Legacy Language Atlas Data Mining: Mapping Kru Languages

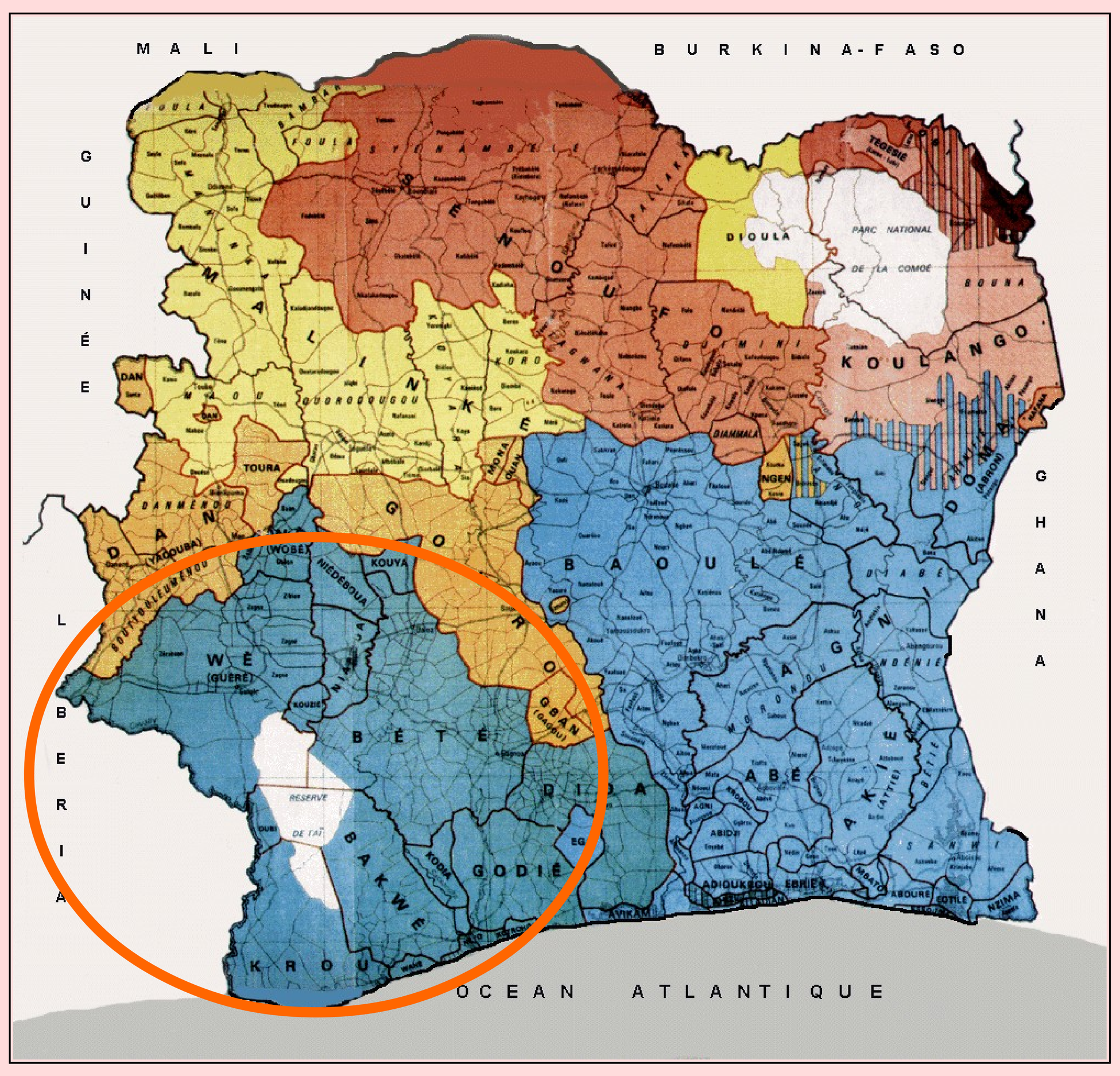
Dafydd Gibbon (Universität Bielefeld, Germany)

This study is dedicated to the memory of our late colleague and Symposium host, Henrike Grohs, Director of Abidjan Goethe Institut, cruelly murdered by terrorists in Grand Bassam, Côte d'Ivoire 13th March 2016.

Resource type: data + online tool 'DistGraph'

* Classification of Kru languages in the context of a 'Language in Context' documentation project of Côte d'Ivoire languages: Stavros Skopeteas (Bielefeld), Firmin Ahoua (Abidjan), Dafydd Gibbon (Bielefeld), in cooperation with François Kipré Blé (Abidjan)
* Data revival: re-use of >30 year old Language Atlas:
  - digital reproduction of data: scanning, retyping, redrawing
  - crosscheck of historical/typological classifications from Language Atlas as basis for new atlas

Data: the languages



South-West Ivory Coast

Ethnologue:
Niger-Congo (1537)
Atlantic-Congo (1440)
Volta-Congo (1367)
Kru (39)
+ Aizi (3)
- Eastern (11)
  + Bakwe (2)
  + Bete (5)
  + Dida (3)
  + Kwadia (1)
  + Kuwaa (1)
+ Seme (1)
- Western (23)
  + Bassa (3)
  + Grebo (9)
  + Klao (2)
  + Wee (9)

The 39 Ethnologue entries have ISO 693-3 codes, but Cedepo, Dewoin, Koyo and Niaboua are not listed.
In some cases more than one language variety is listed. Dida and some Dida varieties are listed, but Atlas varieties named Dida de Lozoua and Dida F are not.

Few Kru languages have ISO 639-3 codes

Data: the language atlas

* Marchese, Lynell. 1984. *Atlas linguistique kru.* Agence de coopération culturelle et technique, Université d'Abidjan, 3ème éd.
* Contents: language sketch tables & maps for 19 languages
* Selection: consonant tables for 19 languages, 44 different consonants
* Why consonants and not lexical items?
  - Lexical items are highly heterogeneous, easily borrowed
  - Consonant systems are relatively stable, slow changing
  - Consonant change laws are well-established for many language families (cf. Grimm's Law, Verner's Law, High German Sound Shift)

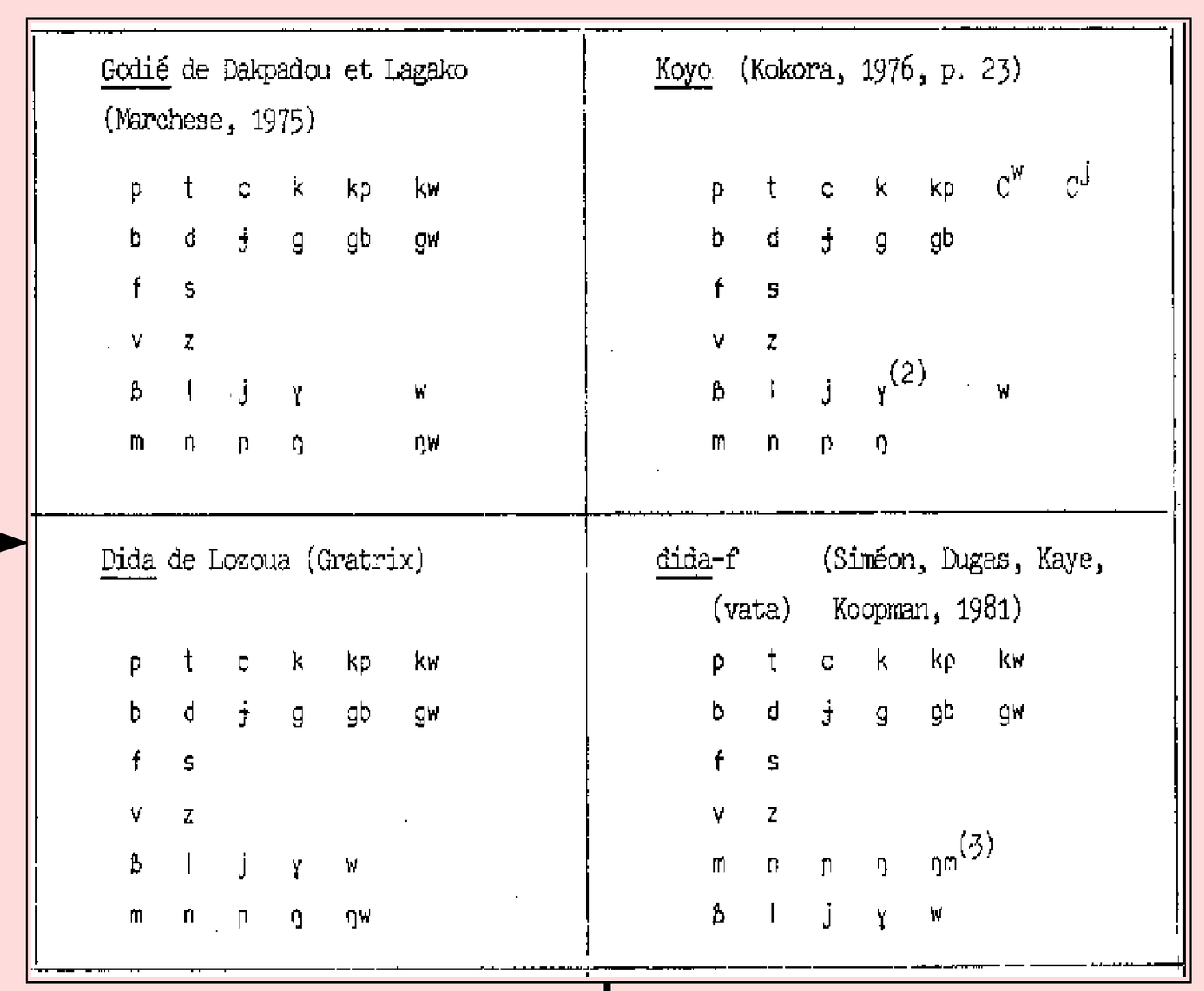
Method ('BLARK' for language typology?)

1. Input:
   19 ordered consonant sets x 44 features (consonants)
2. Outputs:
   1. pairwise difference matrix (Hamming distance)
   2. feature ranking list (variance)
   3. distance distribution histogram
   4. table of average distance/isolation
   5. table of specific pairwise differences

Implementation

* **Server-side web application:**
  - HTML → CGI → HTML+graphics
  - Linux, Windows (public & localhost)
  - Python 2.7
  - GraphViz neato engine (line drawings)
  - SciPy + MatPlotLib (dendrogram)
- **Client:**
  - (almost) any browser
  - resource demo:
    - localhost tablet & laptop
    - internet (see address on footer)

Data flow



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 cʷ cʲ
b d ɟ g gb
f s
v z
ɓ l j ɣ(2) 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 ŋm(3)
ɓ l j ɣ w

Parameter settings (+ CSV input field for consonant table)

IO parameters
Input table CSV separator: semicolon ▾
Graphics format: GIF bitmap graphics (smallest files) ▾

Output type:
- (●) *parametrised LED graph* *(properties of same attributes in same field position)*
- ( ) *parametrised SIRD graph* *(use only if properties in different fields are different, i.e. sets)*
- ( ) *CSV* ( ) *HTML* ( ) *XML formatted input data*
- ( ) *CSV* ( ) *HTML* ( ) *XML output of LED distance matrix*
- ( ) *CSV* ( ) *HTML* ( ) *XML output of LED distance triples*

Graph parameters
Graph engines (from AT&T GraphViz package):
- (●) *neato spring model*
- ( ) *dot undirected graph model*
- ( ) *twopi centred circle model*
- ( ) *circo circle model*

Numerical parameters:
- 0 ... 6 *range of distances to be processed (check distance matrix for full data range)*
- 6 *random seed for neato spring model (trial and error)*
- minimal scaling ▾
- 90 *% graph width (percent of window)*

```
<i>Language Atlas mining:</i><br>Consonant sets of Kru
languages.<br>Data source: Marchese, Lynell. 1984. <i>Atlas
Linguistique des Langues Kru.</i> 3&egrave;me
```
title, comment, etc. (HTML formatting permitted)

Aligned Language x Feature (=consonant) table

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

Hamming distance measure

For binary sequences of equal length:
Feature coding is {1,0}

$$\sum_{i=1}^{n} |x_i - y_i|$$

Length normalisation not used:

$$\frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

Pairwise distance matrix (column headers = row headers)

Marchese's classification

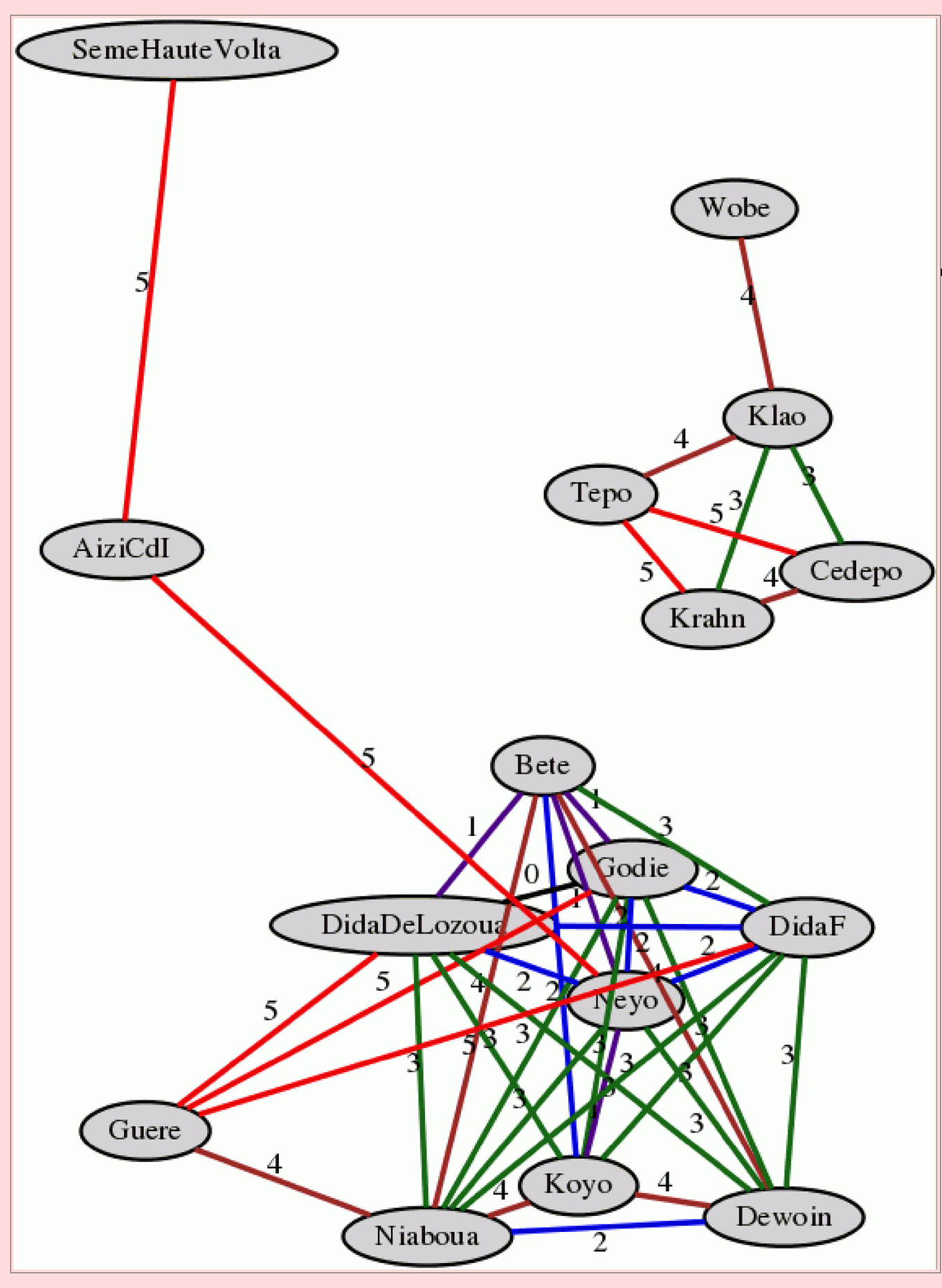
Eastern	Bete	0	1	2	1	1	3	10	6	9	11	8	4	4	7	11	8	12	9	6
	Godie	1	0	3	2	0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
	Koyo	2	3	0	1	3	3	12	8	9	11	8	4	4	9	13	8	12	9	6
	Neyo	1	2	1	0	2	2	11	7	8	10	7	3	3	8	12	7	11	8	5
	DidaDeLozoua	1	0	3	2	0	2	11	5	10	12	9	3	3	8	12	9	13	10	7
	DidaF	3	2	3	2	2	0	11	5	10	10	7	3	3	10	12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

Distance (Difference) Map
(force/spring map)

DIMENSION REDUCTION
CLASSIFICATION
VISUALISATION

Typological Similarity Dendrogram
(hierarchical clustering)

Virtual distance map
(0...5 pairwise differences)



Similarity dendrogram