UPPSALA
UNIVERSITET

Swiss National
Science Foundation

Multilevel phylogenetic inference of harmony in Indo-European

Yingqi Jing, Joakim Nivre and Michael Dunn

13 July, 2024

Background

It has long been observed that languages tend to order the grammatical head and its dependents in a consistent way (Greenberg 1963; Hawkins 1983; Dryer 1992), e.g., VO languages tend to be prepositional while OV languages tend to be postpositional (VO → Prep & OV → Postp).

Background

Over the past decades, new empirical findings and competing theories have constantly been advanced to revisit and explain the syntactic harmony.

Background

Over the past decades, new empirical findings and competing theories have constantly been advanced to revisit and explain the syntactic harmony.

- **Functional theories:** consistent head ordering can facilitate language processing, production and learning (Hawkins 1983; Culbertson, Smolensky, & Legendre 2012)

Background

Over the past decades, new empirical findings and competing theories have constantly been advanced to revisit and explain the syntactic harmony.

- **Functional theories:** consistent head ordering can facilitate language processing, production and learning (Hawkins 1983; Culbertson, Smolensky, & Legendre 2012)
- **Cultural evolution:** Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021)

Background

Over the past decades, new empirical findings and competing theories have constantly been advanced to revisit and explain the syntactic harmony.

- **Functional theories:** consistent head ordering can facilitate language processing, production and learning (Hawkins 1983; Culbertson, Smolensky, & Legendre 2012)
- **Cultural evolution:** Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021)
- **Diachronic origins:** many word order universals can be independently motivated by the grammaticalization processes of syntactic change (Bybee 1988; Cristofaro 2017)

Research questions

It still remains an open question whether there is any systematic constraints of syntactic harmony in language evolution. To better understand this issue, we make a first step towards testing the general hypotheses on the evolution of harmony on corpus data from Indo-European languages.

(1) How can we model the evolution of word order harmony with corpus data of diverse languages?
(2) Is there any systematic evolutionary bias towards harmony in the history of Indo-European, when compared to different random baselines?

Universal Dependencies and Indo-European phylogenies

- 43 Indo-European language corpora from Universal Dependencies version 2.12 (Zeman et al. 2022)

Universal Dependencies and Indo-European phylogenies

- 43 Indo-European language corpora from Universal Dependencies version 2.12 (Zeman et al. 2022)
- 11 dependencies between lexical categories (noun, verb, adjective & adverb)

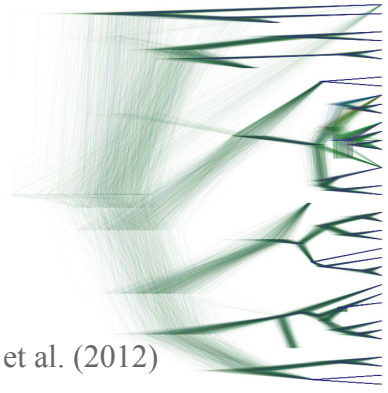
verb→'nsubj'→noun (*the man went away*)	*adjective*→'advmod'→adverb *(very good)*
verb→'obj'→noun (eat the apple)	verb→'advmod'→adverb (*walk slowly*)
verb→'obl'→noun (*finish the work [before the weekend]*)	noun→'acl'→verb (*the man [you love]*)
noun→'amod'→adjective (*a nice shirt*)	verb→'advcl'→verb (*he was happy [when I talked to him]*)
noun→'nmod'→noun (*his mother's friend*)	verb→'ccomp'→verb (*he said [that he knew the man]*)
noun→'advmod'→adverb (*only one choice*)	

Universal Dependencies and Indo-European phylogenies

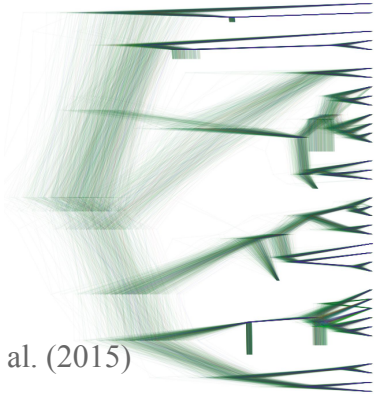
- 43 Indo-European language corpora from Universal Dependencies version 2.12 (Zeman et al. 2022)
- 11 dependencies between lexical categories (noun, verb, adjective & adverb)
- 3 sets of Indo-European phylogenies from the literature

Universal Dependencies and Indo-European phylogenies

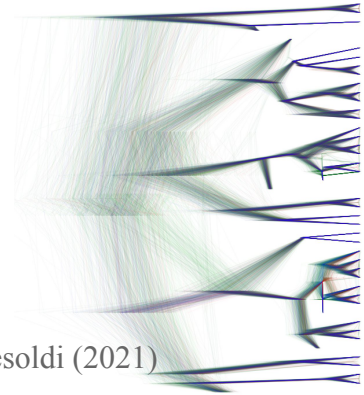
- 43 Indo-European language corpora from Universal Dependencies version 2.12 (Zeman et al. 2022)
- 11 dependencies between lexical categories (noun, verb, adjective & adverb)
- 3 sets of Indo-European phylogenies from the literature



Bouckaert et al. (2012)



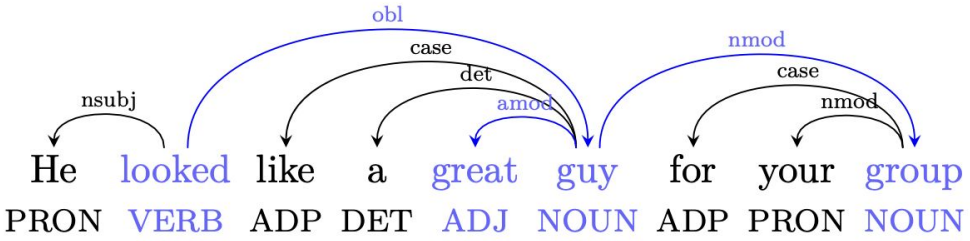
Chang et al. (2015)



Dunn & Tresoldi (2021)

Measuring syntactic harmony

We measure harmony by counting pairs of dependencies that co-occur in the same direction in a sentence.



word order pairs	Harmony	Disharmony
VObl & NGen	1	0
VObl & AdjN	0	1
NGen & AdjN	0	1

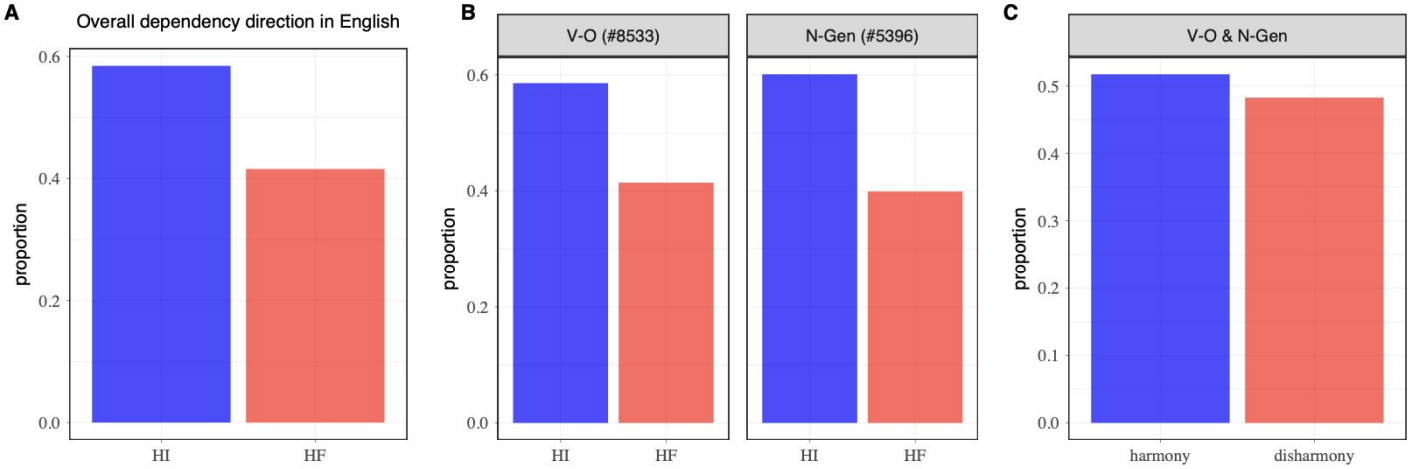
These raw counts will be entered into a Bayesian binomial model to estimate the probabilities of harmony and disharmony, while incorporating the uncertainty due to differences in frequencies and corpus sizes.

Random baselines

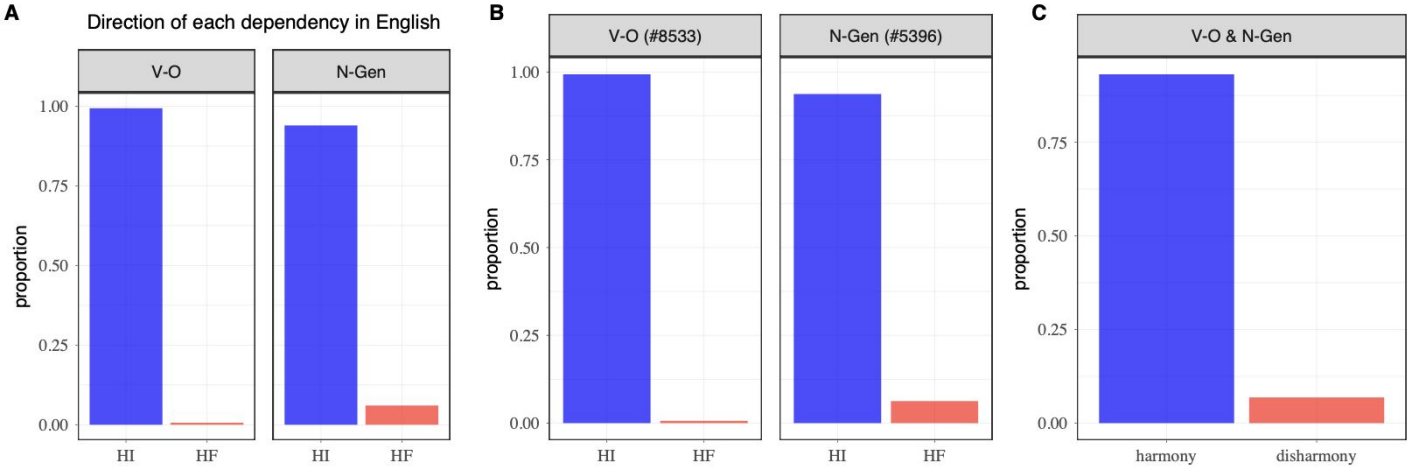
In order to measure the additional constraints of cross-category harmony in real utterances, we need to control for the base distribution of each word order in a language. For this, we introduce two random baselines.

- **Random baseline 1:** we randomly draw an order for each dependency type while holding constant the overall head direction in a language
- **Random baseline 2:** we keep unchanged the order of each dependency type in a language

① Random baseline



② Random baseline



Multilevel phylogenetic model

We developed a novel multilevel phylogenetic Continuous-time Markov Chain model to investigate the evolutionary rates towards harmony vs. disharmony across 55 pairs of word orders in Indo-European (Stan Development Team 2022).

Multilevel CTMC model:

$$\text{tips} \sim \text{TreeLikelihood}(Q, \tau, \pi)$$

$$Q_n = \underbrace{\alpha_0 + \beta_0 * \text{transitions}_n}_{\text{fixed effects}} + \underbrace{\alpha_{\text{type}[n]} + \beta_{\text{type}[n]} * \text{transitions}_n}_{\text{random effects}}$$

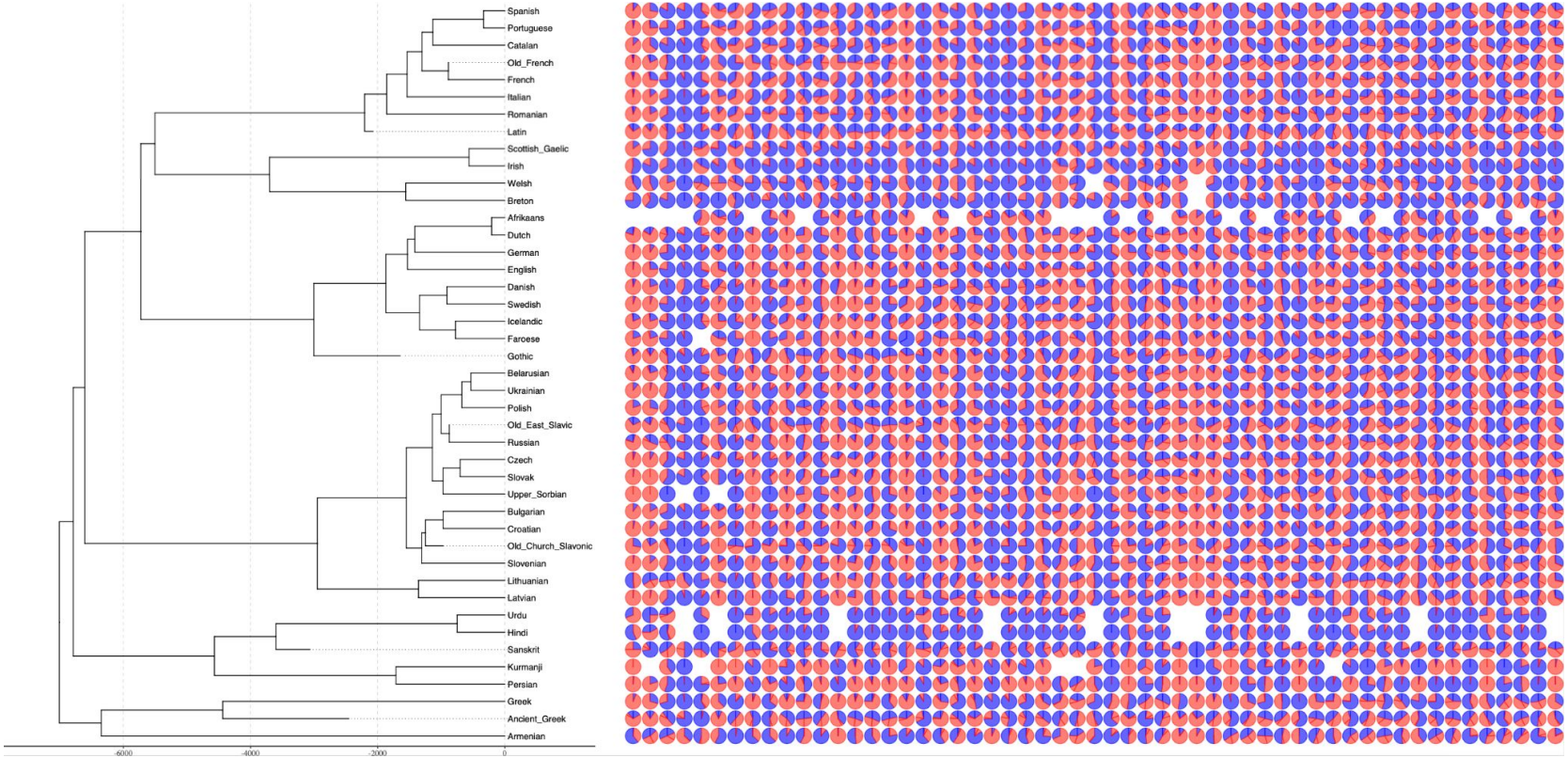


Figure: Probabilistic distributions of pairwise word order combinations (blue: harmony and red: disharmony) mapped onto the summary phylogeny of Indo-European from Bouckaert et al. (2012)

Results

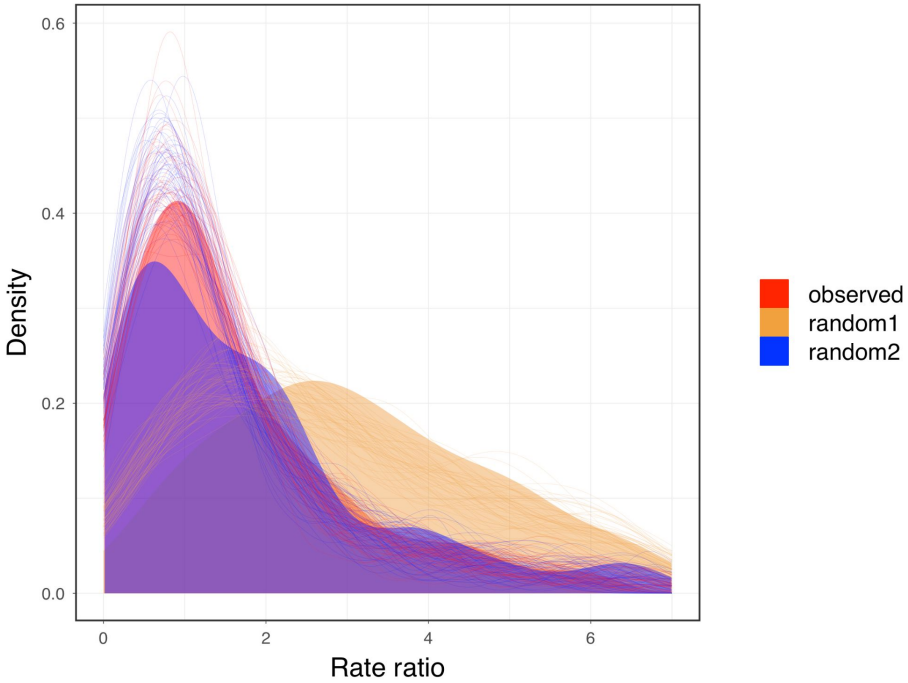


Figure: Posterior rate ratio of harmony to disharmony from the multilevel phylogenetic model.

- Our results reveal no overall differences in the estimated rate ratios for harmony between observed and random baselines.
- There are broad overlaps between observed and the second baseline, suggesting not much room left for cross-category harmony once individual word orders are held constant.
- We also observe a consistently weaker evolutionary bias towards harmony, when compared to the first baseline.

Rate ratio for pairs of orders

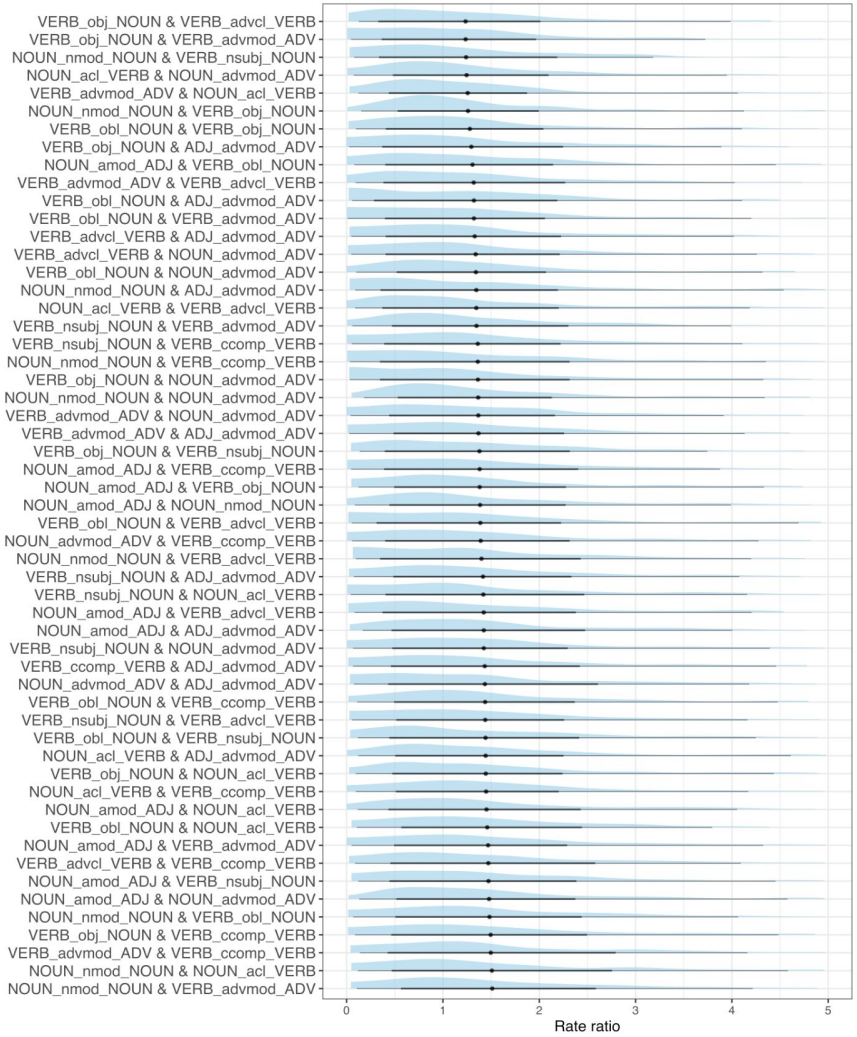


Figure: Distribution of posterior rate ratio for individual pairs of word orders

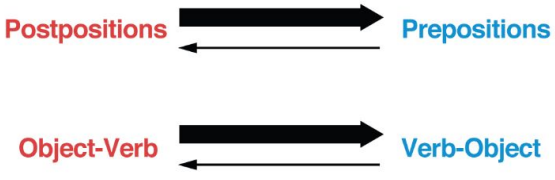
Conclusions

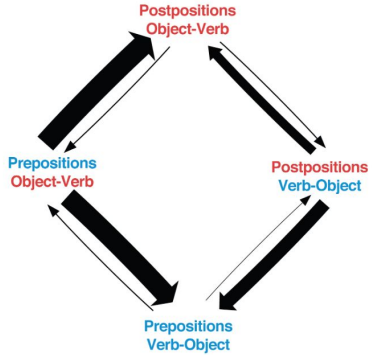
- Using 43 dependency-annotated corpora and Bayesian multilevel phylogenetic inference, we test the selective forces of harmony in language change against random baselines in Indo-European.
- Our results do not support the functional motivations for harmony, instead, we suggest that word order universals might emerge as a side-effect of word order rigidity in language evolution.
- In contrast to previous work that suggests a general head-initial or head-final preference, we show that word orders seem to evolve towards a more mixed configuration at least in Indo-European.

Thanks a lot!

Hypothesis 2: cultural evolution

Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021; Hartung, Jäger et al. 2022 for different positions)

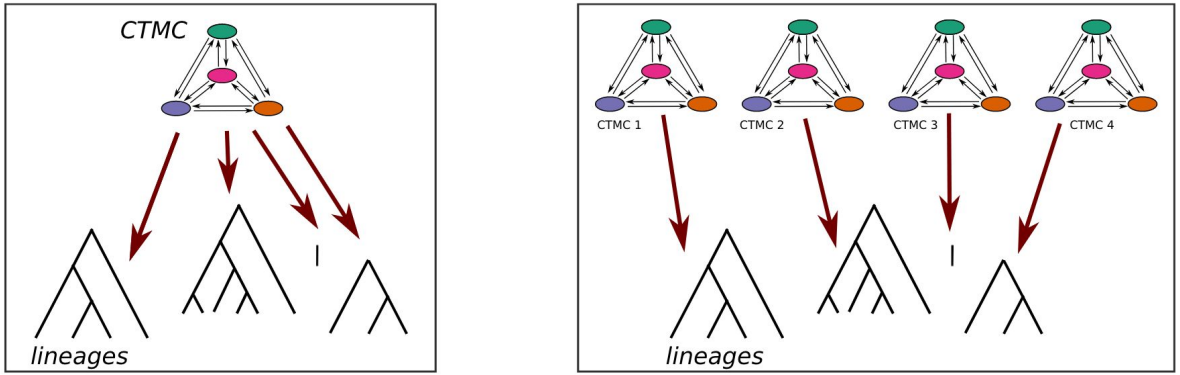




Dunn et al. (2011)

Hypothesis 2: cultural evolution

Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021; Hartung, Jäger et al. 2022 for different positions)



Jäger & Wahle (2021)

Hypothesis 2: cultural evolution

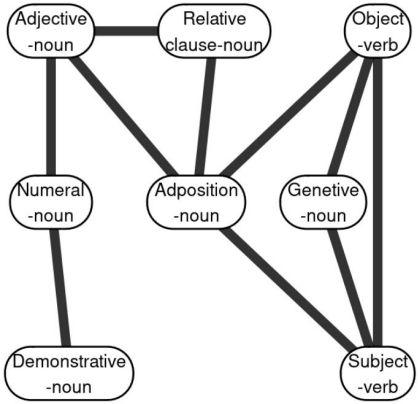
Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021; Hartung, Jäger et al. 2022 for different positions)

$$x \sim Binomial(p)$$

$$logistic(p) \sim MultiNormal(a, V)$$

$$V = R \otimes C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes \begin{pmatrix} t_1 & t_{12} \\ t_{12} & t_2 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 \cdot t_1 & \sigma_{12} \cdot t_1 & \sigma_1^2 \cdot t_{12} & \sigma_{12} \cdot t_{12} \\ \sigma_{12} \cdot t_1 & \sigma_2^2 \cdot t_1 & \sigma_{12} \cdot t_{12} & \sigma_2^2 \cdot t_{12} \\ \sigma_1^2 \cdot t_{12} & \sigma_{12} \cdot t_{12} & \sigma_1^2 \cdot t_2 & \sigma_{12} \cdot t_2 \\ \sigma_{12} \cdot t_{12} & \sigma_2^2 \cdot t_{12} & \sigma_{12} \cdot t_2 & \sigma_2^2 \cdot t_2 \end{pmatrix}$$



Hartung et al. (2022)

Hypothesis 3: diachronic origins

Many word order universals can be independently motivated by the grammaticalization processes of syntactic change (Bybee 1988; Collins 2012; Cristofaro 2017).

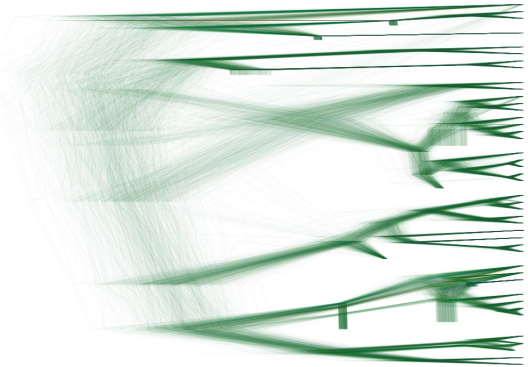
(2) Finnish (N → Postp)

poja-n **kansa**-ssa		poja-n **kanssa**
boy-Gen company-IN	→	boy-Gen with
'with the boy'		

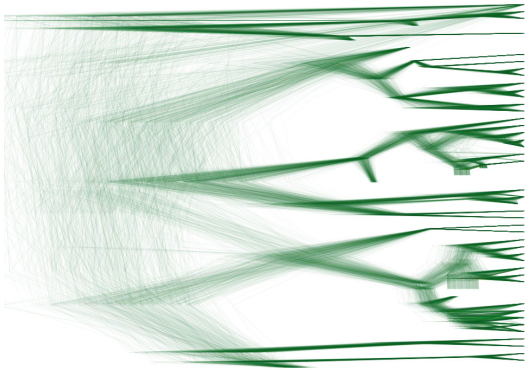
Aristar (1991: 6)

Universal Dependencies and Indo-European phylogenies

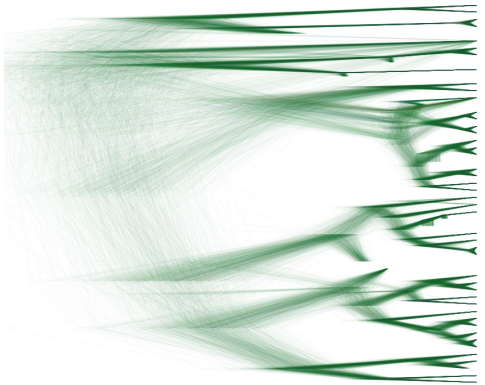
- 54 Indo-European language corpora from Universal Dependencies version 2.14 (Zeman et al. 2024)
- 12 dependencies between lexical categories (noun, verb, adjective & adverb)
- 3 sets of Indo-European phylogenies from the literature



Chang et al. (2015)



Dunn & Tresoldi (2021)



Heggarty et al. (2023)

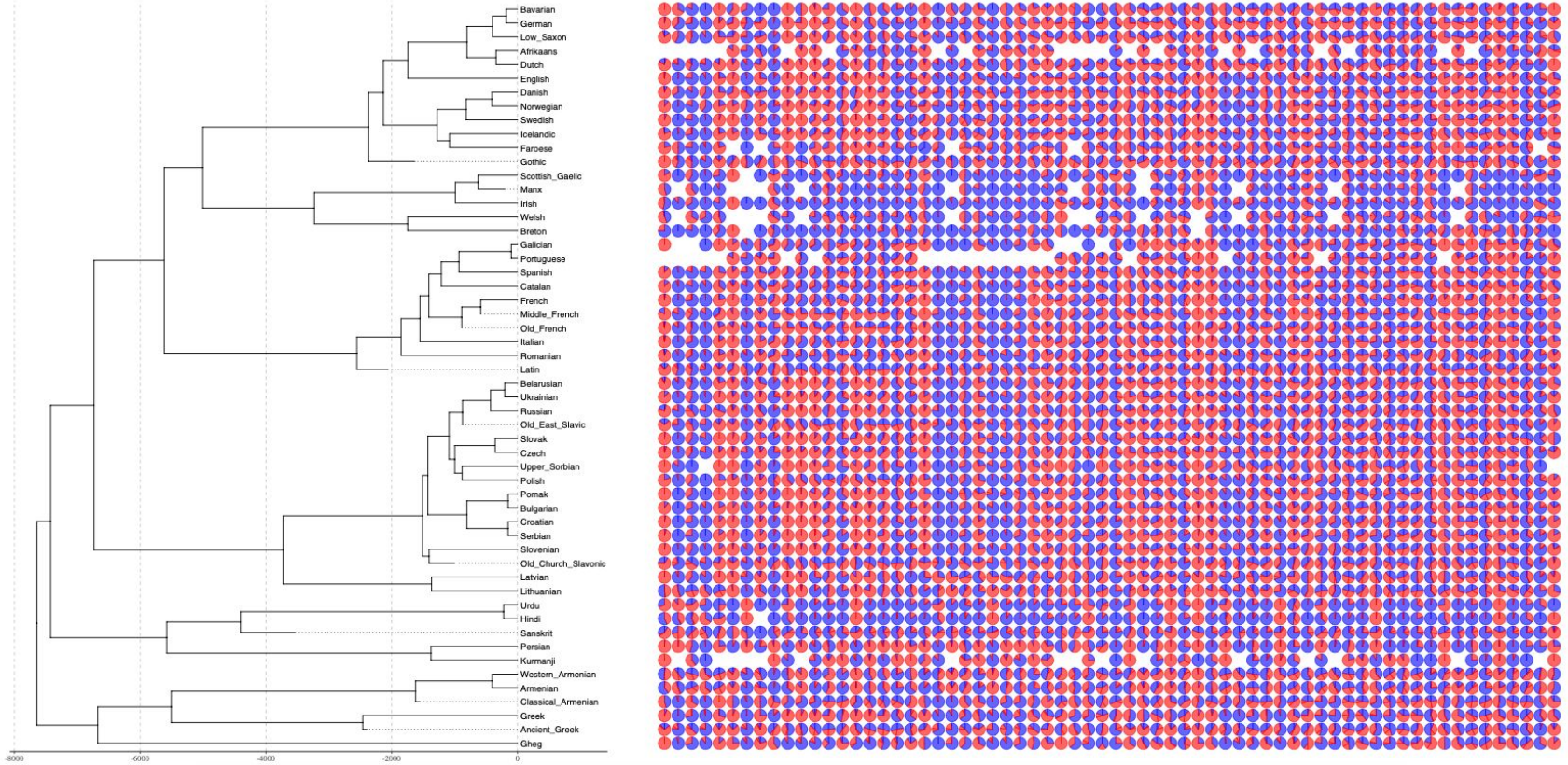


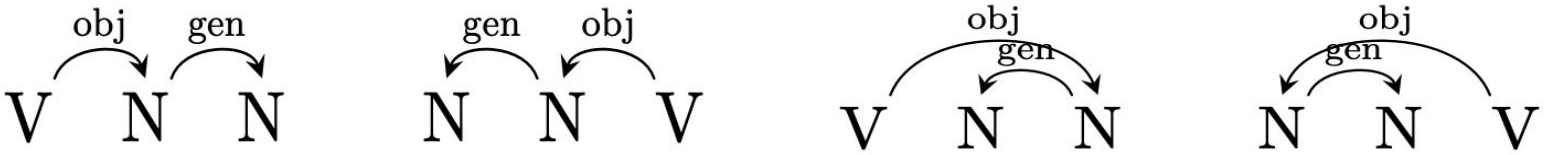
Figure: Probabilistic distributions of pairwise word order combinations (blue: harmony and red: disharmony) mapped onto the summary phylogeny of Indo-European from Heggarty et al. (2023)

Next steps

- Global phylogenetic inference while incorporating family-specific rate variation
- Integrating geographical information (language contact) into the model

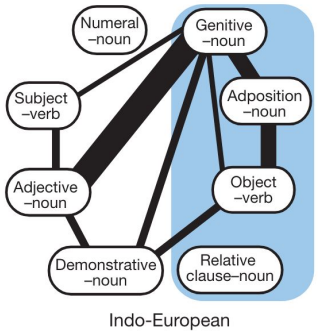
Hypothesis 1: functional theories

Consistent head ordering can facilitate language processing, production and learning (Hawkins 1983; Culbertson, Smolensky, & Legendre 2012; Hahn, Jurafsky, & Futrell 2020)

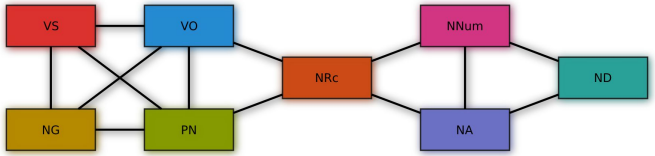


Hypothesis 2: cultural evolution

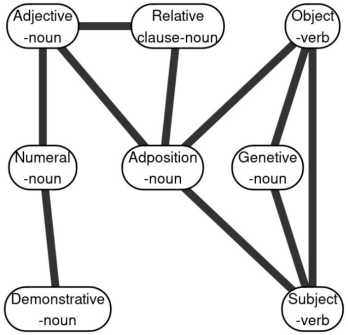
Greenbergian generalizations reflect lineage-specific rather than universal patterns, which are primarily driven by cultural evolution (see Dunn et al. 2011; Jäger & Wahle 2021; Hartung et al. 2022 for different positions)



Dunn et al. (2011)



Jäger & Wahle (2021)



Hartung et al. (2022)

Hypothesis 3: diachronic origins

Many word order universals can be independently motivated by the grammaticalization processes of syntactic change (Bybee 1988; Collins 2012; Cristofaro 2017)

(1) Hakka（V → Prep）

Gia ba **bun** yi kiu tien gi → Gia ba bun yi kiu tien **bun** gi
his father gave one CL field him his father gave one CL field to him
'His father gave a piece of field to him.'

Lai (2001: 141)