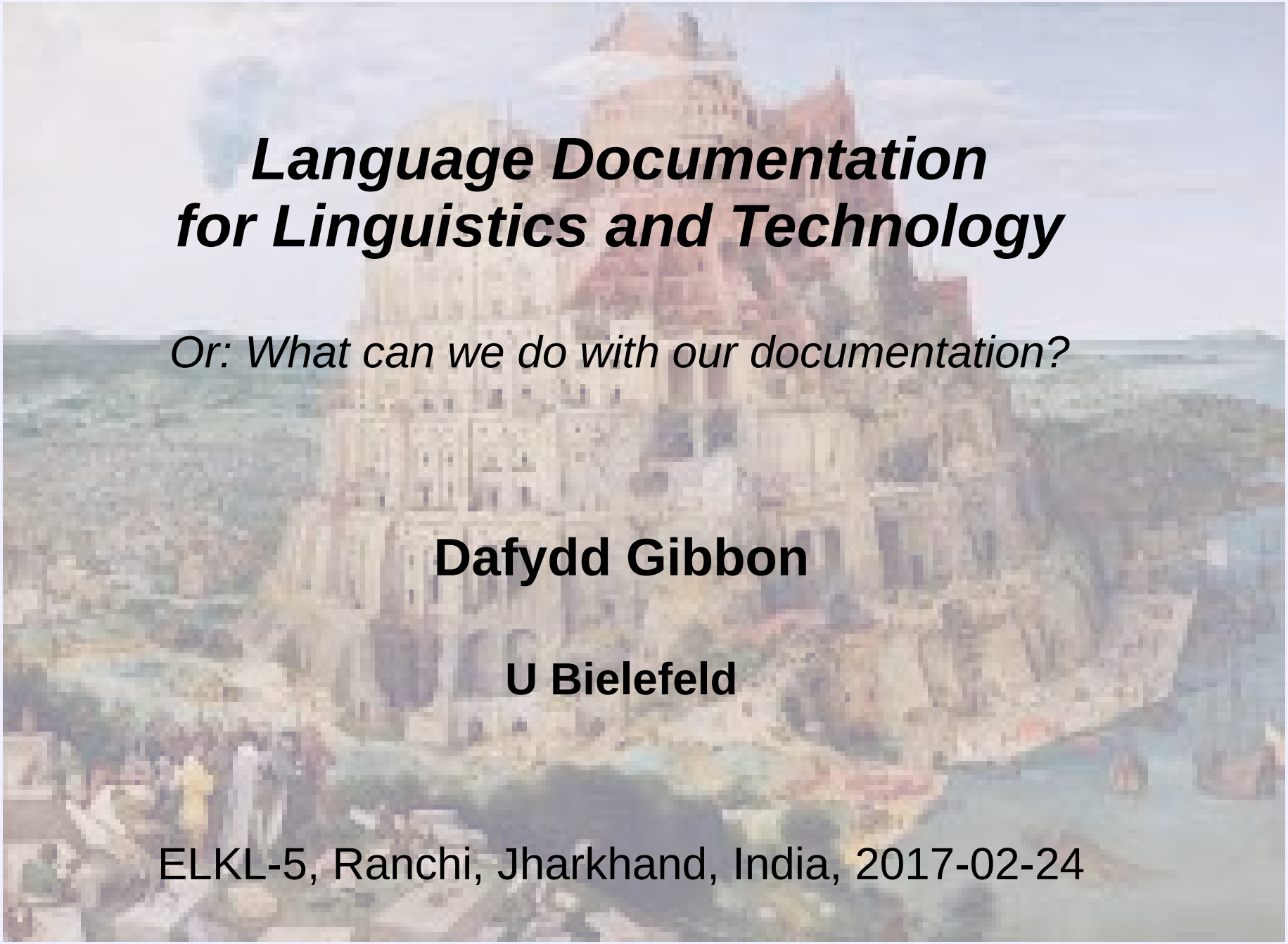


Language Documentation for Linguistics and Technology

Or: What can we do with our documentation?

Dafydd Gibbon

U Bielefeld

ELKL-5, Ranchi, Jharkhand, India, 2017-02-24

Focus: role reversal

Not just:

What can the Human Language Technologies <u>offer</u> to endangered and less resourced languages?

But:

What can Human Language Engineering <u>learn</u> from endangered and less resourced languages?

And:

What does documentation of endangered and less resourced languages <u>require</u> from the Human Language Technologies and their data and tool resources?

Roles for computational technologies in language documentation / language resources

1. **Documentation technologies**
2. **Enabling technologies**
3. **Productivity technologies**

There is a rapidly growing number of language learning and documentation apps for many languages on the various smartphone and tablet app stores – and some kinds are easy to make!

In Africa: from Amharic and Bambara to Yoruba and Zulu
Specifically in Nigeria: Yoruba, Hausa, Igbo, Ibibio, ...

In India there is a huge amount of relevant work going on with national, regional and local languages.

Project planning tools

Data collection tools

- scenario support for
  - elicitation
  - recording (multimodal) and annotating
  - metadata collection
- document scanning, OCRing and annotating

Data archiving and access

- standardized database and search models
  - relational, object-oriented, ...

Multilinear annotation

- for search, re-use analysis, application:
  - sharable (sustainable, interoperable) standards
  - annotation categories for phonetics, grammar, discourse, …
  - semi-automatic annotation methods

Resource construction tools

- phonetic analysis
- lexicon induction from data
  - word lists
  - word frequency lists (and other word statistics)
  - concordances
  - collocations
- grammar induction from data
  - Part of Speech (POS) tagging
  - grammar induction
  - parsing and generation
- translation
  - multilingual dictionaries
  - terminologies
  - processing of parallel or comparable texts
  - translator's workbench

Recognition techniques

- Automatic Speech Recognition (ASR)
- Visual scene and object recognition
- Information retrieval from text

Identification techniques

- Speaker identification
- Language identification
- Authorship attribution

Generation techniques

- Text-to-Speech Synthesis (TTS)
- Written text geneneration from databases

Products

- Dictation and information applications
- Translation applications

Some of the language Technologies

Engineering:

- **speech**: ASR, TTS, speaker id / recognition, ...
- **language**: Natural Language Processing (NLP), NL parsing, Q&A, text mining, text classification, lexicon and grammar induction, machine translation …
- **multimodal**: speech I/0 (dictation, process control, speech computer UI), speech avatars (Siri, Cortana), gesture (touchpad, waving), biometric systems

Computational linguistics & Computer Science:

- **domain models** of natural language syntax, semantics, pragmatics, language typology and genesis
- **formalisms and algorithms** for induction, parsing, generation of language
- **corpus analysis** for lexicon and grammar induction

A selection of definitions

Preliminary definitions

The terminology is specific to disciplines:

- In linguistics: *language documentation*
- In the language technologies: *language resources*

Preliminary definitions:

- documentation as *result*, *outcome* or *product*
  - Corpora of text inscriptions and speech recordings, plus metadata and basic description (transcriptions, translations and annotations of basic categories and spatial or temporal structure).
- documentation as *activity, workflow* or *process*
  - Use of standardised tools, research methodology, data formats, testing procedures for creating documentation products

The *product* model is useful because it suggests that there are *uses* and *users* of documentation for a *purpose.*

Quotes on "language resources"

Resources in speech technology: "appropriate infrastructure in terms of standardised tools, research methodology, data formats, testing procedures"

Gibbon, Dafydd, Roger Moore, Richard Winski, eds. (1997). Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter
http://wwwhomes.uni-bielefeld.de/gibbon/Handbooks/gibbon_handbook_1997/

"Language resources are the collective materials used by those engaged in language-related education, research and technology development. Spanning data collections, corpora, software, research papers and specifications, these vital tools aid and inspire scientific progress."

Linguistic Data Consortium
https://www.ldc.upenn.edu/language-resources

Quotes on "language documentation"

Language documentation (also known by the term 'documentary linguistics') is the subfield of linguistics that is 'concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties' (Himmelmann 2006:v)

A similar definition is given by Woodbury (2010) as 'the creation, annotation, preservation, and dissemination of transparent records of a language'.

Language documentation is by its nature multidisciplinary, and as Woodbury (2010) notes, it draws on 'concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts, and more' (see Harrison 2005, Coelho 2005, Eisenbeiss 2005 for examples).

Peter K. Austin (2010). Current issues in language documentation.

In Peter K. Austin (ed.) *Language Documentation and Description*, vol 7. London: SOAS. pp. 12-33

Language documentation / resources are required for ...

In the language technologies, statistical methods are dominant so huge data resources are needed (e.g. billions of words daily):

Search engines *(cf. Google, Bing, Baidu, ...)*

100% of major players are trained and probabilistic. Their operation cannot be described by a simple function.

Speech recognition *(cf. Siri, Cortana, Alexa, Google)*

100% of major systems are trained and probabilistic, mostly relying on probabilistic hidden Markov models.

Machine translation *(cf. Google, Bing, Baidu)*

100% of top competitors in competitions such as NIST use statistical methods. Some commercial systems use a hybrid of trained and rule-based approaches. Of the 4000 language pairs covered by machine translation systems, a statistical system is by far the best for every pair except Japanese-English, where the top statistical system is roughly equal to the top hybrid system.

Question answering *(cf. Siri, Cortana, Alexa, Google)*

this application is less well-developed, and many systems build heavily on the statistical and probabilistic approach used by search engines. The IBM Watson system that recently won on Jeopardy is thoroughly probabilistic and trained, while Boris Katz's START is a hybrid. All systems use at least some statistical techniques.

Source: Peter Norvig, norvig.com

From documentation-as-a-product to a products

Bing MT:	Greetings!	नमस्ते!
Google MT:	Greetings!	घणी खम्मां!

Speech recognition
Speech synthesis (e.g. screen readers)
Information Systems
All are heavy 'big data' resource users

From documentation-as-a-product to a products

Bing MT: Greetings!

Google MT: Greetings!

Up to you to evaluate!

नमस्ते!

घणी खम्मां!

Up to you to evaluate!

Up to you to evaluate!

Speech recognition
Speech synthesis
Information Systems:
heavy data resource users

Generic Tools

From documentation-as-a-product to documentation products

Aikuma2

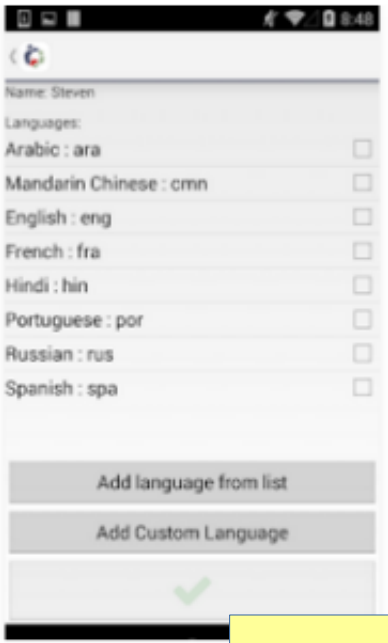
LP20 Education

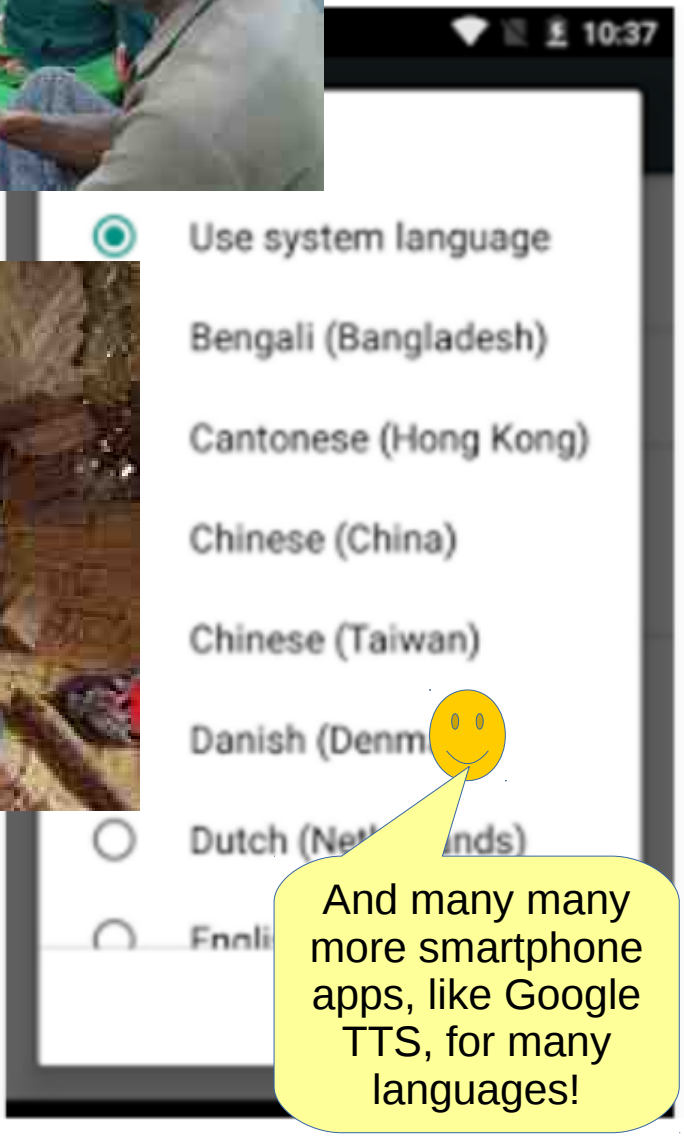
Unrated

This app is compatib

The Hyderabad Simputer CDA

Steve Bird and his Android fieldwork app "Aikuma"

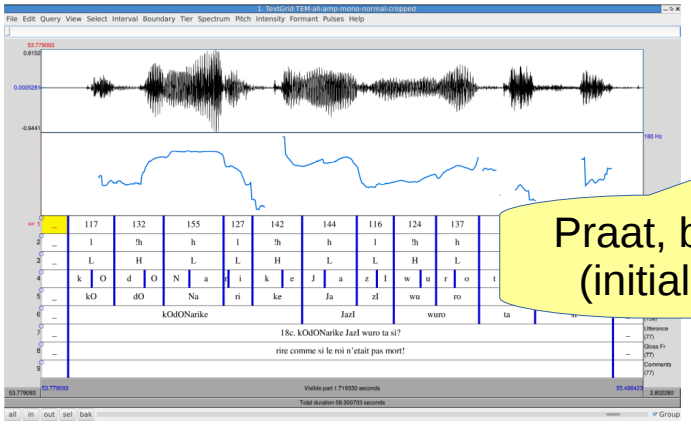




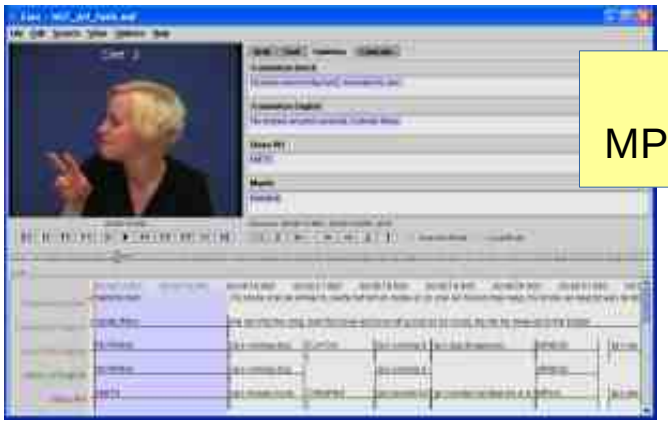
The meeting of the disciplines: Aikuma, Simputer, Text-to-Speech

And many many more smartphone apps, like Google TTS, for many languages!

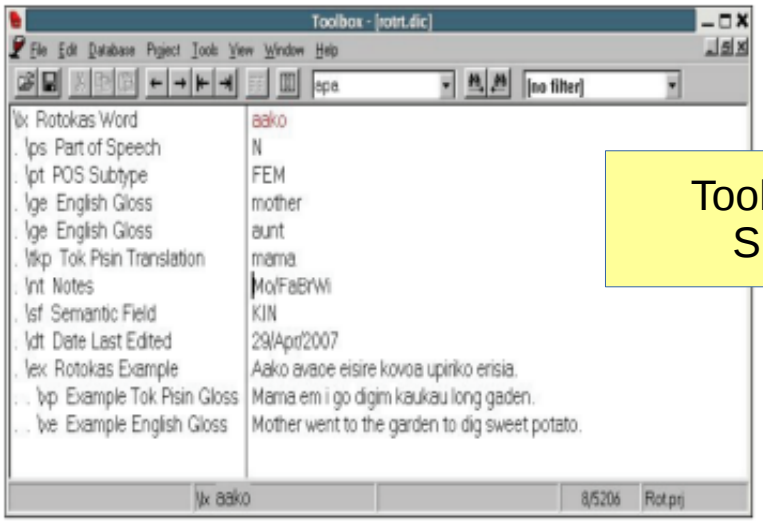
Generic tools for documentation-as-a-process



Praat, by Paul Boersma & David Weenink
(initially a speech technology resource)



ELAN
MPI Nijmegen



Toolbox
SIL

SPPAS
Brigitte Bigi
Aix-en-Provence

What is SPPAS?

SPPAS is a **multi-platform and public** annotation software tool. It is able to produce automatically speech annotations from a recorded speech sound and its orthographic transcription. Some special features are also offered for the analysis of any kind of annotated files.

SPPAS is compatible with Praat, Elan, Transcriber, Annotation Pro, Phonedit, and many others...

Custom Tools

Purpose-designed tools for documentation-as-a-process - text

Linguistics, particularly computational linguistics & Natural Language Processing (NLP), with applications in language and speech technologies:

- Word sense disambiguation:

  100% of top competitors at the SemEval-2 competition used statistical techniques; most are probabilistic; some use a hybrid approach incorporating rules from sources such as Wordnet.

- Coreference resolution:

  The majority of current systems are statistical, although we should mention the system of Haghighi and Klein, which can be described as a hybrid system that is mostly rule-based rather than trained, and performs on par with top statistical systems.

- Part of speech tagging:

  Most current systems are statistical. The Brill tagger stands out as a successful hybrid system: it learns a set of deterministic rules from statistical data.

- Parsing:

  There are many parsing systems, using multiple approaches. Almost all of the most successful are statistical, and the majority are probabilistic (with a substantial minority of deterministic parsers).

Source: Peter Norvig, norvig.com

Purpose-designed tools for documentation-as-a-process - speech

PV – Prosody Visualizer

Comparison: Language Documentation – Language Resources

Comparison of DocLing and LangTech scenarios

- The documentary linguistic scenarios are:
  - rather individual, extremely heterogeneous
  - rather hard to define and delimit
  - *de facto* standards: Praat, ELAN, Wordsmith, TypeCraft,...
  - somewhat *ad hoc* - 'what you can get'

- Language, speech, multimodal technology scenarios are:
  - highly standardised, rather coherent
  - tendentially easy to define and delimit
  - very application / product oriented
    - especially in speech technology: highly product specific
    - text technology is more generic
  - regulated standards:
    - statistical evaluation procedures
    - institutional standards (e.g. ISO)

Documentary linguistics

Language Documentation

- texts, audio, video corpora
- dictionaries
- language structure
- language context

Motivation

- heritage preservation
- education
- linguistic insight
- ethics

Methods

- data collection
- categorial description
- tools

Language technologies

Language resources

- text, audio, video corpora
- dictionaries, wordnets
- language models
- language scenarios

Objectives

- system development
- software applications
- new algorithms
- marketing

Methods

- data collection
- statistical modelling
- tools

Language documentation

Structure

- structural grammars
- functional grammars
- semantics
- pragmatics

Context

- genres
- activities
- historical background

Tools

- Data manipulation
  - Praat, Typecraft, ...
- Database management
  - corpora
  - dictionaries
- Repositories
  - databases

Language Resources

Speech/language models

- statistical models
- stochastic grammars
- databases
- user interaction

Scenarios

- speech, text, modalities
- task orientation
- product innovation

Tools

- Data manipulation
  - custom; Praat, ...
- Database management
  - corpora
  - dictionaries
- Repositories
  - databases

Motivation

Maintenance, revitalisation

- spoken language
- text
- culture

Social payback

- language teaching
- immersive teaching
- health, marketing

Linguistic insight

- language classification
- language typology
- language and cognition

Ethics

- identity
- taboo
- human rights

Objectives

System development

- speech, spoken language
- text
- multimodal

Software products

- speech recognition
- speech synthesis
- speaker recognition

Efficient algorithms

- Hidden Markov Models
- Neural Networks
- Machine Learning

Marketing

- branding
- customer satisfaction
- consumer regulations

Methods

Data collection

- interview
- fieldwork

Data description

- manual annotation
- dictionary
- sketch grammar

Tools

- annotation tools
- databases, repositories
- formats

Data collection

- experiment
- fieldwork

Data modelling

- automatic annotation
- dictionary
- speech & language models

Tools

- custom annotation tools
- databases, repositories
- formats

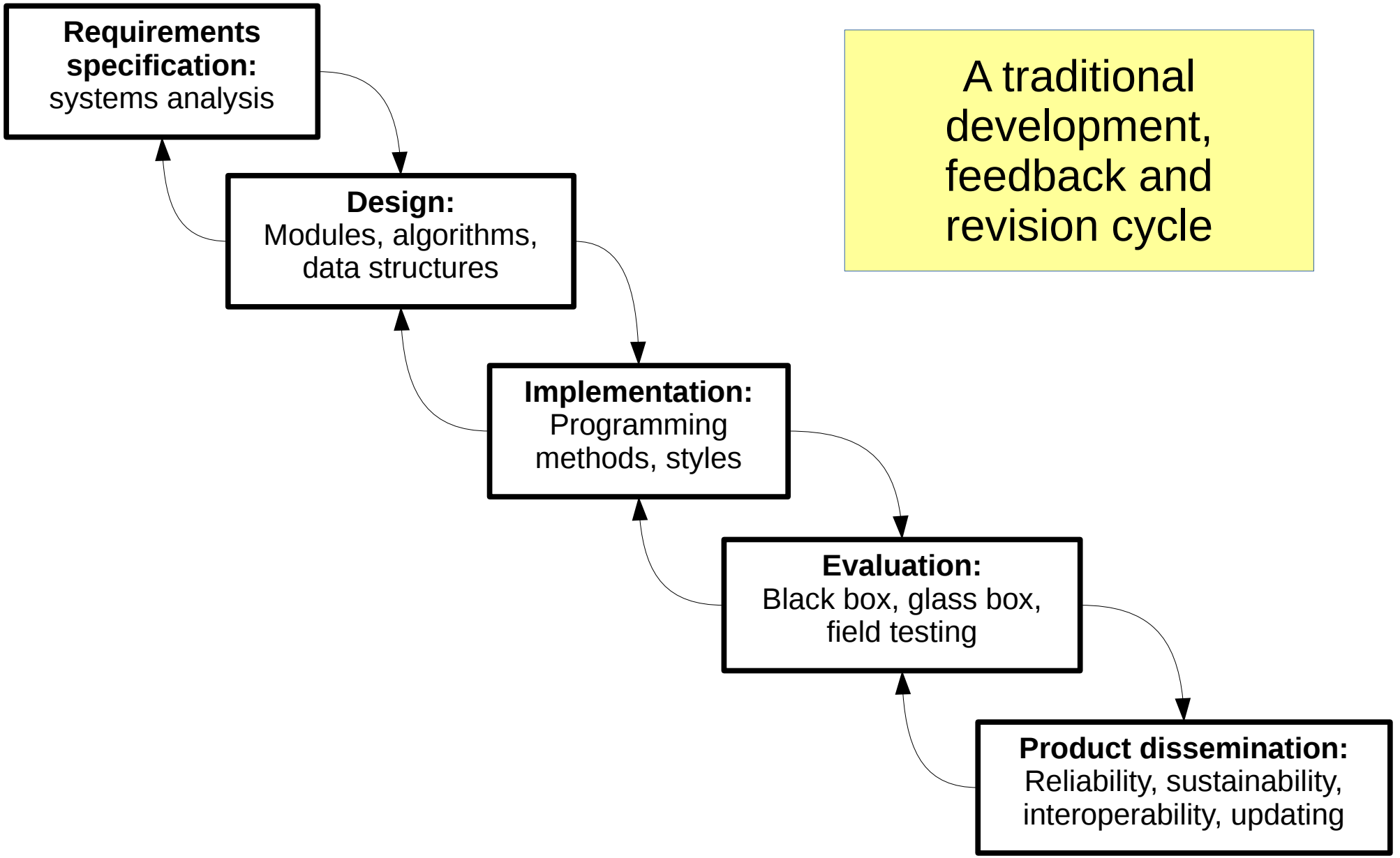
Available tools

- annotation tools
  - Praat, Elan, ...
  - Praat, Perl Shell scripting
- databases, repositories
  - Toolbox; MPI, SOAS, ...
- formats
    - *ad hoc*
    - XML, Unicode
    - IPA
    - novel orthography

Standardized tools

- custom annotation tools
  - (semi-)automatic
  - HMMs, Deep Learning
- databases, repositories
  - custom; LDC, ...
- formats
  - *custom*
  - XML, Unicode
  - SAMPA (IPA)
  - standard orthography

Applying technologies – a reminder



Endangered languages as teachers: some outlets

- Speech Assessment Methodologies (EC Project)
- EAGLES: Expert Advisory Groups for Language Engineering Standards
- Many other resources oriented European Projects, including
  - MATE
  - IMDI
  - ...
- LREC – Language Resources and Evaluation Conference
- Language Resources Map
  - https://en.wikipedia.org/wiki/LRE_Map
- Krauwer's BLARK: *Basic Language Resource Kit*

Endangered languages as teachers: some models

- Steven Krauwer's BLARK:
  - Goal of equal status of European languages
  - Generalisable to the world at large?
  - *Basic Language Resource Kit initial specification*
    - *written language corpora*
    - *spoken language corpora*
    - *mono- and bilingual dictionaries*
    - *terminology collections*
    - *grammars*
    - *modules (e.g. taggers, morphological analysers, parsers, speech recognisers, text-to-speech)*
    - *annotation standards and tools*
    - *corpus exploration and exploitation tools*
    - *bilingual corpora*
    - *etc*

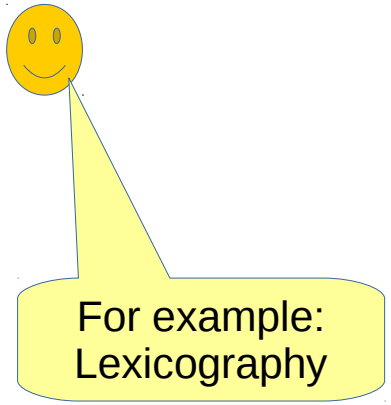
Some models: Basic Language Resource Kit (BLARK)

Modules	Data									Applications							
	monoling lex	multilin lex	thesauri	anno corp	unanno corp	speech corp	multi ling corp	multi mod corp	multi media cor	CALL	access control	speech input	speech output	dialog systems	doc prod	info access	transla-tion
Language Technology																	
Grapheme-phon. conv	++			++						+			++	++	+	+	
Token detection	++			+	++					+		+		+	+	+	+
Sent boundary detection	+			++	++					+		++	++	+	++	++	++
Name recognition	+	+	+	++	++	++				+		++	++	+	++	++	++
Spelling correction										+							
Lemmatising	++			++	+					+		+	+	+	+	+	+
Morphological analysis	++			++	+					+		+	++	+	++	++	++
Morphological synthesis	++			++	+					+			++	+	++		++
Word sort disambig.	++			++	+					+		++	+	++	++	++	++
Parsers and grammars	++			++						+		++	++	++	++	++	++
Shallow parsing	++			++	++					+		++	++	++	++	++	++
Constituent recognition	++			++	+					+		++	++	++	++	++	++
Semantic analysis	++		++	++				++	++	+		++	++	++		++	++
Referent resolution	+		++	++	+					+		++		++	++	++	++
Word meaning disambig.	+		++	++	+					+		++	+	+	+	++	++
Pragmatic analysis	+		+	++				++	++	+		++	++	++		+	++
Text generation	++		++	++				++	++	+			++	++	++		++
Lang. dep. translation		++	++	++			++			+						++	++

Some models: Basic Language Resource Kit (BLARK)

Modules	Data									Applications							
	monoling lex	multilin lex	thesauri	anno corp	unanno corp	speech corp	multi ling corp	multi mod corp	multi media cor	CALL	access control	speech input	speech output	dialog systems	doc prod	info access	transla-tion
Speech Technology																	
Complete speech recog.	++	+		++	+	++	+	++	++	++	++	++		++	++	++	++
Acoustic models	++	+		++	+	++	+	+	+	++	+	++		++	+	+	+
Language models	+			++	+	+	+	+	+	++	+	++		++	++	++	++
Pronunciation lexicon	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Robust speech recog.	+			+	+	+	+	+	++	+	+	++		++	+	+	+
Non-native speech recog.	+	++		+		++	++	+	+	++	+	+		+		+	+
Speaker adaptation	+			+	+	++	+	+	++	+	+	++		+	+	++	+
Lexicon adaptation	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Prosody recognition	+	+		++	+	++	+	+	+	++	+	++		++	++	++	++
Complete speech synth.	++	+		+		+		+		+			++	++	+	+	++
Allophone synthesis	+	+		+		+		+		+			+		+	+	+
Di-phone synthesis	++	+		+		+		+		+			++	++	+	+	+
Unit selection	++	+		+		+		+		+			++	++	+	+	+
Aut. phon. transcription	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Aut. phon. segmentation	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Phoneme alignment	+	+		+		++	+	+	+	++	+	+		+			+
Distance calc. phonemes	+	+		+		++	+	+	+	++	+	+		+			+
Speaker identification	+			++	++	++	+	++	+	+	++	+		+		+	+
Speaker verification	+			++	++	++	+	++		+	++	+		+		+	+
Speaker tracking	+			++		++			++	+	++	+		+	+	+	+
Language identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Dialect identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Confidence measures	+			+	+	++	+	++	+	++	++	++		++	+	+	+
Utterance verification	+			+	+	++	+	+	+	+	+	++		++	+	+	+

Documentation and Description: a Scale of Abstraction



LEXICOGRAPHIC COMPLEXITY

LEXICON

Fourth order lexicon:
– maximally declarative generalisation network

Third order lexicon:
– procedurally optimised local generalisations

Second order lexicon:
– flat tabular lexicon.

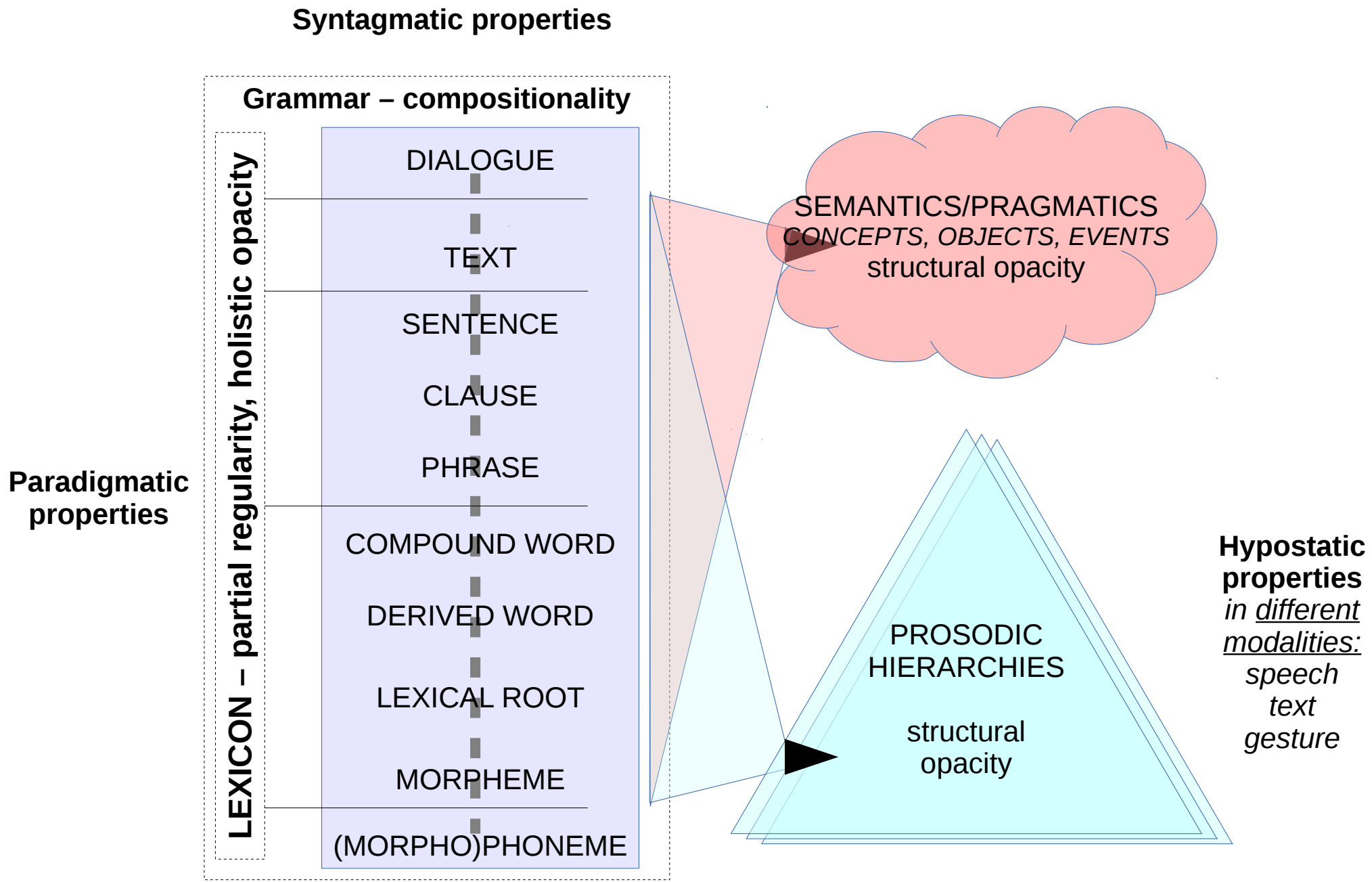
First order lexicon:
– wordlist, concordance, HMM

CORPUS

Secondary corpus:
– transcription, annotation

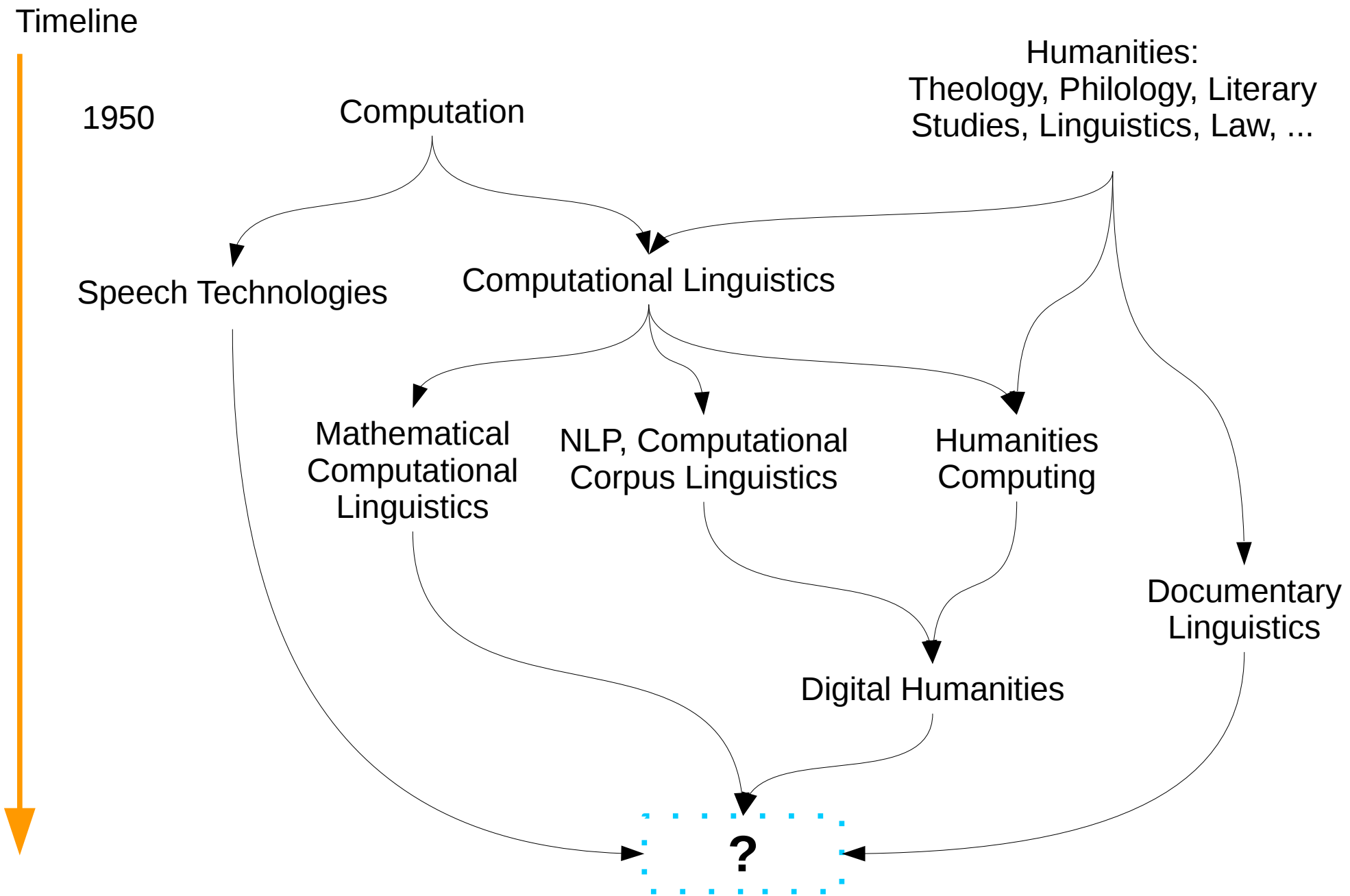
Primary corpus:
– recorded audio–visual corpus

An integrative model is needed: Rank-Interpretation Architecture



The broader interdisciplinary context

Doc Ling and Digital Humanities – a personal view of history



More on Enabling Technologies

Annotation and Annotation Mining

Language Similarity Analysis

Data Capture and Storage

Enabling technologies

Annotation, preferably (semi-)automatic

- associating text labels and time-stamps with speech recordings
- annotation mining (preferably automatic)
- information extraction from annotations:
  - text label list, text label frequencies, text label duration statistics
  - visualisation of text label duration patterns, rhythm patterns

Classification, similarity analysis

- e.g. virtual distance mapping
  - Which languages have been (almost) documented and can be easily related to already documented languages?
  - Geographic (areal contact)
  - Typological (paradigmatic and syntagmatic structural similarity)
  - Genealogical (history of language families)

Quasi-commercial applications

- e.g. text-to-speech synthesis for automatic indigenous information services

The case of Annotation

and Annotation Mining

for linguistics and technology

Why annotation? And how?

The primary reason for the annotation of text and speech data is to enable efficient search procedures

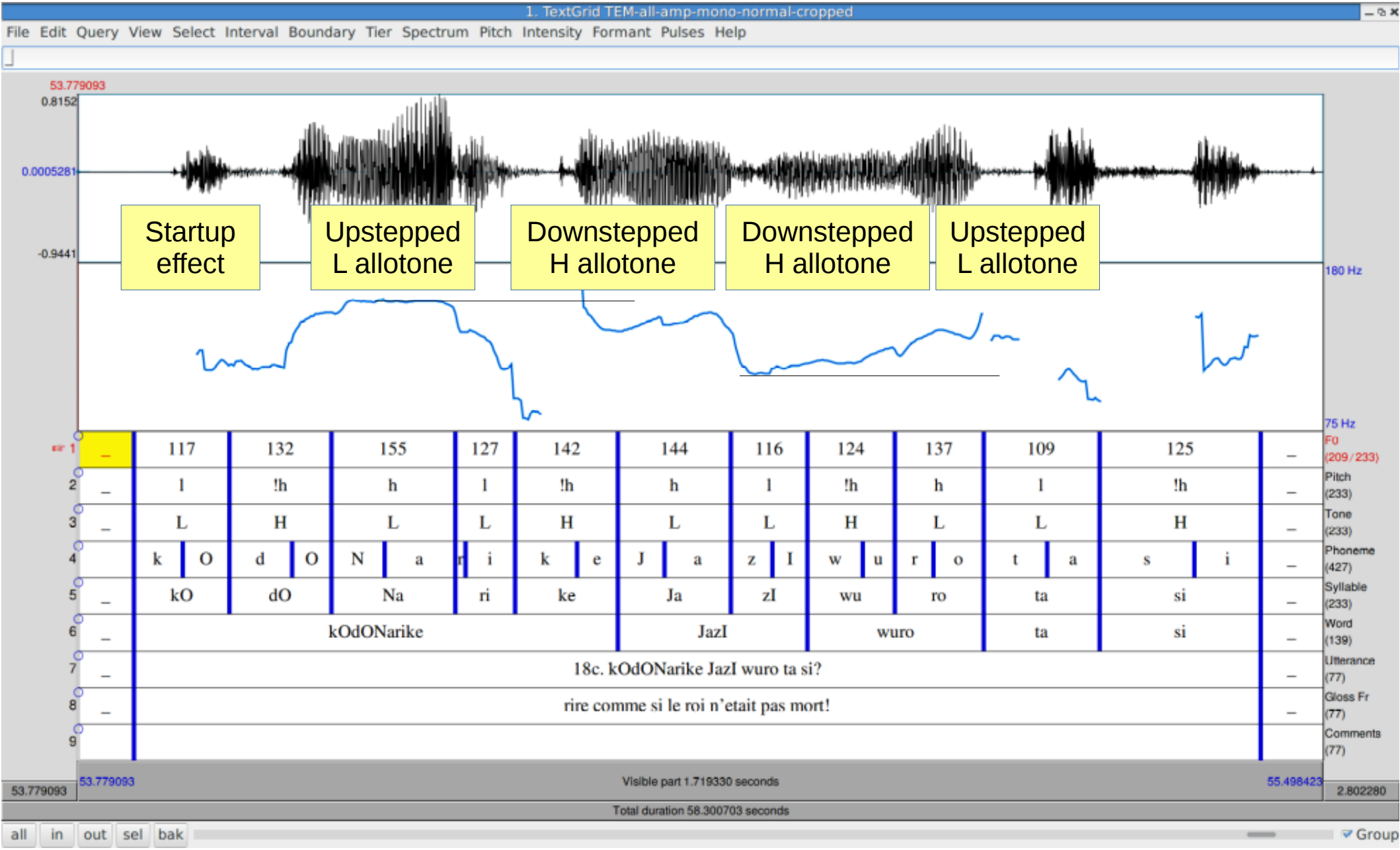
- to provide structure
- in order to make data systematically searchable
- by assigning perceptual/hermeneutic categories to data
- for the purposes of
  - finding archived media
  - linguistic and phonetic analysis
- development of speech and language systems

And searching unstructured data is difficult

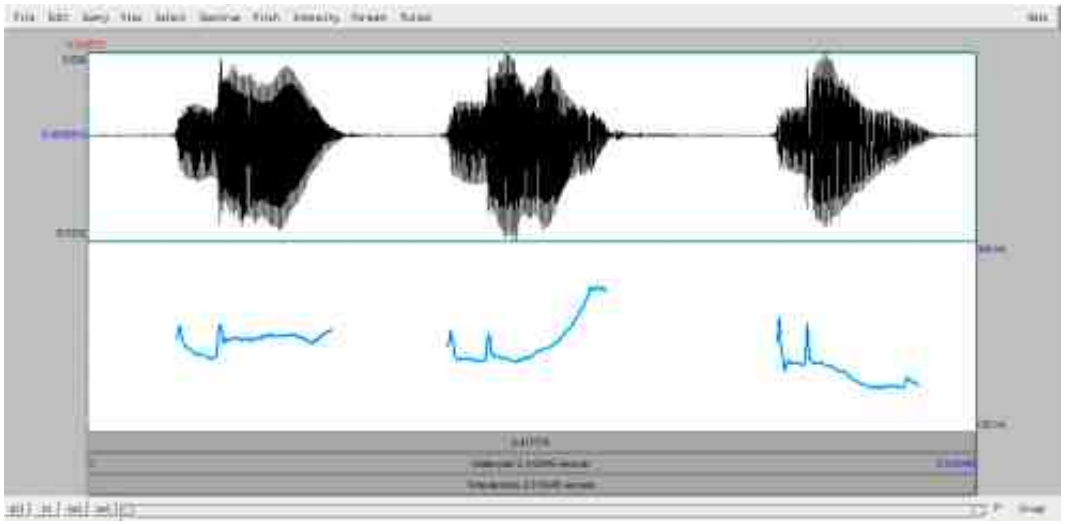
but improving with the help of machine learning

- example: search on free text data
  - Google Bing search as on-the-fly concordance construction from web data)
  - example: Google's image search

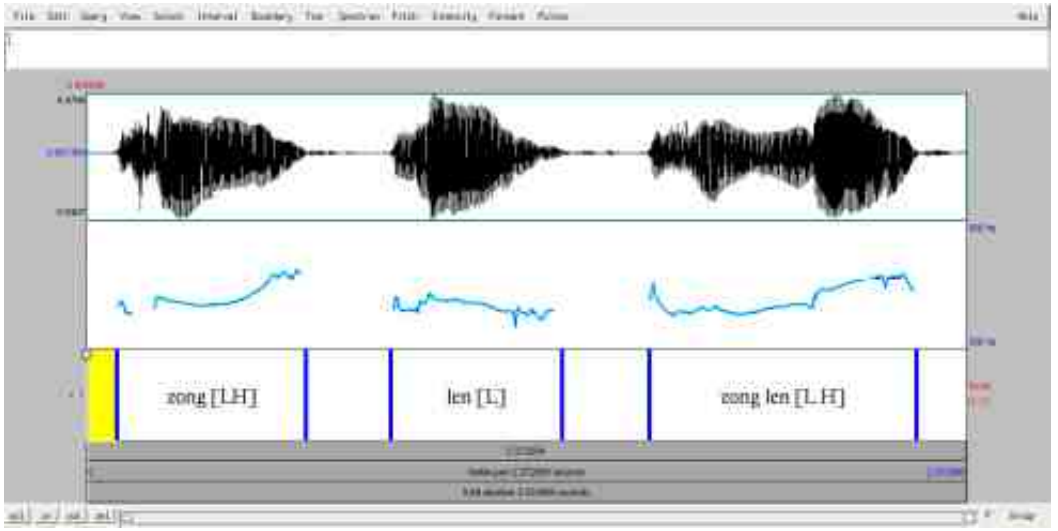
Annotation Mining: tone automaton induction for TTS



Tone: Kuki-Thadou

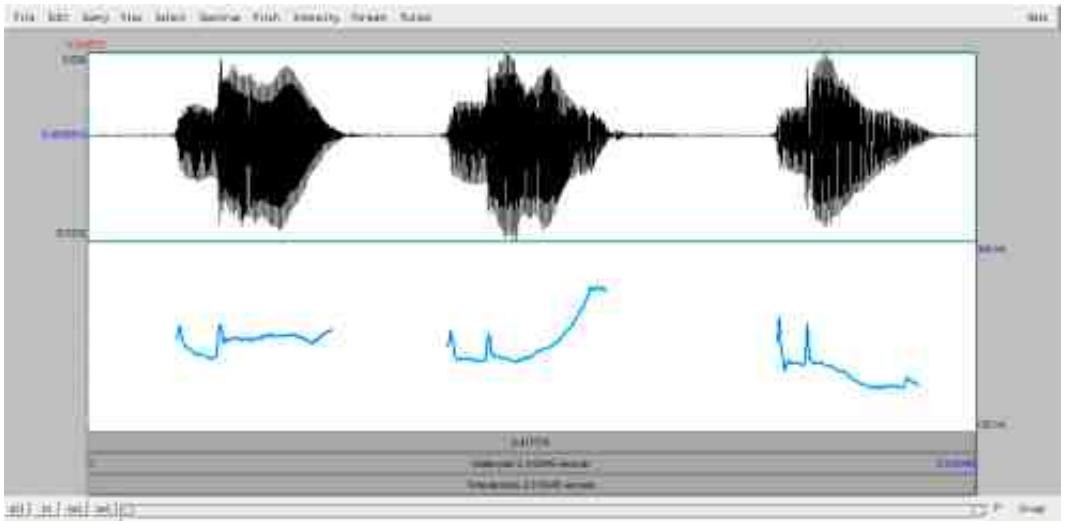


Thadou tones:
lów (H) ‘field’,
lŏw (LH) ‘medicine’,
lòw (L) ‘negative marker’.



LH *zŏng* ‘monkey’
L *lèn* ‘big’ tones in isolation
L+H *zòng lén* 'bit monkey' tone sequence
Note H tone shift and L deletion.

Tone: Kuki-Thadou



Thadou tones:
lów (H) ‘field’,
lŏw (LH) ‘medicine’,
lòw (L) ‘negative marker’.

The descriptive statistics are over averages of 16 pitch samples for each of 3 occurrences of each vowel with which each tone is associated (e.g. 864=18x16x43).
The values over all measurement sets per tone are in parentheses.

Tone	N	min	max	mean	sd	offset	slope
H	18 (864)	200 (220)	244 (222)	221	0.29	221	-0.03
LH	17 (816)	215 (198)	237 (268)	220	7.07	209	1.3
L	18 (864)	192 (178)	213 (227)	203	6.3	215	-1.31

The case of language classification:

Documentation beyond individual languages

Language similarity and difference as virtual distance

Aspects of Machine Learning for Language Documentation

Documentation of languages and language varieties

As for individual languages:

insights into and results concerning

- diversity vs. normalization of speech and language
- the intricacy, complexity of speech, language and languages
- similarities and differences between languages
  (typology, history, dispersion of language)
- scenario-dependent properties of languages
- gender, age, social role
- task orientation
- public vs. informal vs. intimate styles
- diversity of expression of emotion
- political status of languages wrt dominance, minorities

Comparative studies

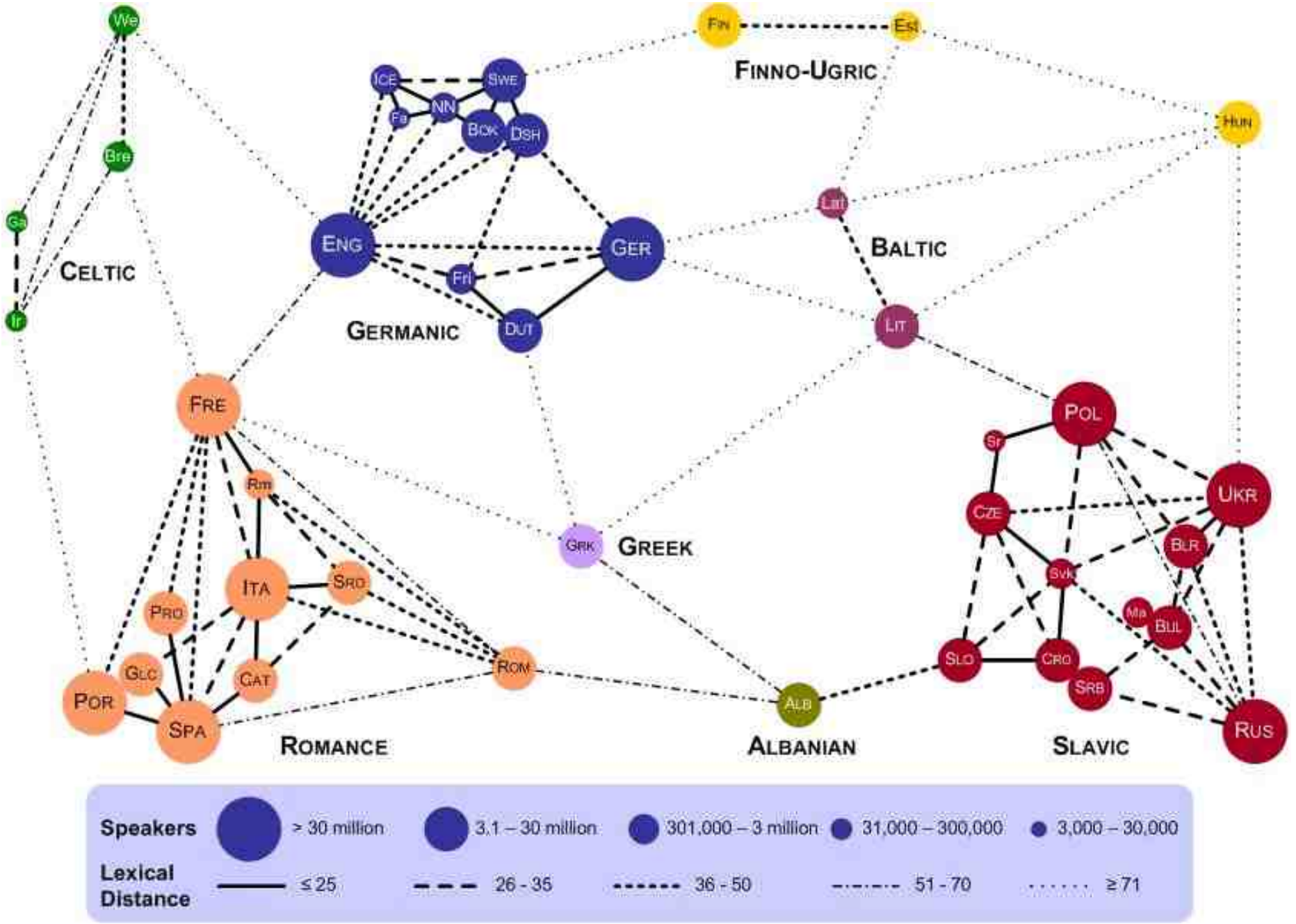
Language similarity

- for priority setting in language selection
  - for funding (usually: chance)
  - for adaptation from documentation of existing languages
- for
  - language history
  - language typology

Language typology:

- structural:
  - similarity/difference in speech sound systems
  - similarity/difference in grammar
  - similarity/difference in the lexicon
- functional
  - similarity/difference in discourse conventions
  - similarity/difference in general cultural conventions

How similar are languages? - A little similar, but not very similar

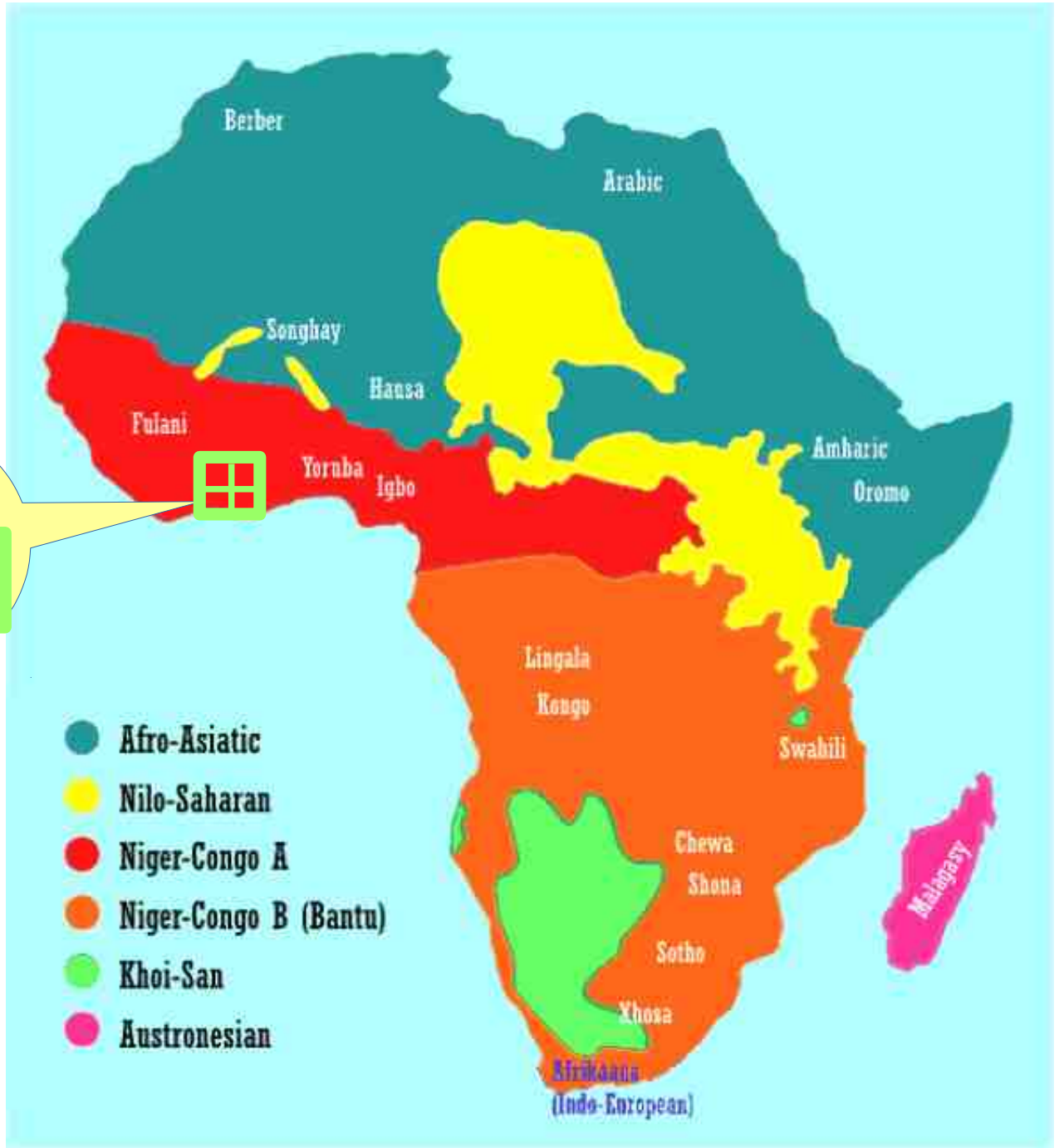


https://elms.wordpress.com/2008/03/04/lexical-distance-among-languages-of-europe/

Selecting features for similarity distances

- Lexical
  - Swadesh word list
  - West African Lexical Dictionary Set (WALDS)
  - …
- Phonetic / phonological
  - vowel set comparison
  - consonant set comparison (more stable – used for many traditional initial classifications, e.g. of Indo-European languages)
- Grammatical features
  - World Atlas of Language Structures (WALS)
- An example:
  - Consonants in the Kru language family

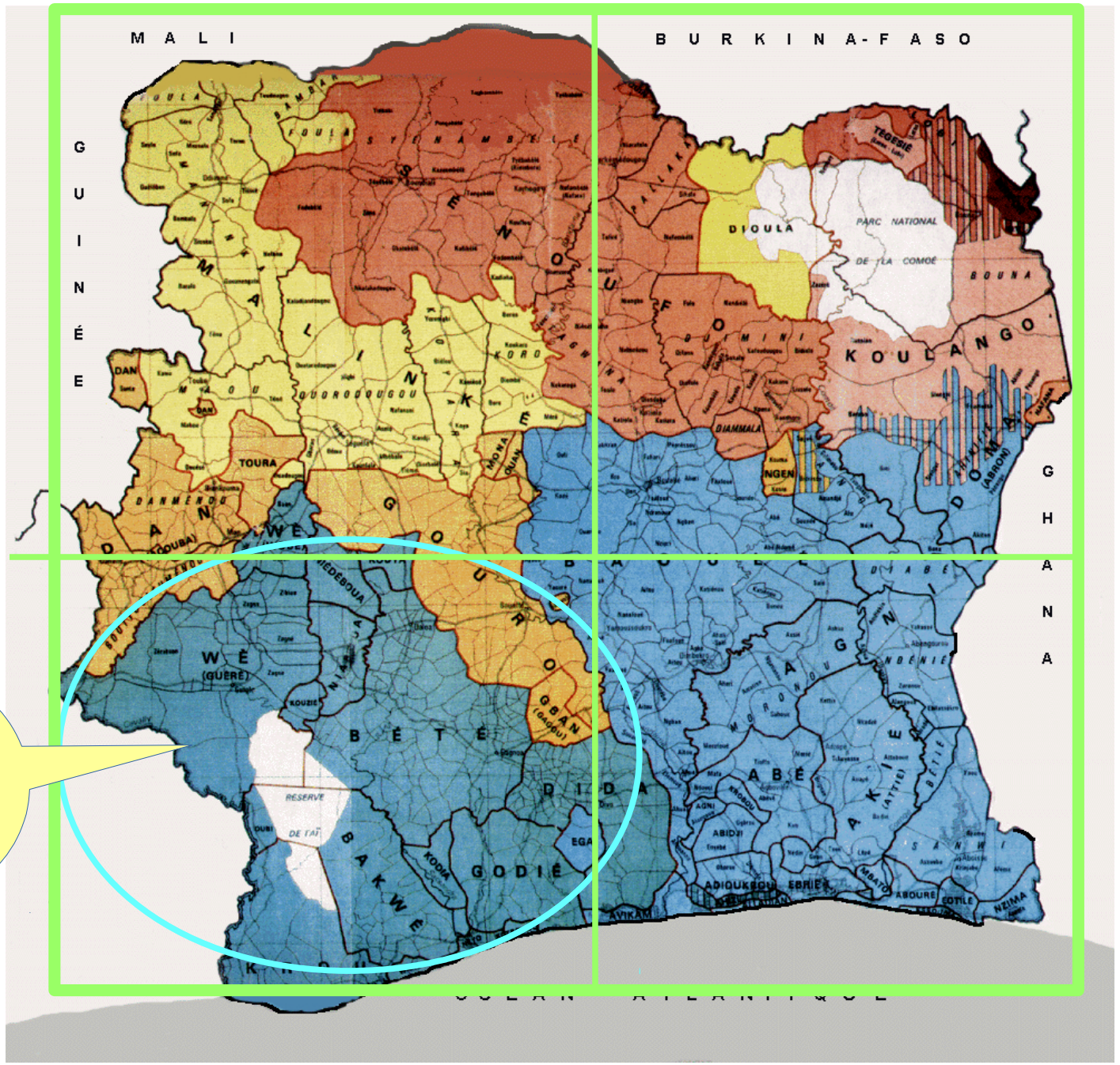
Kru languages – Ivory Coast



Ivory Coast

Mande	Gur
Kru	Kwa

Kru languages – Ivory Coast



Kru languages – Ivory Coast: Feature – consonant systems

Bete	p t c k kp kw _ b d C _ g gb _ f s _ v z _ _ _ B _ l j x w m n J N Nw _ _ _ _ _ _ _ _ _ _ _
Godie	p t c k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B _ l j x w m n J N Nw _ _ _ _ _ _ _ _ _ _ _
Koyo	p t c k kp kw kj b d C _ g gb _ f s _ v z _ _ _ B _ l j x w m n J N _ _ _ _ _ _ _ _ _ _ _ _
Neyo	p t c k kp kw _ b d C _ g gb _ f s _ v z _ _ _ B _ l j x w m n J N _ _ _ _ _ _ _ _ _ _ _ _
DidaDeLozoua	p t c k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B _ l j x w m n J N Nw _ _ _ _ _ _ _ _ _ _ _
DidaF	p t c k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B _ l j x w m n J N _ Nm _ _ _ _ _ _ _ _ _ _
Wobe	p t c k kp kw _ b d C _ _ gb _ f s _ _ _ _ _ _ _ _ _ _ w m n J _ Nw Nm km _ _ _ _ _ _ _ _ _
Guere	p t c k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B D l j _ w m n J _ Nw Nm km _ _ _ _ _ _ _ _ _
Krahn	p t c k _ kw _ b d C _ _ gb _ f s _ _ _ _ _ _ _ _ l _ _ w m n J _ _ _ _ _ _ _ _ _ _ _ _ _
Cedepo	p t c k kp kw _ b d C _ _ gb _ f s _ _ _ _ h _ _ _ l _ _ _ m n J _ _ Nm _ _ _ _ _ _ _ _ _ _
Klao	p t c k kp kw _ b d C _ _ gb _ f s _ _ _ _ _ _ _ _ l j _ w m n J _ _ Nm _ _ _ _ _ _ _ _ _ _
Niaboua	p t c k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B _ l j _ w m n J _ _ _ _ _ _ _ _ _ _ _ _ _
Dewoin	p t _ k kp kw _ b d C _ g gb gw f s _ v z _ _ _ B _ l j _ w m n J N _ _ _ _ _ _ _ _ _ _ _ _
Bassa	p t c k kp _ _ b d C dj g gb _ f s _ v z _ h hw B _ l _ _ w m n J _ Nw _ _ _ _ _ _ _ _ _ _ _
Grebo	p t c k kp _ _ b d C _ g gb _ f s _ _ _ _ h hw _ _ l j _ w m n J N Nw Nm _ _ hm hn hl _ _ _ _ _
Tepo	p t c k _ kw _ b d C _ g gb _ f s _ _ _ _ h _ _ _ l j _ w m n J N _ Nm _ _ _ _ _ _ _ _ _ _
KuwaaLiberia	p t _ k kp kw _ b d C _ _ _ _ f s _ _ _ _ _ _ _ _ l j x w m n J N _ _ _ _ _ _ _ _ mb nd nC Ng Nmgb
SemeHauteVolta	p t c k kp _ _ b d C _ g gb _ f s S v _ _ h _ _ _ l j _ w m n J _ _ _ _ gm _ _ _ _ _ _ _ _
AiziCdI	p t c k kp _ _ b d C _ g gb _ f s S v z Z _ _ _ _ l j _ w m n J N _ _ _ _ _ _ _ _ _ _ _ _

Method:

1. compare all consonant systems pairwise:

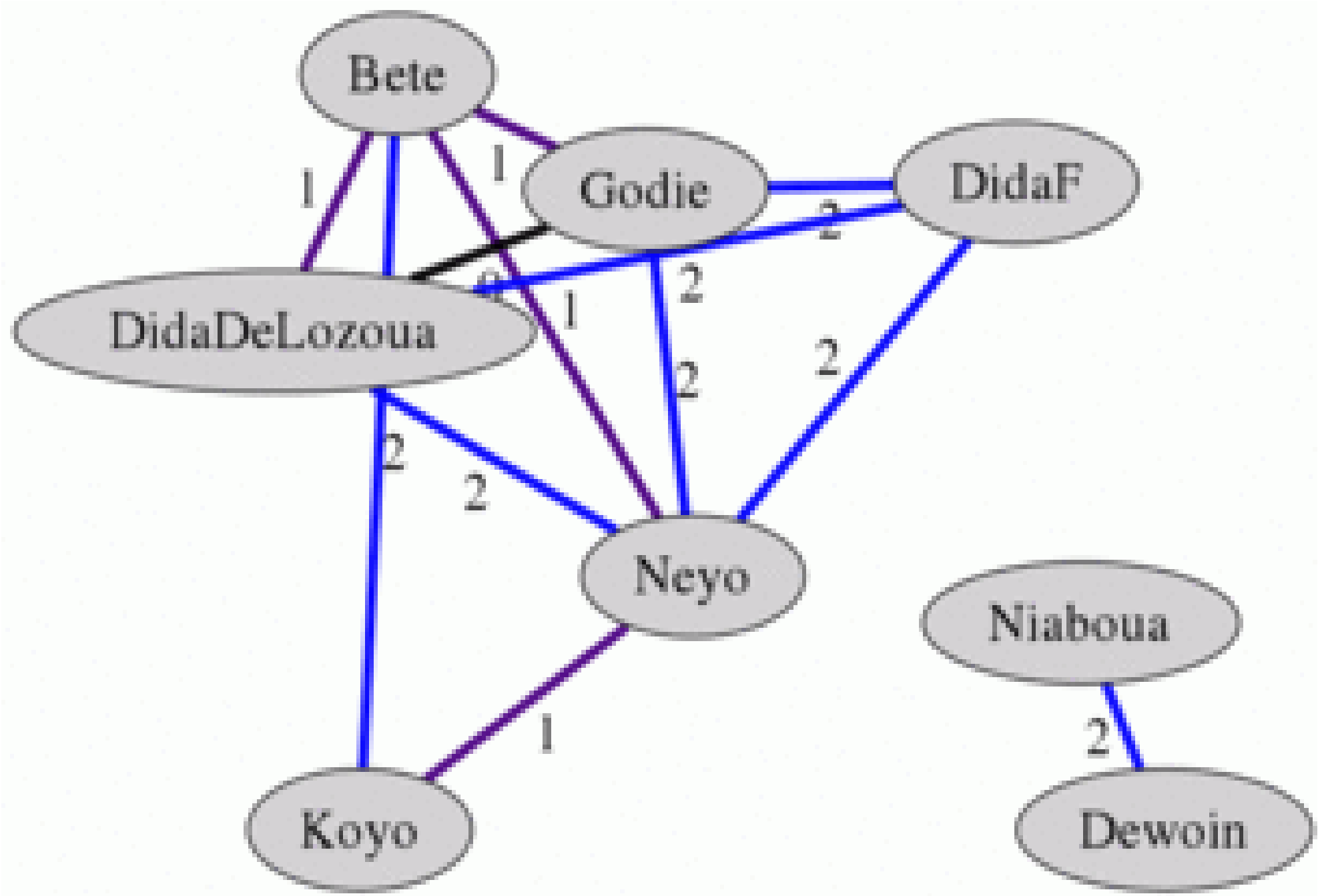
   Levenshtein Edit Distance / Hamming Distance

2. visualise differences as ‘virtual distances’ in a chart

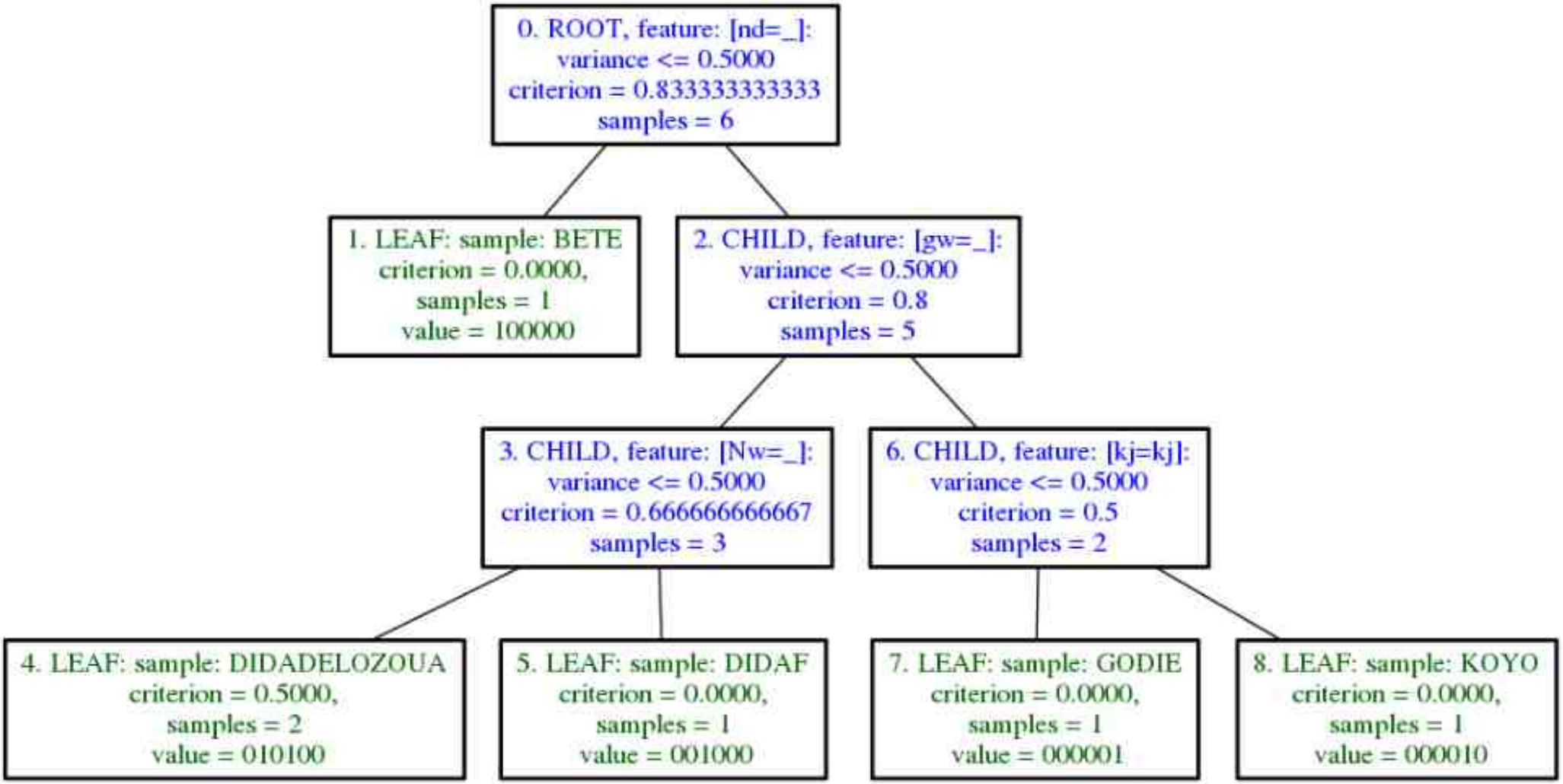
Kru languages – Ivory Coast: Feature – consonant systems

Bete	0	1	2	1	1	3	10	6	9	11	8	4	4	7	11	8	12	9	6
Godie		0	3	2	0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
Koyo			0	1	3	3	12	8	9	11	8	4	4	9	13	8	12	9	6
Neyo				0	2	2	11	7	8	10	7	3	3	8	12	7	11	8	5
DidaDeLozoua					0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
DidaF						0	11	5	10	10	7	3	3	10	12	7	13	10	7
Wobe							0	8	6	6	4	10	12	12	11	8	14	11	12
Guere								0	11	11	8	4	6	9	13	10	18	11	10
Krahn									0	4	3	7	9	10	12	5	11	8	9
Cedepo										0	3	9	11	10	10	5	13	8	11
Klao											0	6	8	11	9	4	10	7	8
Niaboua												0	2	7	13	8	14	7	6
Dewoin													0	9	13	8	12	9	6
Bassa														0	10	11	19	8	9
Grebo															0	7	17	10	11
Tepo																0	12	7	8
KuwaaLiberia																	0	15	14
SemeHauteVolta																		0	5
AiziCdI																			0

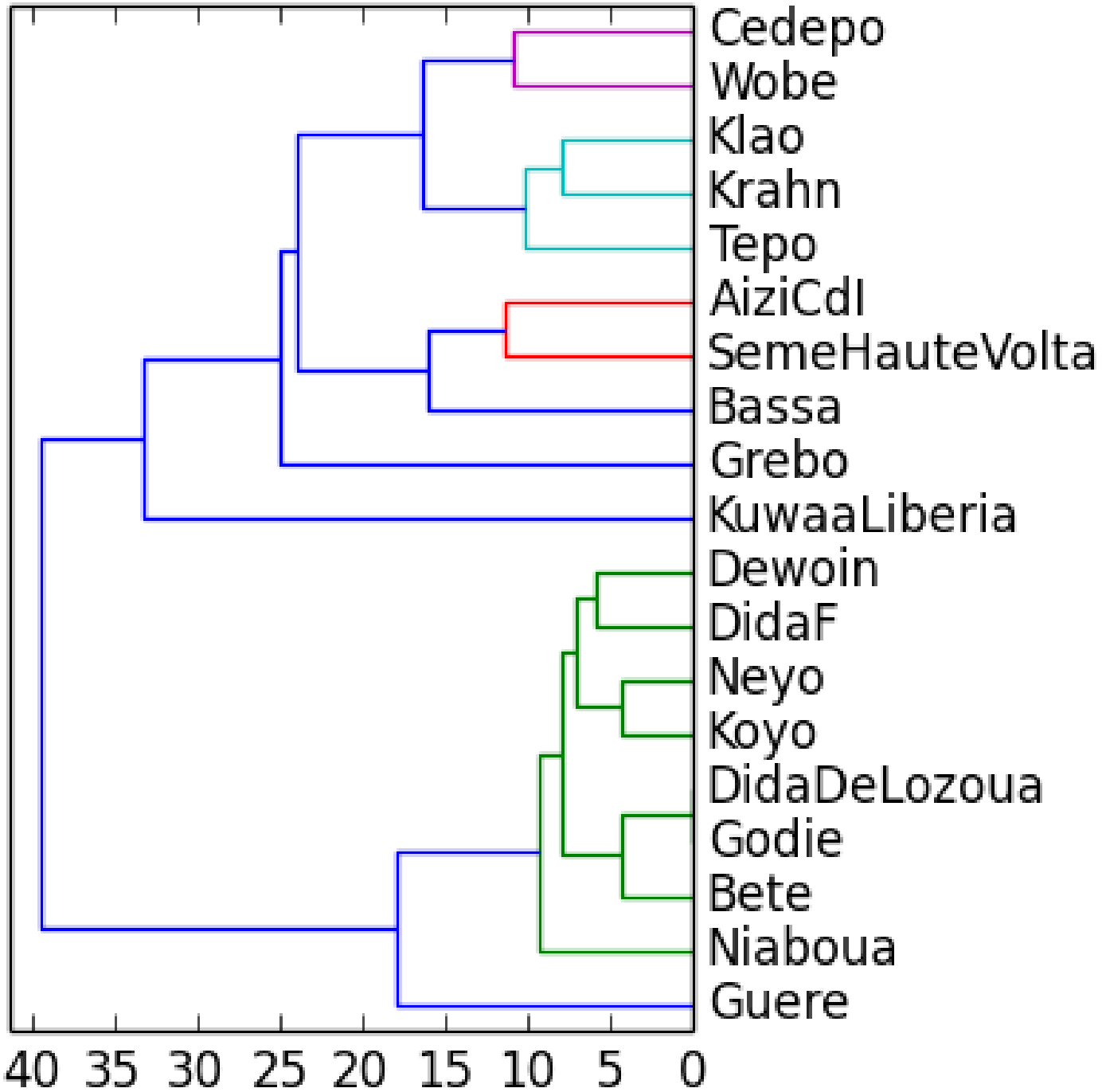
Kru languages – Ivory Coast: Feature – consonant systems



Which features are most useful? - Help from Machine Learning (Decision Tree Induction)



Kru languages – Ivory Coast: Feature – consonant systems



Summary and Conclusion

Summary

- Reversed roles:
  - How can language documentation and the human language technology resources relate to each other?
- Human Language Technologies:
  - Documentary technologies
  - Enabling technologies
  - Product technologies
- Enabling technologies in language documentation:
  - Annotation and annotation mining
    - How do languages differ in speech rhythm?
  - Language classification assisted by Machine Learning (ML)
    - Language differences/distances among languages of Ivory Coast
- Endangered and less resourced languages:
  - Enabling technologies amplify efficiency, speed, size


Diolch yn fawr!