



computer-assisted historical language comparison: an African linguistics perspective



Promise Dodzi Kpoglu

6th June 2024, UCAD, Dakar

promise-dodzi.kpoglu@cnrs.fr/promisedodzi@gmail.com

Outline

- Introduction
 - General introduction
 - Objectives
 - Significance
- From data to cognate sets in automatic approaches
- Data in this study:
 - The raw data
 - Data cleaning steps
- Cognate detection
 - The input
 - The cognate set
- Conclusion

Introduction

- Confronted with a number of languages that show variant degrees of similarity and differences, one of the first tasks of the linguist/anthropologist is to understand the relationships between these languages.
- Historically, two main approaches have been employed:
 - Swadesh-like approaches (Lexicostatistics)
 - Greenberg-like approaches (Mass comparison)
- Whatever the approach, some discrete (non mutually exclusive) steps are often involved in such historical linguistics work.

Introduction

- Discrete (mutually non-exclusive) steps can involve:
 - Cognate identification: words that are related, especially in form – see Jäger 2019.
 - Establishment of sound correspondences –see List et al 2022, Kim et al 2023.
 - Proto-form reconstruction- See Meloni et al 2021 for instance.
 - Time-depth estimation – See Gray & Atkinson 2003, 2006.
 - Phylogenetic relationships – See Rama et al 2018.

Language	Family	Time-depth	Proto-form	Cognates
French	Romance	~1000	cane [kane]	chien [ʃiɛ̃]
Italian	Romance	~1000	cane [kane]	cane [kane]
Spanish	Romance	~1000	cane [kane]	can [kan]

Objectives

- I concentrate on cognate-detection and show how automated methods can be employed in this task.
- The objectives of this presentation are thus two-fold:
 - Show how raw data is processed for automated approaches to comparative historical linguistics work.
 - Demonstrate the methods involved in cognate detection.
 - Present preliminary results from these methods used on data available within the BANG project.

Significance

- The results in this presentation are pertinent for two reasons:
 - Data processing techniques presented here can be relevant for other African languages.
 - Methods used in this study are transferable to workflows defined for different projects that involve linguistic data.
 - Perfected methods are going to be used with data in which Bangime, the object of study of the current project, will be integrated.

Computational methods in cognate detection

- Computationally, it is generally understood that cognate detection involves two major steps (Rama et al 2018: 4):
 - Word similarity calculation: how close, often phonetically, are groups of words?
 - Cognate set partitioning : proposing sets and putting proposed cognates in these sets.
- A third stage induced by this two-step approach is phonetic alignment - cognate sets are aligned (List et al 2022:90).
 - Align the proposed cognate sets for further analysis.

Computational methods in cognate detection

- Alignment analyses have the advantage of :
 - helping to model differences between sequences.
 - allowing us to compare sequences i.e. cognates.

DOCULECT	CONCEPT	ALIGNMENTS	EDIT
Ampari	(e.g. bird) fly	k ì r + ì + y é	156 <input type="checkbox"/>
BankanTey	(e.g. bird) fly	k í s + í + y é	156 <input type="checkbox"/>
BenTey	(e.g. bird) fly	k í s + í + y é	156 <input type="checkbox"/>
Nanga	(e.g. bird) fly	k í s + í + y é	156 <input type="checkbox"/> =
PergeTegu	(e.g. bird) fly	k í r + í + y é	156 <input type="checkbox"/>
TommoSoTongoTongo	(e.g. bird) fly	k íí l + í + y é	156 <input type="checkbox"/>
YandaDom	(e.g. bird) fly	k í l + í + y é	156 <input type="checkbox"/>

DOCULECT	CONCEPT	ALIGNMENTS	EDIT
Ampari	(e.g. bird) fly	k ì - r + ì + y é	156 <input type="checkbox"/>
BankanTey	(e.g. bird) fly	k í - s + í + y é	156 <input type="checkbox"/>
BenTey	(e.g. bird) fly	k í - s + í + y é	156 <input type="checkbox"/>
Nanga	(e.g. bird) fly	k í - s + í + y é	156 <input type="checkbox"/> =
PergeTegu	(e.g. bird) fly	k í - r + í + y é	156 <input type="checkbox"/>
TommoSoTongoTongo	(e.g. bird) fly	k í í l + í + y é	156 <input type="checkbox"/>
YandaDom	(e.g. bird) fly	k í - l + í + y é	156 <input type="checkbox"/>

Data in this study: the raw data

- Data is from the Dogon and Bangime Linguistics project.
- Data similar to data curated by linguists working on the field.

English	français	core	Toro Tegu (Tc)	Ben Tey (Ben)	Bankan-Tey (C)	Nanga (Anda)	Donno So	Jamsay (Doue)
domestic animal (sb's) livestock	animal domestique (ensemble des animaux)	1	àrùzáká ~ àrzi	àrsě:-m\\àrsě:	jáwdi-m\\jáw	gàsègè	bèlú	àsègè, gàsègè
(young herder's animal favorite)	chameau	1	gáfèl	gábèl	gáfèl	gábèri		gápèl
camel	chat	1	ɲó	ò:ɲùw ^m ò-m	ò:ɲè ^m -m\\ò:ɲè ^m	ògò-ɲògò	ògò-yògò	òyò-ɲùgò
cat	bovin (vache ou taureau)	1	niw ^a	ni-niw ^a -m\\n	nùmā-m\\nùm	dòndiyè	gámmà	ni-niw ^a é
cow, bull (any)	veau	2	nàṅá	nā:-m\\nā:	nā:-m\\nā:	nàṅá	nàṅá	nàṅá
calf	vache	2	nàṅà-[ir ^a -í:]	nā: yí-m	nā:-yí-m\\nā:-	nàṅá yí:	nàṅà-í:	nàṅ-í: ^a
cow (female)	vache qui a mis bas au monde	3	nàṅà-yá	nā: yā-m		nàṅá yā:		nàṅà-nā:
cow that has calved	vache (avec un veau)	3	nàṅà-yà-nā:	nā: nā:-m	nā:-yā-m\\nā:-	nàṅá dé:		
cow (with a calf)	génisse, jeune vache (quand elle a mis bas)	2	nàṅà-yà girèy	nā: gùrò-m	wí:gè-m\\wí:g	nàṅá gùrā, nā:	nàṅà-gulò	nàṅá gùrò
heifer, young cow	taureau	2	nàṅà-àrá	nā: ár ^a -m		nàṅá ár ^a	nàṅá ànā	nàṅá ár ^a
bull	taureau (castré)	2	nàṅà-àrà pólù	nā: pòrú-m	nā:-ár ^a -m\\nā:	nàṅá pòrì	nàṅá pòr-ú	nàṅá pòrú
bull (castrated)	taureau (non castré)	2	nàṅà-àrà bòbò	nā: jàṅgú-m	nā:-jì-m\\nā:-j	nàṅá jàṅí	nàṅá kálàndù	nàṅá jàṅ
bull (uncastrate)	taureau (non castré, jeune)	3		nā: já: kírím		nàṅá kóróbòrì		nàṅá kóróbòrì
bull (uncastrate)	chien	1	nènú	injě-m\\injě	ɲjèrú-m\\ɲjèrú	nèr ^a i	idú	ijú
dog								

Data in this study: data cleaning steps

- The steps taken to get the data into computer digestible format are of three main types:
 - All manual process
 - Semi-manual process
- Manual process involved:
 - Removal of non alpha-numeric symbols (‘,’, ‘’), ‘/’)
 - ‘Columnization’ and ‘Rowing’:
 - Columns: ID, language, concepts, singular, plural
 - Rows: 1 ID, 1 language, 1 concept, 1 singular form, 1 plural form
 - Consensus forms, for especially verbs

Data in this study: data cleaning steps

- Data after manual process:

ID	VARID	DOCULECT	GLOSS	SINGULAR	PLURAL
1	0	BankanTey	domestic animal (esp. livestock	jáwdi-m	jáwdi
1	1	BankanTey	domestic animal (esp. livestock	dá:bà-m	dá:bà
2	0	BenTey	domestic animal (esp. livestock	àrsě:-m	àrsě:
3	0	Bunoge	domestic animal (esp. livestock	kómbò	kómbò=gè
4	0	DogulDomBendiely	domestic animal (esp. livestock	bélè-g	bèlè-g-yà
4	1	DogulDomBendiely	domestic animal (esp. livestock	bélè-gù	bèlè-gi-yà
5	0	DonnoSo	domestic animal (esp. livestock	bèlú	bèlú
5	1	DonnoSo	domestic animal (esp. livestock	bèlú	bèlú=mbè
6	0	JamsayGourou	domestic animal (esp. livestock	gàr sègé	gàr sègé
7	0	JamsayDouentza	domestic animal (esp. livestock	à sègé	à sègé=bé
7	1	JamsayDouentza	domestic animal (esp. livestock	gà sègé	gà sègé=bé
8	0	JamsayMondoro	domestic animal (esp. livestock	à sège	à sègé
9	0	Mombo	domestic animal (esp. livestock	dèbù béli	
9	1	Mombo	domestic animal (esp. livestock	béli	
10	0	BonduSoNajamba	domestic animal (esp. livestock	dúmè-ɲgó	dúmè:
10	1	BonduSoKindige	domestic animal (esp. livestock	dùmó:-n	dúmè:
11	0	Nanga	domestic animal (esp. livestock	gà sègé	gà sègé
12	0	Penange	domestic animal (esp. livestock	dá:bá	da:ba=gé
13	0	PergeTegu	domestic animal (esp. livestock	à sègé	
14	0	TebulUre	domestic animal (esp. livestock	à sègé	à sègè-mbó
14	1	TebulUre	domestic animal (esp. livestock	gà sègé	gà sègè-mbó
15	0	TiranigeBoui	domestic animal (esp. livestock	bèl	bel=gè
15	1	TiranigeNingo	domestic animal (esp. livestock	bèli	bèli=ge

Data in this study: data cleaning steps

- Semi-manual process involved:
 - Identifying morphotactic properties
 - Writing segmenting rules in the python programming language
 - Parsing the data based on these rules
- Identifying morphotactic properties:
 - Reading grammars on the languages
- Writing rules:
 - Using python to encode the rules
- Parsing data:
 - Running python script on data
 - Stabilizing graphemes using an orthograph

Grapheme	IPA
^ñi_V\$	n ð
→\$	NULL
^_..._	NULL
é:	é.é
è:	é.è
f:	f.f
ã:	à.á
è:	è.è
è:	é.è
í:	ì.í
ë:	è.é
ì:	ì.ì
ó:	ó.ó
ò:	ò.ò
ù:	ù.ù
ú:	ú.ú
õ:	õ.õ
ü:	ü.ü

```
class VerbParser(NounParser):
    def __init__(self):
        super().__init__()

    def consonant_count(self, item):
        """
        counts number of consonants in words
        """

        if item is None:
            return 0

        consonant_count=0
        for letter in item:
            if letter != " " and letter in self.consonants:
                consonant_count +=1
        return consonant_count
```

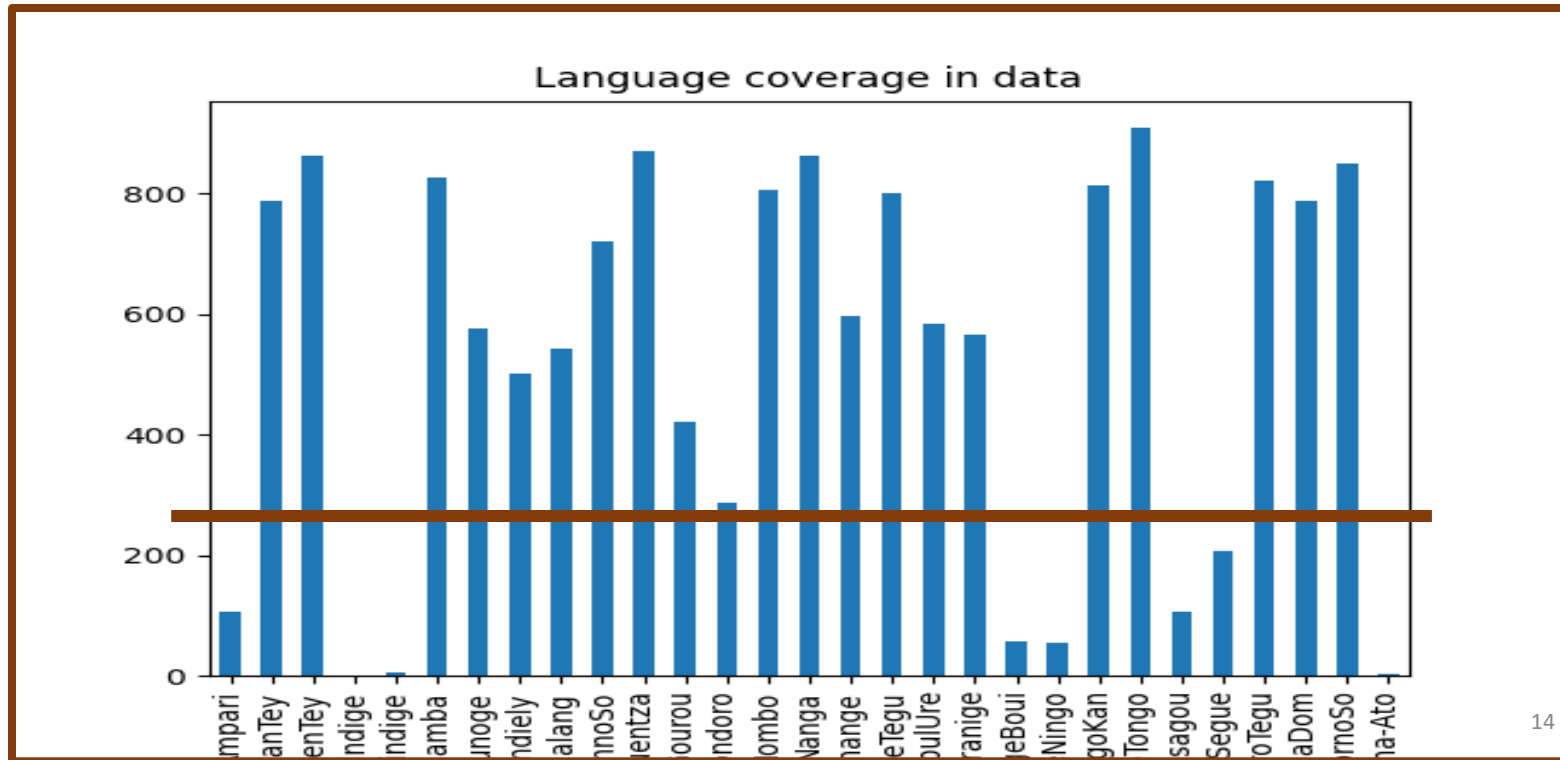
Data in this study: data cleaning steps

- Data after semi-manual process:

1	DOCULECT	GLOSS	IPA
2	BankanTey	domestic animal	dʒáw+dì+m
3	BankanTey	domestic animal	dá.ábà+m
4	BenTey	domestic animal	àrsè.é+m
5	Bunoge	domestic animal	kó+m.bò
6	DogulDomBendi	domestic animal	bélè+g
7	DogulDomBendi	domestic animal	bélè+gù
8	DonnoSo	domestic animal	bèlú
9	DonnoSo	domestic animal	bèlú
10	JamsayGourou	domestic animal	gàr+sè+gé
11	JamsayDouentz	domestic animal	àsè+gé
12	JamsayDouentz	domestic animal	gàsè+gé
13	JamsayMondoro	domestic animal	àsè+ge
14	Mombo	domestic animal	dèbù+bélí
15	Mombo	domestic animal	bélí
16	BonduSoNajamb	domestic animal	dúmé+ŋ.gó
17	BonduSoKindige	domestic animal	dùmó.ó+n
18	Nanga	domestic animal	gàsè+gé
19	Penange	domestic animal	dá.ábá
20	PergeTegu	domestic animal	àsè+gé

Cognate detection: the input

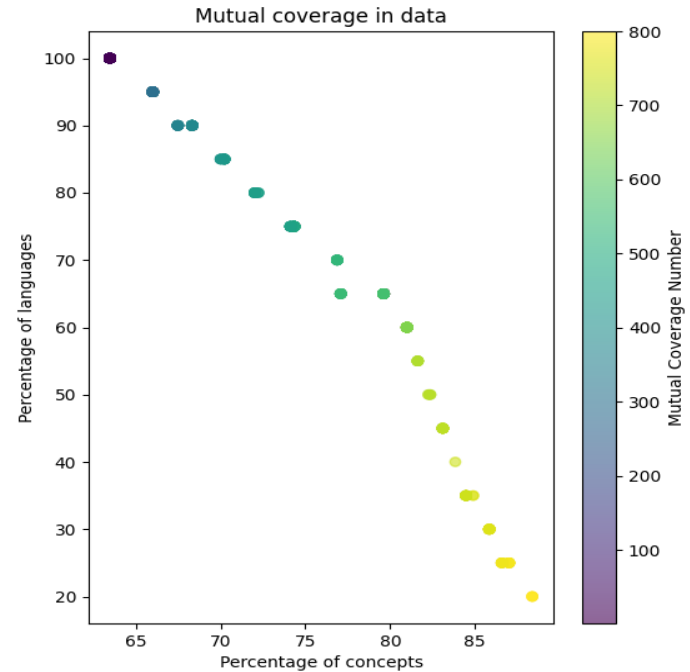
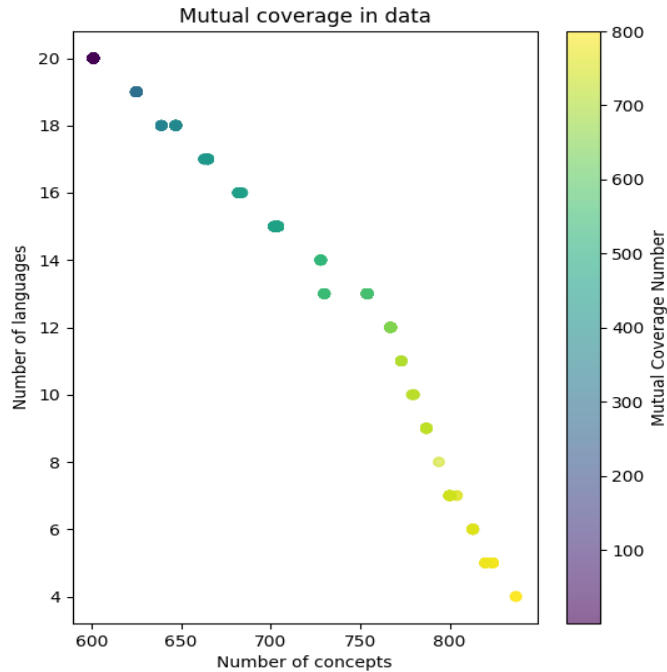
- For cognate detection to be efficient, data must have not just enough words, but also enough comparable word pairs; the higher the better.
- List(2017) suggests a minimum of 100 comparable words i.e. 100 concept pairs.



288 concept pairs

Cognate detection: the input

- Cleaned data had 20 languages with a minimum of 53% concept coverage i.e. 288 concepts.



Cognate detection: the input

- The languages that the cleaned dataset contains are shown below.

language	items	coverage
BankanTey	787	0.83
BenTey	861	0.91
BonduSoNajamba	827	0.87
Bunoge	575	0.61
DogulDomBendiely	501	0.53
DogulDomKundialang	542	0.57
DonnoSo	719	0.76
JamsayDouentza	869	0.92
JamsayGourou	420	0.44
Mombo	805	0.85
Nanga	863	0.91
Penange	596	0.63
PergeTegu	799	0.84
TebulUre	584	0.62
Tiranige	566	0.60
TogoKan	814	0.86
TommoSoTongoTongo	908	0.96
ToroTegu	822	0.87
YandaDom	786	0.83
YornoSo	849	0.90

Cognate detection: the cognate set

- The new data are then fed into the Lexstat model of Lingpy - a model that has been shown to be impressively effective for automated comparative historical linguistics (List et al 2023).
- The Lingpy model outputs a .tsv file that contains the determined cognate sets.
- COGID in the .tsv file is indicative of cognacy.

ID	DOCULECT	CONCEPT	IPA	TOKENS	SONARS	PROSTRINGS	CLASSES	LANGID	NUMBERS	WEIGHTS	DUPLICATES	COGID			
#															
17132	BenTey	(1P1 subject pronominal)		j	j	6	A	J	2	2.J.C	2.0	0	1		
17133	Bunoge	(1P1 subject pronominal)		ŋ	ŋ	4	A	N	4	4.N.C	2.0	0	3		
17134	DogulDomBendiely	(1P1 subject pronominal)		ó	ó	6	ó	7	X	U	5	5.U.V	1.5	0	4
17135	DonnoSo	(1P1 subject pronominal)		ŋ	ŋ	4	A	N	7	7.N.C	2.0	0	3		
17136	DonnoSo	(1P1 subject pronominal)		j.ŋ	j ŋ	6 4	AN	JN	7	7.J.C 7.N.c	2.0 0.8	0	1		
17137	JamsayDouentza	(1P1 subject pronominal)		j	j	6	A	J	8	8.J.C	2.0	0	1		
17138	BonduSoNajamba	(1P1 subject pronominal)		j	j	6	A	J	3	3.J.C	2.0	0	1		
17139	Nanga	(1P1 subject pronominal)		j	j	6	A	J	11	11.J.C	2.0	0	1		
17140	Penange	(1P1 subject pronominal)		ŋ	ŋ	4	A	N	12	12.N.C	2.0	0	3		
17141	PergeTegu	(1P1 subject pronominal)		j	j	6	A	J	13	13.J.C	2.0	0	1		
17142	TebulUre	(1P1 subject pronominal)		j.~	j ~	6 4	AN	J9	14	14.J.C 14.9.c	2.0 0.8	0	1		
17143	Tiranige	(1P1 subject pronominal)		ní	n í	4 7	AX	NI	15	15.N.C 15.I.V	2.0 1.5	0	11		
17144	Tiranige	(1P1 subject pronominal)		nɪ#v	n ì # v	4 7 9 3	AX_A	NI_B	15	15.N.C 15.I.V	15.__. 15.B.C		2.0		
17145	TogoKan	(1P1 subject pronominal)		émé	é m é	7 4 7	XBZ	EME	16	16.E.V 16.M.C 16.E.V	1.5 1.75 0.8	0			
17146	TommoSoTongoTongo	(1P1 subject pronominal)		ú	ú	7	X	Y	17	17.Y.V	1.5	0	13		

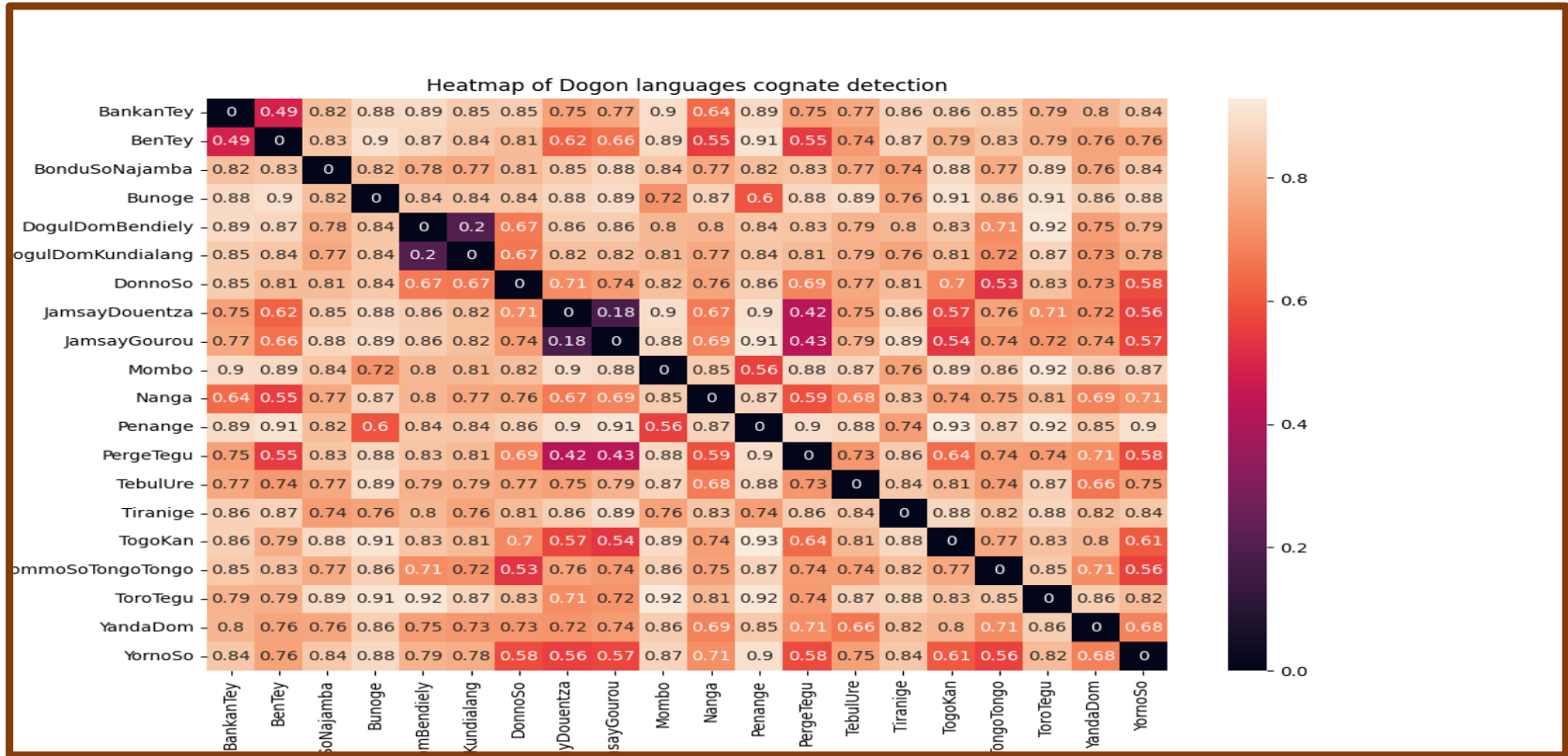
Cognate detection: the cognate set

- Alignments can be obtained using the Alignment class in Lingpy:
- Alignments can then be visualized using Edictor (List et al 2017).

ID	DOCULECT	CONCEPT	IPA	TOKENS	COGID
8954	BankanTey	(man's) sister	pòjâ+m	p ò j â + m	282 ²
8955	BenTey	(man's) sister	pòjà.á	p ò j à á	282 ²
8956	DogulDomBendiely	(man's) sister	sáà	s á à	284 ⁶
8957	DogulDomKundialang	(man's) sister	sà.án̄	s à á ñ	284 ⁶
8958	DonnoSo	(man's) sister	sà.à	s àà	284 ⁶
8959	DonnoSo	(man's) sister	jà.àsà.à	j àà s àà	287 ³
8960	JamsavDouentza	(man's) sister	ièsà.á	i è s à á	287 ³
8962	TebulUre	(man's) sister	sá	s á	284 ⁶
8963	Tiranige	(man's) sister	bà.à	b àà	289
8964	TommoSoTongoTongo	(man's) sister	sáá	s áá	284 ⁶

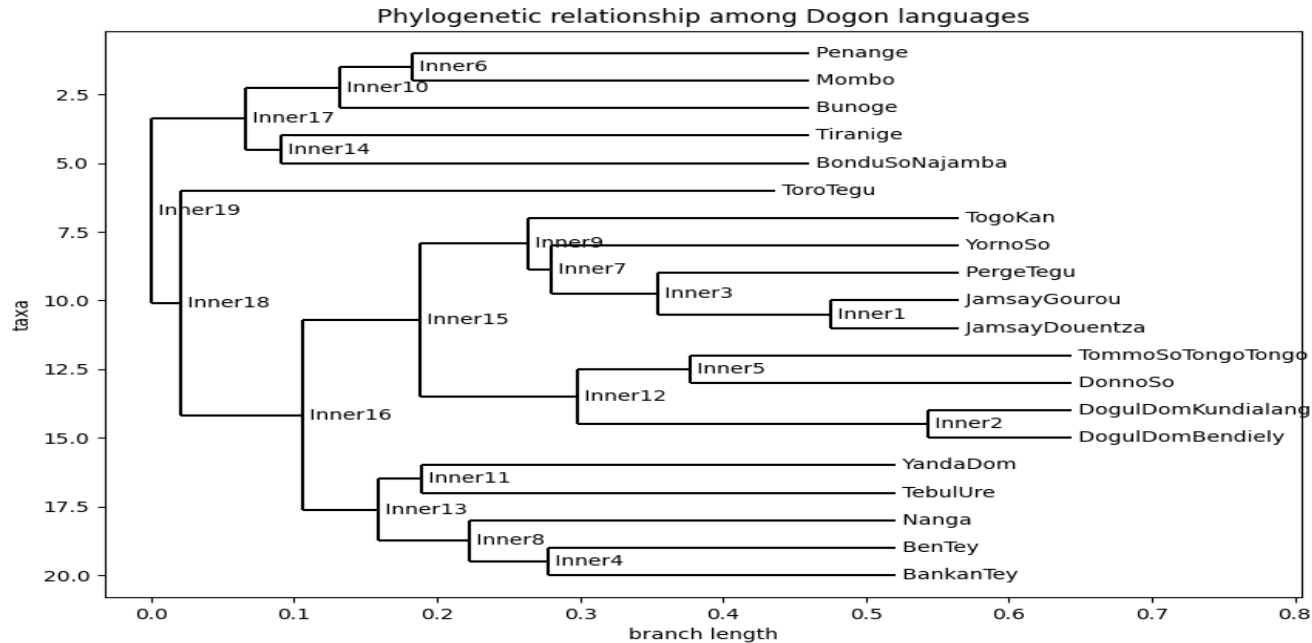
Cognate detection: the cognate set

- Another option is to obtain the distance matrix calculated from the cognate detection algorithm – the lower, the closer.



Cognate detection: the cognate set

- An even more exciting option is to use the distance matrix to obtain a phylogenetic tree (either via clustering algorithms, or the more specialized methods like UPGMA)



Conclusion

- This study has sought to show how raw data, from an African linguistics fieldwork, can be processed for automated computation in historical linguistics.
 - An even more exciting approach involves going through a CLDF workflow
- Three process have been noted as critical:
 - Data formatting: ‘columnization’ and ‘rowing’
 - Data cleaning: deletions and concatenations
 - Data preparation: segmentation and stabilization
- Once data is processed, various algorithms exist to conduct highly specialized modeling that can give critical insights.
- More importantly is the fact that, these methods are applicable to various linguistic landscapes, and is especially useful for the African case, where debates still rage on about some proposed phylogenetic relationships.
- They thus can be used to either confirm these relationships, or call them into question.

Data in this study: the raw data

- Data is from the Dogon and Bangime Linguistics project.
- Data similar to data curated by linguists working on the field.

English	français	core	Toro Tegu (Tc	Ben Tey (Ben	Bankan-Tey (C	Nanga (Anda	Donno So	Jamsay (Doue
domestic anima	animal domestique (surt	1	àrùzàká ~ àrz	àrsě:-m\\àrsě:	jáwdi-m\\jáv	gàsègè	bèlú	àsègè, gàsègè
(sb's) livestock	bétail (l'ensemble des an	3		jáwdù	jáwdi	jáwdi		jáwdi, gàsègè
(young herder's	animal favori (d'un jeun	3	gáfèl	gábèl	gáfèl	gábèri		gápèl
camel	chameau	1	ɲó	ò:ɲùw ^m ò-m	ò:ɲè ^m -m\\ò:ɲè ^m	ògò-ɲòɲò	òɲò-yòɲò	òyò-ɲùɲò
cat	chat	1	niw ^a	ni-niw ^a -m\\n	nùmā-m\\nùm	dòndiyè	gámmà	ni-niw ^a é
cow, bull (any l	bovin (vache ou taureau)	1	nàɲá	nā:-m\\nā:	nā:-m\\nā:	nàɲá	nàɲá	nàɲá
calf	veau	2	nàɲà-[ir ^a -í:]	nā: yí-m	nā:-yí-m\\nā:-	nàɲá yí:	nàɲà-í:	nàɲ-í: ^a
cow (female)	vache	2	nàɲà-yá	nā: yā-m		nàɲá yā:		nàɲà-nā:
cow that has ca	vache qui a mis bas au n	3	nàɲà-yà-nā:	nā: nā:-m	nā:-yā-m\\nā:-	nàɲá dé:		
cow (with a cal	vache (avec un veau)	3		nā: yàyò-m		nàɲá yí: dè:		nàɲà-nā:
heifer, young c	génisse, jeune vache (qu	2	nàɲà-yá girèy	nā: gürò-m	wí:gè-m\\wí:g	nàɲá gürá, nā:	nàɲà-gulò	nàɲá güró
bull	taureau	2	nàɲà-àrá	nā: ár ^a -m		nàɲá ár ^a	nàɲá àná	nàɲá ár ^a
bull (castrated)	taureau (castré)	2	nàɲà-àrà pólù	nā: pòrú-m	nā:-ár ^a -m\\nā:	nàɲá pòrí	nàɲá pòr-ú	nàɲá pòrú
bull (uncastrate	taureau (non castré)	2	nàɲà-àrà bòbò	nā: jàɲgú-m	nā:-jǐ-m\\nā:-j	nàɲá jàɲí	nàɲá kálàndù	nàɲá jàɲ
bull (uncastrate	taureau (non castré, jeun	3		nā: já: kírím		nàɲá kóróbóri		nàɲá kóróbó:r
dog	chien	1	nènú	injě-m\\injé	ɲjèrú-m\\ɲjèrú	nèr ^a í	idú	ijú

Thank you

All code used for this presentation is publicly available on Github at:
<https://github.com/PromiseDodzi/UCAD-presentation>

Bibliography

- Atkinson, Q. D., & Gray, R. D. (2006). How old is the Indo-European language family? Illumination or more moths to the flame. In *Phylogenetic methods and the prehistory of languages* (pp. 91-109).
- Blum, F., & List, J.-M. (2023). Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Computational Typology and Multilingual NLP* (pp. 52-64). Association for Computational Linguistics.
- Creissels, D. (2019). Morphology in Niger-Congo languages. In *Oxford Research Encyclopedia of Linguistics*.
- Dimmendaal, G. J. (2008). Language ecology and linguistic diversity on the African continent. *Language and Linguistics Compass*, 2(5), 840-858.
- Forkel, R., Moran, S., List, J. M., Greenhill, S. J., Ashby, L. C., Gorman, K., & Kaiping, G. (2019). Segments: Unicode standard tokenization routines and orthography profile segmentation [Software Library, Version 2.1.3].
- Forkel, R., List, J. M., Greenhill, S. J., Rzymiski, C., Bank, S., Cysouw, M., ... & Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1), 1-10.
- Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965), 435-439.
- Hantgan-Sonko, A. (2019). Linguistic support for an early Dogon diffusion. Poster presented at the *Peopling History of Africa conference*, Geneva, Switzerland.
- Hantgan-Sonko, A., & List, J. M. (2022). Bangime: Secret language, language isolate, or language island? A computer-assisted case study. *Papers in Historical Phonology*, 7, 1-43.
- Heath, J. (2017). *A grammar of Bunoge (Dogon, Mali)*.
- Heath, J. (2017). *A grammar of Tebul Ure (Dogon, Mali)*.
- Heath, J. (2023). *A grammar of Tebul Ure (Dogon, Mali)*.
- Hochstetler, J. L., Durieux, J. A., & Durieux-Boon, E. I. (2004). *Sociolinguistic survey of the Dogon language area*. SIL International.
- Kim, Y. M., Chang, K., Cui, C., & Mortensen, D. (2023). Transformed protoform reconstruction. *arXiv preprint arXiv:2307.01896*.
- List, J. M. (2012). Multiple sequence alignment in historical linguistics. In *Proceedings of ConSOLE* (Vol. 19, pp. 241-260).
- List, J. M. (2017). Historical language comparison with LingPy and EDICTOR. <https://doi.org/10.5281/zenodo.1042205>.
- List, J. M., Forkel, R., & Hill, N. W. (2022). A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. *arXiv preprint arXiv:2204.04619*.
- List, J. M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2), 130-144.
- List, J.-M. (2012a). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources* (pp. 117-125). Stroudsburg.
- List, J.-M. (2012b). SCA: Phonetic alignment based on sound classes. In M. Slavković & D. Lassiter (Eds.), *New directions in logic, language, and computation* (pp. 32-51). Berlin and Heidelberg.
- List, J.-M., Greenhill, S., & Forkel, R. (2023). *LingPy Documentation Release 2.6*.
- List, J.-M., Walworth, M., Greenhill, S. J., & Tresoldi, T. (2017). The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1), 1-18.
- Moran, S., & Prokić, J. (2013). Investigating the relatedness of the endangered Dogon languages. *Literary and Linguistic Computing*, 28(4), 676-691.
- Moran, S., Forkel, R., & Heath, J. (2016). *Dogon and Bangime linguistics*. Jena: Max Planck Institute for the Science of Human History.
- Rama, T., List, J. M., Wahle, J., & Jäger, G. (2018). Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *arXiv preprint arXiv:1804.05416*.

