# The South African Human Language Technology Audit

Language Resources and Evaluation, Special Issue: African Language Technology

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

Corresponding author email: asharma1@csir.co.za

## ONLINE RESOURCE

This file is the online resource cited in the above-mentioned article. It contains a series of figures and tables that further provide further information for certain sections of the article (as detailed below).

## SECTION 3.1

Figures A, B, C and D illustrate the detailed ontology of HLT components i.e. data, modules and applications developed during the SAHLTA.

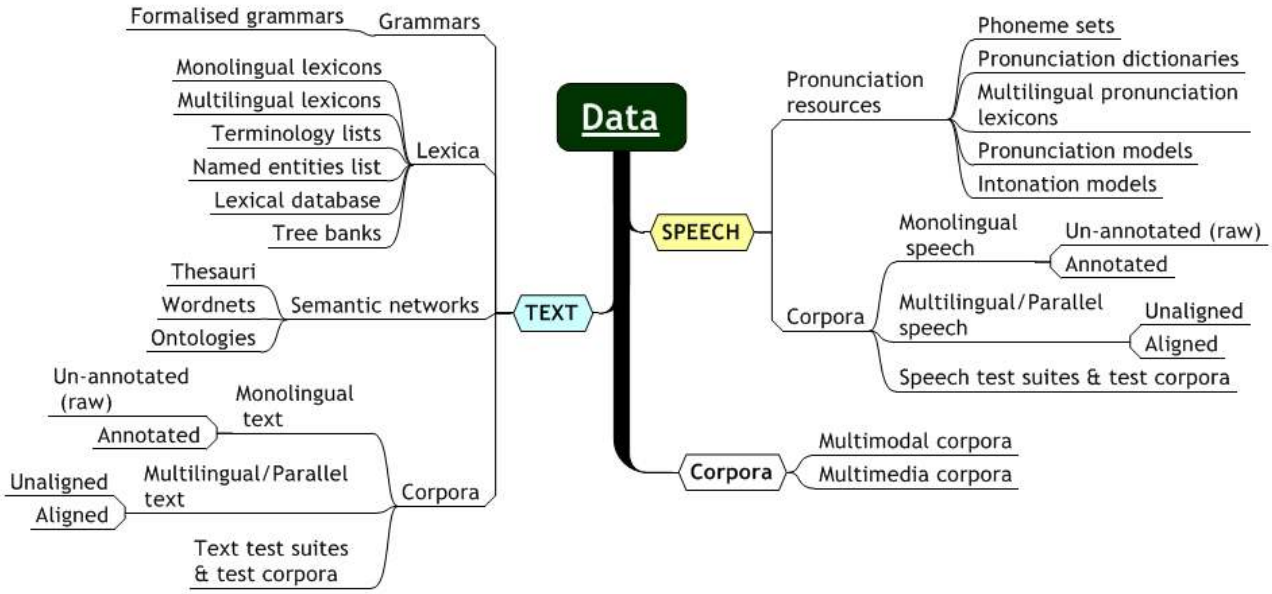


**Fig.A** HLT components: Data ontology

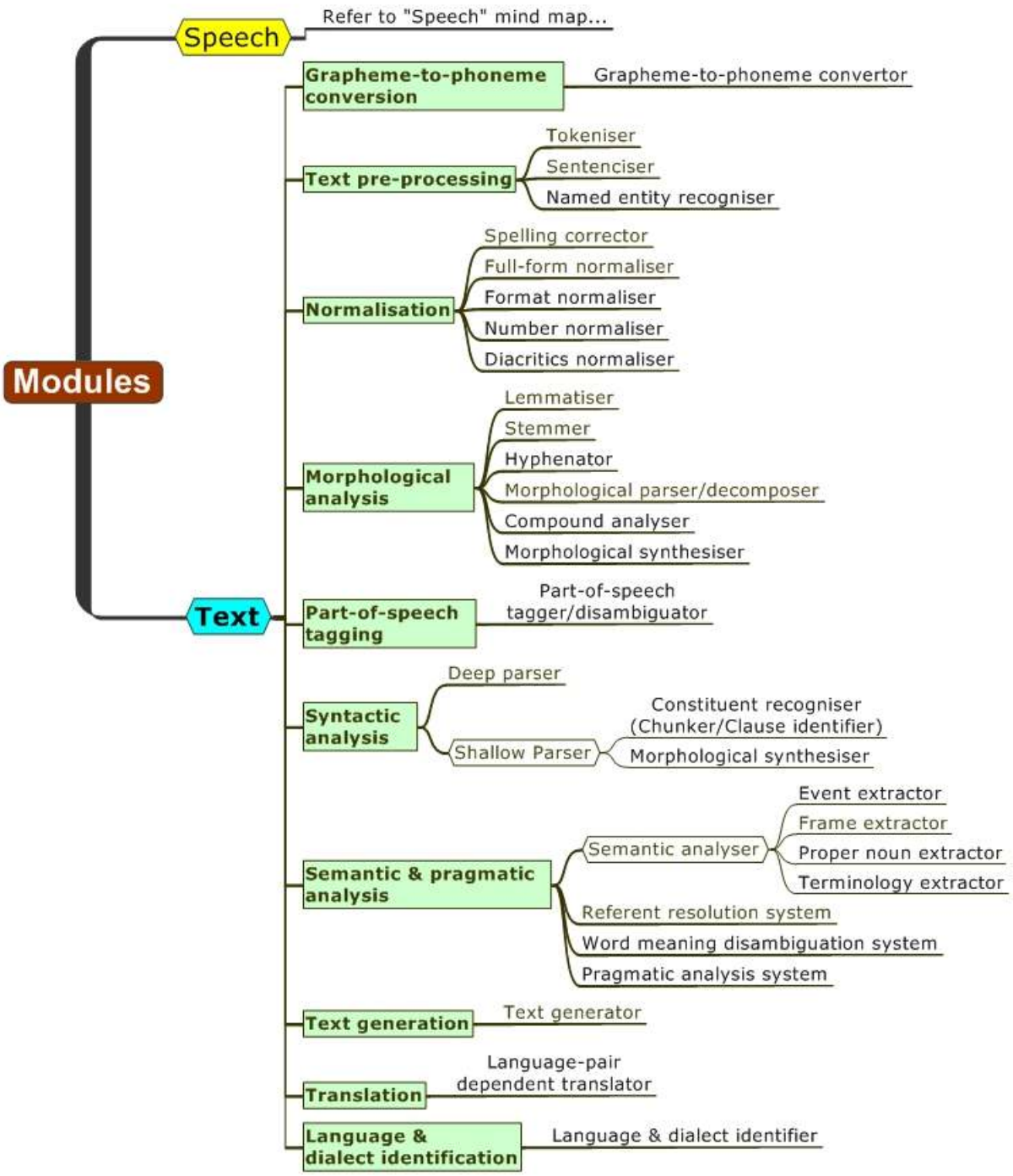


**Fig.B** HLT components: Text modules ontology

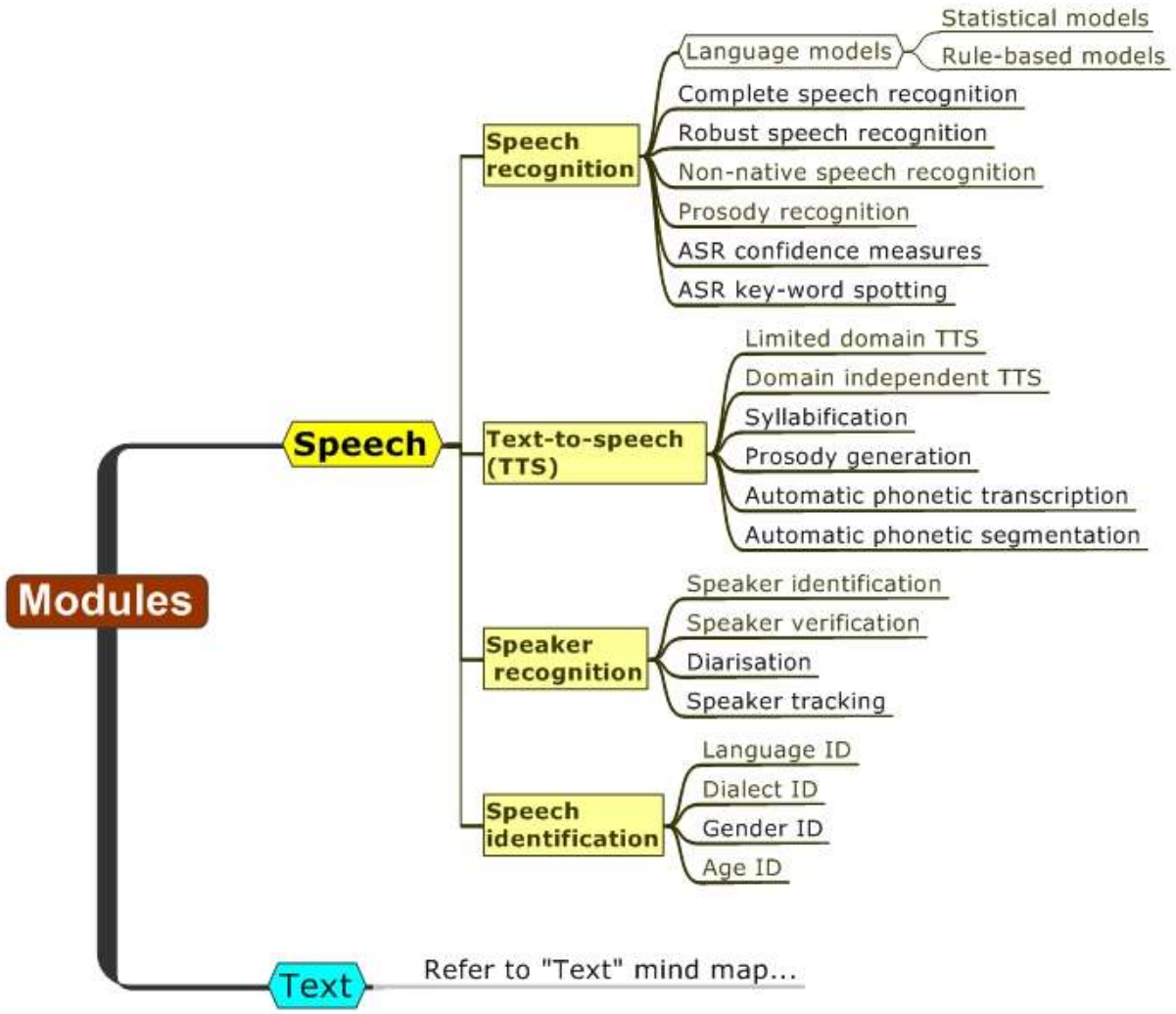


**Fig.C** HLT components: Speech modules ontology

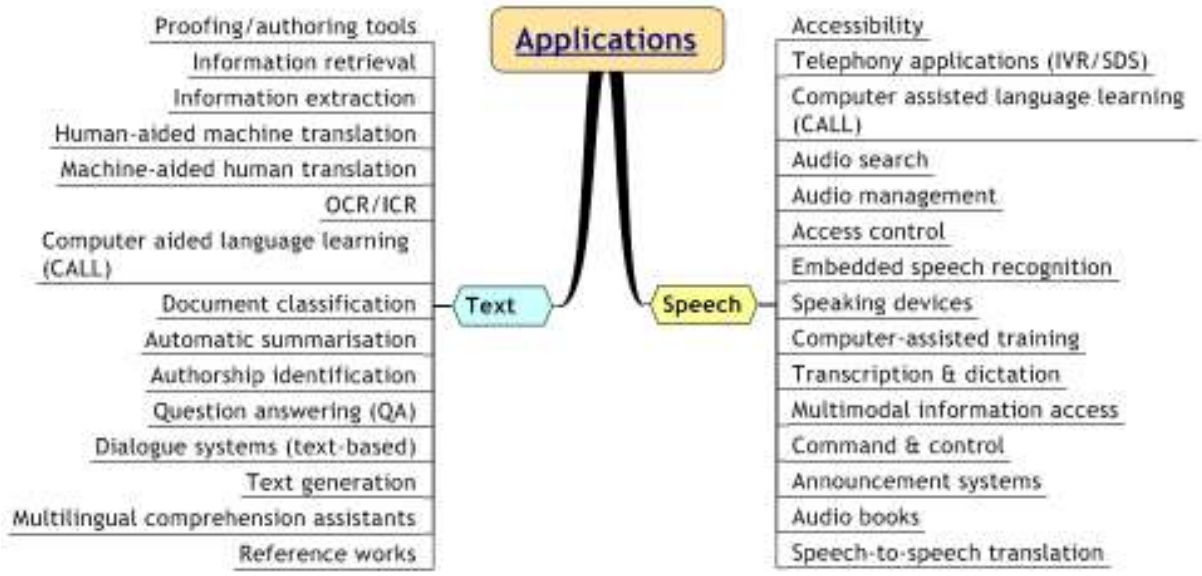


**Fig.D** HLT components: Applications ontology

# SECTION 4.1

Tables A and B respectively show the prioritised HLT modules and data for South Africa determined in the SAHLTA workshop.

**Table A.** Prioritisation of HLT modules for South Africa

| Priority | Text Modules | Speech Modules |
|---|---|---|
| 1 | Grapheme-to-phoneme (G-2-P) convertor | Complete speech recognition (domain-independent and limited)<br>▪ Acoustic models<br>▪ Application based language models. Both statistical (implies large corpus) & rule-based.<br>(Orthographically transcribed representative corpora & related pronunciation resources) |
| | Text pre-processing<br>▪ Tokeniser<br>▪ Sentenciser<br>▪ Named entity recogniser | Non-native speech recognition |
| | Normalisation<br>▪ Spelling corrector<br>▪ Full-form normaliser<br>▪ Format normaliser<br>▪ Number normaliser<br>▪ Diacritics normaliser | Complete TTS (limited domain and domain-independent)<br>▪ G-2-P convertor<br>▪ Pre-processing NLP, POS tagger, Chunker<br>▪ Normalisation<br>▪ Prosody generation<br>▪ Automatic phonetic segmentation<br>▪ Syllabification<br>(Related pronunciation resources) |
| | Morphological analysis<br>▪ Lemmatiser<br>▪ Morphological parser/decomposer<br>▪ Compound analyser<br>▪ Morphological synthesiser | Confidence measures (ASR) |
| | Part-of-speech (POS) tagger | Speaker identification |
| | Syntactic analysis<br>▪ Shallow parser- Constituent recogniser (chunker/clause identifier) | Diarization |
| | Semantic and pragmatic analysis<br>▪ Word meaning disambiguation system | Language identification |
| | Language and dialect identifier | |
| 2 | Morphological analysis<br>▪ Stemmer<br>▪ Hyphenator | Speaker verification |
| | Syntactic analysis<br>▪ Shallow parser (Relation finder) | Dialect identification |
| | Semantic and pragmatic analysis<br>▪ Referent resolution system | Automatic phonetic transcription |
| | Language-pair dependent translator | |
| 3 | Syntactic analysis<br>▪ Deep parser | Prosody recognition |
| | Semantic & pragmatic analysis<br>▪ Semantic analyser (Event, Frame, Proper noun & Terminology extractors)<br>▪ Pragmatic analysis system | Robust speech recognition<br>Speaker tracking |
| | Text generator | Keyword-spotting |

**Table B.** Prioritisation of HLT data for South Africa

| Priority | Text Data | Speech Data |
|---|---|---|
| 1 | Monolingual text corpora<br>▪ Un-annotated<br>▪ Annotated (words+POS, sentence+constituent structure) | Monolingual speech corpora<br>▪ Annotated (with minimum orthographic transcription) |
| | Multilingual text corpora<br>▪ Unaligned<br>▪ Aligned | Corpora of related domains<br>(e.g. for audio search corpus of related searches) |
| | Text test suites & corpora | Speech test suites & corpora |
| | Lexica:<br>▪ Monolingual lexicons<br>▪ Multilingual lexicons<br>▪ Terminology lists<br>▪ Named entity lists | Pronunciation resources<br>▪ Phoneme sets<br>▪ Pronunciation lexicons<br>▪ Pronunciation models<br>▪ Intonation models |
| | Domain/application related text corpora | |
| 2 | Formal grammars | Pronunciation resources<br>▪ Multilingual pronunciation lexicons |
| | Semantic networks<br>▪ Wordnets | Annotated speech corpora with prosodic and phonetic annotations |
| | | Multi-lingual speech corpora<br>▪ Unaligned<br>▪ Aligned (parallel) |
| 3 | Mono-lingual text corpora annotated at sentence+semantic level | Unannotated speech corpora |
| | Semantic networks<br>▪ Thesauri<br>▪ Ontologies | Annotated speech corpora with various types of annotations (age, emotions, speakers, etc.) |
| | Lexica<br>▪ Lexical databases | |
| | Multimodal corpora | |
| | Multimedia corpora | |

## SECTION 4.2.3

Figures E and F show the HLT Component Indexes for data and modules respectively.

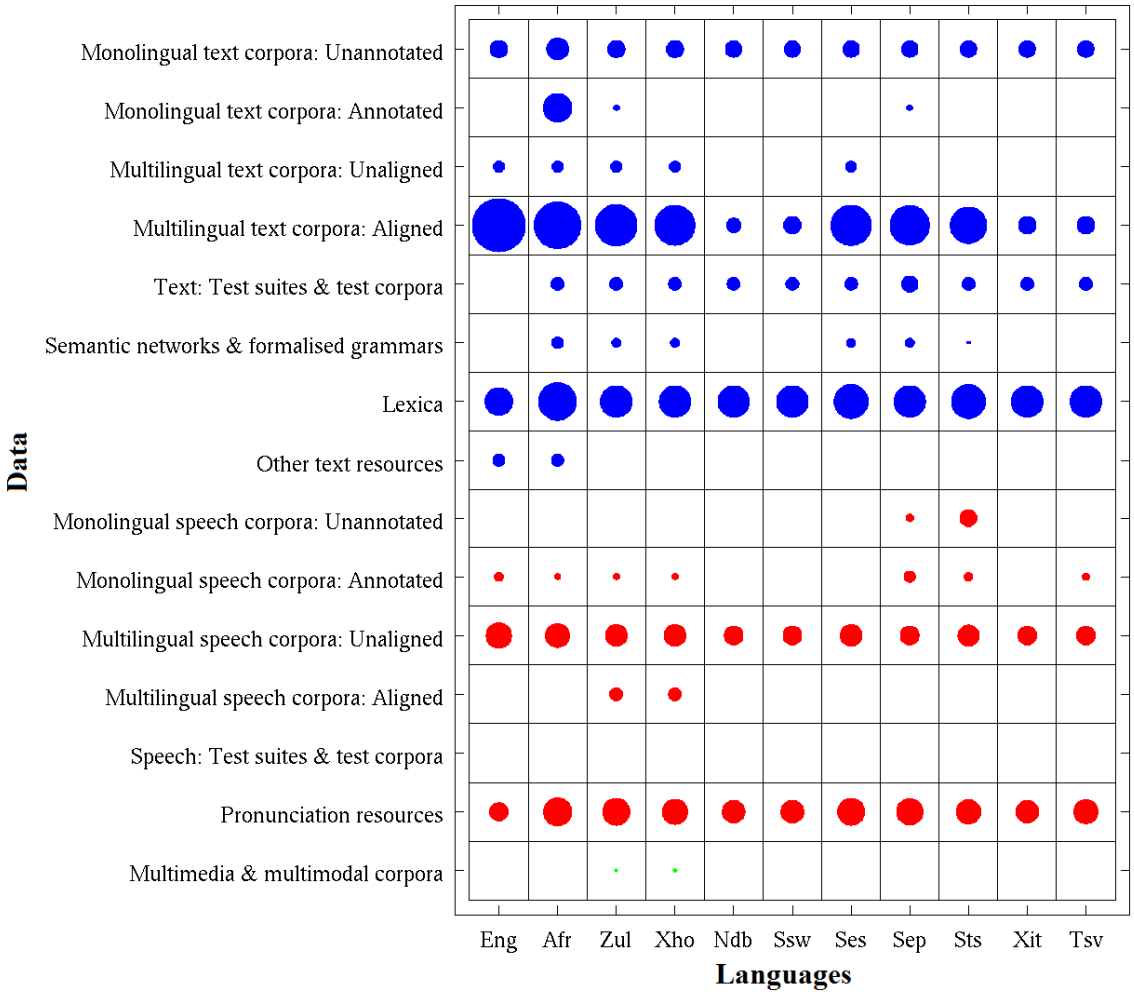


**Fig.E** HLT Component Index for data

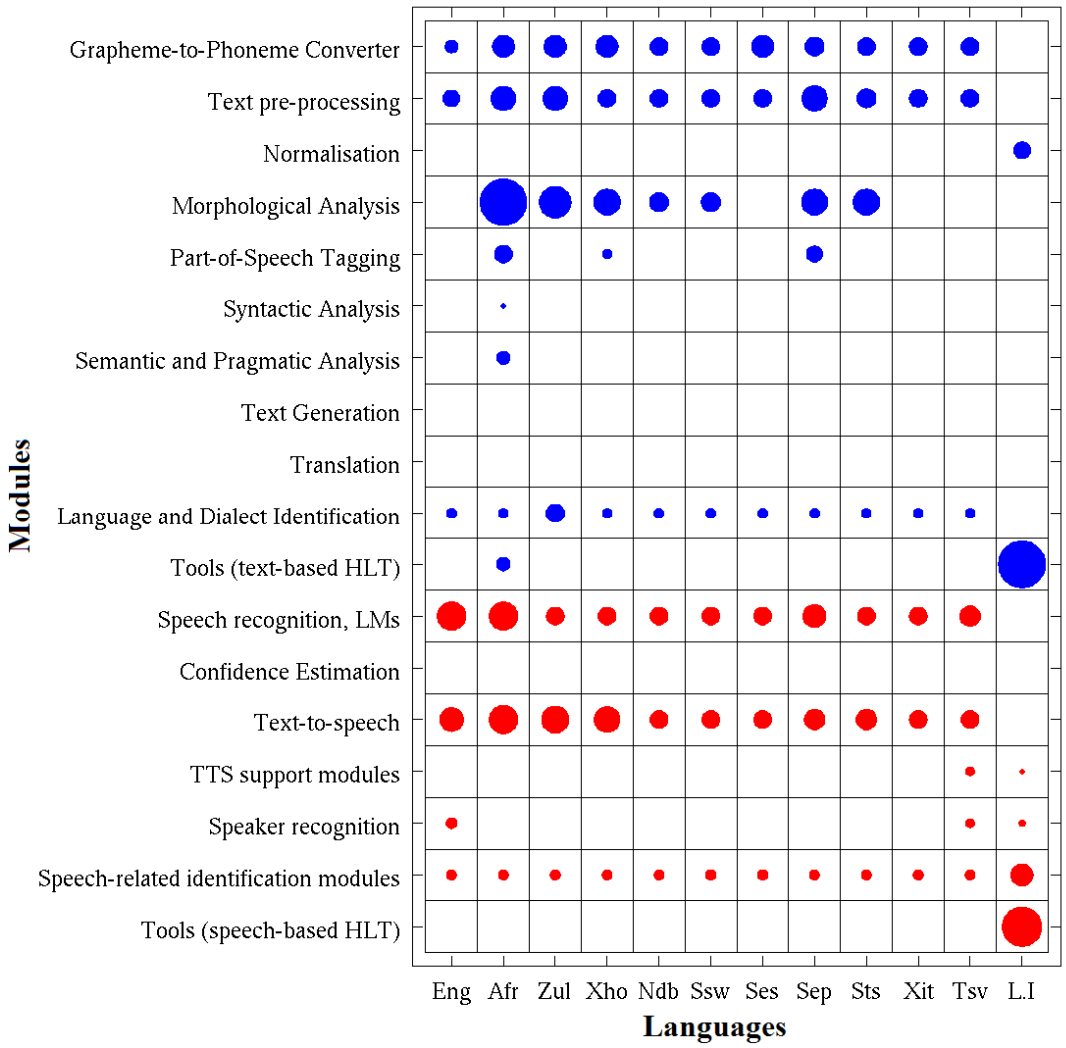


**Fig.F** HLT Component Index for modules

Figures G, H, I, J illustrate a gap analysis, which identifies the gaps between the current status and the prioritised South African HLT components (from the SAHLTA workshop). This information could be highly informative for future road-mapping exercises, as well as to immediately identify areas or languages that should receive particular attention.

The gap analysis for text LRs and applications (Figures G and H) revealed that although all of the data resources identified as priority 1 are somewhat existent, many of them have restricted use due to insufficient maturity or accessibility, and are not always available for all languages. The majority of core HLT modules (priority 1) are somewhat existent, however they are only available in one or a few languages and in most cases are of an uncertain quality. Approximately one-third of the HLT modules that were identified as priority 1 (require definite attention) are currently non-existent. Note, in Figures G and H, components are classified as either 'somewhat existent' if in the detailed inventory analysis they were of fairly adequate quality as ◒, or of somewhat uncertain quality as ●, and if items are 'non existent' as ○.

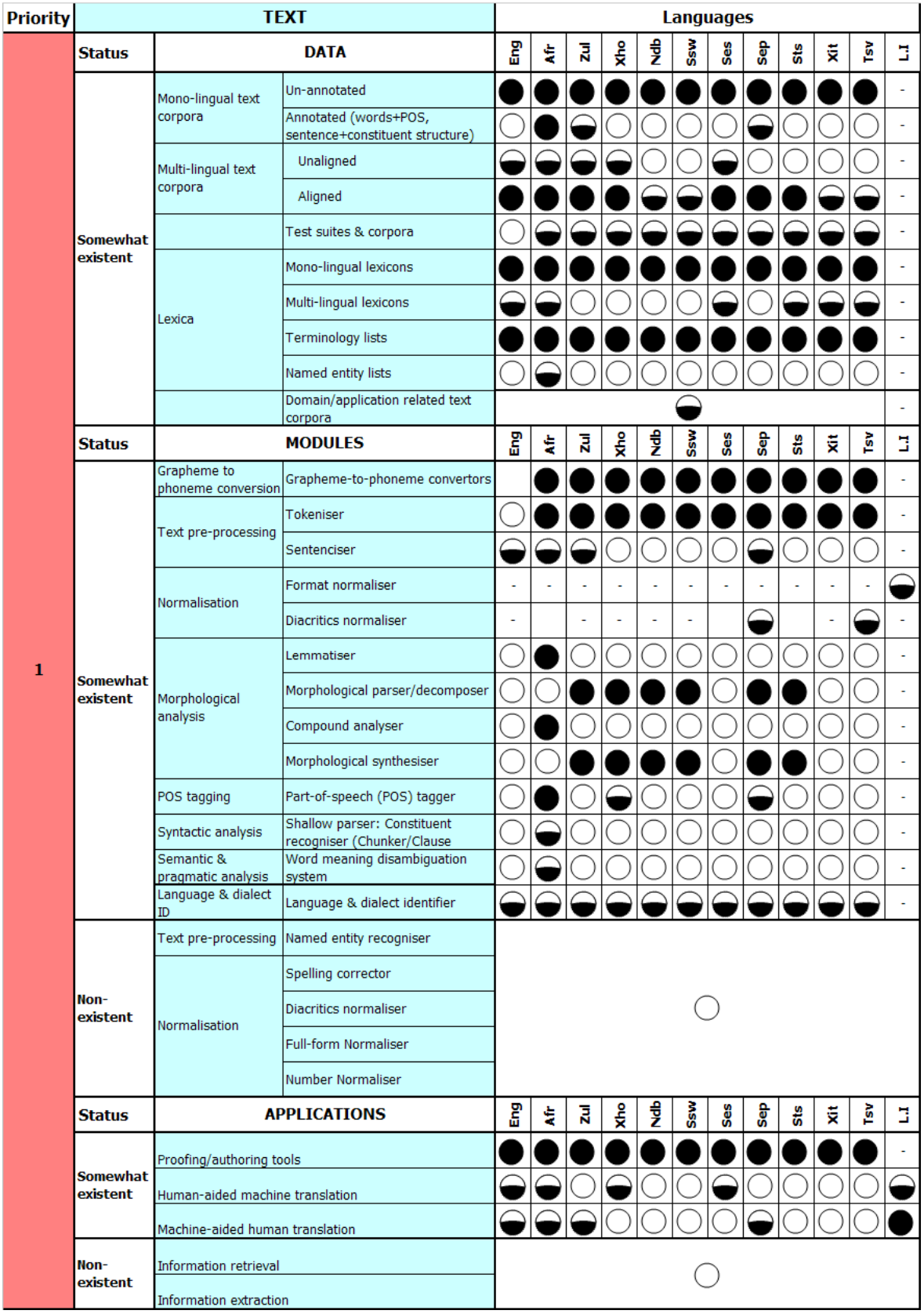


**Fig.G** Text-based HLTgap analysis (priority 1)

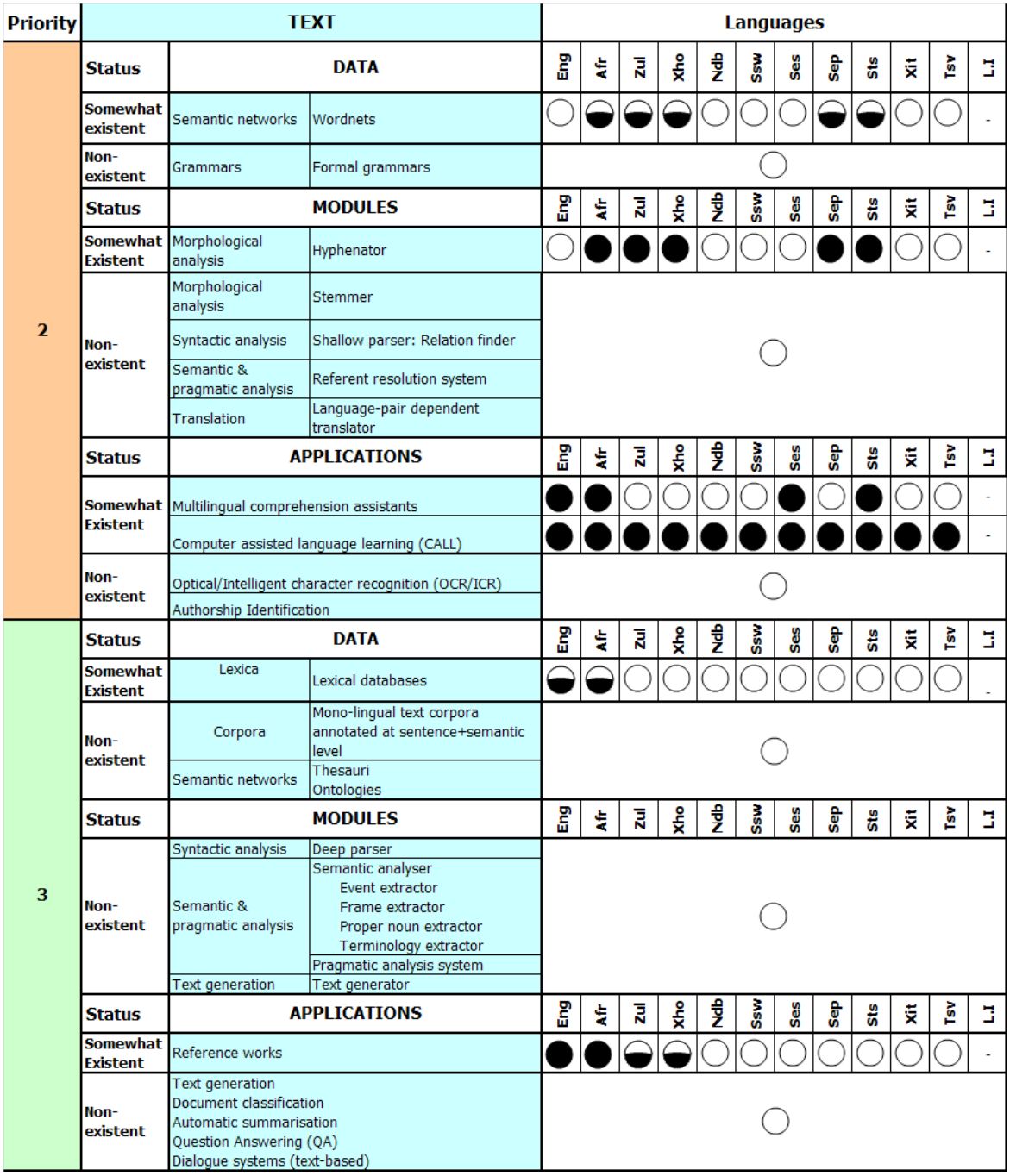


**Fig. H** Text-based HLT gap analysis (priority 2 and 3)

A similar gap analysis was performed for speech-based HLT components across all eleven languages, and is illustrated in Figures I and J. In slight contrast to the text domain, the majority of priority 1 data resources and speech modules (such as pronunciation dictionaries, speech recognition and text-to-speech (TTS) technologies) exist, though mostly in a very basic state. This is as a result of a large, government sponsored project in this domain, which recently made these basic speech domain modules and data available. However, a significant number of core data and modules – such as the more advanced speech recognition and TTS related modules – are still non-existent across the eleven languages. Telephony-based

services (IVR/SDS) is the only priority 1 speech application that has some significant activity (though not equally so in all languages); other applications such as accessibility and audio search are in their very early stages, whilst priority 1 applications like audio management and computer-assisted language learning (CALL) are non-existent.

| 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) | | | | | | | | | | | | | |

**Fig.I** Speech-based HLT gap analysis (priority 1)

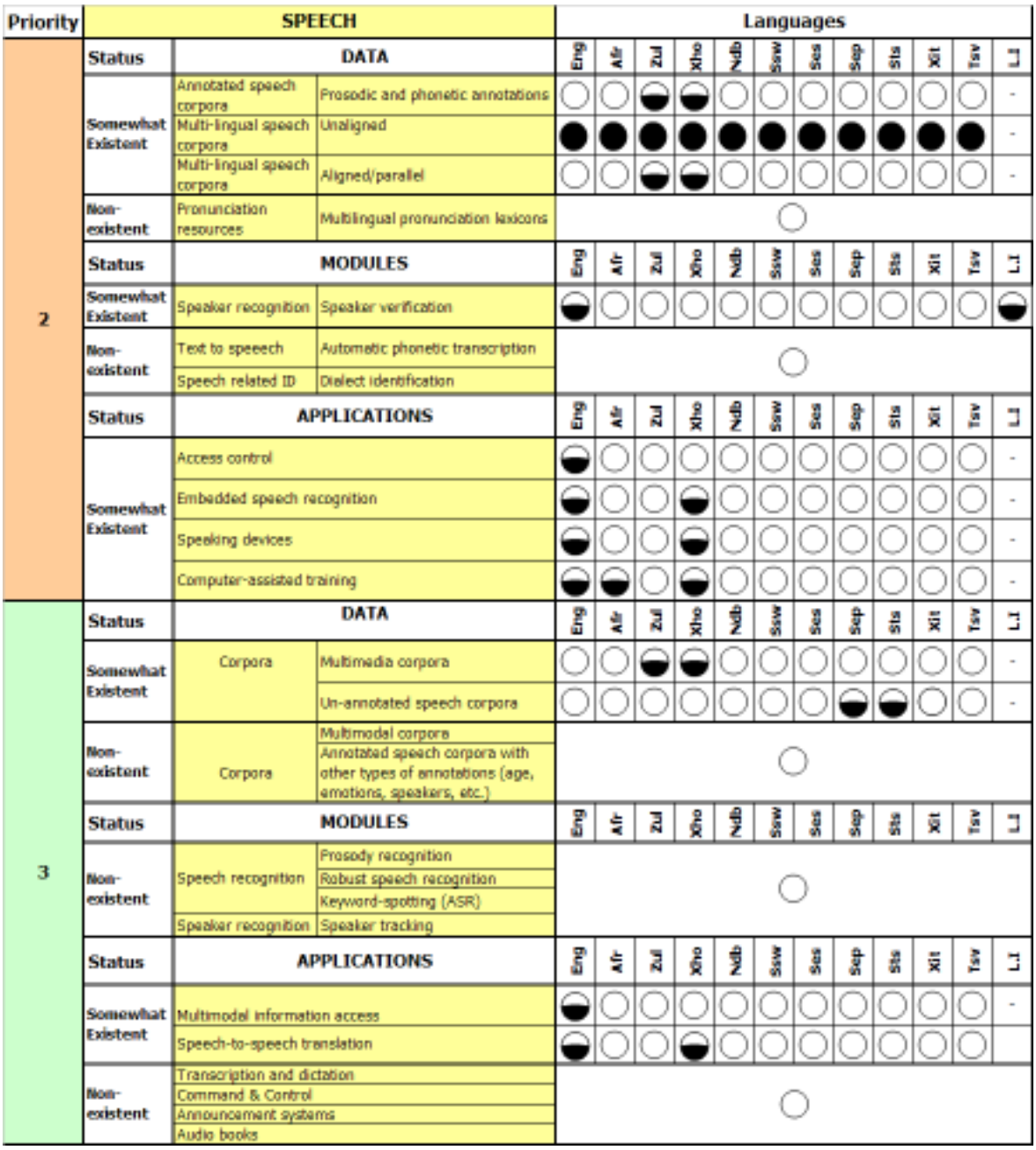


**Fig.J** Speech-based HLT gap analysis (priority 2 and 3)