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Information extraction	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Document classification	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Automatic summarisation	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Question Answering (QA)	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Human-aided MT	◒	◒	◯	◒	◯	◯	◒	◯	◯	◯	◯	◒
	Machine-aided human translation	◒	◒	◒	◯	◯	◯	◯	◒	◯	◯	◯	●
	Multilingual comprehension assistants	●	●	◯	◯	◯	◯	●	◯	●	◯	◯	-
	Computer assisted language learning (CALL)	●	●	●	●	●	●	●	●	●	●	●	-
	Dialogue systems (text-based)	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Reference works	●	●	◒	◒	◯	◯	◯	◯	◯	◯	◯	-
SPEECH	Accessibility	◒	◒	◒	◯	◯	◯	◯	◯	◯	◯	◯	-
	Telephony applications (IVR/SDS)	●	●	●	●	◯	◯	●	●	●	◯	◯	●
	Computer assisted language learning (CALL)	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Computer-assisted training	◒	◒	◯	◒	◯	◯	◯	◯	◯	◯	◯	-
	Audio Search	◒	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Audio Management	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Access control	◒	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Embedded speech recognition	◒	◯	◯	◒	◯	◯	◯	◯	◯	◯	◯	-
	Transcription and dictation	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Command & Control	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Announcement systems	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Audio books	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Speaking devices	◒	◯	◯	◒	◯	◯	◯	◯	◯	◯	◯	-
	Speech-to-speech translation	◒	◯	◯	◒	◯	◯	◯	◯	◯	◯	◯	-
	Multimodal information access	◒	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-

Results

Gap Analysis (speech)

- ● : Item exists, is accessible, released & of fairly adequate quality
- ◒ : Item may exist but available for restricted use or not released/ limited quality
- ◯ : Items do not exist
- '–': Category not applicable to the language

Priority	SPEECH			Languages											
	Status	DATA		Eng	Afr	Zul	Xho	Ndb	Ssw	Ses	Sep	Sts	Xit	Tsv	L.I
1	Somewhat Existent		Monolingual speech corpora (minimum orthographic transcription)	●	●	●	●	●	●	●	●	●	●	●	-
		Pronunciation resources	Phoneme sets	●	●	●	●	●	●	●	●	●	●	●	-
			Pronunciation dictionaries	◒	●	●	●	●	●	●	●	●	●	●	-
			Pronunciation models	◯	●	●	●	●	●	●	●	●	●	●	-
			Intonation models	-	-	◒	◯	◯	◯	◒	◒	◒	◯	◯	-
			Corpora of related domains/apps (orthographically representative)						◒						
	Non-existent		Test suites & corpora						◯						
	Status	MODULES		Eng	Afr	Zul	Xho	Ndb	Ssw	Ses	Sep	Sts	Xit	Tsv	L.I
	Somewhat Existent	Speech recognition	Rule-based language models	●	●	◯	◯	◯	◯	◯	◒	◯	◯	◯	-
			Complete ASR	●	●	●	●	●	●	●	●	●	●	●	-
		Text-to-speech	Grapheme-to-phoneme convertor	◯	●	●	●	●	●	●	●	●	●	●	-
			Complete TTS - limited domain	◒	◒	◯	◒	◯	◯	◯	◯	◯	◯	◯	-
			Complete TTS- domain independent	●	●	●	●	●	●	●	●	●	●	●	-
			Normalisation	◒	◒	◒	◒	◒	◒	◒	◒	◒	◒	◒	◒
			Automatic phonetic segmentation	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◒
			POS tagger (see text LRs)						◯						
			Chunker (see text LRs)												
		Speech related ID	Language identification	◒	◒	◒	◒	◒	◒	◒	◒	◒	◒	◒	-
	Non-existent	Speech recognition	Statistical language models	◯											
			Non-native speech recognition												
			Confidence measures (ASR)												
		Text-to-speech	Pre-processing NLP												
			Syllabification												
			Prosody generation												
		Speaker recognition	Speaker identification												
			Diarization (lang-independent)												
	Status	APPLICATIONS		Eng	Afr	Zul	Xho	Ndb	Ssw	Ses	Sep	Sts	Xit	Tsv	L.I
	Somewhat Existent	Accessibility		◒	◒	◒	◯	◯	◯	◯	◯	◯	◯	◯	-
		Telephony applications (IVR/SDS)		●	●	●	●	◯	◯	●	●	●	◯	◯	●
		Audio Search		◒	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	-
	Non-existent	Audio Management													
		Computer assisted language learning (CALL)							◯						

NHN The National HLT Network

SAHLTA Outcomes

- A SAHLTA online database of LRs and applications (alpha)

www.meraka.org.za/nhnaudit

hlt resources // applications

Below is a listing of all applications available. You can browse forward and backward through the applications, search for specific types of application, search for applications available in specific language(s), or perform a general keyword search of all applications.

Search by Language(s) and/or by Application Type

Choose Specific Language(s)... X | None... | All Speech-based Applications... | Search...

- [x] Afrikaans
- [x] English
- [ ] isiNdebele
- [ ] isiXhosa
- [ ] isiZulu
- [ ] Northern Sotho (Sepedi)
- [ ] Setswana

General Keyword Search

keyword search | Sear

Displaying 1 to 4 of 4 records.

	Background and purpose		Languages	Affiliation
	e and alternate communication (AAC) device synthesised (or pre-recorded) speech as icons. Available as a demonstration. (Uses stival, fre...	More »	Afrikaans, English, isiZulu	Meraka Institute, CSIR
	ber of dedicated devices for the blind. The lly a communication and computing device for the blind and visually impaired which runs on a personal computer. It runs specia...	More »	Afrikaans, English, isiZulu	Meraka Institute, CSIR.
Open Spell (v1.0) »	Open Spell is spelling game that provides spelling exercises (in the language education domain) to teach spelling skills to schoolchildren between the ages 7-12 who live in developing regions; current...	More »	Afrikaans, English, isiNdebele, isiXhosa, isiZulu, Sesotho sa Leboa (Northern Sotho/Sepedi), Sesotho (Southern Sotho), Setswana, Siswati, Tshivenda, Xitsonga	Meraka Institute, CSIR, TEIR, ICSI groups at University of California (Berkeley)
AST Prototype Demonstrators »	Afrikaans prototype of an automated hotel information and reservation system developed during the AST project. The prototype system was subjected to usability tests. South African English prototype of...	More »	Afrikaans, English, isiXhosa	CatchWord (http://www.catchwordlst.co.za/)

Displaying 1 to 4 of 4 records.

Summary

- Few resources available, of basic nature
- Several factors influence this:
  - HLT expert knowledge and interests
  - Availability of data resources
  - Market needs of a language
  - Relatedness to other world languages

Recommendations

- Further resource development based on gap analysis
  - Also of more advanced LRs
- Availability and distribution of existing LRs
  - To enable usage, licensing agreements need to be in place
- Funding: support by government in formative years
  - Also industry stimulation programmes (e.g. support for R&D consortia)
- Collaborations: across SA and internationally, also based on gap analysis
- Human capital development (HCD): scientific & technical, cross silos of academic disciplines, especially for lesser-resourced languages

NHN The National HLT Network

Acknowledgments

- DST – project sponsorship
- Prof Sonja Bosch & Prof Laurette Pretorius – results of the 2008 BLaRK survey
- Audit mini-workshop contributors
  - Prof. Danie Prinsloo (UP), Prof. Sonja Bosch (UNISA), Mr. Martin Puttkammer (NWU), Prof. Gerhard van Huyssteen (CSIR), Prof. Etienne Barnard (CSIR), Dr. Febe de Wet (US), Dr. Marelie Davel (CSIR)
- Numerous audit participants
- Various HLT RG members – guidance and support

www.meraka.org.za/nhnaudit