Dependency-Sensitive Typological Distance

Workshop on Comparing Approaches to
Measuring Linguistic Difference

24-25 October 2011 – University of Gothenburg

Harald Hammarström & Loretta O'Connor

Radboud Universiteit Nijmegen

European Research Council

Linguistic data as matrix of rows and columns

(often languages x features)

	F1	F2	F3	F4	F5	…
Language 1	0	1	0	0	1	
Language 2	0	0	1	1	1	
Language 3	a	b	c	b	a	
Language 4	a	a	a	b	b	
…						

want to calculate the distance between languages
as similarities and differences between features

Typical measure of distance is the GOWER COEFFICIENT

(aka relative Hamming distance)

$$G(L_X, L_y) = \frac{\#_{i \in DEF(L_X, L_Y)} L_x[i] \neq L_Y[i]}{|DEF(L_X, L_Y)|}$$

- counts the **number of features** where the languages have a **different value**

- divides this sum by the **total number of features compared**

- works well if features are **independent** and **of equal weight**
  - true for some types of lexical data
  - **not true for most sets of typological features**

Dependencies in (matrices of) linguistic data

Perspectives in the linguistic literature

- verb-final languages tend to have post-positions
- languages are unlikely to have both strict word order AND case marking

- variety of motivations for dependencies in linguistic approaches

innate and universal properties	→	probabilistic
due to cognitive constraints	→	discourse-functional
family-specific	→	socially-shaped

(Greenberg 1963, Chomsky 1981, Dryer 1992, 2007, Dunn et al 2011)

Computational approaches

- look globally for any correlations between any features

- dependencies must be universal (common to all natural languages)
- they must concern the whole feature (and not just a specific value)
- universal dependencies exist iff no areal or genealogical explanation

OUTLINE of the TALK

1. **describe two dependency-sensitive metrics of linguistic distance**

   - one that addresses **dependencies among features** and
     eliminates these from a standard distance measure

   - another that addresses the **predictability vs. quirkiness**
     of which features are shared

2. **apply each metric to dataset of linguistic features**

3. **evaluate results and illustrate changes** in similarity groupings
   using a distance-based phylogenetic method

DATASET (adapted from Constenla 1991)

81 linguistic features

- 42 morphosyntactic
- 39 phonological

binary coding
only 2 cells missing

35 languages of the Chibcha Sphere (Central and South America)

- 1 Mayan
- 4 Misumalpan
- 15 Chibchan
- 3 Chocoan
- 4 Barbacoan
- 1 Paesan
- 1 Arawakan
- 1 Quechuan
- 1 Xincan
- 1 Lencan
- 3 isolates (Jicaque, Cofán, Camsá)

CALCULATION 1: Gd, a dependency-sensitive measure of linguistic distance

If a feature can be (partly) predicted by another,
then the predictable feature should be (partly) discounted

--> tackles features as a whole, not specific values of features

1. Find feature implications by calculating entropy distribution

2. Distill feature implications by computing a Chu-Liu Tree

3. Modify the Gower coefficient with dependency-sensitive weights

Find feature implications (1)

quantify a predictive relationship by calculating
how much of the **entropy** of one variable
can be predicted from knowing the other
(technique used by Bickel 2010, Daume & Campbell 2007)

entropy = the measure of uncertainty associated with a random variable

$$A \rightarrow B = \frac{MI(A,B)}{H(A)}$$

MI(A,B) = H(A) + H(B) – H(A,B)
= mutual info of A and B

H(A) = Shannon entropy of A

Find feature implications (2)

Rank	Implication	Strength
1	$13 \rightarrow 12$	1.000
649	$39 \rightarrow 67$	0.180
1297	$77 \rightarrow 71$	0.113
1945	$37 \rightarrow 6$	0.079
2593	$50 \rightarrow 19$	0.055
3241	$14 \rightarrow 27$	0.037
3889	$54 \rightarrow 45$	0.026
4537	$38 \rightarrow 29$	0.015
5185	$10 \rightarrow 47$	0.005
5833	$28 \rightarrow 42$	0.000

Some sample feature implications from the Chibcha Sphere dataset

Distill feature implications (1)

Implication set will include **redundancy**

- A → B, B → A
- A → B, B → C, A→ C

Solution:
Keep only the strongest implications in the chains
creating a transparent similarity matrix
with a maximum of ONE predictor for a feature

Distill feature implications (2)

the CHU-LIU algorithm (Chu & Lin 1965)

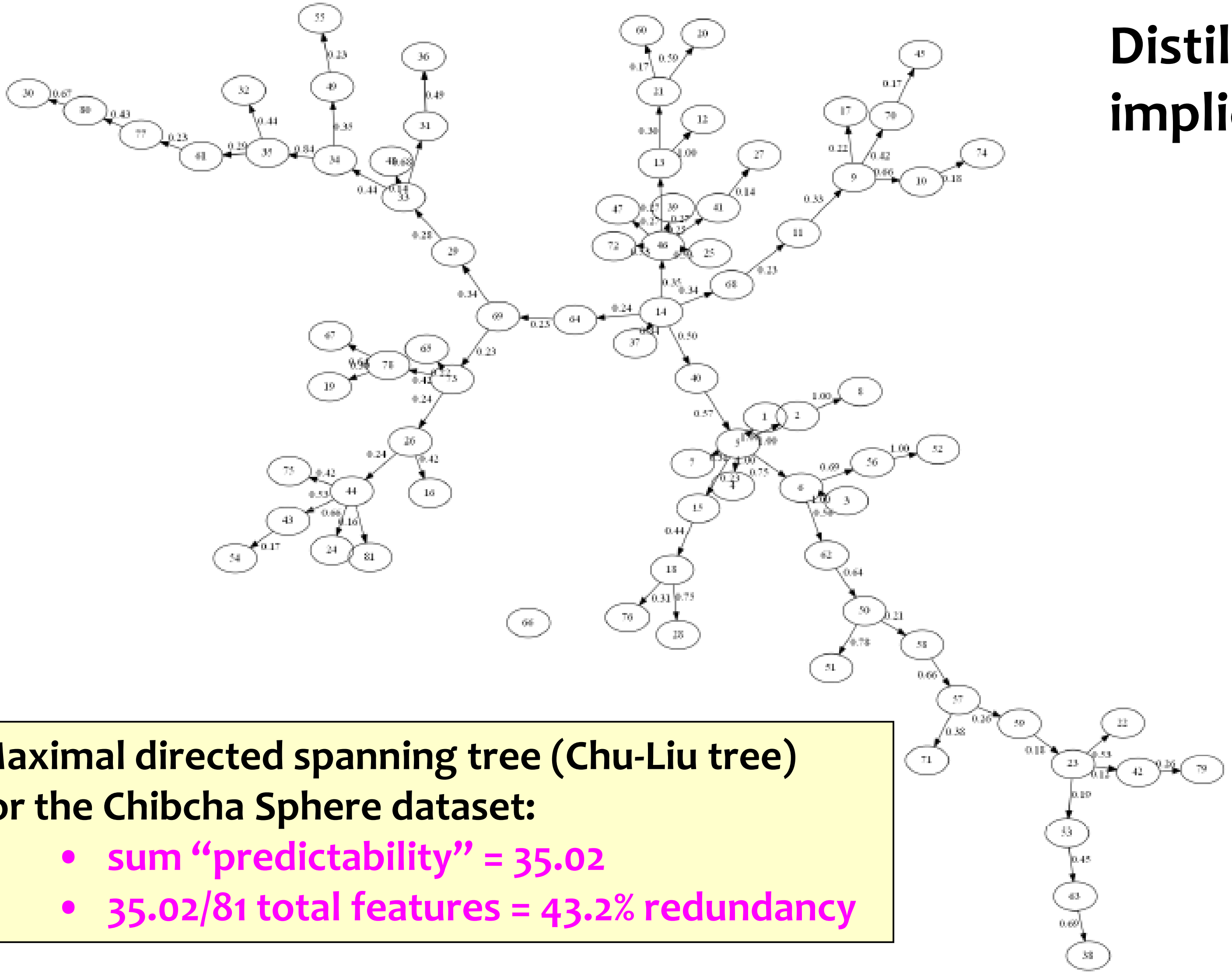
starts with **one node per predicted feature** and
one edge per implication (could be many per node)

processes the set of feature implications
eliminating all but strongest implication

ends with **one node per predicted feature** and
one incoming edge per node
with strongest implication value

→ **can now develop a dependency-sensitive metric**

Distill feature implications (3)



Maximal directed spanning tree (Chu-Liu tree) for the Chibcha Sphere dataset:

- sum “predictability” = 35.02
- 35.02/81 total features = 43.2% redundancy

A dependency-sensitive distance measure Gd (modified Gower)

$$G_d(X,Y) = \frac{\sum_{i \in DEF(L_X, L_Y) \text{ and } L_X[i] \neq L_Y[i]} 1.0 - W(i)}{\sum_{i \in DEF(L_X, L_Y)} 1.0 - W(i)}$$

- W(i) is the weight of the incoming edge that predicts F(i)
  and is 0.0 if there is no such edge
- As in the Gower coefficient, only features which are in both languages
- For each feature, instead of a penalty of 1 for mismatches
  the penalty is the appropriate amount
  reflecting how predictable the feature in question is

CALCULATION 2: Gq, a dependency-sensitive metric that incorporates quirkiness

If languages share unpredictable features
or fail to share predictable features
then these languages are more likely to share a common history

--> tackles specific values of (constellations of) features

Present study:

- **unary quirks**, feature value constellations involving one variable
- **binary quirks**, feature value constellations involving two variables

A dependency-sensitive metric Gq (modified Gower)

1. **Define the quirkiness Q of a feature (or here, of a set of 2 features):**

$$Q(f_i = u, f_j = v) = \frac{\text{The number of languages with values } f_i = u \text{ and } f_j = v}{\text{Total number of languages with } f_i \text{ and } f_j \text{ defined}}$$

2. **Modify Gower coefficient with measure of feature quirkiness Q:**

$$G_q^2(X,Y) = 1.0 - \frac{\sum_{i<j \in DEF(L_X, L_Y) \text{ and } L_x[i]=L_Y[i]} 1.0 - Q(i = L_X[i], j = L_X[j])}{|DEF(L_X, L_Y)|}$$

Feature dependencies and quirky values

Procedure:

1. Enumerate potential unary and binary quirks in the dataset

2. For each pair of languages
   score their matches proportionately to their quirkiness
   with the modified Gower coefficient metrics Gq1 and Gq2

NB: Gq not strictly a distance measure as Gq (X,X) is not necessarily 0

Experimental results: G distances vs. Gd distances (1)

	with dependencies		without dependencies	
Rank	G		G_d	
1	Ulua-Sumo	0.00	Ulua-Sumo	0.00
2	Sumo-Misquito	0.01	Sumo-Misquito	0.02
3	Ulua-Misquito	0.01	Ulua-Misquito	0.02
4	Cabecar-Bribri	0.04	Cabecar-Bribri	0.05
5	Sambu-Catio	0.05	Sambu-Catio	0.07
...	...	...	...	...
591	Quiche-Bocota	0.58	Quiche-Bocota	0.54
592	Quiche-Cabecar	0.58	Xinca-Cabecar	0.55
593	Xinca-Cabecar	0.58	Xinca-Teribe	0.56
594	Teribe-Quiche	0.59	Teribe-Quiche	0.57
595	Quiche-Movere	0.60	Quiche-Movere	0.59

only slight differences
- **values**
- **ranks**

The top-5 and bottom-5 language pairs in terms of G-distance and Gd-distance

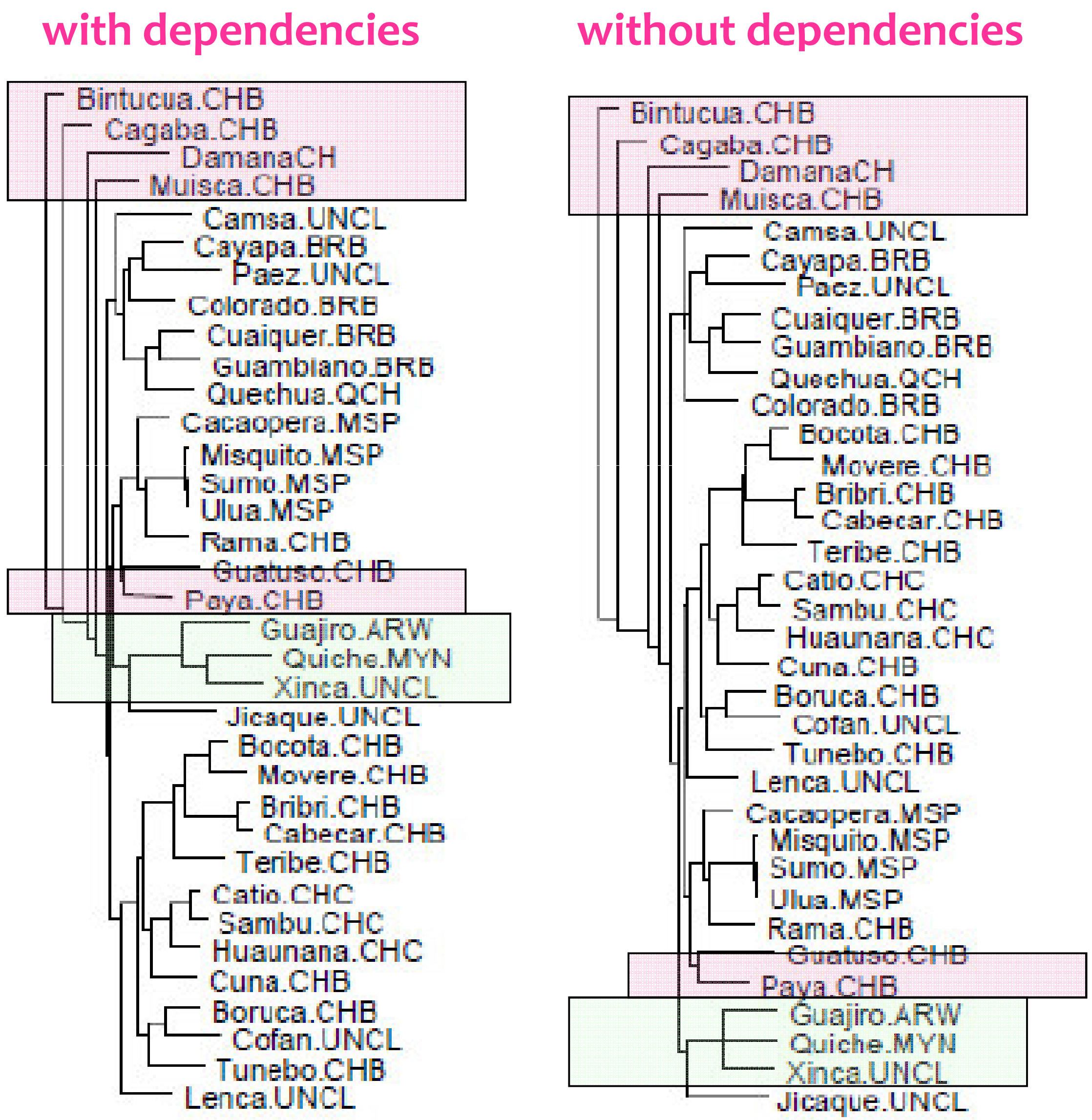
Experimental results: G distances vs. Gd distances (2)

look LESS alike without dependencies | look MORE alike without dependencies

	$G_d - G$	G	G_d		$G_d - G$	G	G_d
Sambu-Cayapa	0.10	0.38	0.48	Quiche-Lenca	-0.08	0.36	0.28
Paya-Bintucua	0.09	0.26	0.35	Quiche-Cayapa	-0.07	0.43	0.36
Paya-Cagaba	0.09	0.22	0.31	Quiche-Paez	-0.06	0.49	0.43
Ulua-Paez	0.09	0.36	0.45	Quiche-Cuna	-0.06	0.41	0.35
Sumo-Paez	0.09	0.36	0.45	Quiche-Boruca	-0.06	0.47	0.41
Cuna-Boruca	0.09	0.26	0.35	Xinca-Camsa	-0.06	0.36	0.30
Paya-Muisca	0.09	0.25	0.33	Xinca-Cofan	-0.06	0.44	0.39
Paez-Bintucua	0.09	0.41	0.49	Quiche-Huaunana	-0.06	0.46	0.40
Huaunana-Boruca	0.08	0.23	0.32	Xinca-Boruca	-0.06	0.44	0.39
Paez-Misquito	0.08	0.35	0.43	Quiche-Colorado	-0.05	0.43	0.38

Language pairs that became more distant (left) or became closer (right) as a result of applying the dependency-sensitive version of the Gower coefficient

Experimental results: G distances vs. Gd distances (3)



FEW DIFFERENCES:

- 43.2% redundancy uniform thru-out

OR

- Is elimination of dependency uninteresting??

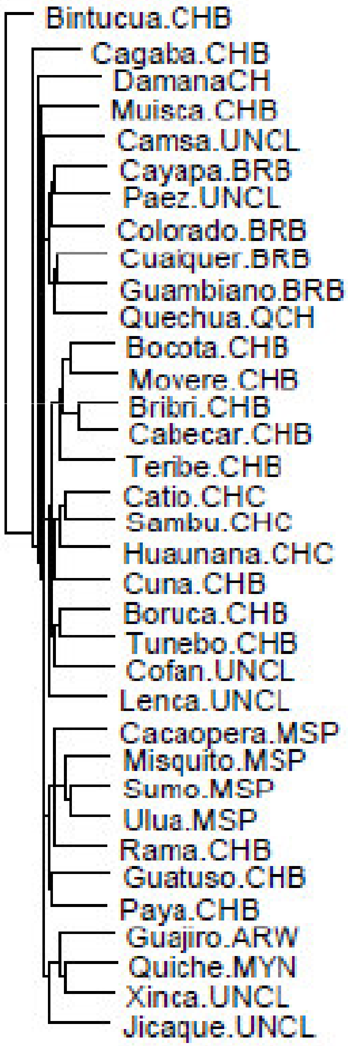
Experimental results: G distances vs. Gq distances (1)

Rank	G all features have equal weight		G_q^1 unary quirks considered		G_q^2 binary quirks considered	
1	Ulua-Sumo	0.00	Cabecar-Bribri	0.49	Cabecar-Bribri	0.48
2	Sumo-Misquito	0.01	Ulua-Sumo	0.53	Ulua-Sumo	0.53
3	Ulua-Misquito	0.01	Sumo-Misquito	0.55	Sumo-Misquito	0.54
4	Cabecar-Bribri	0.04	Ulua-Misquito	0.55	Ulua-Misquito	0.54
5	Sambu-Catio	0.05	Sambu-Catio	0.58	Sambu-Catio	0.58
...	...	...	...	...		
591	Quiche-Bocota	0.58	Xinca-Cabecar	0.93	Xinca-Cabecar	0.93
592	Quiche-Cabecar	0.58	Xinca-Teribe	0.93	Xinca-Teribe	0.93
593	Xinca-Cabecar	0.58	Quiche-Bocota	0.93	Quiche-Bocota	0.93
593	Teribe-Quiche	0.59	Quiche-Movere	0.94	Quiche-Movere	0.94
595	Quiche-Movere	0.60	Teribe-Quiche	0.94	Teribe-Quiche	0.94

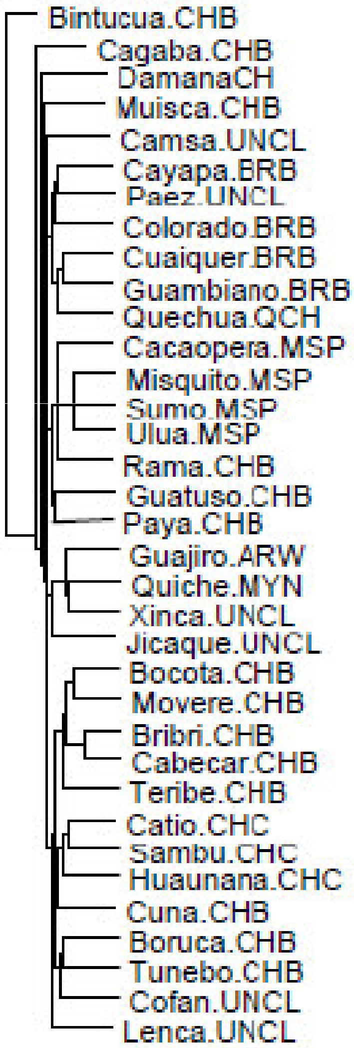
The top-5 and bottom-5 language pairs
in terms of G-distance, Gq1-distance and Gq2-distance

Experimental results: G distances vs. Gq distances (2)

unary quirks considered



binary quirks considered



FEW DIFFERENCES:

- Gq-1 tree= Gd tree topologically
- in Gq-2 tree diffs are de-accentuated, smaller than in Gq-1 tree
  - many possible quirks
  - most not shared

If these results typical of feature set in general, then quirkiness also not interesting??

Conclusions and thoughts on the next steps

1. Presented 2 approaches to factoring out functional dependencies from datasets of typological features – both with **assumption** that **dependencies are of low order** (sets of 1 or 2 features are predictors)

2. Experiments on a dataset of 38 languages of the Chibcha Sphere resulted in **few differences** between **blind and dependency-sensitive distance metrics**

3. Results suggest that dependencies inhabit the feature matrix uniformly, with **no striking effects** between **neighbors or unrelated pairs,** despite the high percentage of redundancy at 43.2%

4. Future tests should involve

   datasets with **more/different languages and families**

   tests of **higher order dependencies**, if tractable methods found

THANK YOU

h.hammarstrom@let.ru.nl
l.oconnor@let.ru.nl

Find feature implications (2)

	F_1	F_2	F_3	F_4
L_1	1	a	1	a
L_2	1	a	0	b
L_3	1	a	1	?
L_4	1	b	0	?
L_5	0	b	1	?
L_6	0	b	0	?
L_7	0	c	1	?
L_8	0	c	0	?
$H(A)$	1.00	1.56	1.00	0.81

	P(A, B)				MI(A,B)	$\frac{MI(A,B)}{H(A)}$	$\rightarrow$
$F_1 \rightarrow F_2$	$P(1,a) = 3/8$	$P(1,b) = 1/8$	$P(0,b) = 2/8$	$P(0,c) = 2/8$	0.65	$\frac{0.65}{1.56}$	0.41
$F_2 \rightarrow F_1$	$P(1,a) = 3/8$	$P(1,b) = 1/8$	$P(0,b) = 2/8$	$P(0,c) = 2/8$	0.65	$\frac{0.65}{1.00}$	0.65
$F_1 \rightarrow F_3$	$P(0,0) = 2/8$	$P(0,1) = 2/8$	$P(1,0) = 2/8$	$P(1,1) = 2/8$	0.00	$\frac{0.00}{1.00}$	0.00

Table 3: A toy example of languages and features, and some example calculations of feature implications