Grounding Syntactic Variation
A plea to develop SSWL

Hilda Koopman and Dominique Sportiche
koopman@ucla.edu
dominique.sportiche@gmail.com

University of California, Los Angeles

Advances in Phylogenetic Linguistics
Ragusa Ibla
July 16 2013

Introduction
(with Robin Ryder, Ceremade, Université Paris Dauphine)

Starting point

Dunn, M., Greenhill, S.J., Levinson, S.C. and Gray, R.D.

Evolved structure of language shows lineage-specific trends in word-order universals.

Nature, 2011, 473, 79-82 (201)

- What was done
- What we like about it
- What was interesting to us
- What was concluded
- Why we are skeptical
- What we would like to do about it

What was done: The following 8 relative word order traits from WALS (the World Atlas of Linguistic Structures) were coded on the leaves of phylogenetic independently generated on the basis of lexical data:

- Adjective-Noun: *big balloon*
- Numeral-Noun: *three balls*
- Demonstrative-Noun: *this/that ball*
- Adposition-Noun Phrase: *in the house*
- Possessor Noun: *Bill's house*
- Subject-Verb: *Bill left, Bill saw Mary*
- Verb-Object: *John saw Mary*
- Noun-Relative Clause: *the person who came*

What was done: This was done in four language families:

- Indo-European
- Uto-Aztecan
- Austronesian
- Bantu

The authors then use statistical methods to reconstruct the most probable path of language evolution leading to these leaves, inferring the kind and direction of changes that occurred as well as functional dependencies between them.

What we like about it:

- Syntactic traits were used: unlike the (classical) cognate (phonology) based methods, this could be considerably enriched by coding all sorts of syntactic properties.

- It could help answer unresolved questions about the deep past of language history (because syntax operates in a more constrained variation space than the lexicon), perhaps even human evolutionary history.

- It could help discriminate among theories (models) of present day syntactic structures, allowable syntactic variation, syntactic change and language learnability questions.

What was interesting to us (in particular):
Are there linguistic invariants, e.g. is there coevolution of word order traits? which traits? how do they coevolve?

What was concluded:
Main Claim: "contrary to the generative account of parameter setting, the evolution of only a few word order features of languages are strongly correlated" and "contrary to the Greenbergian generalizations, most observed functional dependencies between traits are lineage-specific rather than universal tendencies."

Why we are skeptical[1]

- Granting the data, the universality claims were in fact not investigated.
- The notion of word order used is that of Dominant Word Order, a notion that has no currency within generative grammar.
- Questions about the data:
  - Granting the use of dominant order, the data used coming from WALS is of unreliable quality to start with.
  - There are some coding errors and interesting sub patterns suggesting that subsampling would be useful (e.g. to check whether the conclusions depended on a small sample of languages).

---

[1]cf. Koopman, Ryder and Sportiche's 2011 ms. "To find language universals, at least look for them" Comments on "Evolved structure of language shows lineage-specific trends in word-order universals"

The universality claims

- The authors conducted the functional dependencies test (testing for coevolution of word order traits) in a way that does not test for universal functional dependencies. We next explain why.

- Testing for universal dependencies can be done without any additional hypotheses and with (probably) only negligible extra computational time. We next show how.

Say we want to decide between two models M_1 and M_2.
Taking two word-order traits t_1 and t_2 (for example t_1: subject-verb order; t_2: demonstrative-noun order); the associated models are:

- M_1: t_1 and t_2 co-evolve universally
- M_2: t_1 and t_2 do not co-evolve

Here, we are assuming that either M_1 applies universally or M_2 applies universally (there is a third possible scenario, namely that t_1 and t_2 coevolve in some languages but not in others; the calculations below could be adapted to such cases).
When confronted with data D, the Bayes factor K is defined as

$$K = \frac{P[D|M_1]}{P[D|M_2]}$$

A large value of K is indication that the data favor model M_1.
A small value of K is indication that the data favor model M_2.

For each pair of functional dependencies, the authors have calculated four Bayes factors, one per language family. We thus have information about whether the data from each language family support model M_1 or model M_2.

Comparing these four Bayes factors does not give any information about the existence of a universal dependency. We would like to know whether the data as a whole support model M_1 or model M_2. We need to compute a single Bayes factor for each dependency to test whether it is a universal functional dependency.

One way is construct a (independently motivated genealogical) supertree including all languages and use the authors' method to directly compute a global Bayes factor. This unfortunately is hardly possible.

However, a global Bayes factor can nonetheless be calculated.

Let D_{IE}, D_A, D_B and D_{UA} the data restricted to the Indo-European, Austronesian, Bantu and Uto-Aztecan families respectively. Dunn et al. have shown how to compute $P[D_{UA}|M_1]$; it can be viewed as

$$P[D_{UA}|M_1] = \int P[D_{UA}|M_1, \theta] p_0(\theta) \, d\theta$$

where θ represents all the parameters of the model (θ is multidimensional), and p is the prior on θ. Let p_1 be the posterior on θ after this first step.

Then clearly

$$P[D_B|M_1, D_{UA}] = \int P[D_B|M_1, \theta]p_1(\theta)\, d\theta.$$

Similarly, if p_2 is the posterior after this second step, then

$$P[D_A|M_1, D_B, D_{UA}] = \int P[D_A|M_1, \theta]p_2(\theta)\, d\theta$$

leading to posterior p_3 after this third step, and finally

$$P[D_{IE}|M_1, D_{UA}, D_A, D_B] = \int P[D_{IE}|M_1, \theta]p_3(\theta)\, d\theta.$$

By using the posterior for one step as the prior for the next step, we can thus calculate the joint likelihood of all datasets. The same can be done for M_2, giving us everything that is needed to compute K. Of course, p_1, p_2 and p_3 cannot be computed exactly nor can they be written in closed form, but they can be sampled using Monte Carlo methods. This is enough to give an unbiased estimate of the relevant likelihoods.

This method allows the computation of a true Bayes factor to decide on the existence of universal functional dependencies.

WALS Dominant word order ($\neq$ from Greenberg's notion)

From Matthew Dryers's online supplement to WALS:

> "The expression 'dominant order' is used here, rather than the more common expression 'basic order', to emphasize that priority is given here to the criterion of what is more frequent in language use, as reflected in texts. **The reason for assigning priority to this criterion is that for most languages, this is the only criterion for which we have any relevant information** [our emphasis]
> ..... For **some** [our emphasis] languages, the classification of a language in this atlas is based on actual text counts.
> .....When a language allows both orders of adjective and noun, for example, grammars will often mention this but describe one order as the normal order or the more frequent order.

.... The rule of thumb employed is that if text counts reveal one order of a pair of elements to be more than twice as common as the other order, then that order is considered dominant, while if the frequency of the two orders is such that the more frequent order is less than twice as common as the other, the language is treated as lacking a dominant order for that pair of elements.

*.... For some languages, the classification on the map is based on a claim in the source that some order is basic **or** [our emphasis] that it is pragmatically neutral. In the absence of evidence to the contrary, I assume that these are also the dominant orders.*

.....If a grammar indicates that both orders of a pair of elements are possible, without stating that one is more common or without any comment suggesting that one order is more common, then the language will be shown on the map as having both orders without one being dominant.

....Of course, unless one examines a large number and a broad variety of texts, one cannot be sure that differences in frequency may not occasionally reflect the idiosyncratic properties of a particular set of texts. It is likely that in some cases, further text counts would lead to classifying a language differently.

In principle: Dominant order is "most frequent in language use as reflected in texts".

In practice: coded "dominant order" is a heterogeneous measure based on "the most frequent in language use as reflected in **texts**" (our emphasis as not much is known of these test and their relation to actually speech patterns) or on claims of "pragmatic neutrality" or on claims about "basic order".

Even for the few languages in which the strict notion of dominant order is used, an arbitrary "rule of thumb" is used, without any justification for such a cutting point.

Main additional problem: Dominant order yields arbitrarily binned data resulting in important loss of information.

Basic Order

In generative grammar, word order variation of a given set of items (e.g. V, O, S) is derived by postulating a **basic order** from which all orders are derived by word order permutation functions (aka **movement**).

Basic Order

- Empirically, in a given language (in a given proposition), this is based on **sets of fine grained observable patterns**
- Theoretically, it is coded as basic order + movement because two properties need to be explained (coded in the theory of movement):
    - not all patterns are possible
    - different existing patterns have different syntactic properties (constituent structures),
- From the point of view of word order traits coevolution, either basic order (rather than dominant order) could be used as defining traits, or (even better perhaps), the **sets of fine grained surface patterns** on the basis of which basic order is established.

Using Fine grained Word Order Patterns: why?

Hypothetical language: Glishen (semi-mirror image of English)
Adjective/Noun: balloon big
Numeral / Noun: balloons three
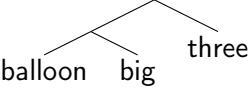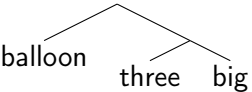
Reporting Glishen as: Adjective>Noun & Numeral>Noun
underreports variation as both Glishen1 and Glishen2 are possible
Adj/Num/Noun:
Glishen1: balloon big three
Glishen2: balloon three big

This also underreports structural differences:

Fine grained Word Order Patterns: Examples

Verb Object order
English: Pretty consistently VO –> Classified as VO
John **saw this movie**
...that John **saw this movie**

Dutch: both VO and OV –> reported as lacking dominant order
Jan **zag deze film** = John **saw this movie**
.. dat Jan **deze film zag** = ...that John **this movie saw**

The importance of syntactic structure

In Dutch
V O is a super pattern of V X O
O V is a super pattern of O Y V

X's and Y's are completely different:
X: subjects, weak pronouns, modal, temporal adverbs, etc...
Y: negation, modal particles (NOT subjects, NOT weak pronouns, etc...

Conclusion:
Not reporting both VO and VO leads to loss of information
Basic Order = OV; and VO is derived by "moving" V leftward
This conclusion is recoverable if we pay close attention to X and Y
in the set of allowed VXO or OYV patterns in Dutch.

What we would like to do about it: Short Term

Redo the study with better controlled, finer grained data:

- For each trait, code sufficient data in each language (e.g. sets of available patterns) so as not to underestimate existing variation

- Code the data for a (sufficiently large) subset of the relevant languages in SSWL:

- Inventory the set of a priori possible patterns (possible languages) and which ones are actually observed (existing languages)

- Formulate theories of possible variation (which subsets of patterns are allowed), and theories of possible change (from one subset to another)

- Use phylogenetic methods on these data to build trees, correlate with lexically build trees, explore patterns of change and coevolution and thus test these synchronic and diachronic theories.

What we would like to do about it: Long Term

As we do not know the atoms of (morpho) syntactic variation, we would like to further develop the open source database (SSWL) recording as many atomic properties as possible from the point of view of (morpho-) syntactic variation to answer the kind of questions we raise above in the domain of word order, in and many others.

What is coming next:

- Some examples
- The type of theories they lead to
- 2 Case studies

Starting with an illustration with word order
Example 1 (subcase of Greenberg's Universal 20)

- Consider combinations of 3 elements: "123"
- $\boxed{\text{Dem}_1 \text{ Num}_2 \text{ N}_3}$ or $\boxed{\text{Dem}_1 \text{ A}_2 \text{ N}_3}$

  (These are subcases of Greenberg's (63/66) Universal 20 (U20) (Dem Num A N)), which will be
  discussed as example 2)
- How many orders are in principle possible? $\boxed{3!=6}$
- How many neutral orders are observed? $\boxed{5/6}$
  (Neutral = without narrow /contrastive focus)
- $\boxed{\text{one pattern is unattested} \boxed{\text{*213}}}$
- Why? What's the theory?[2]

---

[2]cf. Cinque 05

Model: Theory take 1

- Structured basic order (Binary branching and Universal hierarchy)[3]

  Dem=1 Num or A = 2 N =3
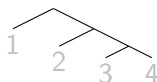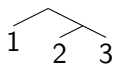
- Reordering:
    - derivations are strictly cyclic and obey the Extension Condition.
    - Moves constituents only;
    - Always includes the N;
    - Is leftwards only
    - No sub-extraction from specifiers[4]:
- Reordering operations yield observed orders ($\boxed{5/6}$ cf. next slide))

$\boxed{\text{The unattested order (*213) cannot be derived (Cinque 2005)}}$

[3]with Num $>$ A, cf. below

[4]This assumption is debated, and not made by Cinque 05

Derivation of attested orders

Dem=1 Num or A = 2 N =3, with a hierarchical structure (1) a.:
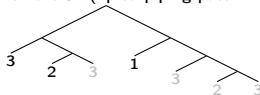
a.   1 2 3 (no reordering):
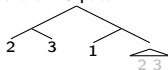

b.   132: Leftward movement of 3(=N) past 2:
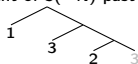
c.   312: leftward movement of 3 (past 1)
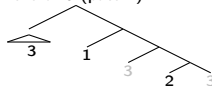
d.   321: Leftward movement 32 ("pied piping past 1

e.   231: Leftward movement of 23 past 1

Why is *213 unattested?

- This order would necessarily involve moving 2 without 3 to the left of 1
- This does not include 3 (=N), since all leftwards movements must contain the N.[5]
- Exclusion is not accidental: it is predicted

---

[5]For a slightly different theory which also gets this result, see Koopman and Szabolcsi (2000). For an evaluation of the different theoretical predictions, see Koopman 13 IGG and Penn colloquium slides

Expectations for 123

Patterns expected to occur ✓; (predicted gaps 0)

123	✓
132	✓
312	✓
321	✓
231	✓
213	0

- 5 possible patterns
- 1 impossible pattern *213.

U20 patterns generalize

- Since 00, 05: same patterns and restrictions are found not just in the domain of U20, but in many other domains.
    - Verbal complexes[6]
    - Morphology[7], "nanosyntax" Starke 2010, 2012, Caha, 2008, 2012
    - Numerous applications by Cinque:

attributive A	(1994), (2010)
A	2010
Adverbs	(1999)
circumstantial PPs	(2002, 2005)
Mood > Tense > Aspect	2008
etc.	

---

[6]Koopman & Szabolcsi 2000, Barbiers, 2005, Abels 2012
[7]Koopman 2005

123	✓	Malagasy..
132	✓	Bantu, Dutch..
312	✓	Bantu, Quechua "scope violations"
		Japanese, Korean (Koopman 2005)
321	✓	English [[red-en]-ed]] , Japanese....
231	✓	Malagasy, Bambara..
213	0	

An example from English morphology:

unlockable: 1 [3 2]] un_1 able $_2$ $lock_3$ [un [lock able]

unlockable: [[23] 1] able $_1$ un_2 $lock_3$ [[unlock] able]

---

[8] cf. Koopman 2003, 2005, Caha 2010, ..

Possible languages

A language does not necessarily exhibit only one pattern. It may exhibit several simultaneously.

Possible languages = Possible sets of patterns

1. Theory take 1 says nothing about which patterns (orders) may co-occur within a language.

2. Needed: Additional hypotheses about sets of patterns ($\rightsquigarrow$ Theory take 2).

3. This will lead to predictions about possible and impossible synchronic (and diachronic) variation.

This theory is highly restrictive

What will be shown first for 123 (combinations of 3 elements), and then for 1234 (combinations of 4 elements).

- Possible and predicted languages (ignoring structural hierarchy)
  1. For 123: Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 0.5$
  2. For 1234: Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 0.0005$
- Possible and predicted languages (taking structural hierarchy into account)
  1. For 123: Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 0.006$
  2. For 1234: Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 10^{-32}$

Possible adjustments

Possible and predicted languages:

- Recall: derivations are strictly cyclic (obey the extension condition)
- Allowing subextraction of a constituent containing an N from specifiers (from inside a left branch) increase the number of predicted orders and structures:
  1. For 1234: orders go from 13 to 14.
     Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 0.001$
  2. For 1234: possible structures go from 13 to 20.
     Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx 10^{-30}$
- But this overestimates the number of predicted languages because there are further constraints on which set of patterns are allowed to cooccur within a language (cf. infra **Topological Connectedness**)

- $S =$ set of possible order combinations: Cardinality of $S = 3! = 6$
- Number of possible languages (sets of patterns)
  $=$ (cardinality of power set of S) -1 (every language has at least one ordering):

$$S = P(S) = 2^6 - 1 = 63$$

- Number of allowed orderings/patterns: 5
- Number of allowed languages/sets of patterns: $2^5 - 1 = 31$
- Ratio: $\frac{predicted}{(in\ principle)\ possible}$ $\frac{31}{63} \approx \frac{1}{2}$

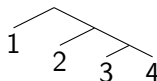# Possible and predicted languages (w/ hierarchy)

A trait (pattern) is an ordering of 123. For each order,
there are 2 possible (binary[9]) trees
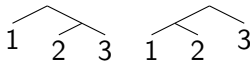
- S= set of possible orderings: Card $S = 3! = 6$
- $S^* =$ Set of possible structured ordering: $2 \times 3! = 12$
- Number of possible languages (sets of patterns) = Card $P(S^*)$
  (minus 1 : every language has at least one ordering)

$$\text{Card } P(S^*) = 2^{12} - 1 = 5096 - 1 = 5095$$

- Predicted possible orderings: 5 (each with a particular binary branching structure)
- Number of possible languages/sets of patterns: $2^5 - 1 = 31$
- Ratio: $\frac{predicted}{(in\ principle)\ possible}$ $\frac{31}{5095} \approx 0.006$
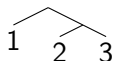
---

[9]Allowing more trees would further shrink the ratio.

Example 2: Extending to 1234 (or 12345 etc)

- So far, a subset of Universal 20 (orders of 3 elements: 123).
- Example 2: the full Universal 20 (orders of 4 elements)

  Dem=1 Num or A = 2 N =3   Dem=1 Num=2 A=3 N=4

- Possible, Predicted and Observed/Unattested languages

The full Universal 20: from Greenberg 63/66[10] to Cinque 05 (and beyond)

- *Prenominally*:
  The order of demonstrative, numeral, and adjective (or any subset thereof) conforms to the order Dem Num A
  "virtually" uncontested

  These three red balloons

- *Postnominally*
  The order of the same elements (or any subset thereof) conforms either to the order Dem Num A or to the order A Num Dem.
  In fact: many more attested orders post nominally. Any constraints? Hawkins, 83: no; Cinque 05: yes

---

[10]Extensively studied Hawkins 83, Croft &Deligianni 01, Rijkhoff 81...

The empirical generalizations (Cinque 05)

- These four basic elements: 1Dem, 2Num, 3 Adj, 4N
- Possible Combinations (4!=24)

Attested ✓; 0 Un-attested [11]			
1234	✓	1324	0
1243	✓	1342	✓
1423	✓	1432	✓
4123	✓	**4132**	0/??[12]
2134	0	2314	0
2143	0	2341	✓
2413	0	2431	✓
4213	0	4231	✓
3124	0	3214	0
3142	0	3241	0
3412	✓	3421	✓
4312	✓	4321	✓

---

[11] Frequency of patterns omitted

[12] Existence dubious but requires subextraction from specifiers

Given these four basic elements:

1Dem, 2Num, 3Adj, 4N.

the patterns $\boxed{\text{*2134, *3124, etc}}$

cannot be derived:

they would involve moving a constituent that does not contain N

Possible and predicted languages (w/o hierarchy) 1234

- S= set of possible order combinations: Cardinality of S= 4!= 24
- Number of possible languages (sets of patterns) = (cardinality of power set of S) -1 (every language has at least one ordering):

$$S = P(S) = 2^{24} - 1 = 25969215$$

- Number of allowed orderings/patterns: 13
- Number of allowed languages/sets of patterns: $2^{13} - 1 = 10192$
- Ratio: $\frac{predicted}{(in\ principle)\ possible} = \frac{10192}{25969215} \approx \frac{2^{13}}{2^{24}} \approx 2^{-11} \approx 0.0005$

Possible and predicted languages (w/ hierarchy) 1234

A trait (pattern) is an ordering of 1234. For each order, there are 5 possible (binary) trees.

- S= set of possible orders : Cardinality of S= 4!= 24

- S*=number of possible structured orders: 5x4!=120

- Number of possible languages (sets of orders/patterns) = (cardinality of power set of S*) −1 (every language has at least one ordering):

$$\text{Card } (P(S^*)) \approx 2^{120}$$

- Number of allowed orderings/patterns: 13

- Number of allowed languages/sets of patterns: $2^{13} - 1$

- Ratio: $\frac{predicted}{(in\ principle)\ possible} \approx \frac{2^{13}}{2^{120}} \approx 2^{-107} \approx 6.2 \times 10^{-33}$

Very highly restrictive Theory!

Theory take 2: Handling Variation

So far we have assumed that any subset of allowed orders is a possible language, but in fact:

1. Which subsets of patterns are attested? Which subsets of patterns are unattested?

2. What is the theory of allowed subsets?
   - Movement Theory (Strict Cyclicity, Locality, Triggers)
   - Coherence: Topological connectedness

3. How is a particular allowed subset characterized
   - Parameter of Pied Piping
   - Parameter of height of moving constituent

4. How does a language change from one subset to another
   - Change in Pied piping possibility
   - One step gain or loss in height of moving constituent

Theory take 2. Handling Variation: the facts

1. Which sets of patterns are attested?
   Which sets of patterns are unattested?
   - two case studies

Theory take 2. Handling Variation: Movement Theory

2. What is the theory of allowed subsets?
   - **Movement Theory**
     - Strict Cyclicity: movement is always to the top (a movement always create a substructure strictly including the substructure without this movement)
     - Locality: Movement is always stepwise
     - Triggers: Movement is of constituent containing a distinguised element (e.g. the Noun)

Theory take 2. Handling Variation: Subset Coherence

2. What is the theory of allowed subsets?
   - **Topological Connectedness**.
     If structures with movement of some trigger to both height p and q are in the subset, movement to height r is too for any r between p and q.

Theory take 2. Handling Variation: Parameters

3. How is a particular allowed subset (language) characterized?
   - Parameter of **Pied Piping**: movement of the trigger is allowed to pied pipe a constituent containing it. For example, N movement may or must carry A along if licit in principle.
   - Parameter of **Height** of moving constituent: the moving constituent may be required to reach a certain height. For example, the Noun may be required to be above Numeral.

Theory take 2. Handling Variation: Parameter Change

4. How does a language change from one subset to another
   - Change in Pied piping possibility: loss or gain of pied piping possibility or requirement
   - One step gain or loss in height of moving constituent: the height requirement on a moving constituent may increase or decrease one step

Example: the Height Parameter[1]

- Languages vary as to whether or how high a constituent containing the N moves up in the universal hierarchical structure.
- We code this as: "N$> X$",($N > 3$=Kurdish, $N > 2$ =Maasai or $N > 1$ = Kiyaka). No pied piping by N here.

(1)   Examples of 4123, 1423, 1243, 1234

---
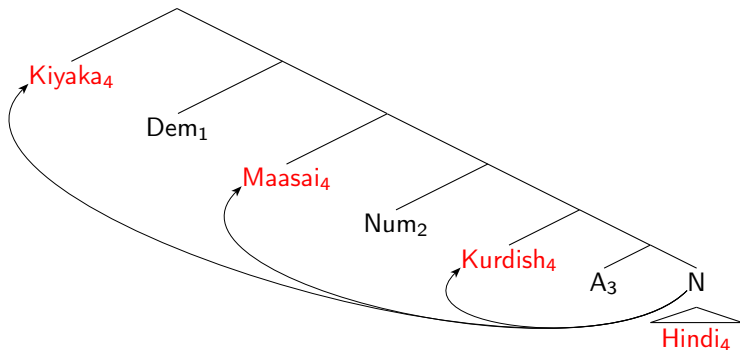
[1]Height of V/N movement, Pollock 89, ..

Pied Piping

From 1234 *to*	12 [43]	*to*	1[43]2
	these two [books blue]	to	these [books blue] two

Move [43] $\rightsquigarrow$ eg. Burmese ...
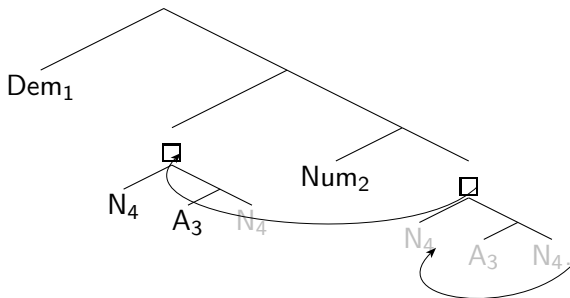
Here is the 1[43] 2 repeated..
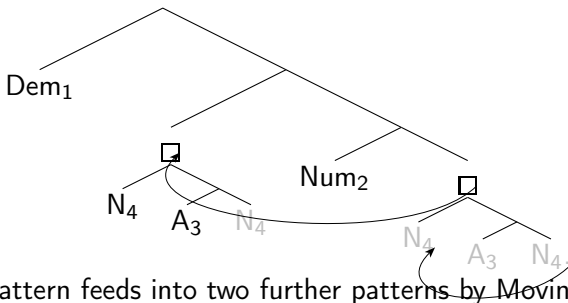

This pattern feeds into two further patterns by Moving a constituent which contains 4 (the N), with or without pied-piping.

- Move 43 leftward ⤳ [43] 1 2 = [N A] Dem Num (cf. Aghem, Grassfied Bantu,..)
- Move 43 pied piping 2, ⤳ [[43] 2] 1= [N A Num] Dem (cf. Gungbe, Kwa, ...)

Locality: Movement is local (stepwise)

- 4123. How does 4 come to precede 1?
  - Movement is local (stepwise).[2]
  - 4123 must have the following derivation:

(2) $1234 \rightsquigarrow 1243 \rightsquigarrow 1423 \rightsquigarrow 4123$
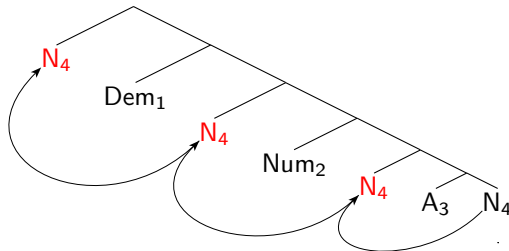
- Topological Connectedness: a language cannot allow e.g. $1243 \rightsquigarrow$ $\boxed{1423} \rightsquigarrow 4123$ without also allowing 1423.

---

[2]cf. Modern incarnation of the Head movement constraint

An example of expected and excluded patterns of variation

Expected languages allowing two word orders in the Universal 20 domain given:

- Locality: stepwise movement.

- Topological connectedness (surface patterns)

  Examples of expected patterns:
  N Dem Num A (4123) & Dem N Num A(1432)[15]
  N Dem Num A (4 [123]) & N Num A Dem ([432]1)[16]

  Examples of excluded patterns of variation:
  N Dem Num A(4123) & Dem Num N A (1243) [17]
  A N Num Dem (3421) & Dem Num N A (1243) [18] etc.

---

[15]N > Num; N may move > 1 Dem

[16]N > Dem, pied-pipes Num

[17]Excluded by Topological connectedness

[18]*Excluded by Local movement and Topological connectedness

Patterns and gaps: SSWL Case studies

SSWL

Syntactic Structures of the World's Languages

http://sswl.railsplayground.net/
http://terraling.com

- SSWL: open access, community based, "(expert) crowd sourced"

- Open ended and built to grow.

- Data: Coded as vectors aka "property definitions" ((ideally) written by community) typical di- or tridimensional (Yes/ No/ (NA)) but flexible (n-dimensional).

- Based on surface patterns/ "raw" data: no loss of information.

- Data values: set by "linguistic experts": native speaker linguists, linguists with deep familiarity with the language,..),
illustrated with examples, comments (where appropriate), references.

- Sophisticated search functionality.

- data can be downloaded in csv format

- Next generation: Terraling/SSWL. Collection of searchable databases. Flexible platform for linguists to tailor their projects. http://terraling.com

SSWL: current state (July 15 2013)

- languages: 207
- number of (vectors) properties: 93. (+ 40 in advanced stages of development)
- Contributors: 307
- Set properties: ≈ 13.000

SSWL case studies

- Check the theoretical predictions against sets of patterns in SSWL:
  Data from February (migrated to TerraLing page SSWL-0212 search of 2/27/13)
- Current word order properties: 30 vector Word order patterns of combinations consisting of 3 elements.
- Two (word order) case studies
  - Dem A N & Dem Num A
  - "SOV"

Expectations for single patterns

Table : Expectations for languages with a single pattern

123	✓
132	✓
312	✓
321	✓
231	✓
213	0

What is found: Languages with a single pattern for Dem A N or for Dem Num N)[19]

Table : Dem A N

	predicted	n/144
123	✓	38
132	✓	4
321	✓	42
312	✓	1
231	✓	3
213	0	0
		89/144

Table : Dem Num N

	predicted	n/103
123	✓	38
132	✓	4
321	✓	13
312	✓	2
231	✓	7
213	0	0
		59/103

213 not found

[19] http://sswl.railsplayground.net/ accessed on 1/31/2013. For languages with all 6 values set, excluding NA (applicable")
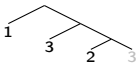
Expectations (subsets of 2)

- Derivations (from left to right, each cell on the left feeds into the cell on its right.)

123	⤳	132	⤳ 321
			⤳ 312
	⤳	231	

- Predicted patterns of subsets of 2 (allowed) patterns: $->$ $\boxed{7}$/10. Predicted gaps: $\boxed{3}$/10

- 7 predicted subsets of 2: 123 & 132; 132 & 321; 132 & 312; 321 & 312; 123 & 231; 321 & 231; 312 & 231.

- 3 predicted gaps

*	123, 312	violates topological connectedness
*	123, 321	violates topological connectedness
*	132, 231	not height characterizable

- +5 additional patterns should not occur: any of the 5 (licit) patterns in combination with the excluded 213.

  This is developed in table format on the next 4 slides.

How to read the table

- horizontal axis:
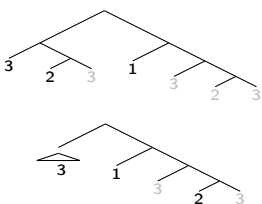 first row: from left to right, height parameters, patterns
  second row: height parameters, out of how many languages in total
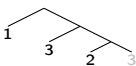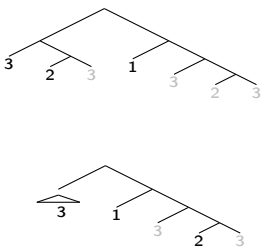
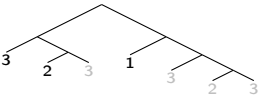|	height	height	height	Pattern Dem A N	Pattern Dem Num N
2>3	3>2			out of 144 lgs	out of 103 lgs
				15	5
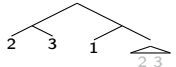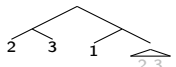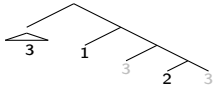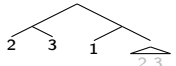
Table (part 1/3)

	height	height	height	Pattern Dem A N	Pattern Dem Num N
	2 >3	3>2		out of 144 lgs	out of 103 lgs
				15	5

	3>2	3>1	Dem A N	Dem Num N
			6	2
			2	2
		3>1		
			2	5

Table continued (part 3/3)

		2>1	Dem A N	Dem Num N
			3	4
		2>1 & 3>1		
			2	1
		3>1		
			0	0
			32/144	19/103

Theoretically Predicted Gaps (considering 2 patterns)

- **Confirmed** Any combination that violates topological connectedness: no skipped patterns

- Inconsistent with findings : Any pattern of 2 with *213 order: (i.e. A Dem N or Num Dem N).
  (This pattern is found, but an implicational search shows the 213 order always implies the existence of 123, see below)*discussed on next slides*

No violations of topological connectedness (predicted and found)

*123, 312	0	0
*123, 321	0	0
*132, 231	0	0

213 always implies 123 order (cf. also Cinque 05, footnote 2)

123, 213	7	5
132, 213	0	0
312, 213	0	0
321, 213	0	0
231, 213	0	0

- 213 is *never* the only pattern: 213 *always* implies 123. (cf. by sswl implicational search)

  For adjectives this correlates with the relative clause pattern: these languages all allow Rel Dem N order.)
  This suggests: As can enter the derivation as (reduced relatives), and as attributive (direct modification) As in these languages. This should correlate with interpretative properties: only intersective readings for Rel Dem N order (as proposed by Cinque 10)).

- leave open for future research 213 for Num Dem N (partially) correlates with RelClauses. It could also overlap with a partitive structure in the languages in question.

Remaining set of patterns...

- Consistent with predictions (not shown here)
- Patterns of *213 are found throughout:
  they always imply 123 order, and correlate with the Rel Dem N
  orders.

Prospects

- First case study support general research program.
- Develop and broaden the empirical base that allows testing theories of variation:
  Expand property definitions into relative clause patterns, different classes of As, and demonstratives, partitives, and definiteness/indefiniteness.
- further populate sswl
  use expert crowdsourcing to build a community database

Second case study: an application to S V O word order (data from SSWL)

- First question: What is the basic order of Merge? (What corresponds to 123)?

- SVO = 123? S=1, V=2, O=3? *No, this would predict VSO is excluded (= 213).*

- SOV = 123? S=1, O=2, V=3? *Yes*

- SOV as 123 is consistent with current theory [20]
  – objects (specific, definite) of transitive verbs occur well above the merge position of V, most likely as the result of internal Merge (movement).
  –Subjects (definite specific) of transitive verbs end up merged above the surface position of O (numerous arguments).

- ⤳ next slides a brief look at single patterns and patterns of 2 elements, with SOV as 123.

---

[20] And widely assumed: Givón 79, Gell-Mann & Ruhlen (11), *The origin and evolution of word order*, (and (possibly) Langus and Nespor 2010).

SOV Results: single patterns (89/164 languages

123	SOV	✓	19
132	SVO	✓	58
312	VSO	✓	10
321	VOS	✓	0
231	OVS	✓	1
213	OSV	*	0
			89/164

- No language with just VOS (so far).[21].
- No *213= OSV As predicted[22]

---

[21]Dryer (Wals) out of 1377 lgs: 25 VOS dominant languages; so far only 3 of these are entered in SSWL: all 3 are VOS *and* VSO

[22]Dryer lists 4 languages as "dominant" OSV in Wals; at least one language (Tobati, Donahue) seems to have neutral SOV as well, judging from the grammar making this potentially another case where 213 would imply 123.
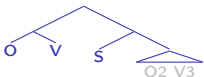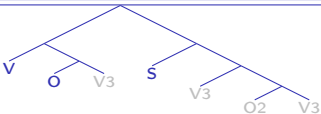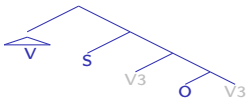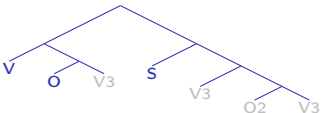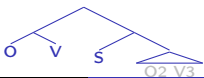
Sets of expected patterns of 2 [23]

O2> V3	height V3> O2	height	n languages 164
			22
	V3>O2	V3>S1	
			4
			3

---

[23]4/7 attested

Table SVO *continued part2 ..*

		O2>S1		
			0	
		V3>S1		
			7	
		O2>S1		
			0	

Table continued (part 3/3)

		3>1		
			0	0

- 4/7 attested

Gaps (sets of 2 patterns.. all expected, except for 123 and 213)

123, 213	SOV, OSV	0
132, 213	SVO, OSV	4
312, 213	VSO,OSV	0
321, 213	VOS,OSV	0
231, 213	OVS,OSV	0
123, 312	SOV, VSO	0
123, 321	SOV, VOS	0
132, 231	SVO,OVS	0

- Unexpected: SVO (132) and OSV 213: found where? Hanga, Palue, Thai, Dholuo.
- Possibly topicalization of object? cf Dholuo example

  what happened? Dholuo (Nilotic):

  Mary(O) dog(S) bite.
  a dog bit Mary/ Mary was bitten by a dog.
  appears to depend on relative animacy and definiteness

Conclusion Second case study

- General theory can be applied to quite coarse variation, and connects well to general literature on this topic. (See appendix for a further brief evaluation of Gell-Mann& Ruhlen 11's data on the evolution of word order.)
- problematic cases should be further investigated
- Future expansion in depth and breath should lead to a much better empirical testing ground

Conclusion

Short term plan: redo Dunn et al (11) with finer grained data

1. Inventory the attested/allowable sets of patterns in a language
2. Formulate theories regarding why these sets are allowed and not others, and formulate theories of possible change (from one set to another).
   Put the theories to test on available data SSWL (2/12)
3. Code the data (sets of patterns) for a (sufficiently large) subset of the relevant languages in SSWL.[24]
4. Use phylogenetic methods to evaluate patterns of co-evolution of properties thus these synchronic and diachronic theories, with sufficiently fine grained data.

---

[24]http://sswl.railsplayground.net/

Long term prospects

Since we do not know the atoms of (morpho) syntactic variation, we would like to further develop the open source database (SSWL) recording as many atomic properties as possible from the point of view of (morpho-) syntactic variation, to answer the kind of questions we raise above in the domain of word order.

This is a community enterprise: hence our plea to use, support, and help develop SSWL/Terraling.

Thanks!

The members of the current SSWL/Terraling team
Dennis Shasha, NYU. System architect
Marco Liberati, system developer
Chris Collins and Richard Kayne creators and consultants

The property authors who wrote the property definitions that generated the data in this presentation
Andrea Cattaneo, Chris Collins and Jim Wood
Cristina Guardiano and Hilda Koopman

UCLA undergraduate research assistants: Hannah Kim, Arwa Rangwala (winter 13)

The SSWL contributors and collaborators who contributed their time and shared their insights on their
languages for the U20 data under discussion. For the full list of contributors see
http://sswl.railsplayground.net/

Appendix. More about the change from SOV

- Figures from Gell-Mann & Ruhlen 2011: the evolution of word order
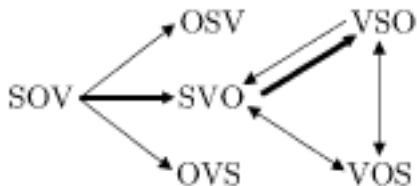
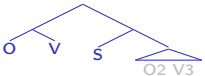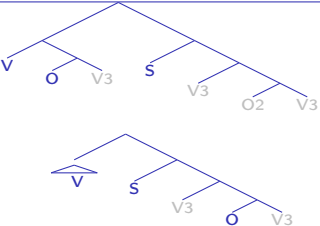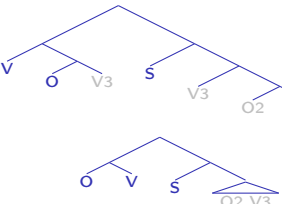# Languages with two patterns of dominant order, Gell-Mann & Ruhlen 11

Table 1. Languages with mixed word order

SOV/SVO	46
SVO/VSO	24
VSO/VOS	17
SVO/VOS	11
SOV/OVS	9
SOV/OSV	6
SVO/OVS	4
SOV/VOS	2
SOV/VSO	2
VOS/OVS	2
SVO/OSV	1
VOS/OSV	1

- cf. criticism of dominant order..(not known if these are patterns of 2 or more)
- A quick look how this fare w.r.t. the theory developed in this paper: 12 listed patterns. see next slides

Sets of patterns of 2: Gellman& Ruhlen's data for languages with 2 patterns of dominant order

O2> V3	height V3> O2	height	n languages
			46
	V3>O2	V3>S1	
			11
			24

Table continued part 2

Table continued (part 3/3)

		3>1		
			0	0

- 4 of the predicted 7 are attested...

Gaps, sets of 2 patterns.. Gellman & Ruhlen ..

- The following are predicted to be 0, but 5 are found in their table. These should all be further investigated.

- Important for testing the predictions: include all neutral orders, not just dominant orders.

- As in the SSWL case study, the interaction with topicalization of the object/ "passivization" (SVO 132 and OSV 213) should be must be looked into further.

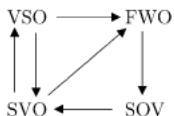| 123, 213	SOV, OSV	0
132, 213	SVO, OSV	1
312, 213	VSO,OSV	6
321, 213	VOS,OSV	1
231, 213	OVS,OSV	0
123, 312	SOV, VSO	0
123, 321	SOV, VOS	2
132, 231	SVO,OVS	4

Venneman's evolution of word order

Fig. 2. Possible word order changes (Vennemann).