Basic Word Order and Language Contact

Harald Hammarström

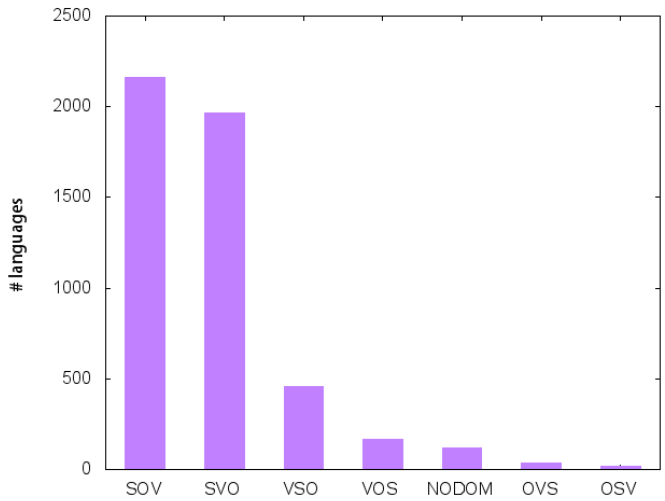
15-16 Jan 2014, Århus

Basic Word Order and Language Contact

- Today's Questions:
  - How much language contact is there?
  - Is more within family or more between families?
  - How much is contact-induced retention and how much is contact-induced change?
  - How much is substrate effects and how much is superstrate effects?
- To answer these questions we first have to separate other language contact from other factors influencing basic word order!

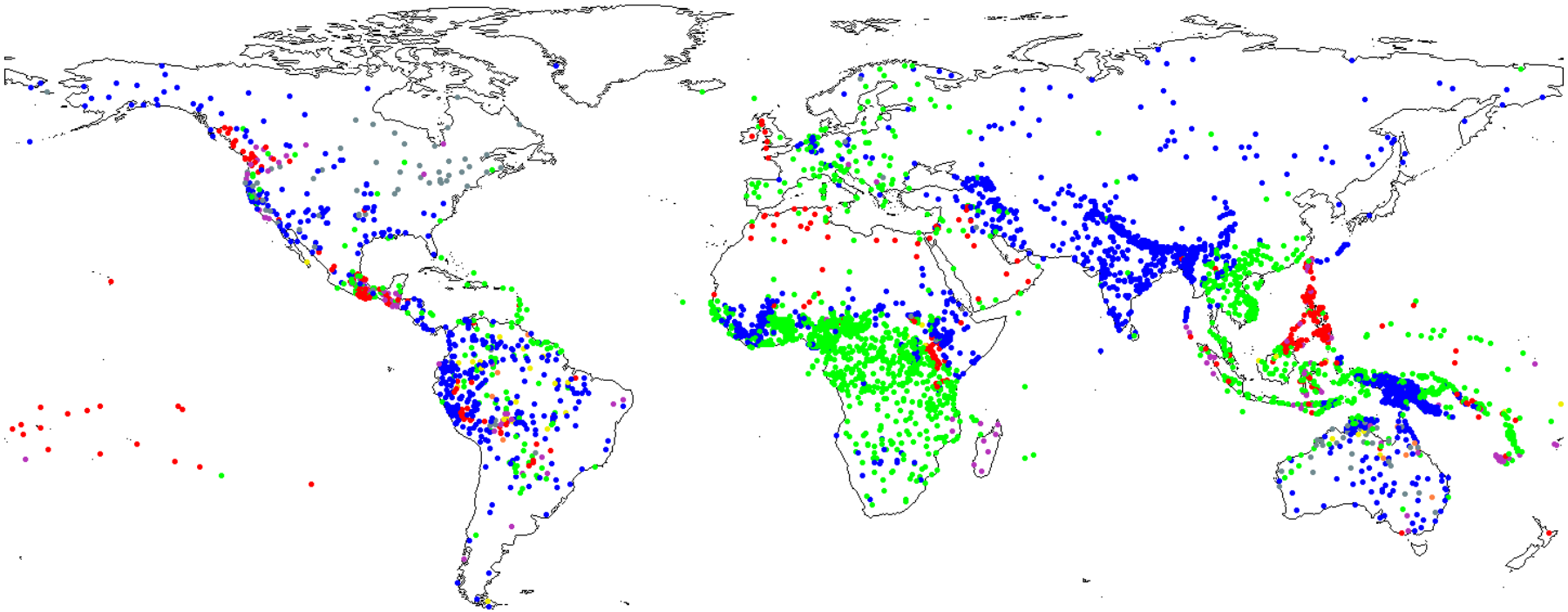
Data: 4916 Languages

	# languages	
SOV	2160	**43.9%**
SVO	1963	**39.9%**
VSO	457	**9.2%**
VOS	165	**3.3%**
NODOM	120	**2.4%**
OVS	35	**0.7%**
OSV	16	**0.3%**
	4916	
No published data	2412	
No access to data	150	
	7478	



All data from first hand descriptive sources

Geographical Distribution



SOV	blue	VOS	purple	OSV	orange
SVO	green	NODOM	slate gray		
VSO	red	OVS	yellow		

Data and Explanations

There are three reasons why a lot of languages have the same value:

Genealogical History	$<$	Languages stem from a common ancestor
$+$		
Areal History	$<$	Languages in contact have influenced each other
$+$		
Universals	$<$	Innate/Communicative/Cognitive constraints?
$=$		
Data		

Disentangling Explanations

- For each one of 4916 languages
  - Basic Word Order is known
  - Classification family/subfamily is known `glottolog.org`
  - Geographical coordinate is known
- This means we can measure the relative impact of areality and genealogy
- ... and attribute any residue to universal tendencies

UGA Decomposition

Explain every datapoint as a mix of weighted factors $\alpha \cdot P_U + \beta \cdot P_G + \gamma \cdot P_A$ with weights

$$\alpha + \beta + \gamma = 1$$

U(niversal): The BWO is drawn from an assumed universal distribution P_U

G(enealogical): The probability P_G of the observed BWO for the most likely projected BWO of its immediate ancestor

A(real): The BWO is drawn from the BWO distribution P_A of its neighbours

Try all α, β, γ and see which fits the observed data best. If $\alpha > 0$ there is evidence for universals!

Universal

- If there is a universal tendency at play, it should be close to the one achieved by areal & genealogical stratification i.e.

SOV	242	**65.7%**
SVO	56	**15.2%**
VSO	26	**7.0%**
NODOM	25	**6.7%**
VOS	12	**3.2%**
OVS	4	**1.0%**
OSV	3	**0.8%**

- (We could try other universal tendencies, but it is already intuitively clear that this will give a poorer fit)

Genealogical

- Given a set of languages $\{L_1, L_2, \ldots, L_n\}$ and their latest common ancestor A
- We usually do not know what the BWO of A was
- But given the BWO values of $\{L_1, L_2, \ldots, L_n\}$ we can pick a *most likely* value to infer for A
- For example, if there were no Universal or Areal factors, the most likely value for A is just the majority value for $\{L_1, L_2, \ldots, L_n\}$

Areal

- Every language L has a number of neighbours $\{N_1, N_2, \ldots, N_n\}$

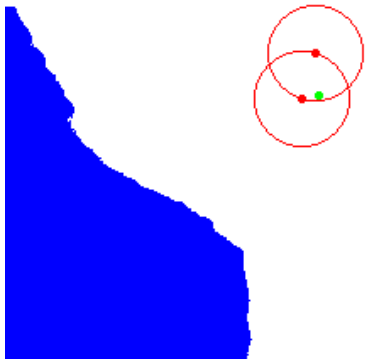
  See next slide for definition

- We may model areal influence such that L picks a random value from its neighbours' values
- (This is oblivious to asymmetries often present in real contact situations where one of two neighbours influences the other, but not vice versa)

Neighbouring Languages

- Two languages A and B are neighbours iff there is no language C located between them
- C is between A and B if C is both closer to A and closer to B, than A and B are to each other

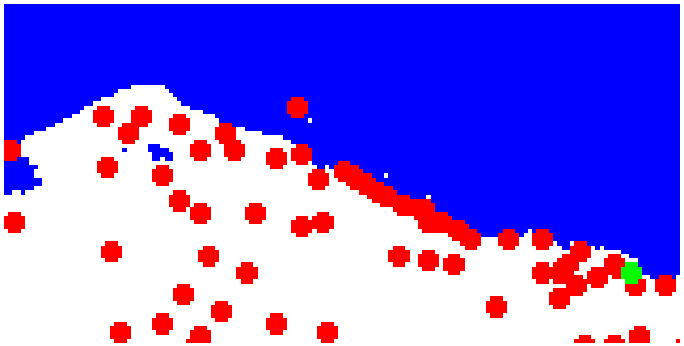
$$N(A,B) = \begin{array}{l} \neg\exists C \\ d(A,C) < d(A,B) \wedge \\ d(B,C) < d(A,B) \end{array}$$



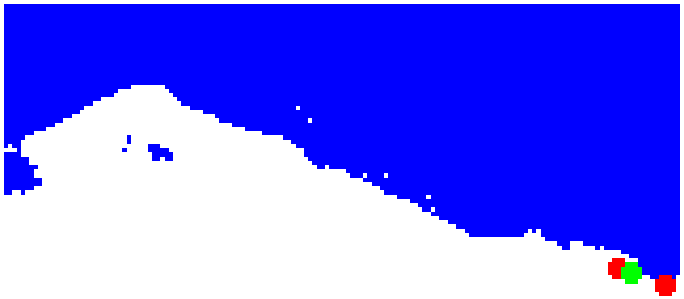
- This is equivalent to checking if the intersection of circles centered at A and B with radius $d(A,B)$ is inhabited

Example: Kayupulau

- Kayupulau is an SOV Austronesian language on the North Coast of Papua



- Kayupulau has 2 neighbours: Skou [set] A SOV Sko family language Tobati [tti] A OSV Austronesian language



Kayupulau belongs to the Sarmi coast AN subgroup

Tobati [tti]	tti	OSV
Tarpia [tpf]	tpf	SOV
Kaptiau [kbi]	kbi	SOV
Bonggo [bpg]	bpg	SOV
Yamna [ymn]	ymn	SVO
Sobei [sob]	sob	SVO
Liki [lio]	lio	SVO
Wakde [wkd]	wkd	SVO
Anus [auq]	auq	SVO
Podena [pdn]	pdn	SVO
Ormu [orz]	orz	SOV
Kayupulau [kzu]	kzu	SOV

	# lgs	
SVO	6	**50.9%**
SOV	5	**41.6%**
OSV	1	**8.3%**
	12	

What Caused Kayupulau to be SOV?

- UGA model says $\alpha \cdot U + \beta \cdot G + \gamma \cdot A$ generated Kayupulau's BWO
- U here is SOV: 0.646, SVO: 0.13, etc.
- A here is SOV: 1/2, OSV: 1/2
- Suppose we are told what α, β, γ are **and** what the BWO proto-Sarmi, e.g., $\alpha = 0.2, \beta = 0.3, \gamma = 0.5$ and proto-Sarmi was SVO

Kayupulau	$\alpha \cdot U + \beta \cdot G + \gamma \cdot A$		P
SOV	$0.2 \cdot 0.646 + 0.3 \cdot 0 + 0.5 \cdot 1/2$	$=$	0.379
SVO	$0.2 \cdot 0.13 + 0.3 \cdot 1 + 0.5 \cdot 0$	$=$	0.326
...			

- This would predict Kayupulau should have been SOV even if proto-Sarmi is SVO!
- And if proto-Sarmi was SOV

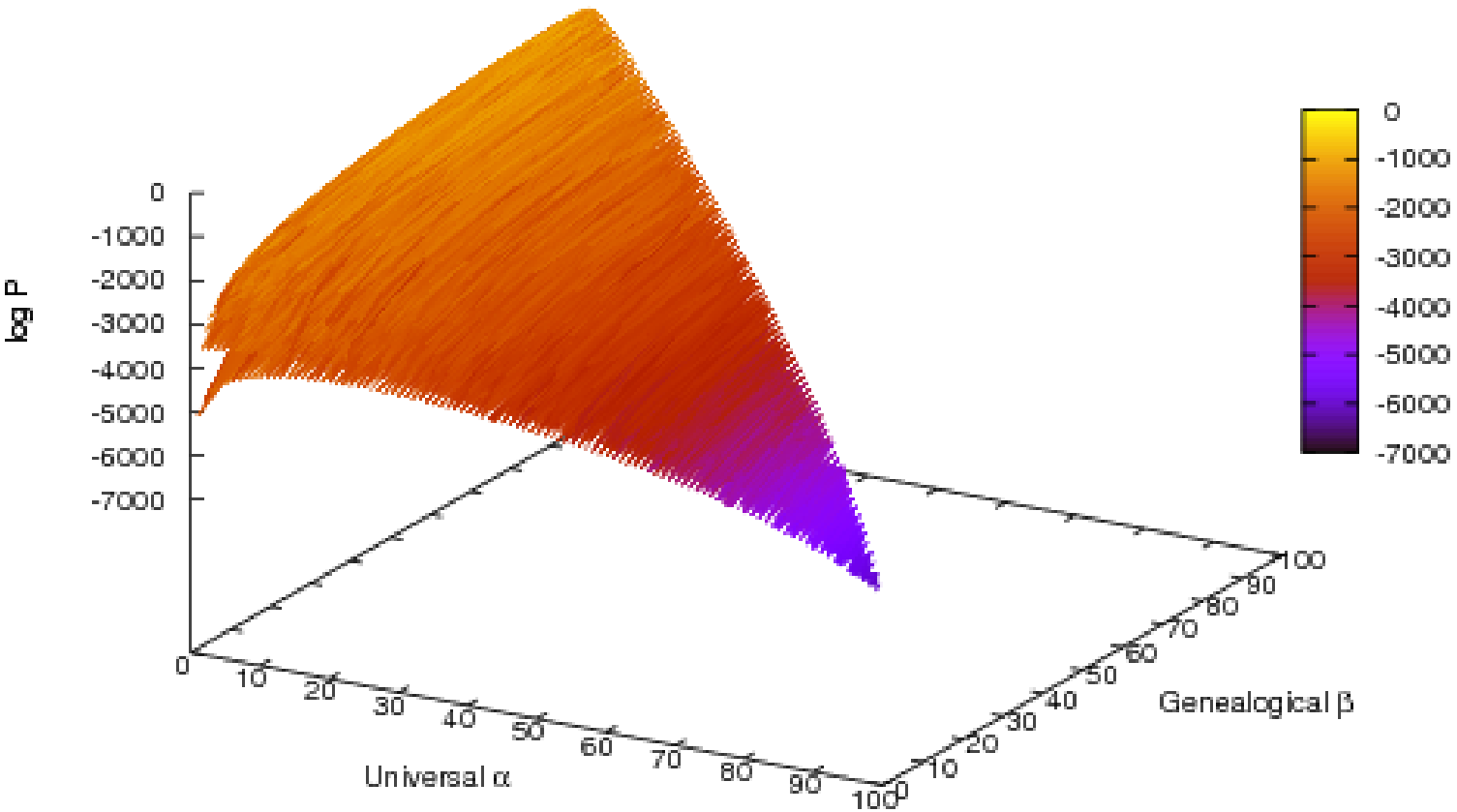
Kayupulau	$\alpha \cdot U + \beta \cdot G + \gamma \cdot A$		P
SOV	$0.2 \cdot 0.646 + 0.3 \cdot 1 + 0.5 \cdot 1/2$	$=$	0.679
SVO	$0.2 \cdot 0.13 + 0.3 \cdot 0 + 0.5 \cdot 0$	$=$	0.026
...			

- Then Kayupulau is predicted to be SOV with even higher probability

Outline of Computation

- Step through all values for α, β, γ
- For each setting, calculate the observed-world probability
  - α, β, γ are given
  - For every terminal branch (interior node with only leaves), set its proto value as the value which maximizes the probability of the values of its children
  - Calculate the observed-world probability as the product of all the terminal branch probabilities
- Find the α, β, γ-setting which maximizes the likelihold of the observed world

Results



The best fit is:

Universal $\alpha \approx 0.14$ Genealogical $\beta \approx 0.78$ Areal $\gamma \approx 0.08$

When is Contact Needed as an Explanation?

- The best fit history $\alpha \approx 0.14$, $\beta \approx 0.78$, and $\gamma \approx 0.08$ comes with an explicit history of changes from proto-terminal branches to leaves
- The best fit history does not say which specific changes/retentions were due to what factor, it only gives the "overall" number that an areal weight of 0.08 is needed
- Thus, we need to check exactly which cases benefit from a contact scenario:
  - Compare the best fit history to a contact-less history (where the areal weight $\gamma = 0.0$)
  - The transitions which have a higher probability in the best fit history than in the contact-less history can be called the **likely cases of contact**
  - 3054 out of 4916 (61%) transitions are likely cases of contact (to various degrees)

Likely Cases of Contact

- By far the biggest effect of contact is contact-induced **retention**

From	To	# Cases	
SOV	SOV	1380	**45.1%**
SVO	SVO	1210	**39.6%**
VSO	VSO	202	**6.6%**
SVO	SOV	56	**1.8%**
VOS	VOS	43	**1.4%**
NODOM	NODOM	27	**0.8%**
SOV	SVO	19	**0.6%**
VSO	SVO	12	**0.3%**
...			
		3054	

- In the likely cases of contact, 64% of the neighbour pairs are of the same family
- This is roughly the same as in general: for all neighbour pairs, 67% are of the same family

Direction of Word Order Borrowing

- In general:
  - Most languages (78.5%) have at least one demographically larger neighbour
  - 2/3-rds of languages (66.4%) are smaller than the average of their neighbours
- In the likely cases of contact:
  - Most languages (71.2%) have at least one demographically larger neighbour
  - Almost 2/3-rds of languages (58.1%) are smaller than the average of their neighbours
- If this means anything, it means that the larger languages are proportionately more influenced by (substrate?) contact, rather than vice versa

Conclusions

- The biggest effect of language contact in basic word order is contact induced retention
- Languages are oblivious as to who they borrow from, but since 2/3-rds of the neighbours of a language belongs to the same family, 2/3-rds of their borrowings is intra-family
- If anything is discernable with present-day population figures, it is, on average, the larger languages that are proportionately influenced by the smaller languages around them, rather than vice versa
- Presumably this comes about as a language becomes large by subuming other languages, whose word order, to some extent, remains as a substrate effect