

Computational language documentation as software development

Procedures and Standards

Dafydd Gibbon
Universität Bielefeld

Bolzano, LULCL 13-14 November 2008

Lesser Used Languages and Computational Linguistics

Objectives of this contribution

- Overview of requirements for
 - Local Language Human Language Technologies (LL-HLT)
 - in an interdisciplinary context:
 - in system Research and Development
 - in Computational Language Documentation
- What we can do
 - HLT = NLP ('language', 'writing') + SLT ('speech')
- Standards and Resources
 - development standards
 - less-resourced languages
- Responsibilities of computational linguists
 - appropriate models
 - education
 - cooperation

Remember

- Science is a language activity, with
 - SYNTAX
 - formal theories, premises, argumentation)
 - SEMANTICS
 - truth, models, absolute/fuzzy, relation to the real world
 - PRAGMATICS
 - the elbows of science – colleagues, money, equipment, infrastructure, education, the kids ...
- Responsibility
 - is not only of physicists, agricultural engineers, medics
 - also linguists, computational linguists
 - payback via
 - applications (texts, dictionaries, grammars)
 - networks (e-cooperation, 'adoption')
 - education (multiplier factors)

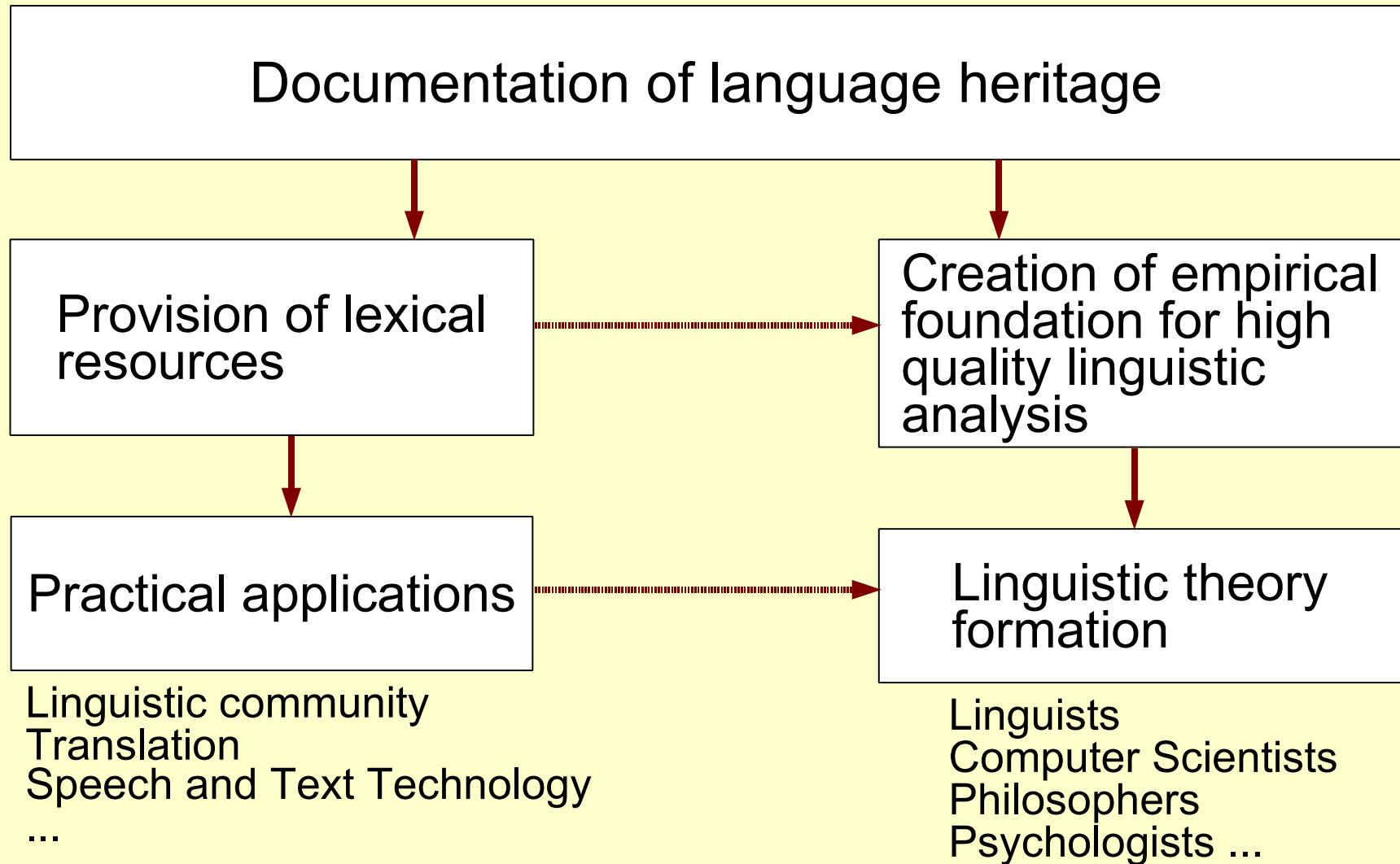
My background: comp ling & lang doc

- Resources:
 - for German & English (Verbmobil project)
 - for West African languages (Côte D'Ivoire, Nigeria)
 - for heritage documentation and speech technology
- Organisations, projects, for example:
 - COCOSDA
 - Coordinating Committee for Speech Databases and Assessment
 - Trippel: member
 - Gibbon: Chairman
 - EU SAM project, contributor to SAM-PA (SAMPA) alphabet
 - EU EAGLES projects 1 & 2:
 - Gibbon & al. 1997: Handbook of Standards and Resources ...
(Gibbon & al. 1997 hypertext version)
 - Gibbon & al. 2000: Handbook of Multimodal and Spoken Dialogue ...
 - EMELD (worldwide consortium led by *LinguistList* coordinators)

General specifications for resources

- Criteria for local language resources, tools and systems - CESAF:
 - Comprehensive (with respect to application domain)
 - Effective (in terms of human and computing resources)
 - State-of-the-art (intellectually, not necessarily the latest internet-dependent software and hardware)
 - Affordable (for example older computing facilities may be available)
 - Fair (for example: does software localisation benefit the producer or the community or both?)
- One consequence: open archives, open software

Functionality of Computational Linguistics



BLARK: an important step forward

- Joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) [Krauwert, 1998]
 - Launched with Dutch initiative “Dutch Human Language Technologies Platform” (April 1999).
 - ENABLER thematic network (European National Activities for Basic Language Resources -Action Line: IST-2000-3.5.1)
 - ELDA report: (minimal) set of LRs to be made available for as many languages as possible, mapping the actual gaps that should be filled in so as to meet the needs of the HLT field.
- Later initiatives:
 - Arabic BLARK
 - EthioBLARK
 - extensions to speech resources

BLARK: standardised basic requirements

- Requirements specification for
 - basic applications
 - basic resources for HLT technology
 - matrices:
 - use-cases X modules
 - applications
 - resources
- Krauwer/ELDA BLARK (Basic Language Resource Kit)
 - BLARK aims: <http://www.elda.org/blark/>
 - BLARK matrices:
 - Applications: http://www.elda.org/blark/matrice_app_mod.php
 - Resources: http://www.elda.org/blark/matrice_res_mod.php
- Further refinement necessary:
 - links between resources & applications
 - updating of speech resources

Compare legacy and state-of-the-art content ...

- Content
 - traditional:
 - collections of texts and recordings
 - card-indexed and printed lexicons
 - traditional printed sketch grammars and full grammars
- State of the art:
 - interoperability: annotated for archiving & application, standards
 - metadata characterisations:
 - attribute-value sets (DC, IMDI, OLAC), ontologies
 - formal and empirical coherence criteria
 - completeness (no false negatives), soundness (no false positives)
 - task-specific, ergonomic criteria:
 - naturalness, comprehensibility: acceptability, fitness for purpose

Compare legacy and state-of-the-art methods ...

- Legacy:
 - pencil and paper note-taking
 - discontinued electronic data formats
 - unsupported software
 - discontinued operating systems
- State of the art:
 - extensions of BLARK
 - professionally designed tools:
 - professional archiving (XML, Unicode)
 - ubiquitous computing:
 - software interoperability / hardware compatibility
 - physical portability: laptops, PDAs, telecommunication
 - standardisation of interchange specifications
 - as far as possible
 - without prejudicing creativity and innovation

System R&D and Documentary Linguistics

- Surprisingly to some:
 - System R&D and Computational Documentary Linguistics share a wide range of goals
- These shared, BLARK-like goals, include:
 - development of
 - necessary resources
 - Human Language Technology systems
 - for problem-solving in processes of
 - communication infrastructure
 - language learning
 - language and cultural heritage documentation ...
 - for applications in the areas of
 - education
 - health services
 - trade ...

The conclusion seems rather obvious:

transdisciplinary cooperation
is necessary
at all relevant locations in all continents

in both resource development for R&D
and in Language Documentation resourcing

Transdisciplinary R&D context for LL-HLT (1)

- TEXT:
 - Text technology, Natural Language Processing for
 - information retrieval
 - abstracting, term extraction
 - text generation
 - machine translation
- SPEECH:
 - Speech technology for
 - assistive technologies for barrier-free access (TTS, ASR)
 - information and control front ends in visually difficult environments
 - security (speaker verification & identification)

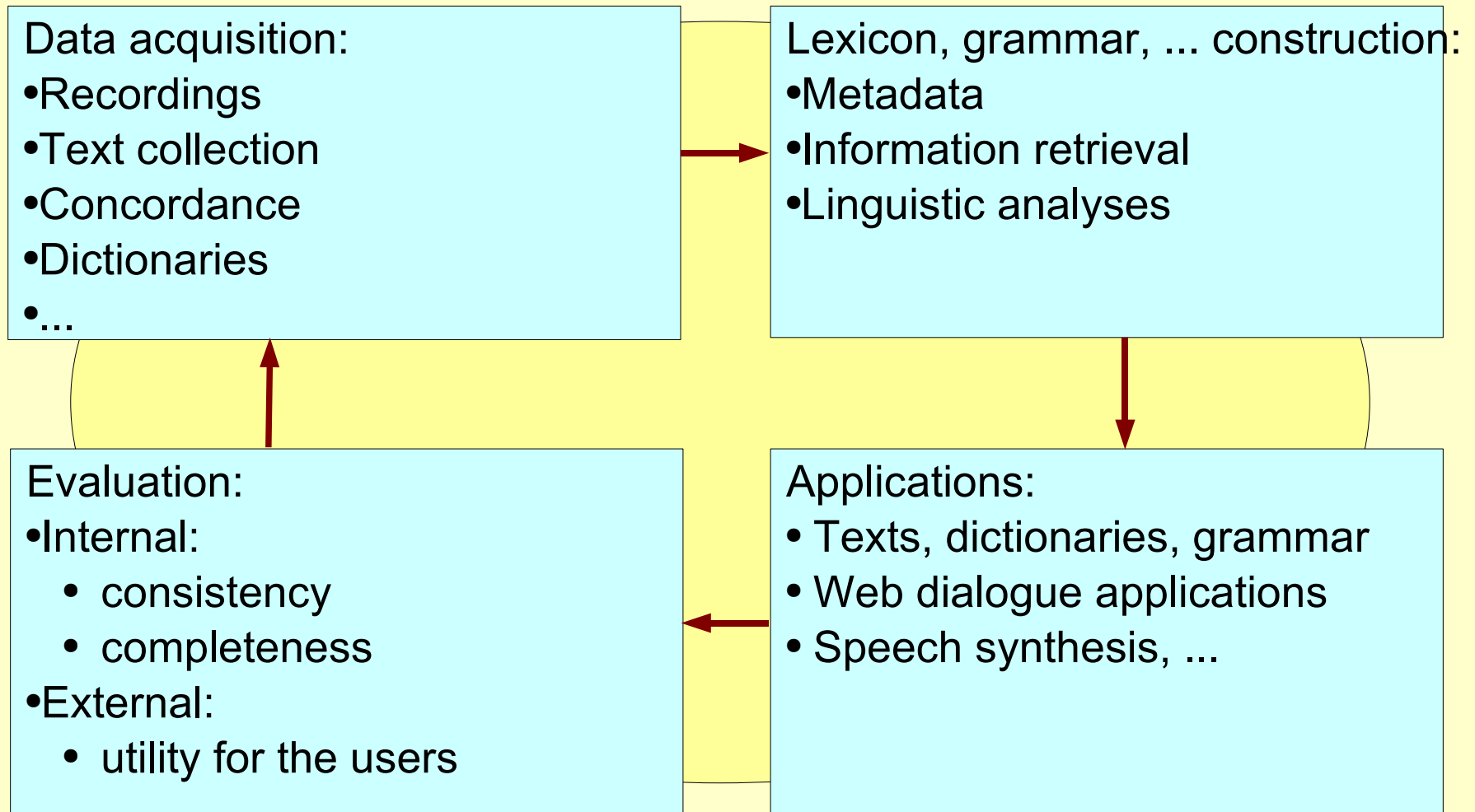
Transdisciplinary R&D context for LL-HLT (2)

- What do TEXT and SPEECH resources BOTH need?
 - requirements for
 - annotation standards:
 - phonetic labelling (complete: IPA, SAMPA)
 - linguistic tagging (partial: language specific)
 - linguistic descriptions as a basis for
 - annotation tools
 - lexicon construction tools
 - grammar induction tools
 - easy to use software
 - comprehensive, efficient, state-of-the-art, affordable, fair

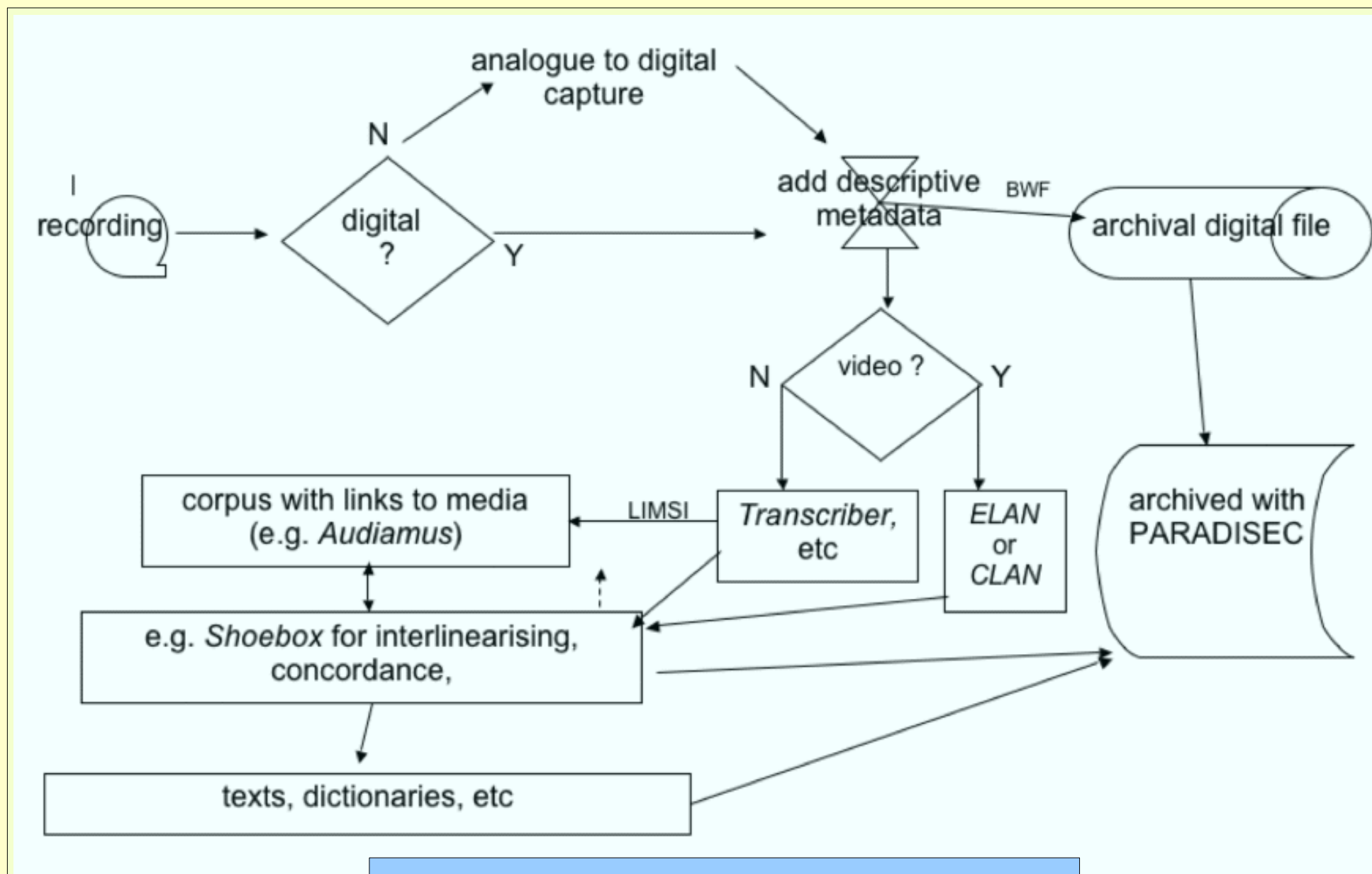
Transdisciplinary R&D context for BLARK (3)

- Which disciplines are involved in this cooperation?
 - Linguistics and phonetics:
 - phoneticians:
 - phonetic, phonemic, prosodic analyses
 - linguists:
 - word, sentence, text, dialogue grammars
 - semantic and pragmatic world + user models
 - Computer science
 - software engineers
 - text and speech pattern recognition specialists
 - human-machine interface specialists

Basic workflow

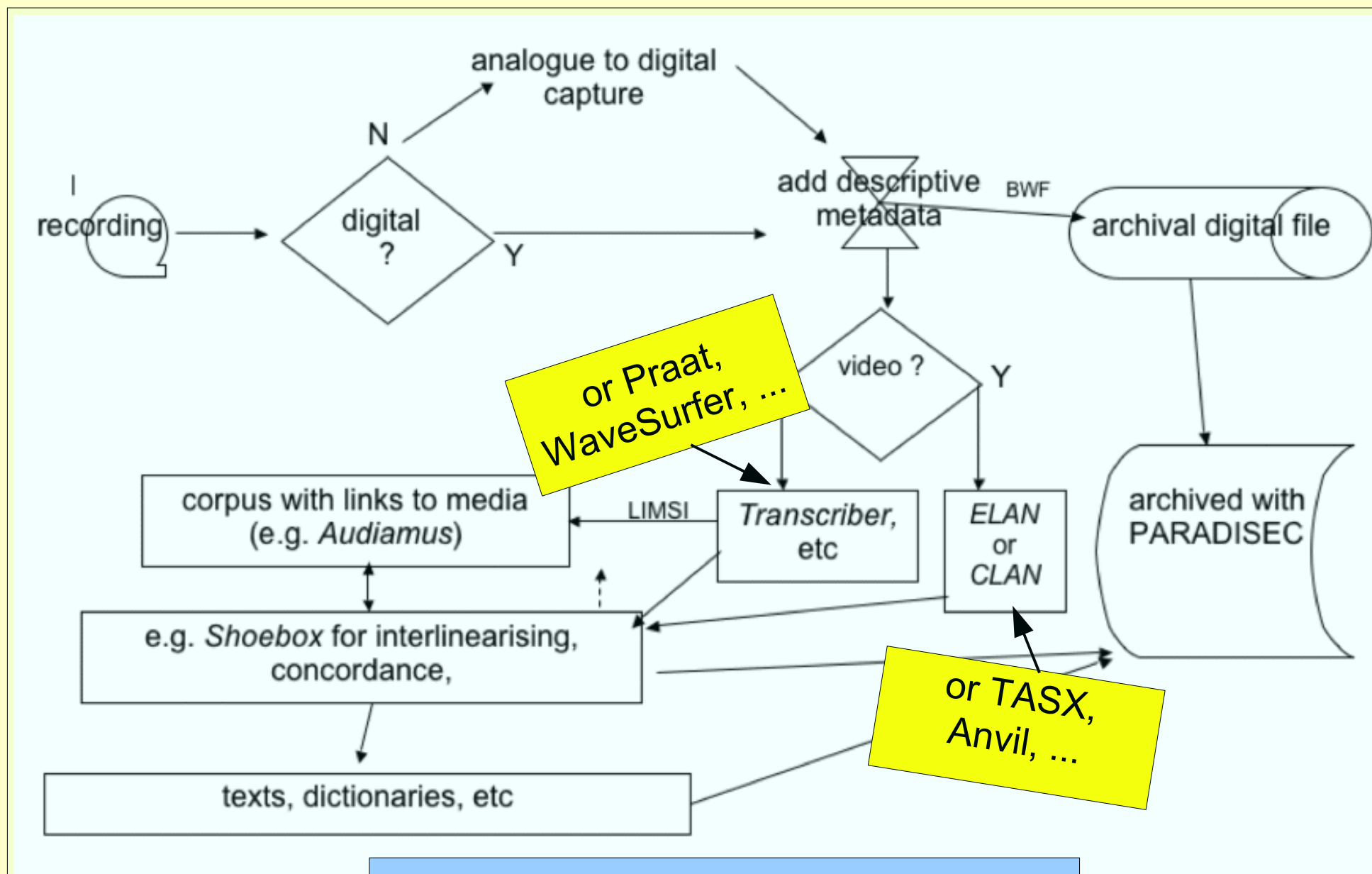


Resource creation in linguistics



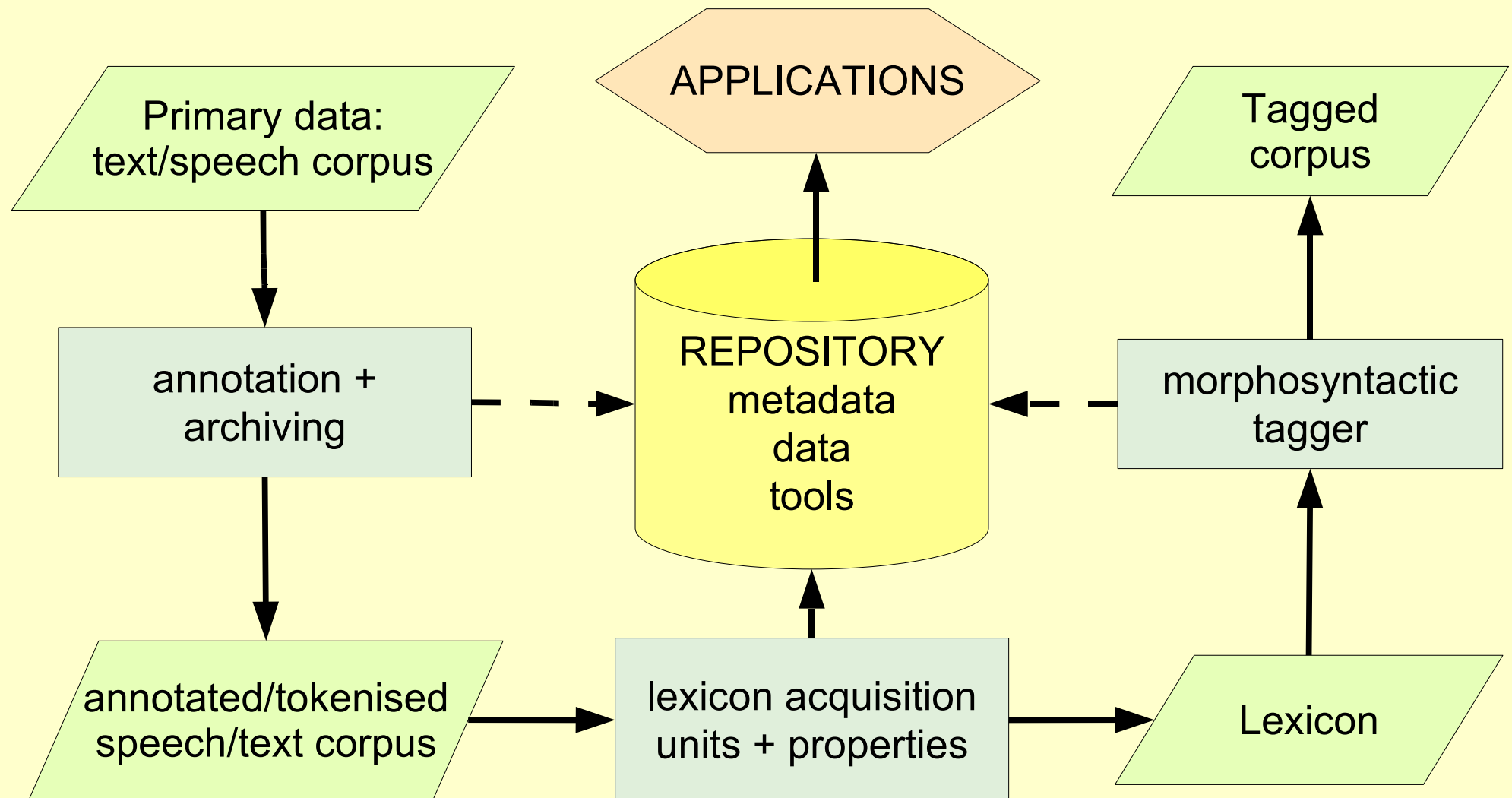
Nick Thieberger, David Nash (Australia)

Resource creation in linguistics

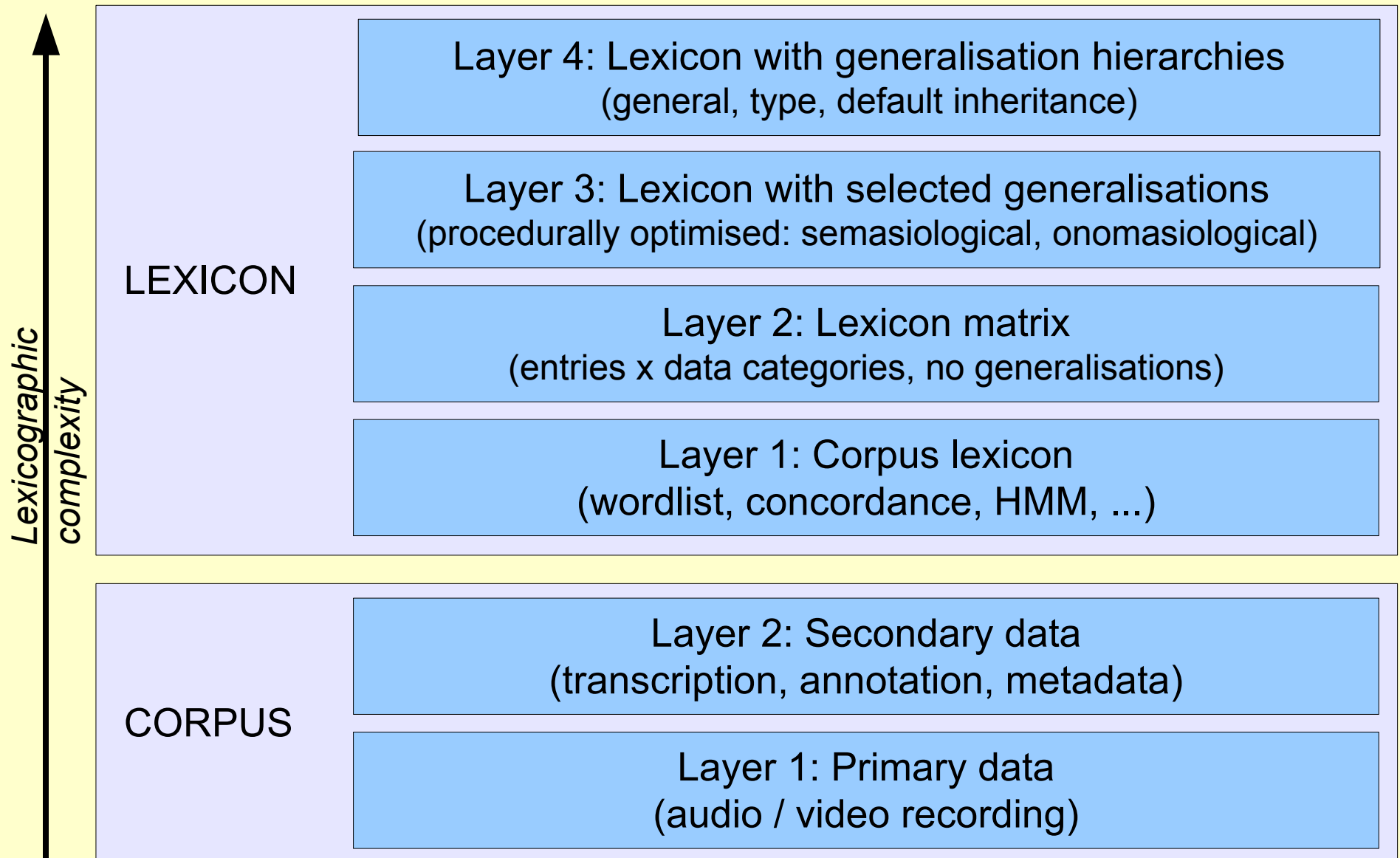


Nick Thieberger, David Nash (Australia)

Resource creation model: TEXT and SPEECH



Coherent lexicography



Still: the two domains of text and speech
each have their special requirements.
For example, speech technology development
requires training of and cooperation between

ENGINEERS

(system design and development; ...)

PHONETICIANS

(phonetic categories; annotation; evaluation; ...)

COMPUTATIONAL LINGUISTS

(lexicons, language models, parsers, taggers, text I/O; ...)

Speech-specific resources

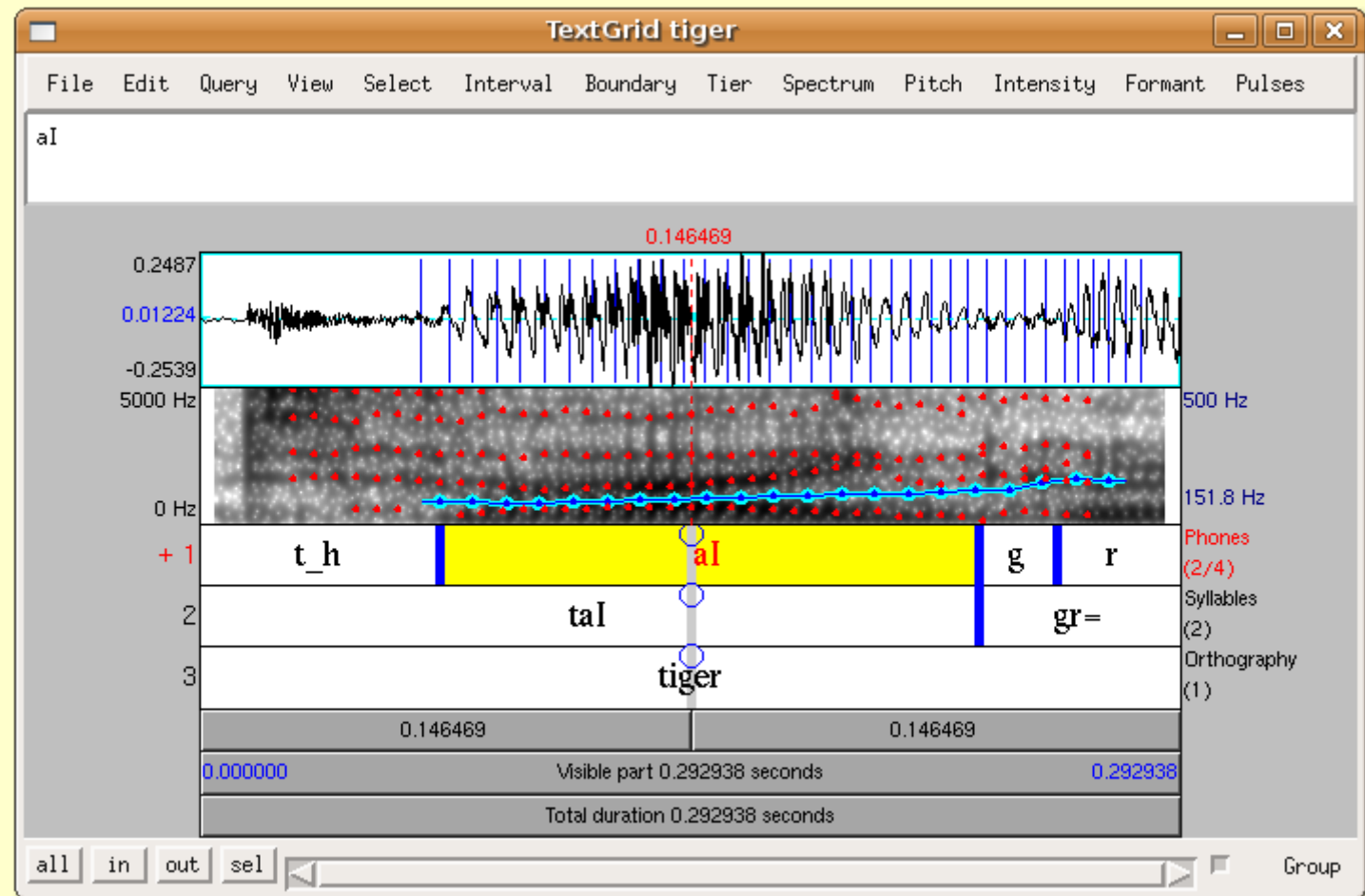
- Automatic Speech Recognition (ASR):
 - semi-automatically time-aligned speech + transcription
 - extraction of corpus lexicon
 - specific statistical methods, for example:
 - acoustic feature detection (bottom up pattern recognition)
 - language modelling (for top-down prediction)
 - stochastic word-graph search
- Text To Speech synthesis (TTS):
 - time-aligned speech + transcription for
 - identification of units (diphone and larger) in corpus
 - extraction of phoneme / duration / pitch values

Cooperation with phoneticians

Cooperation with phoneticians

waveform,
spectrogram,
formants, pitch track:

time-aligned transcription
(annotation, labelling):



- The signal + annotation contains all the information required for corpus input to speech recognition and automatic speech synthesis
- Praat has a scripting language for performing time-alignment automatically

MBROLA speech synthesis

phonemes

durations

MBROLI editor for MBROLA

ThankYouForYourAttention.pho - Mbroli

File Edit Tools View Help

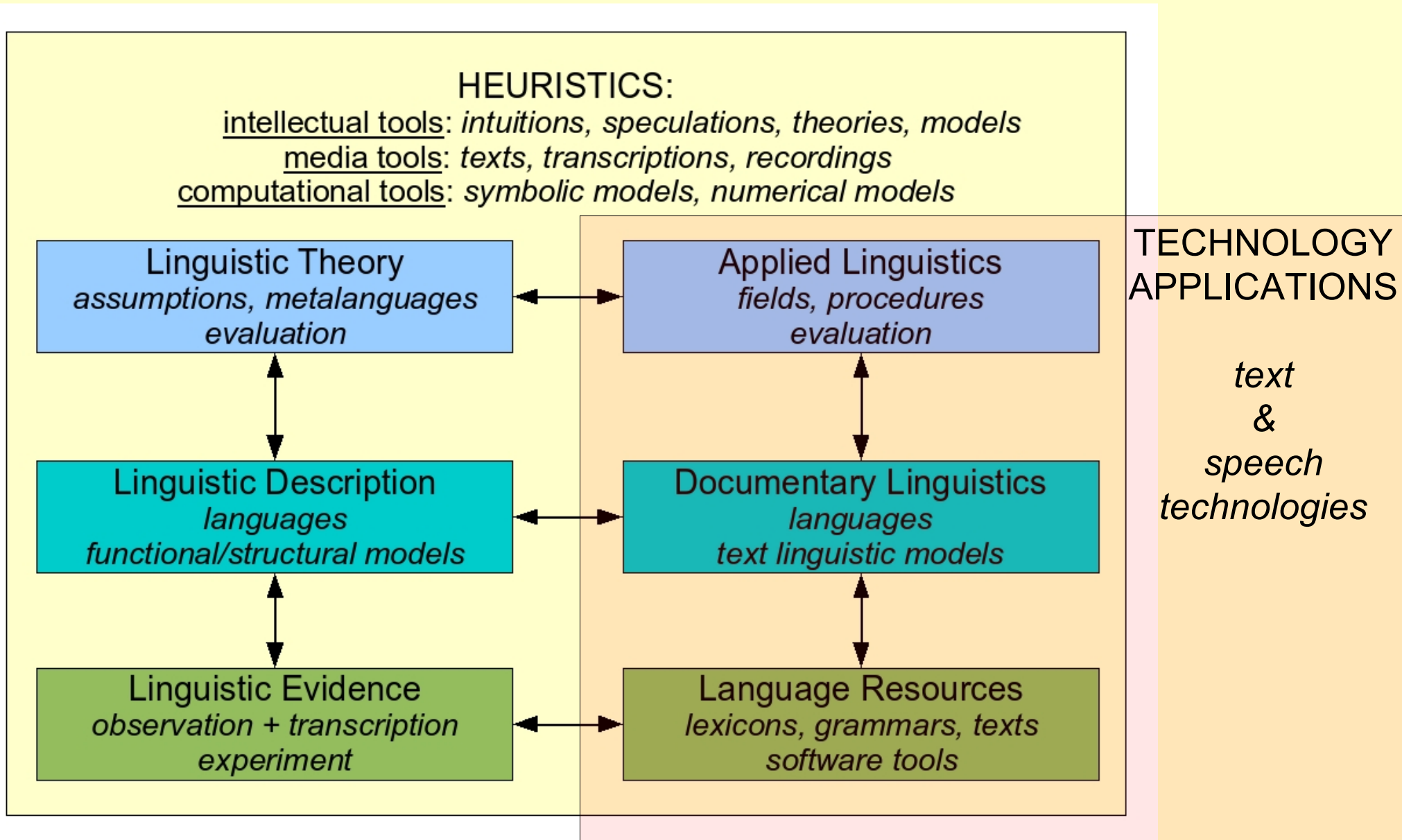
en1

THANK YOU FOR YOUR ATTENTION
DG

T	50		
{	50		
N	100	50	210
K	50	50	200
J	50		
U:	50	50	170
F	100	50	170
@	70		
J	100	50	150
O:	70	50	150
R	100	50	140
@	50	50	130
T	50		
E	70		
N	120	50	150
S	70	50	130
@	70		
N	50	50	120
	150	50	110
T	1000		
{	50		
N	100	50	410
K	50	50	400
J	50		
U:	50	50	360
F	100	50	360
@	70		
J	100	50	340
O:	70	50	330
R	100	50	320
@	50	50	300
T	50	50	280
E	70		
N	120	50	330
S	70	50	300
@	70		
N	50	50	260
	150	50	220
T	50		
#			

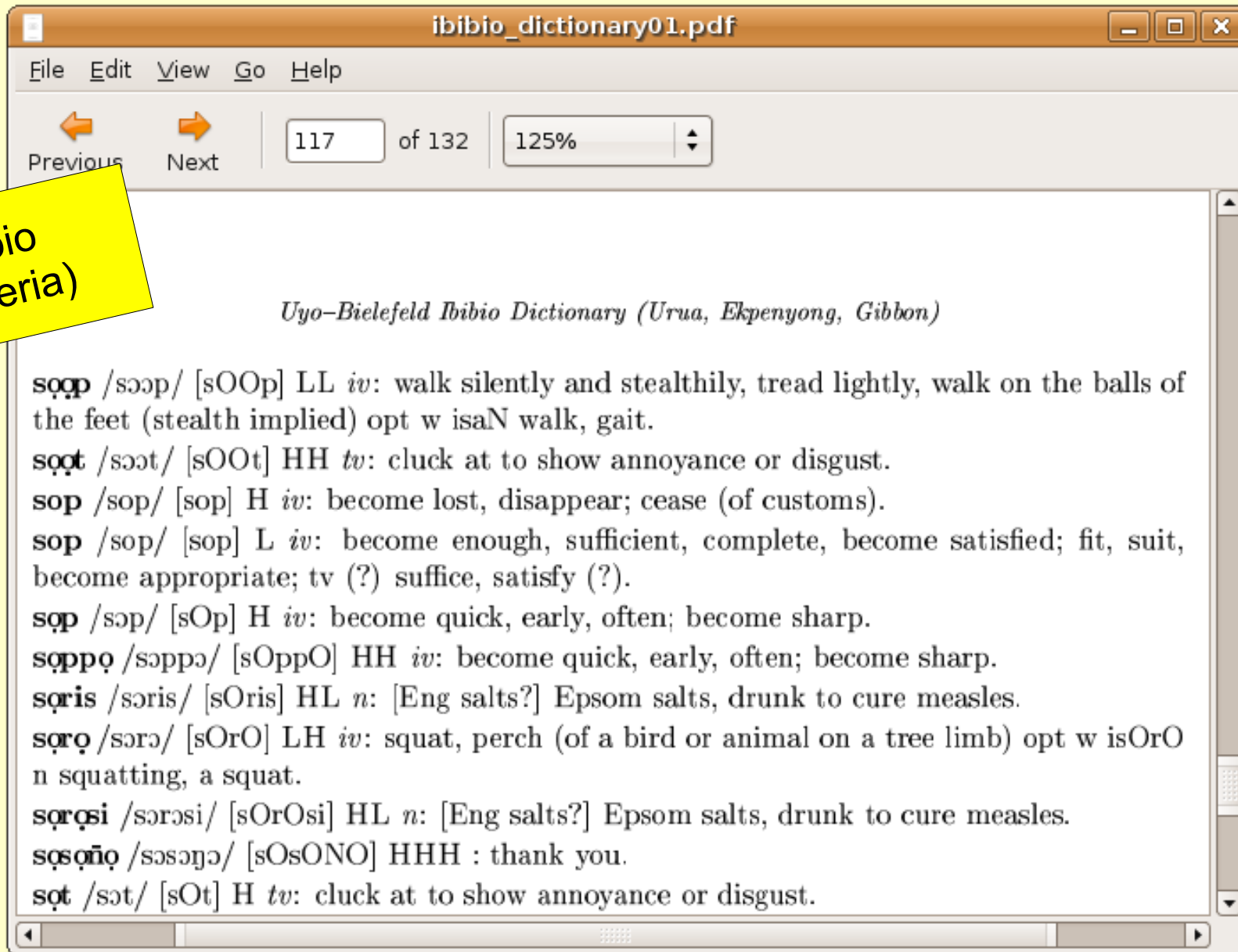
Ready

Cooperation with computational linguists



Cooperation with CL: lexical databases

**Ibibio
(Nigeria)**



The screenshot shows a PDF viewer window titled 'ibiblio_dictionary01.pdf'. The window has a menu bar with 'File', 'Edit', 'View', 'Go', and 'Help'. Below the menu bar is a toolbar with 'Previous' and 'Next' buttons, a page number '117 of 132', and a zoom level '125%'. The main content area displays the title 'Uyo-Bielefeld Ibibio Dictionary (Urua, Ekpenyong, Gibbon)' and a list of Ibibio words with their phonetic transcriptions and meanings. A yellow sticky note with the text 'Ibibio (Nigeria)' is placed over the left side of the PDF content.

Uyo-Bielefeld Ibibio Dictionary (Urua, Ekpenyong, Gibbon)

sɔp /sɔp/ [sOOp] LL *iv*: walk silently and stealthily, tread lightly, walk on the balls of the feet (stealth implied) opt w isaN walk, gait.

sɔt /sɔt/ [sOOt] HH *tv*: cluck at to show annoyance or disgust.

sop /sop/ [sop] H *iv*: become lost, disappear; cease (of customs).

sop /sop/ [sop] L *iv*: become enough, sufficient, complete, become satisfied; fit, suit, become appropriate; *tv* (?) suffice, satisfy (?).

sop /sɔp/ [sOp] H *iv*: become quick, early, often; become sharp.

soppo /sɔppɔ/ [sOppO] HH *iv*: become quick, early, often; become sharp.

sɔris /sɔris/ [sOris] HL *n*: [Eng salts?] Epsom salts, drunk to cure measles.

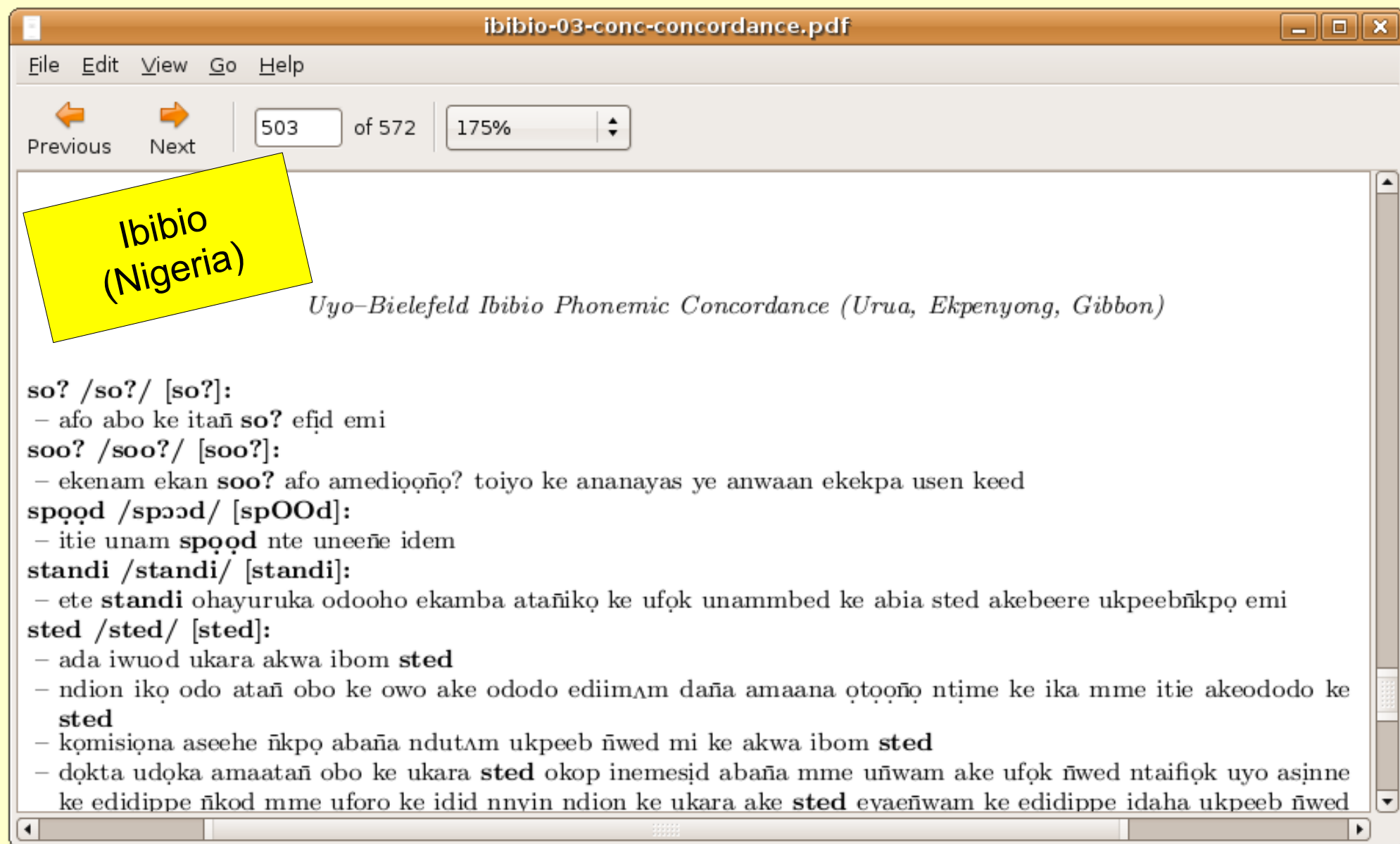
sɔro /sɔrɔ/ [sOrO] LH *iv*: squat, perch (of a bird or animal on a tree limb) opt w isOrO *n* squatting, a squat.

sɔrosi /sɔrɔsi/ [sOrOsi] HL *n*: [Eng salts?] Epsom salts, drunk to cure measles.

sɔsɔno /sɔsɔɲɔ/ [sOsONO] HHH : thank you.

sɔt /sɔt/ [sOt] H *tv*: cluck at to show annoyance or disgust.

Cooperation with CL: concordancing



ibibio-03-conc-concordance.pdf

File Edit View Go Help

Previous Next 503 of 572 175%

**Ibibio
(Nigeria)**

Uyo-Bielefeld Ibibio Phonemic Concordance (Urua, Ekpenyong, Gibbon)

so? /so?/ [so?]:
– afo abo ke itañ **so?** efiḍ emi

soo? /soo?/ [soo?]:
– ekenam ekan **soo?** afo amediọọñọ? toiyo ke ananayas ye anwaan ekekpa usen keed

spọd /spɔd/ [spOOd]:
– itie unam **spọd** nte uneefi idem

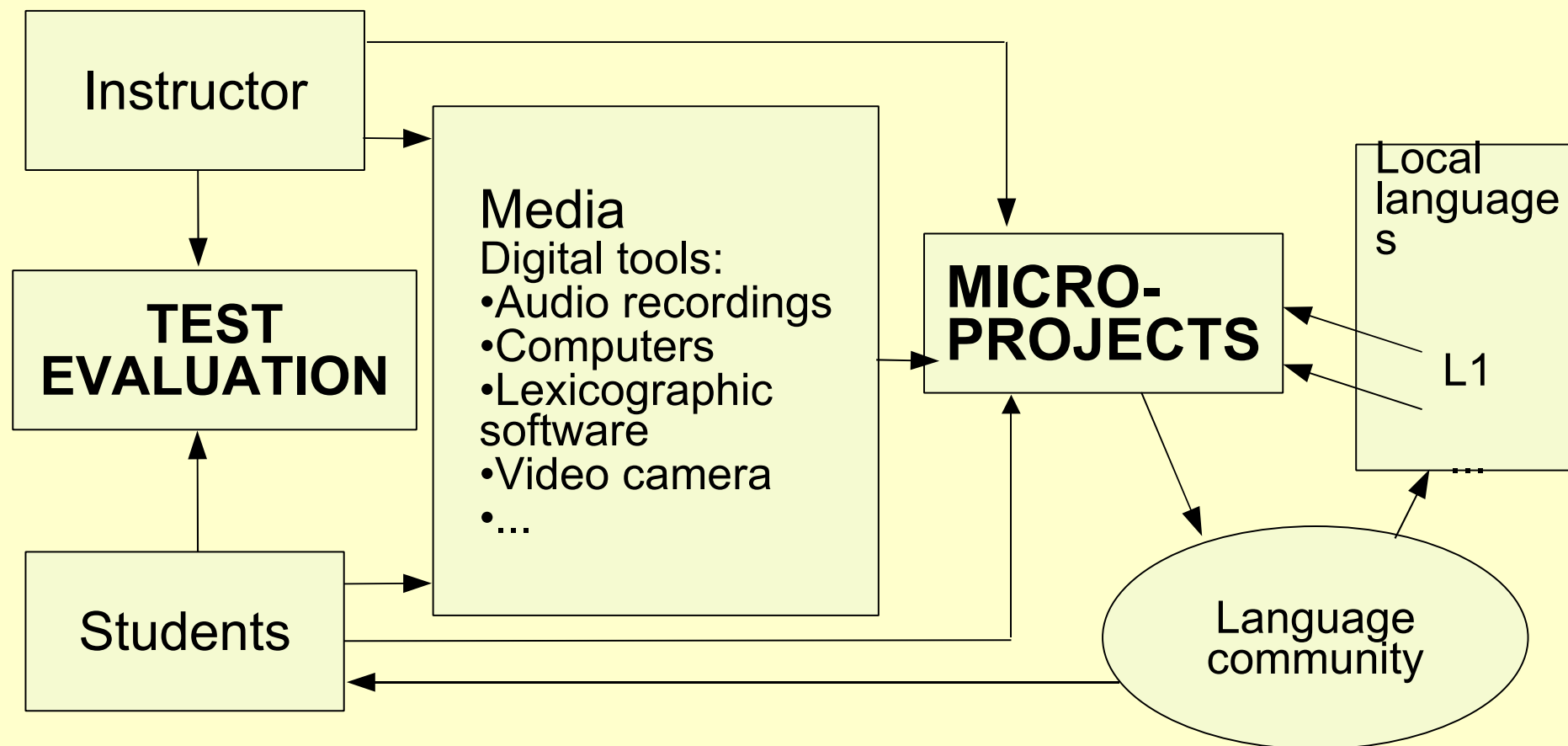
standi /standi/ [standi]:
– ete **standi** ohayuruka odooho ekamba atañikọ ke ufọk unammbed ke abia sted akebeere ukpeebnkpọ emi

sted /sted/ [sted]:
– ada iwuod ukara akwa ibom **sted**
– ndion ikọ odo atañ obo ke owo ake ododo ediimam daña amaana ọtọtọ ntime ke ika mme itie akeododo ke **sted**
– kọmisiọna aseehe nkpọ abaña ndutam ukpeeb nīwed mi ke akwa ibom **sted**
– dọkta udọka amaatañ obo ke ukara **sted** okop inemesiḍ abaña mme uñwam ake ufọk nīwed ntaifiok uyo asinne ke edidippe nīkod mme uforo ke idid nnyin ndion ke ukara ake **sted** evaeñwam ke edidippe idaha ukpeeb nīwed

Education

- Cooperation with local linguists & computer scientists
- As with any technical activity, a good understanding of procedures is necessary.
- Therefore:
 - basic training in computation
 - General:
 - Python
 - Simple web-based dialogue systems (CGI)
 - Natural Language Toolkit (NLTK)
 - web text formatting / deformatting
 - wordlist extraction from corpora, concordance construction
 - finite automata, XFST
 - Speech:
 - Praat annotation
 - MBROLA speech synthesis interface manipulation

Teaching methods



Results of cooperation with Abidjan & Uyo

- A lexical databases for use in
 - language and speech systems
 - production of dictionaries for general use
- A prototype TTS synthesiser for Ibibio
 - Check: <http://www.llsti.org/>
- An MA course “Computational documentation of Local Languages” at
 - Université de Cocody, Abidjan (Côte d'Ivoire)
 - University of Uyo, Akwa Ibom State (Nigeria)

**In conclusion, some practical examples
and tentative recommendations**

A few selected examples of blarkish work

- Worldwide:
 - The Local Languages Speech Technology Initiative (LLSTI): speech synthesis in Africa & India:
 - <http://www.llsti.org/>
 - India:
 - [Simputer](#)
 - South Africa:
 - ASR & TTS initiatives
 - Lexical databases for the 11 official languages
 - Free software (basic resources, overlooked in this context):
 - Operating systems (Linux based)
 - Applications (e.g. OpenOffice, Mozilla; Praat, MBROLA)
 - The Open Archive Initiative (OLAC)
- There are numerous other examples (cf. Kenya, South Africa)
 - some of which are rather well networked internationally
 - some of which are local and not sufficiently networked

Recommendations

- R&D:
 - Standards for annotation, tagging, archival formats
 - Interoperable tools for specific tasks:
 - Praat for annotation creation, MBROLA for checking annotations via speech re-synthesis
 - Generic text tools: taggers, concordancers, ...
- NETWORKING
 - network of cooperation and exchange:
 - local
 - regional
 - continental (esp. Africa: “African COCOSDA”?)
 - transcontinental (nucleus: COCOSDA – NB: LREC)
 - because our best resources are colleagues and students ...

THANKS!

