An empirical and quantitative reassessment of morphological autonomy

Sandra Auderset
University of Bern, Switzerland

Adam J.R. Tallman
Friedrich-Schiller-University Jena, Germany

24.10.2024

Morphological autonomy

refers to the thesis that morphology follows different principles from syntax / that it is autonomous from syntax and cannot be subsumed under it

> "the structure of the word form must be supplied by statements of a wholly morphological nature" (Matthews 1972: 107)

- distinction not discrete, i.e. whatever properties support the 'morphological nature' of a phenomenon are also found in syntax but to a lesser degree
- → basically a statistical argument

Morphological autonomy and wordhood

- statistical arguments for MA presume “fuzzy words” – two positions:

> “No criterion [for wordhood] is either necessary or sufficient,... But they are relevant insofar as, in particular languages, they tend to coincide”
>
> (Matthews 2002: 274)

“In order to show that a fuzzy concept of a word is theoretically significant one would have to demonstrate that grammatical units are not randomly distributed over the continuum between fully bound and fully independent units, but that they cluster significantly” (Haspelmath 2011)

Outline

1. Background on morphological autonomy and the morphology-syntax divide and research questions
2. Languages, Variables, & Coding
3. Correlations as a measure of boundary strength
4. Random Forests for assessing classification variability
5. Clustering for assessing the morphology-syntax distinction
6. Conclusions and outlook

Introduction: What is morphological autonomy?

Morphological autonomy (MA):

Morphological phenomena follow different principles of organization from syntax.

which is a different concept than

Word-based morphology:

Morphology refers to the organization of formatives and/or morphemes word-internally.

Introduction: Perspectives on MA

Blevins (2006: 555): *"In some systems, it is true that formatives may realize stable properties in all of the contexts in which they occur. Yet this can be seen to be a **limiting rather than a normative case**, and in many systems it is only recently morphologized formatives that can be described in this way."*

Haspelmath (2011: 63): *"In the first hypothetical situation (clustering distribution), there are three clearly discernible clusters. If the dimension along which the units differ (the boundedness scale) can be quantified, the clustering can be demonstrated by **statistical techniques**. There are intermediate cases between the clusters of affixes, clitics, and independent words, but these are few and just exceptions to the rule."*

Maiden (2004: 140): *"(...) autonomous morphological structure may be present even at the level of simple, linear, formative in word structure, and therefore potentially present cross linguistically, given that **all languages possess morphological structure of this kind**."*

Introduction: Perspectives on MA

- Encapsulation (Lexical integrity)
- Head-dependent order (Morphotactics)
- Types of operations (Notation)
- Morphological status / wordhood criteria (boundedness, freedom…)
- Deviations from biuniqueness (morphological complexity)

Deviations from biuniqueness

- common to claim that morphology is autonomous by virtue of displaying deviations from biuniqueness (cf. Matthews 1991)
- assumption:
  - syntax: one to one relation between meaning and form
  - morphology: not so much

Pattern	Morphemic	Cumulative	Extended	Empty	Zero
Properties	P	$P_I \ldots P_n$	P	—	P
	|	|	|		
Morphs	μ	μ	$\mu_I \ldots \mu_n$	μ	—

FIGURE 3.3 Types of exponence (cf. Matthews 1991: 170ff)

Deviations from biuniqueness: Example

paradigm (indicative present) of the verb ‘come’ in Bernese German

1SG	xʊmə
2SG	xʊnʃ
3SG	xʊnt
1PL	xœmə
2PL	xœmət
3PL	xœmə

1 form to many meanings: -ə

1 meaning many form: xʊm ~ xʊn ~ xœm

Three problems from a variationist approach

1) Boundary strength problem:

Languages may vary in the degree to which morphology and syntax are distinct. Some languages might display more indeterminacy than others.

2) Composition problem:

Languages may differ with respect to how the distinction between morphology and syntax is made.

3) Architecture problem:

Languages may vary with respect to whether morphology and syntax are distinct at all.

Languages, Variables, Coding

Study design

Language sample

- 7 languages from southwestern Amazonia
  - no inflection classes → typical arguments for MA based on paradigm complexity do not apply
  - described as displaying ‘syntax-like’ morphology → ideal test case for architecture problem
  - grammars of Amazonian languages often have many clitic and indeterminate categories
  - author expertise (Tallman 2018) & previous qualitative study (Tallman & Epps 2020)
- Central Alaskan Yupik for comparison because it is claimed to be canonically polysynthetic

Language sample

Language	Glottocode	Family	Source	Data points
Movima	movi1243	Isolate	Haude (2006)	153
Wãnsöjöt [Puinave]	puin1248	Isolate	Girón Higuita (2008)	99
Tariana	tari1256	Arawak	Aikhenvald (2003)	119
Ashéninka [Alto] Perené	ashe1272	Arawak	Mihas (2015)	98
Chácobo	chac1251	Pano	Tallman (2018a)	96
Cavineña	cavi1250	Takana	Guillaume (2008)	56
Hup	hupd1244	Naduhup	Epps (2004)	65
Central Alaskan Yupik	cent2127	Eskimo-Aleut	Miyaoka (2012)	81

Variables: Overview

we only code closed class items – for each item, we determine:

Variable	Name	Description	Type
Boundedness	FREE	Can the morph stand on its own as an (elliptical) utterance?	Binary
Interruptability	INTERone	Can the morph be separated from the verb/noun/adjective root by a free form?	Binary
Fixedness	PRfixed	Does the morph display a fixed order with respect to the verb/noun/adjective root?	Binary
Coding elaboration	CODelab	Does the morph display inflectional elaboration independent of the base with which it combines semantically?	Binary
Prominence projection	PRM	Does the morph always/sometimes/never project its own stress domain?	Ordinal
Exponence complexity	EXPcomplex	A metric that aggregates various types of deviations from biuniqueness that a morpheme can display	Continuous

Boundedness and interruptability

- Can the element stand alone as a complete utterance?
  - yes = free
  - no = bound

Chácobo:

tsaya-ʔaka =yáma=tɨkɨ(n)=ʔitá=kɨ

see-PASS. =NEG=again=REC.PST=DEC:PST

‘He was never seen again.’

a.	**tsaya* (intended: ‘see’)	(but *tsaya=kɨ* ‘s/he saw’)
b.	**-ʔaka*	(coded: bound)
c.	**=tɨkɨ(n)*(intended: ‘again’)	(coded: bound)
d.	**=ʔitá* (intended: ‘recently’)	(coded: bound)
e.	**=kɨ*	(coded: bound)
f.	*(=)yáma* ‘there is nothing/no one ’	(coded: free)

Interruptability/contiguity

- Can the element be interrupted from its head/host by a free form?
  - yes = interruptable
  - no = contiguous

Chácobo

a. ***tsaya****-ʔaka* *honi* *siri* *=yama=tɨkɨ (n)=ʔitá=kɨ*
see-PASS man old =NEG=again=REC:PST=DECL:PST
'The old man was not seen again yesterday.'

b. ****tsaya*** *honi* *siri* ***-ʔaka*** *=yama=tɨkɨ (n)=ʔitá=kɨ*
see man old -PASS =NEG=again=REC:PST=DECL:PST
'Intended: The old man was not seen again yesterday.'

Asháninka Perené

karini-taki *incha-**panki***
smooth-INTNS plant-CLT:long.rigid
'The wood planks are very smooth.'

Fixedness/Variable order

- Does the element display a fixed with respect to the head/host?
  - yes = fixed
  - no = variable

Hup

a. *dĕh* *wɔ́ç-ɔ́y*
water boil-DYN
'The water is boiling' (Epps 2004: 517)

b. *pĕd* *dĕh* *d'oʔ-wɔ́ç-ɔ́y*
Ped water take-boil-DYN
'Ped is boiling water' (Epps 2004: 517)

c. *yág* *ʔãh* *d'óʔ-óy* *ʔãh* *g'et-ni-tæ̃ʔ-ní-h*
hammock 1SG take-DYN 1SG stand-be-CNTRFCT-INFR-DECL
'(…) I took (was given) a hammock; I would have stayed there (but these days it's impossible).' (Epps 2004: 614)

Coding elaboration

- Does the element display inflectional elaboration independent of the semantic head/host?
  - yes = has coding elaboration
  - no = has no coding elaboration

Paresi

*ha=**moka***	*natyo*	*hoka*	*n=aotya-**ki**-tsa*	*xitso*
2SG=CAUS	1SG-put	CON	1SG=remember-CAUS-TH	2PL
haliti	*ni=rai-ne*			
Paresi	3SG=talk-POSSD			

‘You made me teach you all the Paresi language.’ (Brandão 2014: 269)

Prominence projection

- Does the element project its own stress domain?
  - always
  - sometimes
  - never

Kokama

a. *penu* *yawachima-ka-t=**utsu*** *uyarika* *awa=pura*
1PL.F arrive-REI-CAUS=FUT again person=FOC
ukuata-ri-n=pura=nu
pass-PROG-NMLZ=FOC-PL
'We will reach again the people (who are crossing the street)'
(Vallejos Yopán 2010: 603)

b. *yanamata* *kari-ri=tsui* *y=itika-ka* *y=utsu*
bush scrape-PROG=ABL 3SG.F=throw-REI 3SG.F=FUT
'After scarping the bushes, he goes to throw it.'
(Vallejos Yopán 2010: 480)

Exponence complexity

- defining feature for advocates of MA
- there are different types with different criteria
- aggregated the codable criteria into a single measure:

Criterion		Description	Value
a	number of allomorphs	how many allomorphs are there?	1-n
s	suppletive allomorphy	is there suppletion?	yes = 1, no = 0
m	multiple exponence	does the meaning distribute over several forms?	yes = 1, no = 0
f	fossilization	does it combine with an empty/opaque root?	yes = 1, no = 0

ec = a + s + m + f

Boundary strength problem

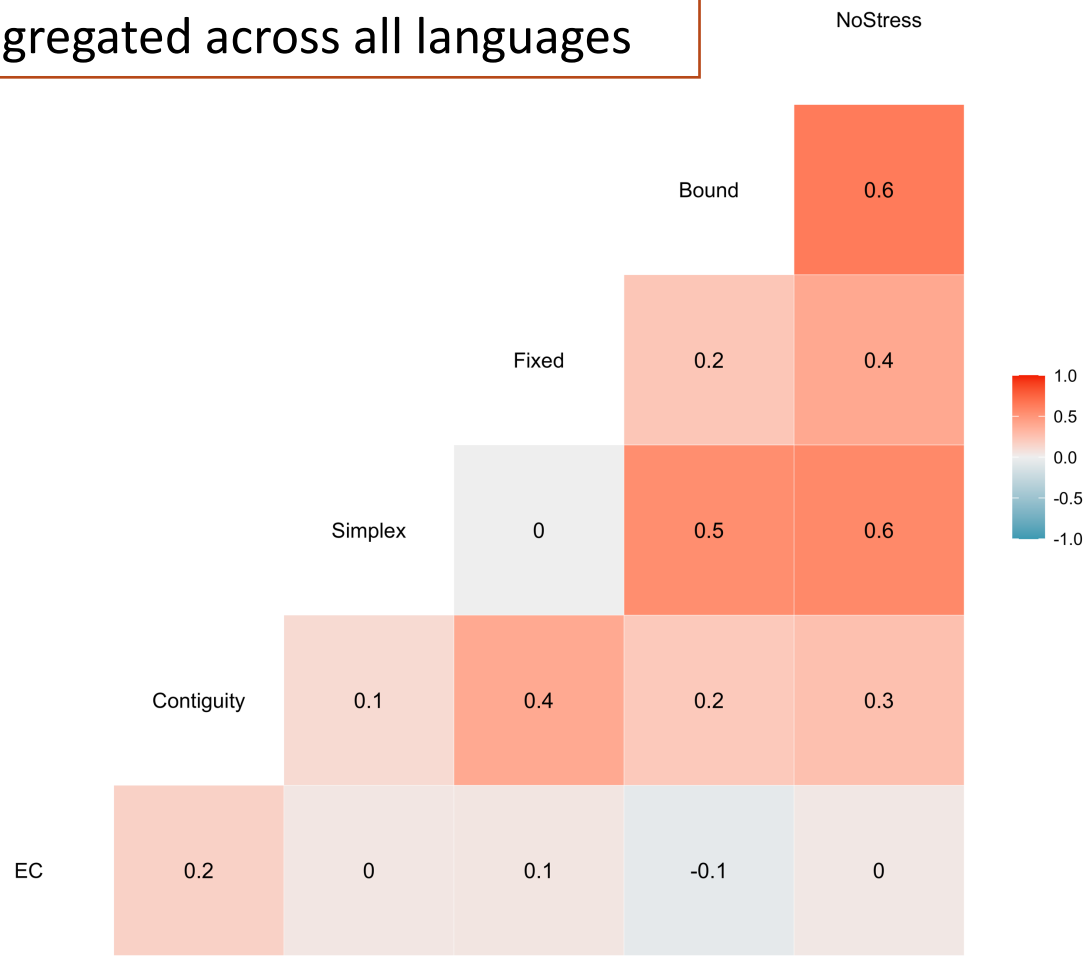
Correlation matrices

The boundary strength problem

- Problem: Languages may vary in the degree to which morphology and syntax are distinct. Some languages might display more indeterminacy than others.
- Our approach: Proposal for visualizing and measuring the degree of distinctiveness of morphology and syntax as a typological index
  - morphological autonomy is associated with exponence complexity in the literature → is exponence complexity a good proxy for MA?

Correlations

correlation matrix of all variables aggregated across all languages



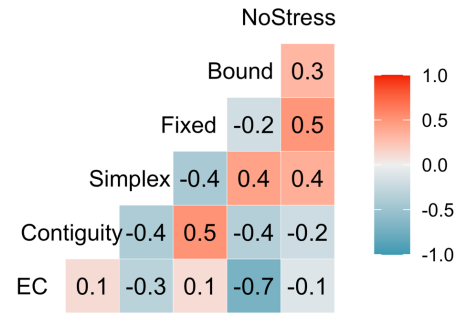
if exponence complexity is a good measure of MA, there should be positive correlations with most variables

Variable	Coded as 0 (synt.)	Coded as 1 (morph.)
Interruptability	Interruptable	Not interruptable
Coding elaboration	Present	Absent
Fixedness	Variable	Fixed
Boundedness	Free	Bound
Prosodic prominence	Present	Absent = 2, both = 1 (clitics)

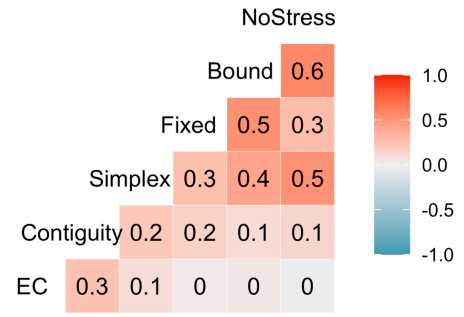
Correlations

correlation matrices of all variables aggregated by language

Movima

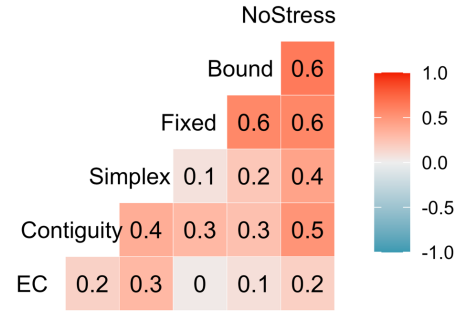


Puinave

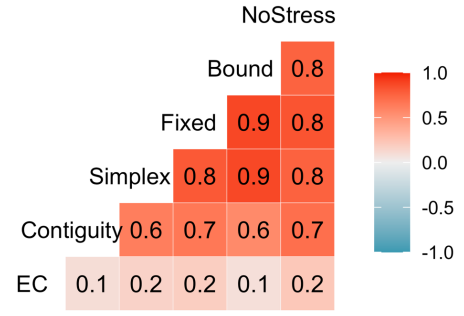


Tariana

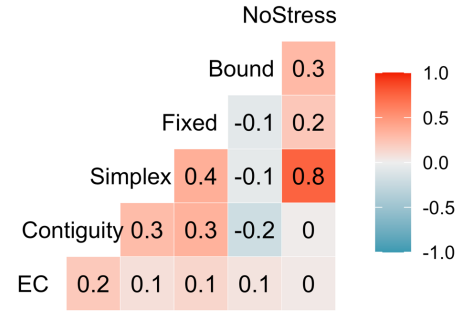
AsheninkaPerene



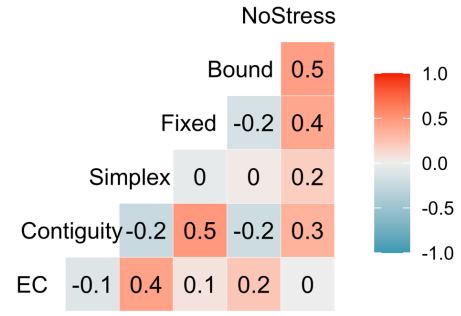
Chacobo



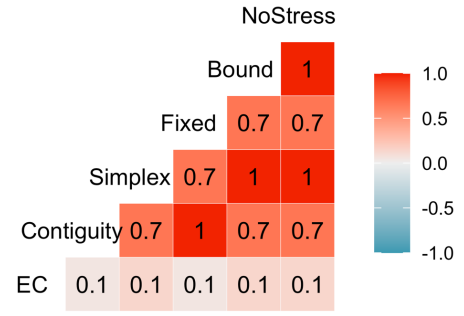
Cavineña



Hup



CentralAlaskanYupik



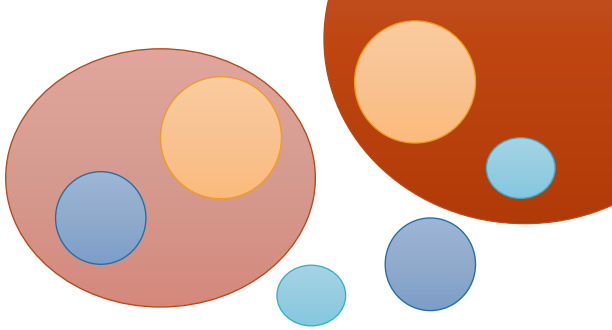
The boundary strength problem

- Problem: Languages may vary in the degree to which morphology and syntax are distinct. Some languages might display more indeterminacy than others.
- Interim conclusions:
  - languages vary in the degree to which MA-variables correlate
  - exponence complexity does not strongly correlate with other variables
  - some variables are unimportant/unlikely to contribute to this distinction, but which ones varies by language

Composition problem

Random Forests

The composition problem



- Problem: Languages may differ with respect to how the distinction between morphology and syntax is made. Certain properties (e.g. high degree of allomorphy) may distinguish morphology and syntax in one language and not another. Languages may also vary in the degree to which certain properties help distinguish morphology and syntax.
- Our approach: Random Forests (classification algorithm)
  - based on author classifications: reflects the intuition that grammar authors are mostly consistent in applying wordhood criteria internally
  - based on exponence complexity: reflects ‘theoretical grounding’, as it is assumed to be particularly important for MA

Background on Random Forests

- a classification algorithm that aggregates over a multitude of decision trees
- number of variables out of all dependent variables that are tried at each split in each decision tree for best classifying the data has to be defined beforehand.
- determined by running multiple RF models with different numbers of variables to find the one producing the best results
- unlike regression, RF models make no assumptions about the data
- but they still need a dependent variable (on which the classification is based)

Background on Random Forests

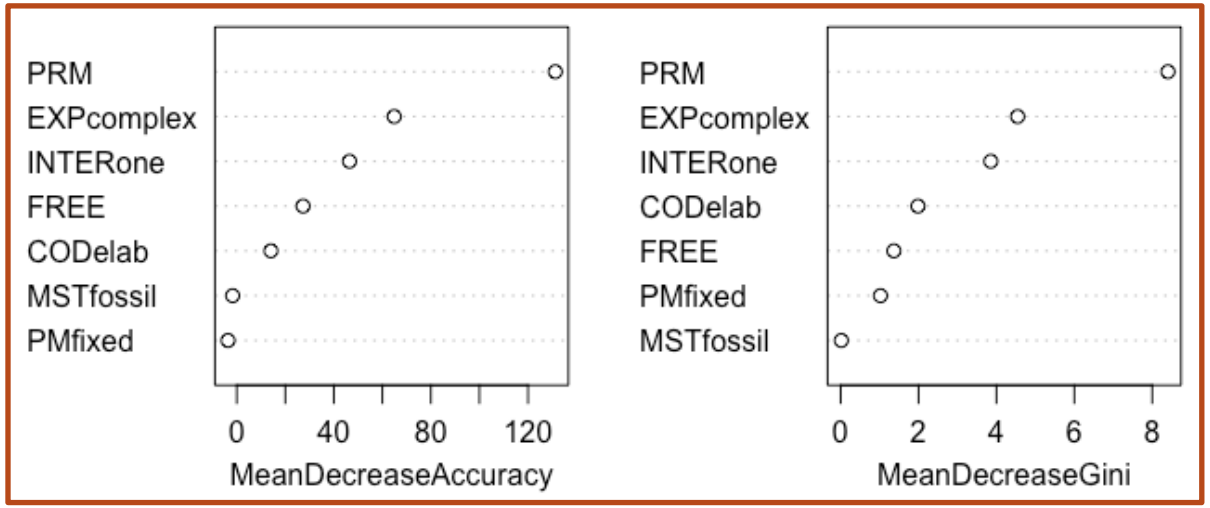
- output:
  - out-of-bag error rate (OOB): how much the model classified correctly
  - relative importance variable plot: how much each variable contributes to the classification
- evaluation measures:
  - baseline: skewness of the data
  - accuracy: sum of correct predictions
  - difference: accuracy-baseline (how much better than change the RF model performs)

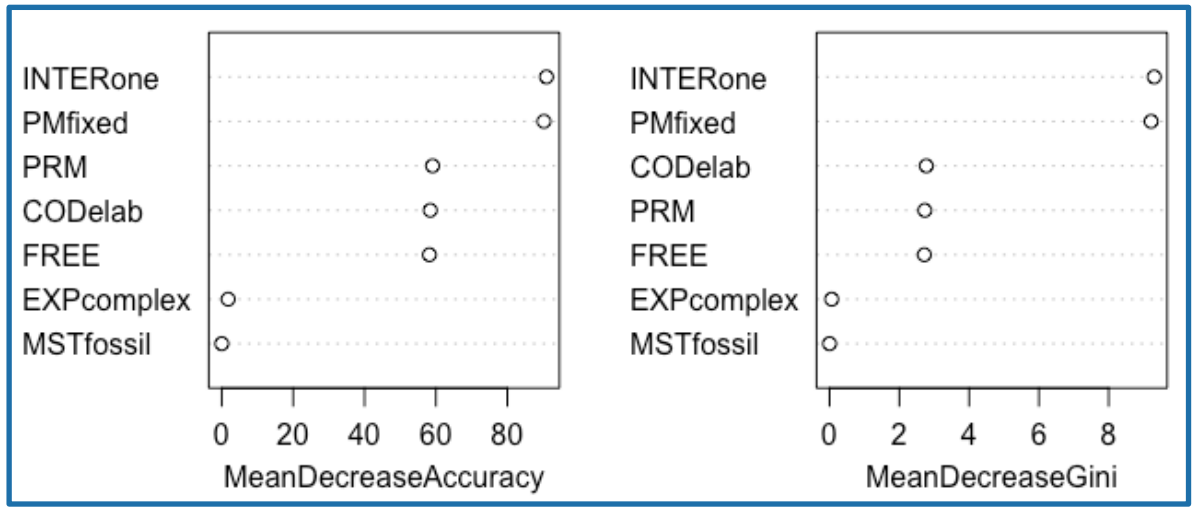
RF with Author Classification

Central Alaskan Yupik

```
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0%
Confusion matrix:
       affix clitic word class.error
affix     65      0    0           0
clitic     0      8    0           0
word       0      0    8           0
```



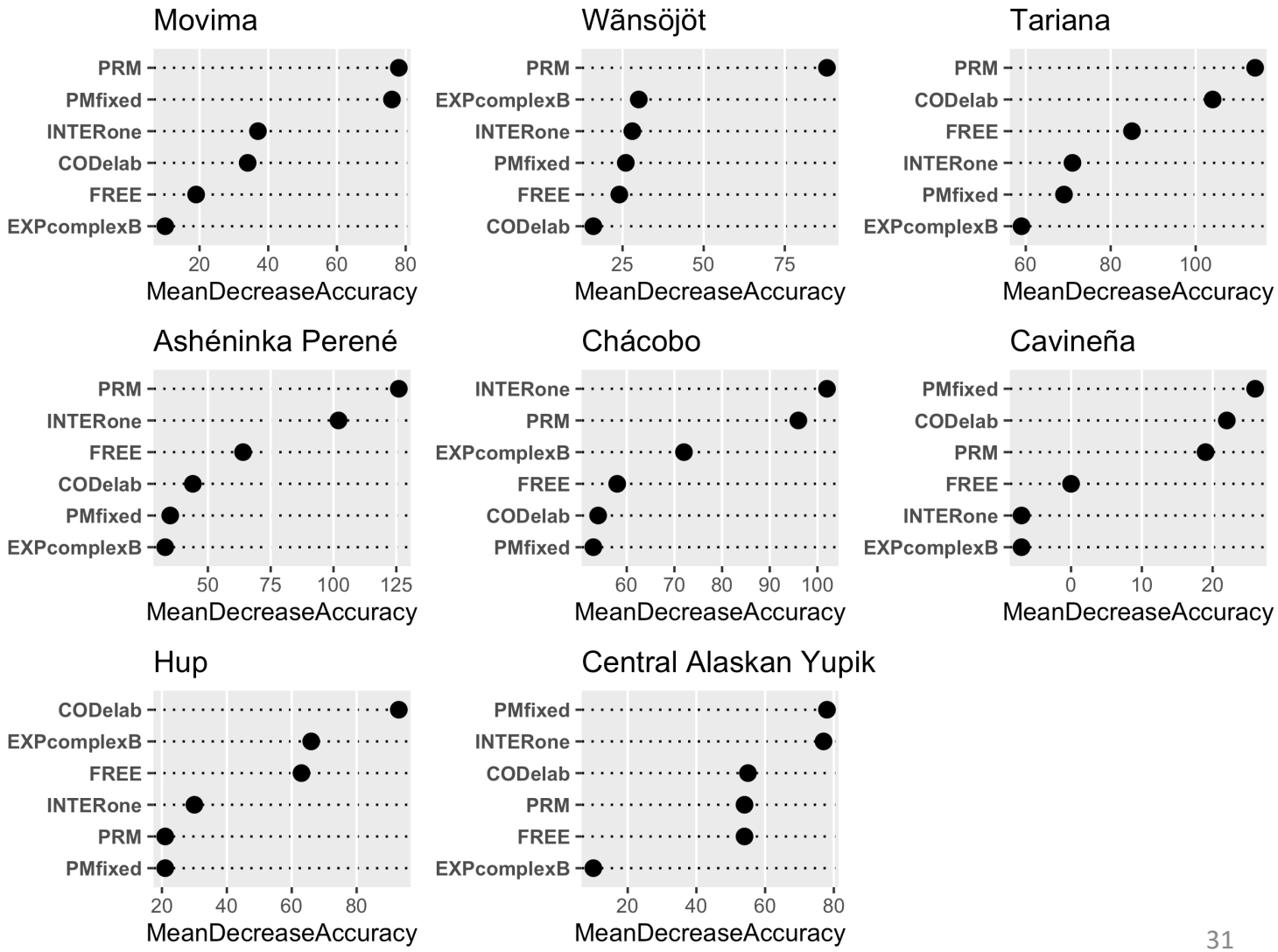



```
No. of variables tried at each split: 7

        OOB estimate of  error rate: 18.18%
Confusion matrix:
       affix clitic word class.error
affix     72      2    2  0.05263158
clitic     8      4    0  0.66666667
word       5      1    5  0.54545455
```

Puinave

RF with Author Classification



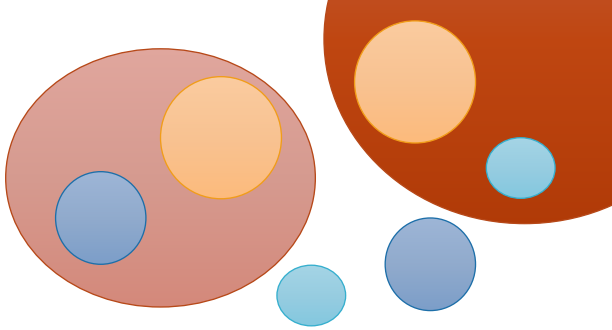
variable importance plots by language

RF with Exponence Complexity

Movima						**Puinave**					
OOB estimate of error rate: 6.43%						OOB estimate of error rate: 29.29%					
	High	Low	class.error	bl.	57.14		High	Low	class.error	bl.	72.73
High	52	8	0.134	acc.	93.57	high	0	27	1.0	acc.	70.71
Low	1	79	0.013	diff.	36.43	low	2	70	0.028	diff.	−2.02
Tariana						Ashéninka Perené					
OOB estimate of error rate: 10.08%						OOB estimate of error rate: 11.22%					
	High	Low	class.error	bl.	90.76		High	Low	class.error	bl.	88.78
High	0	11	1	acc.	89.92	High	0	11	1	acc.	88.78
Low	1	107	0.009	diff.	−0.08	Low	0	87	0	diff.	0
Chácobo						Cavineña					
OOB estimate of error rate: 35.42%						OOB estimate of error rate: 19.64%					
	High	Low	class.error	bl.	67.71		High	Low	class.error	bl.	80.36
High	0	31	1.0	acc.	64.58	High	0	11	1	acc.	80.36
Low	3	62	0.046	diff.	−3.13	Low	0	45	0	diff.	0
Hup						Central Alaskan Yupik					
OOB estimate of error rate: 13.85%						OOB estimate of error rate: 16.05%					
	High	Low	class.error	bl.	90.77		High	Low	class.error	bl.	83.95
High	0	6	1	acc.	86.15	High	0	13	1	acc.	83.95
Low	3	56	0.051	diff.	−0.046	Low	0	68	0	diff.	0

OOB error rates and evaluation measures by language

The composition problem



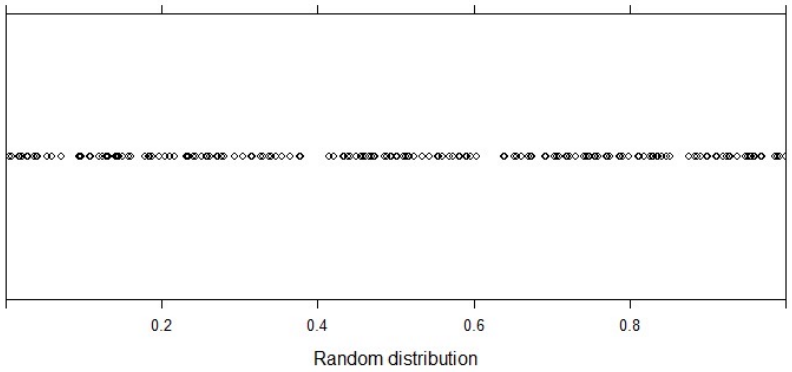
- Problem: Languages may differ with respect to how the distinction between morphology and syntax is made. Certain properties (e.g. high degree of allomorphy) may distinguish morphology and syntax in one language and not another. Languages may also vary in the degree to which certain properties help distinguish morphology and syntax.
- Interim conclusions:
  - there is a lot of variation in terms of which variables are important for classifying morphemes into wordhood categories
  - languages also seem to vary in how much wordhood variables reflect important structural generalizations (morphology-syntax distinction)
  - RFs can be used to describe variation in wordhood variables, but they need a ‘baseline’
  - general issue: exponence complexity displays weak correlations with other variables in all languages of our sample → probably need another baseline, but this is tricky

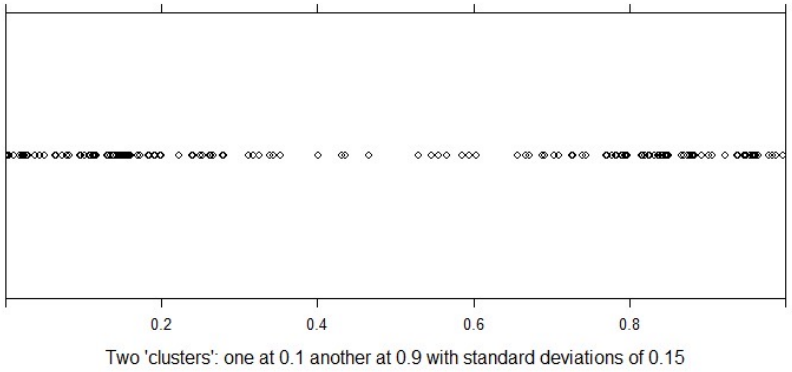
Architecture problem

Cluster models and validation

The architecture problem

- Problem: Languages may vary with respect to whether morphology and syntax are distinct at all.
- Our approach: clustering with validation techniques
  - if there is MA, we would expect the variables to cluster either into two groups (words vs. affixes/clitics) or three groups (words vs. clitics vs. affixes)
  - comparison with simulated data sets





The ‘clustering problem’

- Haspelmath’s formulation of the issue suggests that should cluster
- clustering models can show that, but there are limitations:
  - there is no standard of definition of the term ‘cluster’
  - there are many algorithms and models (it’s not *a priori* clear which one is most appropriate)
  - clustering is not inferential, i.e. it does not test hypotheses – it’s exploratory
  - clusters need to be validated to show that they are not arbitrary partitions – validations techniques are still domain-specific
- we use hierarchical clustering and the height difference between the first and second partition
- we compare the clusters of the languages to a set of simulated data

Simulated data

Type	Exp.Comp	Fixed	Free	Inter	Prom	Cod.El.	Obl	Fossil
“affix”	4 (1-16)	yes	bound	no	no	no	40/40	no
”clitic”	3 (2-12)	23/17	21/19	23/17	14/15/11	7/33	9/31	4/36
“word”	2 (1-5)	no	free	yes	yes	63/17	no	13/67

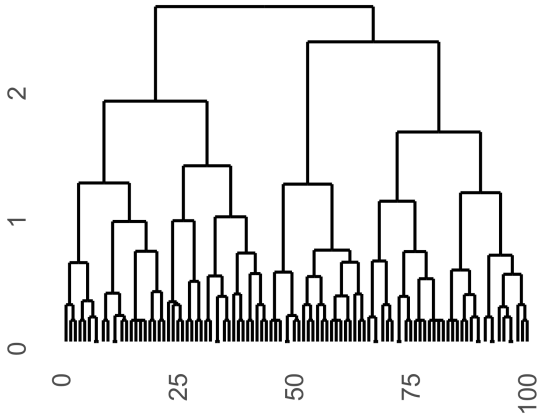
Example of a dummy set with some variation

Hierarchical clustering in a nutshell

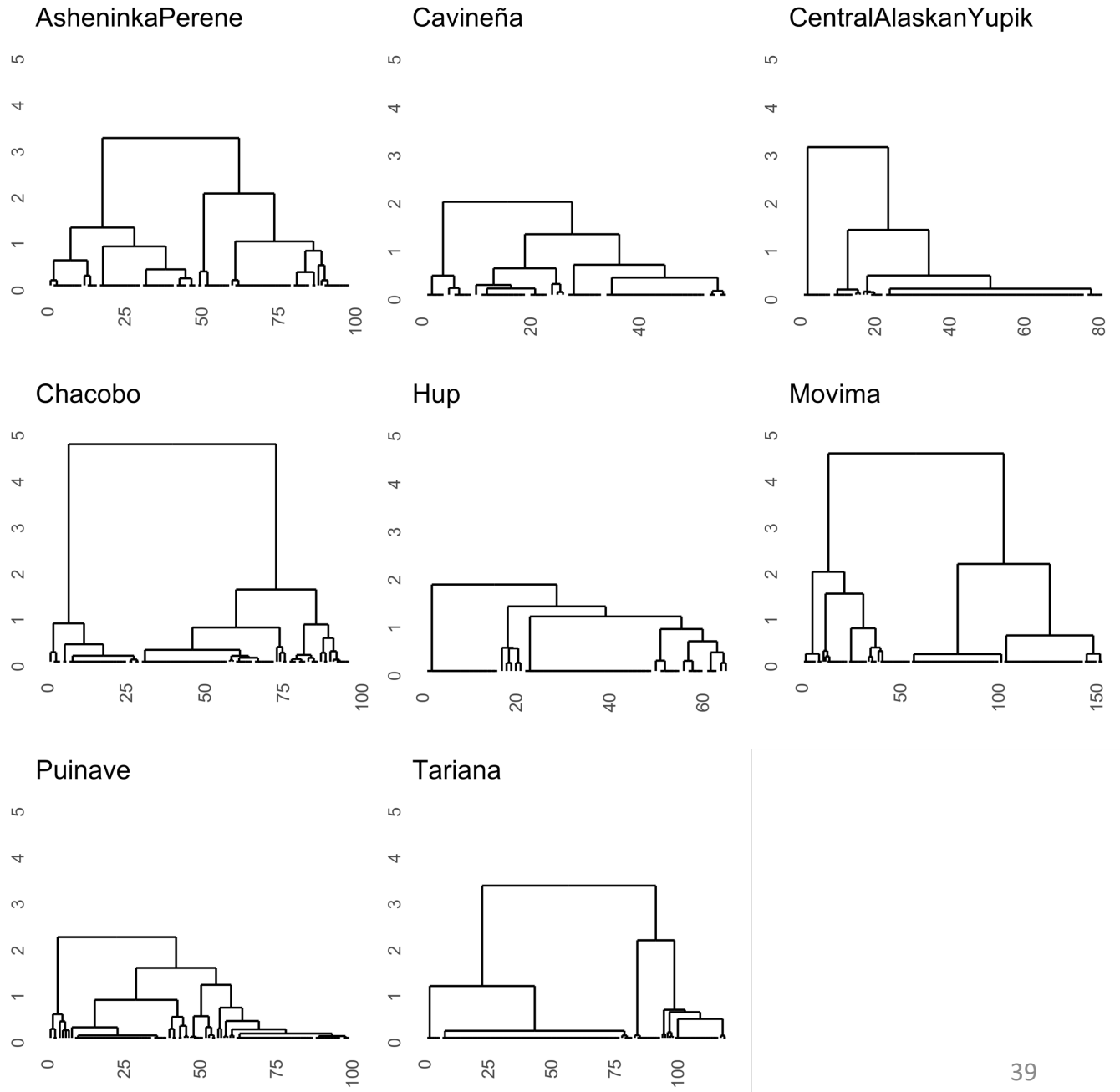
1) construct a distance matrix for each language (based on the numeric versions of the variables)
2) apply hierarchical cluster model (Ward's minimum variance method)
3) inspect the dendrograms and compare to the simulated data
4) look at the cophenetic distance and height difference between the first and second partition
   - cophenetic distance: measure how (dis)similar elements need to in order to be grouped into the same cluster (scale 0-1)

Clustering

Dendrogram of HC for each language of the sample

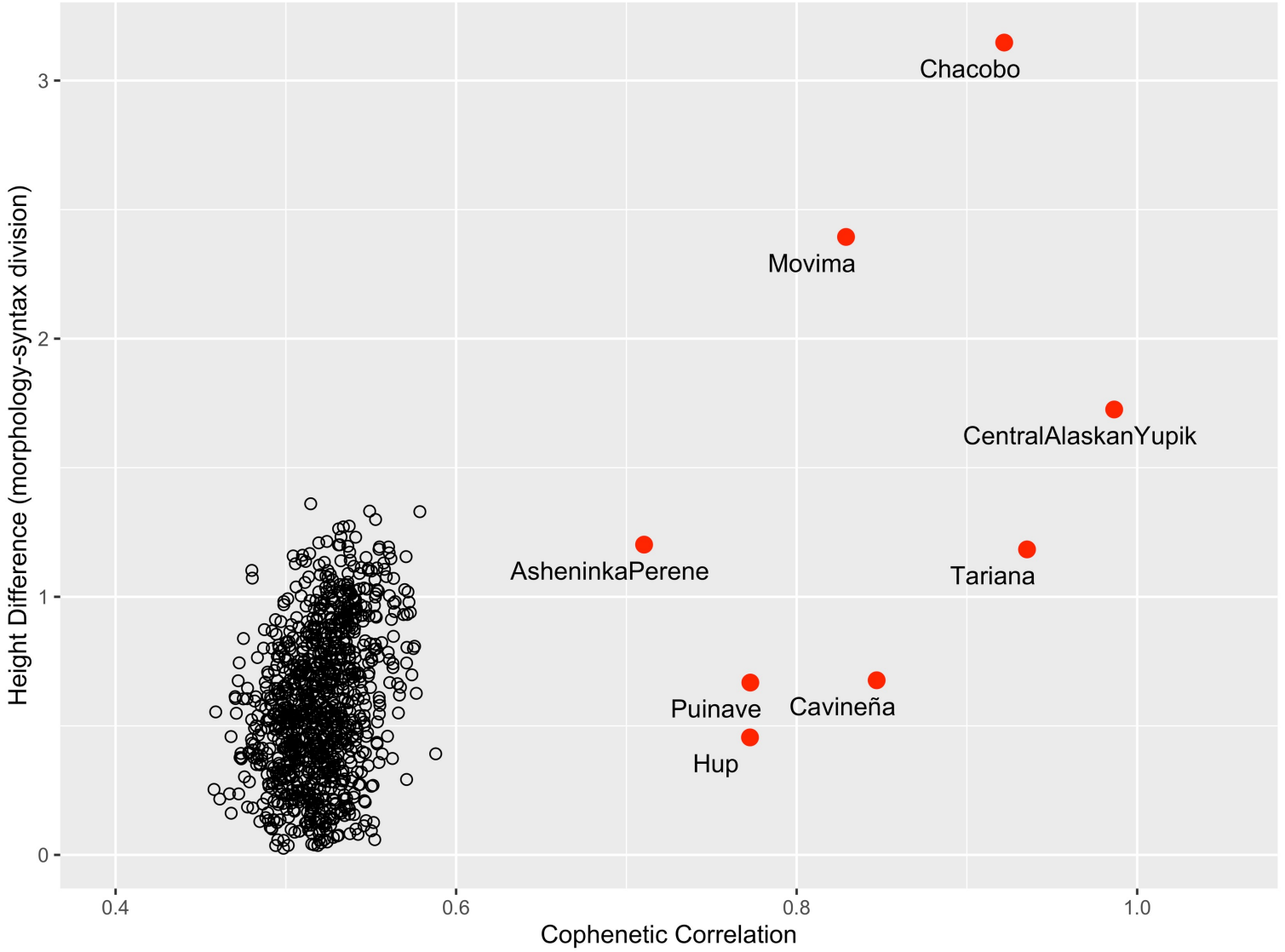


Dendrogram of HC on simulated data



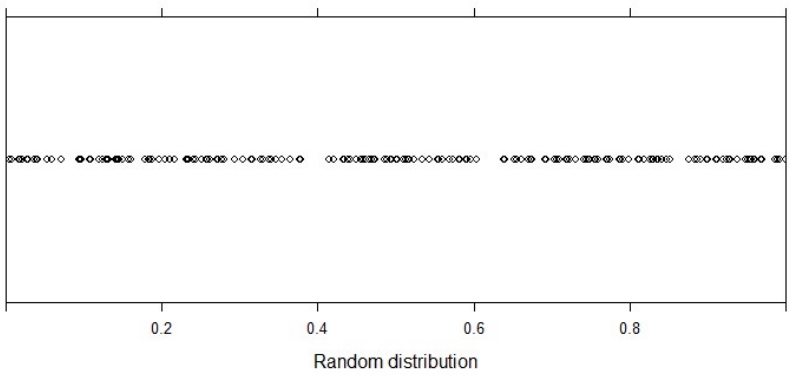
Clustering

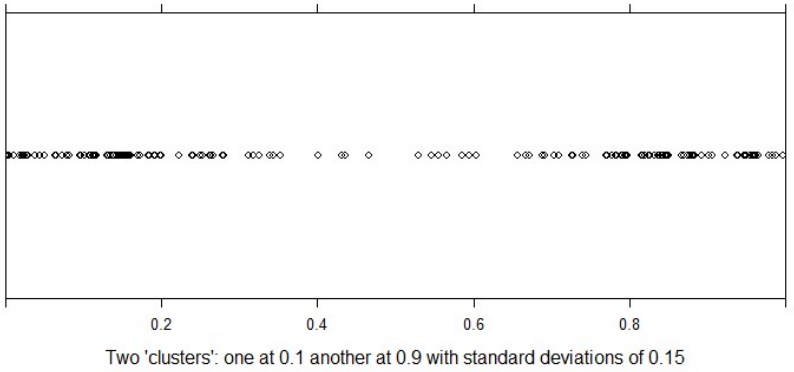
Cophenetic correlation and first/second partition height differences of 1,000 simulated data sets (black circles) and the languages of our sample (red dots).



The architecture problem

- Problem: Languages may vary with respect to whether morphology and syntax are distinct at all.
- Interim conclusions:
  - languages vary in terms of whether they display a morphology-syntax distinction
  - we can measure the degree to which such a distinction is valid (enganging with the notion of ‘fuzzy’ wordhood empirically)





Conclusions and outlook

What we can say about MA so far

- more data necessary (but coding is time-intensive)
- further explorations needed of statistical techniques
- more awareness of the empirical issue

BUT:

- everything we tried so far provides little (or no) support for the “clustering distribution“
- rather, it suggests there is huge variation between languages regarding to what degree the variables „bunch“ together

Desiderata and further research

- larger sample both in terms of languages and in terms of morphemes
  - would allow for generalizations
  - difficult to implement because coding is very time-consuming and can only be done in collaboration with language experts
- developing and refining the variables
  - more sophisticated measure of exponence complexity
- focusing on the global architecture problem (from an empirical perspective) could add to the ‘continuity’ debate
  - emphasis on continuity in the grammaticalization literature (gradual development of words into affixes, etc.)
  - but little engagement on what this means for the language system as a whole

Thanks for listening!

Full paper: https://doi.org/10.1515/lingty-2021-0041

Supplementary materials: https://doi.org/10.5281/zenodo.6008054

Contact: sandra.auderset@unibe.ch