A Proposal for Data Interface Formats for Cross-Linguistic Data

Harald Hammarström

30 Apr 2015, Leipzig

Desideratum: Cross-Linguistic Data Format

- For us who do *Diversity Linguistics* with the aid of *Databases*
- It would be handy with a **C**ross-**L**inguistic **D**ata **F**ormat (CLDF)
- Desirable properties
  - Capture the bulk of the kinds of data figuring in Diversity Linguistics
  - Be editable by hand and readable by a computer
  - Simple to concatenate/merge/aggregate
  - Fit in one single file
  - ...

A CLDF Proposal

- tab-separated (.tsv) with rows x columns
- One row = one datapoint
- Columns

Col	Name	Contents	
0	Row	Row/datapoint id	Optional
1	Language	Language id	Mandatory, non-empty
2	Feature	Feature id_1-Feature id_n	
3	Value	Value	
4	Source	Source	Mandatory, possibly empty
5	Comment	Comment	
6	Example	Example	Optional
...	...	...	
...	...	...	

- With header line at top
- Columns have a conventional order as above, but are ID:d by name
- UTF-8

Column 0: Row ID

- Present if and only if there's a column named `Row` in the header, e.g.

Row	Language	Feature	Value
DP1995_01	swe	Adjective-Noun Order	1
DP1996_07	eng	Adjective-Noun Order	1
...	...	...	...

- Otherwise Row implicitly read as the row-numbers of the file
- Thus the below two are equivalent

Ex 1

Row	Language	Feature	Value
1	swe	F1	A
2	eng	F1	B
...	...	...	...

Ex 2

Language	Feature	Value
swe	F1	A
eng	F1	B
...	...	...

Column 1: Language

- All cross-lingistic data is about languages, so this field is mandatory
- Use whatever id:s you like, e.g., `swe`, `lg58`, `Dialect A`, `idiolect Z`, `experient20150801-5`, ...
- BUT, if you use standard identifiers (iso-639-3 or glotto_id:s) the application is licensed to interpret them as such
- Two ways to invoke
  - Plain id: `swe`, `swed1254`, ...
  - Bracketed id: `Swedish [swe]`, `Swedish [swed1254]`, ...
- The latter way allows a CLDF-data language to make more distinctions than the standard identifiers while making use of information derivable from the standard identifiers, e.g.
  - `Swedish A [swe]` and `Swedish B [swe]` would be two distinct language_id:s in a CLDF-file
  - but the application reading it is licensed to interpret both as varieties of iso-639-3 [swe], thus possibly taking metadata such a macro-area, classification, etc relating to iso-639-3 `swe`

Column 2: Feature

- Feature_id:s can be anything, e.g., `156275`, `http://wals.info/feature/2A`, `What is the order of Numeral and Noun`, ...
- As a lot of linguistic data is multidimensional, CLDF must be able to encompass it
- Thus, Feature(_id:)s can be composite but are necessarily baked into **one** column
- Use special symbol sequence $\sim$ to indicate the break-down, e.g.

Language	Feature	Value
deu	1$\sim$ SG	ich
deu	2$\sim$ SG	du
deu	1$\sim$ PL	wir
...		

- Feature metadata: see later slides

Column 3: Value

- The value can be any kind of data: `string`, `int`, `float` etc
- Standard database convention of not having alternatives or multiple values encoded in the same cell

Wrong:

Language	Feature	Value
deu	F1	A/B
deu	F2	{B, C}

Right:

Language	Feature	Value
deu	F1	A
deu	F1	B
deu	F2	B
deu	F2	C

- Three conventional ways to indicate missing data
  - `<blank>` or `-`: Not known (because never checked)
  - `?`: Not known (in the source indicated)
  - `N/A` or `n/a`: Not applicable (this language can never have data on the feature in question)
- Curators of data often want to distinguish these kinds of missing data, so and its saves a frequent translation if they are understood by CLDF
- No CLDF way to encode uncertainty in values, e.g., notation like 1? (for "probably value 1") **not** supported

Column 4: Source

- As free text

Language	Feature	Value	Source
tmy	F1	A	Bamler, Georg. (1900) Bemerkungen zur Grammatik der Tamisprache. Zeitschrift für afrikanische und orientalische Sprachen V. 198-253.
tmy	F2	A	Experiment in Zambia 2015-10-01

- With Glottolog ID

deu | F1 | A | `http://glottolog.org/resource/reference/id/318814`

- With local bibtex key as per supplied bibtex file (see later slides)

deu | F1 | A | meier2015

- Multiple sources semicolon separated

deu | F1 | A | meier2015; forkel2010

- Source context (typically page numbers in square brackets)

deu | F1 | A | meier2015[2]; fokel2010[ix-x, 2-3, 7]

CLDF Special Characters used for parsing

- $< TAB >$ and $< CR >$ inherited from .tsv
- $\sim$ to break composite features
- square brackets [] in the language id field
- ; and square brackets [] in the source field

That's all, unless we allow inlining of metadata/bibtex (see later slides)

Feature Metadata/BibTex Libraries/Example Specifications

- Features in linguistics typically have a lot of structured associated metadata that cannot or should not be packed into the Feature ID
  - Text explanation of the meaning the of the feature(s)
  - Specification/explanation of legal values (to human or computer)
  - Who designed the feature and when etc.
  - Links to related materials
  - ...
- Sources are typically specified in standalone structured bibliographic databases
- Linguistic examples typically have formatting and structured conventions such as IGT

  Supply as separate associated files, or bake into CLDF somehow?

Metadata Inlining or Separate Files?

- Forkel solution:
  - Have three extra files with a recognizable name convention. If the dataset is called `harald.cldf`
    - ★ `harald.cldf-metadata.json`: For feature metadata
    - ★ `harald.bib`: For bibliographical data
    - ★ `harald.igt.cldf`: For formatted examples
  - Preserves modulization!
- Other possible (complementary solution)
  - Allow inlining of a dictionary (key-value pairs) with some special formatting, e.g.

Language	Feature	Value
deu	1{'creator': "Harald", "Legal values": [1, 2, 3], ...}~SG{'creator': "Harald", "Legal values": [SG, DU, PL], ...}	ich
deu	2{'creator': "Harald", "Legal values": [1, 2, 3], ...}~SG{'creator': "Harald", "Legal values": [SG, DU, PL], ...}	du

  - Violates principles of modulization and duplication
  - Possibility to bake everything into **one** CLDF-file

Example 1: WALS in CLDF

Language	Feature	Value
alo	103A. Third Person Zero of Verbal Person Marking	2
alo	16A. Weight Factors in Weight-Sensitive Stress Systems	1
alo	14A. Fixed Stress Locations	6
alo	104A. Order of Person Markers on the Verb	2
alo	48A. Person Marking on Adpositions	3
alo	9A. The Velar Nasal	2
alo	17A. Rhythm Types	3
alo	15A. Weight-Sensitive Stress	8
alo	102A. Verbal Person Marking	5
alo	100A. Alignment of Verbal Person Marking	4
nwi	16A. Weight Factors in Weight-Sensitive Stress Systems	4
nwi	14A. Fixed Stress Locations	1
nwi	17A. Rhythm Types	1
nwi	15A. Weight-Sensitive Stress	3

Example 2: Paradigms in CLDF

- Suppose your language features are paradigms (= matrices)

L1

P/N	1	2	3
Sg	ich	du	er/es/sie
Pl	wir	ihr	sie

L2

P/N	1	2	3
Sg	na	a	mene
Pl	ni	i	mene

L3 ...

- Represent each cell in the matrix as one CLDF datapoint/row: Language x Row Col x Cell Value

Language	Person~ Number	Value
deu	1~ SG	ich
deu	2~ SG	du
deu	1~ PL	wir
moq	1~ PL	ni
...	...	...

Example 3: Constructions in CLDF #1

Blad 1: *Emerillon*

DP	*question*	*key*	*C1*	*C1*	*C2*
			CltzSub	CltzSub	PrefSub
			score	remarks	score
1	Can subject-verb agreement be expressed independently on the subordinate EDU?	Y/N	Y		Y
1.1	Are the same forms used as in the main clause?	Y/N	Y		Y
2	Can object-verb agreement be expressed independently for the subordinate EDU?	Y/N	Y		Y
2.1	Are the same forms used as in the main clause?	Y/N	Y		Y
3	Can tense categories be expressed independently for the subordinate EDU?	Y/N	?		?
3.1	Are the same forms used as in the main clause?	Y/N	?		?
4	Can aspectual categories be expressed independently for the subordinate EDU?	Y/N	Y		Y
4.1	Are the same forms used as in the main clause?	Y/N	Y		Y
5	Can event-modal categories be expressed independently for the subordinate EDU?	Y/N	Y		Y
5.1	Are the same forms used as in the main clause?	Y/N	Y		Y
6	Can epistemic and/or evidential categories be expressed independently for the subordinate EDU?	Y/N	N/A		N/A

Example 3: Constructions in CLDF #2

- Suppose a language consists of a set of constructions: $L = \{C_1, C_2, \ldots, C_n\}$
- Suppose a construction is defined in terms of features of form and features of meaning

	form				*meaning*			
Construction	$[f_1$	f_2	$\ldots$	$f_n]$	$[m_1$	m_2	$\ldots$	$m_k]$
$C_1 =$	1	0	1	A	0	0	1	1
$C_2 =$	1	0	1	A	0	0	1	1
$\ldots$					$\ldots$			

- Represent each datapoint definition a construction as one CLDF datapoint/row: Language x Construction Feature x Cell Value

Language	Person$\sim$ Number	Value
eme	$C_1 \sim m_1$	1
eme	$C_1 \sim m_2$	0
eme	$C_1 \sim f_2$	0
eme	$C_2 \sim f_2$	0
$\ldots$		

Generic Applications of CLDF data

- No matter exactly what the data is, we know the CLDF represents *cross-linguistic* data
- This means fairly generic applications can be built which take CLDF-data and produce some appreciable outcome
- This has good inspirational and pedagogical value (but probably not sufficient for a research paper)
- Some examples to follow

Drawing Isogloss Lines: CLDF Input

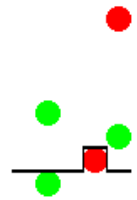
- Input: A CLDF-file with a binary feature and language id:s that maps to iso-639-3 or glottocodes

Language	Feature	Value
Imbabura Highland Quichua [qvi]	Numeral system	NON-RESTRICTED
Pataxó Hã-Ha-Hãe [pth]	Numeral system	?
Pai Tavytera [pta]	Numeral system	RESTRICTED
Salamãi [mnd]	Numeral system	RESTRICTED
San Martín Quechua [qvs]	Numeral system	NON-RESTRICTED
Tawandê [xtw]	Numeral system	RESTRICTED
Waorani [auc]	Numeral system	RESTRICTED
Lakondê [lkd]	Numeral system	RESTRICTED
Caribbean Javanese [jvn]	Numeral system	NON-RESTRICTED
Aurê y Aurá [aux]	Numeral system	RESTRICTED
Mamaindé [wmd]	Numeral system	RESTRICTED
Canichana [caz]	Numeral system	RESTRICTED
Qawasqar [alc]	Numeral system	RESTRICTED
Makuráp [mpu]	Numeral system	RESTRICTED
Matís [mpq]	Numeral system	RESTRICTED
Jaqaru [jqr]	Numeral system	NON-RESTRICTED
Umotína [umo]	Numeral system	RESTRICTED
...	...	...

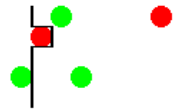
Drawing Isogloss Lines: Definition

- Put a dot coloured by the feature value on a 2D map using coordinates implied by the iso/glotto-codes
- Draw the line which
  - is non-circular, i.e.
    - ★ Runs from the west end to the east end on the map, crossing each column at exactly once OR
    - ★ Runs from the north end to the south end on the map, crossing each row at exactly once
  - and proportion-optimal, i.e., maximizes the proportion of correctly classified points to the total number of points, on both sides of the line

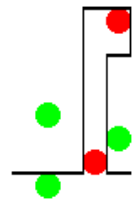
- Legal



- Legal

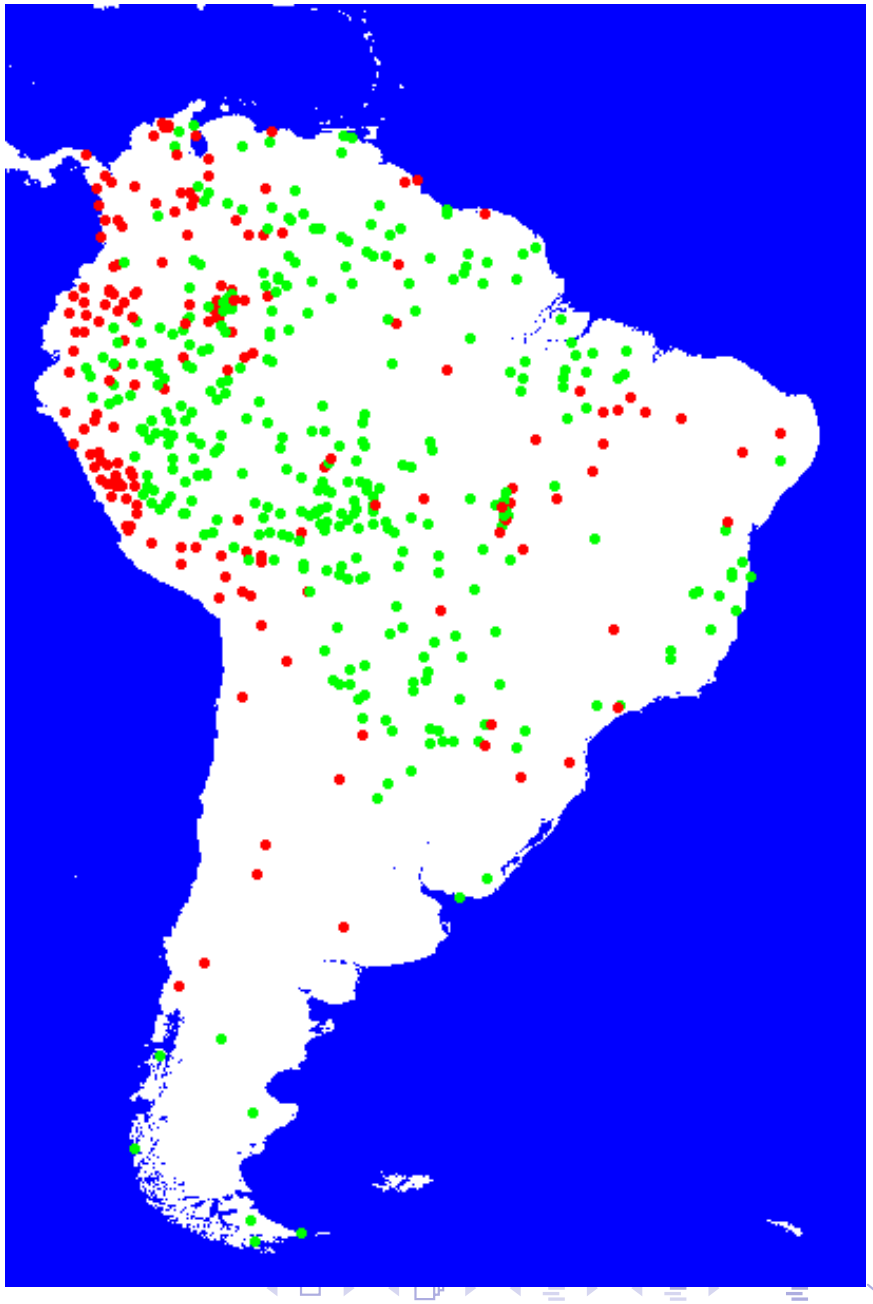


- NOT Legal

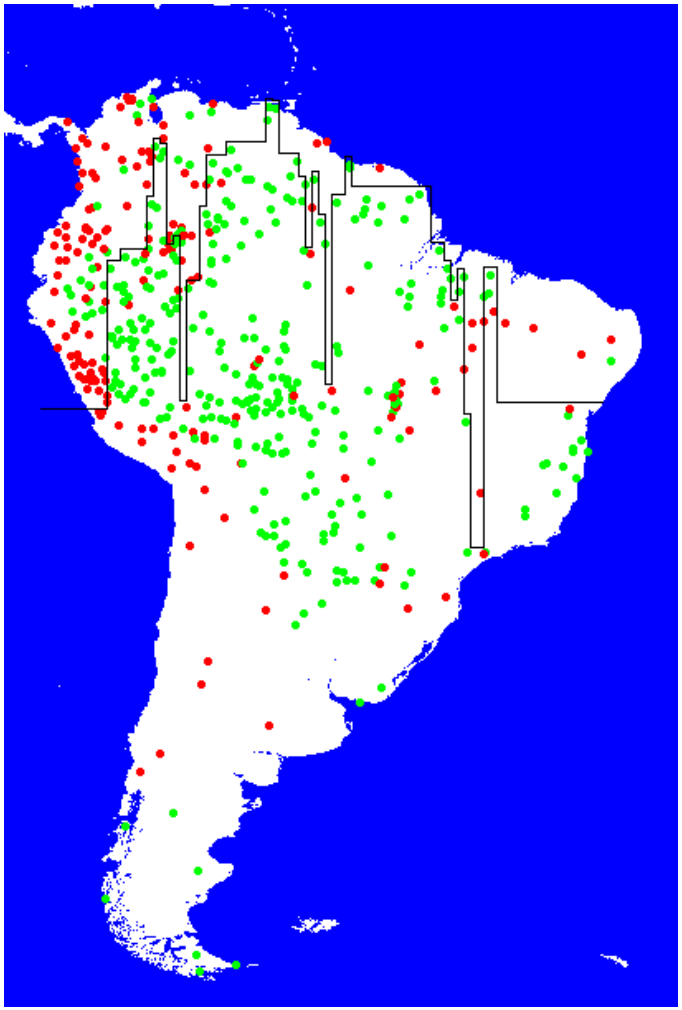


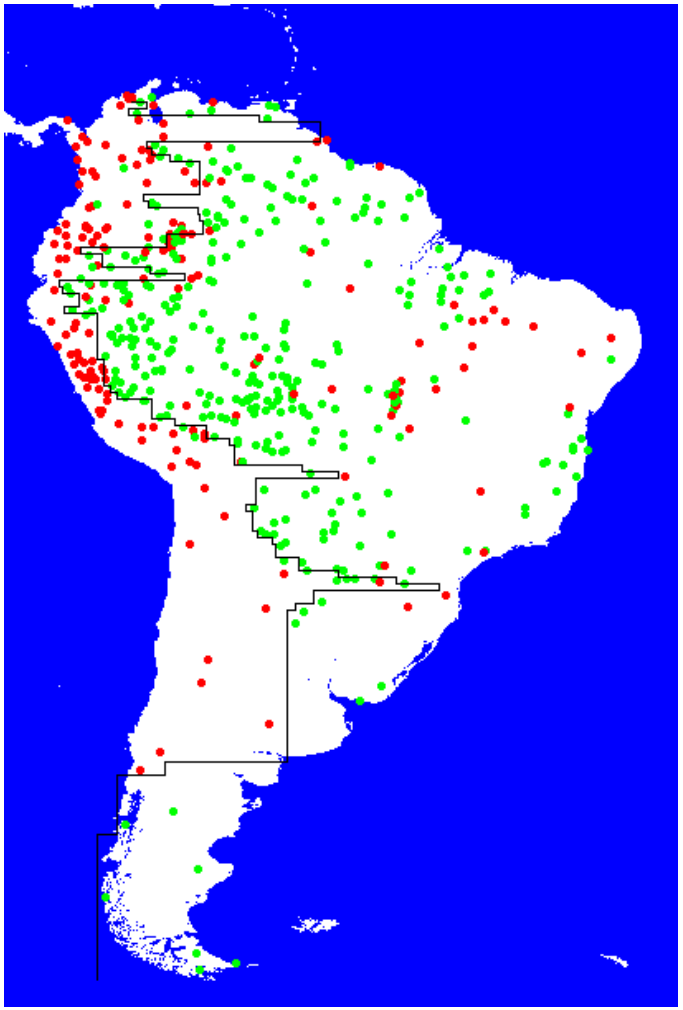
Drawing Isogloss Lines: Full Input

- A numeral system is "restricted" iff
  - Monomorphemic numerals exist only up to 2 or 3 AND
  - Higher quantities are expressed orally only inexactly, or up to ca 10 with additions of 1, 2 and 3 (possibly including ad hoc use of 'hand' for 5).
- Suppose we want to know if the border of restricted numeral system coincides with the extent of the Amazon forest



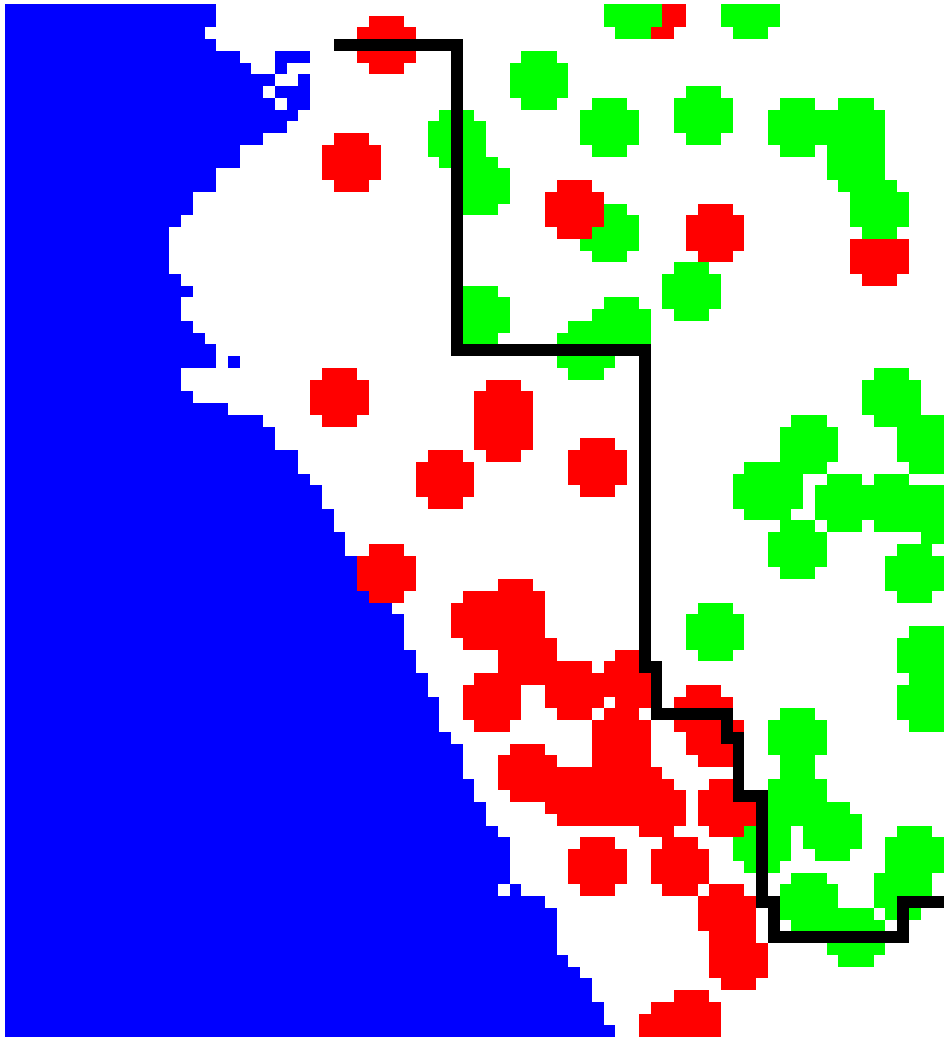
Optimal Isogloss Lines: Outcome





- The optimal line east-west has error 513.125 but
- the optimal line north-south only has 241.875

Zoom on Andean-Amazon Divide



- The isogloss lines for the South American numeral data gives a fairly consistent Andean-Amazonian boundary
- But includes the Chaco and Southern Cone regions of South America in "Amazonian" part
- The border may be studies more closely for non-linguistic correlates to the boundary line

Parsimony Reconstruct: CLDF Input

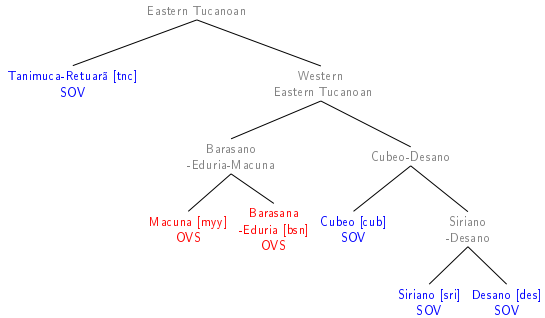
- Input: A CLDF-file with a discrete feature and language id:s that maps to iso-639-3 or glottocodes

Language	Feature	Value
Cuyamecalco Mixtec [xtu]	Word Order	?
Tacahua-Yolotepec Mixtec [xtt]	Word Order	?
Sindihui Mixtec [xts]	Word Order	?
Tumshuqese [xtq]	Word Order	SOV
San Miguel Piedras Mixtec [xtp]	Word Order	?
Yoloxochitl Mixtec [xty]	Word Order	VSO
Transalpine Gaulish [xtg]	Word Order	SVO
Ketengban [xte]	Word Order	SOV
Diuxi-Tilantongo Mixtec [xtd]	Word Order	VSO
San Juan Teita Mixtec [xtj]	Word Order	VSO
Sinicahua Mixtec [xti]	Word Order	?
Warji [wji]	Word Order	?
Waja [wja]	Word Order	SVO
Toto [txo]	Word Order	SOV

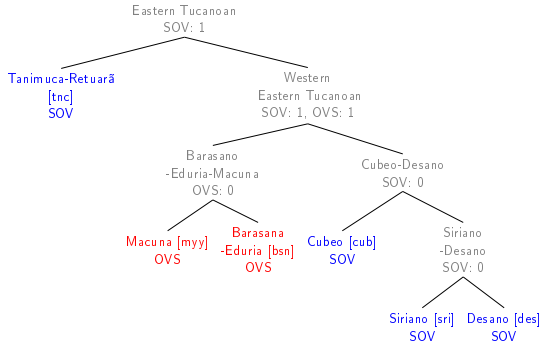
Parsimony Reconstruct: Procedure & Output

To each internal node, reconstruct the value that **minimizes** *the* **total number of changes** *required*

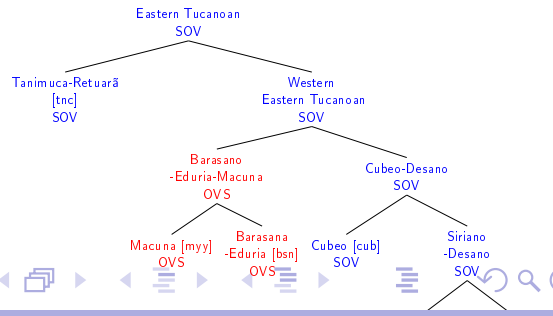
1. Input (a tree and values at the leaves)



2. For each internal node, starting near the leaves, calculate the minimum number of changes required below it for each possible reconstructed value



3. Reconstruct that which yield the minimum total number of changes in the tree



From Parsimony to Stability

- With the (parsimony-) tree reconstruction, you have vertical transitions from node to node

From	To	From	To
Tucanoan	Eastern Tucanoan	SOV	SOV
Tucanoan	Coreguaje-Siona	SOV	SOV
Coreguaje-Siona	Siona-Secoya	SOV	SOV
Coreguaje-Siona	Koreguaje [coe]	SOV	VSO
Western Eastern Tucanoan	Cubeo-Desano	SOV	SOV
Siriano-Desano	Desano [des]	SOV	SOV
Siona-Secoya	Siona-Tetete [snn]	SOV	SOV
Siona-Secoya	Secoya [sey]	SOV	SOV
...	...	...	...

- Counting the number of changes on such transitions gives (an upper bound) estimate of the historical stability of the feature in question

From	To	#
SOV	SOV	19
OVS	OVS	2
SOV	VSO	1
SOV	OVS	1

Stability Rankings

	Feature	p-stability	p-stability n lgs	v-p-stability
NTS100	Is number ever marked separately from person on the verb?	0.87	246	0: 0.92 (240.9), 1: 0.79 (40.9)
NTS101	Are person, number and any TAM category (i.e. 3 or more categories in all) marked by portmanteau morphemes on verbs?	0.87	258	0: 0.92 (225.2), 1: 0.81 (66.6)
NTS102	Are categories such as person, number, gender related to a single participant discontinuously marked on a verb?	0.93	253	0: 0.97 (277.2), 1: 0.73 (15.2)
NTS103	Are benefactive nominals marked on the verb?	0.87	263	0: 0.91 (226.9), 1: 0.84 (76.9)
NTS104	Can instruments be marked on the verb?	0.91	224	0: 0.95 (240.7), 1: 0.73 (22.0)
NTS105	Can recipients be treated as a transitive object, i.e. as Direct Object?	0.84	240	0: 0.84 (81.8), 1: 0.88 (180.4)
NTS106	Are there syntactically ditransitive verbs?	0.87	245	0: 0.86 (96.6), 1: 0.91 (173.6)
NTS107	Is negation marked morphologically on the verbs?	0.83	287	0: 0.87 (223.3), 1: 0.85 (102.7)
NTS108	Can locative or direction be morphologically marked on the verb?	0.83	231	0: 0.88 (188.9), 1: 0.83 (64.2)

Conclusions

- CLDF-format proposal
  - Nothing magic
  - A simple convention to exchange cross-linguistic data
  - Need your input on a few design issues, let's then agree
- Some examples of generic applications to CLDF-data
  - There are more, e.g., language distance matrix/MDS, infer genealogical tree, correlation, ...
  - Let's make a whole battery of them along with CLDF (the format)
- Thanks to Robert Forkel for many items in this CLDF proposal