Digital Humanities

Processing Text and Speech

Dafydd Gibbon

Universität Bielefeld
Germany

AMU Poznań, 2015-12-01

Overview

- Digital Humanities: “Metamethodology for the Humanities”
- Technologies and Languages in Africa
- From fieldwork to Digital Humanities
- Tongue in cheek retrospective ...
- A personal view of the ancestry of Digital Humanities
- Humanities + Technology common ground: high quality data - ‘big data’
- Case Study 1: Types of readers
  Based on data provided by Maya Nikolova
- Case Study 2: Phonological typology of Kru languages (Ivory Coast)
- Case Study 3: Evaluation of spoken discourse transcribers
  Based on data provided by Jolanta Bachan
- Case Study 4: Recovery and technological application of legacy data
  Based on data provided by Zakari Tchagbale
- Conclusion

Digital Humanities:

"Computational Metamethodology for the Humanities"

(Prof. Harold Short, King's College, London)

with applications ranging from literary style analysis through manuscript recovery and preservation to archaeology and geography.
The applications involve recovery and enhancement of images, visualization of stylistic and other relationships between texts and authors, machine learning in order to find generalization hypotheses efficiently.

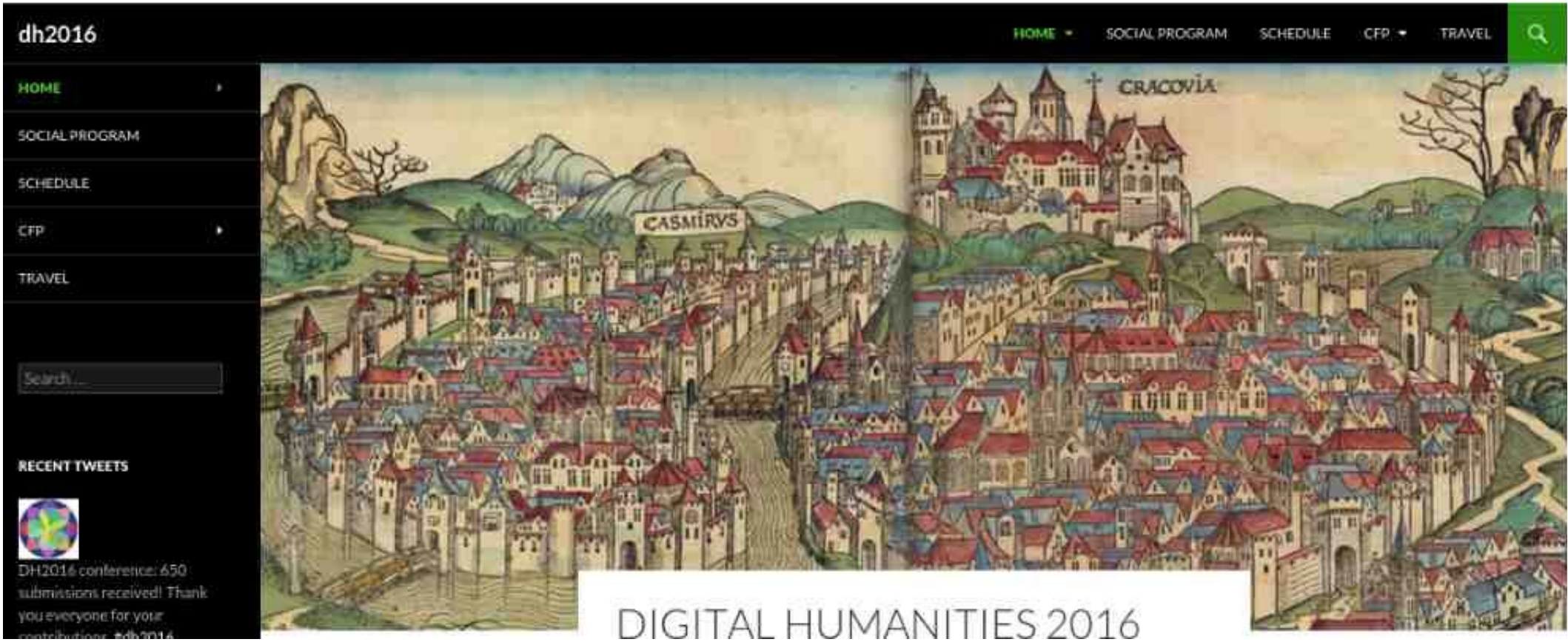
Association of Digital Humanities Organisations (ADHO):

- The European Association for Digital Humanities (EADH)
- Association for Computers and the Humanities (ACH)
- Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
- centerNet
- Australasian Association for Digital Humanities (aaDH)
- Japanese Association for Digital Humanites (JADH)

International conference every two years

Digital Humanities 2016

Kraków, 12–16 July



ICT [Information and Communication Technologies] can either spell doom for our languages and the cultures they bear, knocking the death knell or they can provide new avenues and media to invigorate them by opening up new windows of orality on one hand and redefining literacy on the other. (p. 8)

Adegbola, ‘Tunde. 2006. Globalization and the future of African Languages. In: F. Egbokhare and Clement Kolawole, eds. *Globalization and the Future of African Languages*. Ibadan: Ibadan Cultural Studies Group, pp. 2-10.

Dr. Adetunde Adegbola is a research scientist, consulting engineer and culture activist. He is Director of the African Language Technology Initiative (Alt-i), Ibadan, Nigeria, with degrees in Engineering, Computer Science and Computational Linguistics. His team has made many contributions to the digital empowerment of Yoruba, Igbo and other Nigerian languages through text and speech technologies.

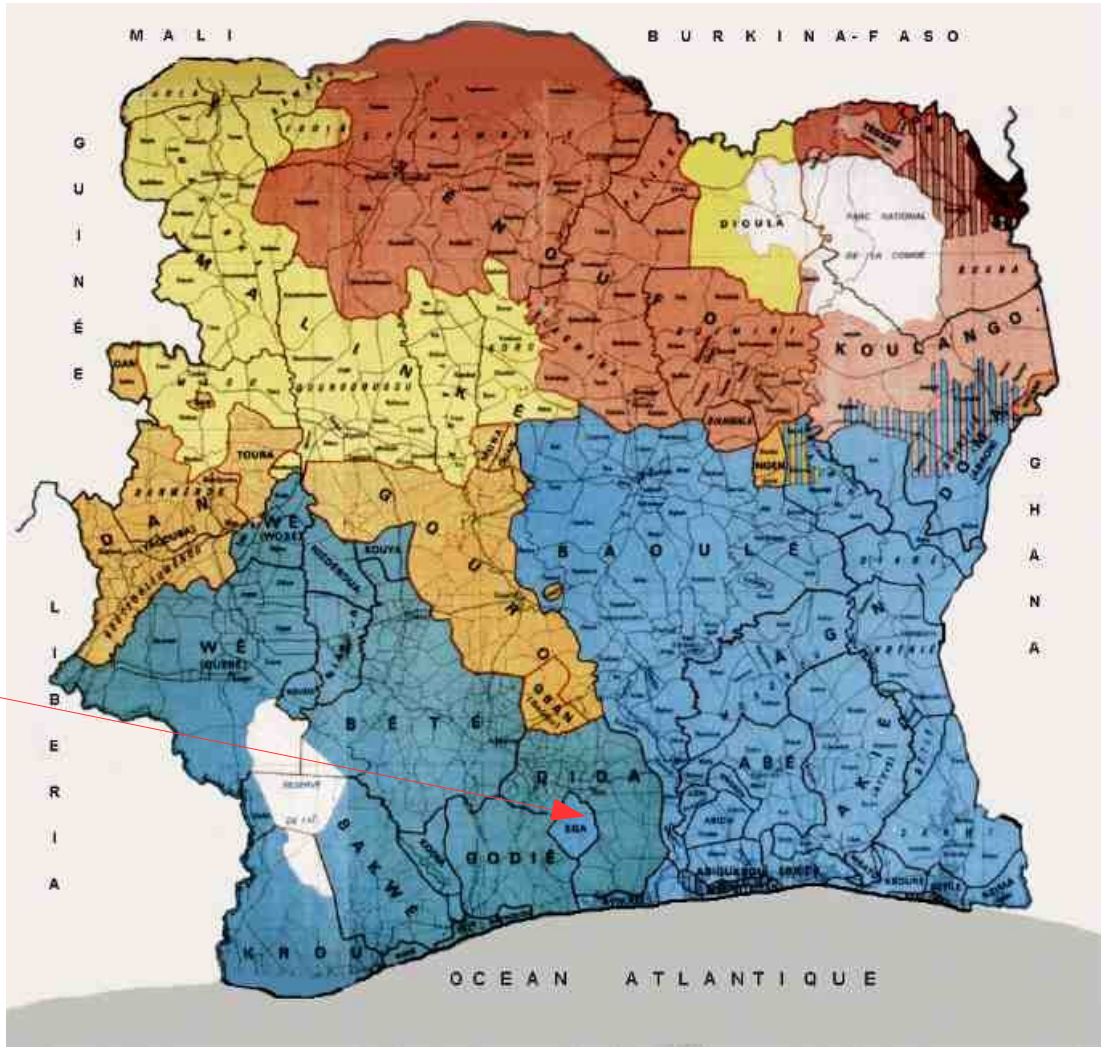
Background

Much of my work in joint projects with African (and other) colleagues and students over the past 30 years has been guided by the issues noted by Tunde Adegbola, particularly in the form of:

- Scientific cooperation and facilitation:
  - prosodic features of Niger-Congo languages in West Africa, from a computational linguistic perspective
- Infrastructural cooperation and facilitation:
  - training in modern language and speech technologies
  - development of computational tools for resource creation:
    - speech analysis – timing and tone
    - speech synthesis
    - lexicography

This presentation will reflect these interests and commitments in the framework of Digital Humanities.

Background



My work in teaching and in the field in West Africa has been mainly mainly in Côte d'Ivoire and Nigeria.

From fieldwork to Digital Humanities

A medium term goal:
Analysis of oral literature in a Digital Humanities framework, at present in a DAAD project with colleagues in Abidjan and Bielefeld.

Current status:
Phonetic analysis; sound-gesture synchronisation in text and music.

Tongue in cheek retrospective ...

I specialise in computational linguistics.

Apparently I have been a Digital Humanities specialist for the past 35 years – unknowingly ☺

- 1979: visualisation of the rhythm and melody of speech
- 1980: visualisation of English tense semantics, for teaching
- 1988: conference contribution 1991 book chapter on software as text
- 1990s: online lexicography and concordance building
  - http://www.spectrum.uni-bielefeld.de/VM-HyprLex/
  - Eliot, Old Possum's Practical Cats: 'And'
  - Samuel Beckett, A Piece of Monologue: 'Night'
- 2000s: documentation projects on endangered languages
- 2010s: multimodality; gesture, text and music; online tools for similarity analysis, text alignment, syllable visualisation

1988

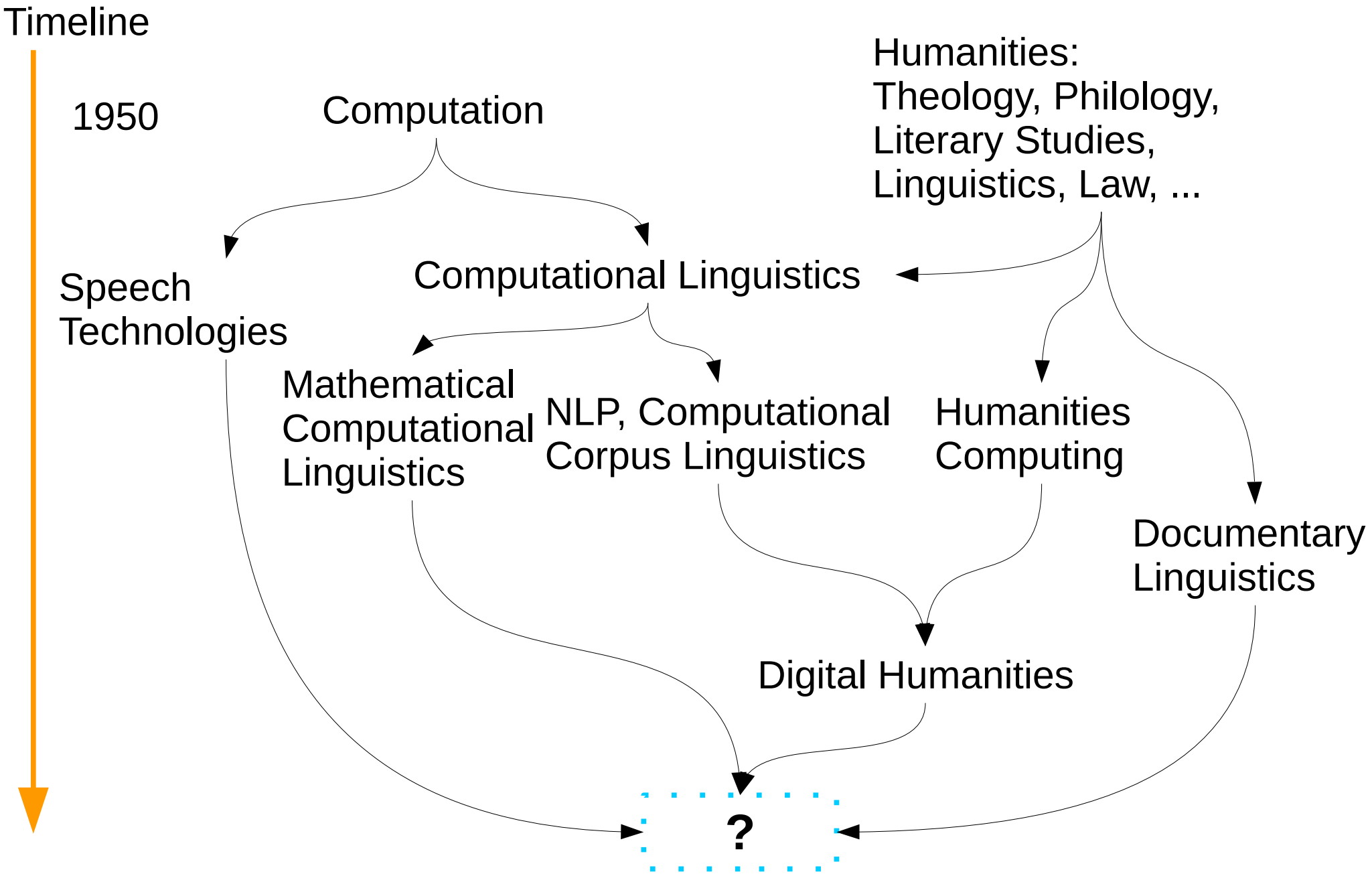
Christiane Floyd Heinz Züllighoven
Reinhard Budde Reinhard Keil-Slawik
(eds.)

Software Development and Reality Construction

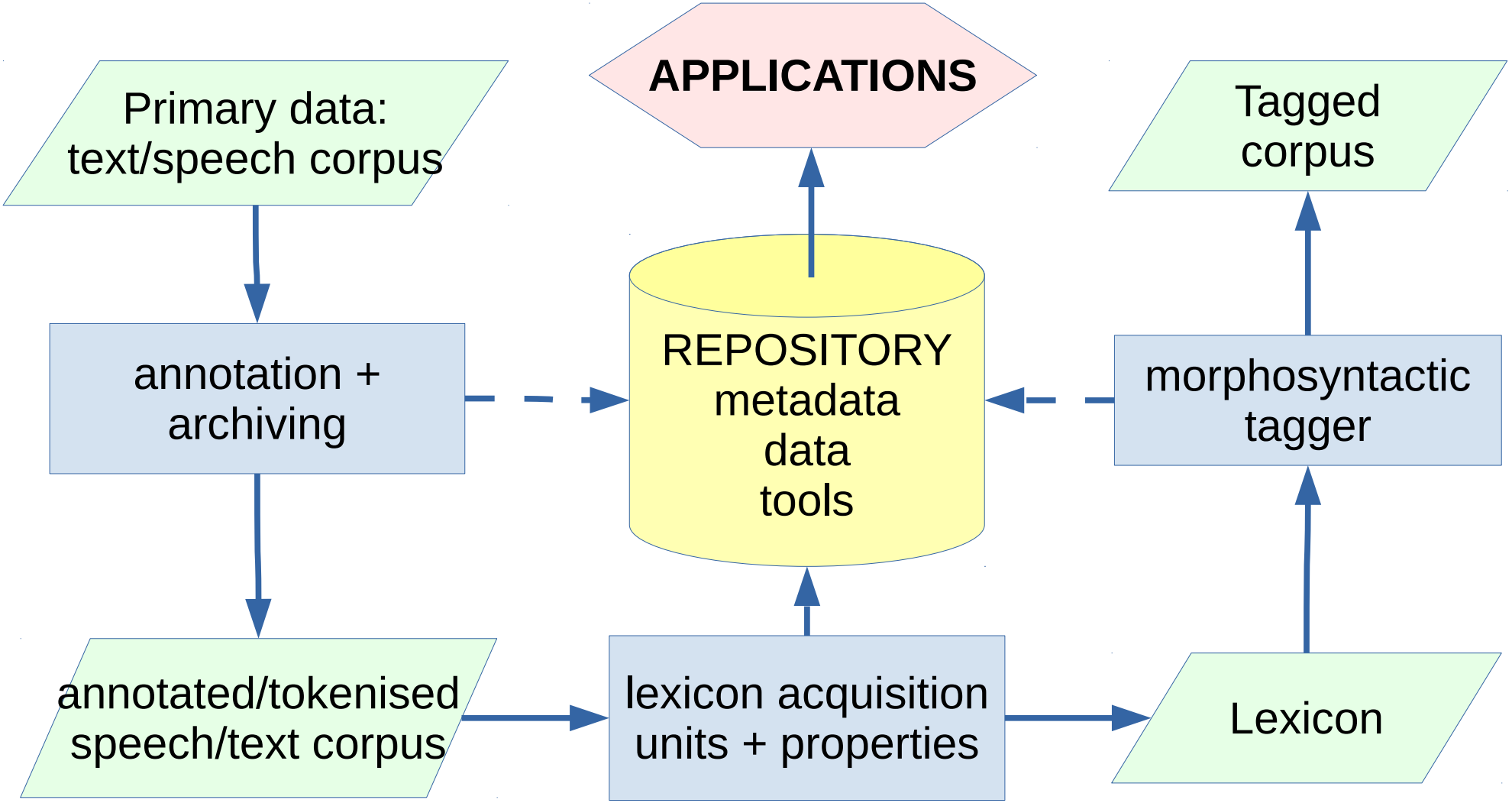
With contributions by

Klaus Amann, Gro Bjerknes, Rodney M. Burstall, Rafael Capurro, John M. Carroll, Wolfgang Coy, Bo Dahlbom, Wolfgang Dzida, Heinz von Foerster, Klaus Fuchs-Kittowski, Dafydd Gibbon, Joseph A. Goguen, Thomas F. Gordon, Pentti Kerola, Heinz K. Klein, Donald E. Knuth, Klaus-Peter Löhr, Kalle Lyytinen, Susanne Maaß, Markku Nurminen, Kristen Nygaard, Horst Oberquelle, Arne Raeithel, Fanny-Michaela Reisin, Douglas T. Ross, Dirk Siefkes, Jouini Similä, Walter Volpert

A personal view of the ancestry of Digital Humanities

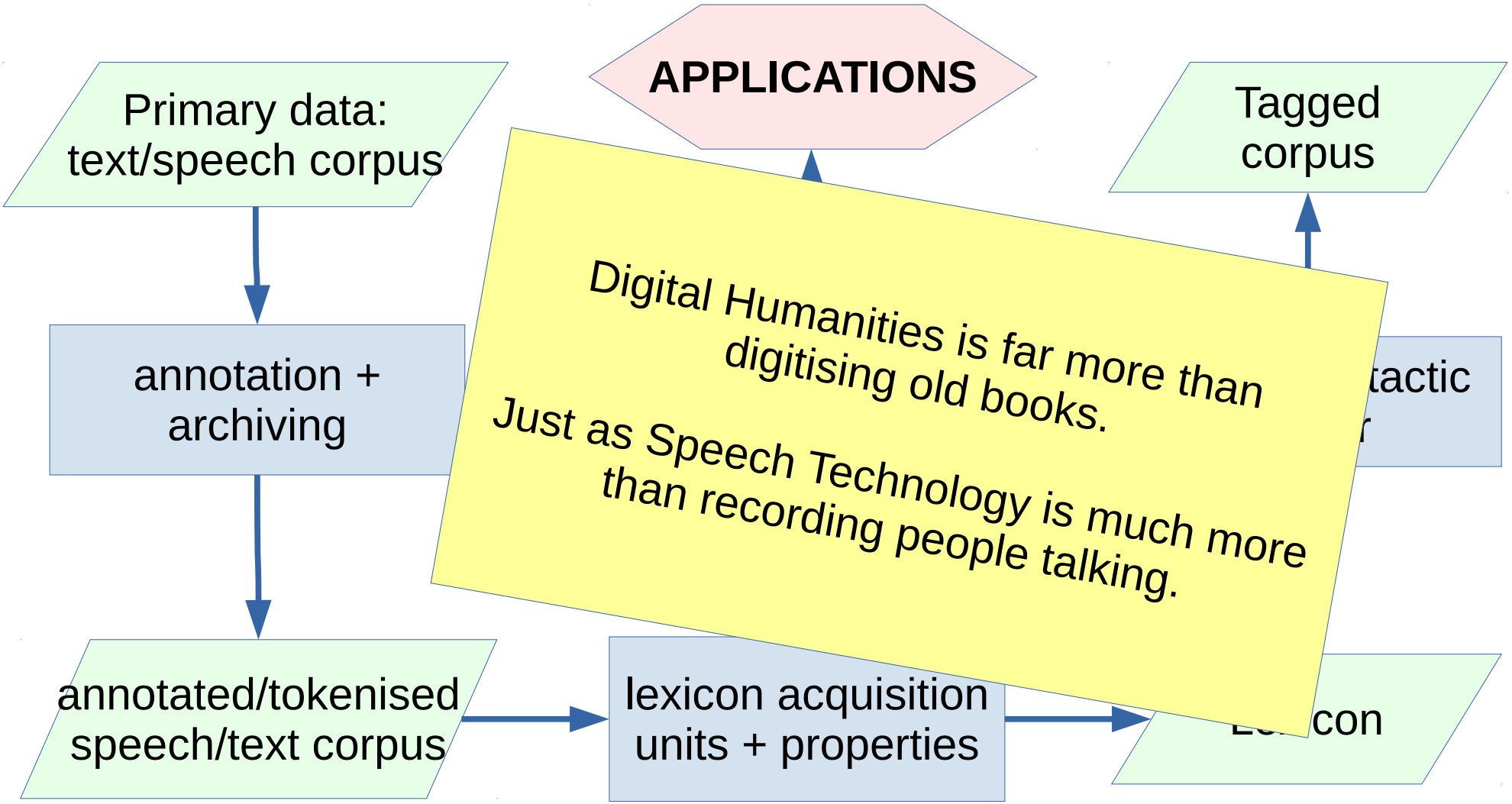


Humanities + Technology common ground: high quality data - 'big data'



One of the task areas of a Resource Management Agency

Common ground: high quality data - 'big data'



One of the task areas of a Resource Management Agency

The algorithm doesn't care!

Central tasks in Digital Humanities are *classification* and *comparison*:

- in both linguistics and literary studies:
  - between versions and parts of texts
    - whole texts - 'big data'
    - microanalyses of sentences
  - between characteristics of
    - Writers / speakers
    - Readers / hearers

Some online applications:

MN: Reading preferences

DistGraph: similarities between Kru consonant systems

WordAlign: edit differences between texts

Case Study 1: Types of readers

Based on data provided by Maya Nikolova

Project: comparison of reading preferences

Hypothesis:

There are (no) gender-specific differences in reading preferences

Method:

Interview and Focus Group analysis

Tabulation of functional and metadata property values

Results:

Forthcoming

Here:

Visualisation of relations induced from the data

(*by permission of Maya Nikolova*)

Gender-specific differences between readers (functional properties)

Information value

Entertainment value

Character: likeable

Character: interesting

Character: plausible

Character: role model

Character: conversation partner

Humanism

Misanthropy

Fast

Slow

Isolation, escapism

Intruding/present author/narrator

Realism

Fantasy

Active interest in arts

Passive interest in arts

Identification

Non-identification

Reader-writer relationship

Book choice: rational

Book choice: semi-rational

Book choice: non-rational

Simplicity

Complexity

Intellectual

Formulaic

Style

Plot

Irony

Topic

Techniques

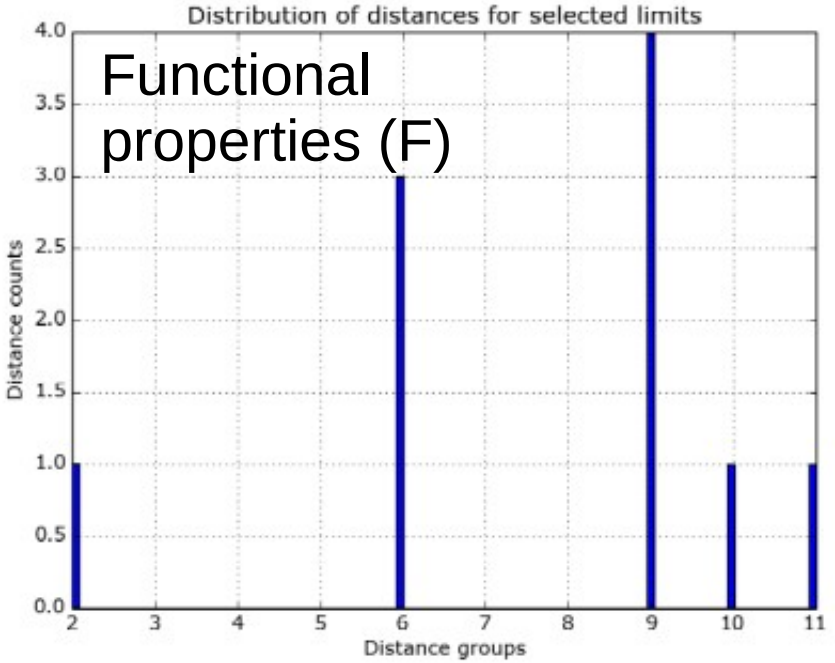
Paper book

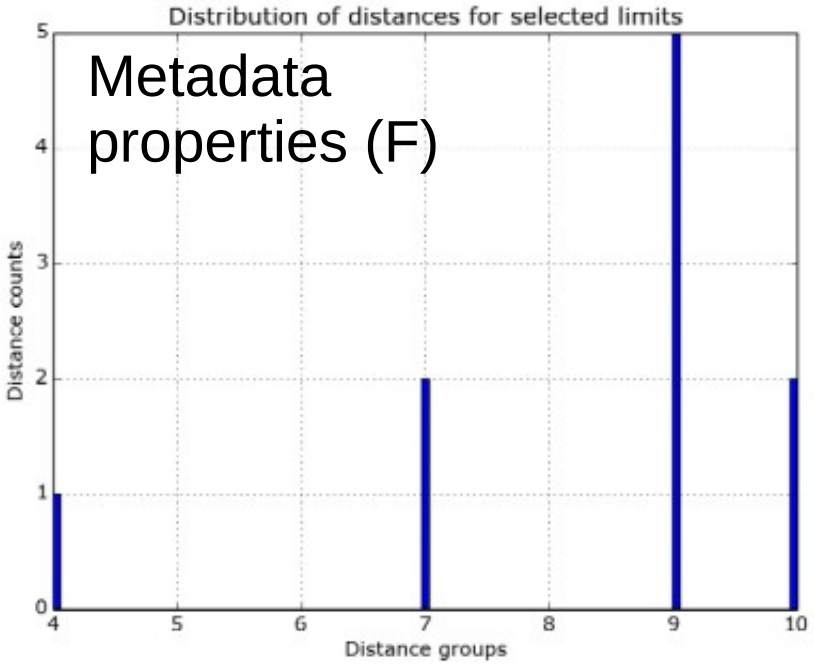
E-book

Differences between readers

	ABf	AKf	BOm	CCm	DPf	ICf	JSm	JVm	MLf	SSM
ABf	0	6	11	11	10	6	7	9	6	8
AKf		0	9	10	11	9	9	9	9	12
BOm			0	5	12	10	12	7	12	11
CCm				0	11	11	11	9	11	11
DPf					0	9	10	11	9	8
ICf						0	7	7	2	5
JSm							0	9	8	8
JVm								0	9	7
MLf									0	6
SSm										0

Differences between readers (F)



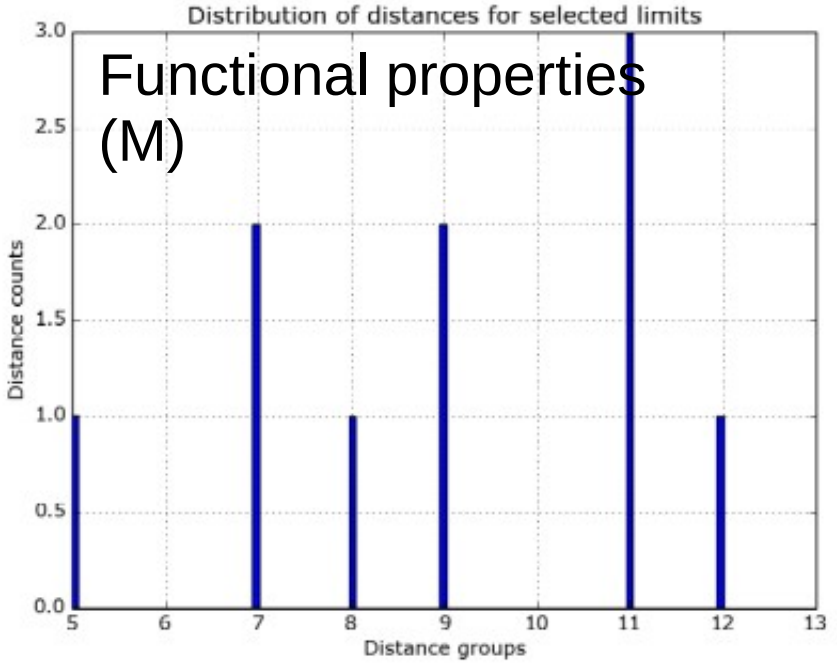


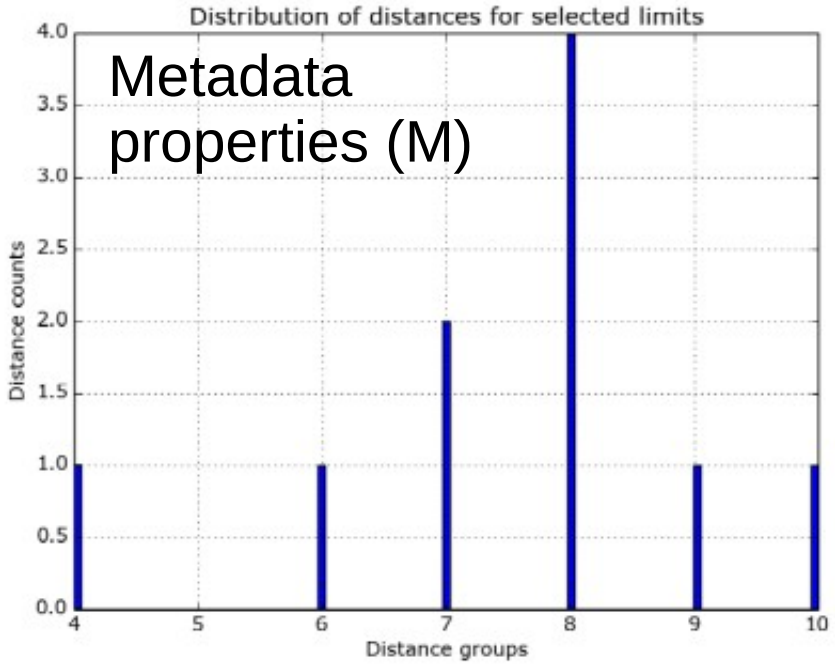
1. Functional properties (top left) represent classifications of reading preferences and habits.
2. Metadata properties (bottom left) represent classifications according to gender, education, etc.

The histograms represent sizes of groups of subjects (Y) as a function of the average number of differences observed for each subject (X).

In general, the graphs show that reading habits are quite diverse, with two exceptions (on the left).

Differences between readers (M)



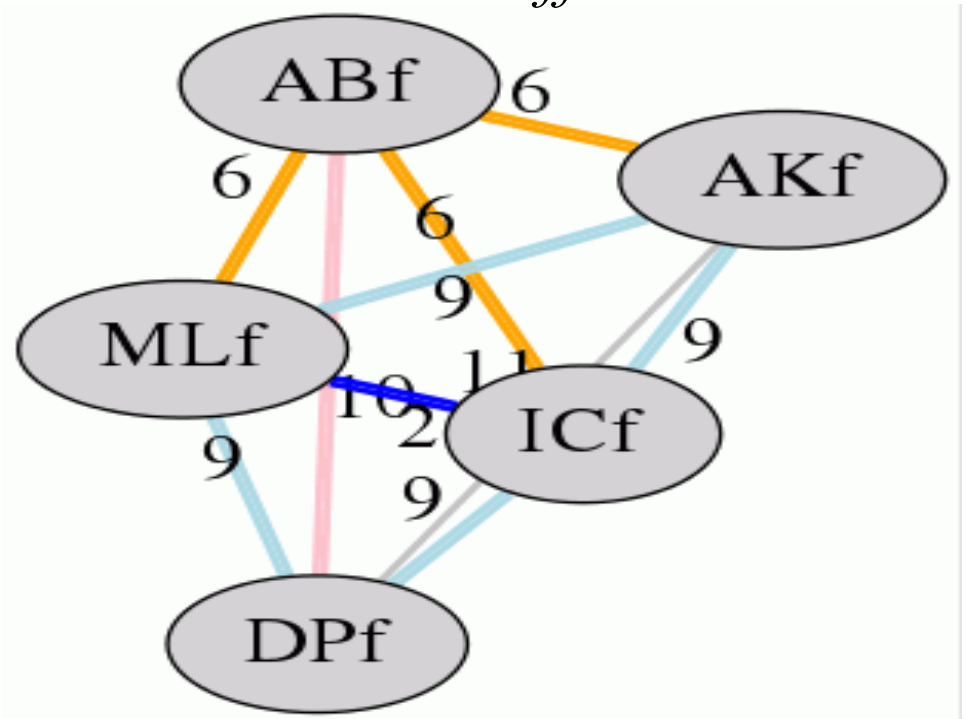


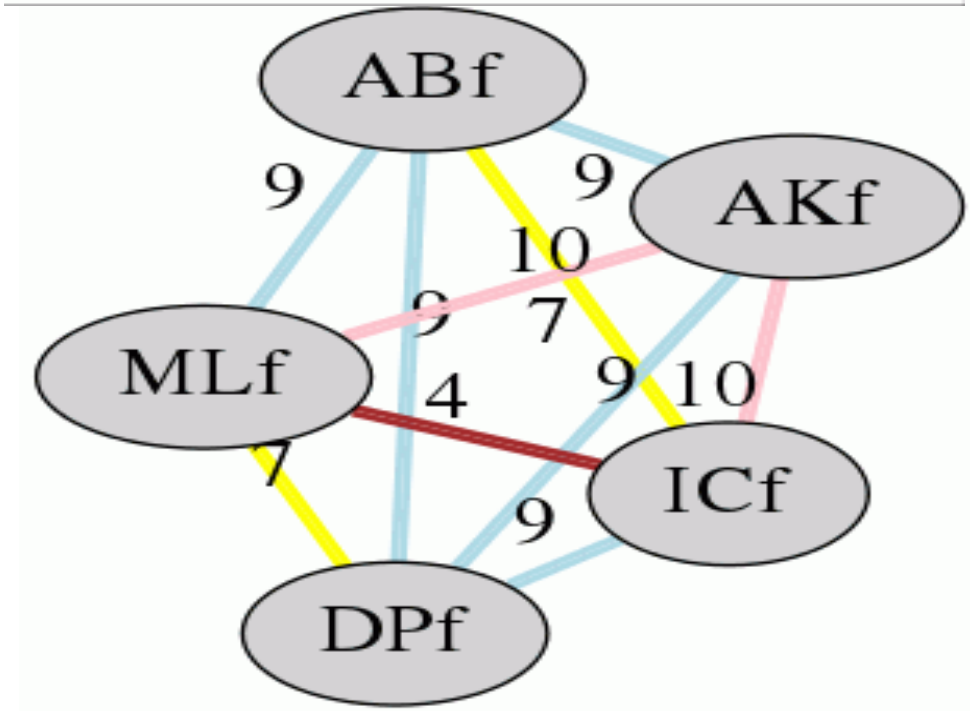
1. Functional properties (top left) represent classifications of reading preferences and habits.
2. Metadata properties (bottom left) represent classifications according to gender, education, etc.

The histograms represent sizes of groups of subjects (Y) as a function of the average number of differences observed for each subject (X).

Differences among males were less heterogeneous, reflecting variation in reading habits.

Differences between readers (F)



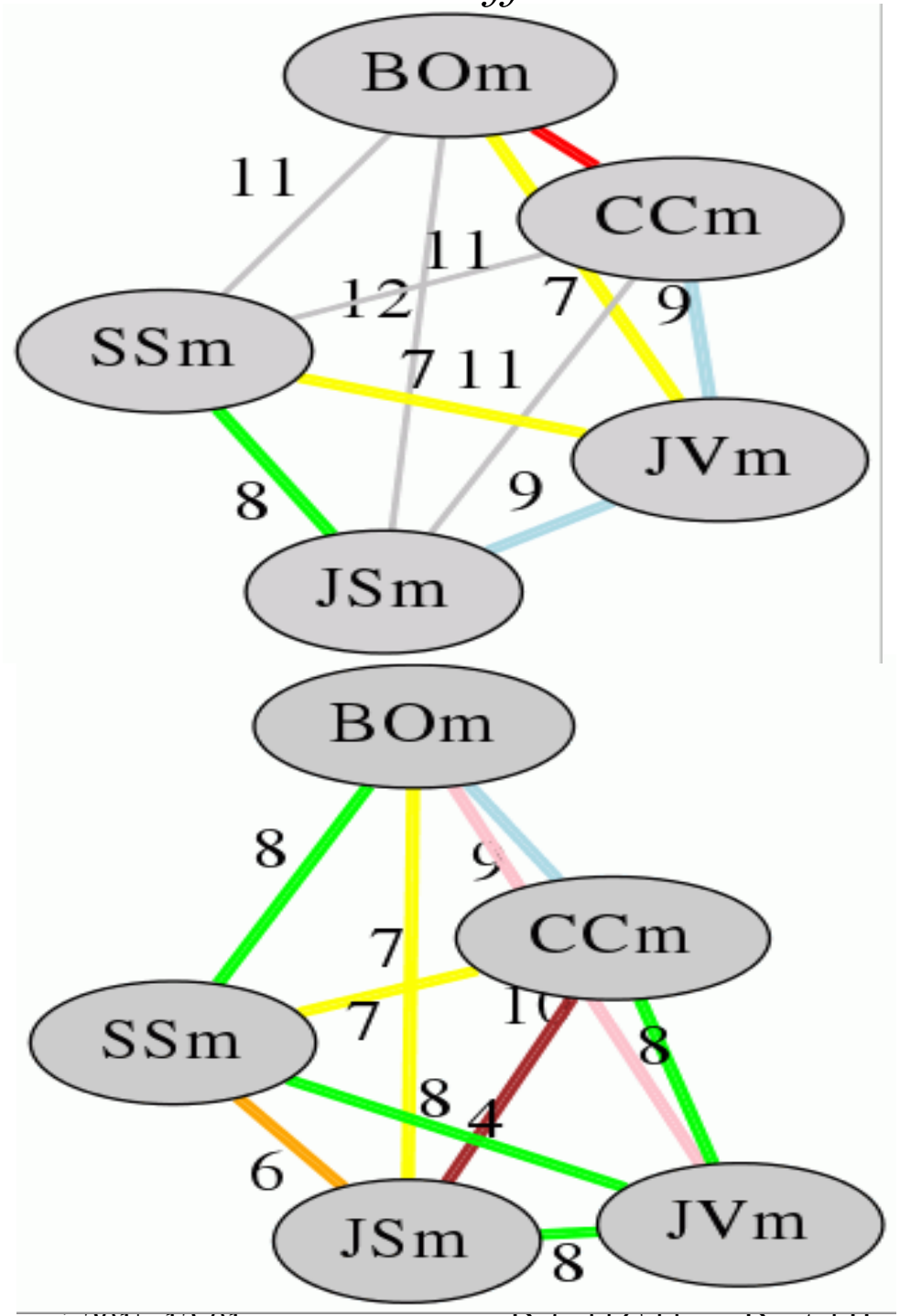


1. Functional properties (top left) represent classifications of reading preferences and habits.
2. Metadata properties (bottom left) represent classifications according to gender, education, etc.

The graphs represent differences between subjects as distances (not to scale), based on the Levenshtein Edit Distance function of the properties.

These graphs represent female subjects.

Differences between readers (M)

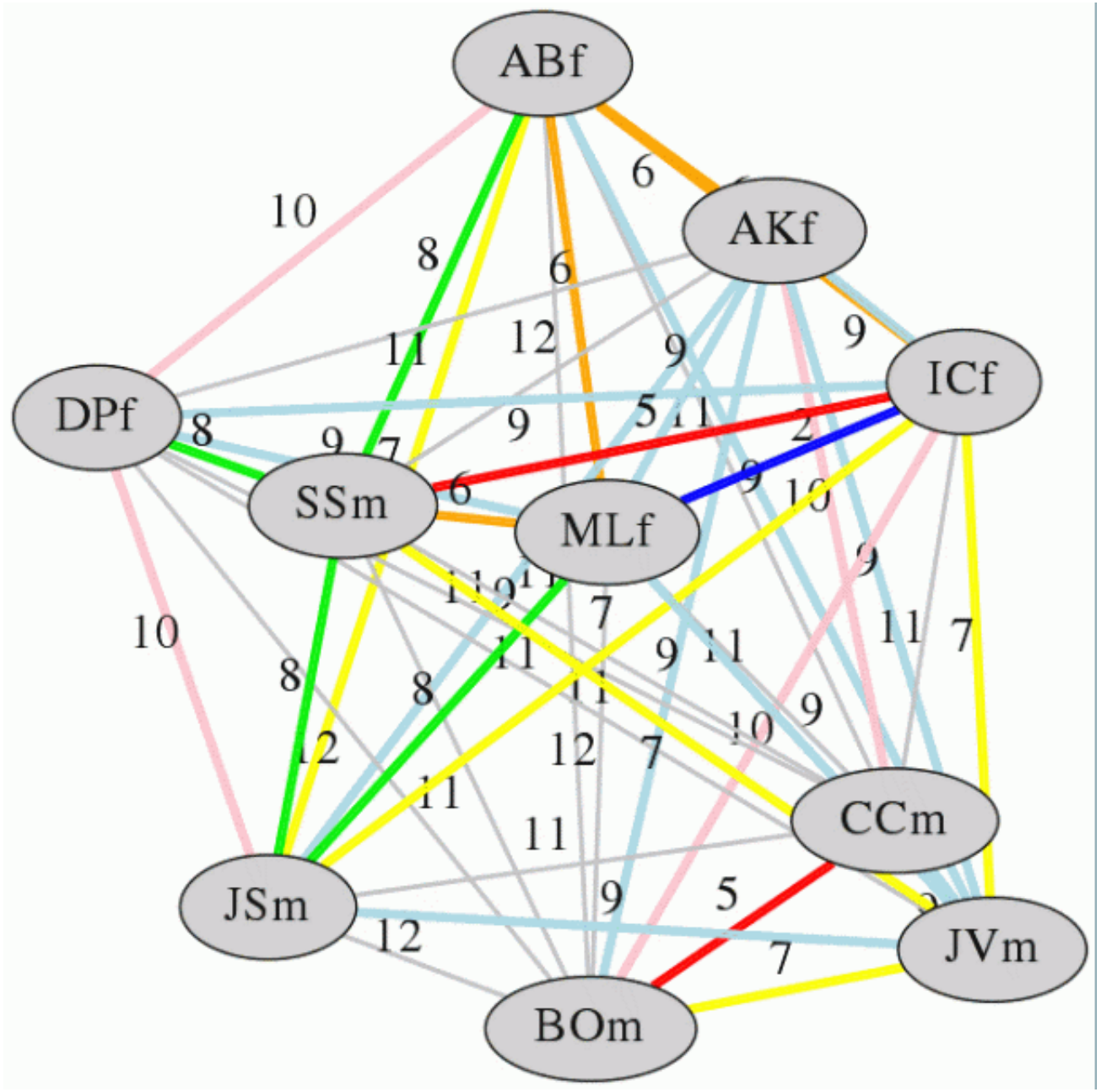


1. Functional properties (top left) represent classifications of reading preferences and habits.
2. Metadata properties (bottom left) represent classifications according to gender, education, etc.

The graphs represent differences between subjects as distances (not to scale), based on the Levenshtein Edit Distance function of the properties.

These graphs represent male subjects.

Differences between readers (F and M combined)



All subjects, male and female.

Note that (with one exception) the males group together in their reading habits, and so do the females.

DistGraph (Maya Nikolova Reading Data, functional properties)

Differences between readers (ordered by mean distance to neighbours)

Name	Gender	Mean dist
IC	F	7.333
ML	F	8.000
AB	F	8.222
SS	M	8.444
JV	M	8.556
JS	M	9.000
AK	F	9.333
BO	M	9.889
CC	M	10.000
DP	F	10.111

Of course there is far more to say about qualitative methods of analysing interview data than these quantitative analyses, and the quantitative methods can be taken much further, but the methods illustrated here are a very useful starting point.

The moral of this story is that

Qualitative interviews can be given a solid quantitative foundation in addition to any further qualitative argumentation which may follow.

Standard arrangements of quantitative information (e.g. tables) may be useful.

Graphical visualisations are helpful in either suggesting or underlining lines of investigation.

Case Study 2: Phonological typology of Kru languages (Ivory Coast)

Project: comparison of Kru languages

Hypothesis:

The geographical distances between the Kru languages are reflected in the differences of their consonant systems.

Method:

Data mining with legacy language atlases

Results:

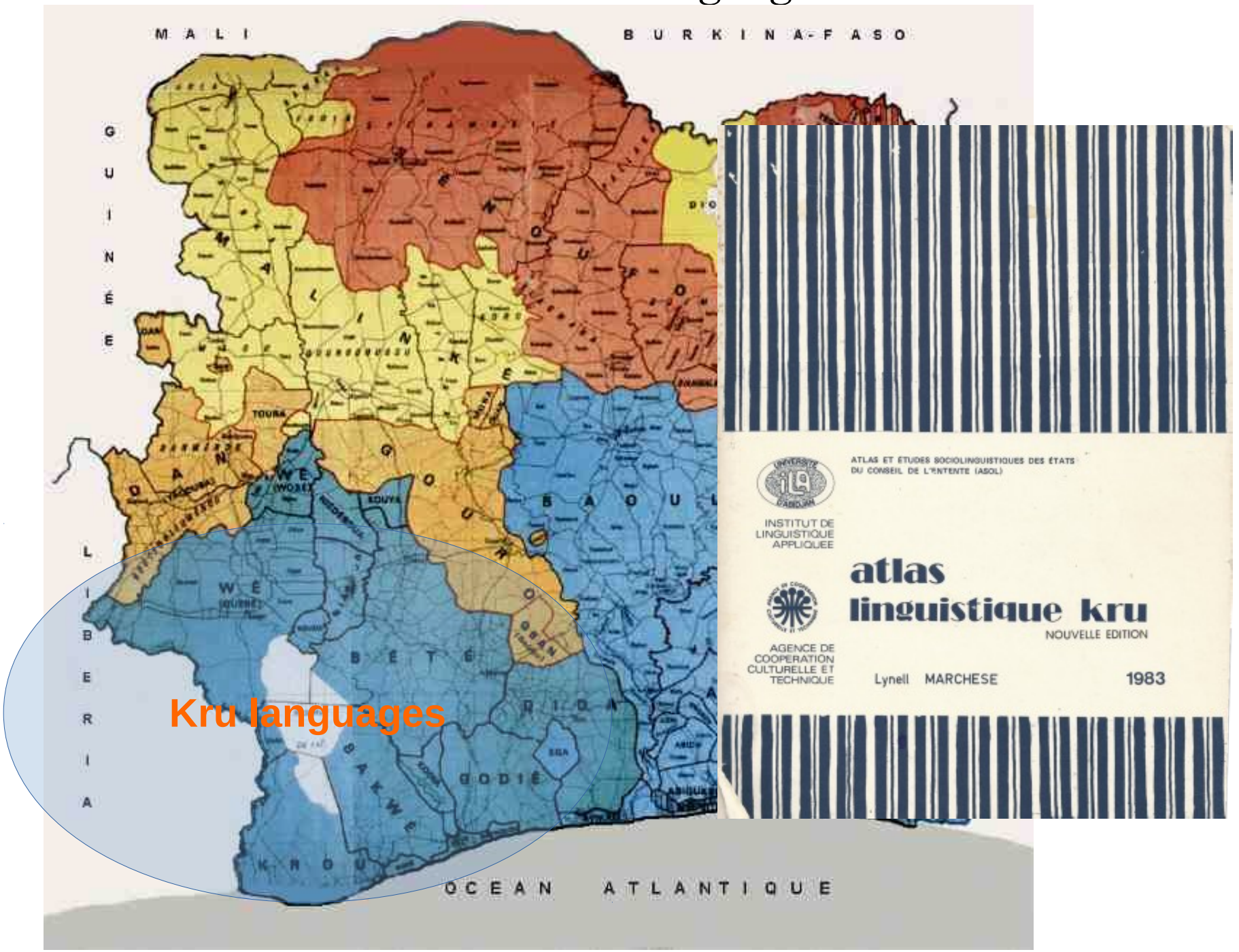
Coming up

Here:

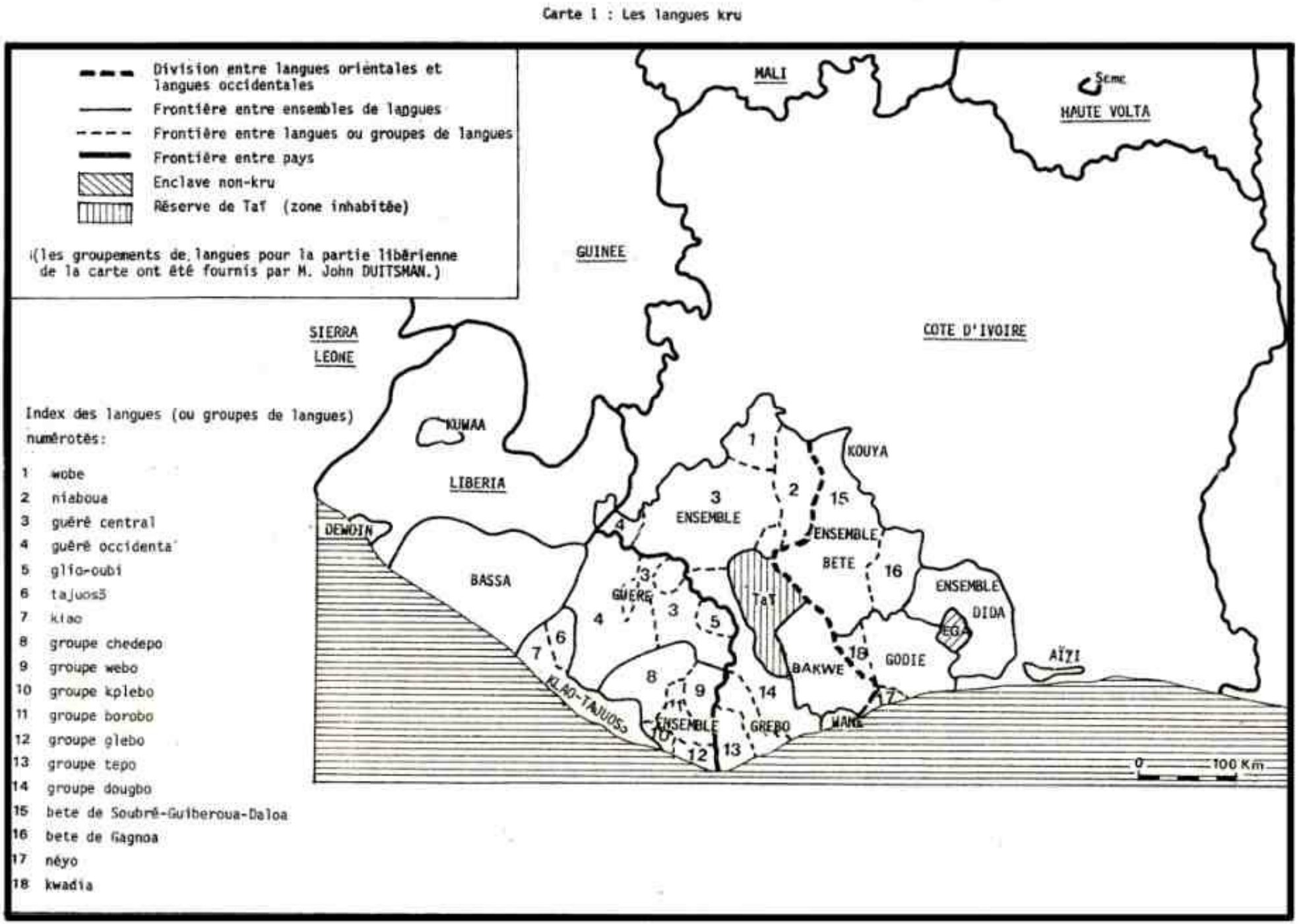
Visualisation of relations induced from the legacy data for 19 Kru languages

(of the 39 in the Ethnologue database of language metadata)

Côte d'Ivoire: Kru languages



Côte d'Ivoire: Kru languages



Côte d'Ivoire: Kru languages – consonants

- SYSTEMES CONSONANTIQUES DE QUELQUES LANGUES ORIENTALES -

bété de Guibéroua (Werle, 1976)

p	t	c	k	kp	C^{w}(1)
b	d	ɟ	g	gb	
f	s				
v	z				
ɓ	l	j	ɣ	w	
m	n	ɲ	ŋ	ŋw	

Godié de Dakpadou et Lagako (Marchese, 1975)

p	t	c	k	kp	kw
b	d	ɟ	g	gb	gw
f	s				
v	z				
ɓ	l	j	ɣ		w
m	n	ɲ	ŋ		ŋw

Koyo (Kokora, 1976, p. 23)

p	t	c	k	kp	C^{w}	C^{j}
b	d	ɟ	g	gb		
f	s					
v	z					
ɓ	l	j	ɣ(2)		w	
m	n	ɲ	ŋ			

Néyo (Grah)

p	t	c	k	kp	C^{w}
b	d	ɟ	g	gb	
f	s				
v	z				
ɓ	l	j	ɣ		w
m	n	ɲ	ŋ		

Dida de Lozoua (Gratrix)

p	t	c	k	kp	kw
b	d	ɟ	g	gb	gw
f	s				
v	z				
ɓ	l	j	ɣ	w	
m	n	ɲ	ŋ	ŋw	

dida-f (vata) (Siméon, Dugas, Kaye, Koopman, 1981)

p	t	c	k	kp	kw	
b	d	ɟ	g	gb	gw	
f	s					
v	z					
m	n	ɲ	ŋ	$ŋ^{m}$(3)		
ɓ	l	j	ɣ	w		

(1) Voir section 2.1.4.4.

(2) En Koyo, [ɣ] n'apparaît que dans quelques lexèmes dont la plupart sont des emprunts.

(3) "Le ŋm résulte d'une assimilation nasale devant les consonnes labio-vélaires" (Siméon, et. al. 1980:107).

Côte d'Ivoire: Kru languages – consonants

- SYSTEMES CONSONANTIQUES DE QUELQUES LANGUES KRU OCCIDENTALES -

Wobé[1] (Link, 1975 p. 206)	Guéré[2] (Fisher, 1976 p. 96)	Krahn (Duitsman)	cedepo (Laesch, c.p.)[4]	Klao (Duitsman, et.al, 1975, p. 92)
p t c k kp kw b d ɟ gb f s w m n ɲ ŋm km ŋw	p t c k kp kw b d ɟ g gb gw f s v z ɓ l j w ɗ m n ɲ ŋm km ŋw	p t c k kw b d ɟ gb f s l w m n ɲ	p t c k kp kw b d ɟ gb f s h l m n ɲ ŋm	p t c k kp kw b d ɟ gb f s l j w m n ɲ ŋm
Niaboua (Bentinck, 1975 p.8)	**Dewoin** (Welmers)	**Bassa**[3] (Bertkau, et al.	**Grebo** (Innes, p. 14	**Tépo** (Dawson, MS)
p t c k kp kw b d ɟ g gb gw f s v z ɓ l j w m n ɲ	p t k kp kw b d ɟ g gb gw f s v z ɓ l j w m n n ŋ	p t c k kp b d ɟ dj g gb f s v z ɓ l w m n ɲ gw hw n hw	p t c k kp b d ɟ g gb f s l j w m n ɲ ŋ ŋm nw hm hn hw h hl	p t c k kw b ɗ ɟ g gb f s h l j w m n ɲ ŋ ŋm

(1) *La série de nasales en wobé, guéré, tépo et bassa n'est pas phonémique (voir section 2.2.1.2.)*

(2) *D'après Fisher, il y a une opposition entre* l *et* ɗ *en guéré :* ɟɗí *singes / ɟlì' gallons*

(3) *Il semble y avoir une opposition entre* ɟ *et* dj *en bassa. Cette opposition n'a pas été relevée dans d'autres langues kru.*

(4) *Nous ne savons pas pourquoi les semi-voyelles* y *et* w *ne figurent pas sur les tableaux de cedepo et de bassa.*

- 109 -

Côte d'Ivoire: Kru languages – consonants

- 111 -

- SYSTEMES CONSONANTIQUES DES LANGUES KRU ISOLEES -

(LIBERIA)	(HAUTE-VOLTA)	(COTE-D'IVOIRE)
kuwaa (Thompson, p. 12)	Sɛmɛ (Prost, p. 346)	Aĩzi (Hérault p. 10)
p t k kp kw	p t c k kp	p t c k kp
b d ɟ	b d ɟ g gb	b d ɟ g gb
f s	f s (ă)	f s š
	v	v z ž
l j ɣ w	l l j w	l j w
m n ɲ ŋ	m n ɲ gm	m n ɲ ŋ
mb nd nj ŋg ŋmgb	(h semble être un allophone de f)	

Côte d'Ivoire: Kru languages – consonants

A practical systematisation procedure for a machine-readable database:
Step 1: A word processor or spreadsheet or DBMS table.
Step 2: Export as CSV (character/comma/tab separated value) table.
Step 3: Process manually or automatically: analyse and format as desired.

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

Côte d'Ivoire: Kru languages – consonants

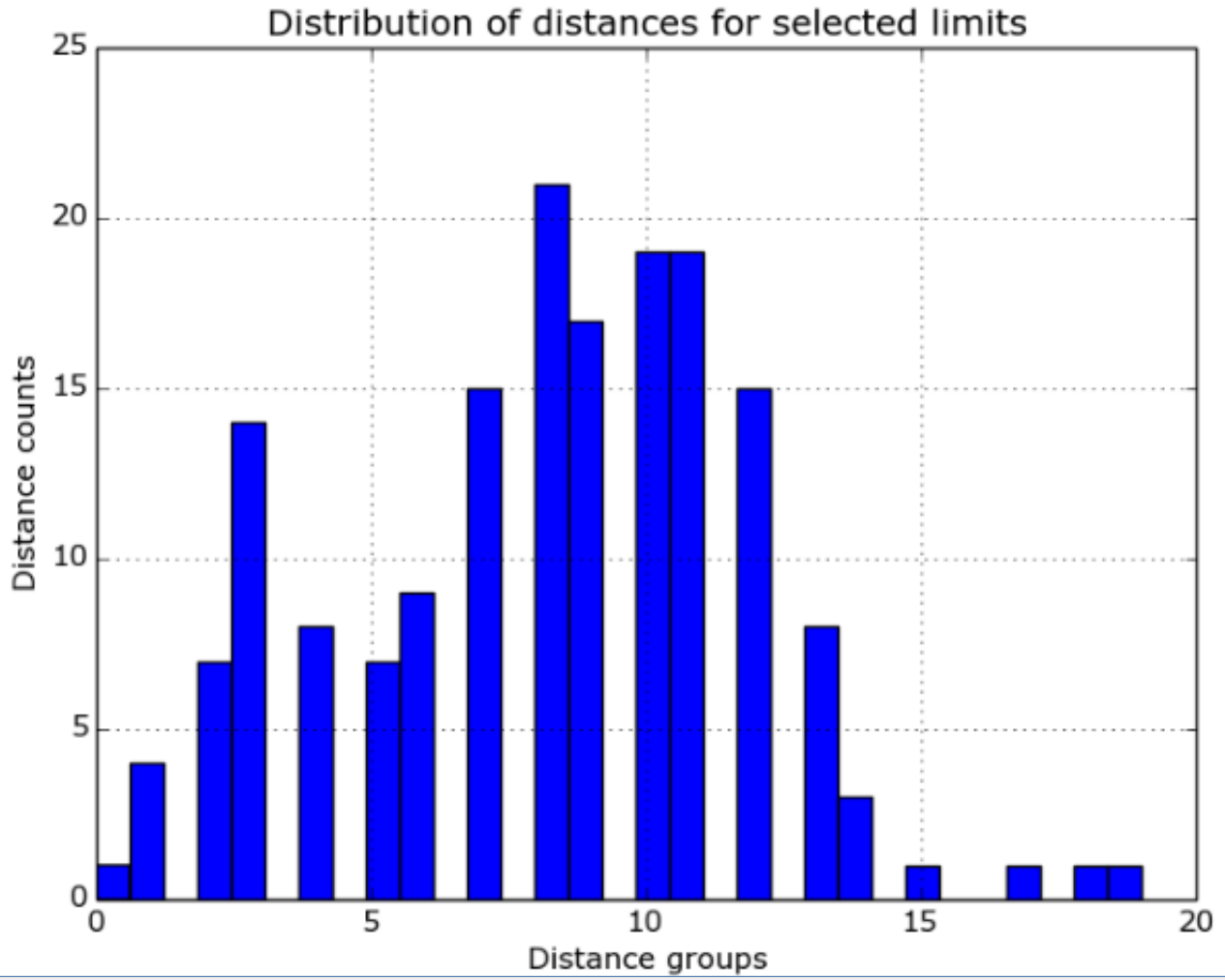
A practical systematisation procedure for a machine-readable database:
Step 1: A word processor or spreadsheet or DBMS <u>table</u>.
Step 2: Export as <u>CSV</u> (character/comma/tab separated value) table.
Step 3: Process <u>manually</u> or <u>automatically</u>: analyse and format as desired.

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 … 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 _ … J N … Nm _ _ _ _ _ _ _ _ _ _ _
Wobe	p t c k kp kw _ b d C _ _ gb _ f s _ … _ _ _ _ _ _
Guere	p t c k kp kw _ b d C _ g gb gw f s …
Krahn	p t c k _ kw _ b d C _ _ gb _ f s …
Cedepo	p t c k kp kw _ b d C _ _ gb _ f s …
Klao	p t c k kp kw _ b d C _ _ gb _ f …
Niaboua	p t c k kp kw _ b d C _ g gb gw f …
Dewoin	p t _ k kp kw _ b d C _ g gb gw f …
Bassa	p t c k kp _ _ b d C dj g gb _ … _
Grebo	p t c k kp _ _ b d C _ g gb _ … _
Tepo	p t c k _ kw _ b d C _ g gb _ f s _ _ _ _ h _ _ _ … _
KuwaaLiberia	p t _ k kp kw _ b d C _ _ _ _ f s _ _ _ _ _ _ _ _ _ l j x w m n J N _ … g 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 _ _ _ _ _ _ _ _ _ _ _ _ _

Then: 171 language comparisons to do:
$(n^2 - n) / 2 = (19^2 - 19) / 2 = 171$
for 44 features each time: 7524.

That's a helluva lot.
So in comes Digital Humanities

Côte d'Ivoire: Kru languages – consonants



Spread of differences between 19 Kru consonant inventories for 44 features, which we want to visualise.

Useful strategy: interpret and map differences as distances in quality space.

Côte d'Ivoire: Kru languages – consonants

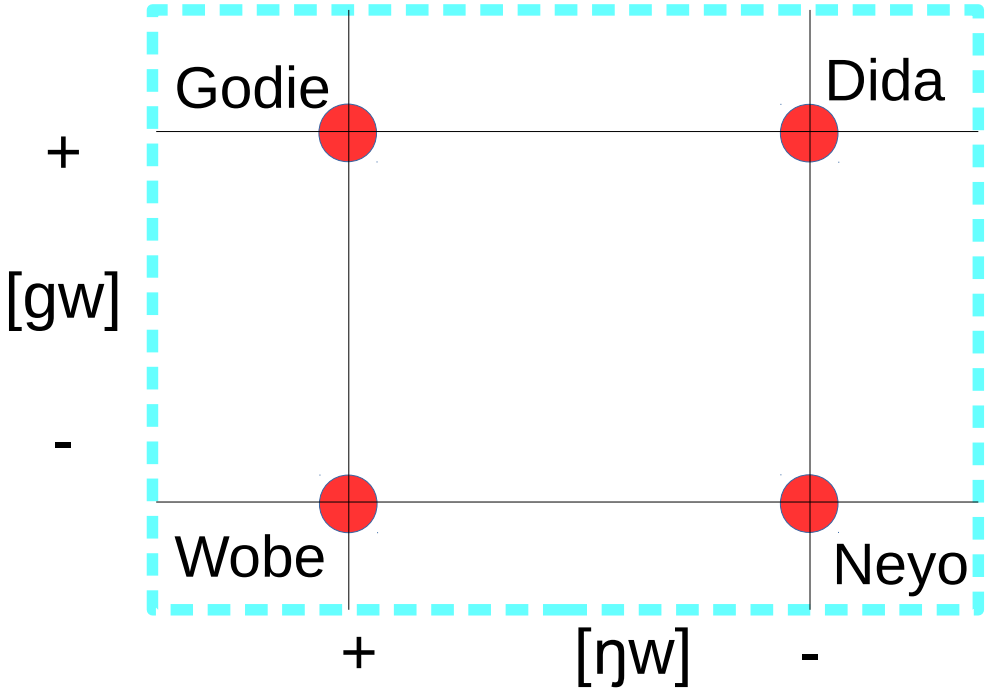
Bete	0	1	2	1	1	3	10	6	9	11	8	4	4	7	11	8	12	9	6
Godie		0	3	2	0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
Koyo			0	1	3	3	12	8	9	11	8	4	4	9	13	8	12	9	6
Neyo				0	2	2	11	7	8	10	7	3	3	8	12	7	11	8	5
DidaDeLozoua					0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
DidaF						0	11	5	10	10	7	3	3	10	12	7	13	10	7
Wobe							0	8	6	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

Côte d'Ivoire: Kru languages – consonants

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

Côte d'Ivoire: Kru languages – consonants

- But: the phoneme data matrix is deceptively 2-dimensional: 19 languages x 44 consonants:
- The 19 objects are actually located in a 44 dimensional quality space. Here are 2 of these dimensions, applied to the 4 languages Godie, Dida, Wobe and Neyo:



- Even distinctive features involve around 12 dimensions.
- How to visualise all 44 dimensions in 2 dimensions?

Strategy #1:

Squash to 2 dimensions!

- Differences are interpreted as distances
- Distances are represented spatially as a distance map
- The dimensions are squashed – like a system of springs – into 2 dimensions
- Further dimensions may be represented by colours, etc.

• Strategy #2:

Select elite features!

- Check the features for their importance in distinguishing objects
- Randomly start with an important feature and build a hierarchy of features distinguishing between sets of objects until all are distinguished
- Different choices lead to different results, different insights

Strategy #1:

Squash to 2 dimensions!

- Differences are interpreted as distances
- Distances are represented spatially as a distance map
- The dimensions are squashed – like a system of springs – into 2 dimensions
- Further dimensions may be represented by colours, etc.

And visualise the results!

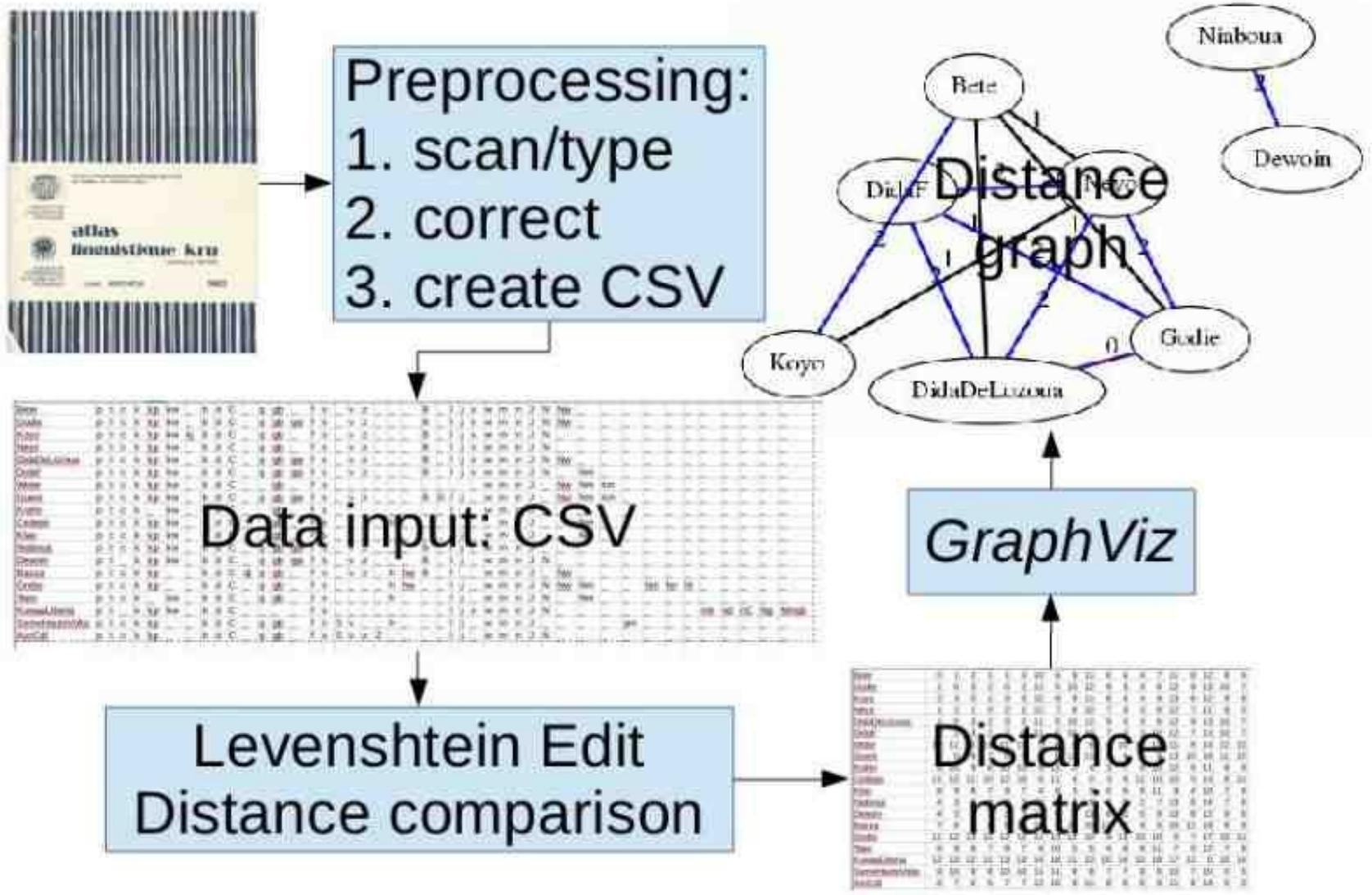
Strategy #2:

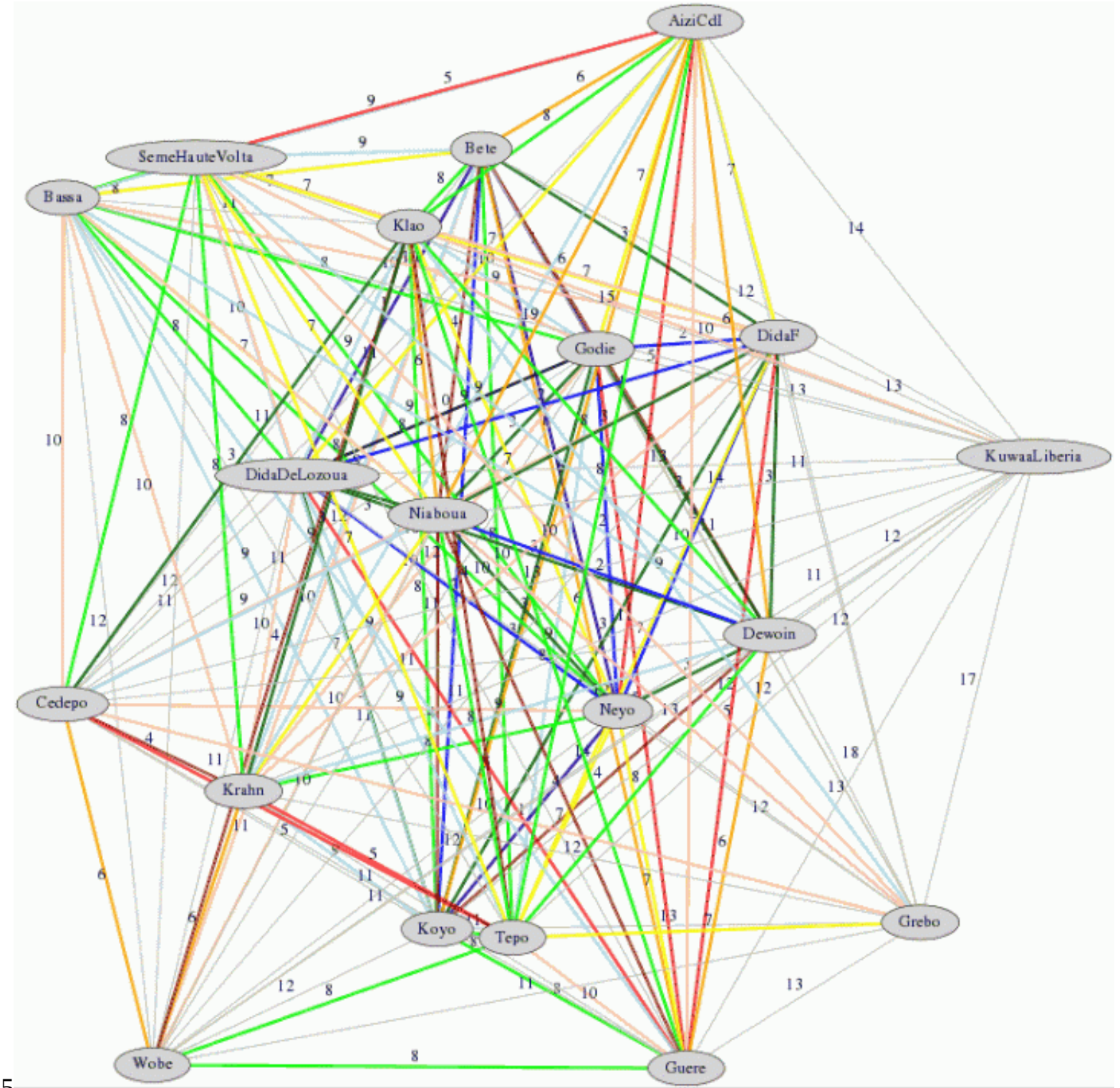
Select elite features!

- Check the features for their importance in distinguishing objects Randomly start with an important feature (SD of feature values), build a hierarchy of features distinguishing between sets of objects until all are distinguished
- Different choices lead to different results, different insights

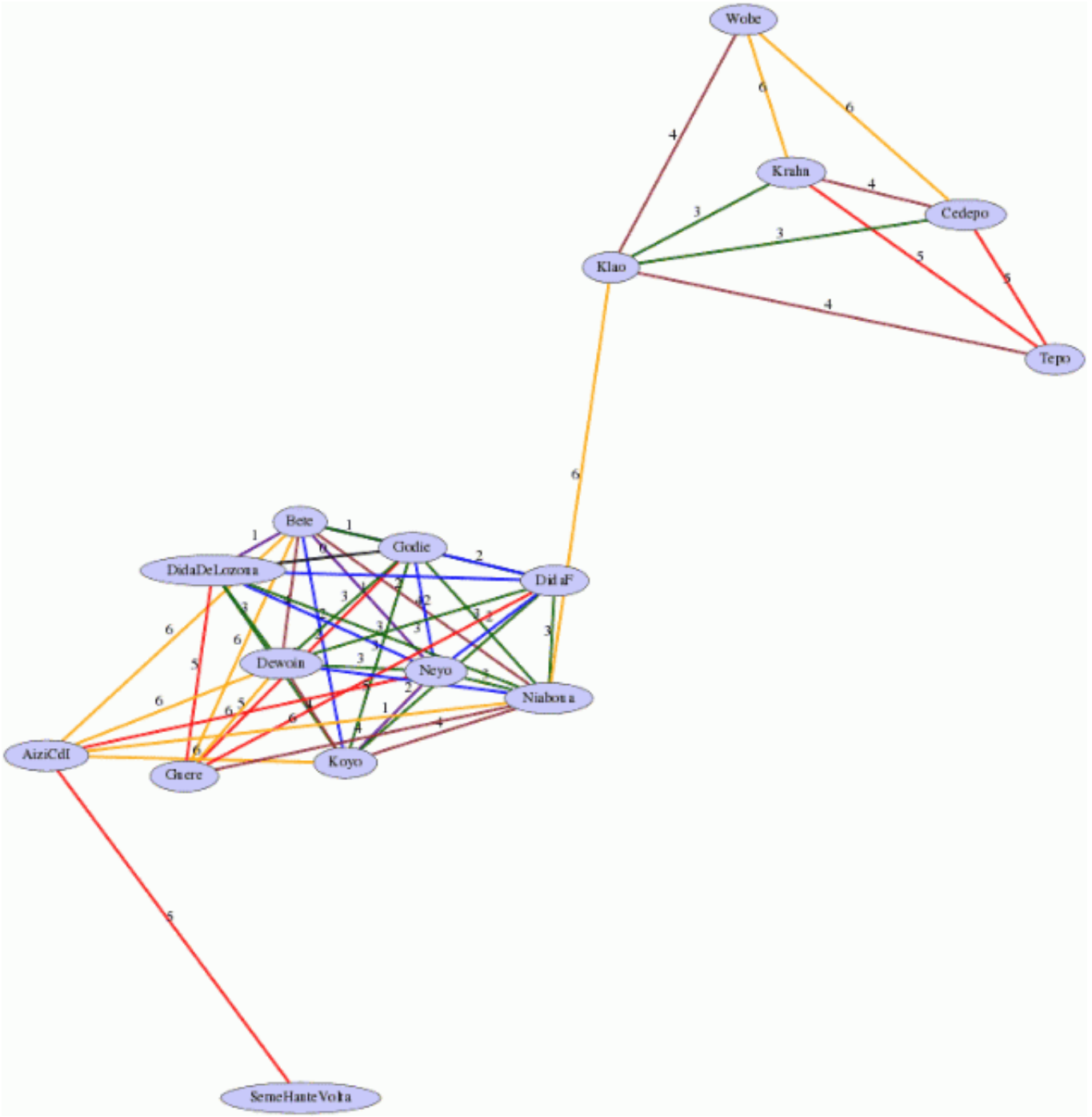
Côte d'Ivoire: Kru languages – consonants

Visualisation workflow:

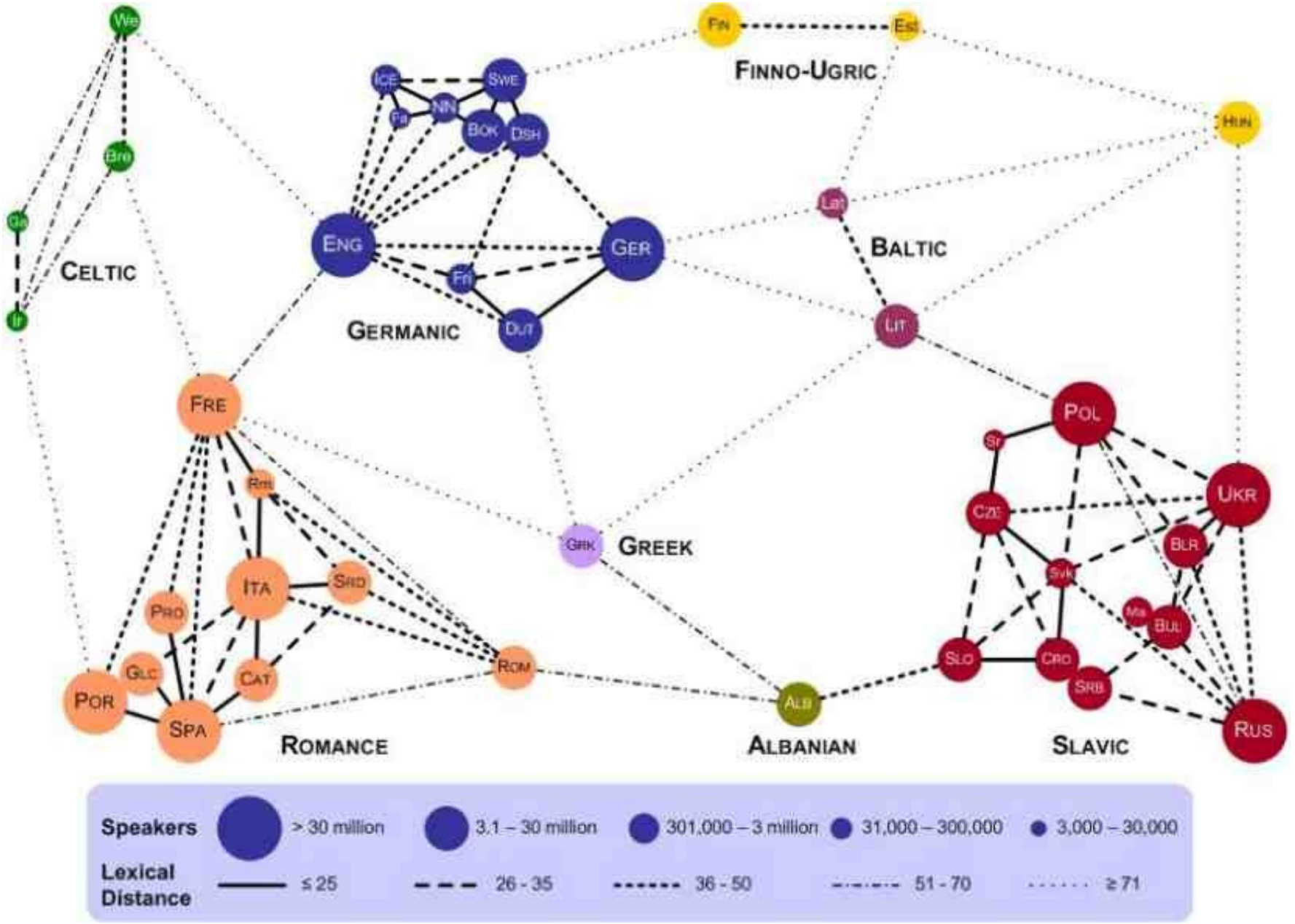



AiziCdl
SemeHauteVolta
Bete
Bassa
Klao
DidaF
Godie
KuwaaLiberia
DidaDeLozoua
Niaboua
Dewoin
Cedepo
Neyo
Krahn
Grebo
Koyo
Tepo
Wobe
Guere

Strategy 1: Squash those dimensions!



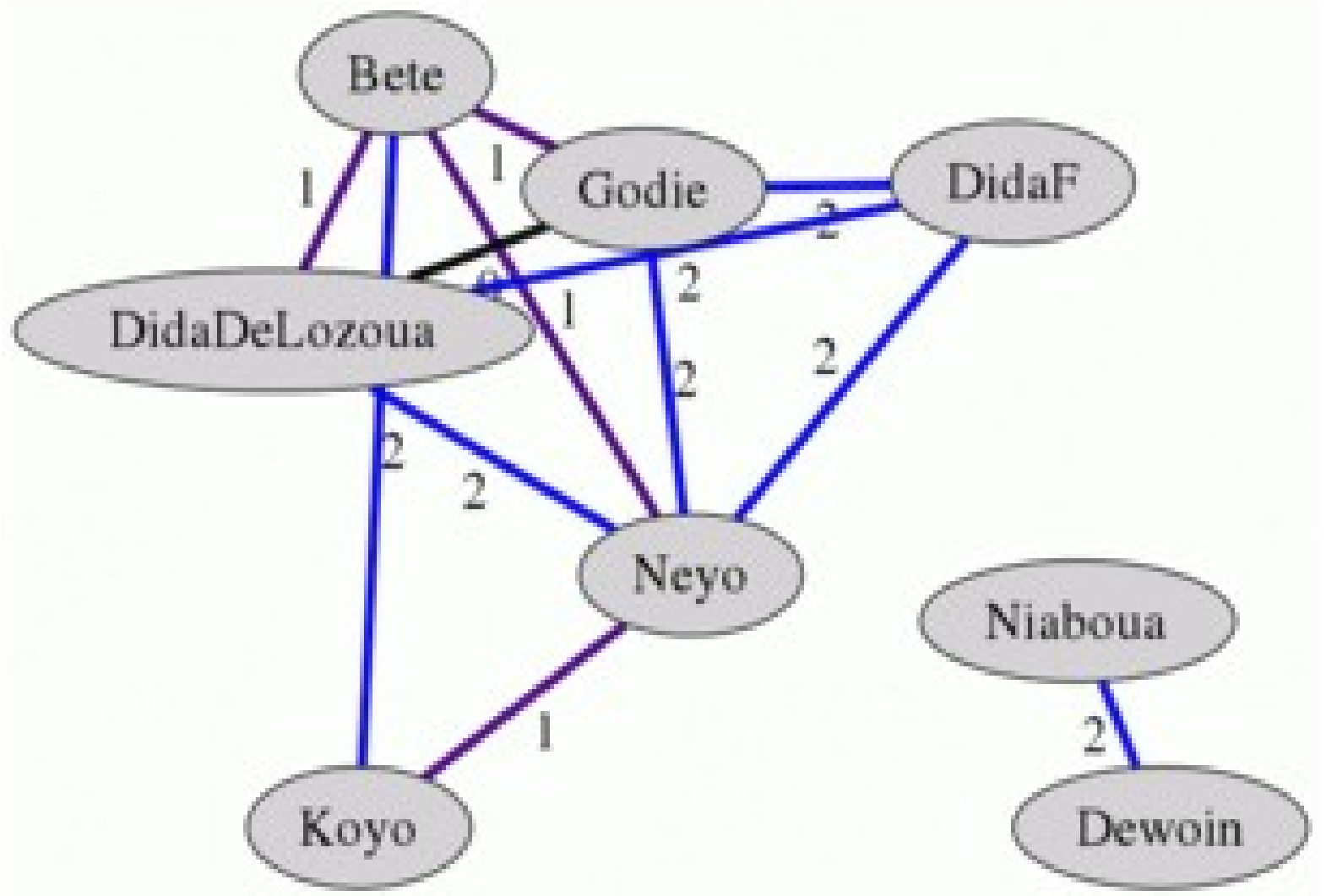
Strategy 1: Squash those dimensions!



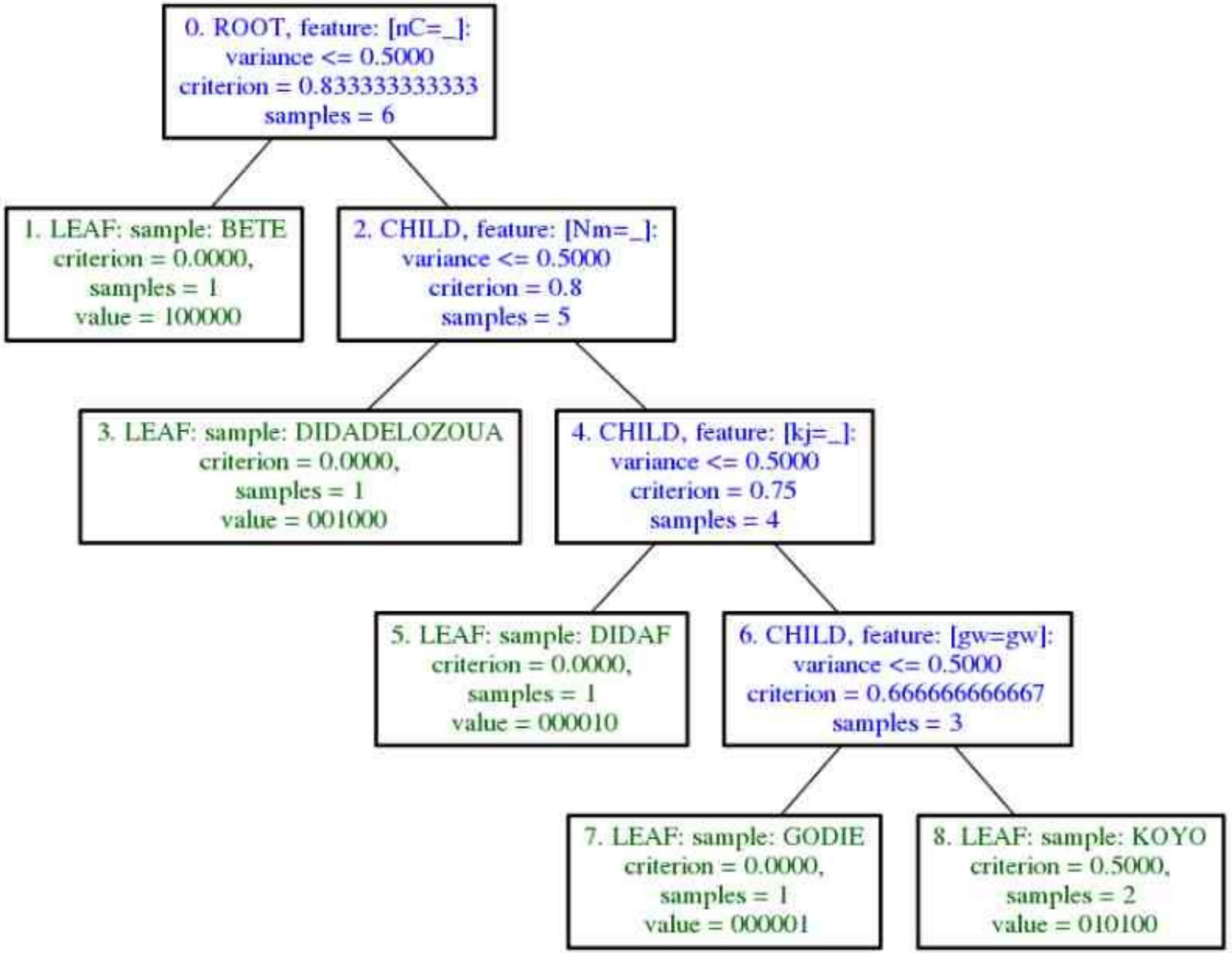
Speakers: > 30 million; 3.1 – 30 million; 301,000 – 3 million; 31,000 – 300,000; 3,000 – 30,000

Lexical Distance: ≤ 25; 26 - 35; 36 - 50; 51 - 70; ≥ 71

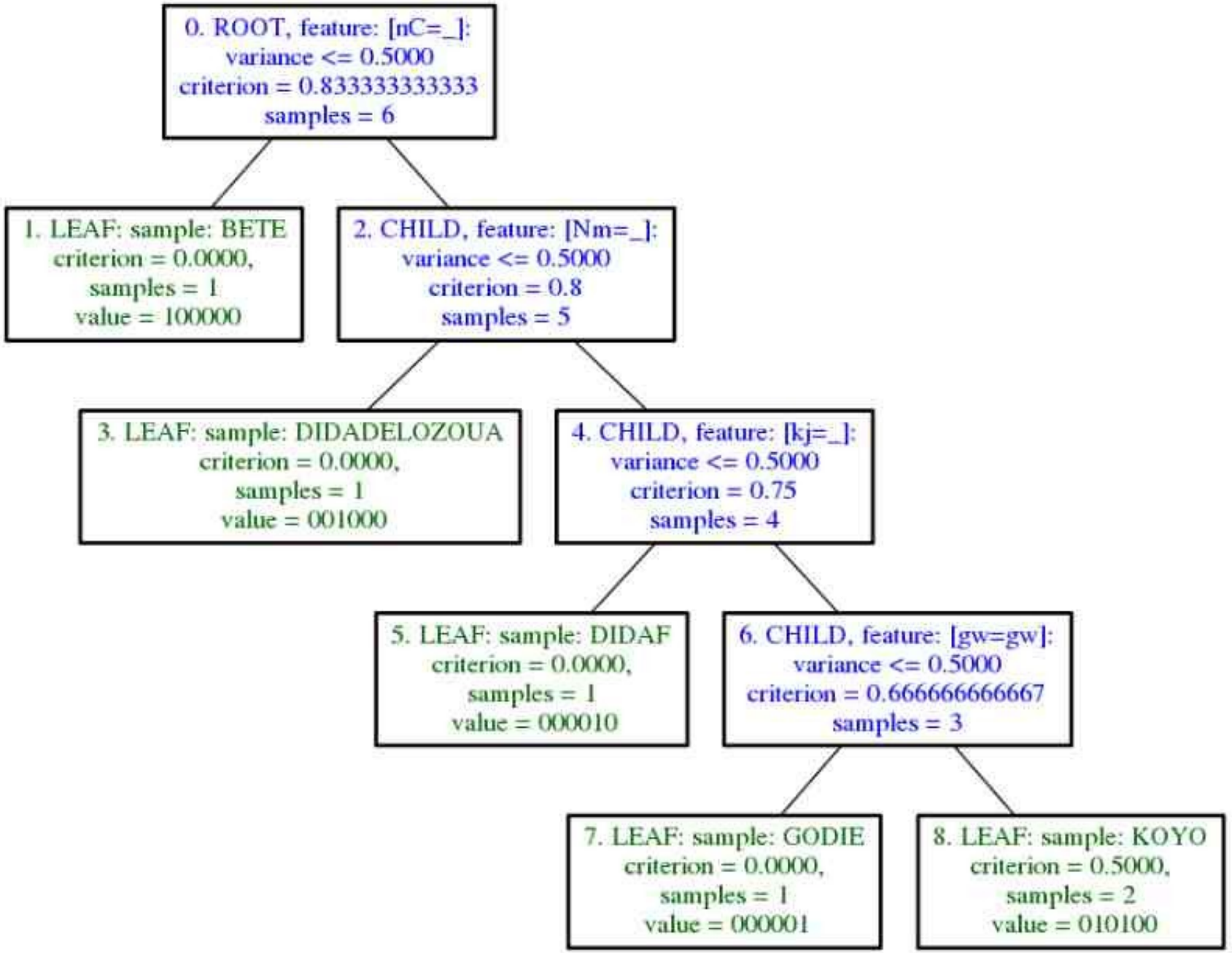
Strategy 1: Squash those dimensions!



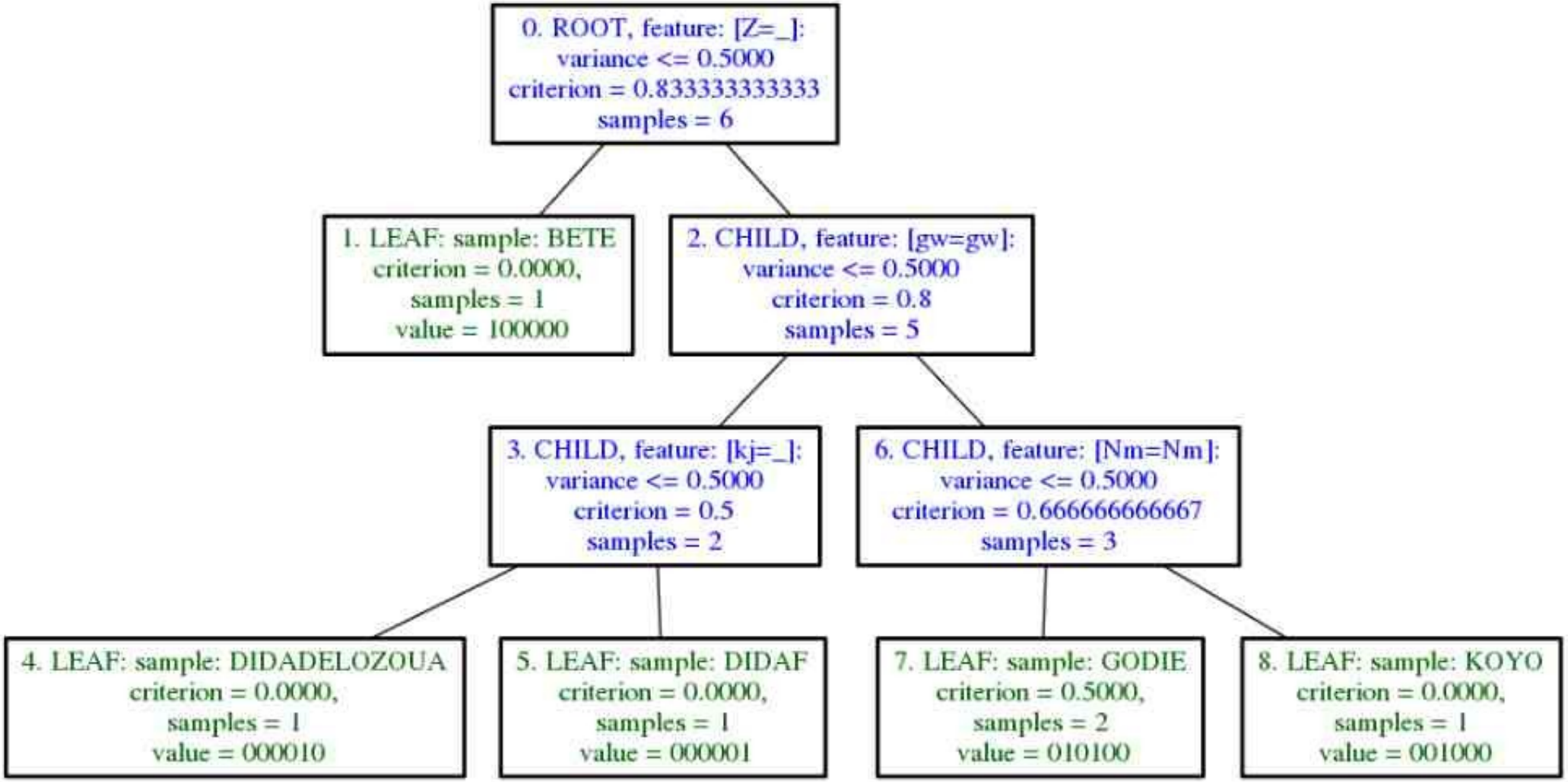
Strategy 2: Pick out the best features!



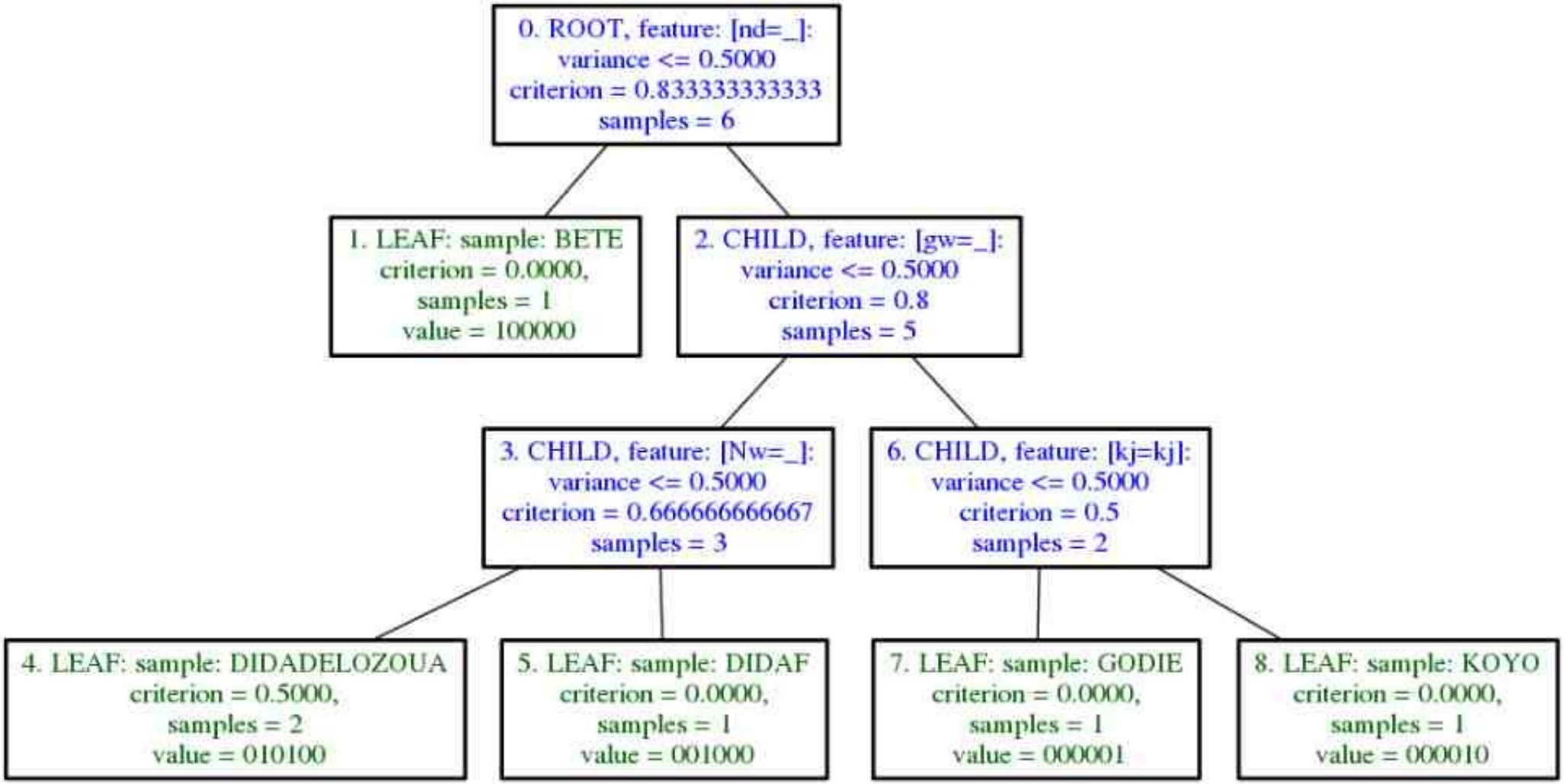
Strategy 2: Pick out the best features!



Strategy 2: Pick out the best features!



Strategy 2: Pick out the best features!



The moral of this story is that

Legacy data in linguistic atlases can be given a new lease of life and a solid quantitative foundation in addition to any further research on dialect relations and history which may be pursued.

Standard arrangements of quantitative information (e.g. tables) may be useful.

Graphical visualisations are helpful in either suggesting or underlining lines of investigation.

Case Study 3: Evaluation of spoken discourse transcribers

Based on data provided by Jolanta Bachan

Project: Evaluation of discourse transcribers

Hypothesis:

The best transcribers can be selected on the basis of comparing differences between transcriptions by different transcribers.

Method:

1. Normalisation of transcriptions to equal length
2. Distance analysis of normalised transcriptions

Results:

Coming up

Here:

Visualisation of relations induced from transcriptions

Project: Evaluation of discourse transcribers

Hypothesis:

The best transcribers can be selected on the basis of comparing differences between transcriptions by different transcribers.

Method:

1. Normalisation of transcriptions to equal length
2. Distance analy malised transcriptions

Results:

Coming up

Here:

Visualisation of relations induced fr

Note that this procedure is related to other fields such as:
1) Spell-checking
2) Text correction in language teaching
3) Basic grammar comparison in stylometry

Original transcriptions (extracted from speech annotations):

Ali: jest to podzial taki nasz umowny &, mozna powiedziec &,

Mal: jest to podzial taki nasz & umowny &, mozna powiedziec &,

Mic: jest to podzial taki nasz umowny &, mozna powiedziec &,

Pio: jest to poglad taki nasz umowny & mozna powiedziec &.

Mat: jest to podzial taki nasz & umowny mozna powiedziec &,

Ola: jest to podzial & taki nasz umowny & mozna powiedziec &,

Method:

Extraction of transcriptions from annotations

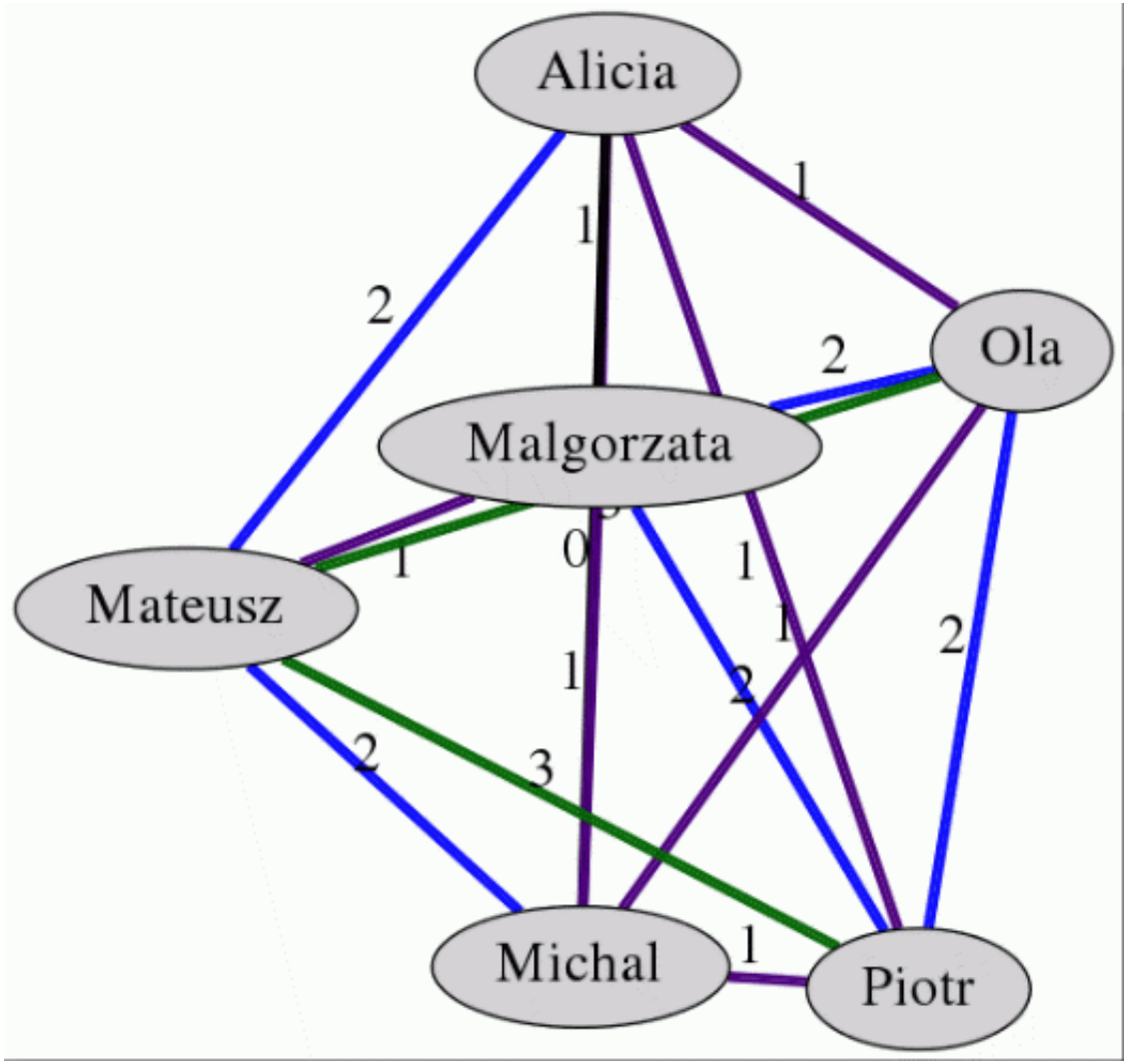
Definition of substitutions as token equivalences

e.g. ‘&’, &,’, ‘&.’

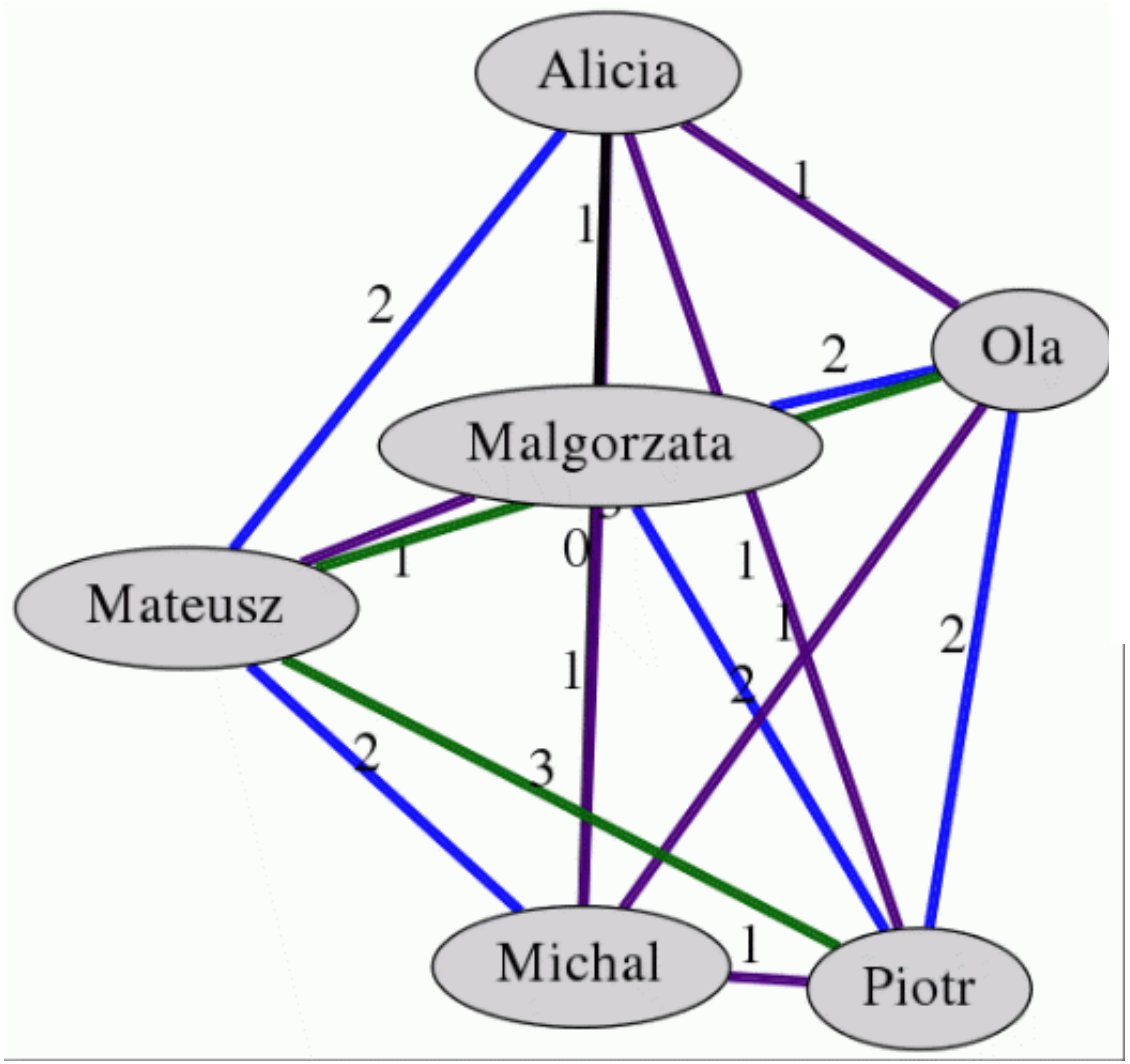
Processing of insertions, deletions

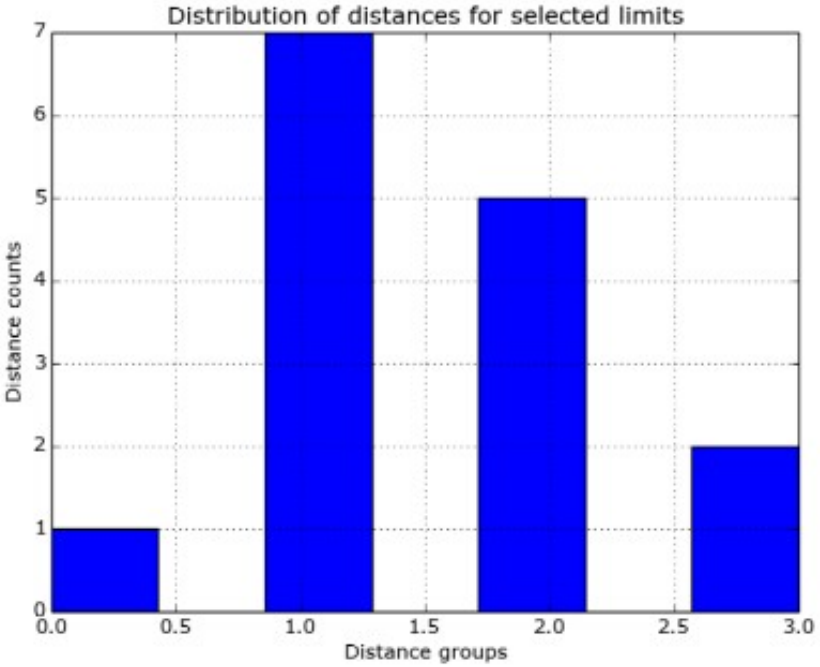
jest	to	podzial	_	taki	nasz	_	umowny	&,	mozna	powiedziec	&,
jest	to	podzial	_	taki	nasz	&	umowny	&,	mozna	powiedziec	&,
jest	to	podzial	_	taki	nasz	_	umowny	&,	mozna	powiedziec	&,
jest	to	poglad	_	taki	nasz	_	umowny	&	mozna	powiedziec	&.
jest	to	podzial	_	taki	nasz	&	umowny	_	mozna	powiedziec	&,
jest	to	podzial	&	taki	nasz	_	umowny	&	mozna	powiedziec	&,

Text editing – comparison of transcribers of Polish

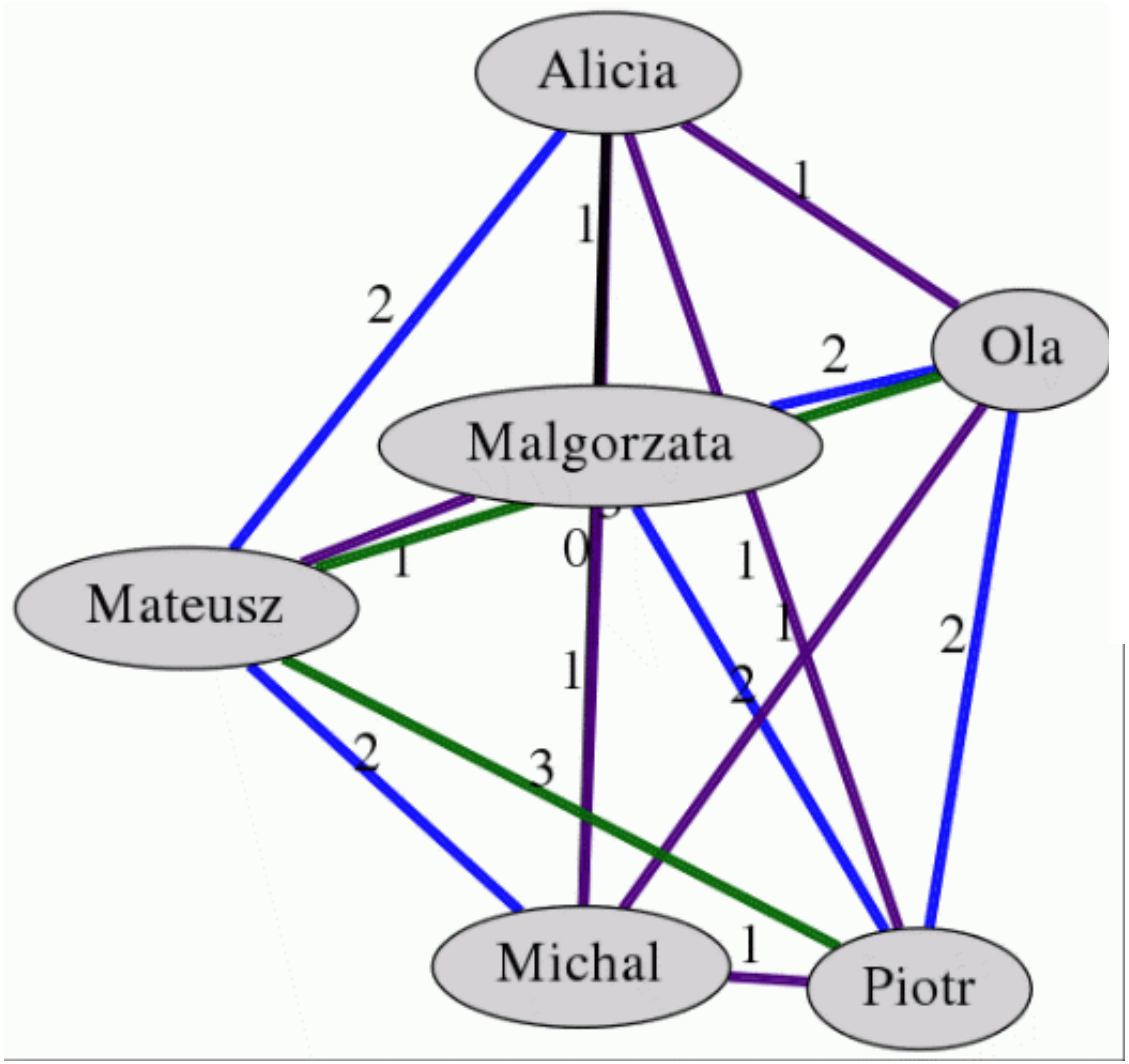


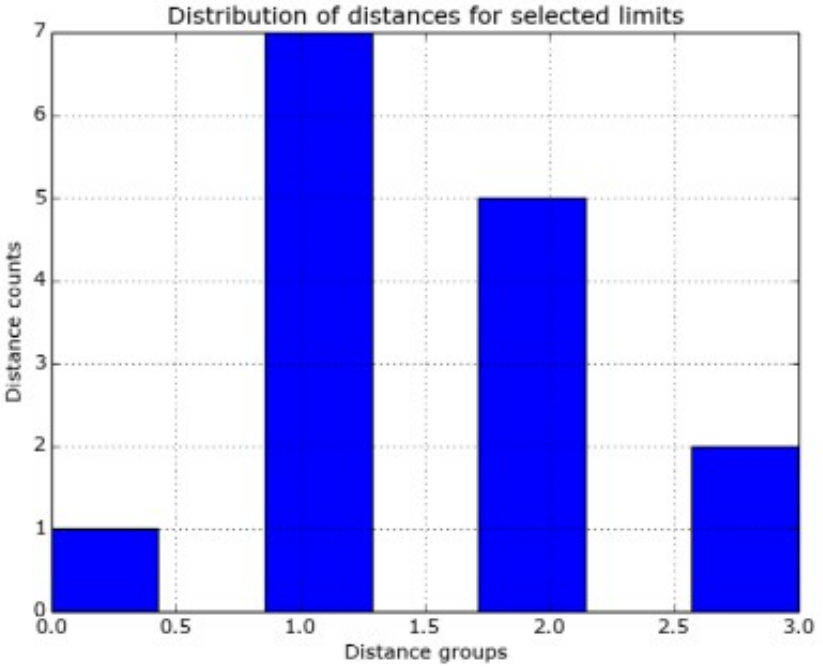
Text editing – comparison of transcribers of Polish





Text editing – comparison of transcribers of Polish





Name	Gender	Mean dist
Alicia	F	1.000
Michal	M	1.000
Malgorzata	F	1.400
Ola	F	1.800
Piotr	M	1.800
Mateusz	M	2.200

The moral of this story is that

Comparative text editing is facilitated by quantitative methods

- e.g. the Levenshtein Edit Distance algorithm and variants such as the Hamming Distance algorithm

for identifying differences in texts.

This particular application is in speech technology, but there are many other applications of this techniques.

Graphical visualisations are helpful in either suggesting or underlining lines of investigation.

Case Study 4: Recovery and technological application of legacy data

Based on data provided by Zakari Tchagbale

Legacy Tem data – recovered and applied to speech technology

Project: From legacy data to speech technology

Hypothesis:

Legacy data from sparsely documented languages can be updated and given speech technology applications.

Method:

1. Speech recording from legacy data
2. Creation of diphone database
3. Speech synthesis application

Results:

Gibbon, Dafydd, Eno-Abasi Urua and Moses Ekpenyong (2006). Problems and solutions in African tone language Text-To-Speech. In: Justus Roux, ed., Proceedings of the Multiling 2006 Conference, Stellenbosch, South Africa.

Gibbon, Dafydd, Ugonna Duruibe and Jolanta Bachan (2012). 'Market Speak' in Igbo: A speech synthesis training project. In: Hugues Steve Ndinga-Koumba-Binza and Sonja E. Bosch, eds., Language Science and Language Technology in Africa. Festschrift for Justus Roux. Stellenbosch: Sun Press, 339-359.

Here:

Not only 'sonification' of transcriptions, but practical application

Legacy Tem data – recovered and applied to speech technology

Tem

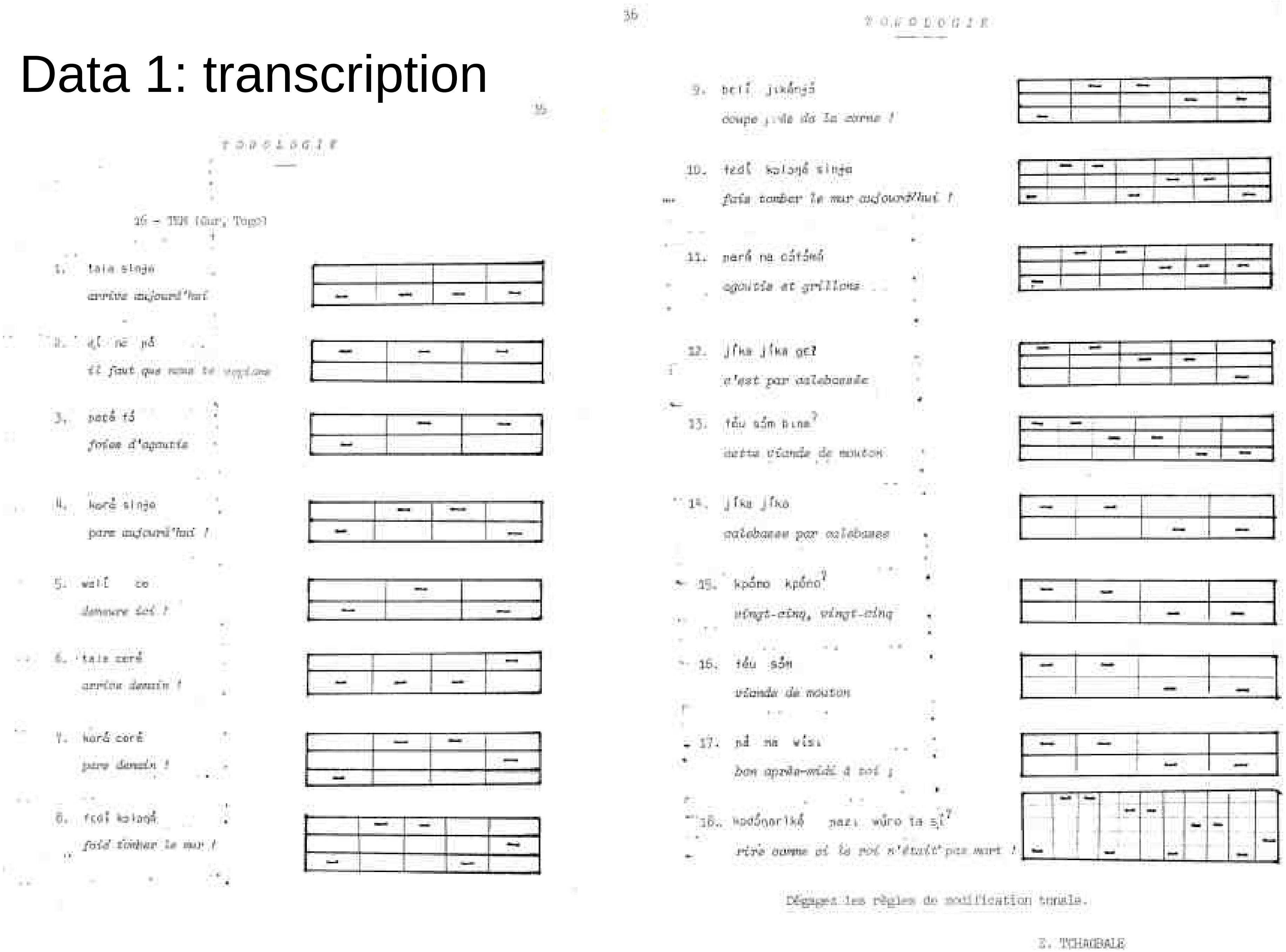
ISO 639-3 *kdh*

- Togo

- Niger-Congo
  > Atlantic-Congo
  > Volta-Congo
  > North
  > Gur
  > Tem

Tchagbale, Zakari. 1984. T.D. de Linguistique: exercises et corrigés. Institut de Linguistique Appliquée, Université Nationale de Côte d'Ivoire, Abidjan, No. 103.

Data 1: transcription


35

TONOLOGIE

16 – TEM (Gur, Togo)

1. taia singa
arrive aujourd'hui

2. ŋɩ́ nɛ pɔ́
il faut que nous te voyions

3. pecɛ́ tɔ́
foies d'agoutis

4. korɛ́ singa
pars aujourd'hui !

5. walɩ́ cɔ
demeure ici !

6. taia cerɛ́
arrive demain !

7. korɛ́ cerɛ́
pars demain !

8. fɛdɩ́ kɔlɔŋɔ́
fais tomber le mur !

36

TONOLOGIE

9. bɛlɩ́ jikɔ́ŋɔ́
coupe ... de la corne !

10. fɛdɩ́ kɔlɔŋɔ́ singa
fais tomber le mur aujourd'hui !

11. nerá na cátɔ́mɔ́
agoutis et grillons

12. jɩ́ka jɩ́ka gɛ́?
c'est par calebassée

13. fɛ́u sɔ́m bɩna?
cette viande de mouton

14. jɩ́ka jɩ́ka
calebasses par calebasses

15. kpɔ́nɔ kpɔ́nɔ?
vingt-cinq, vingt-cinq

16. fɛ́u sɔ́m
viande de mouton

17. nɔ́ na wísɩ
bon après-midi à toi !

18. kodɔ́ŋarɩkɔ́ nazɩ wúro ta sɩ́?
rire comme si le roi n'était pas mort !

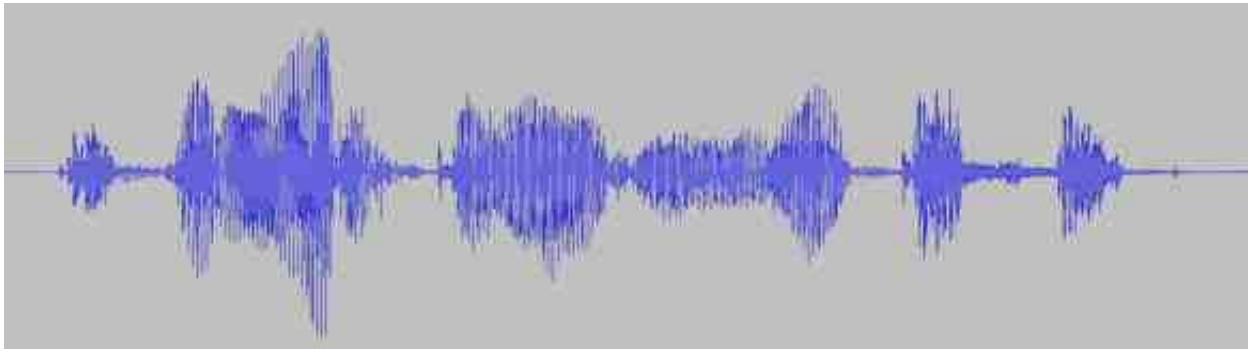
Dégagez les règles de modification tonale.

E. TCHAGBALE

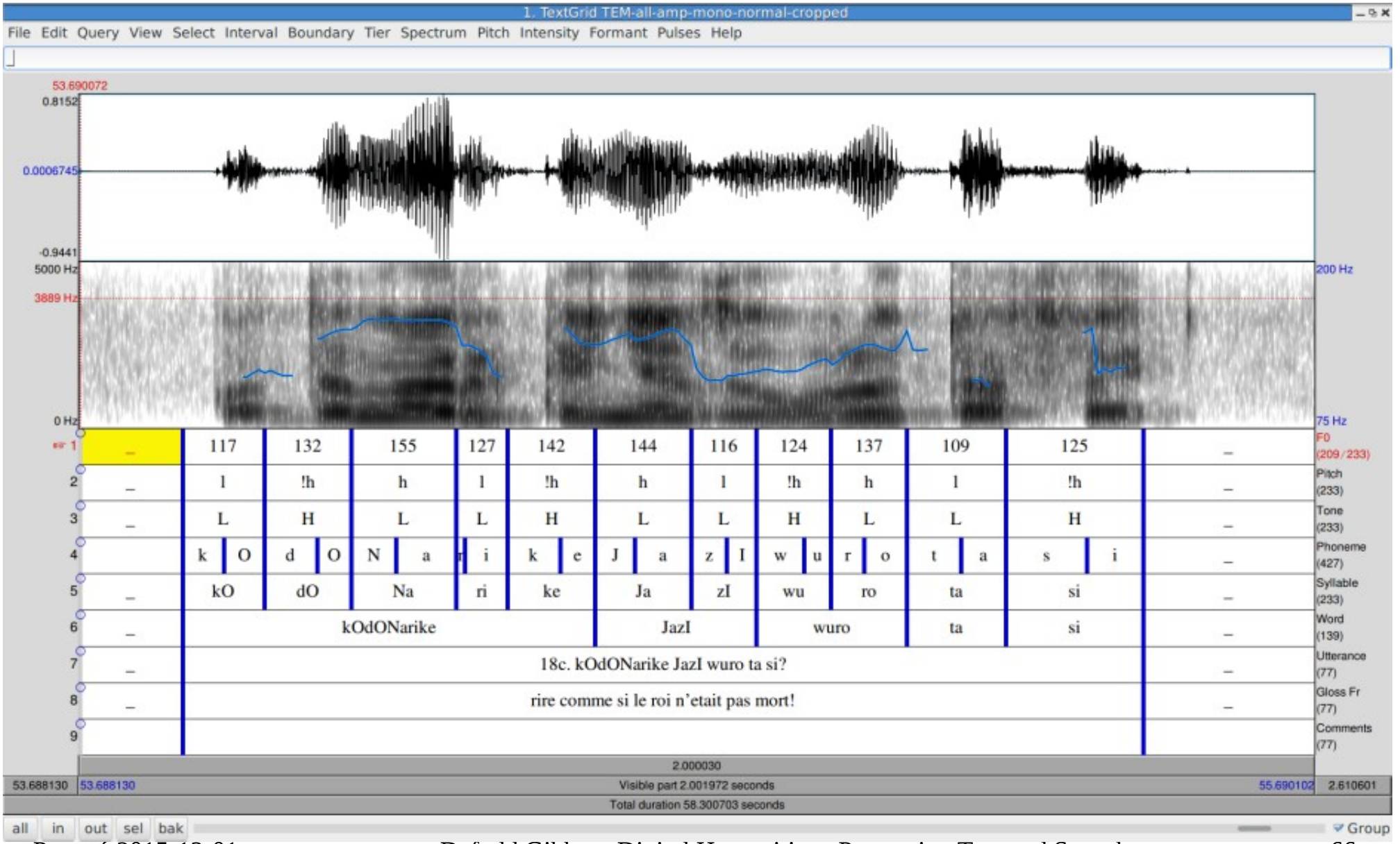
Data 2: speech recording

‘Facebook Fieldwork’: ☺

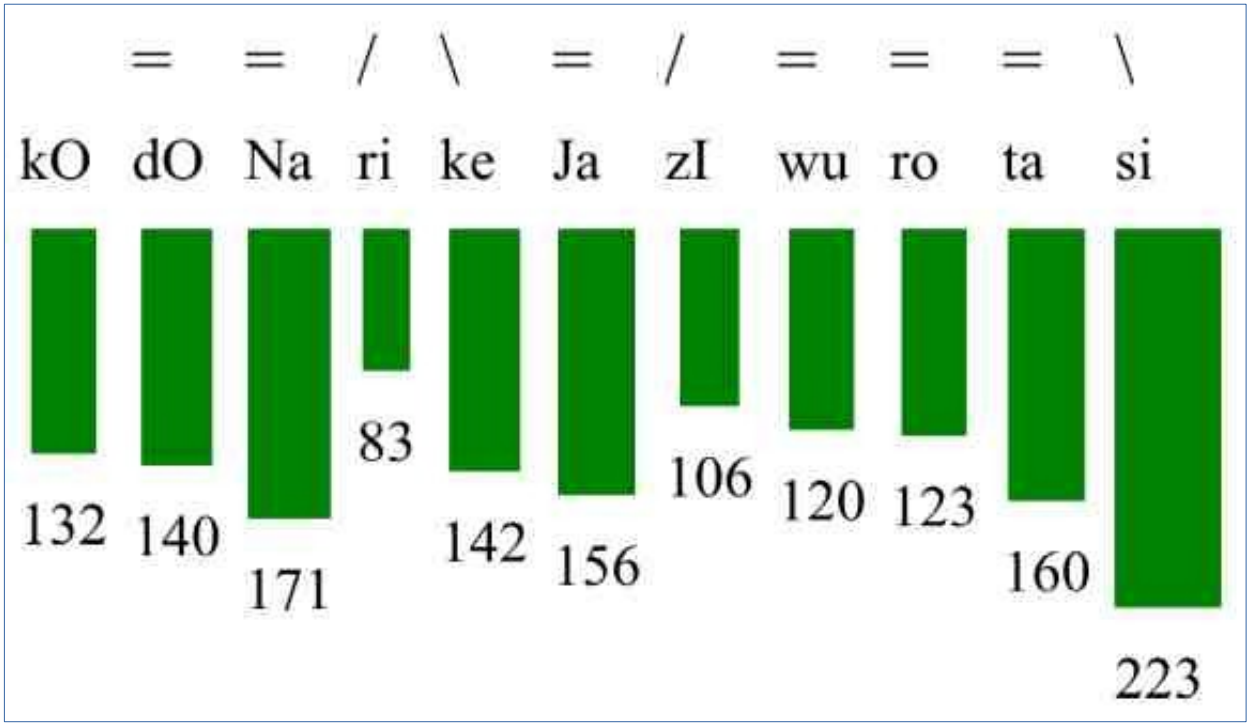
- Zakari Tchagbale’s “Exercices et corrigés” (1984)
  - resources for many languages, including Tem (1984)
  - DG had already developed a tone model for Tem (1987)
- DG asks ZT via internet for recording of Tem data (2012)
- recording by ZT with mobile phone (2012)
- ZT emails recording to DG (2012)



Data 3: annotation



Legacy Tem data – timing patterns



TGA online tool: visualisation of syllable time relations

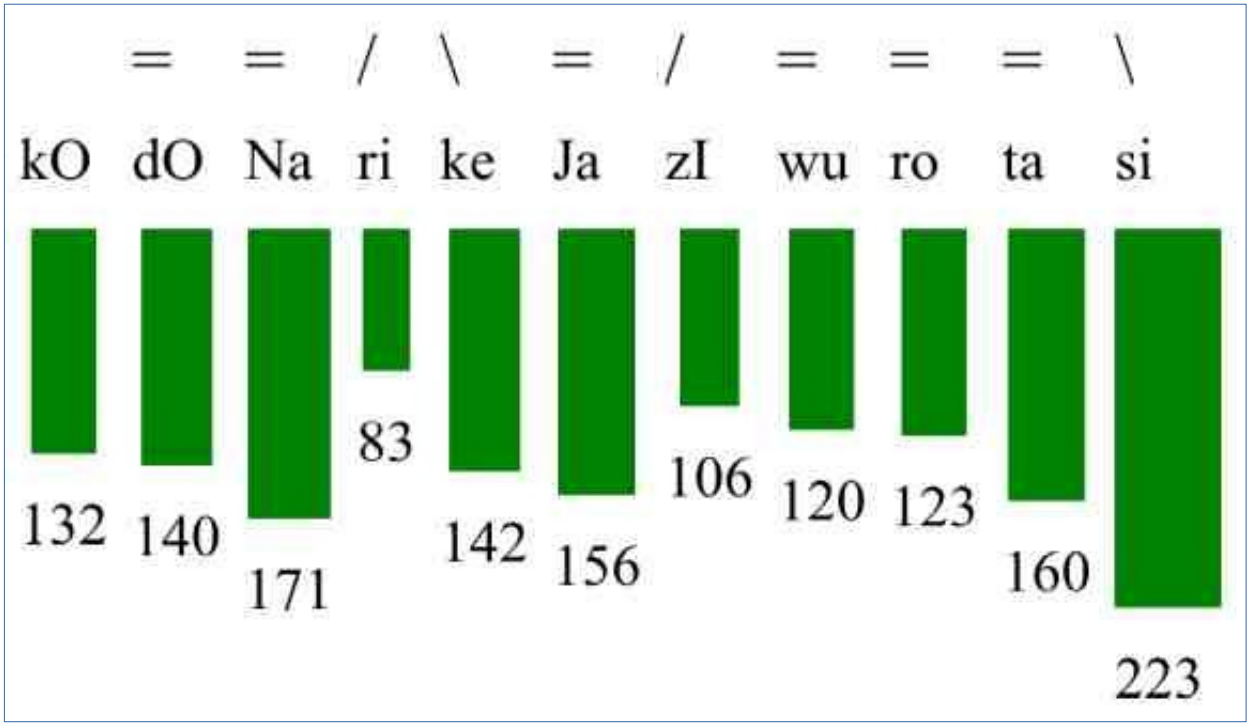
http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

Duration difference tokens:

Keyboard friendly transcription:

2D visualisation of durations:

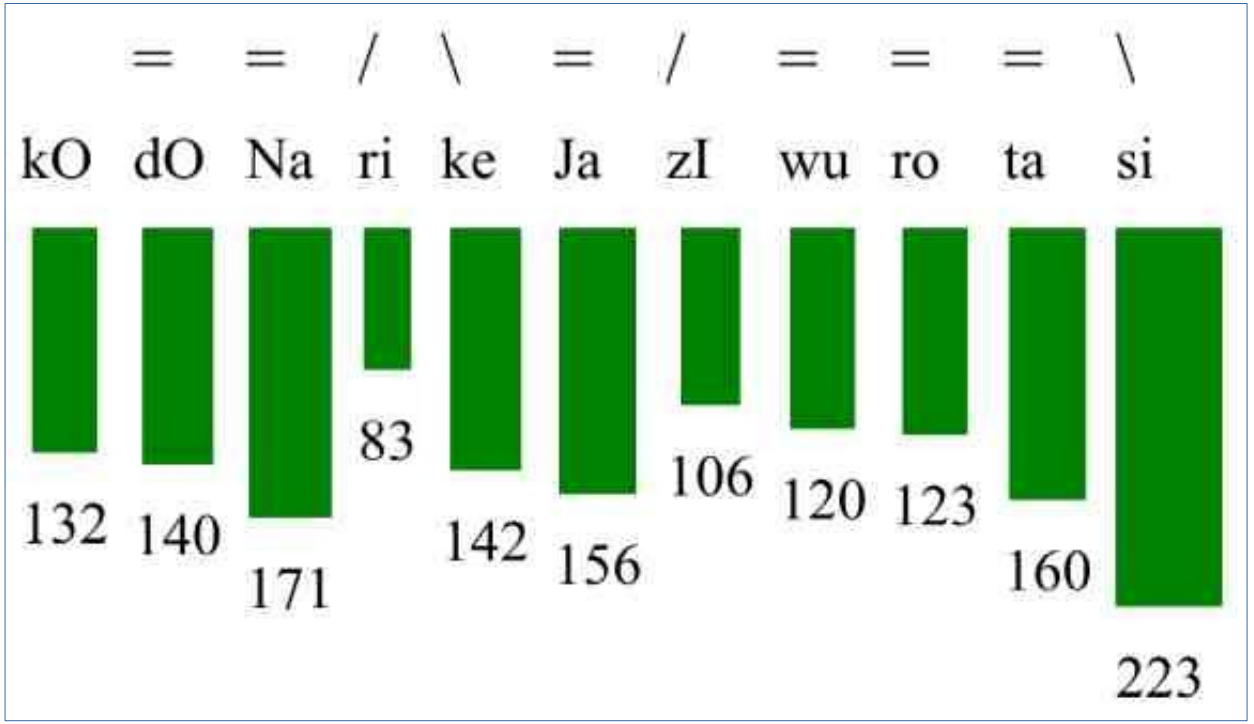
Syllable durations:

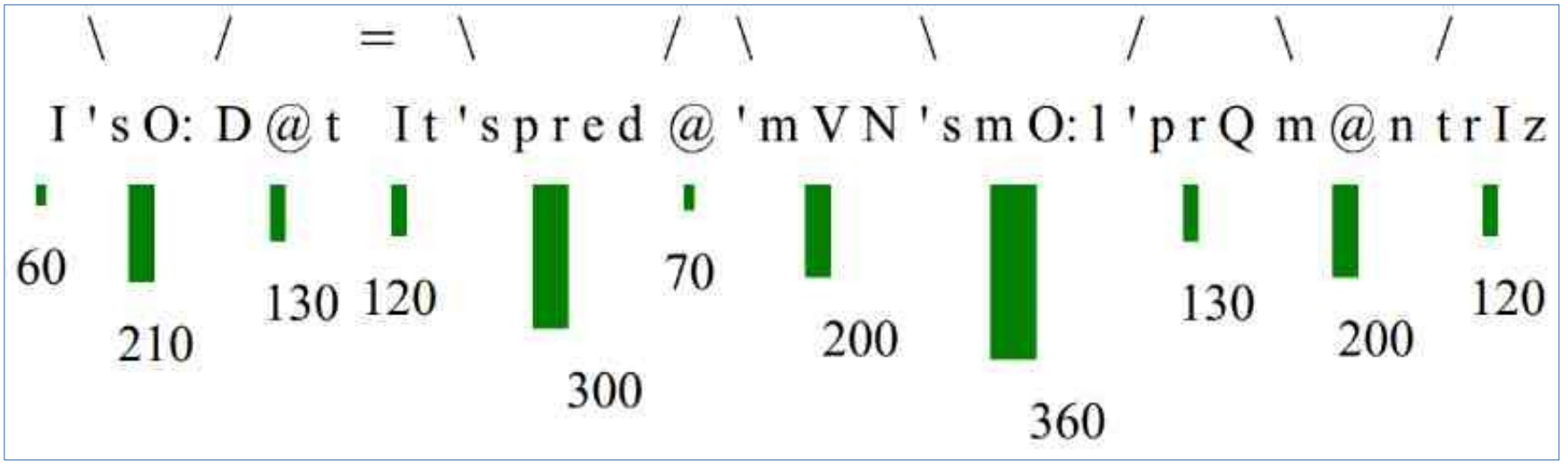


Duration difference tokens for utterance #37:

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

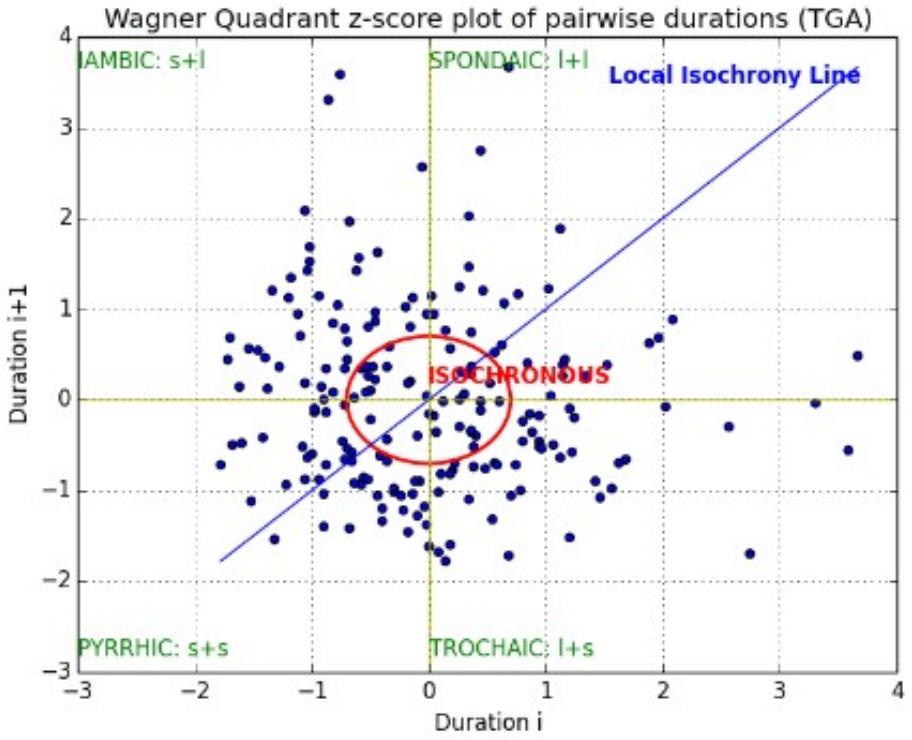
Legacy Tem data – timing patterns





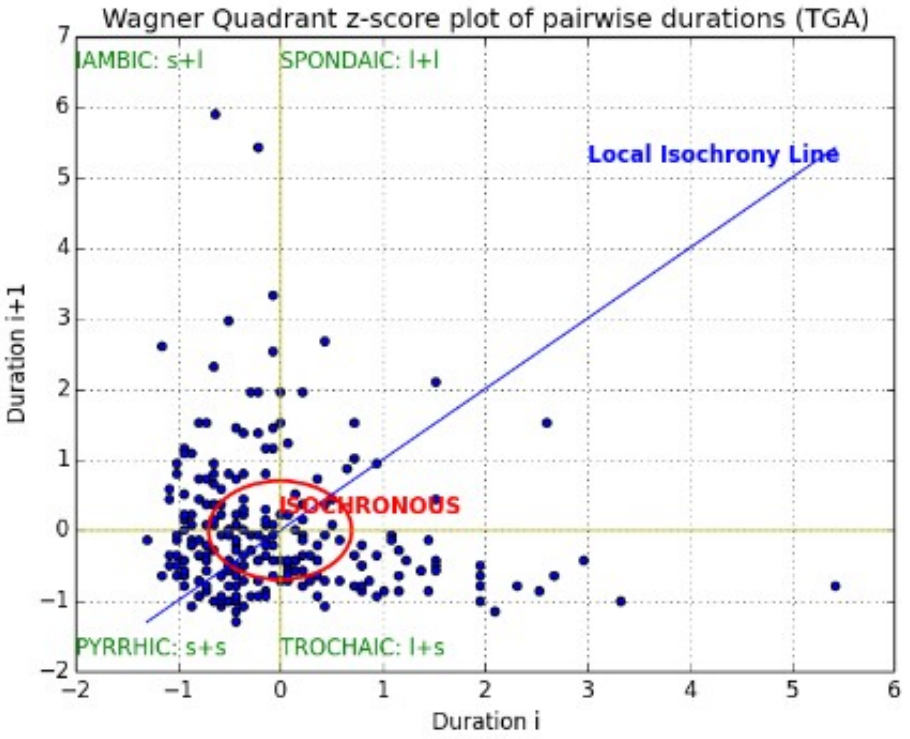
Compare with English!

Legacy Tem data – timing patterns



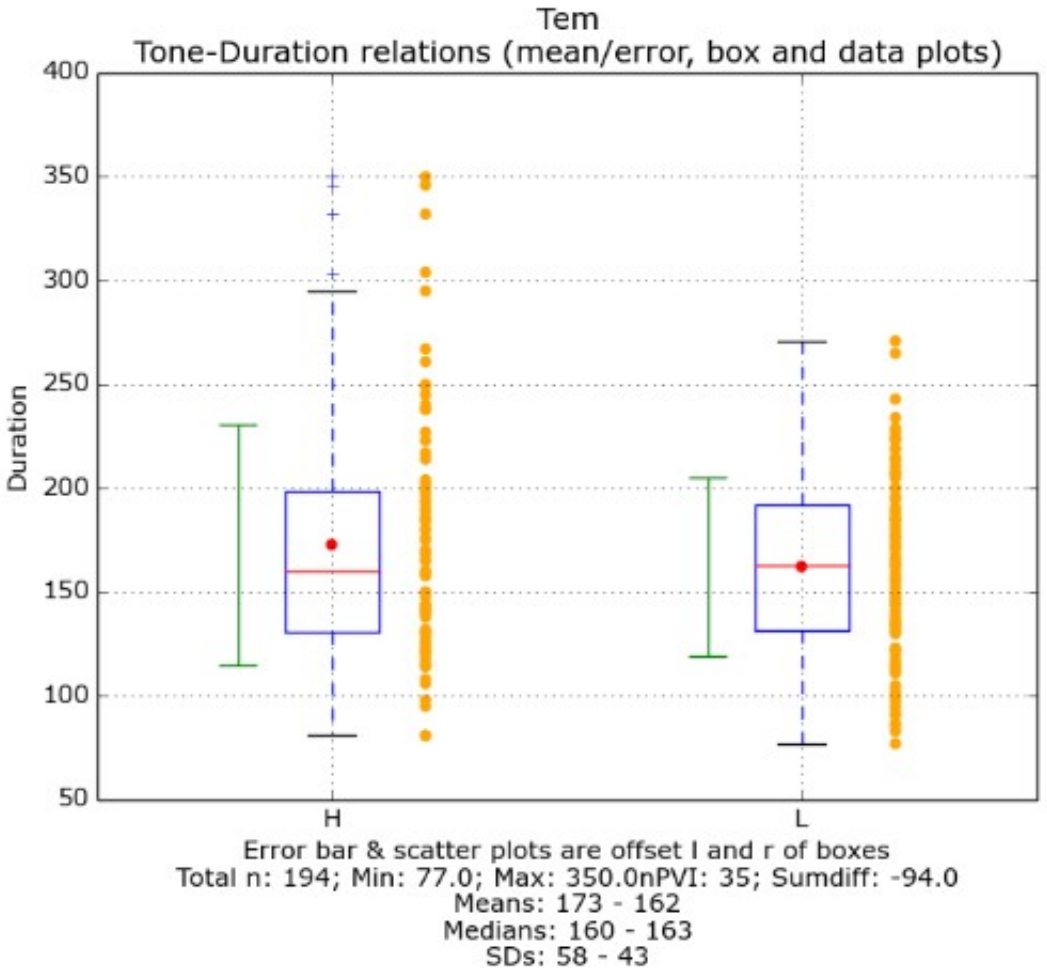
Wagner Quadrant graphs (scatter plots for duration z-scores).

Note the distributions of points in the four quadrants around the x and y zero values.

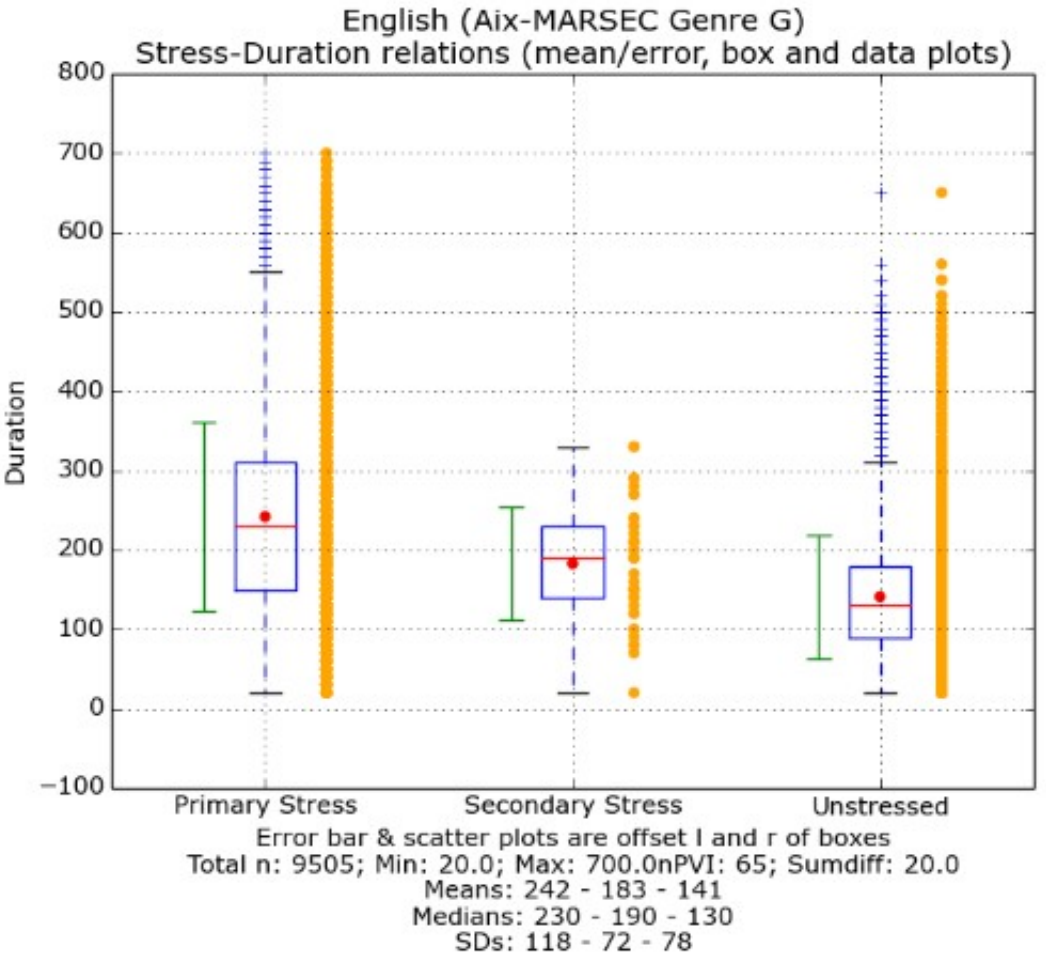


Compare with English!

Legacy Tem data – timing patterns



Box plots for tone/accent and duration relations



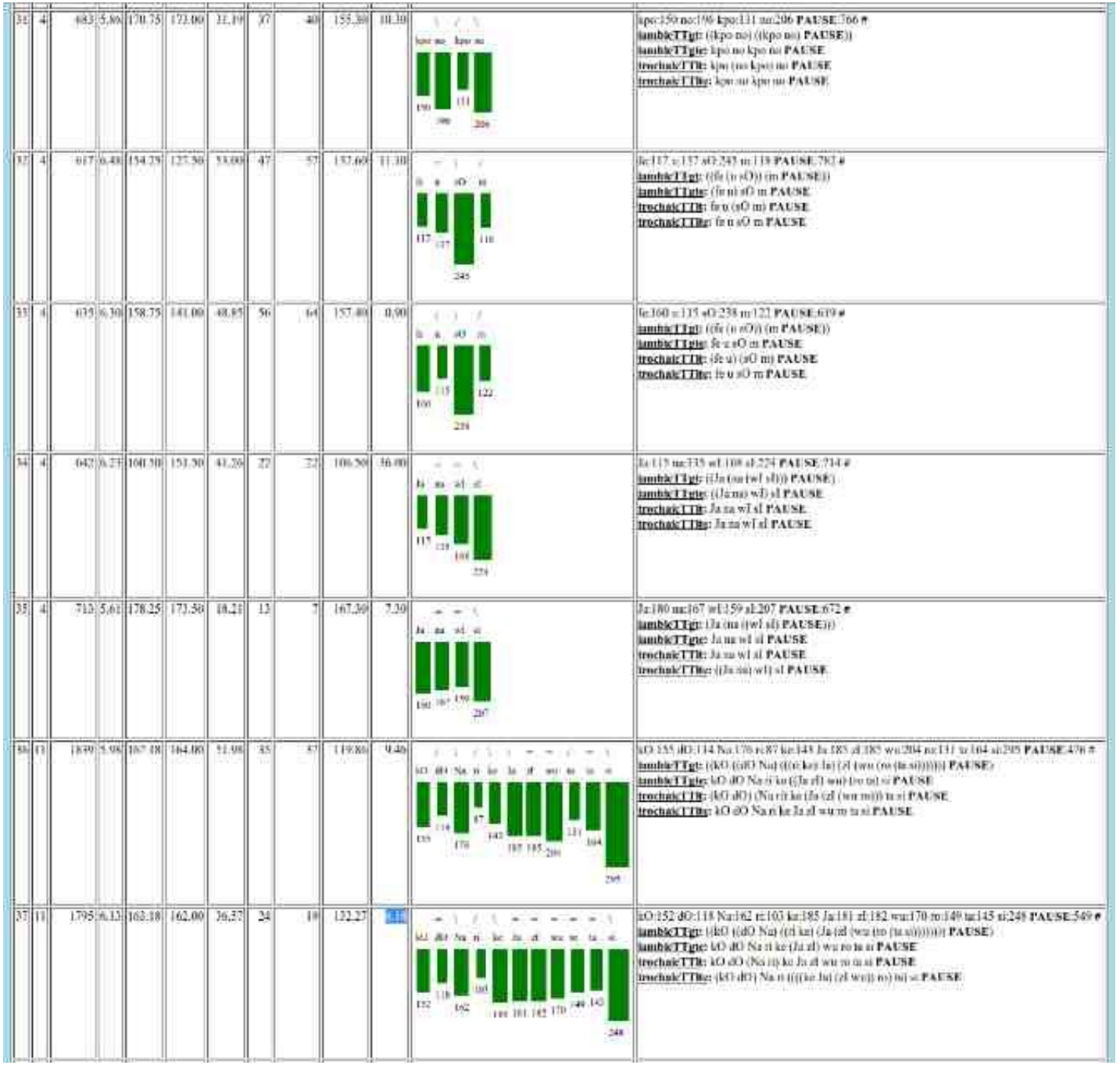
Compare w

Legacy Tem data – timing patterns

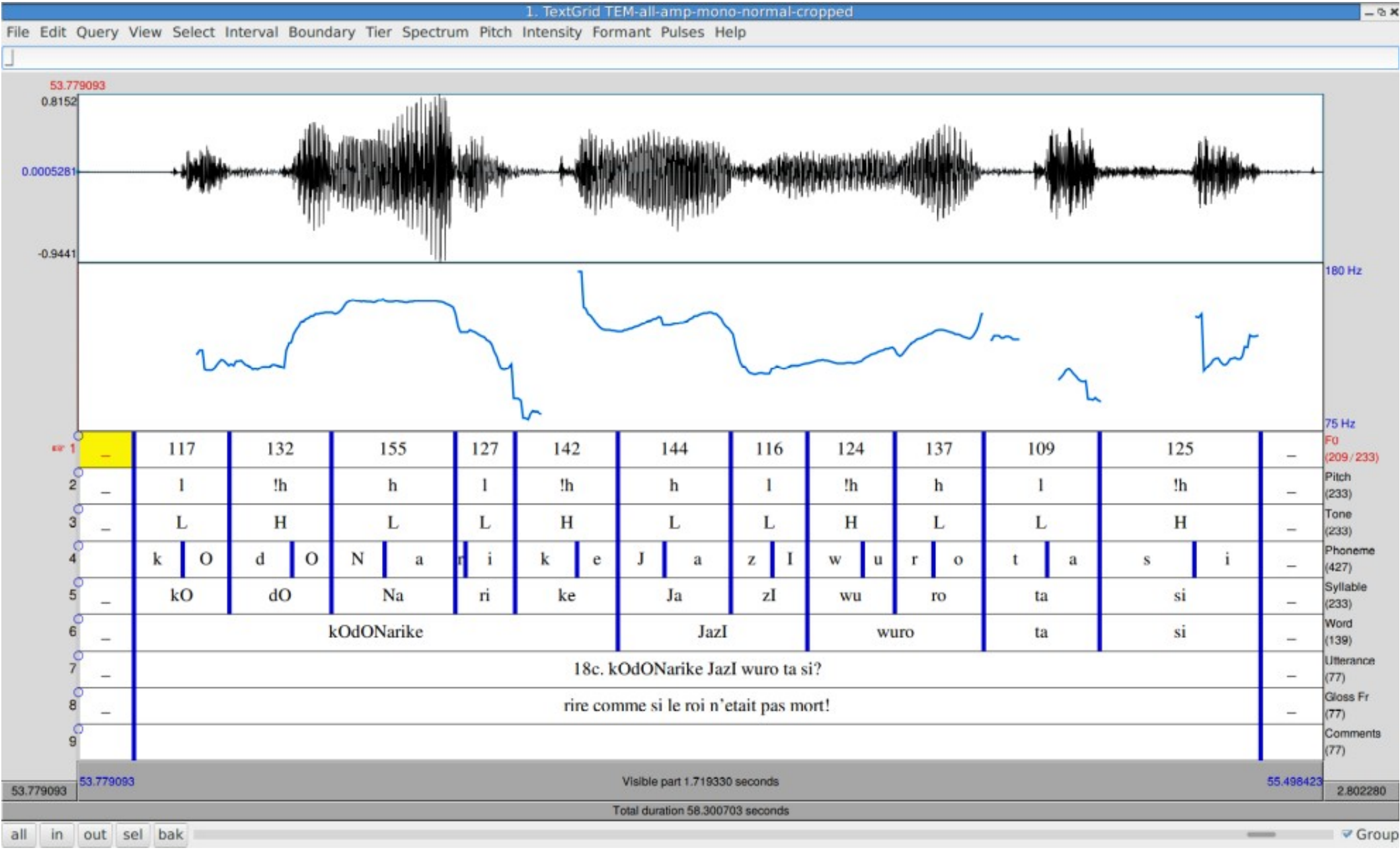
TGA output sample:

Visualisation of interpausal units

(screenshot)



Legacy Tem data – terraced tone sequences

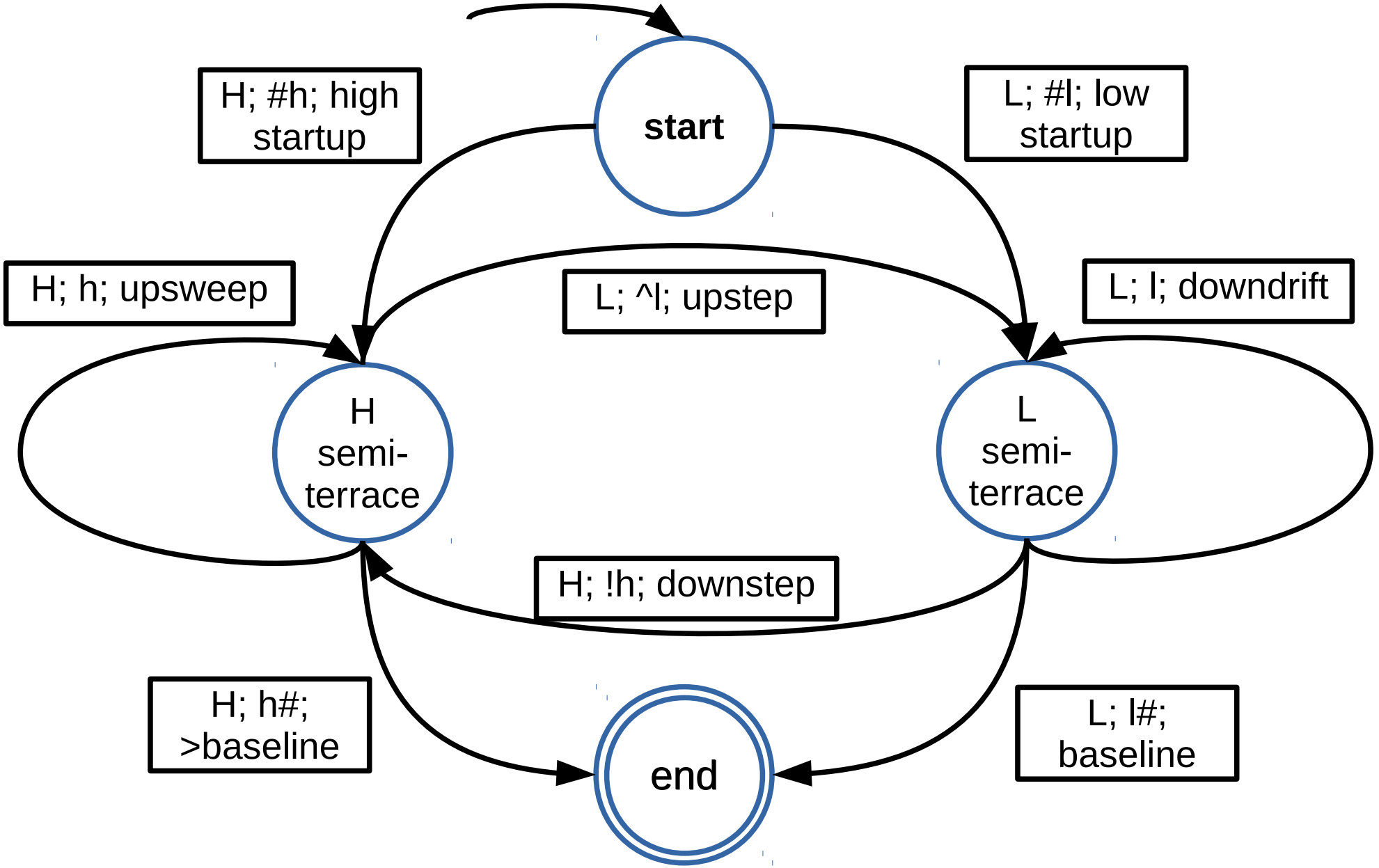


Legacy Tem data – terraced tone sequences

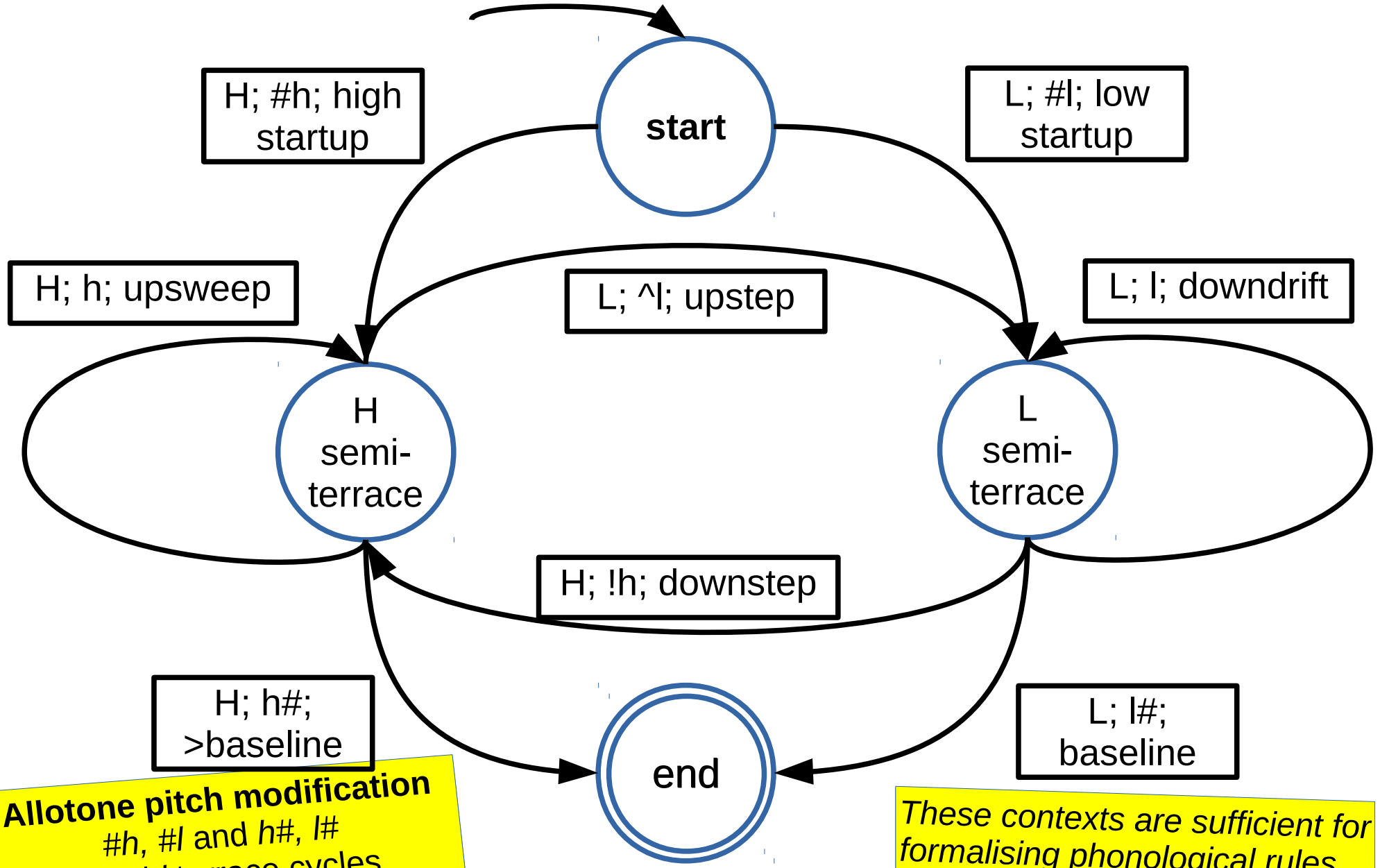
Tone	Pitch	*N*	mean (Hz)
H	!h#	10	128
	h#	9	129
	#h	14	154
	!h	56	131
	h	28	144
L	^l#	9	93
	l#	10	98
	#l	24	115
	^l	50	139
	l	60	113

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

Legacy Tem data – terraced tone sequences



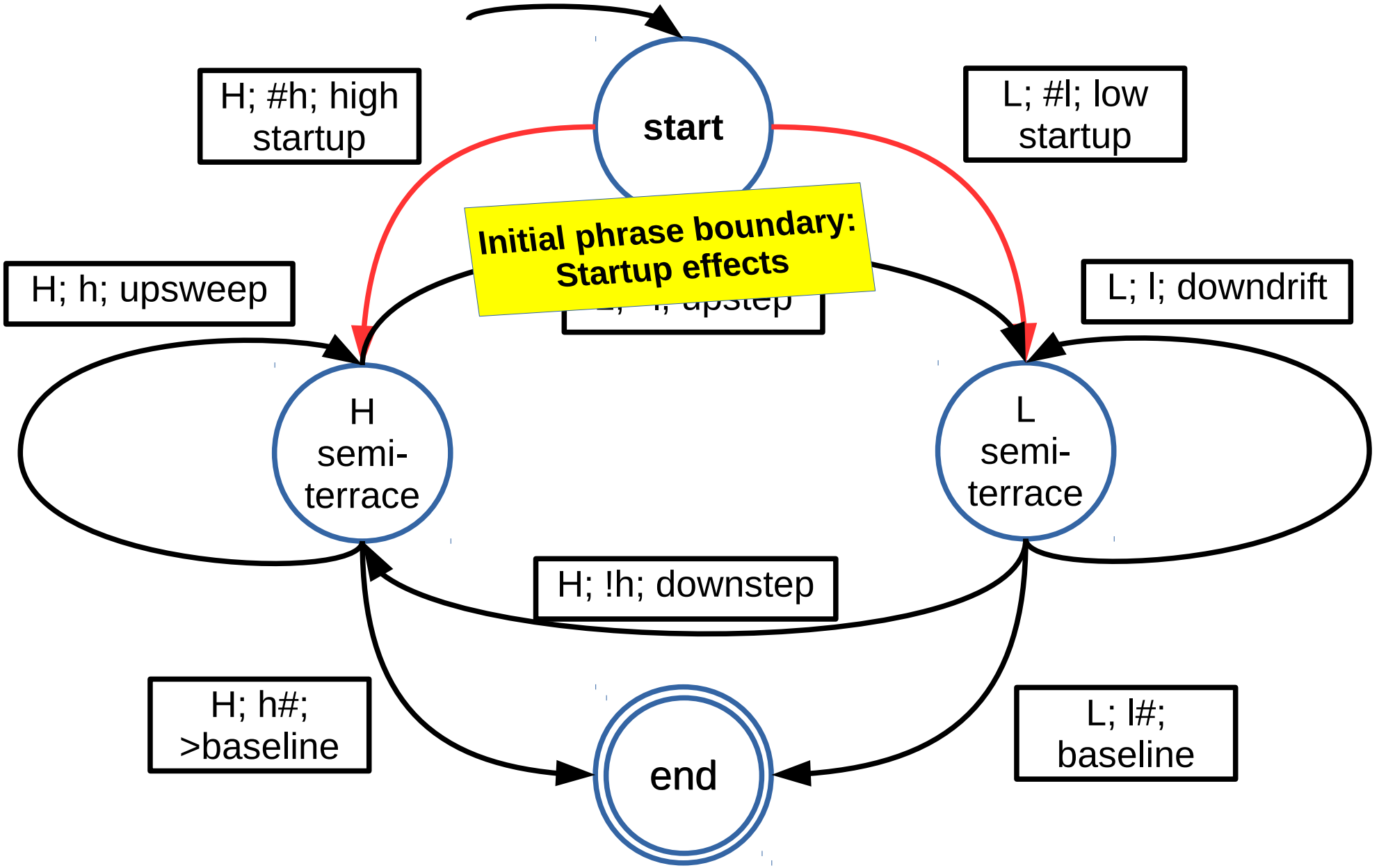
Legacy Tem data – terraced tone sequences



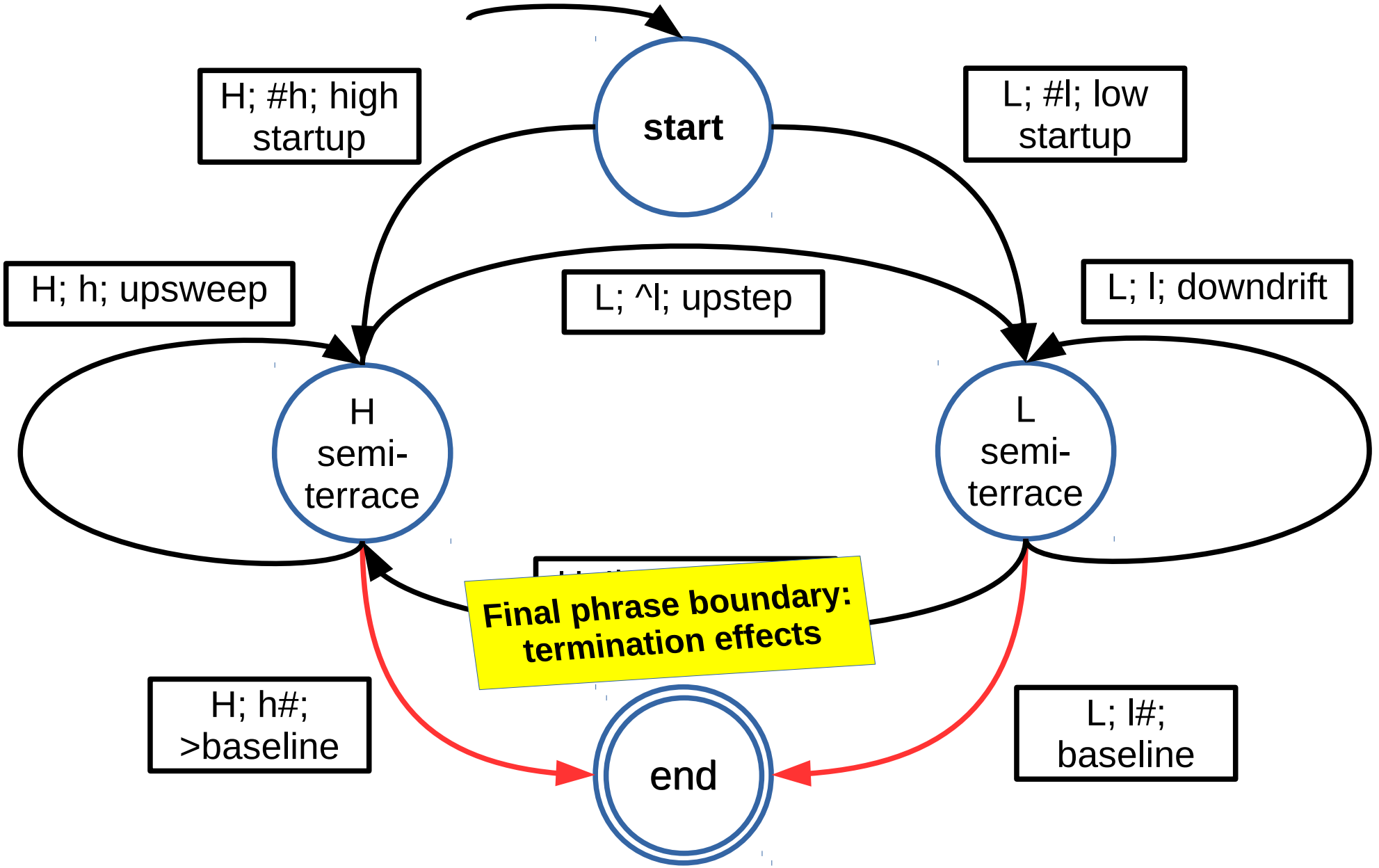
Allotone pitch modification
#h, *#l* and *h#*, *l#*
h and *l* terrace cycles
^l and *!h* terrace transitions

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

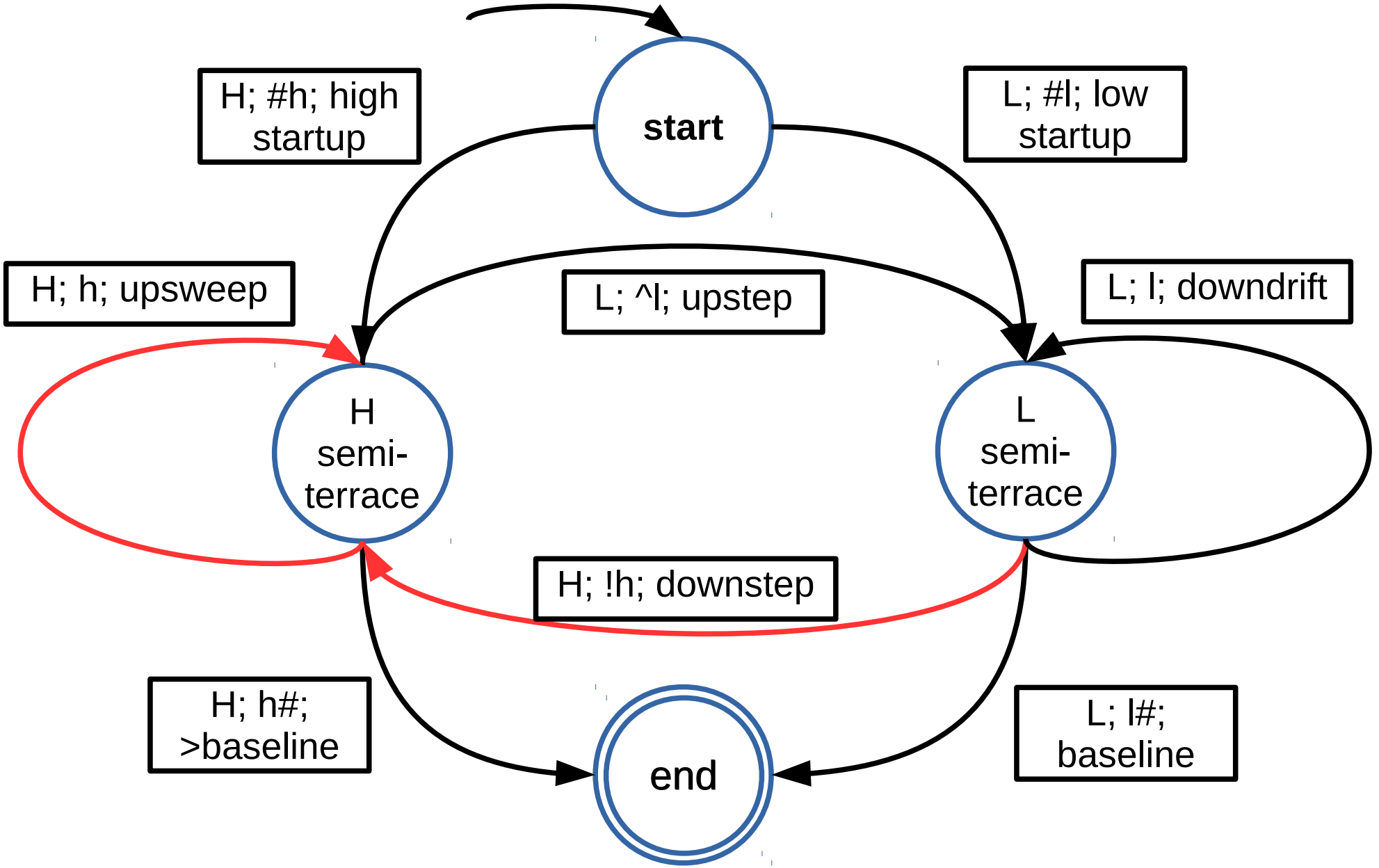
Legacy Tem data – terraced tone sequences



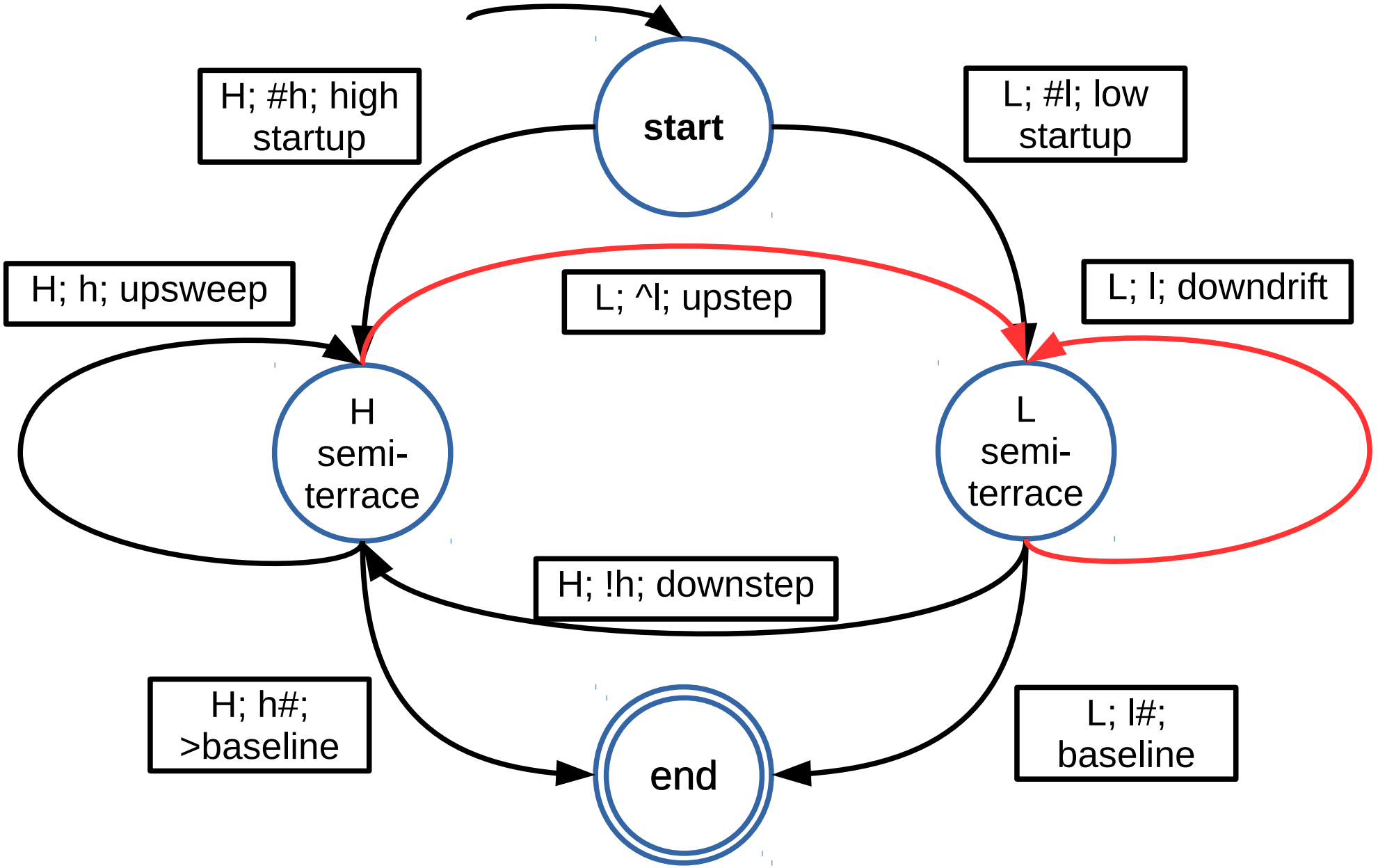
Legacy Tem data – terraced tone sequences



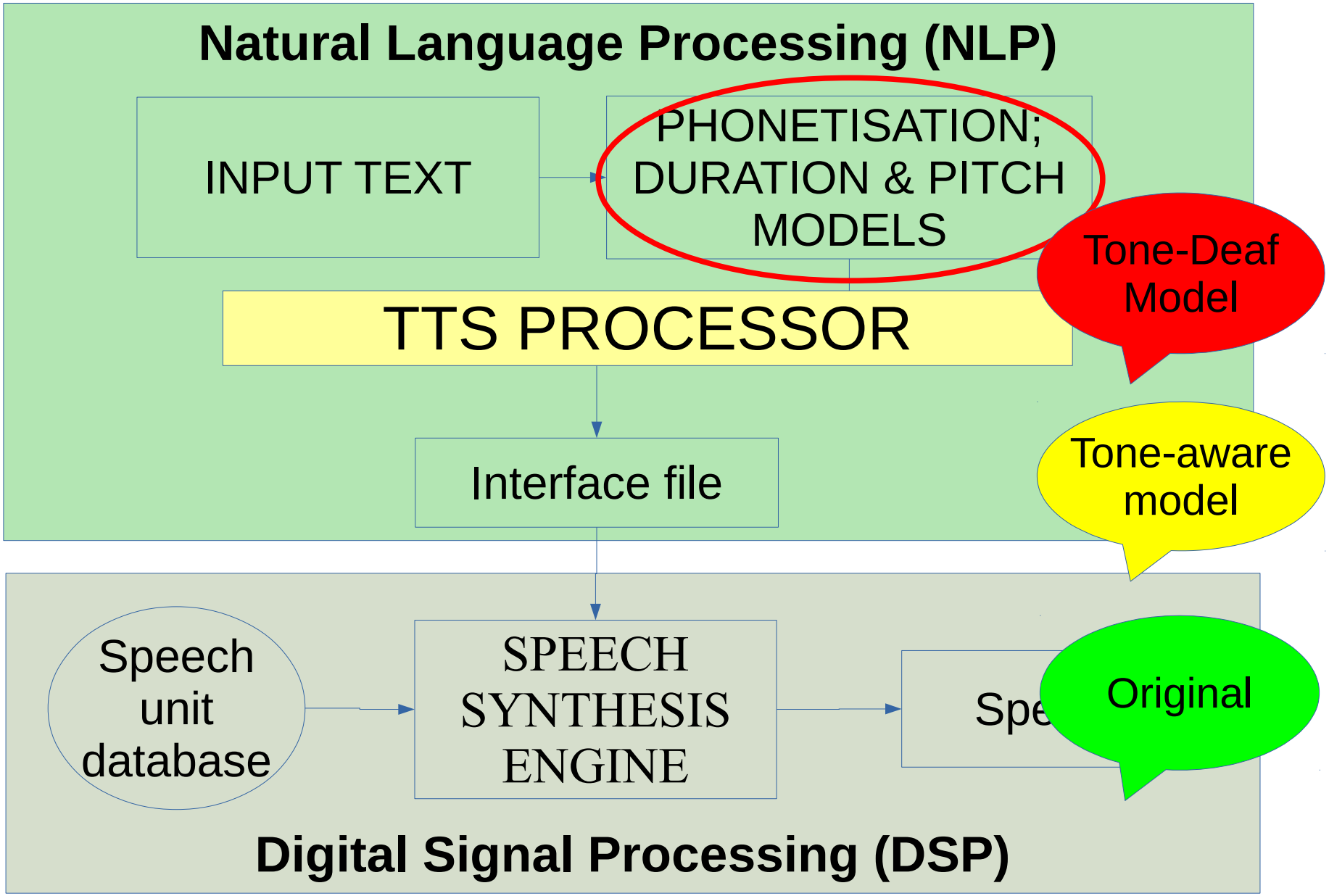
Legacy Tem data – terraced tone sequences



Legacy Tem data – terraced tone sequences



Legacy Tem data – modded and applied to speech technology



Legacy Tem data – modded and applied to speech technology

The moral of this story is that

Legacy data in linguistic atlases can be given a new lease of life and a solid quantitative foundation in addition to any further research on dialect relations and history which may be pursued.

Standard arrangements of quantitative information (e.g. tables) may be useful.

Graphical visualisations are helpful in either suggesting or underlining lines of investigation.

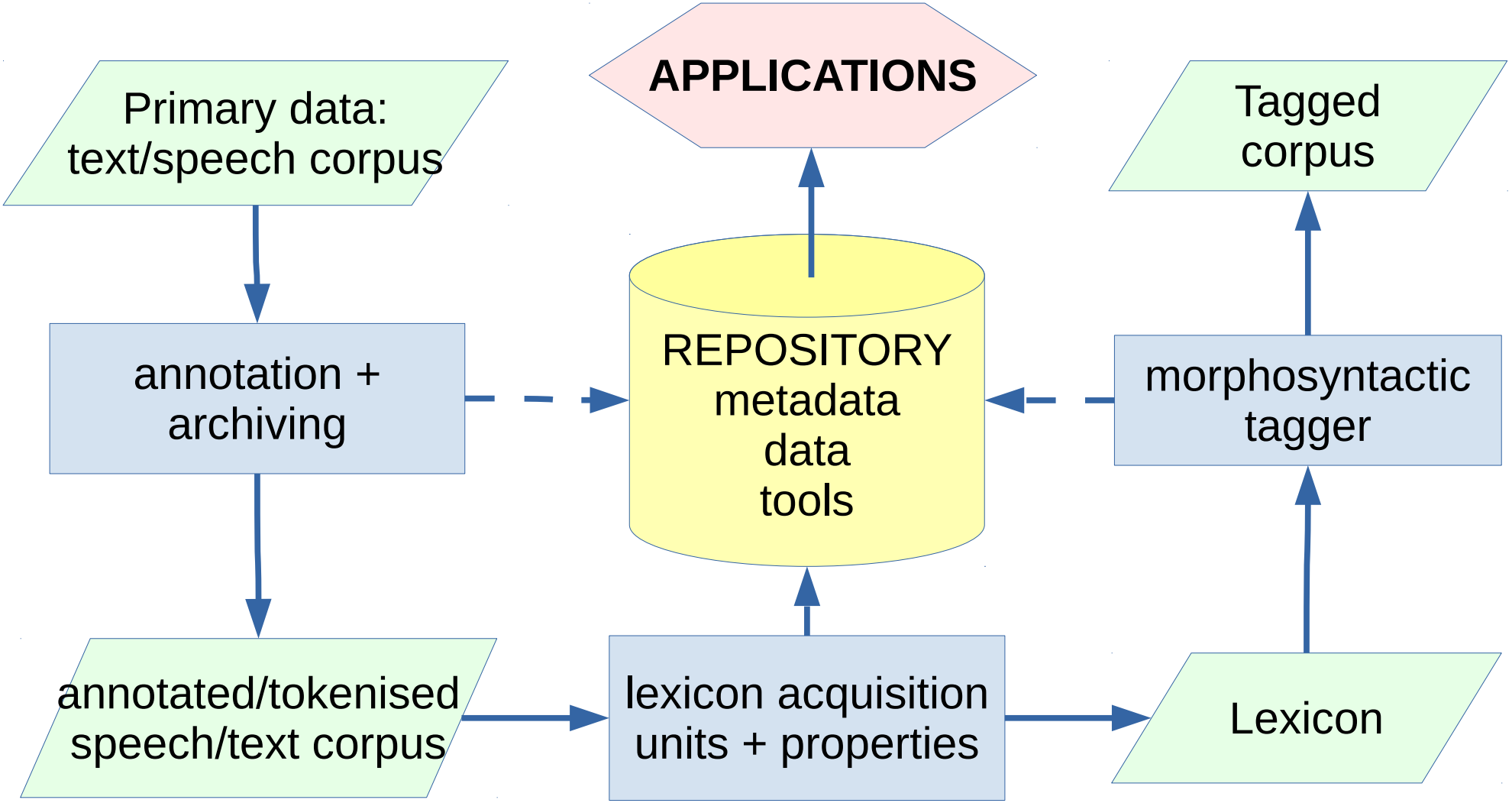
Conclusion

What endangered (and other) languages can contribute to HLT

So what do all these activities in literary computing, language typology and dialectology, text editing and speech technology have in common?

1. Cross-disciplinary team research, teamwork
2. Application of computing to oral and written human languages.
3. Use of high quality digitisation of resources:
    1. Socio-literary interview data
    2. Textual data
    3. Legacy textual language description
    4. Speech data

What endangered (and other) languages can contribute to HLT



RIM:

Rank Interpretation Model

Syntagmatic r

GRAMMAR
compositic

Paradigmatic properties

LEXICON – partial regularity, holistic opacity

DIALO
TE
SENTE
CLAUSE
PHRASE
COMPOUND WORD
DERIVED WORD
LEXICAL ROOT
MORPHEME
(MORPHO)PHONEME

Many properties of language do not become apparent until endangered and other less resourced languages are modelled.

These activities are urgently needed!

utic
ies

MULTIMODAL "PROSODIC" HIERARCHIES

Hypostatic properties *in different modalities*

RIM:

Rank Interpretation Model

Syntagmatic properties

GRAMMAR: compositionality

Paradigmatic properties

LEXICON – partial regularity, holistic opacity

DIALOGUE

TEXT

SENTENCE

CLAUSE

PHRASE

COMPOUND WORD

DERIVED WORD

LEXICAL ROOT

MORPHEME

(MORPHO)PHONEME

SEMANTICS/PRAGMATICS
CONCEPTS, OBJECTS, EVENTS

Hermeneutic properties

semiotic relation

MULTIMODAL "PROSODIC" HIERARCHIES

Hypostatic properties *in different modalities*

There are many infrastructural initiatives which support RMA types of work:

- Europe, North America, Far East, Australia:
  - projects and other initiatives (*too many to mention*)
  - conferences (e.g. LREC, Oriental COCOSDA)
  - repositories (e.g. LDC, ELRA/ELDA, HRELP, PARADISEC)
  - SALTMIL, AfLaT; African Languages Technology Institute
- My experience in HLT projects:
  - SAM, EAGLES, DoBeS, EMELD, LEGO, VERBMOBIL
  - DAAD project: linguistic education in West Africa
  - COCOSDA (International COordinating COmmittee for Speech Databases and Assessment)
    - Oriental COCOSDA highly successful, annual conference in different Asian countries
    - African COCOSDA?

There are many infrastructural initiatives which support RMA types of work:

- Europe, North America, Far East, Australia:
  - projects and other initiatives (*to be mentioned*)
  - conferences (e
  - repositories (e. ARADISEC)
  - SALTMIL, AfLa y Institute
- My experience in HL ects:
  - SAM, EAGLES, DoE MELD, LEGO, VERBMOBIL
  - DAAD project: lingu education in West Africa
  - COCOSDA (Interna al COordinating COmmittee for Speech Databases nd Assessment)
    - Oriental COCOSDA highly successful, annual conference in different Asian countries
    - African COCOSDA?

Cooperation between HLT specialists and local linguists is urgently needed!

Many thanks!