# Sampling in Typology
# Some potential pitfalls

Michael Cysouw
ZAS Typology-circle
29 january 2003

# Large-sample typology:

- Comparison of many languages (50 and up)
- Sampling genetically independent
- Abstraction of variation into a typology
- Explain asymmetries in frequency of types

# Possible problems:

- How large is the actual variation?
- What does genetic independence mean?
- Large-areal consistency
- Statistical interpretation of numbers

# Actual or possible languages?

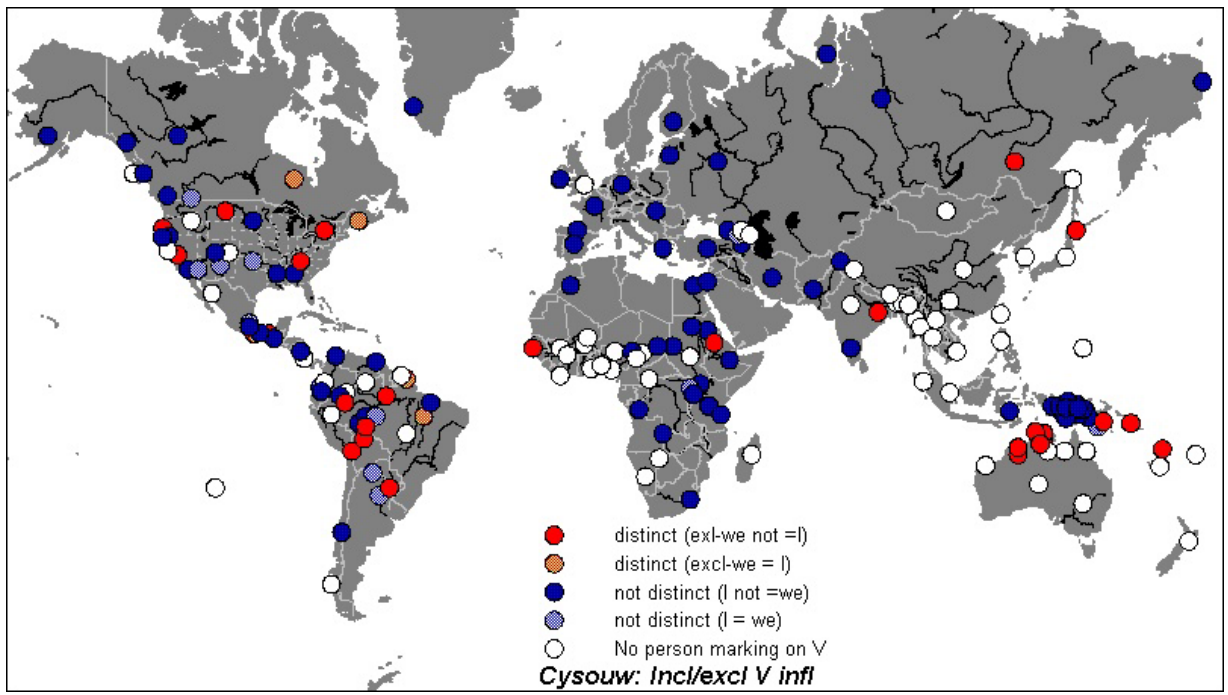
- The world's languages might not represent all possible types
- If true, than a sample would only represent the actual languages, not the possible ones
- E.g. Nichols (1992) assumes this
- Maslova (2000) gives some theoretical backing to this idea

# Why sample by families?

- Genetic families are defined by particular criteria (sound change, non-borrowed cons.)
- The feature of the typological investigation does not have to be distributed accordingly
- E.g. Haspelmath (1997) finds a large variation in indefinites in Europe alone.
- Taking only one language per genetic unit is only a bottom-line criterium for succes.

# Large-areal consistencies

- Many typological distribution show large uniform geographical areas (not genetic!)



# Head/Dependent marking
## (Nichols 1986, 1992)

- For each language, she counts overtly marked H(ead) or D(ependent) constructions. A construction can be both H and D marked!

- **Noun phrase possession (maximal two H and two D points):**

  Pronominal: *my book* (English: one D point, as *my* is marked)

  Nominal: *John's book* (English: one D point, as *John* is marked)

- **Noun phrase modification (maximal one H and one D point):**

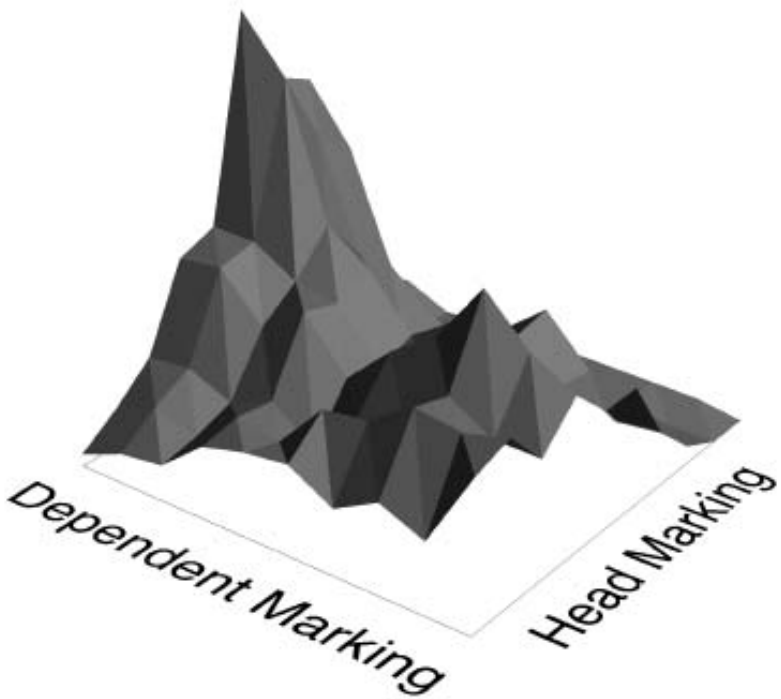
  *the red book* (English zero points, no marking)

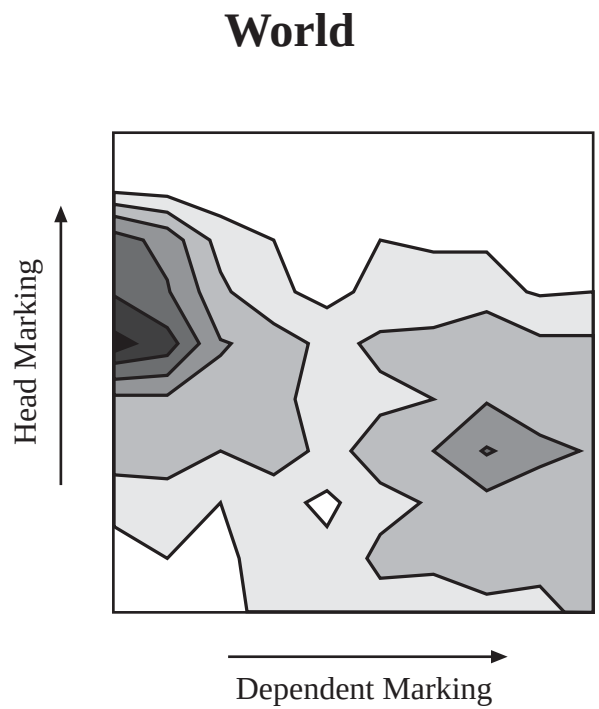
- **Sentence arguments (maximal six H and six D points):**

  Pronominal: *I gave it to you.* (English two D points, as *I* and *it/you* are case marked)

  Nominal: *John gave the book to Mary*. (English zero points: no case marking on nouns)

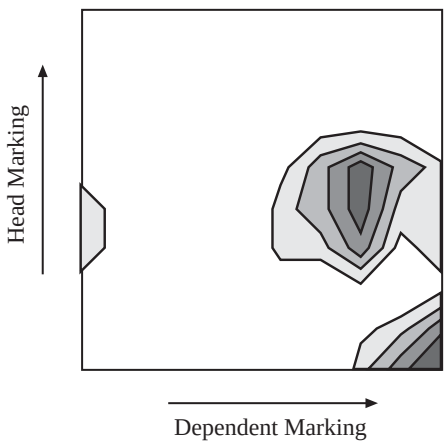
# Graphical analysis of Nichols' data



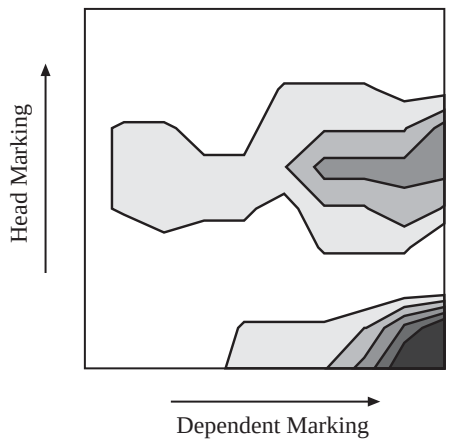



# Some areas are typically D-marked

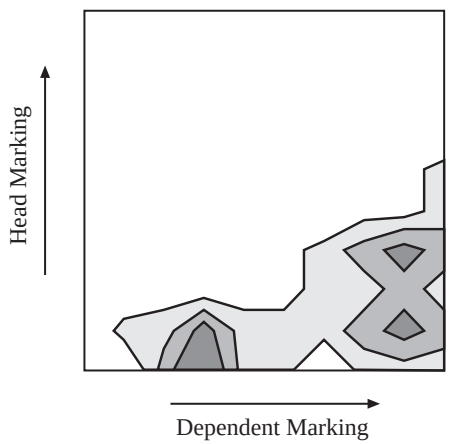
**Northern Eurasia (East)**



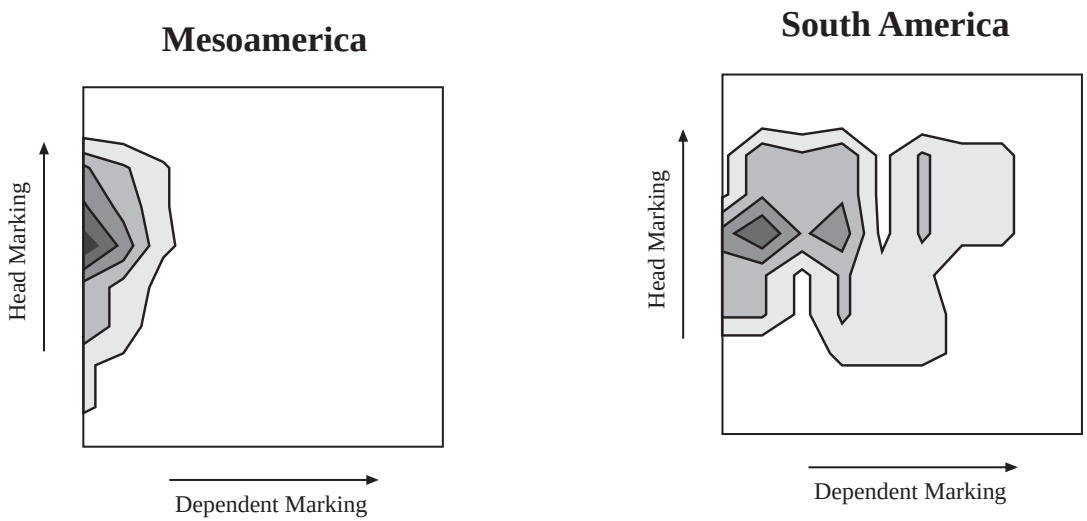

**Australia**



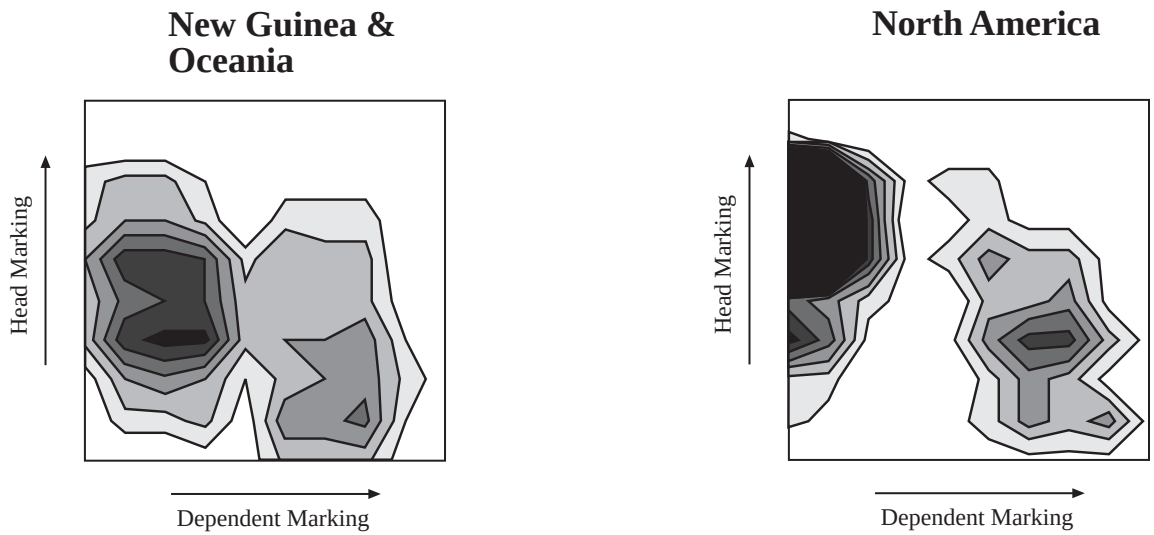

**South and Southeast Asia**



# Some areas are typically H-marked



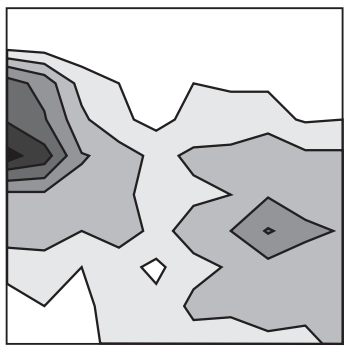
# Some areas are alike to whole world



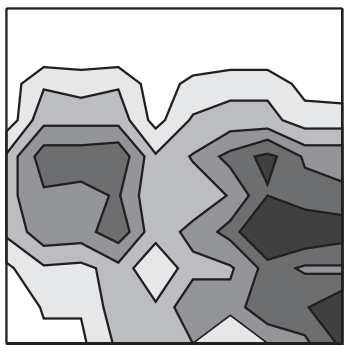

Note: the D-marked languages are geographically restricted (NE New Guinea and American Westcoast)

# Worldwide H-cluster is regionally determined

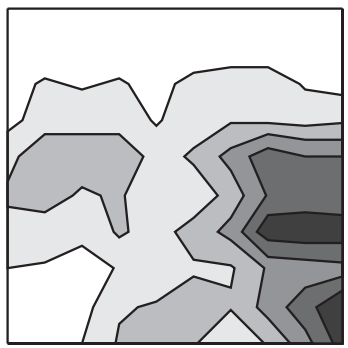
**all languages in Nichols' sample**



**without North America and Mesoamerica**



**without North America, Mesoamerica, New Guinea and Oceania**



# Solutions?

- Typological patterns can only be interpreted as universals if most areas show the same pattern as the whole world (Dryer 1989, 1991, 1992)
- Ideal solution: compare within-genus variation with between-genus variation (very labour intensive!)

# What do numbers mean?

- Basic analytic tool in typology is the implicational universal
- A $\rightarrow$ B iff the combination (A+, B-) is (almost) unattested:

| | | A | |
|---|---|---|---|
| | | + | Š |
| B | + | **$X_1$** | **$X_2$** |
| | Š | ***Æ*** | **$X_3$** |

# However, low frequency does not necessarily mean anything

| | | A: + | A: Š | total |
|---|---|---|---|---|
| B | + | **10** | **31** | 41 |
| | Š | **2** | **12** | 14 |
| | total | 12 | 43 | 55 |

| | | A: + | A: Š | total |
|---|---|---|---|---|
| B | + | $\frac{41}{55} \cdot \frac{12}{55} \cdot 55 = 8.9$ | $\frac{41}{55} \cdot \frac{43}{55} \cdot 55 = 32.1$ | 41 |
| | Š | $\frac{14}{55} \cdot \frac{12}{55} \cdot 55 = 3.1$ | $\frac{14}{55} \cdot \frac{43}{55} \cdot 55 = 10.9$ | 14 |
| | total | 12 | 43 | 55 |

| | | A: + | A: Š | total |
|---|---|---|---|---|
| B | + | **+ 1.1** | **- 1.1** | 41 |
| | Š | **- 1.1** | **+1.1** | 14 |
| | total | 12 | 43 | 55 |

This difference is not statistically significant (e.g. Fisher's Exact $p = 0.71$)

# What do typologists say?

| Smallest number | Kind of universal | Hypothetical distributions of a 100-language sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Zero | Exceptionless universal | 33 | 34 | 26 | 48 | 20 | 60 | 14 | 72 |
| | | **0** | 33 | **0** | 26 | **0** | 20 | **0** | 14 |
| Five | Strong tendency | 36 | 23 | 31 | 33 | 27 | 41 | 22 | 51 |
| | | **5** | 36 | **5** | 31 | **5** | 27 | **5** | 22 |
| Ten | Statistical tendency | 38 | 14 | 33 | 24 | 30 | 30 | 25 | 40 |
| | | **10** | 38 | **10** | 33 | **10** | 30 | **10** | 25 |
| Fifteen | Maybe something | | | 35 | 15 | 31 | 23 | 28 | 29 |
| | | | | **15** | 35 | **15** | 31 | **15** | 28 |
| Nineteen | Nothing | | | | | 31 | 19 | 27 | 27 |
| | | | | | | **19** | 31 | **19** | 27 |

# What do statisticians say?

Hypothetical distributions of a 100-language sample

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 33 | 34 | 26 | 48 | 20 | 60 | 14 | 72 | | |
| **0** | 33 | **0** | 26 | **0** | 20 | **0** | 14 | | |
| 36 | 23 | 31 | 33 | 27 | 41 | 22 | 51 | | |
| **5** | 36 | **5** | 31 | **5** | 27 | **5** | 22 | | |
| 38 | 14 | 33 | 24 | 30 | 30 | 25 | 40 | | |
| **10** | 38 | **10** | 33 | **10** | 30 | **10** | 25 | | |
| | | 35 | 15 | 31 | 23 | 28 | 29 | | |
| | | **15** | 35 | **15** | 31 | **15** | 28 | | |
| | | | | 31 | 19 | 27 | 27 | | |
| | | | | **19** | 31 | **19** | 27 | | |

| **Kind of interaction** | Very strongly significant | Strongly significant | Significant | No interaction |
|---|---|---|---|---|
| **Fisher's Exact two-tailed** | $p < 0.000001$ | $p < 0.001$ | $p < 0.05$ | $p > 0.2$ |

# almost orthogonal interpretations:

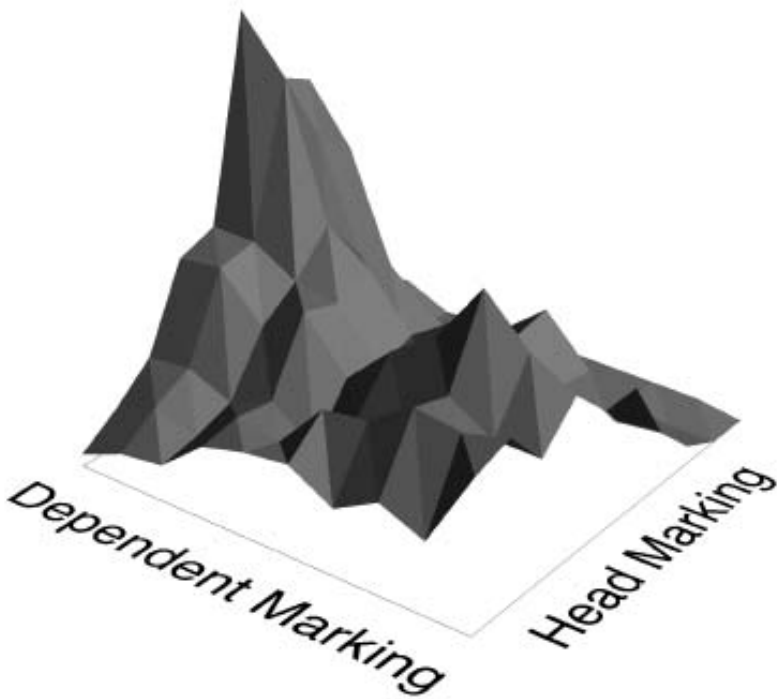
| Smallest number | Kind of universal | Hypothetical distributions of a 100-language sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Zero | Exceptionless universal | 33 | 34 | 26 | 48 | 20 | 60 | 14 | 72 |
| | | **0** | 33 | **0** | 26 | **0** | 20 | **0** | 14 |
| Five | Strong tendency | 36 | 23 | 31 | 33 | 27 | 41 | 22 | 51 |
| | | **5** | 36 | **5** | 31 | **5** | 27 | **5** | 22 |
| Ten | Statistical tendency | 38 | 14 | 33 | 24 | 30 | 30 | 25 | 40 |
| | | **10** | 38 | **10** | 33 | **10** | 30 | **10** | 25 |
| Fifteen | Maybe something | | | 35 | 15 | 31 | 23 | 28 | 29 |
| | | | | **15** | 35 | **15** | 31 | **15** | 28 |
| Nineteen | Nothing | | | | | 31 | 19 | 27 | 27 |
| | | | | | | **19** | 31 | **19** | 27 |

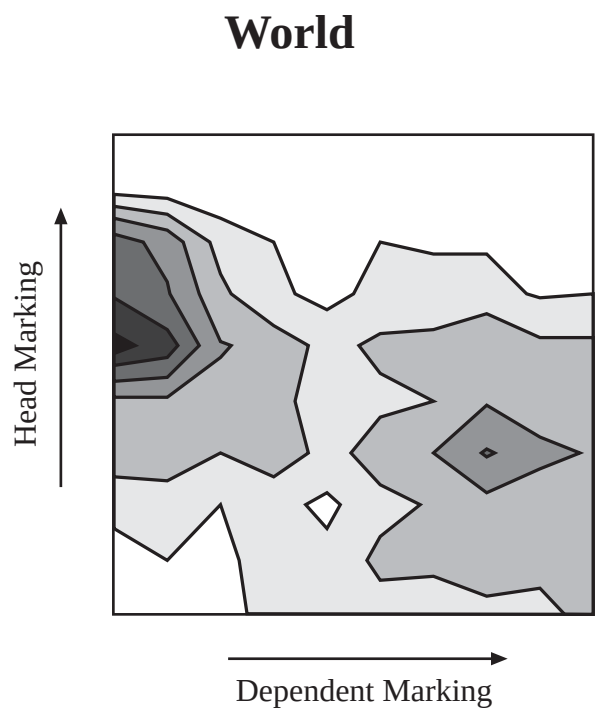
| **Kind of interaction** | Very strongly significant | Strongly significant | Significant | No interaction |
|---|---|---|---|---|
| **Fisher's Exact two-tailed** | $p < 0.000001$ | $p < 0.001$ | $p < 0.05$ | $p > 0.2$ |

# Against implicational universals

- The zero in the table is not important
- The statistical significance of the distribution is important
- As most typologist are well-thinking human being, errors are not widespread
- However, in complicated typologies with many variables, it might easily go wrong

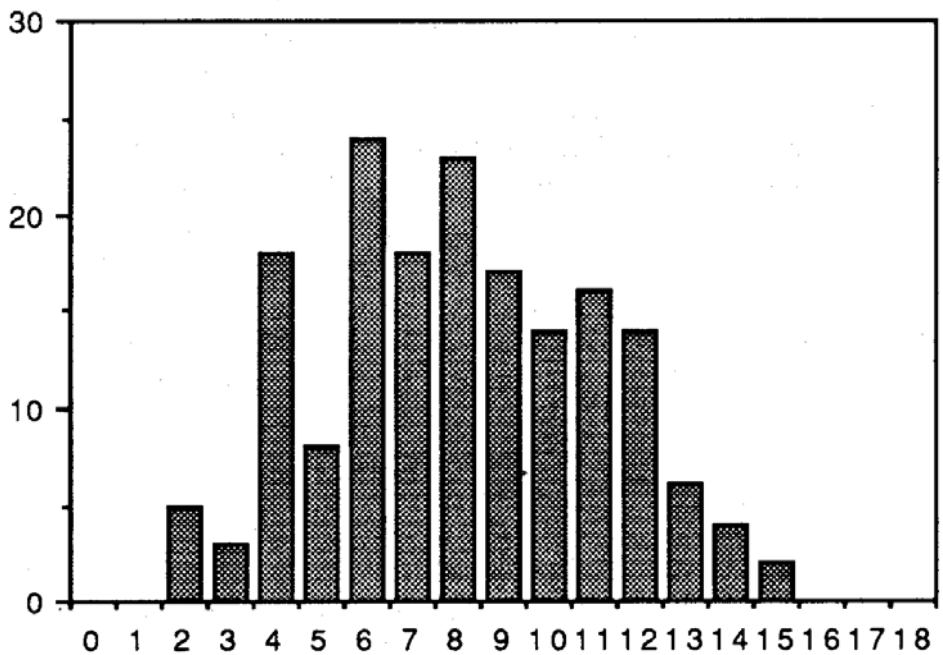
# Nichols' Head/Dependent typology



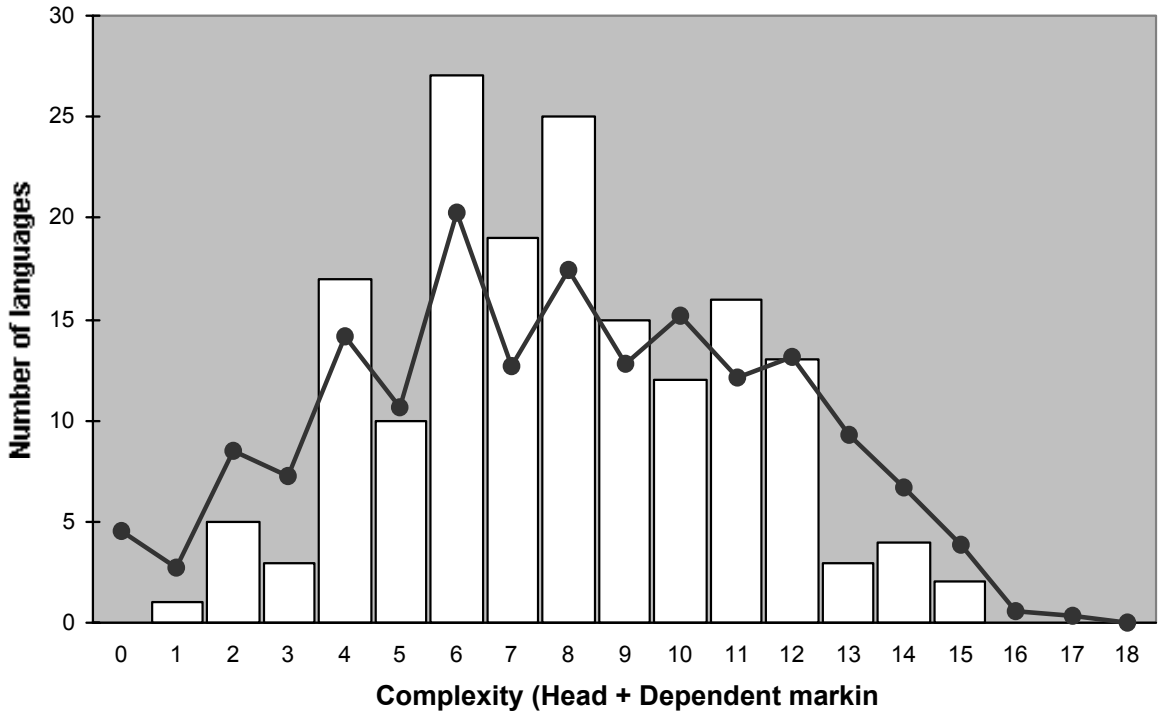



# Complexity (Head+Dependent)

'… the complexity (Dependent points plus Head points …) has a roughly normal distribution. Neither zero complexity nor the theoretical maximum complexity of [18] points (9 Head points plus 9 Dependent points …) occurs. the highest attested complexity is 15, found in only two languages. Figure 4 shows the complexity values attested in my sample. … The normal distribution and preference for moderate complexity shown in the overall sample are echoed in most … areas, with high complexity predominating in only two.' (Nichols 1992: 88-89)



# Statistically, this in not correct

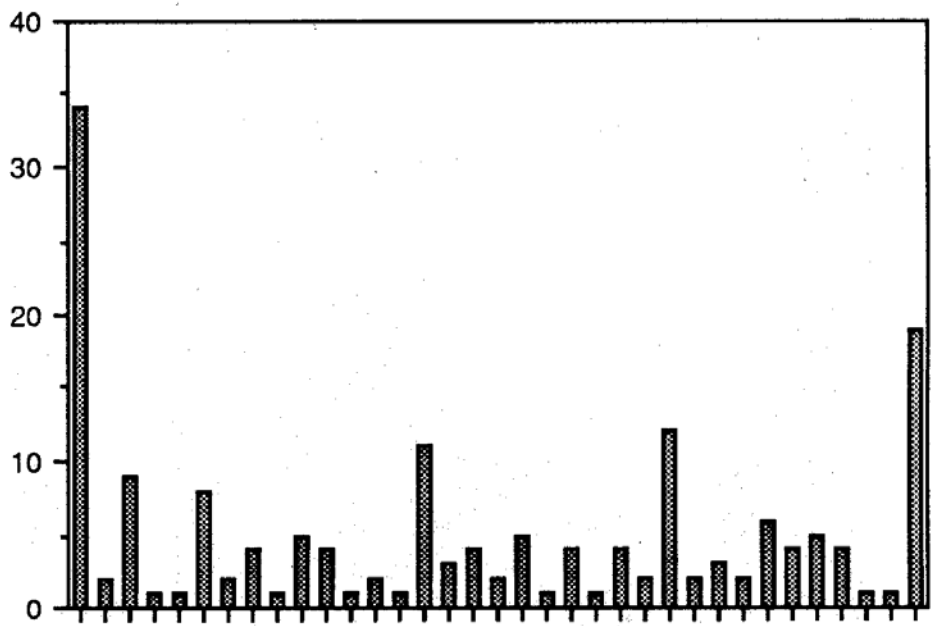



Bars: Actual values from Nichols

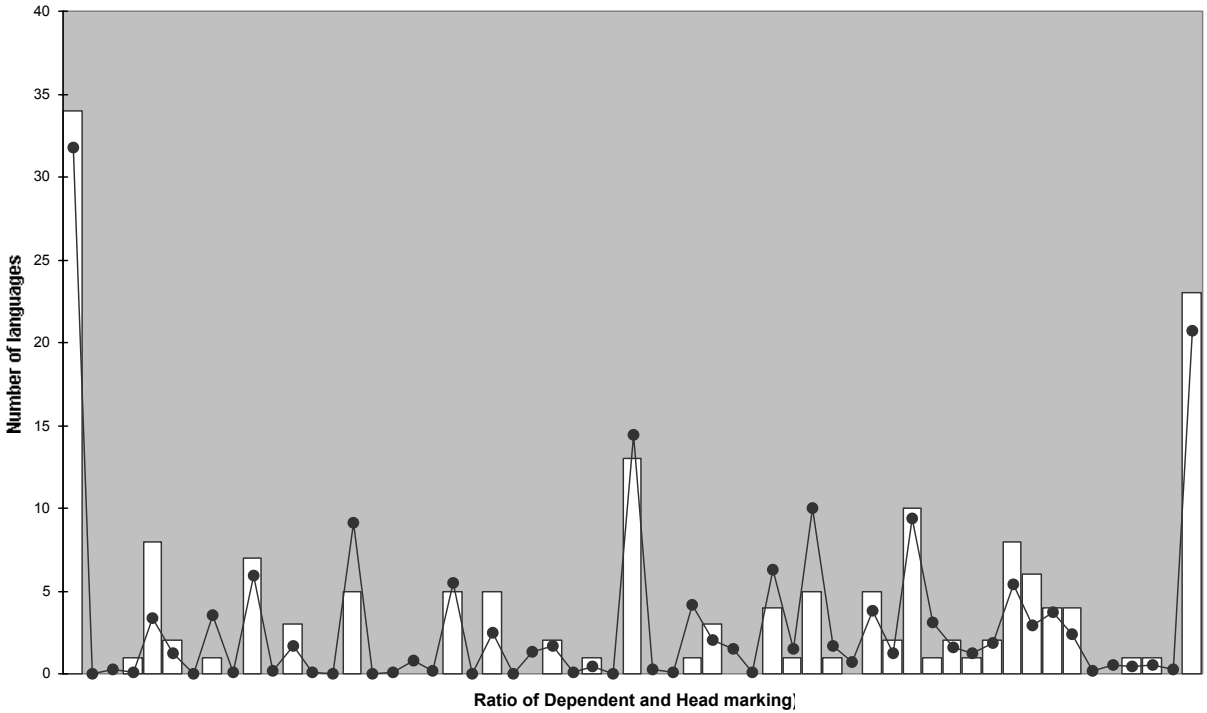
Line: Statistically expected values

# Ratio Head/Dependent marking

'… computing the ration of dependent to head marking … gives us 35 different ratios among the 174 sample languages. Their distribution is shown in figure 1. It is bimodal, with the greatest peaks at the extremes of exclusive head marking (ration of zero since D = 0) and exclusive dependent marking (since H = 0, an actual ratio cannot be computed as it has a zero denominator). The other ratios, whose without zeroes, run from 0.14 (two languages) to 8.00 (one language). … The other three frequency peaks suggest that preferred patterns cluster at perceptually simple ratios: two to one, one to one, and one to two. Overall, then , we have a preferecne for neatness of some sort: polar types, two-to-one ratios and even splits.' (Nichols 1992: 72-73)



# Statistically, this is not correct




Bars: Actual values from Nichols

Line: Statistically expected values

# Implicational Hierarchies

- $A \rightarrow B$
  $B \rightarrow C$
  $C \rightarrow D$

- $A \rightarrow (B \rightarrow (C \rightarrow D)) \equiv (A \wedge B \wedge C) \rightarrow D$

- $(A \rightarrow B) \wedge (B \rightarrow C) \wedge (C \rightarrow D)$

- $A > B > C > D$

- | | A | B | C | D |
  |---|---|---|---|---|
  | type 1: | + | + | + | + |
  | type 2: | – | + | + | + |
  | type 3: | – | – | + | + |
  | type 4: | – | – | – | + |
  | type 5: | – | – | – | – |

# Apparently, a hierarchy A > B > C > D

| | A | B | C | D | |
|---|---|---|---|---|---|
| 1 | + | + | + | + | **26** |
| 2 | š | + | + | + | **78** |
| 3 | š | š | + | + | **99** |
| 4 | š | š | š | + | **20** |
| 5 | š | š | š | š | **21** |
| 6 | + | š | + | + | **3** |
| 7 | š | + | š | + | **12** |
| 8 | š | š | + | š | **4** |
| 9 | + | š | š | + | 1 |
| 10 | š | + | + | š | 0 |
| 11 | + | + | š | + | 0 |
| 12 | + | š | + | š | 0 |
| 13 | š | + | š | š | 0 |
| 14 | + | + | + | š | 1 |
| 15 | + | + | š | š | 0 |
| 16 | + | š | š | š | 0 |
| Total + | 31 | 117 | 211 | 239 | |

| | A | B | C | D | deviation / standard dev. | |
|---|---|---|---|---|---|---|
| 1 | + | + | + | + | **+ 5.2** | more common than expected |
| 5 | š | š | š | š | **+ 11.5** | |
| 2 | š | + | + | + | + 0.5 | no significant deviation from expectation |
| 3 | š | š | + | + | + 0.7 | |
| 4 | š | š | š | + | - 0.9 | |
| 14 | + | + | + | š | - 0.1 | |
| 15 | + | + | š | š | - 0.6 | |
| 16 | + | š | š | š | - 0.6 | |
| 12 | + | š | + | š | - 1.2 | less common than expected |
| 7 | š | + | š | + | - 1.4 | |
| 9 | + | š | š | + | - 1.5 | |
| 13 | š | + | š | š | - 1.5 | |
| 11 | + | + | š | + | - 1.6 | |
| 8 | š | š | + | š | **- 2.0** | |
| 6 | + | š | + | + | **- 2.8** | |
| 10 | š | + | + | š | **- 2.9** | |

# From hierarchies to markedness

- **Independent frequency of the four parameters**

| | | | |
|---|---|---|---|
| A+ | 31 | A- | 234 |
| B+ | 117 | B- | 148 |
| C+ | 211 | C- | 54 |
| D+ | 239 | D- | 26 |

- **Hierarchy of frequencies**
  A+ < B+ < C+ < D+

- **High frequency interpreted as low markedness**
  A > B > C > D

**There is a markedness hierarchy iff**

- There is a significant interaction between the parameters, AND
- The differences in frequency between the independent frequencies are large

# The ‘large’ criterium is important

- **E.g. Hawkins word order data**
- **There is a significant interaction**
  (VO ~ Pr ~ NG ~ NA) versus (OV ~ Po ~ GN ~ AN)
- **Independent frequencies**

| | | | |
|---|---|---|---|
| VO | 162 | OV | 174 |
| Pr | 148 | Po | 188 |
| NG | 145 | GN | 191 |
| NA | 187 | AN | 149 |

- **Hierarchy of frequencies**
  NG < Pr < VO < NA
- **But no markedness hierarchy, because the differences between the frequencies are not large!**

# Reinterpreting implicational universals

- **Implicational universals can be seen as small hierarchies**

- **An implicational universal A → B**

| | | A | |
|---|---|---|---|
| | | + | Š |
| B | + | **$X_1$** | **$X_2$** |
| | Š | **Æ** | **$X_3$** |

- **An implicational universal as a hierarchy with two parameters**

| | A | B | |
|---|---|---|---|
| type 1: | + | + | attested with frequency $X_1$ |
| type 2: | – | + | attested with frequency $X_2$ |
| type 3: | – | – | attested with frequency $X_3$ |
| type 4: | + | – | unattested |

**An implicational universal A → B is a markedness hierarchy A > B iff:**

- There is a significant interaction between the parameters A an B, AND
- A+ is much smaller than B+.

# The typological versus the statistical view

The traditional logic of the implicational universal stressed the frequency difference. The statistical interpretation stresses the significant interaction, and thereby possibly declares a distribution as interesting, although there is no frequency difference and thus no implicational universal.

**No significant interaction,**
**but a large frequency difference A+ << B+**

| | | A | | |
|---|---|---|---|---|
| | | + | Š | total |
| B | + | **10** | **31** | 41 |
| | Š | **2** | **12** | 14 |
| | total | 12 | 43 | 55 |

**A significant interaction,**
**but no frequency difference between A+ and B**

| | | A | | |
|---|---|---|---|---|
| | | + | Š | total |
| B | + | **17** | **10** | 27 |
| | Š | **9** | **19** | 28 |
| | total | 26 | 29 | 55 |

# A problem for the interpretation

**A typological distribution with (apparently) 4 major types (dark grey) and 4 minor types (light grey).**

| | Independent pronouns | | | | | |
|---|---|---|---|---|---|---|
| | no *we* | *we* identical to *I* | unified *we* | only inclusive *we* | inclusive+ exclusive *we* | |
| no person marking | 1 | 5 | 36 | 1 | 27 | 70 |
| *we* identical to *I* | 1 | 1 | 9 | 0 | 1 | 12 |
| unified *we* | 0 | 2 | 75 | 0 | 2 | 79 |
| only inclusive *we* | 0 | 0 | 0 | 4 | 5 | 9 |
| inclusive and exclusive *we* | 0 | 2 | 0 | 0 | 28 | 30 |
| | 2 | 10 | 120 | 5 | 63 | 200 |

# However, common is not necessary interesting!

**Major deviations from expectation. The positive deviations are shaded dark grey (highly significant) and light grey (slightly significant)**

| | Independent pronouns | | | | |
|---|---|---|---|---|---|
| | no *we* | *we* identical to *I* | unified *we* | only inclusive *we* | inclusive+ exclusive *we* |
| no person marking | | + 1.5 | - 6.0 | | + 5.0 |
| *we* identical to *I* | | | + 1.8 | | - 2.8 |
| unified *we* | | - 1.9 | + 27.6 | - 2.0 | - 22.9 |
| only inclusive *we* | | | - 5.4 | + 3.8 | +2.2 |
| inclusive and exclusive *we* | | | - 18.0 | | + 18.5 |

# Summary

- The actual variation is not necessarily related to the possible variation
- Genus-based sampling is only a bottom-line assurance of variability
- The existence of large uniform areas show that there are super-genetic consistencies, which devaluate genus-based samples
- Beware of numbers! High frequencies do not necessarily mean that the feature is important for a theory of linguistic structure