Language Technologies

What endangered languages can teach us

Dafydd Gibbon

U Bielefeld

ELKL-4, Agra, 2016-02-

Focus: role reversal

Not just:

What can the Human Language Technologies <u>offer</u> to endangered (and other) languages?

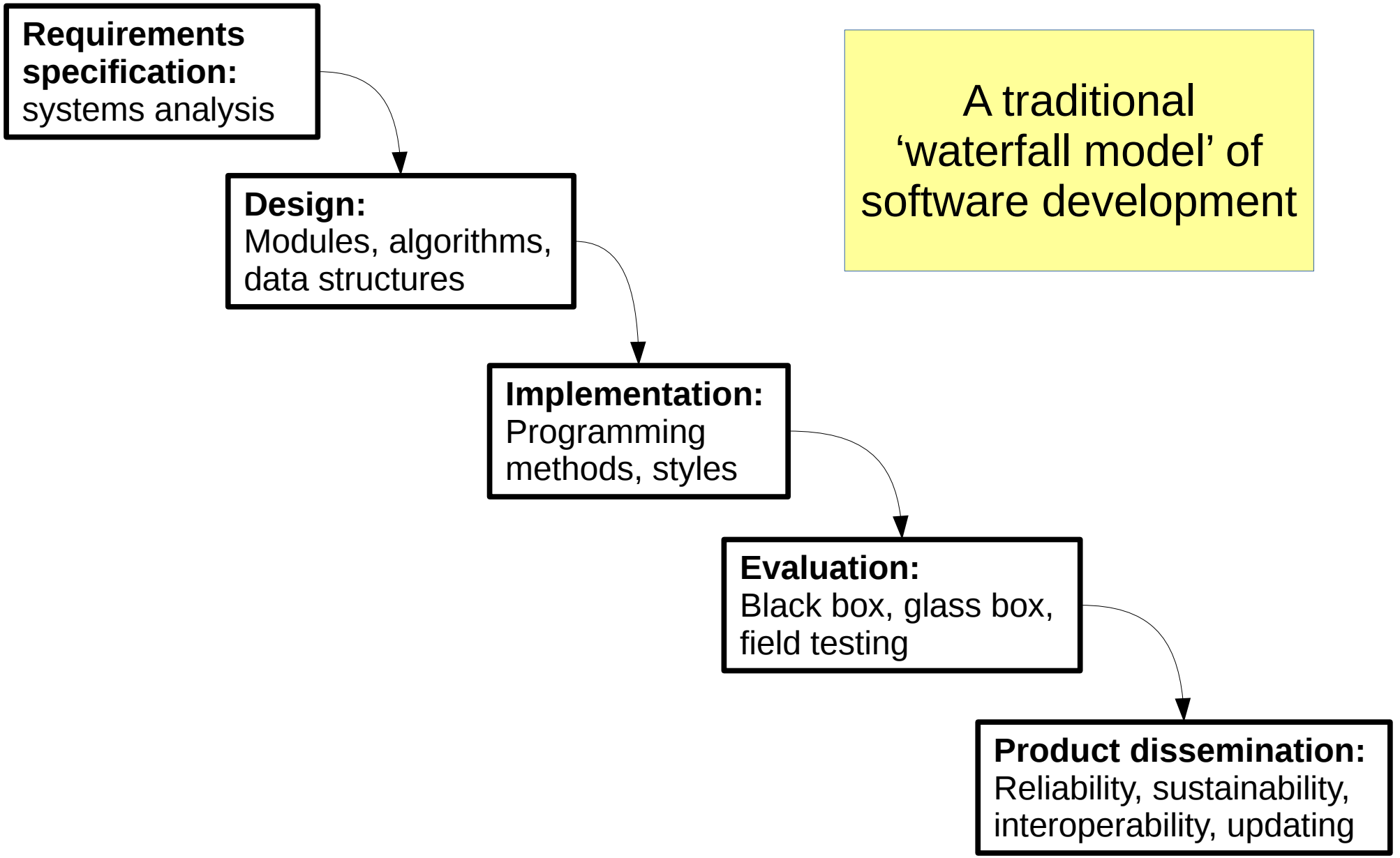
But:

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

Or:

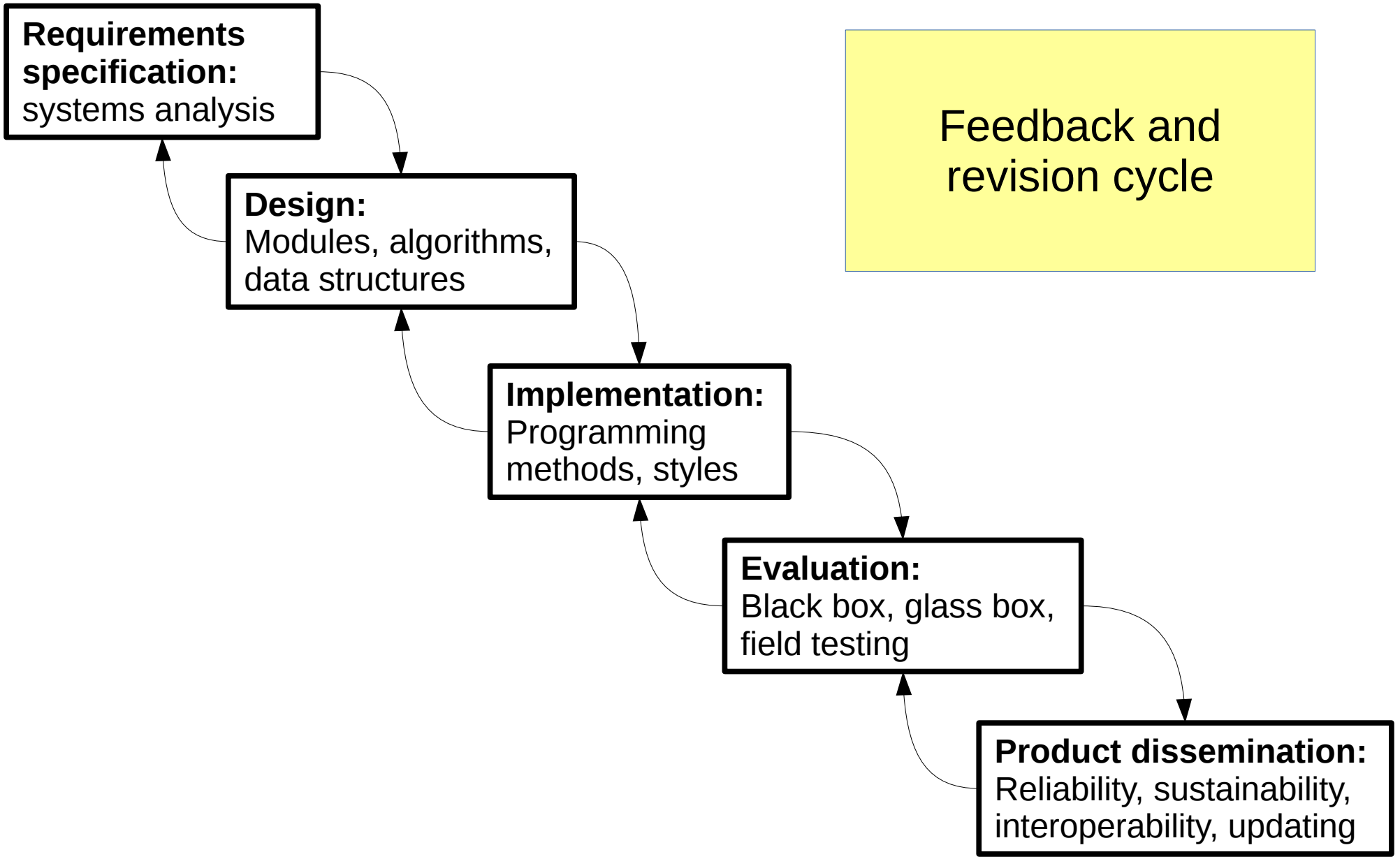
What do endangered languages <u>require</u> from the Human Language Technologies?

Applying technologies – a reminder



A traditional 'waterfall model' of software development

Applying technologies – a reminder



Summary

- Reversed roles:
  - What can endangered languages teach lang tech?
- Background to Language Endangerment
- Technology roles:
  - Documentary technologies
  - Enabling technologies
  - Productivity technologies
- Enabling technologies
  - Annotation and annotation mining
    - How do languages differ in speech rhythm?
  - How similar / different are languages?
    - Language differences/distances among languages of Ivory Coast
- Endangered languages:
  - What do science and technology (human knowledge in general) lose when languages die?

Language Technologies

- **Engineering:**
  - **speech**: ASR, TTS, speaker id / recognition, ...
  - **language**: 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

Language Technologies

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

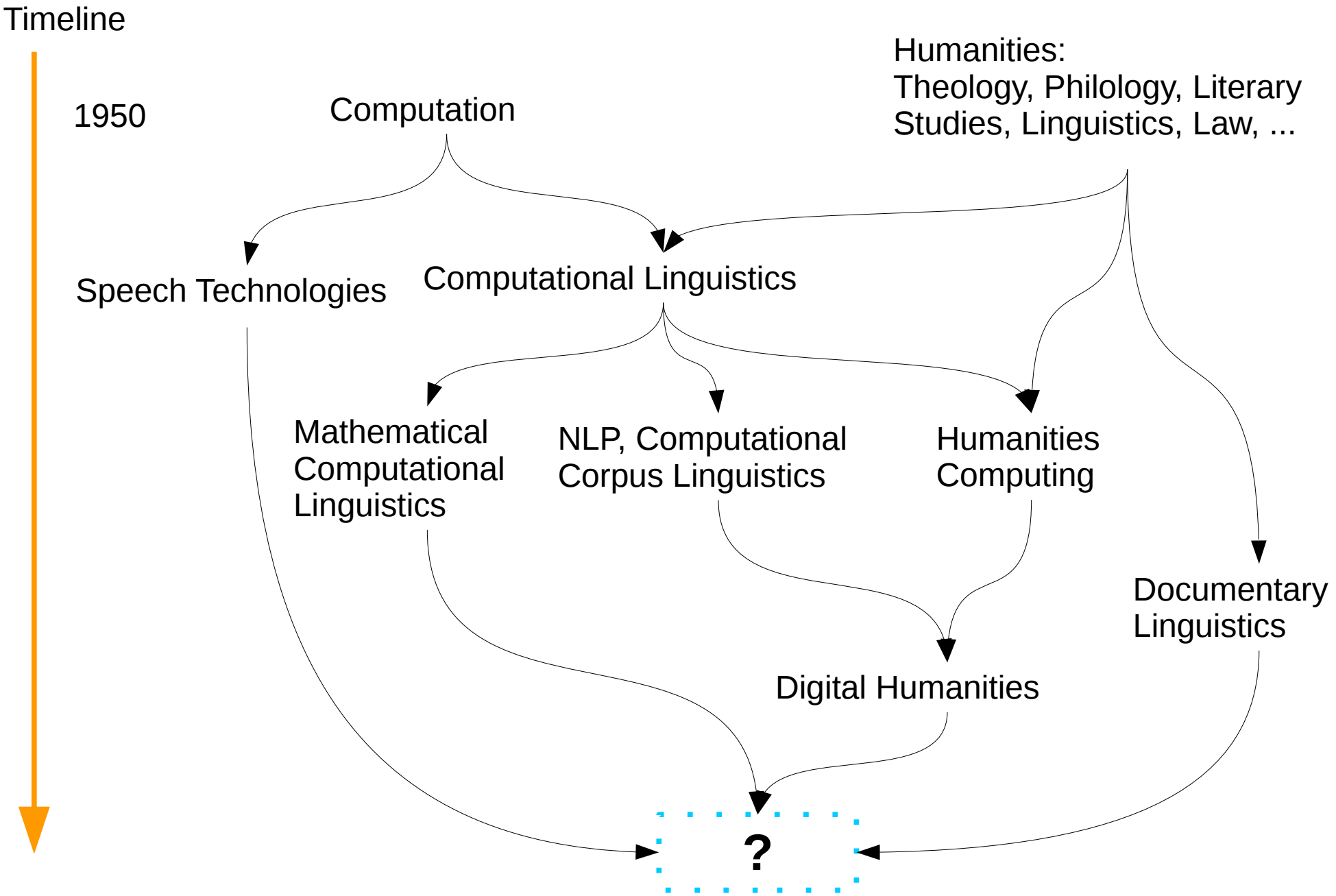
OVERLAP

(also with linguistics)

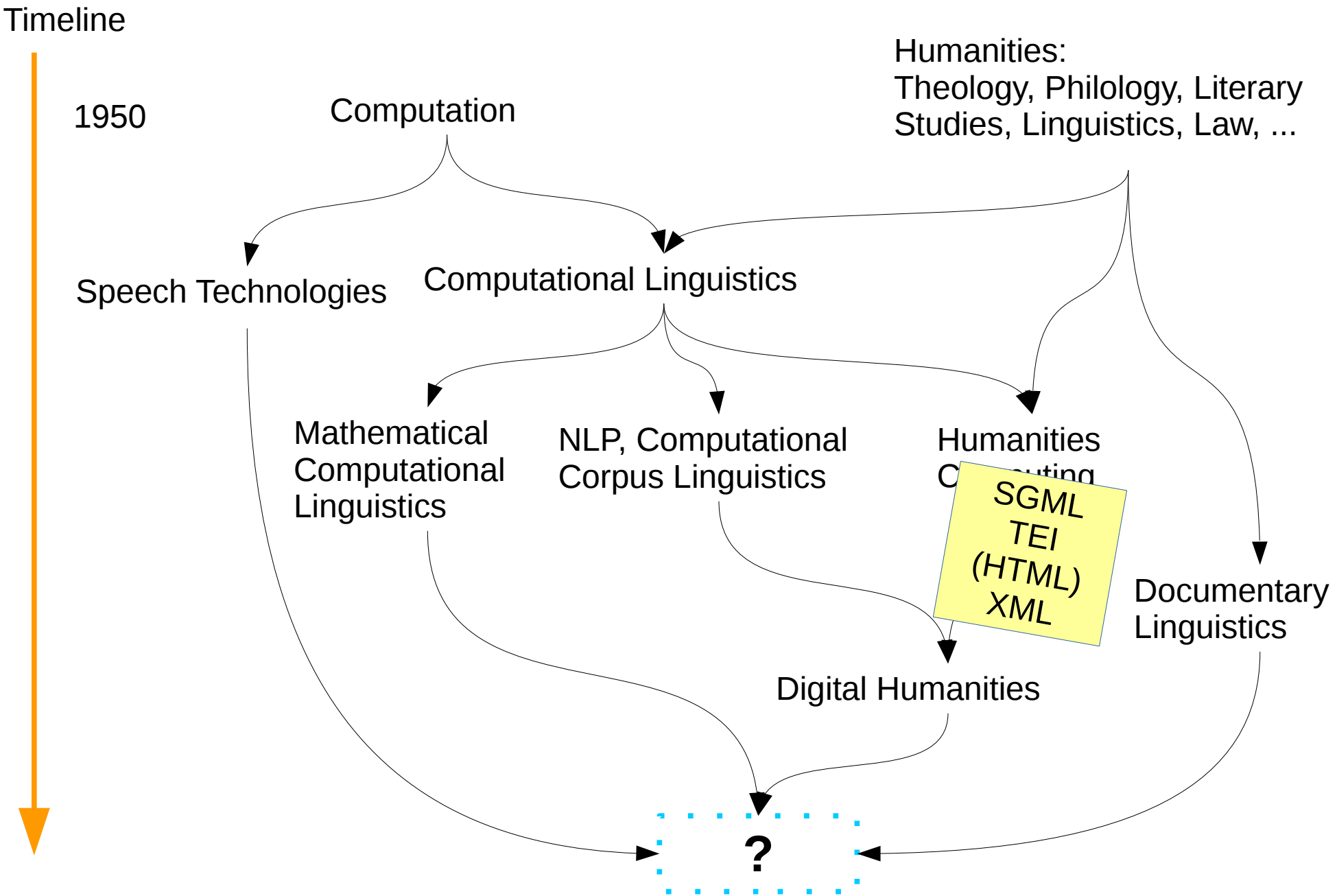
- **Computational linguistics & Computer Science:**
  - **domain models** of natural language syntax, semantics, pragmatics, language typology and genesis
  - **formalisms and algorithms** for induction, parsing, generation of language
  - **corpus analysis** for lexicon and grammar induction

The broader interdisciplinary context

Doc Ling and Digital Humanities



Doc Ling and Digital Humanities



Roles for computational technologies

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

1. Documentation Technologies

- Project planning tools
- Data collection tools
  - scenario support for
    - elicitation
    - recording (multimodal)
    - metadata collection
  - document scanning
- Data archiving and access
  - database and search model
    - 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

2. Enabling Technologies

- 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

3. Productivity Technologies

- Recognition
  - Automatic Speech Recognition (ASR)
  - Visual scene and object recognition
  - Information retrieval from text
- Identification
  - Speaker identification
  - Language identfication
  - Authorship attribution
- Generation
  - Text-to-Speech Synthesis (TTS)
  - Written text geneneration from databases
- Products
  - Dictation and information applications
  - Translation applications

Endangered languages as teachers: some topics

Endangered languages as teachers: some topics

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

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	+			+	+	++	+	+	+	+	+	++		++	+	+	+

Many more topics …

- A closer look at language typology reveals many more requirements for the technogies, for example:
  - concurrent data and parallel processing:
    - phonological tone
    - morphological tone vs. pitch accent vs. stress; prosodic negation
    - grammatical stress/accent, focus, intonation, word order
    - gesture, sign language, discourse participants
  - language similarity for priority setting in language selection for funding (usually: chance) and for system adaptation:
    - role of language typology:
      - similarity/difference in speech sound systems
      - similarity/difference in grammar
      - similarity/difference in the lexicon
    - similarity/difference in discourse conventions
    - similarity/difference in general cultural conventions

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'

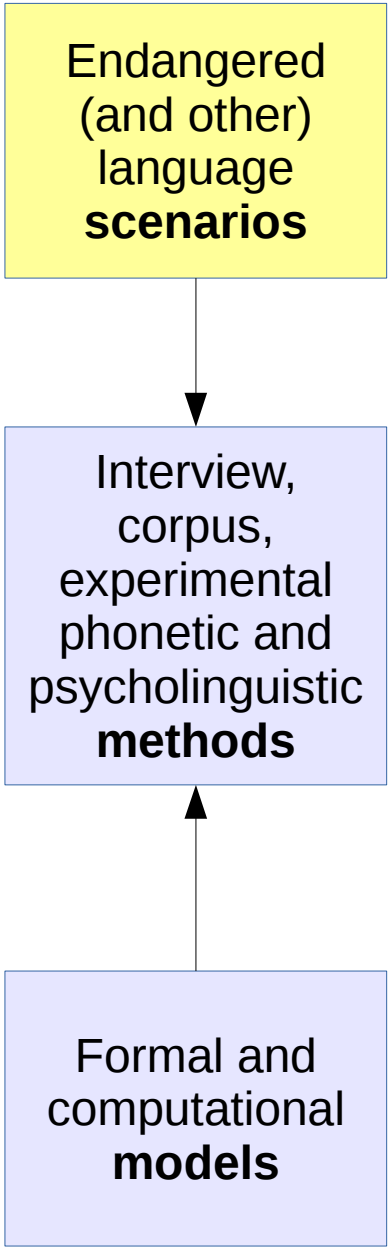
- Language, speech, multimodal technology scenarios are:
  - highly standardised, rather coherent
  - tendentially easy to define and delimit
  - very application / product oriented
    - especially in speech technology: highly product specific
    - text technology is more generic
  - regulated standards: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

Relations between the technologies: a simplified 'workflow'

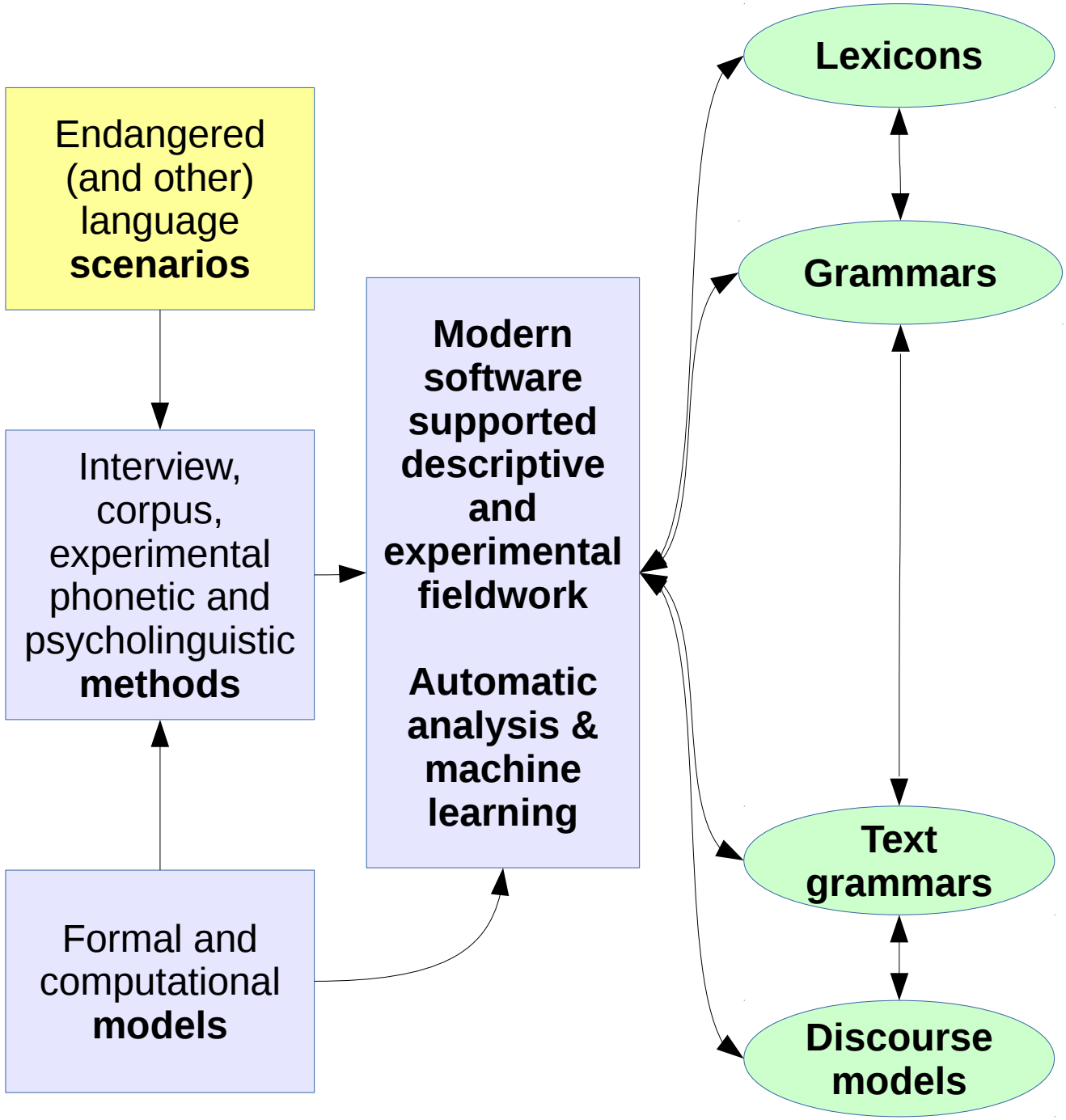
Reversed roles: what endangered languages offer to engineering

Endangered (and other) language **scenarios**

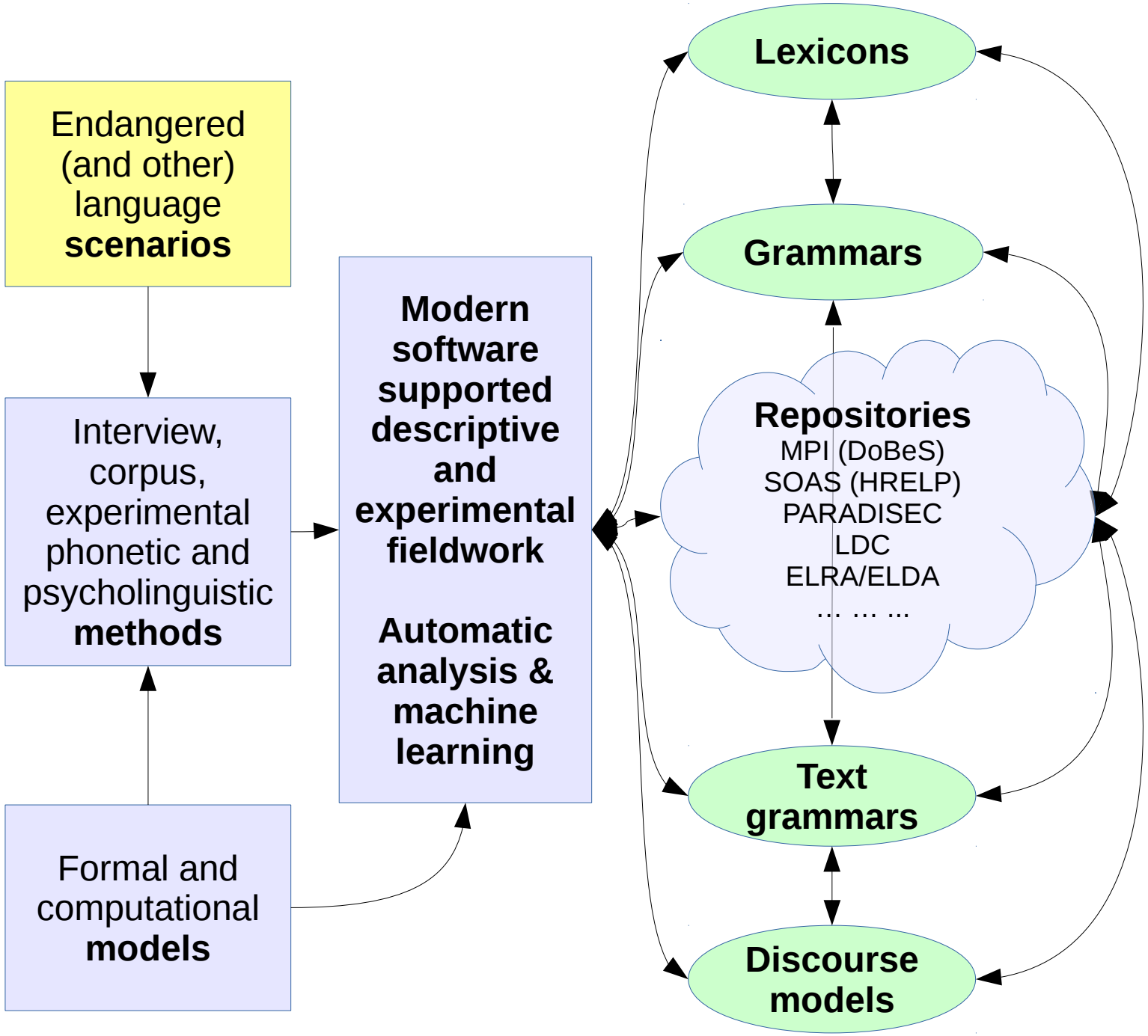
Reversed roles: what endangered languages offer to engineering



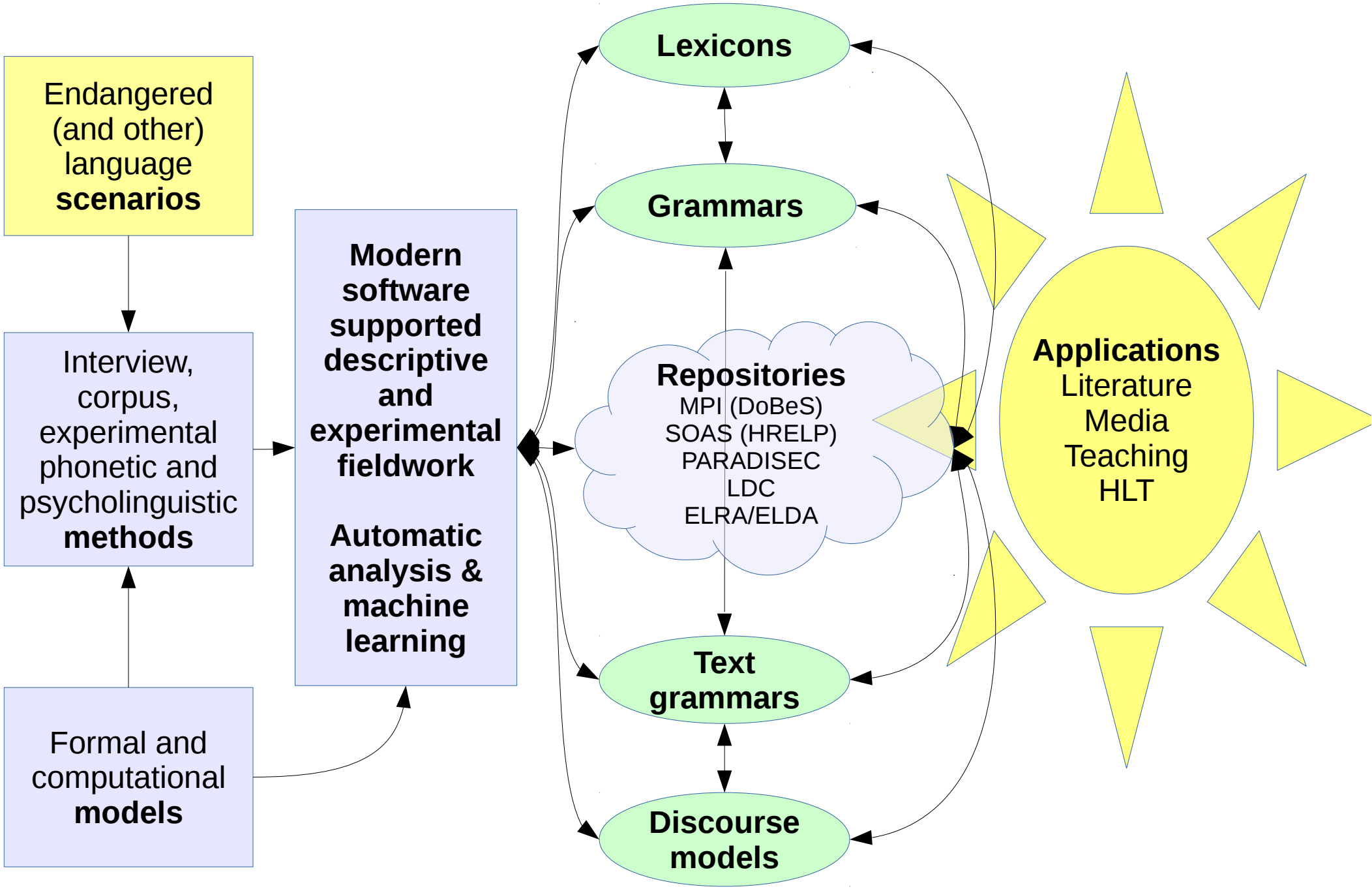
Reversed roles: what endangered languages offer to engineering



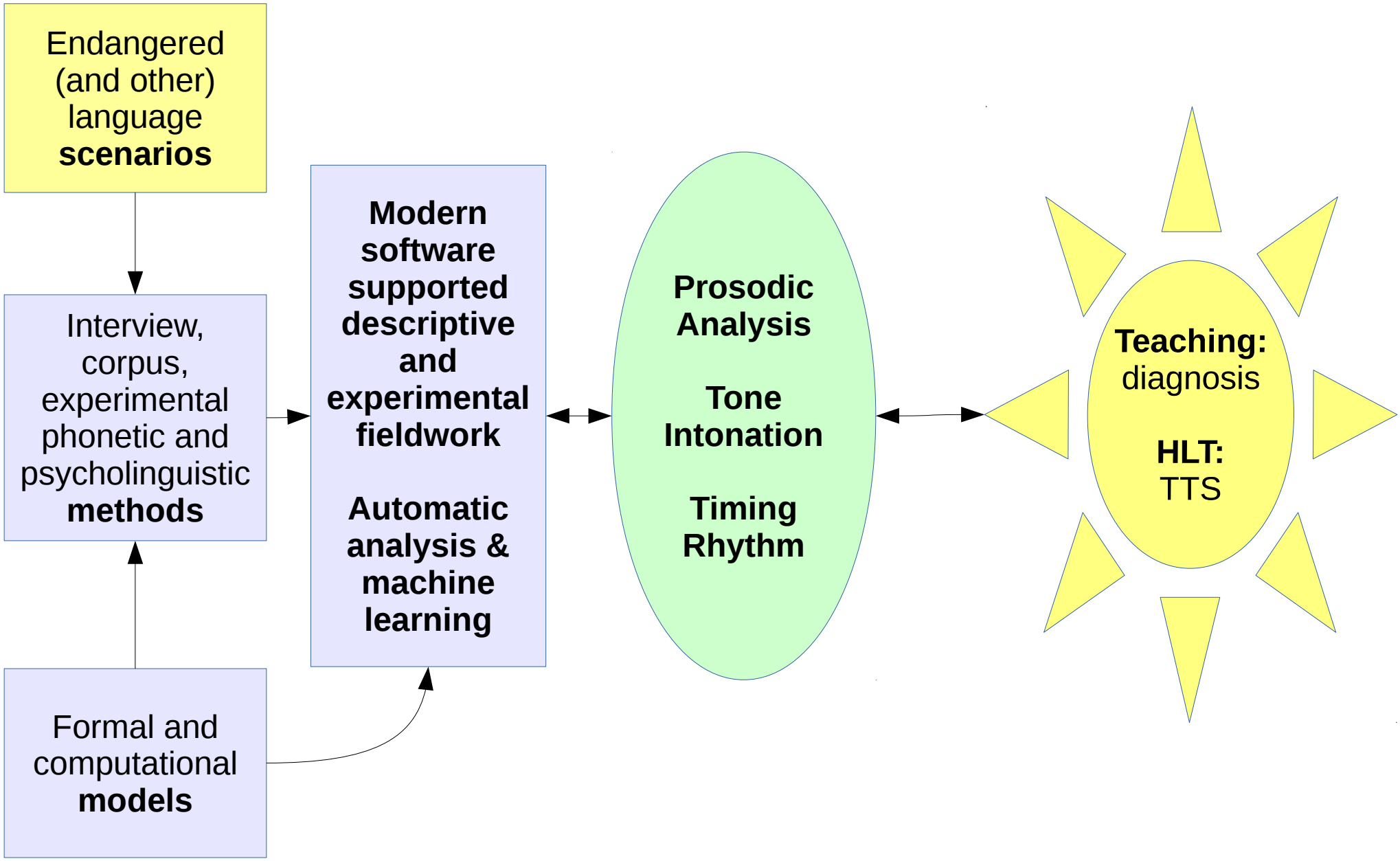
Reversed roles: what endangered languages offer to engineering



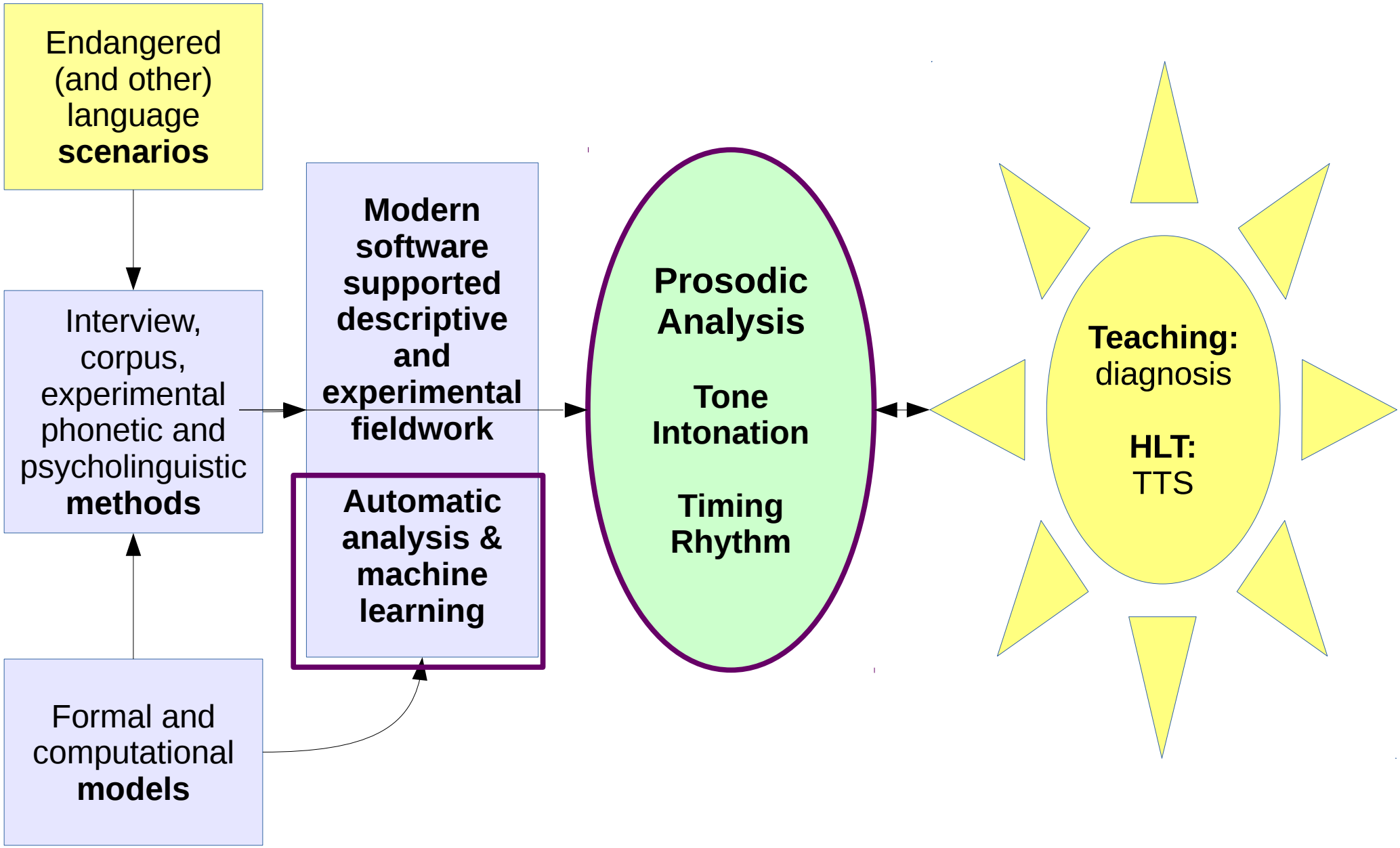
Reversed roles: what endangered languages offer to engineering



What I will be looking at – a modest selection



What I will be looking at – a modest selection



Background to Language Endangerment:

A Global Perspective

Ethnologue metadata repository of all known languages

Background: Global Perspective on language endangerment

- Globalization
  - travel, trade, economy, politics
    → language dominance
    Trade languages, pidgins:
    Greek → Latin → Arabic → Portuguese → Spanish → French → English
    → religion, culture
  - crises
    - Example: climate deterioration (Africa, West Asia)
      → lack of water → famine → poverty → hunger → disease
      → migration → diaspora

- "Clash of Civilizations" (Samuel Huntington)
  → conflicts: both local and global

- The role of language stereotypes, discrimination?

Background: Global Perspective on language endangerment

- Globalization
  - trav
    - → l
      T
      c
    - → r
  - crise
    - E
      –
- “Clash
  - →
- The role of language stereotypes, discrimination?

→ English

The ‘explanation’ for language endangerment as

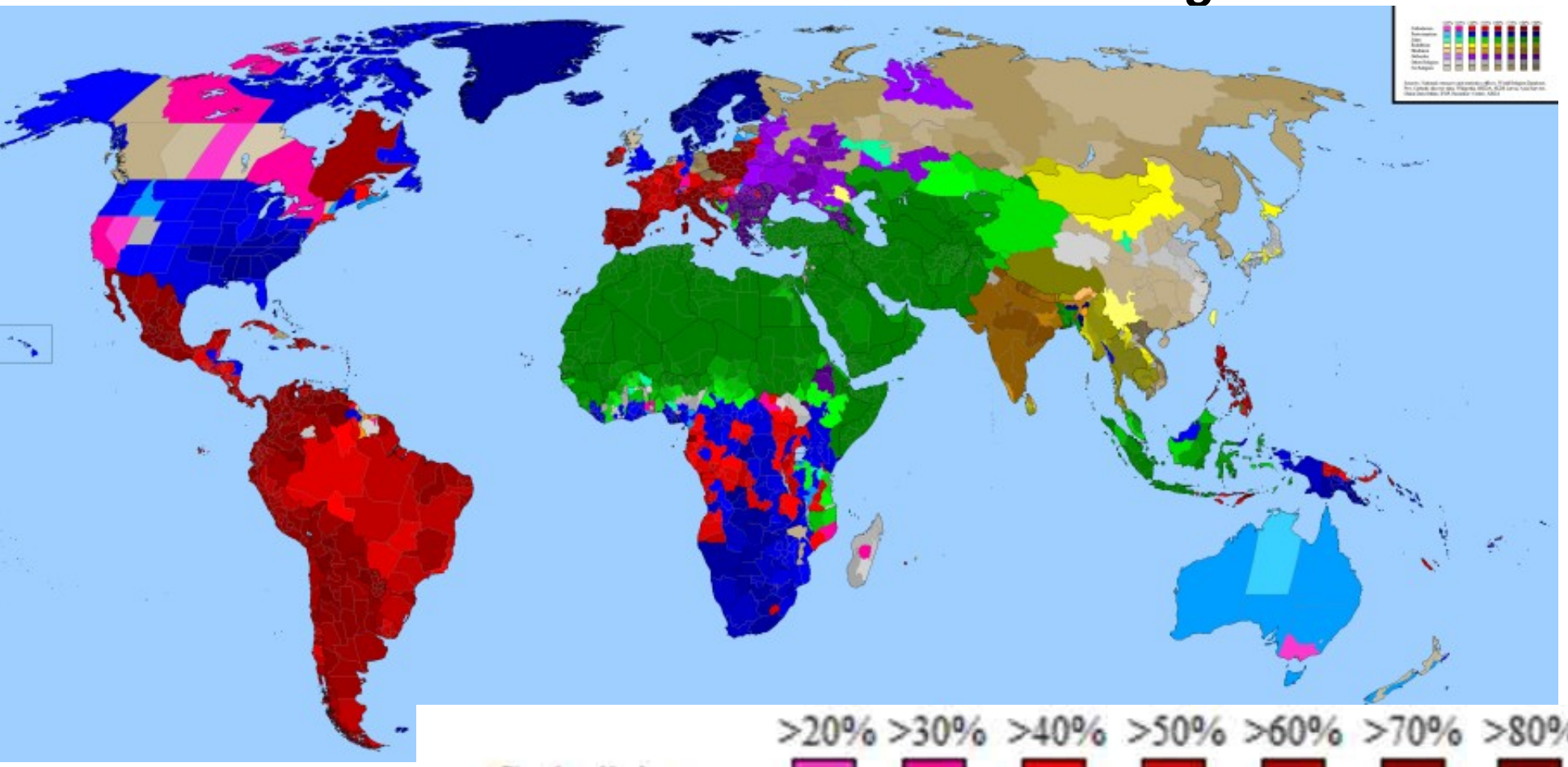
LOSS OF INTERGENERATIONAL TRANSMISSON OF LANGUAGES, CULTURE, RELIGION, SKILLS, …

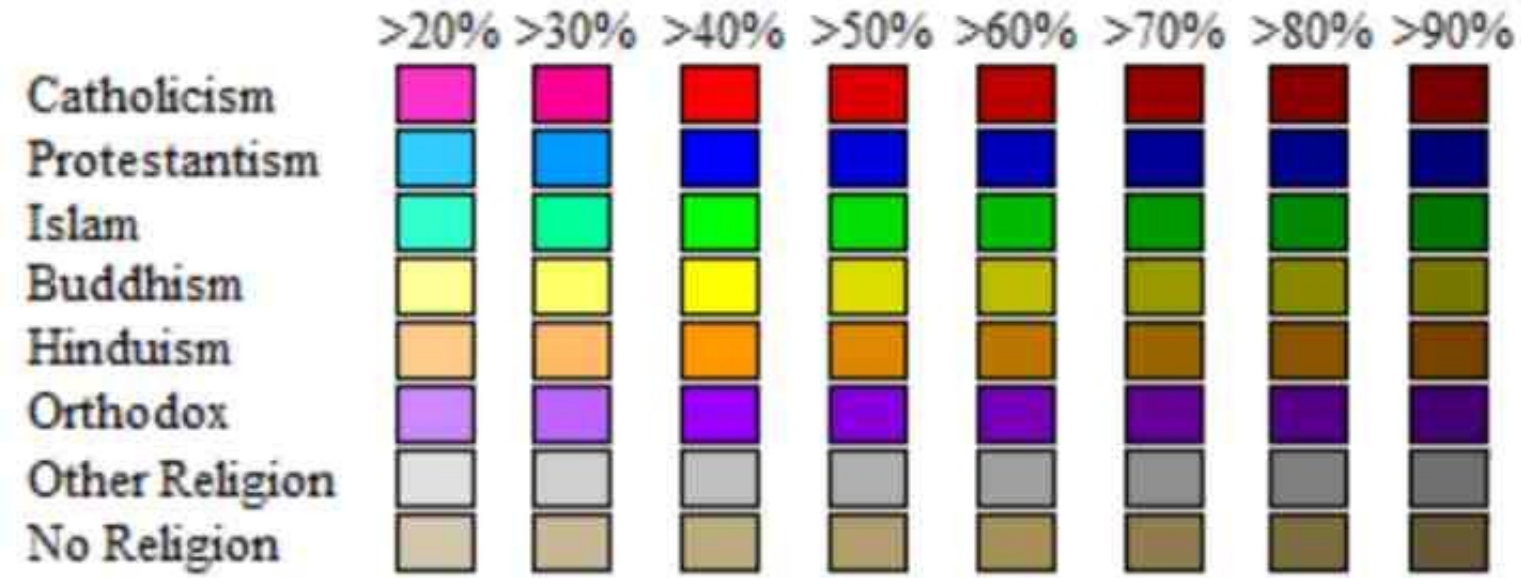
is too simple.

It begs the question:

Why is there loss of intergenerational transmission?

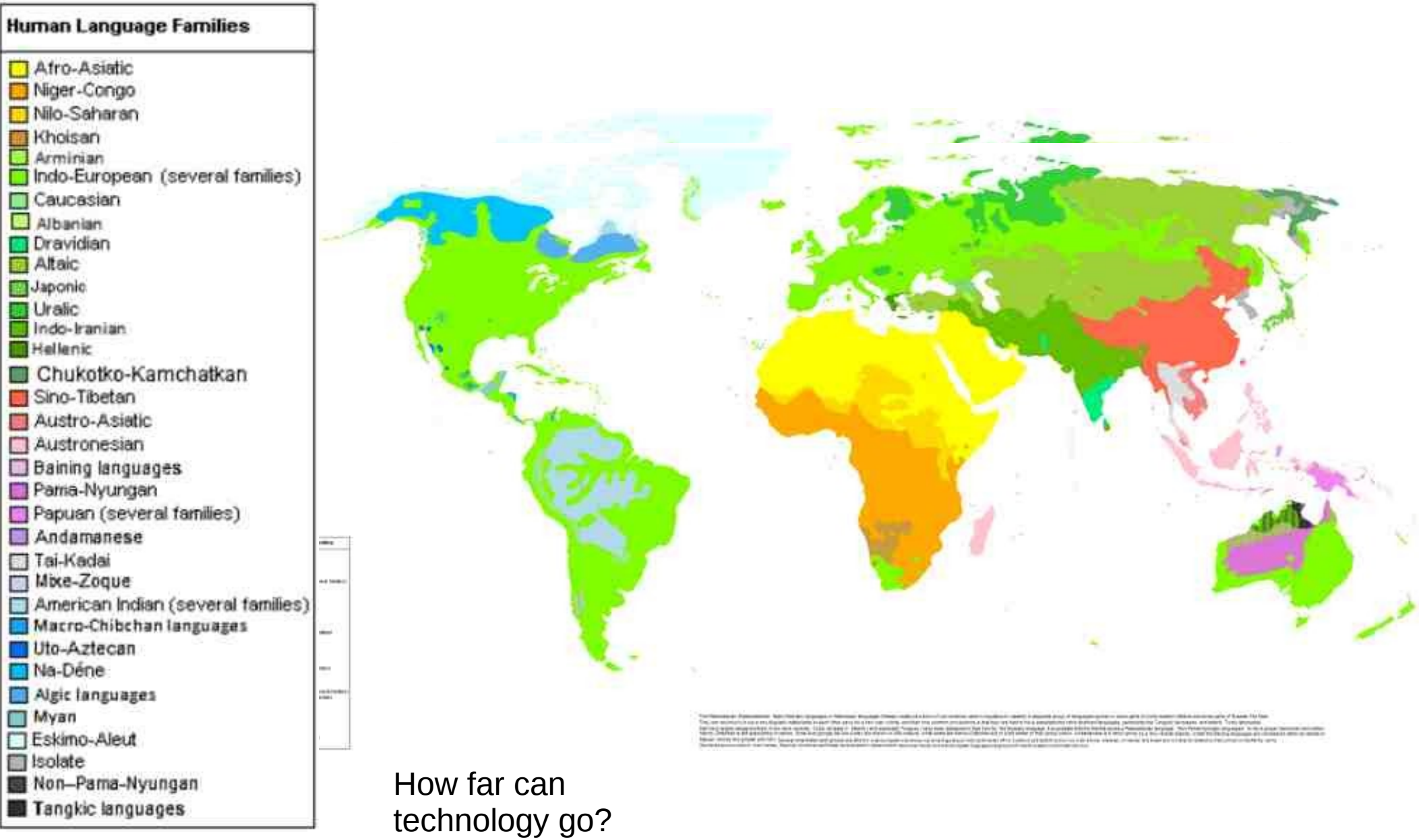
Clash of Civilizations? World religions





How far is technology relevant?

Clash of Civilizations? World language families

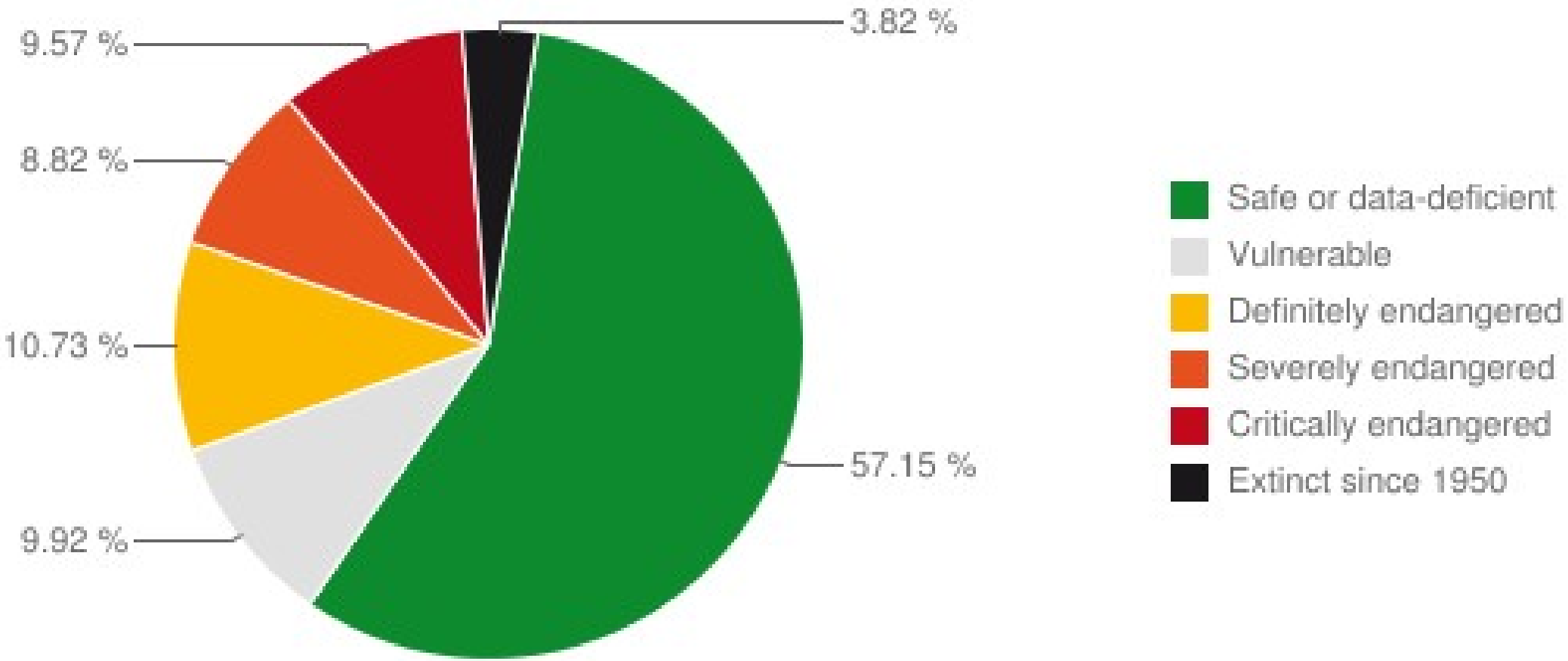


How far can technology go?

ISO 639-3 Language Code Standard – 5 language types

- **Living languages:** A language is listed as living when there are people still living who learned it as a first language. Based on intelligibility not politics.
- **Extinct languages:** no longer living, gone extinct in recent times. (e.g. in the last few centuries), distinct languages identified based on intelligibility (as defined for individual languages).
- **Ancient languages:** If it went extinct in ancient times (> 1000 years), must have had a distinct literature and be treated distinctly by the scholarly community, must have an attested literature or be well-documented as a language known to have been spoken by some particular community at some point in history, may not be areconstructed language inferred from historical-comparative analysis.
- **Historic languages:** Must be distinct from any modern languages that are descended from it; for instance, Old English and Middle English. Here, too, the criterion is that the language have a literature that is treated distinctly by the scholarly community.
- **Constructed languages:** Must have a literature and it must be designed for the purpose of human communication. Specifically excluded are reconstructed languages and computer programming languages.

UNESCO chart of endangered languages



UNESCO map of endangered languages

Endangered languages - 'in danger of disappearing'

- Endangered languages:
  - Languages in danger of disappearing
  - UNESCO
  - The language birth and death cycle
- Why and how do new languages appear / disappear?
- Why 'danger'?
  - The economics of language endangerment
  - The politics of language endangerment
  - Social consequences of language endangerment
- What can the Human Language Technologies do?
  - Documentation Technologies?
  - Enabling Technologies?
  - Productivity Technologies?
- Which languages can be handled by HLT?

Enabling Technologies

Annotation and Annotation Mining

Language Similarity Analysis

Enabling technologies

- Annotation (preferably 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

- Similarity analysis – virtual distance mapping
  - Which languages have been (almost) documented?
  - Which languages are likely to respond to adaptive techniques for modelling a language starting with a known model for another language.
- Similarity feature definition
  - Geographic (areal contact)
  - Typological (similarity of paradigmatic and syntagmatic structures)
  - Genealogical (history of language families)

Enabling technologies

Annotation and Annotation Mining

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

- Searching unstructured data is difficult
  - but improving with the help of machine learning
    - example: search on free text data (e.g. Google search as a machine for on-the-fly concordance construction from web data)
    - example: Google's image search

Why annotation? And how?

- Th eech
  dat
  -
  -
  - ta
  -
- Se
  - machine

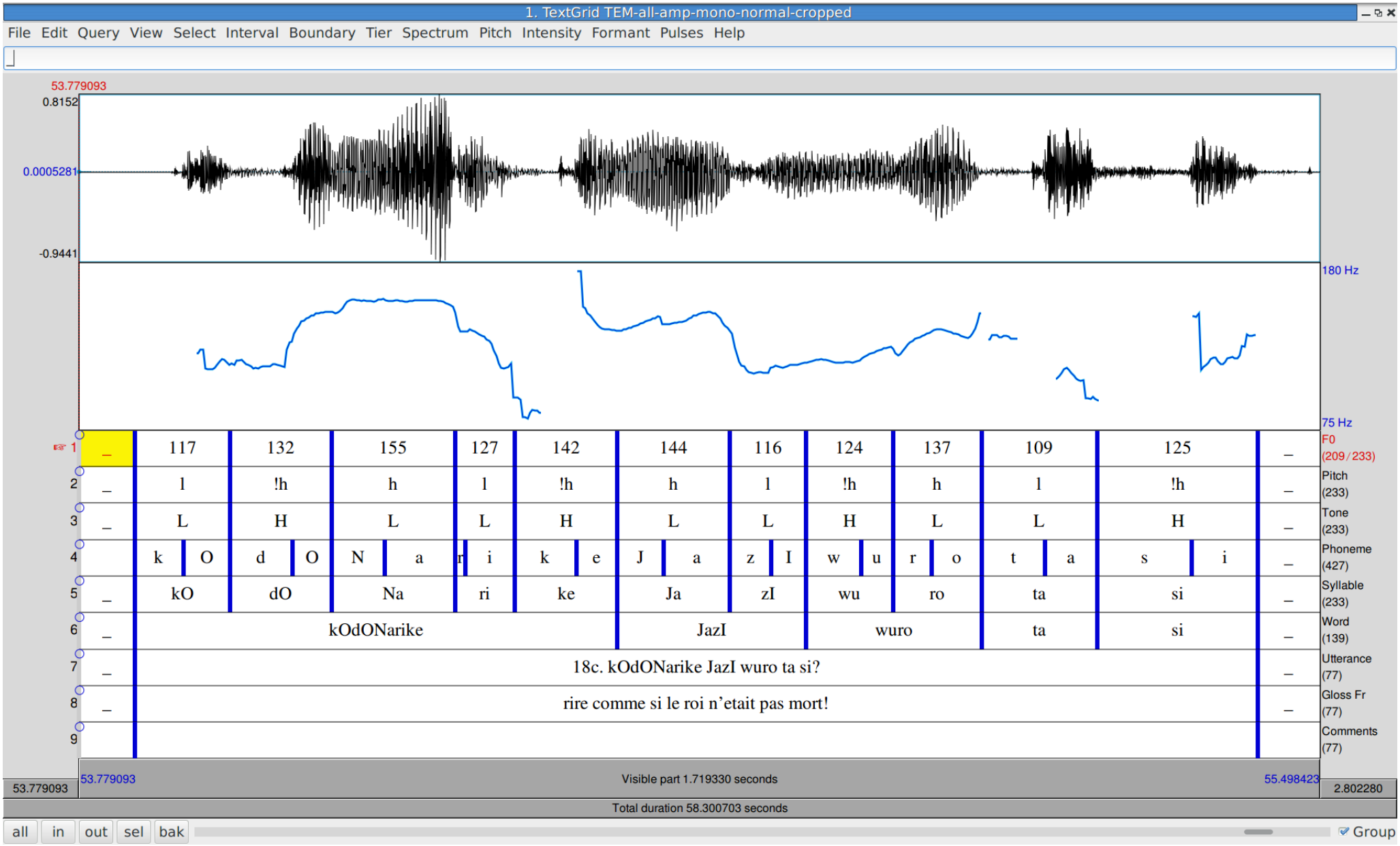
ALL THESE PROCEDURES
ARE VERY PRECISELY DEFINABLE
AND THEREFORE CAN BE (AND ARE OFTEN)
AUTOMATISED WITH APPROPRIATE SOFTWARE

(POSTAGGERS, CONCORDANCERS,
AUTOMATIC SPEECH ANNOTATORS, ...)

THIS IS STANDARD PRACTICE
IN THE HUMAN LANGUAGE TECHNOLOGIES (HLT)
BUT NOT (YET) IN DOCUMENTARY LINGUISTICS

WHEN THIS HAPPENS, THE EFFICIENCY
OF MANY ASPECTS OF DL
WILL BE REVOLUTIONLSED

Speech annotation example: graphical display with Praat (Niger-Congo>Gur>Tem, ISO 639-3 kdh)



Annotation mining: the syllable timing space

- A standard meme in speech timing descriptions is that the spoken forms of languages fall into 3 categories:
  - stress (or foot) timing
  - syllable timing
  - mora timing
- What kind of speech rhythm – 1/1, 2/4, 3/4, 4/4 ... ?
  - Embarrassing problem: physical correlates are elusive
  - Measures of relative regularity or irregularity of timing relative to the syllable or the foot - global numbers, no detailed information
  - Example:
    - *normalised Pairwise Variability Index* (*nPVI*):
    - average duration differences between neighbouring items
    - English: *nPVI* = 70 (*mean*=186ms, *SD*= 127ms, *coeffvar*=69%)
    - Mandarin: *nPVI* = 40 (*mean*=174ms, *SD*= 59ms, *coeffvar*=34%)

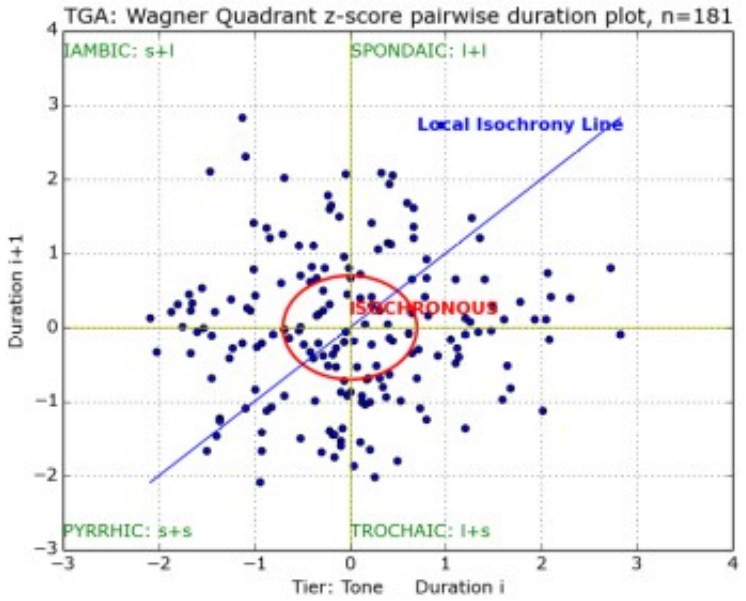
Automatic Annotation Mining: looking at the details (global descriptive statistics for British English recordings)

Duration properties (without pauses)			
Attributes	**Values**	**Attributes**	**Values**
n	227	*intercept*	198.289
min	10	*slope*	-0.106
max	940	*SD*	127.737
mean	186.26	*coeff var (%)*	68.582
median	150.0	*nPVI*	70
mean rate	5.37	*rPVI*	140
median rate	6.67		
total	42280		
range	930		

Automatic Annotation Mining: looking at the details (KWIC concordance of text labels in context, with statistics)

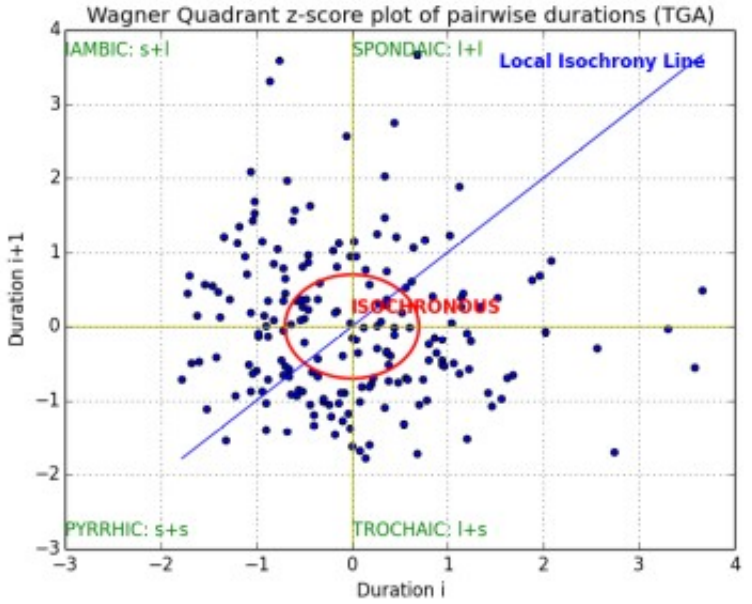
Type	N	min	max	median	mean	SD	CV%	Durations	Concordance
' @U	2	80	80	80.0	80.0	0.0	0.0	[80, 80]	I(60) ' w e n t(190) ' aU t(220) ' f A: s t(460) **' @U(80)** v @(90) D @(160) ' g l i:(130) m I N(160) ' s { n d(940) _(29) **' @U(80)** v @(60) r @(100) ' m I(150) d l(150) ' r i: (180) dZ @ n(200) w e(130) ' r Q k s(390) _(34)
' @U n	1	200	200	200.0	200.0	0.0	0.0	[200]	' w e n(130) I(30) w @ z(100) ' s @U(200) ' f A:(220) ' r aU t(210) D @ t(130) I(10) k @ d(170) ' l U k(170) ' b { k(210) ' n Q t(260) **' @U n(200)** l I(90) Q n(140) D @(40) ' l I(120) t l(130) ' b eI(290) _(33)
' D e n	1	250	250	250.0	250.0	0.0	0.0	[250]	@ n(80) **' D e n(250)** I(40) w @ z(140) I n(120) D @(80) ' r I@ l(400) ' s i:(480) _(612)
' D e@	2	130	350	240.0	240.0	110.0	45.0	[130, 350]	**' D e@(130)** S I(150) ' w Q z(460) _(569) @ t(110) ' b i:(60) I N(110) ' S U@(240) S I(130) w @ z(170) **' D e@(350)** _(724)
' D { t	1	320	320	320.0	320.0	0.0	0.0	[320]	@ n(100) b I(50) ' j Q n(320) **' D { t(320)** _(30)
' I n	2	90	170	130.0	130.0	40.0	30.0	[170, 90]	@ n(50) **' I n(170)** l e t s(300) _(29) I(100) ' r { n(280) ' s t r eI t(310) **' I n(90)** t @(70) D @(40) ' w O:(270) t @(90) _(37)

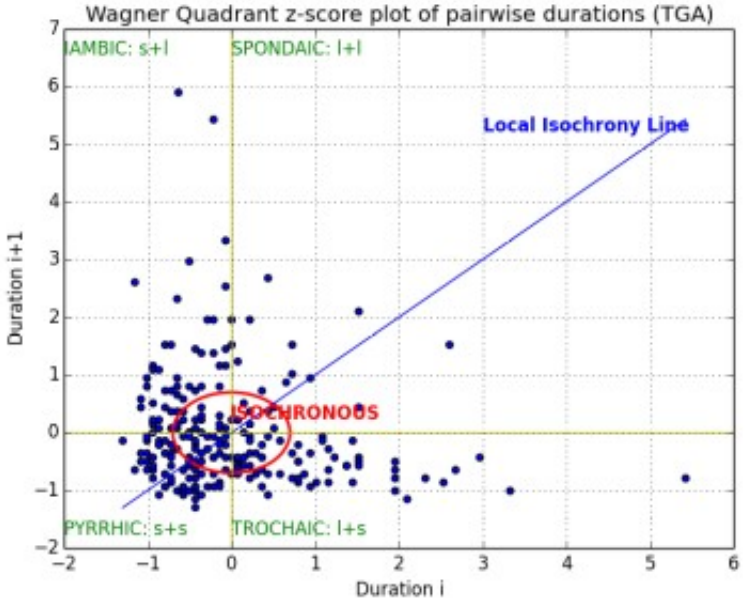
Automatic Annotation Mining: looking at the details (Wagner Quadrant Graphs for Mandarin, Tem and English)



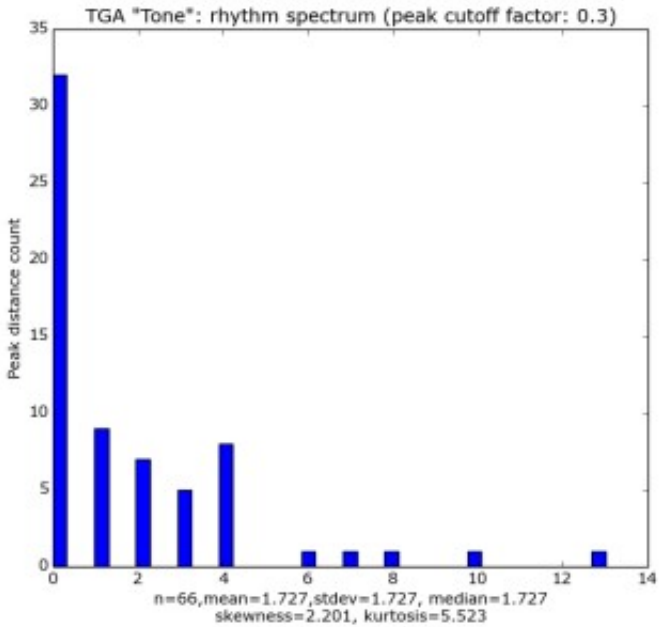
Duration relations between neighbouring syllables:

- $duration_n$ x $duration_{n+1}$
- note the distribution:
  - regular?
  - random?
  - clustering?
  - position?



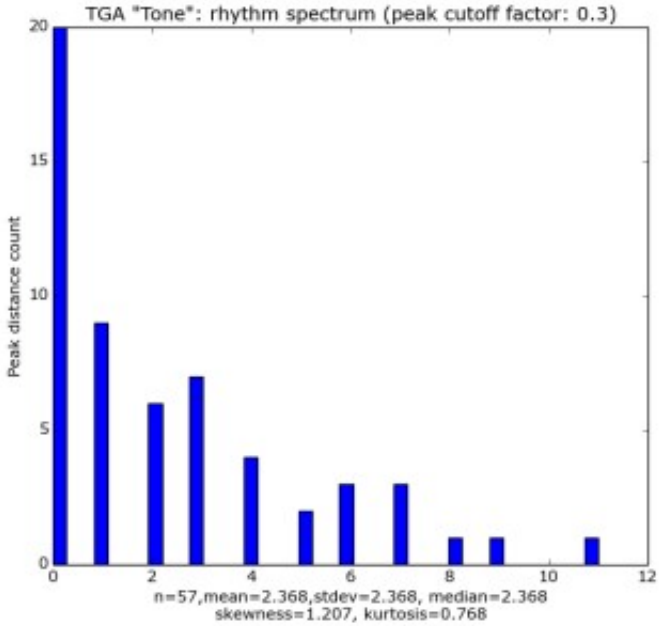


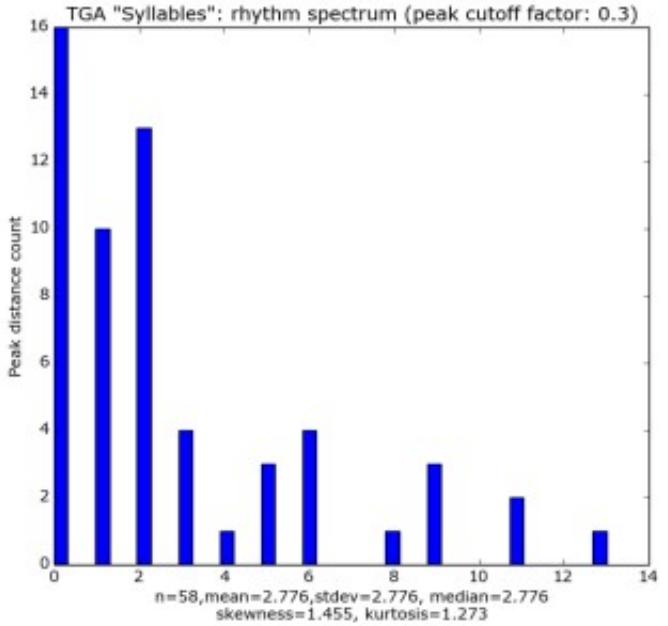
Automatic Annotation Mining: looking at the details (rhythm spectrum graphs for Mandarin, Tem and English)



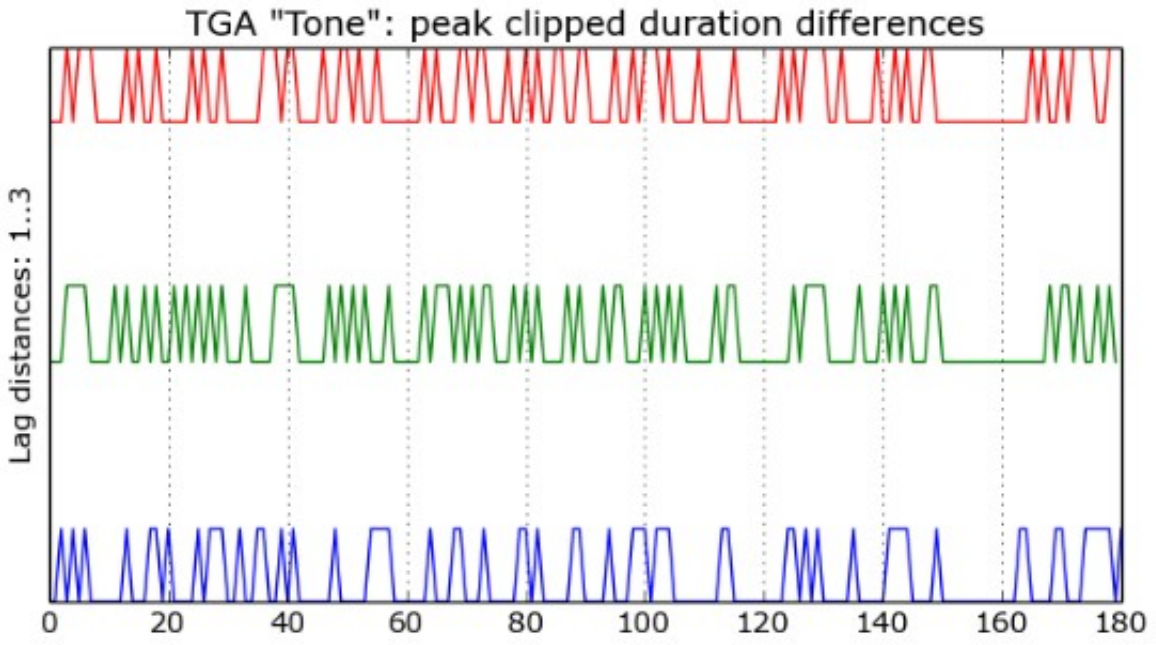
Numbers of duration maxima (peaks) at different item distances:

x = peak count
y = peak distance



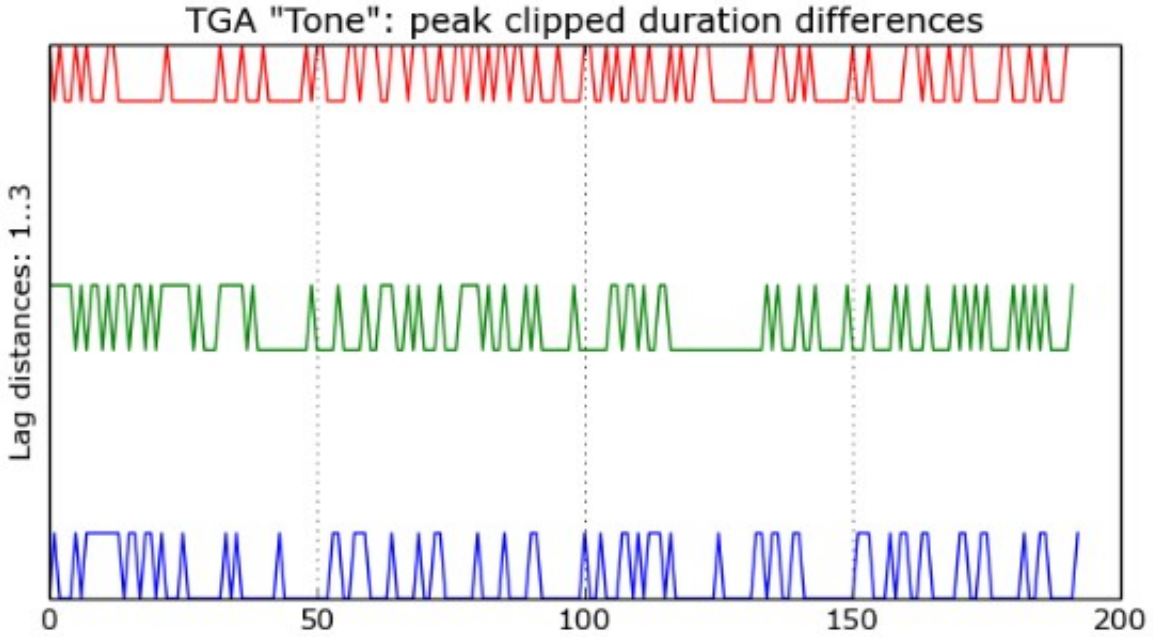


Automatic Annotation Mining: looking at the details (stylised duration patterns – the reality of rhythm)



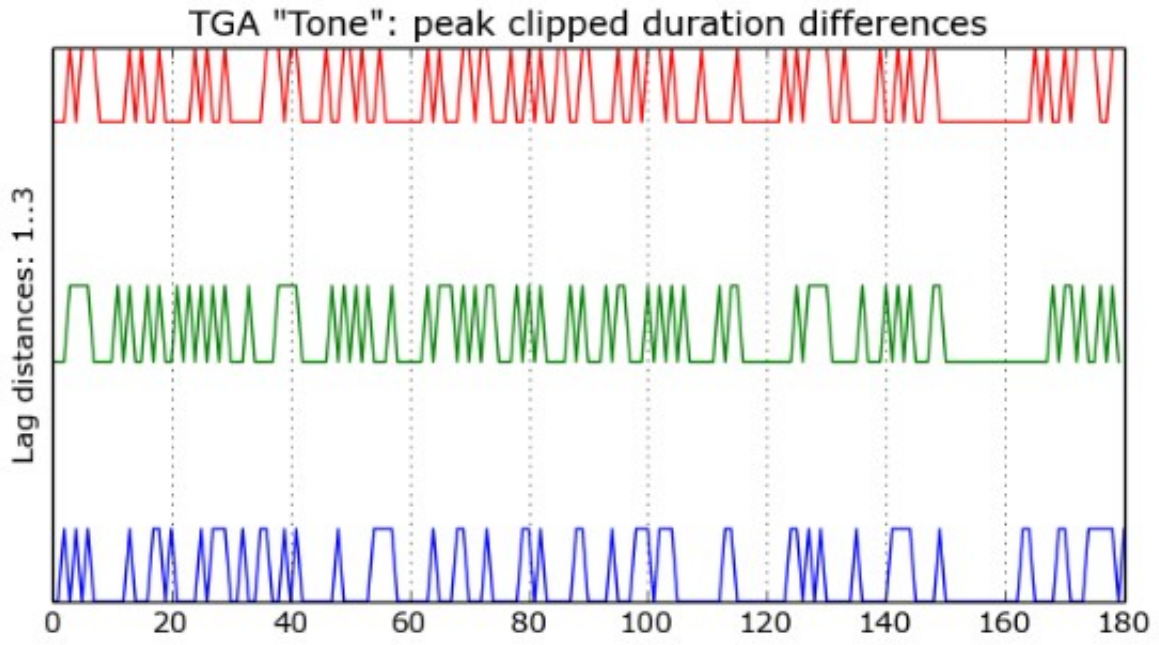
Mandarin

Peak sequences at distances 1,...,3



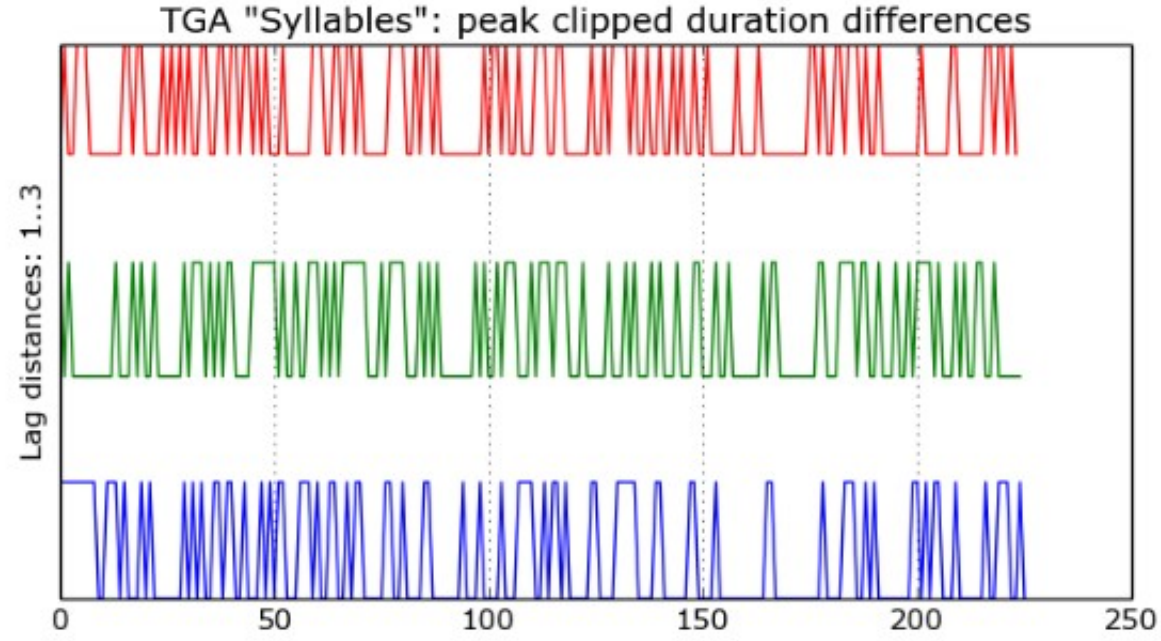
Tem

Automatic Annotation Mining: looking at the details (stylised duration patterns – the reality of rhythm)



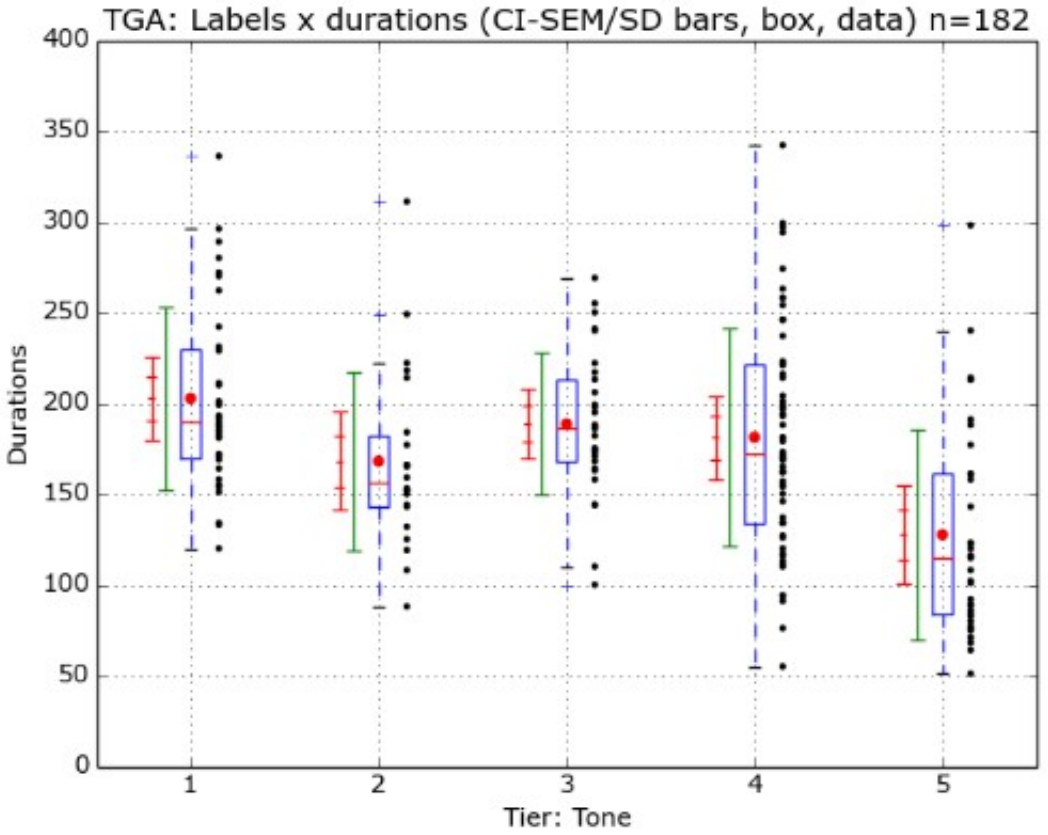
Mandarin

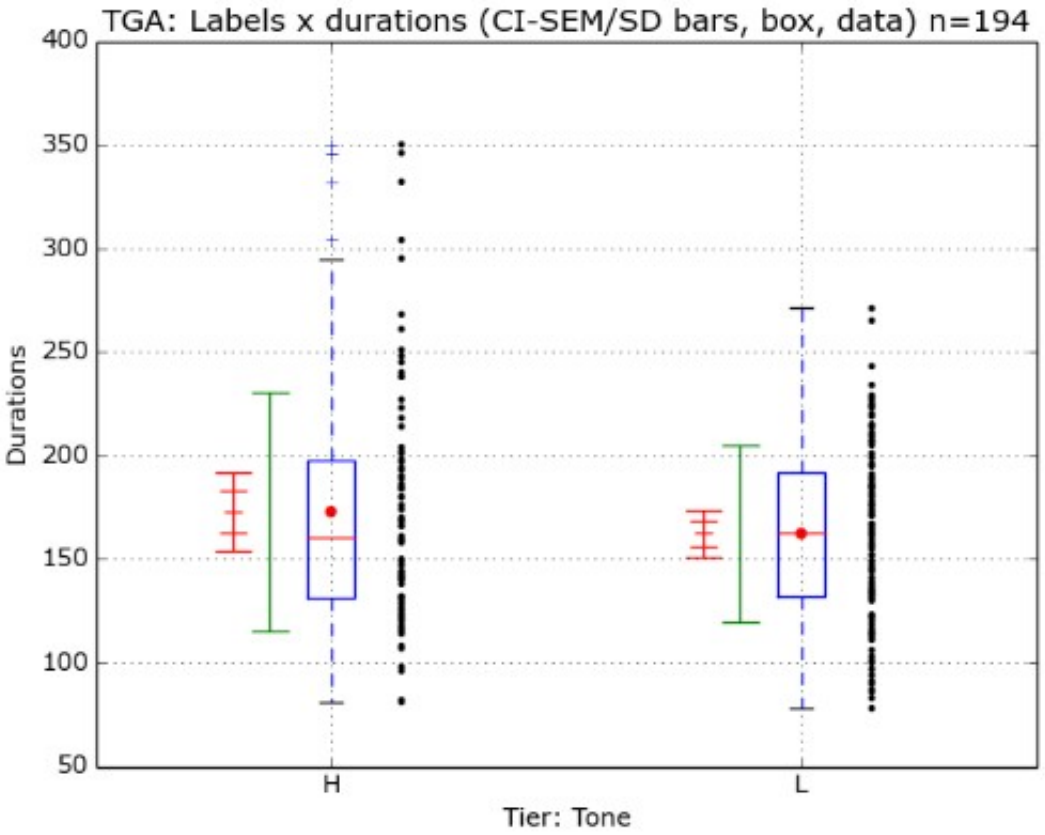
Peak sequences at distances 1,...,3



English

Automatic Annotation mining: tone in Mandarin and Tem



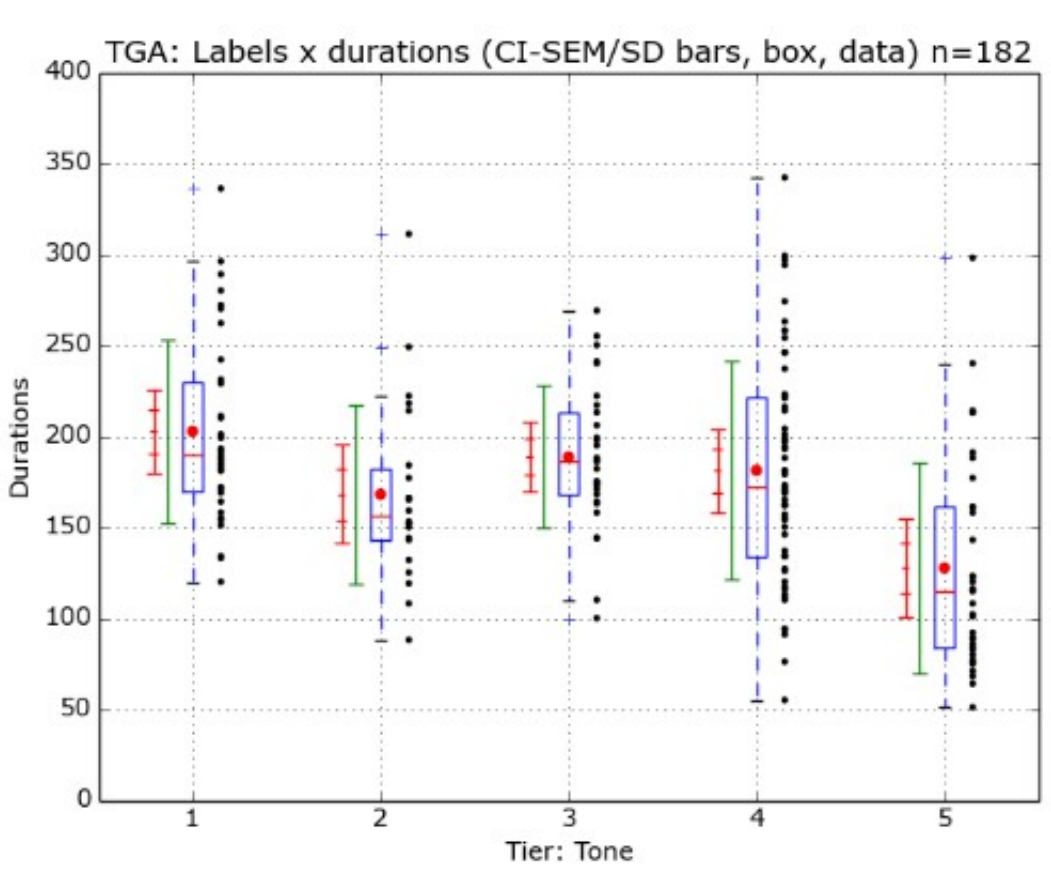


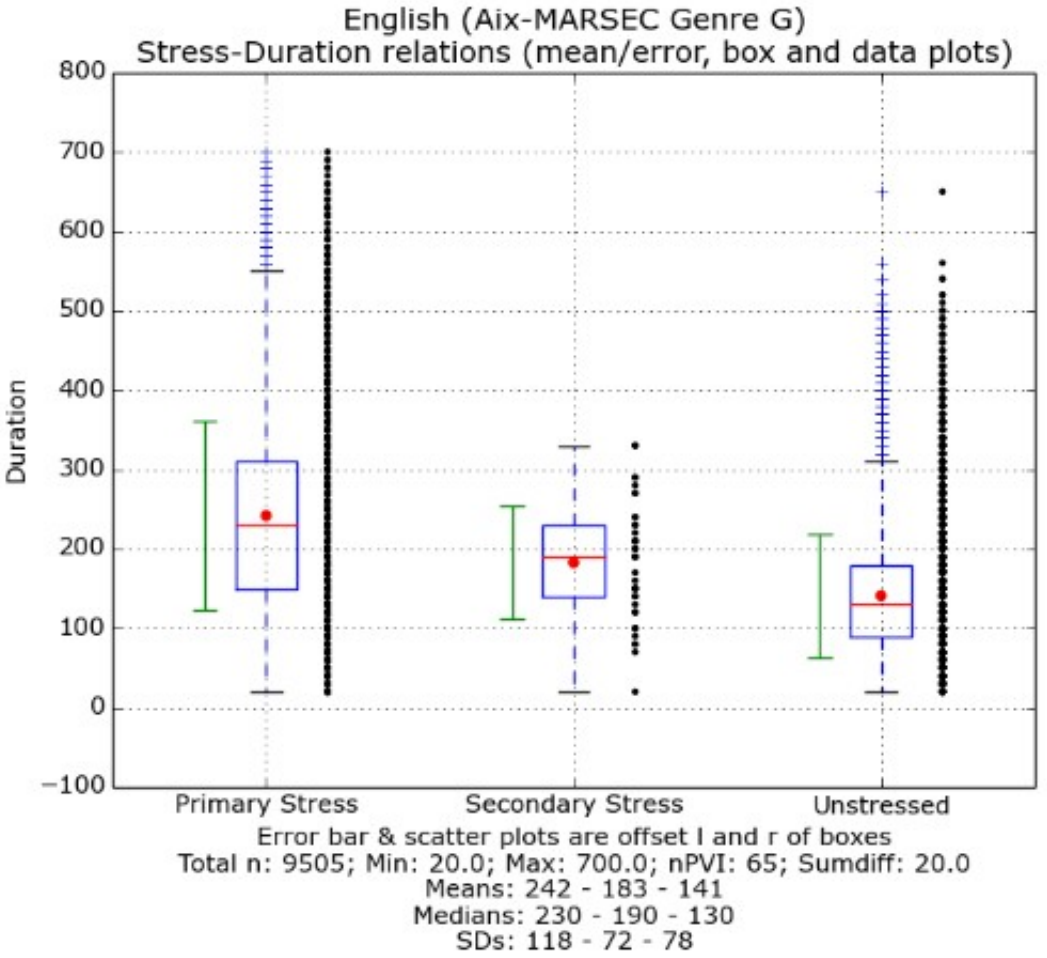
Mandarin
(difference for Tone 5, neutral tone)

Tem
(no difference)

Significance: natural TTS, e.g. reading functionality for the blind, voice information services (e.g. Apple Siri, Microsoft Cortana, Google Now)

Automatic Annotation mining: tone/stress in Mandarin, English





Mandarin
(difference for Tone 5, neutral tone)

English
(gradient difference tendency)

Significance: natural TTS, e.g. reading functionality for the blind, voice information services (e.g. Apple Siri, Microsoft Cortana, Google Now)

Time Group Analysis

TGA online annotation mining tool

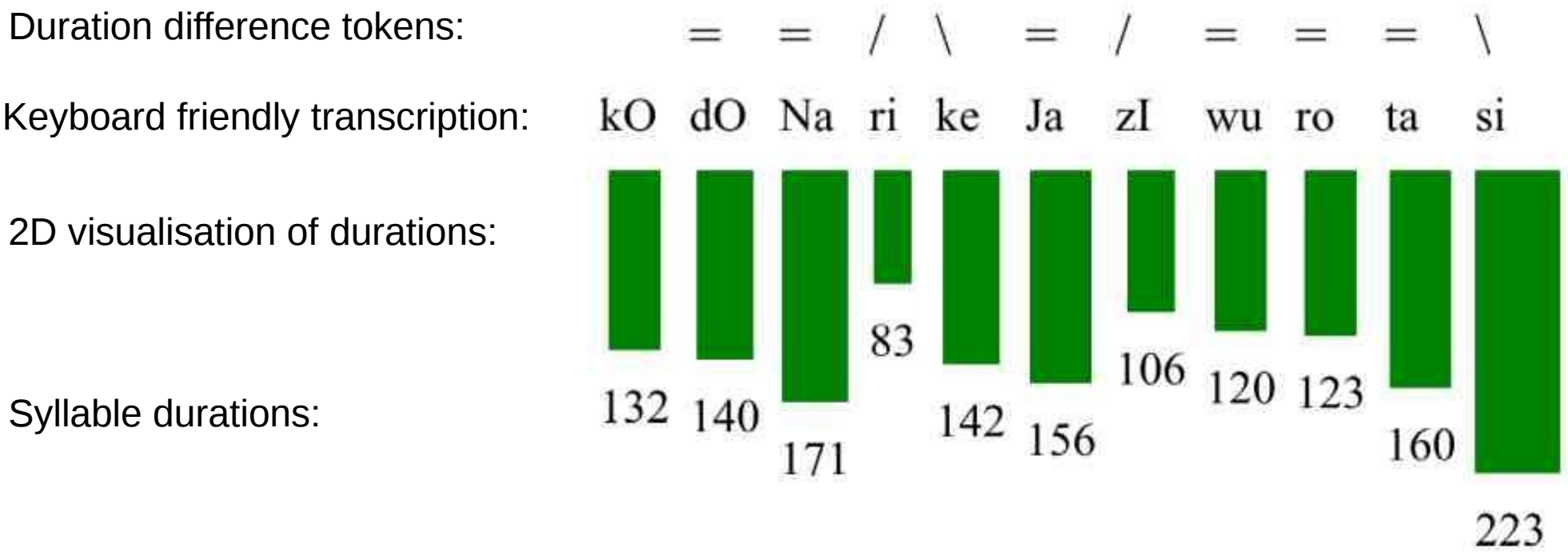
http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

http://localhost/TGA/

Annotation Mining: rhythm visualisation

TGA online tool: visualisation of syllable time relations

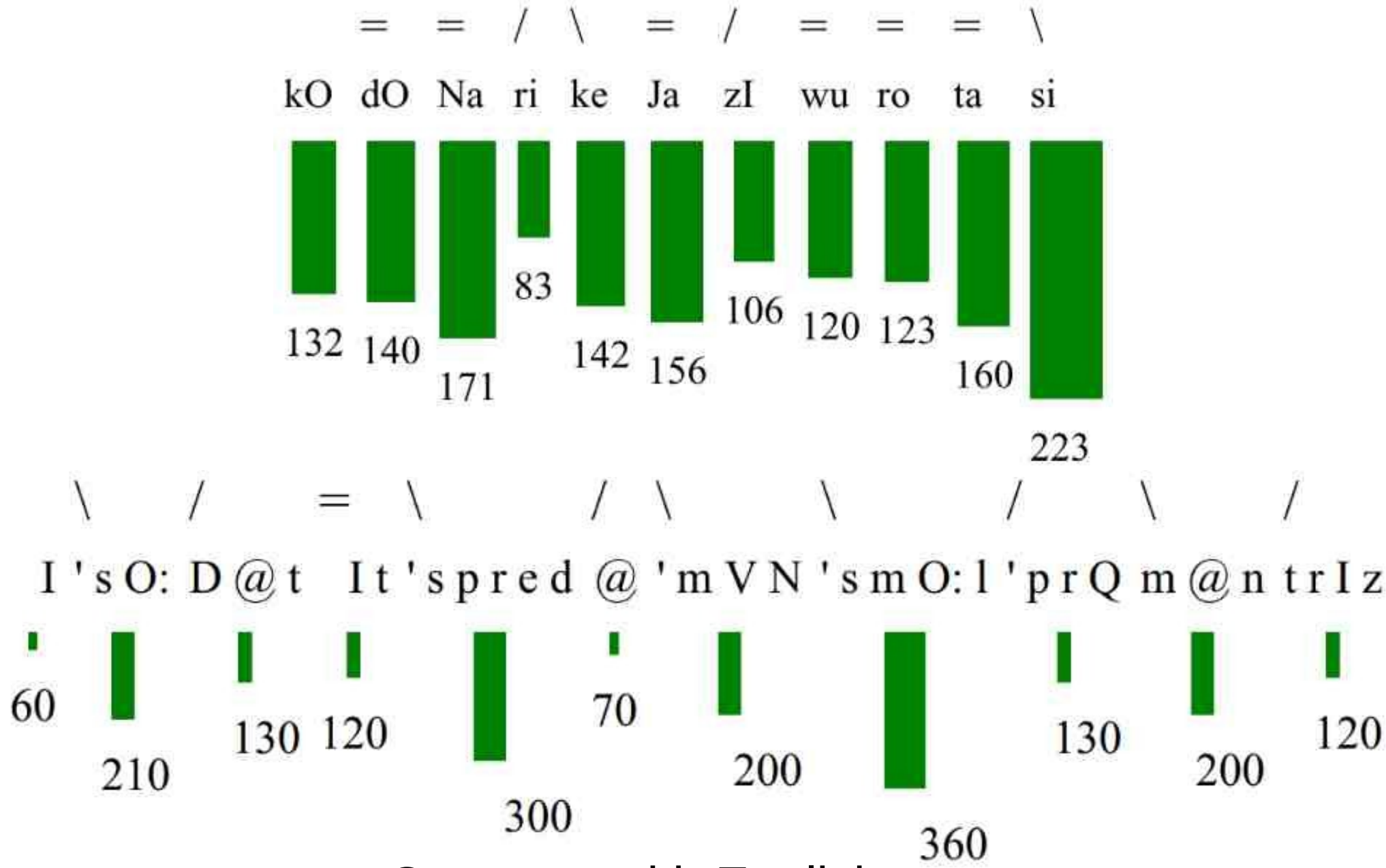
http://wwwhomes.uni-bielefeld.de/gibbon/TGA/



Duration difference tokens for utterance #37:

- pos, neg, equal differences between neighbours: /, \, =
- difference threshold: 40ms.
- clear indication of syllable isochrony

Annotation Mining: rhythm visualisation



Compare with English pattern

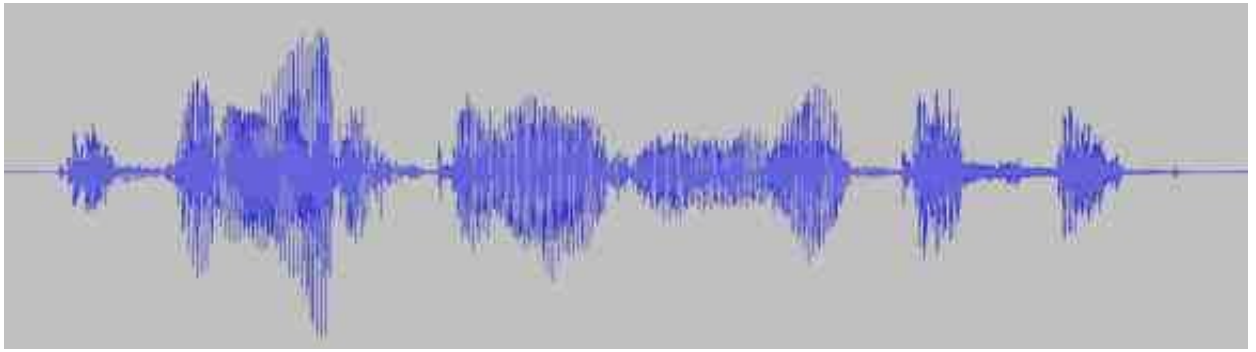
Enabling Technologies

Modelling tones: Niger-Congo > Gur > Tem (Togo)

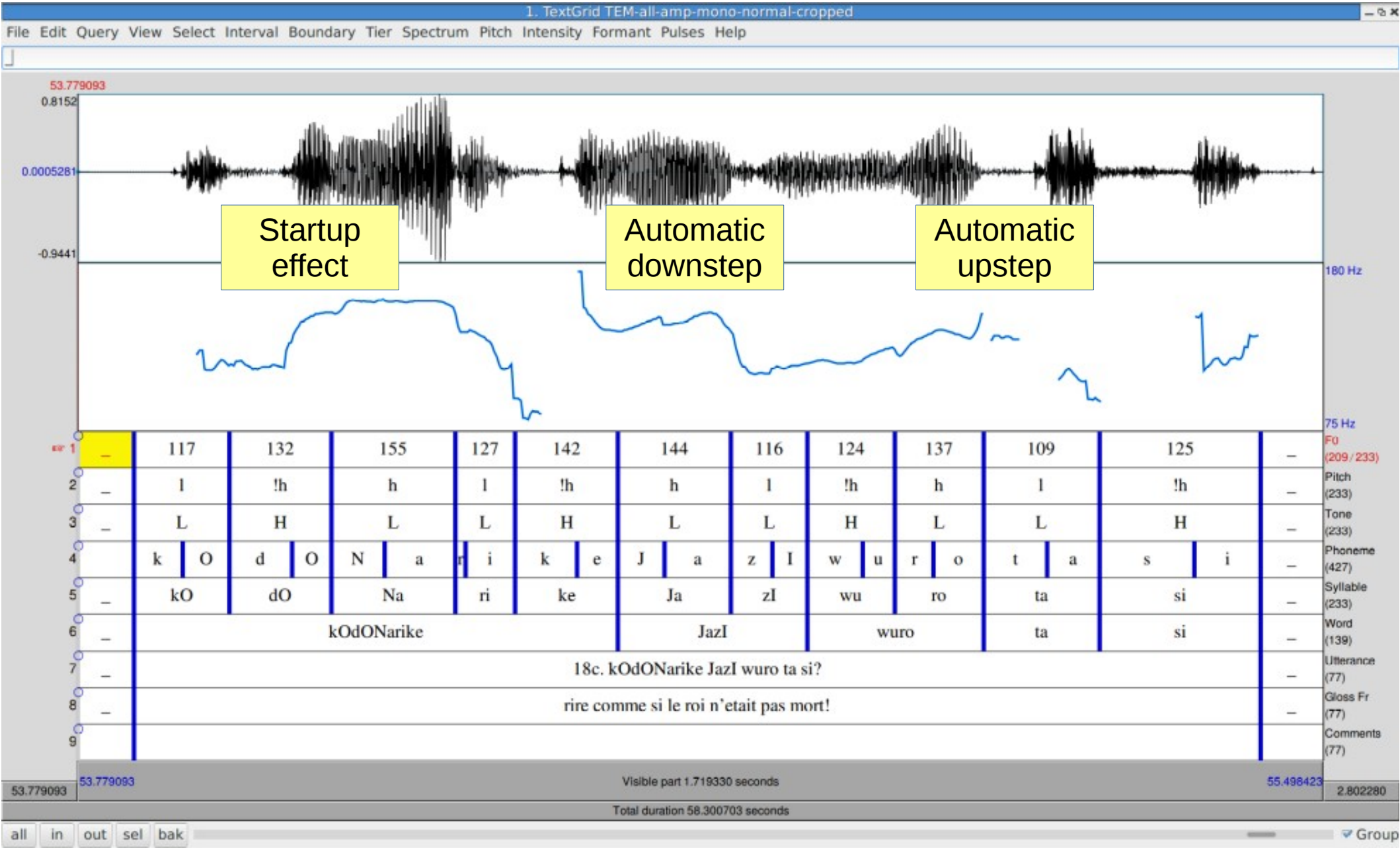
Annotation Mining: tone automaton induction for TTS

Resources:

- Zakari Tchagbale: “Exercices et corrigés” (1984)
- Dafydd Gibbon: finite state model of tone sandhi (1987)
- Zakari Tchagbale: recording (2012)



Annotation Mining: tone automaton induction for TTS



Annotation Mining: tone automaton induction for TTS

Tone	Pitch	*N*	mean (Hz)
H	!h#	10	128
	h#	9	129
	#h	14	154
	!h	56	131
	h	28	144
L	^l#	9	93
	l#	10	98
	#l	24	115
	^l	50	139
	l	60	113

Tone	Mean F0 (Hz) in sequential contexts				
	initial	overall	step	final	step final
H	154	144	131	129	128
L	115	113	139	98	93

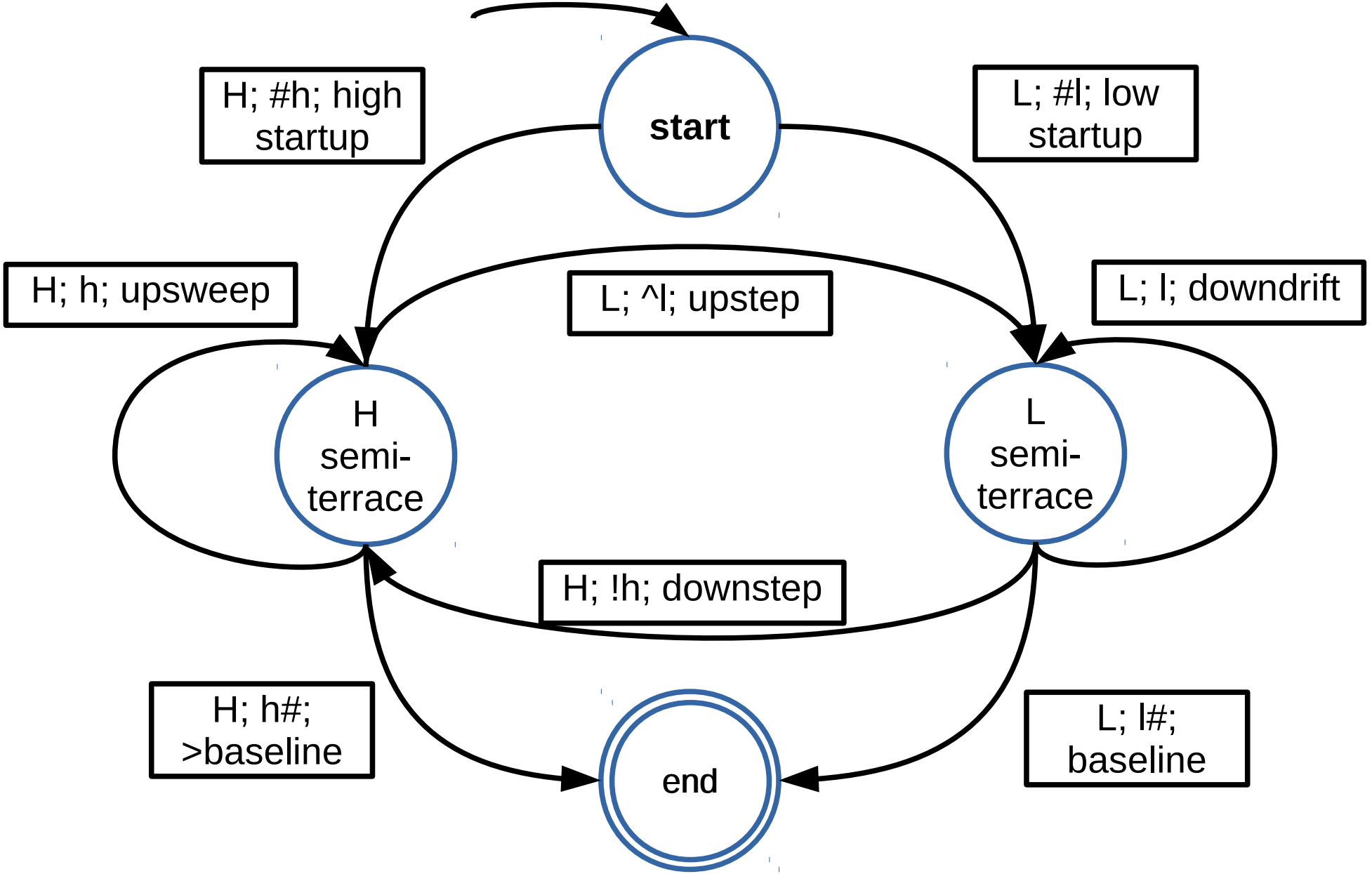
Annotation Mining: tone automaton induction for TTS

Tone co-occurrences

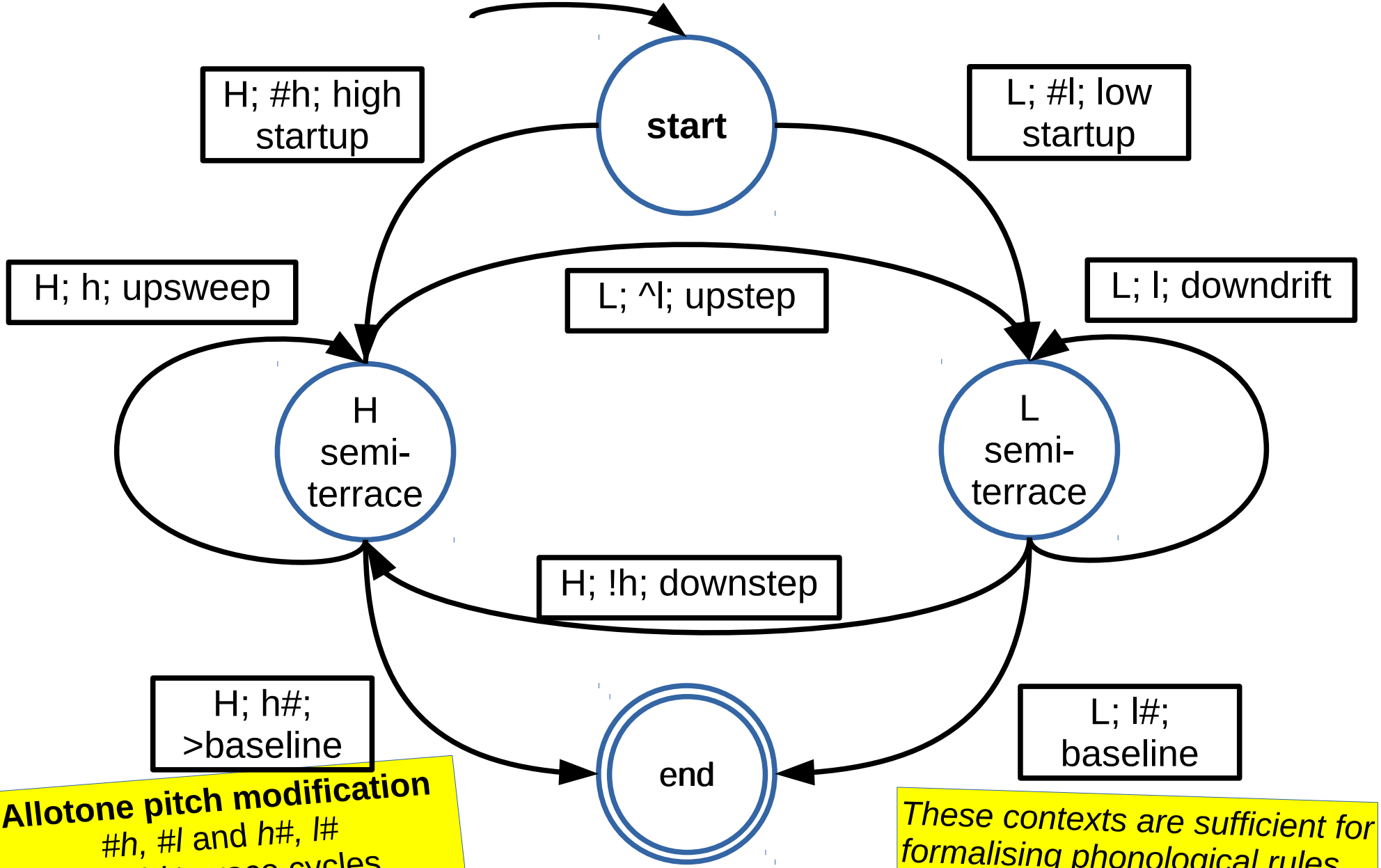
Tone	Pitch	*N*	mean (Hz)
H	!h#	10	128
	h#	9	129
	#h	14	154
	!h	56	131
	h	28	144
L	^l#	9	93
	l#	10	98
	#l	24	115
	^l	50	139
	l	60	113

Tone	Mean F0 (Hz) in sequential contexts				
	initial	overall	step	final	step final
H	154	144	131	129	128
L	115	113	139	98	93

Case 2: Allotones: Tem (Gur; ISO 639-2 kth) resources



Case 2: Allotones: Tem (Gur; ISO 639-2 kth) resources



Allotone pitch modification
#h, *#l* and *h#*, *l#*
h and *l* terrace cycles
^l and *!h* terrace transitions

These contexts are sufficient for formalising phonological rules. More contexts are needed for natural phonetic detail!

Enabling technologies

Language similarity and difference

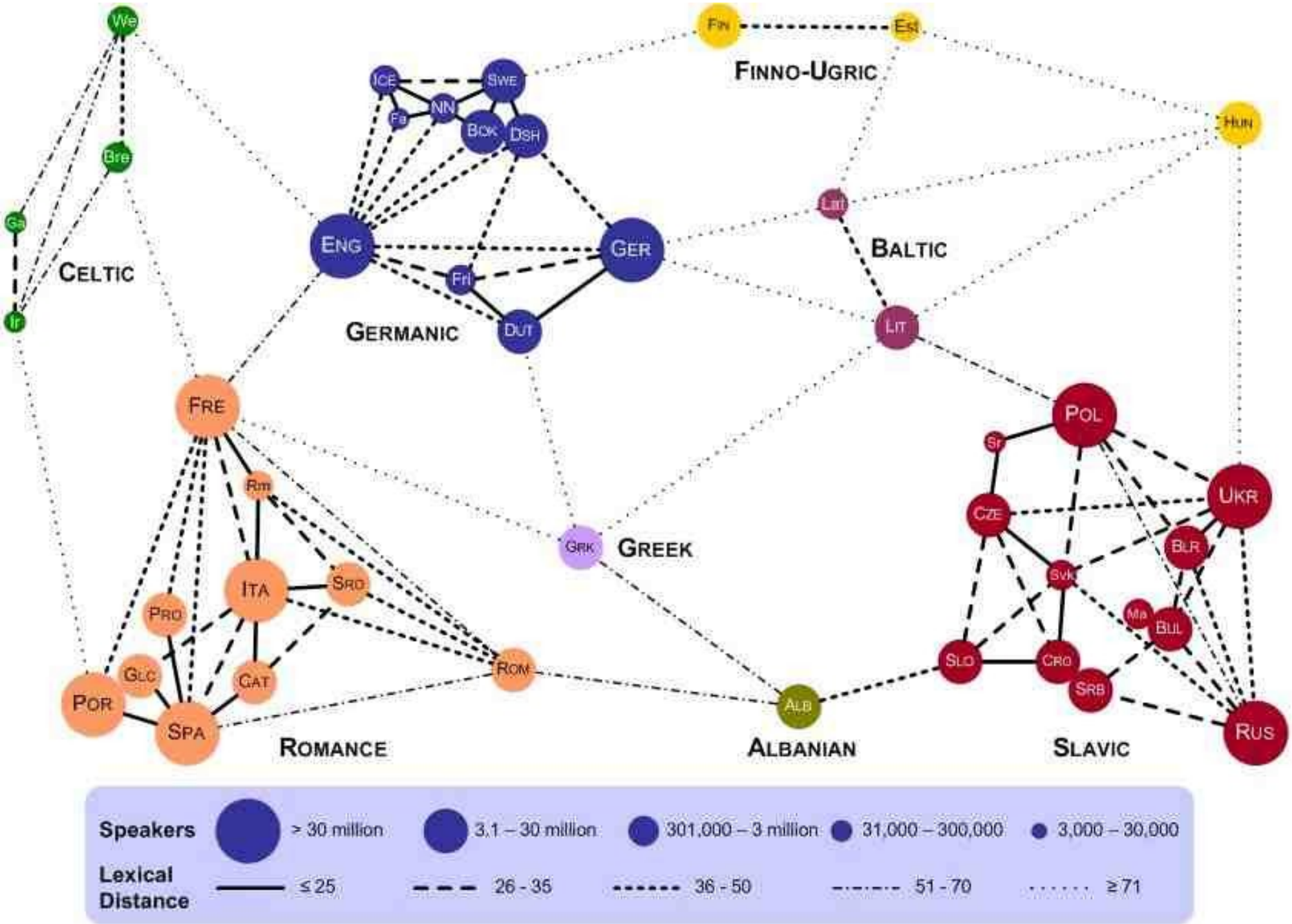
as virtual distance

Towards the use of Machine Learning

Enabling technologies: prioritisation

- Importance of determining language similarities and dissimilarities:
  - re-use of existing materials in different communities
    - beware of problems (e.g. Ibibio / Efik)
  - adaptation of existing productivity technologies
  - prioritisation of documentation funding
  - ...

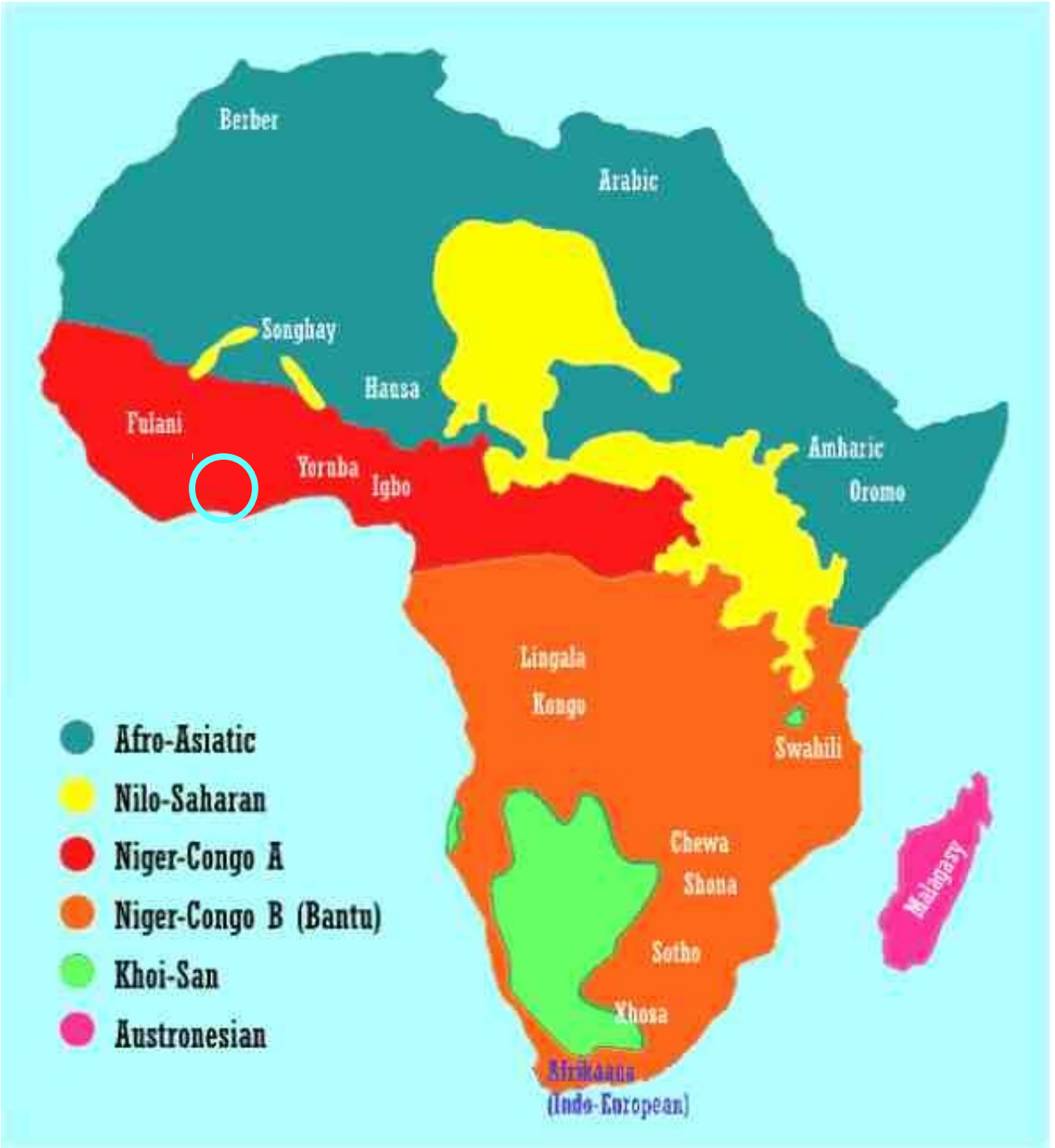
How similar are languages? - A little similar, but not very similar



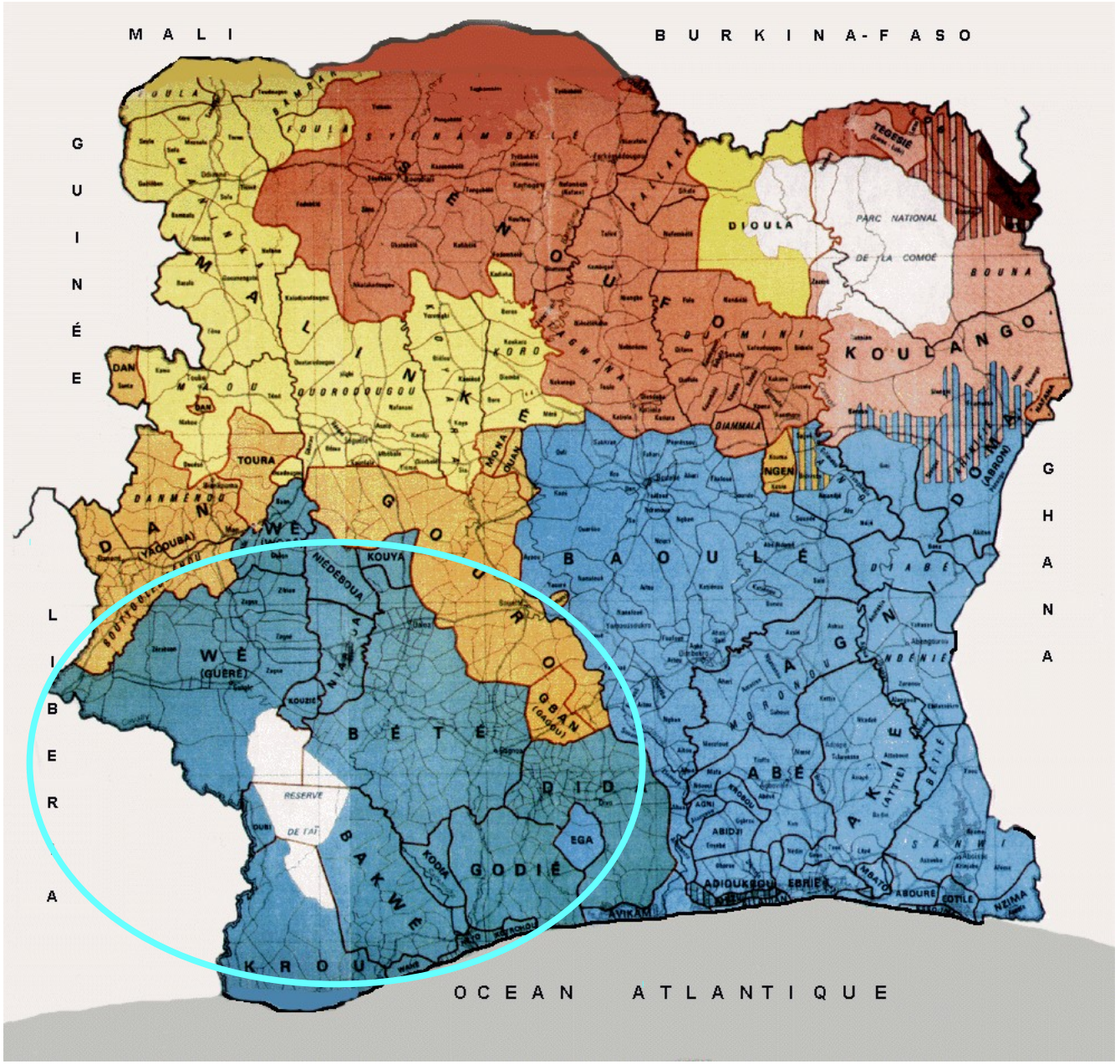
Selecting features for similarity distances

- 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)

Kru languages – Ivory Coast: Method – consonant systems



Kru languages – Ivory Coast: Method – consonant systems



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 _ _ _ _ _ _ _ _ _ _ _ _

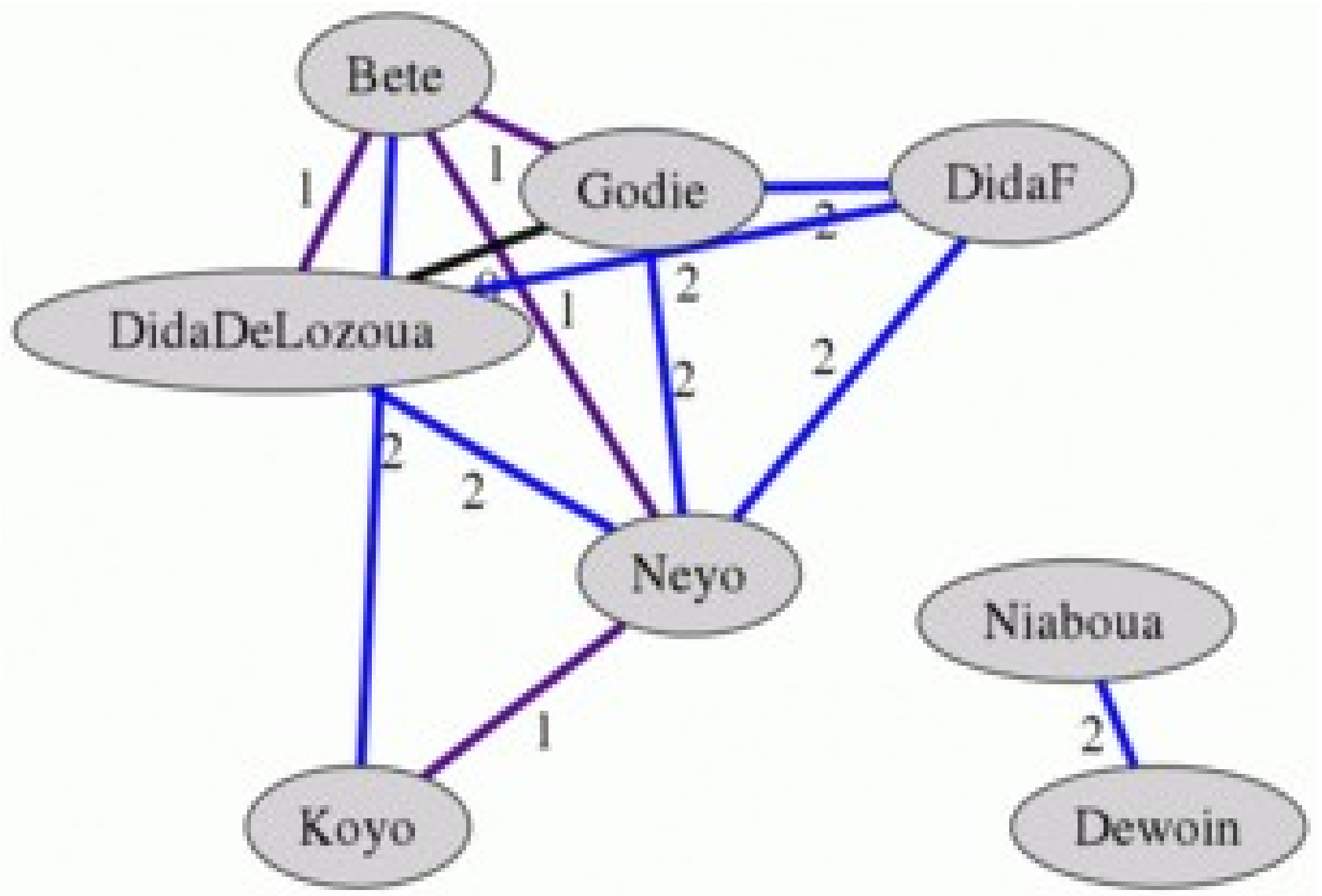
Method:

1. compare all consonant systems pairwise:
   Levenshtein Edit Distance / Hamming Distance
2. visualise differences as ‘virtual distances’ in a chart

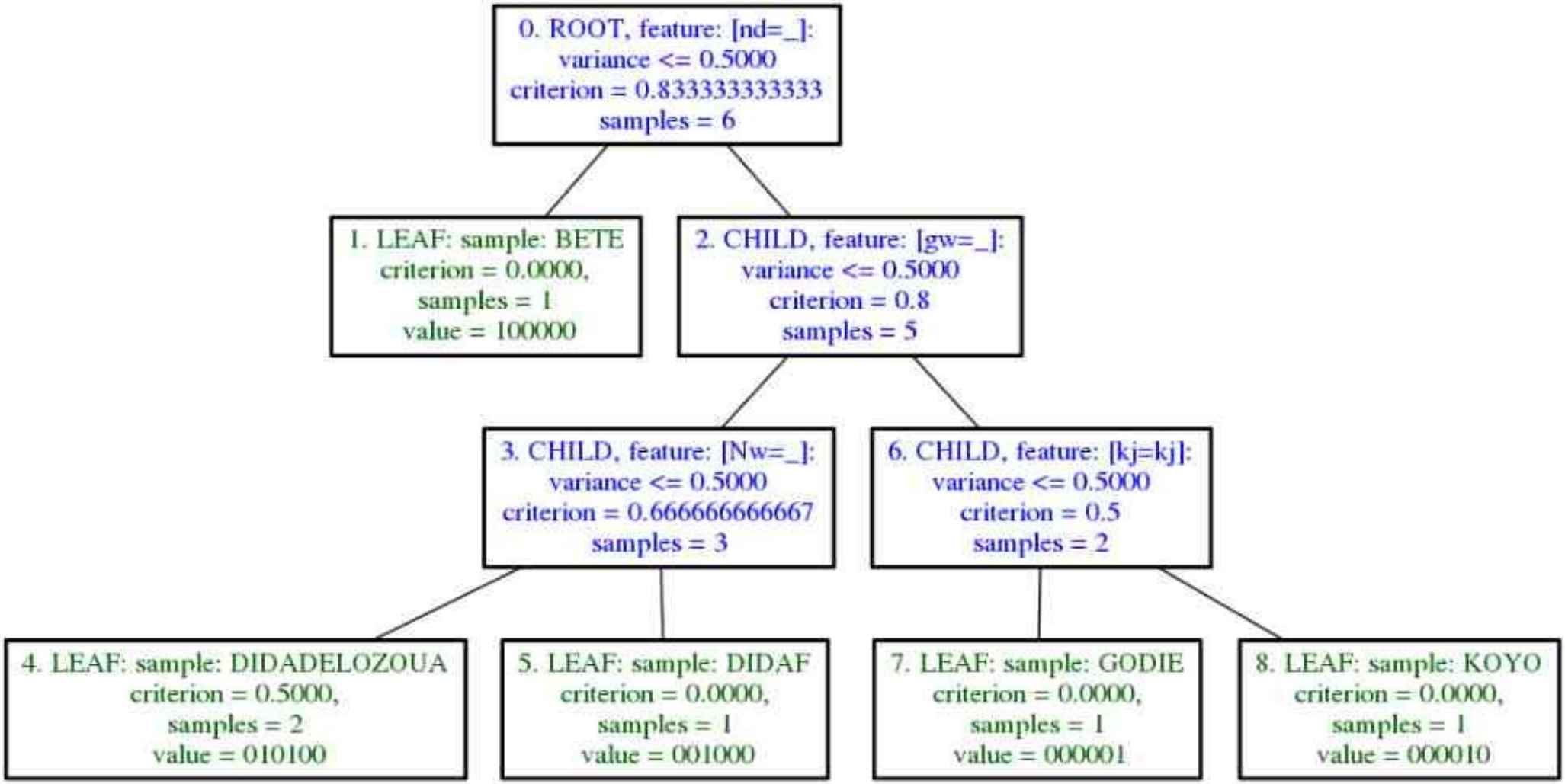
Kru languages – Ivory Coast: Feature – consonant systems

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)



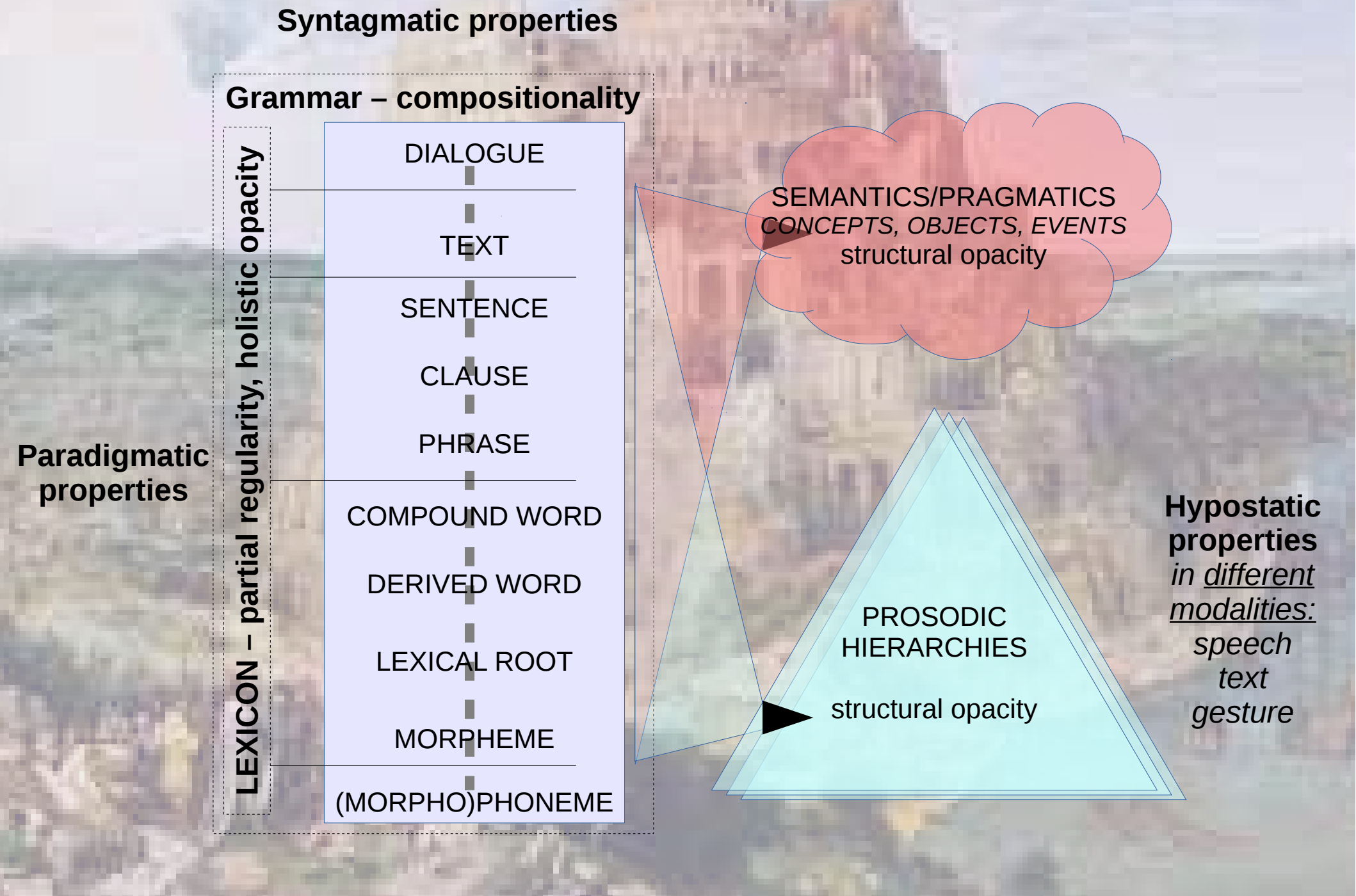
Conclusion

What do we lose if we lose languages?

What do we lose if we lose languages?

- Why is this an issue? What danger?!
  - Imagine what we lose if a language disappears...
  - Why not just use one language worldwide?
  - Remember what language is for!
- Imagine what we lose if a language disappears:
  - Structure: understanding of the architecture of the human mind – each language has a unique combination of cognitive patterns
  - Language semantics: a language is a repository of knowledge about the world – skills, health
  - Pragmatics: narratives of history and identity of a language community, law, literature (with music and art), religion, ...
- Why not just use one language worldwide?
  - Lack of interregional communication means that the shared language will develop local differences anyway ...
  - It's boring ...

Language is complex: Ranks, Interpretations, Languages, Context




Diolch yn fawr!