The Basic Word Order Typology: An Exhaustive Study

Harald Hammarström

3 May 2015, Leipzig

The Basic Word Order Typology

- One of typology's most celebrated themes, popularized by Greenberg (1963) in a study comprising 30 languages.
- Since then, basic-word-order statistics from ever wider arrays of languages have been presented
    - Hawkins 1983: 336 languages
    - Tomlin 1986: 402 languages
    - Haarmann 2004: 636 languages
    - Dryer 2005: 1228 languages
- Today we will look at statistics from 5230 languages

Defintion of Basic Word Order #1

- Transitive declarative main clause
- Both subject and object involve an overt noun phrase (not just a pronoun)

[The woman]	chased	[the man]
S	V	O

Note:

- Subject: The more agent-like of the arguments
- Object: The more patient-like of the arguments

  So SVO is really better labeled AVP

Defintion of Basic Word Order #2

Language Has Basic Word Order:

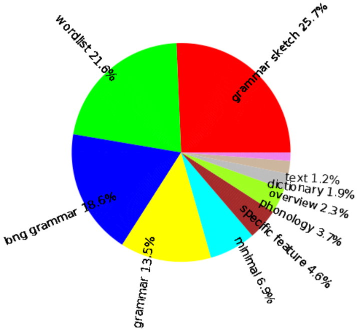
- Only one order is grammatically possible OR
- Several orders are possible AND
  - There is a difference in meaning and one of the orders can be considered neutral
  - There is no difference in meaning but there one order is MUCH more frequent than the others

Language Doesn't Have Basic Word Order:

- Several orders possible and common/neutral
- Several orders occur, not freely, but conditioned by morphosyntax (e.g., the presence of an auxiliary)

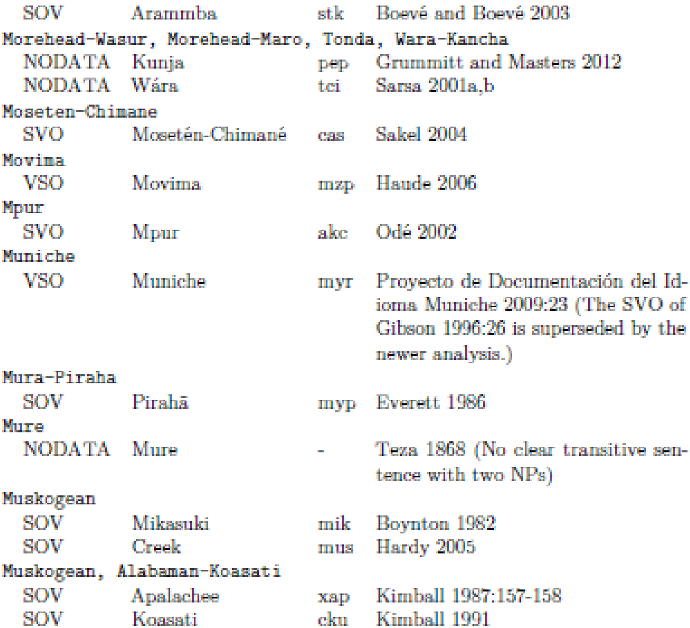
Status of Documentation of the World's Languages

MED type	# lgs	
long grammar	1403	**18.6%**
grammar	1015	**13.4%**
grammar sketch	1931	**25.6%**
specific feature	346	**4.5%**
phonology	277	**3.6%**
dictionary	143	**1.8%**
text	93	**1.2%**
wordlist	1631	**21.6%**
minimal	516	**6.8%**
overview	174	**2.3%**
	7529	



- I was able to get word order data from 5230 languages
- For 82 lgs there is data but I have not been access it (yet)
- For the remaining 2219 lgs there is no published data on word order

Example Page of Database



SOV	Arammba	stk	Boevé and Boevé 2003
Morehead-Wasur, Morehead-Maro, Tonda, Wara-Kancha			
NODATA	Kunja	pep	Grummitt and Masters 2012
NODATA	Wára	tci	Sarsa 2001a,b
Moseten-Chimane			
SVO	Mosetén-Chimané	cas	Sakel 2004
Movima			
VSO	Movima	mzp	Haude 2006
Mpur			
SVO	Mpur	akc	Odé 2002
Muniche			
VSO	Muniche	myr	Proyecto de Documentación del Idioma Muniche 2009:23 (The SVO of Gibson 1996:26 is superseded by the newer analysis.)
Mura-Piraha			
SOV	Pirahã	myp	Everett 1986
Mure			
NODATA	Mure	-	Teza 1868 (No clear transitive sentence with two NPs)
Muskogean			
SOV	Mikasuki	mik	Boynton 1982
SOV	Creek	mus	Hardy 2005
Muskogean, Alabaman-Koasati			
SOV	Apalachee	xap	Kimball 1987:157-158
SOV	Koasati	cku	Kimball 1991

Comparison of Data Sources

1. **Haarmann:** 636 data points
   - Sources not systematically indicated
   - Convenience selection
2. **Ethnologue 17th ed.:** 1281 data points
   - Sources for the data points are not indicated
   - It is not clear how the data points/languages were selected
3. **WALS:** 1302 data points
   - Sources for the data points are indicated
   - It is not clear how the data points/languages were selected, but it may be guessed that it is some kind of convenience sample
4. **Hammarström:** 5230 data points
   - Sources for the data points are indicated
   - Every language checked

Dataset Agreement

	HAAR	HH	WALS
E17	**97.0%**	**90.0%**	**83.5%**
	840/866	1125/1250	338/405
HAAR		**80.6%**	**79.0%**
		965/1197	362/458
HH			**86.5%**
			1117/1292

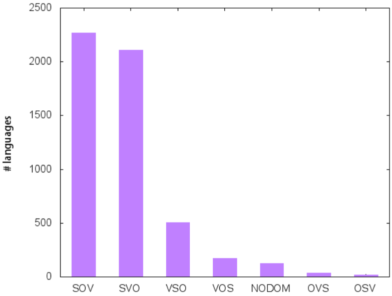
It turns out that the bulk Haarmann is lifted from (an earlier edition of) Ethnologue!

Dataset Differences

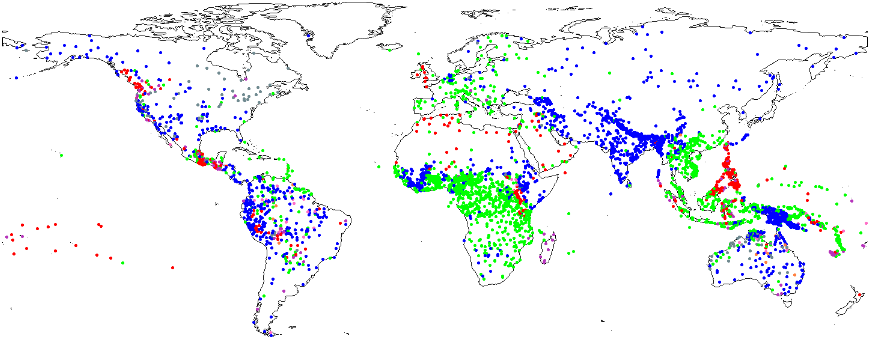
WALS	E17	#	WALS	HH	#	E17	HH	#
NODOM	SVO	9	NODOM	SOV	28	SVO	VSO	19
NODOM	VSO	7	NODOM	SVO	22	SVO	SOV	8
SVO	SOV	6	NODOM	VOS	13	VOS	VSO	6
NODOM	SOV	5	SVO	SOV	9	SVO	VOS	6
VSO	VOS/VSO	3	NODOM	VSO	9	SOV	NODATA	6
SVO	SVO/VSO	3	SVO	VSO	8	SOV	SVO	5
SVO	SOV/SVO	3	SOV	NODATA	8	OSV	SOV	5
SOV	SVO	3	SVO	NODATA	5	SOV	NODOM	2
NODOM	OVS	3	VSO	VOS	2	SVO	NODOM	2
...	...	...	...	...	...	...	...	...
		67			175			125

Basic Word Order Statistics

	# lgs	
SOV	2267	**43.3%**
SVO	2107	**40.2%**
VSO	502	**9.5%**
VOS	174	**3.3%**
NODOM	123	**2.3%**
OVS	38	**0.7%**
OSV	19	**0.3%**
	5230	



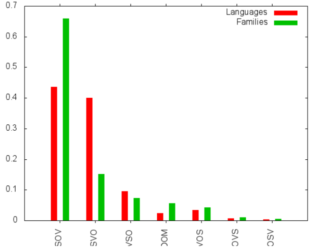
Geographical Distribution



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

Genealogically Stratified

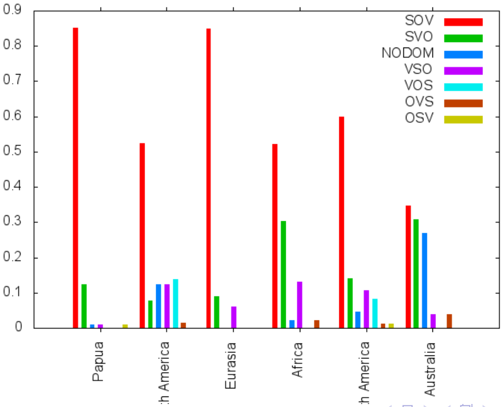
	All languages		One per family		Isolates		Majority per family	
SOV	2260	**43.3%**	241	**65.1%**	114	**67.1%**	131	**65.8%**
SVO	2101	**40.3%**	60	**16.2%**	23	**13.5%**	32	**16.1%**
VSO	498	**9.5%**	26	**7.0%**	14	**8.2%**	11	**5.5%**
VOS	174	**3.3%**	16	**4.3%**	6	**3.5%**	10	**5.0%**
NODOM	123	**2.3%**	21	**5.7%**	9	**5.3%**	14	**7.0%**
OVS	38	**0.7%**	5	**1.4%**	3	**1.8%**	2	**0.5%**
OSV	19	**0.3%**	1	**0.3%**	1	**0.6%**	0	**0.0%**
	5213		370		170		200	



- When we remove family bias, the ratio of SOV goes up, on the expense of SVO
- Some large families are responsible for the proliferation of SVO

Genealogically & Areally Stratified

	Papua		Australia		Eurasia		Africa		North America		South America	
SOV	96	84.2%	9	34.6%	28	84.8%	24	52.2%	32	49.2%	52	61.2%
SVO	15	13.2%	7	26.9%	3	9.1%	15	32.6%	8	12.3%	11	12.9%
VSO	1	0.9%	0	0.0%	2	6.1%	6	13.0%	11	16.9%	8	9.4%
NODOM	1	0.9%	7	26.9%	0	0.0%	1	2.2%	8	12.3%	4	4.7%
VOS	0	0.0%	1	3.8%	0	0.0%	0	0.0%	5	7.7%	8	9.4%
OVS	0	0.0%	2	7.7%	0	0.0%	0	0.0%	1	1.5%	1	1.2%
OSV	1	0.9%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	1.2%
	114		26		33		46		65		85	
	31.2%		7.1%		9.0%		12.6%		17.8%		23.3%	



Universals of Basic Word Order

- There is variation in the 6 continent size areas, but
- Some results recur in the 6 macro-areas and when family bias is removed
  - SOV is the most common
  - Object-inital is the least common
  - ...
- Thus, they are universal tendencies (Dryer 1992)!

  We have found the precious "linguistic preferences" that give insight to possible innate specification, processing preferences, communicative needs, ...

Universals of Language in the Brain, or?

Are the universal tendencies of word order "linguistic preferences" explainable by innate specification, processing preferences or communicative needs?

- Effects of speech-community size?
  - Object-initial word order in small speech communities (Trudgill, 2011, 100-101)
  - SVO word order associated with large speech communities (e.g., David Gil in this conference)
- Historical contingencies, after all?
  - Because every large family is different internally, the word order tendencies cannot be universal (Dunn et al., 2011)
- The reflection of proto-world SOV order?
  - Proto-world had SOV and we are now in the middle of a drift towards SVO (Gell-Mann and Ruhlen 2011, Maurits and Griffiths 2014)

Population Size Influence on Word Order?

Sociolinguistic Typology: Social Determinants of Linguistic Complexity *2011 Peter Trudgill, pp 100-101:*

- Speech community size matters
- Object-initial word order occur only with small speech communities

The claim is based on:

- Nettle (1999:139) who observes that the set of object-initial languages known to him had a median speaker number of 750
- But this set was not a *random* sample
- Only for *random* samples can we generalize and draw statistically sound conclusions

Word Order and Community Size?

- Sampling *one* language *at random* from every family for which there is data (367 families) turns up
  - 7 object-initial languages:

Language	Order	Population
Panare [pbh]	OVS	3540
Ona [ona]	OVS	Extinct, though 3500-4000 around 1900 which is when children ceased to learn the language.
Urarina [ura]	OVS	3 000
Waikuri [-]	OVS	Extinct, they were supposed to be a "small tribe" and I've been unable to find a specific population estimate. For the sake of the argument, let's assume it had 1 speaker.
Ngarinyin [ung]	OVS	82
Warao [wba]	OSV	28309
Macuna [myy]	OVS	1110

  - 64 SVO languages
- The speakers numbers compare as follows

		Sampling one per family		
	SVO	Object-Initial	Any Order	All E17
Median # speakers	2000	3 000	1 100	7 270
Mean # speakers	24 879	20 711	5 649	697 626

- Neither SVO&large (median $p \approx .367$, mean $p \approx .16$) nor Object-Initial&small (median $p \approx .259$, mean $p \approx .11$) is statistically significant

Every Family Is Different?

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. (2011) **Evolved structure of language shows lineage-specific trends in word-order universals***. Nature 13 April. pp, 1-4.*

- There seems to be a lot of word order variation *within* families
- If there are universals, shouldn't *every* family drift towards the distribution demanded by the universal?
  - The bigger the family, the closer to the universal distribution
- Except if the family is very shallow or a lot of branching happened very recently
  - But then the older the family, the closer it should be the universal distribution

Intra-Family Divergence: Raw Frequencies

	#	SOV	SVO	VSO	NODOM	VOS	OVS
Austronesian	802	**8.7%**	**57.4%**	**21.8%**	**0.0%**	**11.2%**	**0.6%**
Atlantic-Congo	787	**5.8%**	**93.9%**	**0.1%**	**0.0%**	**0.1%**	**0.0%**
Indo-European	517	**60.7%**	**35.2%**	**2.9%**	**0.8%**	**0.4%**	**0.0%**
Sino-Tibetan	320	**89.1%**	**10.9%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Afro-Asiatic	262	**24.4%**	**51.9%**	**22.1%**	**0.4%**	**1.1%**	**0.0%**
Nuc. Trans New Guinea	164	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Pama-Nyungan	133	**69.9%**	**8.3%**	**0.0%**	**12.0%**	**3.0%**	**2.3%**
Otomanguean	119	**1.7%**	**9.2%**	**79.8%**	**0.0%**	**9.2%**	**0.0%**
Austroasiatic	99	**20.2%**	**74.7%**	**2.0%**	**0.0%**	**3.0%**	**0.0%**
Mande	63	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Tai-Kadai	62	**6.5%**	**93.5%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Dravidian	52	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Tupian	51	**62.7%**	**23.5%**	**2.0%**	**0.0%**	**2.0%**	**7.8%**
Arawakan	47	**12.8%**	**40.4%**	**42.6%**	**0.0%**	**4.3%**	**0.0%**
Uto-Aztecan	45	**44.4%**	**20.0%**	**22.2%**	**8.9%**	**4.4%**	**0.0%**
Quechuan	44	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Nilotic	41	**4.9%**	**31.7%**	**51.2%**	**0.0%**	**0.0%**	**12.2%**
Turkic	40	**92.5%**	**7.5%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Central Sudanic	40	**22.5%**	**77.5%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Athapaskan-Eyak-Tlingit	37	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**

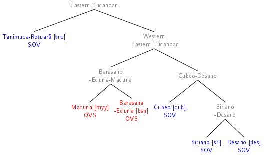
Zooming in on the History of Changes

- Raw-intra family distributions do not take the family tree topology into account
- We know quite a lot about the history of languages when knowing family-tree internal classifications (source `glottolog.org`)
- We can check this knowledge of the history of languages to what is predicted by the existence of universals
- Method: Estimate *transition probabilities* in a family tree

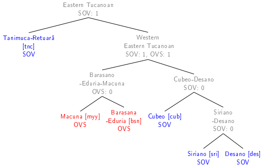
Example: Parsimony Reconstruct

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

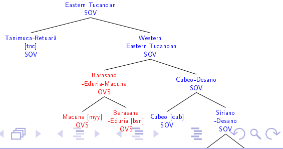
1. Input (a tree and values at the leaves)



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



Example: Transitions

From	To	From	To
Tucanoan	Eastern Tucanoan	SOV	SOV
Tucanoan	Coreguaje-Siona	SOV	SOV
Eastern Tucanoan	Eastern Eastern Tucanoan	SOV	SOV
Eastern Tucanoan	Tanimuca-Retuarã [tnc]	SOV	SOV
Eastern Tucanoan	Western Eastern Tucanoan	SOV	SOV
Coreguaje-Siona	Siona-Secoya	SOV	SOV
Coreguaje-Siona	Koreguaje [coe]	SOV	VSO
Eastern Eastern Tucanoan	Eastern Eastern Tucanoan II	SOV	SOV
Eastern Eastern Tucanoan	Eastern Eastern Tucanoan I	SOV	SOV
Western Eastern Tucanoan	Barasano-Eduria-Macuna	SOV	OVS
Western Eastern Tucanoan	Cubeo-Desano	SOV	SOV
Eastern Eastern Tucanoan II	Guanano [gvc]	SOV	SOV
Eastern Eastern Tucanoan II	Tuyuca [tue]	SOV	SOV
Eastern Eastern Tucanoan I	Waimaha [bao]	SOV	SOV
Eastern Eastern Tucanoan I	Tucano [tuo]	SOV	SOV
Barasano-Eduria-Macuna	Macuna [myy]	OVS	OVS
Barasano-Eduria-Macuna	Barasana-Eduria [bsn]	OVS	OVS
Cubeo-Desano	Siriano-Desano	SOV	SOV
Cubeo-Desano	Cubeo [cub]	SOV	SOV
Siriano-Desano	Siriano [sri]	SOV	SOV
Siriano-Desano	Desano [des]	SOV	SOV
Siona-Secoya	Siona-Tetete [snn]	SOV	SOV
Siona-Secoya	Secoya [sey]	SOV	SOV

Example: Transition Probabilities

- Transition frequencies

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

- Transition probabilities

	SOV	VSO	OVS
SOV	0.905	0.048	0.048
OVS	0.000	0.000	1.000

Family-Variation in Transition Probabilities: From SOV

	#	SOV	SVO	VSO	NODOM	VOS	OVS
All	7755	94.8%	3.0%	0.3%	1.0%	0.1%	0.4%
Austronesian	1258	94.5%	4.5%	0.0%	0.0%	0.0%	0.0%
Atlantic-Congo	1288	79.2%	20.8%	0.0%	0.0%	0.0%	0.0%
Indo-European	779	94.3%	4.6%	0.4%	0.2%	0.3%	0.0%
Sino-Tibetan	495	98.5%	1.5%	0.0%	0.0%	0.0%	0.0%
Afro-Asiatic	419	90.9%	6.1%	2.6%	0.4%	0.0%	0.0%
Nuc. Trans New Guinea	259	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Pama-Nyungan	216	84.5%	4.8%	0.6%	6.8%	1.1%	0.0%
Otomanguean	170	38.2%	32.3%	14.7%	0.0%	14.7%	0.0%
Austroasiatic	158	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mande	112	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Tai-Kadai	99	60.0%	40.0%	0.0%	0.0%	0.0%	0.0%
Dravidian	83	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Tupian	76	81.0%	12.1%	0.9%	0.0%	0.0%	4.3%
Arawakan	68	41.7%	41.7%	16.7%	0.0%	0.0%	0.0%
Uto-Aztecan	72	79.0%	3.9%	6.1%	10.3%	0.7%	0.0%
Quechuan	60	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Nilotic	68	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%
Turkic	60	94.8%	5.2%	0.0%	0.0%	0.0%	0.0%
CentralSudanic	66	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%
Athapaskan-Eyak-Tlingit	52	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Family-Variation in Transition Probabilities: From SVO

	#	SOV	SVO	VSO	NODOM	VOS	OVS
All	7755	**3.3%**	**92.6%**	**2.3%**	**0.4%**	**1.3%**	**0.1%**
Austronesian	1258	**1.5%**	**91.0%**	**3.8%**	**0.0%**	**3.5%**	**0.1%**
Atlantic-Congo	1288	**1.9%**	**98.0%**	**0.1%**	**0.0%**	**0.0%**	**0.0%**
Indo-European	779	**7.7%**	**89.4%**	**1.5%**	**1.1%**	**0.3%**	**0.0%**
Sino-Tibetan	495	**5.4%**	**94.6%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Afro-Asiatic	419	**2.9%**	**87.6%**	**8.0%**	**0.2%**	**1.3%**	**0.0%**
Nuc. Trans New Guinea	259	-	-	-	-	-	-
Pama-Nyungan	216	**29.7%**	**56.3%**	**0.0%**	**12.7%**	**0.0%**	**0.6%**
Otomanguean	170	**15.9%**	**42.7%**	**20.7%**	**0.0%**	**20.7%**	**0.0%**
Austroasiatic	158	**0.9%**	**95.3%**	**2.1%**	**0.0%**	**1.7%**	**0.0%**
Mande	112	-	-	-	-	-	-
Tai-Kadai	99	**2.2%**	**97.8%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Dravidian	83	-	-	-	-	-	-
Tupian	76	**14.1%**	**67.9%**	**0.0%**	**0.0%**	**7.7%**	**10.3%**
Arawakan	68	**16.9%**	**66.9%**	**13.0%**	**0.0%**	**3.2%**	**0.0%**
Uto-Aztecan	72	**5.8%**	**60.1%**	**25.6%**	**0.5%**	**8.0%**	**0.0%**
Quechuan	60	-	-	-	-	-	-
Nilotic	68	**3.9%**	**83.3%**	**4.9%**	**0.0%**	**0.0%**	**7.8%**
Turkic	60	**50.0%**	**50.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
CentralSudanic	66	**5.6%**	**94.4%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Athapaskan-Eyak-Tlingit	52	-	-	-	-	-	-

Every Family Is Different, But

- In terms of raw internal frequencies

  Every family **is different**

- In terms change patterns

  Every family **is the same** *(with few exceptions)*

- Thus, language families do behave the same, it is simply that
  - The proto-languages of families started out with different word orders
  - Changes are relatively uncommon, i.e., word order is relatively stable

Proto-World Word Order?

The origin and evolution of word order *2011 Murray Gell-Mann and Merritt Ruhlen, pp 1-6:*

The idea goes:

- Too many spontaneous changes *towards* SVO, not SOV
- Proto-World had SOV
- We are now in the middle of a drift towards SVO
- If we play time ahead X millenia the world will have converged towards SVO
- Problem with the data presented in Gell-Mann & Ruhlen (2011): spontaneous changes distinguished from contact-induced changes in a rigged manner
    - Whenever there is a change to SOV, it is blamed on contact
    - Whenever there is a change to SVO, it is deemed a spontaneous change

Global Transition Probabilities

- 8119 transitions in total

	NODOM	OSV	OVS	SOV	SVO	VOS	VSO
NODOM	0.73	0.00	0.01	0.12	0.10	0.01	0.03
OSV	0.09	0.59	0.09	0.23	0.01	0.00	0.00
OVS	0.02	0.01	0.60	0.21	0.09	0.00	0.06
SOV	0.01	0.00	0.00	0.95	0.03	0.00	0.00
SVO	0.00	0.00	0.00	0.03	0.92	0.01	0.02
VOS	0.01	0.00	0.00	0.01	0.11	0.75	0.11
VSO	0.00	0.00	0.01	0.01	0.09	0.05	0.84

- Markov Theory: Every aperiodic irreducible transition matrix determines a *stationary distribution*!

$M \times M$ One Step

	NODOM	OSV	OVS	SOV	SVO	VOS	VSO
NODOM	0.53	0.00	0.01	0.21	0.18	0.02	0.05
OSV	0.12	0.34	0.10	0.38	0.04	0.00	0.01
OVS	0.03	0.01	0.37	0.33	0.15	0.00	0.10
SOV	0.02	0.01	0.01	0.91	0.05	0.00	0.01
SVO	0.01	0.00	0.00	0.06	0.86	0.02	0.04
VOS	0.01	0.00	0.00	0.03	0.20	0.57	0.18
VSO	0.00	0.00	0.01	0.02	0.16	0.09	0.72

$M \times M \times M$ One More Step

	NODOM	OSV	OVS	SOV	SVO	VOS	VSO
NODOM	0.40	0.00	0.01	0.27	0.22	0.04	0.05
OSV	0.08	0.13	0.00	0.62	0.07	0.07	0.02
OVS	0.04	0.00	0.18	0.43	0.22	0.02	0.11
SOV	0.02	0.01	0.01	0.87	0.08	0.00	0.01
SVO	0.01	0.00	0.00	0.09	0.80	0.03	0.06
VOS	0.01	0.00	0.00	0.06	0.29	0.39	0.24
VSO	0.01	0.00	0.01	0.04	0.21	0.10	0.63

$M \times M \times M \times M$ Fourth Step

	NODOM	OSV	OVS	SOV	SVO	VOS	VSO
NODOM	0.30	0.00	0.01	0.31	0.26	0.04	0.06
OSV	0.07	0.07	0.01	0.65	0.10	0.06	0.03
OVS	0.04	0.00	0.10	0.46	0.25	0.03	0.11
SOV	0.03	0.01	0.01	0.83	0.10	0.01	0.02
SVO	0.01	0.00	0.00	0.11	0.76	0.04	0.07
VOS	0.02	0.00	0.01	0.08	0.34	0.29	0.26
VSO	0.01	0.00	0.01	0.06	0.26	0.11	0.55

After Many Steps

	NODOM	OSV	OVS	SOV	SVO	VOS	VSO
NODOM	0.02	0.00	0.01	0.42	0.40	0.04	0.10
OSV	0.02	0.00	0.01	0.42	0.40	0.04	0.10
OVS	0.02	0.00	0.01	0.42	0.40	0.04	0.10
SOV	0.02	0.00	0.01	0.42	0.40	0.04	0.10
SVO	0.02	0.00	0.01	0.42	0.40	0.04	0.10
VOS	0.02	0.00	0.01	0.42	0.40	0.04	0.10
VSO	0.02	0.00	0.01	0.42	0.40	0.04	0.10

Transition Predictions vs. Reality

- Even assuming the least possible amount of change (parsimony reconstruction)
- Even assuming that these changes are independent (many are actually to to one and the same historical accident namely European colonialism)
- While there are “too many” transitions to SVO
- Transitions still predict SOV to be most common

	Predicted by Transitions	Observed in isolates
SOV	42.2%	65.1%
SVO	40.0%	16.2%
VSO	10.0%	7.0%
VOS	4.2%	4.3%
NODOM	2.4%	5.7%
OVS	0.7%	1.4%
OSV	0.4%	0.3%

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	**56.8%**
SVO	**13.1%**
VSO	**6.5%**
NODOM	**6.3%**
VOS	**2.5%**
OVS	**0.9%**
OSV	**0.2%**

- (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 $\{N_1, N_2, \ldots, N_n\}$

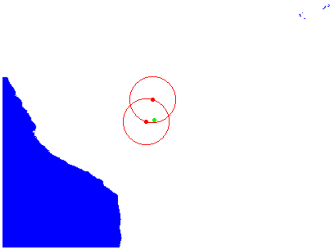
  See next slide for definition

- 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

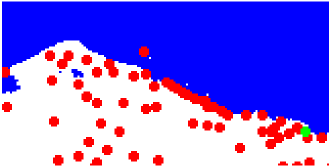
$$N(A,B) = \begin{array}{l} \neg\exists C \\ d(A,C) < d(A,B) \wedge \\ d(B,C) < d(A,B) \end{array}$$



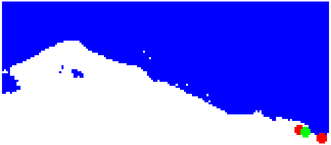
- This is equivalent to checking if the intersection of circles centered at A and B with radius $d(A,B)$ is inhabited

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, VSO: 0.06, NODOM: 0.06 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

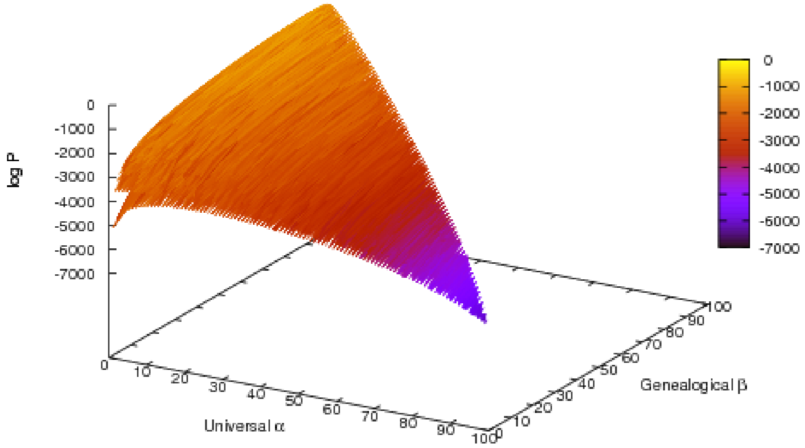
Kayupulau	$\alpha \cdot U + \beta \cdot G + \gamma \cdot A$		P
SOV	$0.2 \cdot 0.646 + 0.3 \cdot 0 + 0.5 \cdot 1/2$	=	0.379
SVO	$0.2 \cdot 0.13 + 0.3 \cdot 1 + 0.5 \cdot 0$	=	0.326
...			

- 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

Results



The best fit is:

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

Conclusion

- Essentially, every family is different in its internal composition, but is the same with respect to change patterns
- With 5230 languages which can take language contact seriously
- The data are best explained by the existence of a universal tendency

SOV	**56.8%**	SVO	**13.1%**
VSO	**6.5%**	VOS	**2.5%**
OVS	**0.9%**	OSV	**0.2%**
	NODOM **6.3%**		

- But the universal is not the only, nor the most important, factor:
    - Most important (78%): the order of the immediate ancestor
    - 2nd most imporant (14%): the order governed by a universal tendency
    - 3rd most important (8%): the order favoured by neighbouring languages

Thank you

Dryer, M. S. (2005). Order of subject, object, and verb. In Comrie, B., Dryer, M. S., Gil, D., and Haspelmath, M., editors, *World Atlas of Language Structures*, pages 330–333. Oxford University Press.

Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 13 April:1–4.

Gell-Mann, M. and Ruhlen, M. (2011). The origin and evolution of word order. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, October 10:1–16.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of language: report of a conference held at Dobbs Ferry, New York, April 13-15, 1961*, pages 73–113. Cambridge, Massachusetts: MIT Press.

Haarmann, H. (2004). *Elementare Wortordnung in den Sprachen der Welt: Dokumentation und Analysen zur Entstehung von Wortfolgemustern*. Hamburg: Helmut Buske.

Hawkins, J. A. (1983). *Word order universals*, volume 3 of *Quantitative Analyses of Linguistic Structure*. San Diego: Academic Press.

Maurits, L. and Griffiths, T. L. (2014). Tracing the roots of syntax with bayesian phylogenetics. *PNAS*, 111(37):13576–13581.

Tomlin, R. S. (1986). *Basic word order: functional principles*. London: Croom Helm.

Trudgill, P. (2011). *Sociolinguistic Typology*. Oxford: Oxford University Press.