

Efficient Language Documentation: creation of local multipliers

Dafydd Gibbon
Universität Bielefeld

LSA 2008 Annual Meeting Tutorial, January 2008:

Mobilizing linguistic resources within speaker communities

Overview

- Goals
- By way of explanation – my background
- Specifications for resources
- Interim consequences
- Transdisciplinary shared goals for resources
- Transnational & transdisciplinary cooperation
- Current resource creation in doc ling
- Integrated computational resource development
- Example of local multiplier oriented cooperation
- Cooperation with Computational Linguistics
- Cooperation with speech technology
- The future...

Goals (1)

- General goal:
 - Over view and sketch of programme elements for Documentary Linguistics in the 21st century
- Why?
 - Documentary Linguistics is coming of age
 - Heterogeneous influences from corpus-based work in
 - field linguistics
 - Speech Technology
 - Natural Language Processing
 - computational (statistical) corpus linguistics
 - word processor development
 - (spell checkers, grammar checkers, lexicons, document models)
 - document modelling for automatic document classification, text mining and information retrieval
 - machine learning:
 - automatic induction of grammars, lexicons
 - this is not fantasy – a well-established field in R&D

Goals (2)

- Specific goal:
 - Workable, efficient language documentation
- How?
 - Social:
 - collaborative multi-level partnership (for creating local multipliers and “human payback”)
 - Empirical:
 - valid empirical procedures (for descriptive accuracy, soundness and completeness)
 - Formal:
 - suitable linguistic models (for consistency, archiving, search, lexicon and grammar induction, ...) - GOLD
 - Operational:
 - suitable archiving and processing data structures: annotation conventions – accepted tagsets for different levels of description
 - productive uses of speech and language technology (for validating formal and empirical resources and for applications) - BLARK

By way of explanation – my background

- Core research:
 - fieldwork methods: West African Languages
 - computational phonology/prosody/morphology/lexicography
- Applications: machine processable resources
 - for heritage documentation & speech technology
 - for various languages:
 - for German & English (Verbmobil project)
 - for West African languages (Côte D'Ivoire, Nigeria)
- Involvement in organisations, projects, e.g.:
 - COCOSDA
 - Coordinating Committee for Speech Databases and Assessment
 - 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)

SPECIFICATIONS FOR RESOURCES

General specifications for language resources

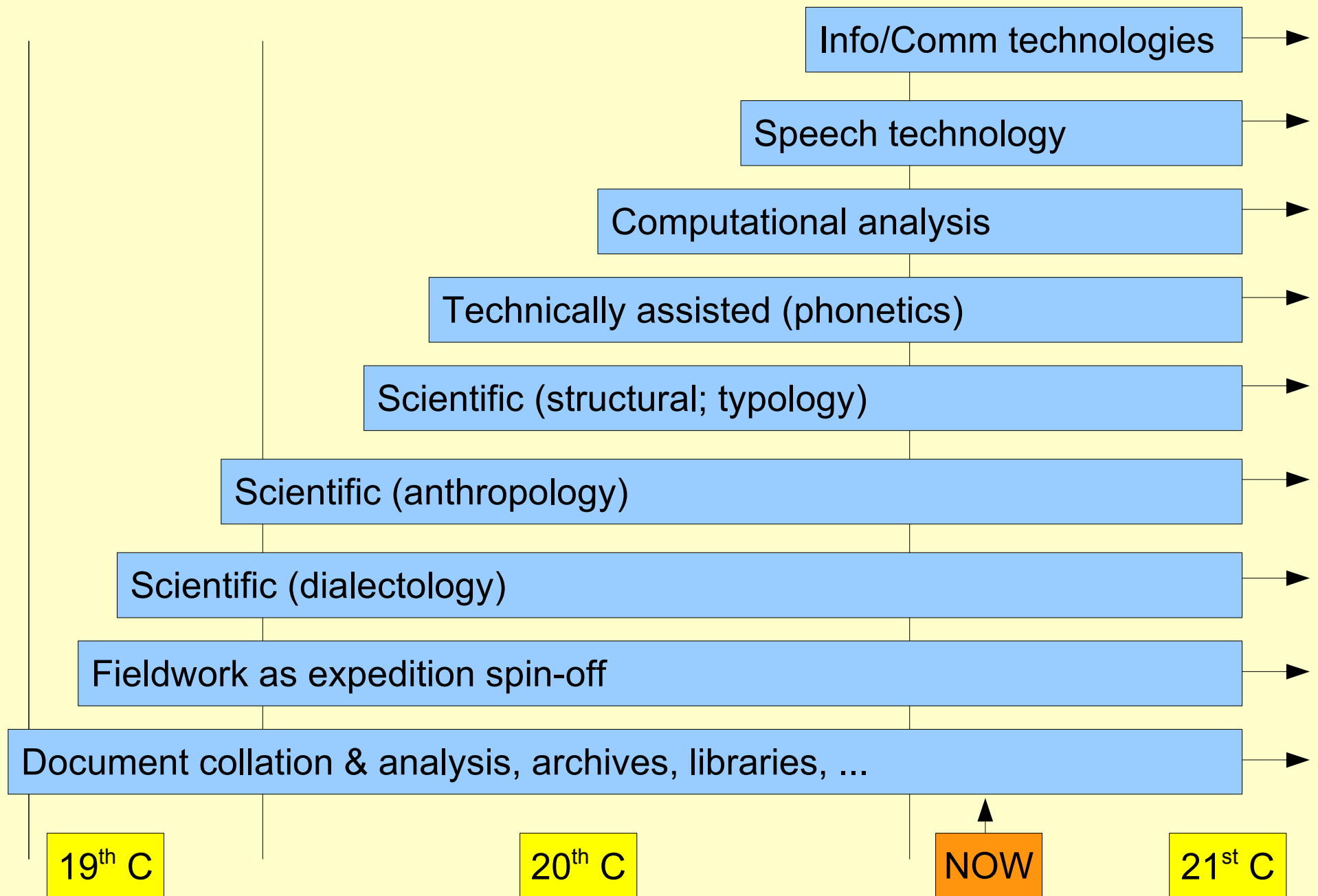
- Criteria for local language resources, tools and systems - CESAFA:
 - Comprehensive (with respect to application domain)
 - Effective (in terms of human and computing resources)
 - State-of-the-art (intellectually, not necessarily the latest internet-dependent software and hardware)
 - Affordable (for 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 (OAI, OLAC)
 - open software
 - generic (operating systems; office; archive formatting & display)
 - specific (speech and text analysis & lexicography tools)

COOPERATIVE RESOURCE CREATION

Cooperative resource creation

- No longer (if it still exists) the
 - HEROIC LONE FIELDWORKER MODEL
 - fieldworker – community
- Multi-level cooperative models:
 - TRANSDISCIPLINARY MODELS
 - linguistics (descriptive; corpus linguistics)
 - phonetics
 - speech technology
 - text technology
 - computer science / computational linguistics
 - TRANSNATIONAL INFRASTRUCTURAL MODELS
 - linguistics student – local linguistics student
 - linguistics department – local/regional linguistics department
 - funding organisation – regional funding organisation
 - political institution – regional political institution

Development of corpus resource methods



Transdisciplinary shared goals for resources

- Wide range of shared goals of
 - Speech and language processing R&D and
 - Computational Documentary Linguistics:
- Examples:
 - Development of
 - necessary resources
 - Human Language Technology systems
 - for research-oriented problem-solving in
 - empirically based linguistic theory development
 - communication infrastructure
 - language learning
 - language and cultural heritage documentation ...
 - for practical applications in the areas of
 - education
 - health services
 - trade ...

Transdisciplinary R&D context

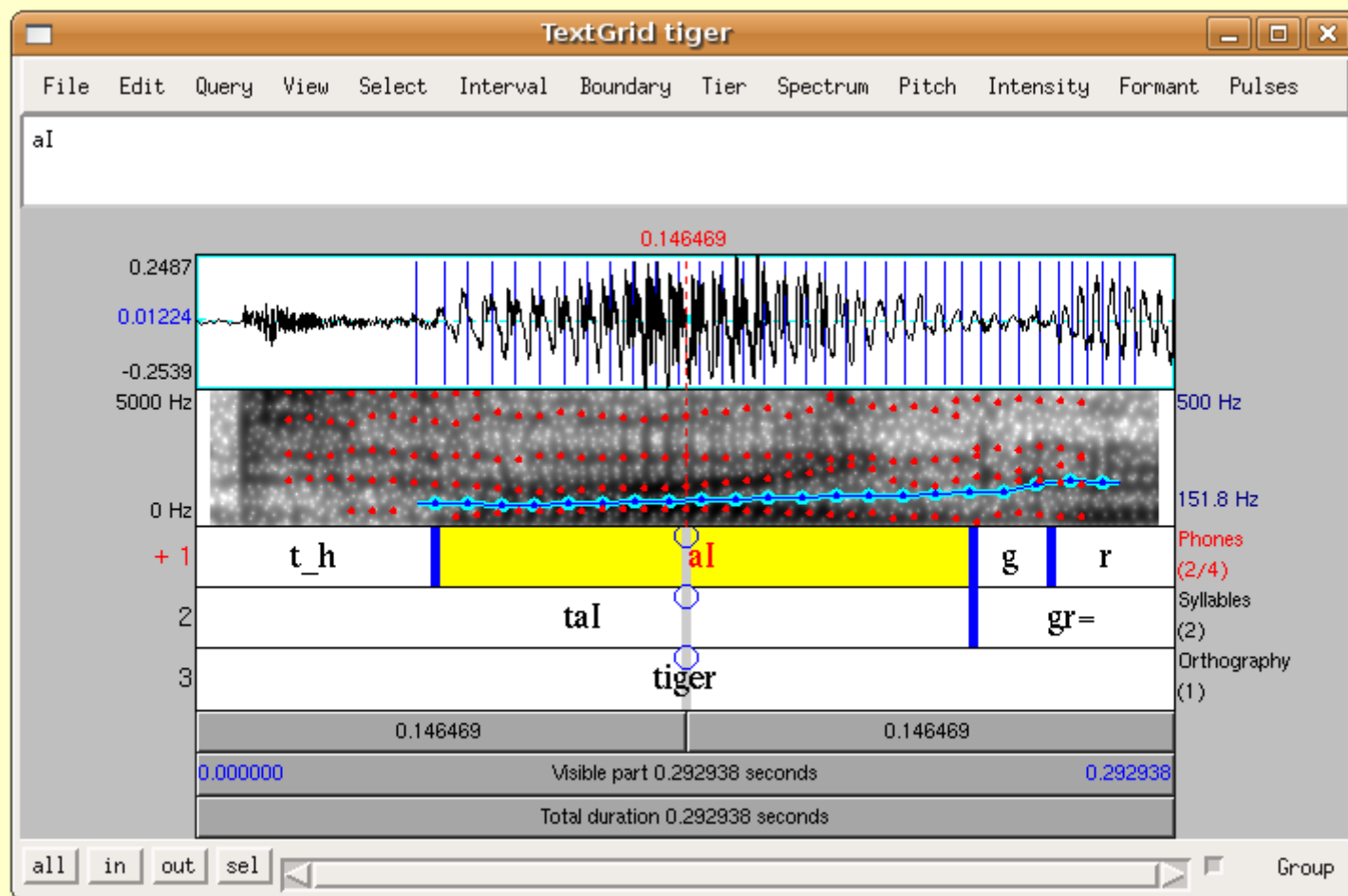
- Which disciplines are involved in this cooperation?
 - Linguistics and phonetics:
 - fieldworkers
 - text technologists (archiving, search)
 - phoneticians:
 - phonetic, phonemic, prosodic analyses
 - linguists:
 - word, sentence, text, dialogue grammars
 - semantic and pragmatic world + user models
 - Computer science
 - software engineers
 - text and speech pattern recognition specialists
 - Human-Computer Interface specialists

Speech resources: technology AND linguistics

Cooperation with phoneticians

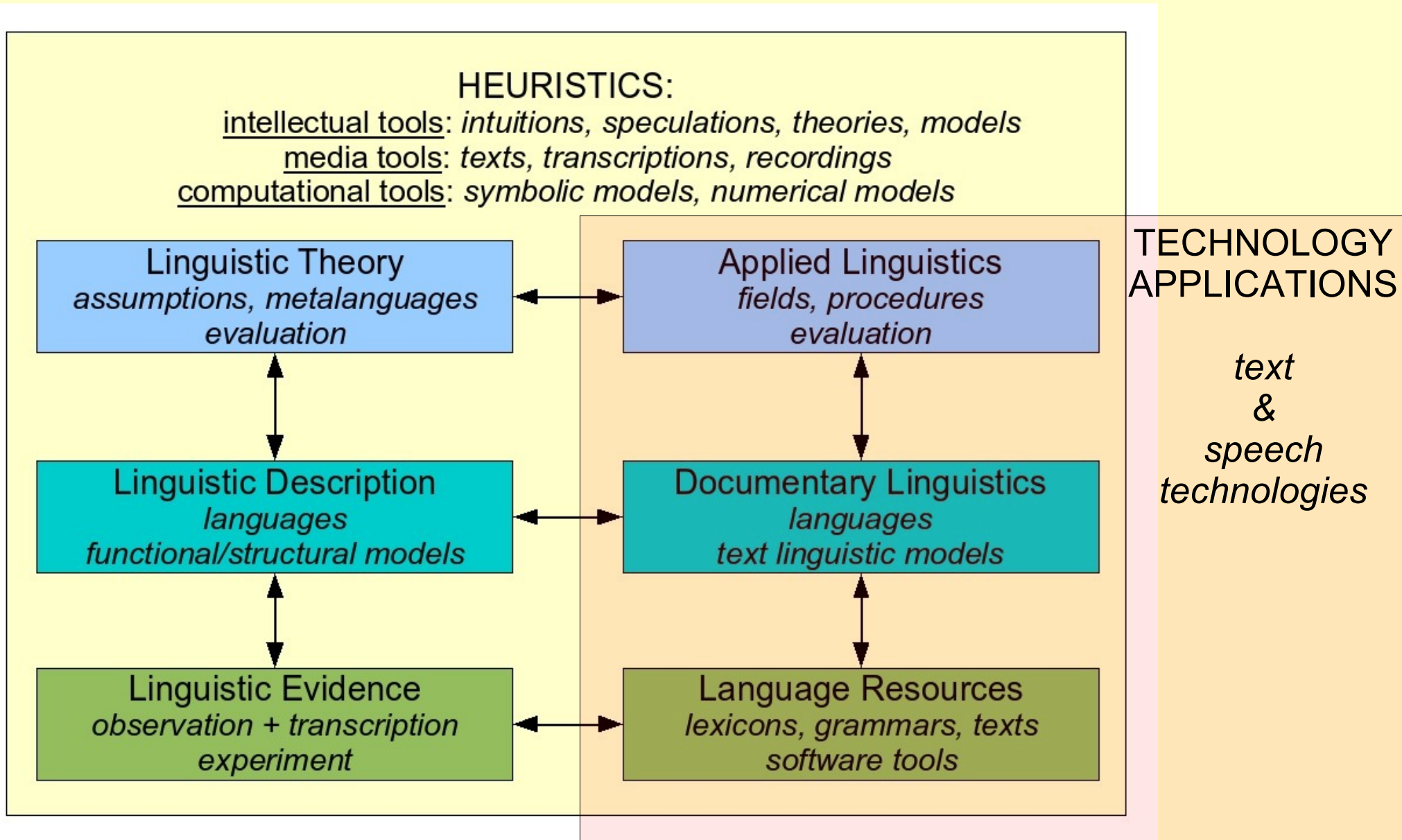
waveform,
spectrogram,
formants, pitch track:

time-aligned transcription
(annotation, labelling):

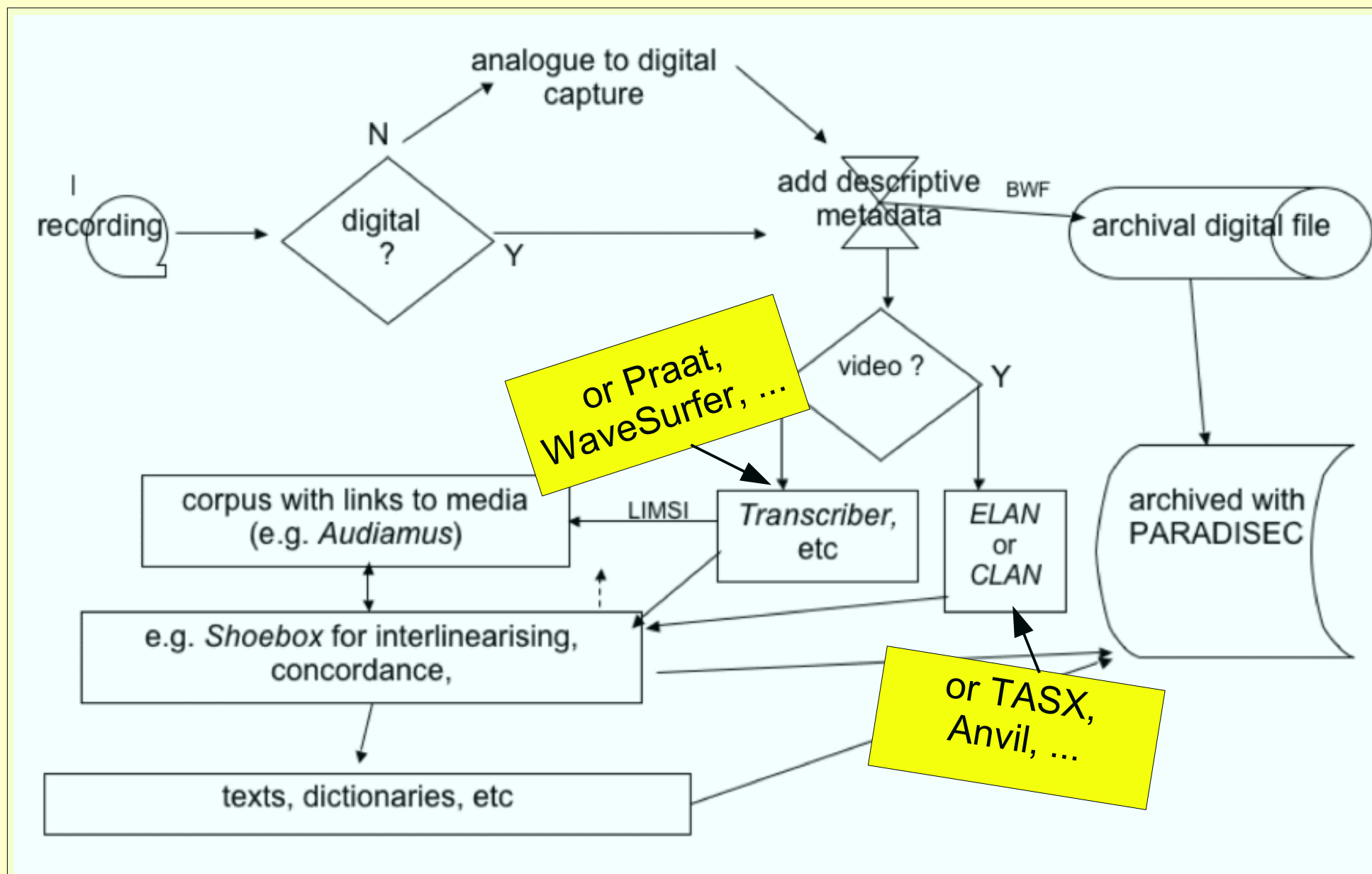


- The signal + annotation contains all the information required for corpus input to speech recognition and automatic speech synthesis
- Praat has a scripting language for performing time-alignment automatically

Cooperation with Computational Linguistics

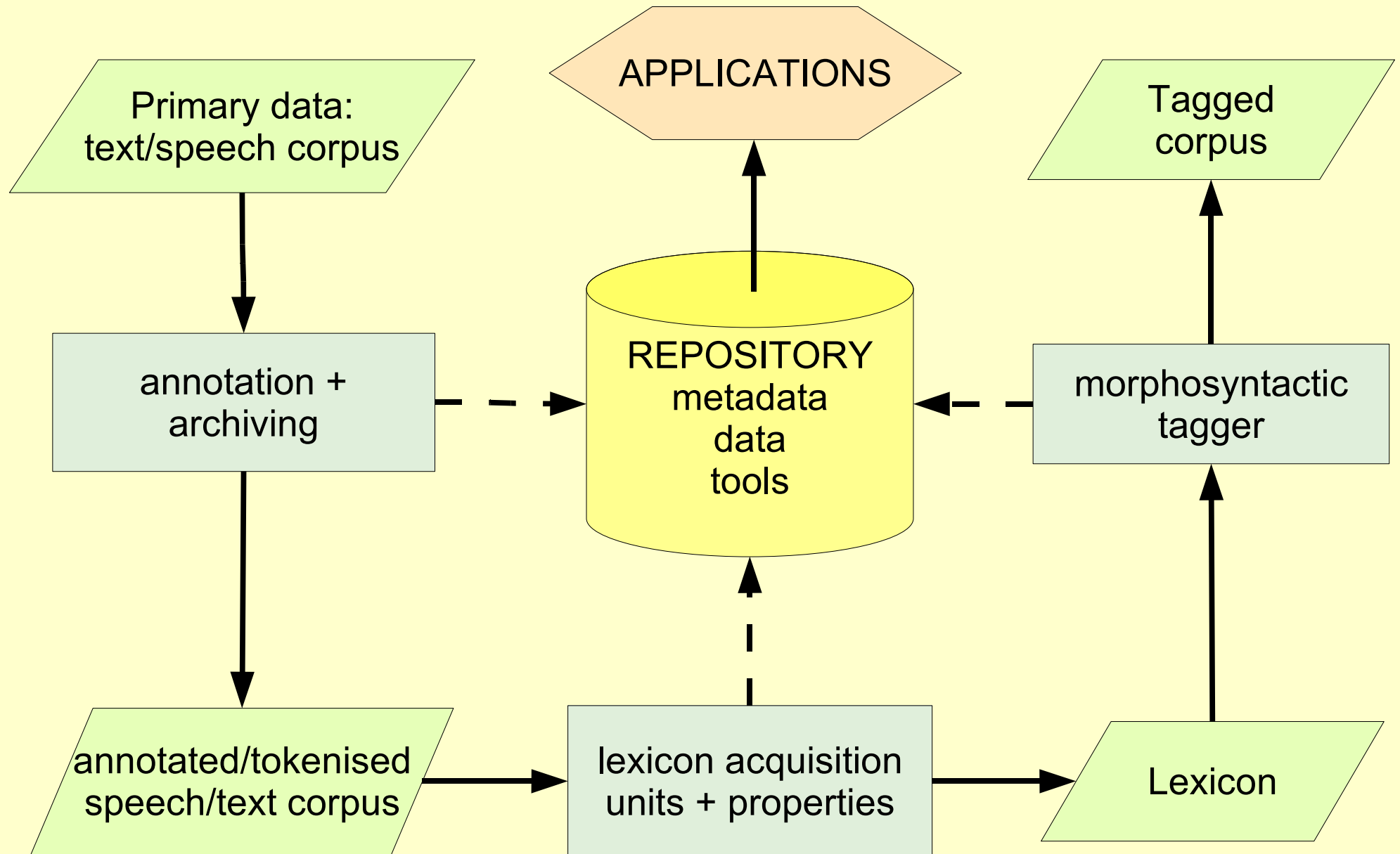


Current resource creation in doc ling



Nick Thieberger, David Nash (Australia)

Integrated computational resource development



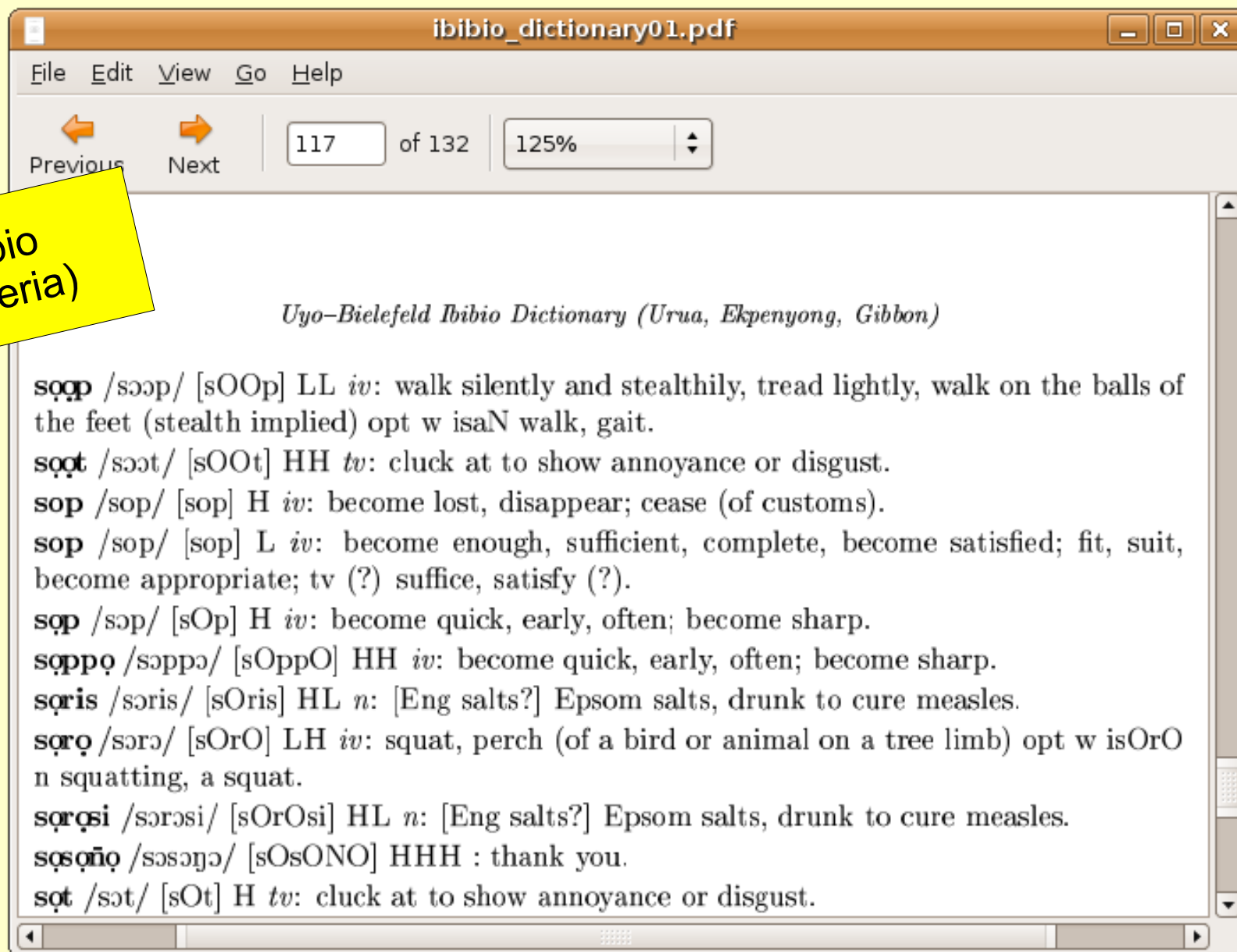
FOR EXAMPLE ...

Some transnational resource projects

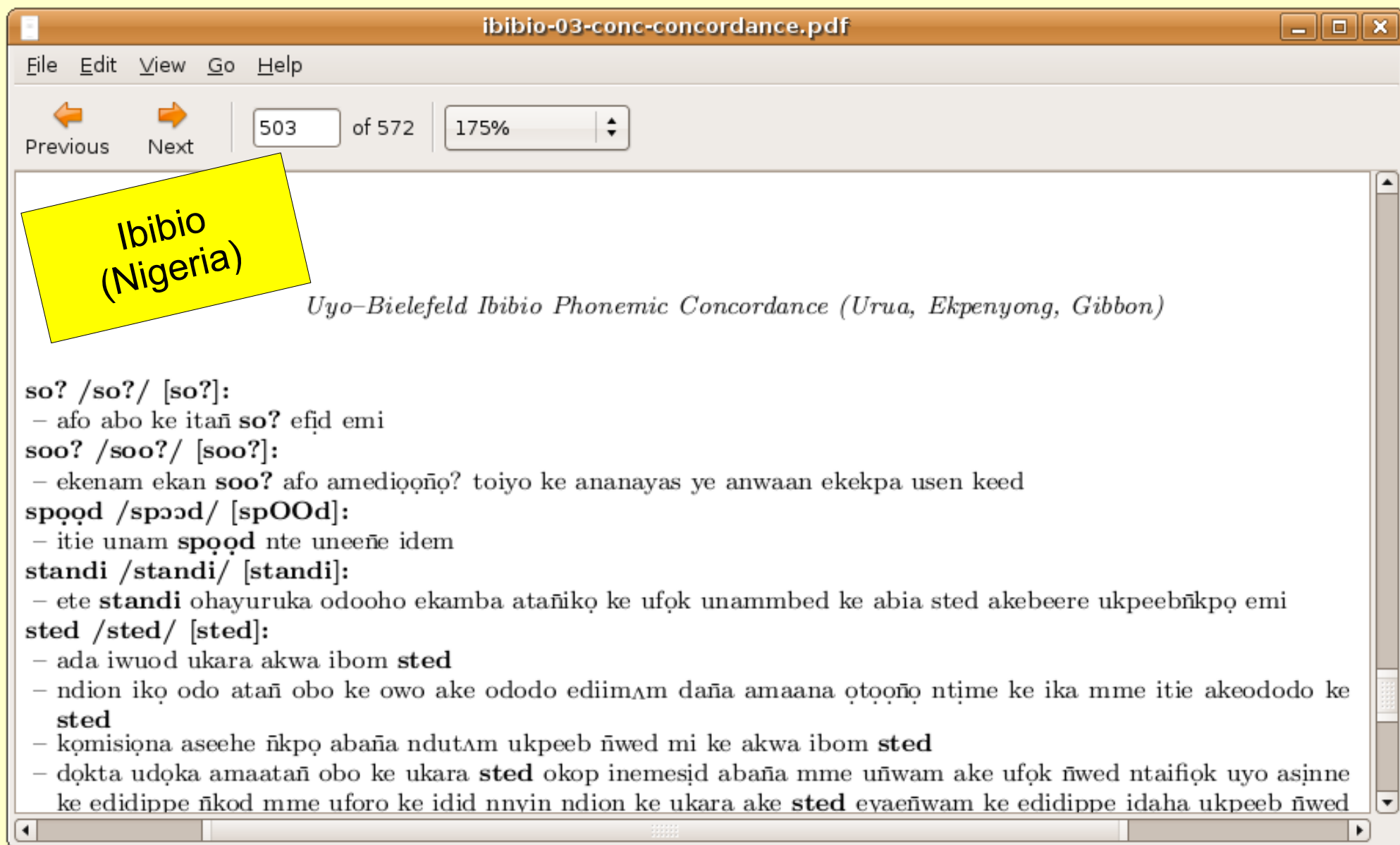
- DAAD funded international projects:
 - 1990s: Design for a new atlas of Ivory Coast languages
 - with Christian Lehmann
 - Université de Cocody, Côte d'Ivoire
 - 2001-2005: Development of an MA curriculum for Computational Language Documentation
 - Université de Cocody, Abidjan, Côte d'Ivoire
 - University of Uyo, Akwa Ibom State, Nigeria
 - 2002: DoBeS pilot project: EGA: A Documentation Model for an endangered Ivorian Language.
 - 2002-2003: Data Mining on Large Spoken Language Corpora
 - University of Campinas, Brazil
- Outside Echo funded international project:
 - 2002-2003: Speech Synthesis for Ibibio
 - University of Uyo, Akwa Ibom State, Nigeria
 - also: Nairobi, Johannesburg, Hyderabad partners

Cooperation with CL: lexical databases

Ibibio
(Nigeria)



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?** efið emi

soo? /soo?/ [soo?]:

- ekenam ekan **soo?** afo amediõõñõ? toiyõ ke ananayas ye anwaan ekekpa usen keed

spõd /spõd/ [spOOd]:

- itie unam **spõd** nte uneeñe idem

standi /standi/ [standi]:

- ete **standi** ohayuruka odooho ekamba atañikõ ke ufõk unammbed ke abia sted akebeere ukpeebñkpõ emi

sted /sted/ [sted]:

- ada iwuod ukara akwa ibom **sted**
- ndion ikõ odo atañ obo ke owo ake ododo ediimam daña amaana õtõõñõ ntĩme ke ika mme itie akeododo ke **sted**
- kõmisiõna aseehe ñkpõ abaña ndutam ukpeeb ñwed mi ke akwa ibom **sted**
- dõkta udõka amaatañ obo ke ukara **sted** okop inemesid abaña mme uñwam ake ufõk ñwed ntaifiõk uyo asĩne ke edidippe ñkod mme uforo ke idid nnyin ndion ke ukara ake **sted** evaeñwam ke edidippe idaha ukpeeb ñwed

Some results of cooperation

- A lexical database 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)
- Continuation:
 - the MA course has aroused the interest of UNESCO
 - the speech technology work and the educational work have influence the establishment of
 - Chair in Documentary Linguistics, Johannesburg
 - PhD course, Addis Ababa

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

Where can we go from here?

- 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)
 - SPICE (Tanja Schulz):
 - Web-based, any language, automatic
 - Text to speech synthesis & Automatic Speech Recognition
 - Link:
 - <http://csl.ira.uka.de/index.php?id=29&L=1>

Recommendations

- Cooperate with HLT 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)
 - transdisciplinary (relevant departments)
 - because our best resources are colleagues and students – wherever they may be ...

THANKS!