Power laws in linguistic typology

Gerhard Jäger
gerhard.jaeger@uni-tuebingen.de

March 12, 2010

11th Szklarska Poreba Workshop

Eberhard Karls Universität Tübingen

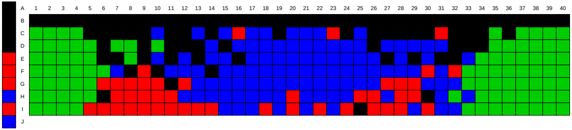
sfs
seminar für sprachwissenschaft

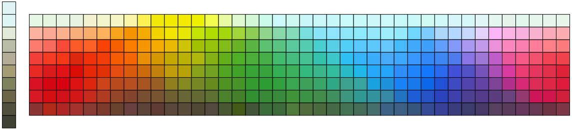
The World Color Survey

- started by Paul Kay and co-workers; traces back to Berlin & Kay 1969
- investigation of color vocabulary of 110 non-written languages from around the world
- around 25 informants per language
- two tasks:
  - the 330 Munsell chips were presented to each test person one after the other in random order; they had to assign each chip to some basic color term from their native language
  - for each native basic color term, each informant identified the prototypical instance(s)
- data are publicly available under http://www.icsi.berkeley.edu/wcs/

Raw data

- are irregular and noisy
- example: randomly picked test person (native language: Piraha)
- 1,771 such data points in total





Statistical feature extraction

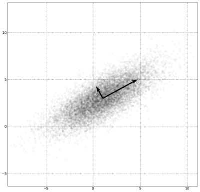
- first step: representation of raw data in *contingency matrix*
    - rows: color terms from various languages
    - columns: Munsell chips
    - cells: number of test persons who used the row-term for the column-chip

	A0	B0	B1	B2	⋯	I38	I39	I40	J0
red	0	0	0	0	⋯	0	0	2	0
green	0	0	0	0	⋯	0	0	0	0
blue	0	0	0	0	⋯	0	0	0	0
black	0	0	0	0	⋯	18	23	21	25
white	25	25	22	23	⋯	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
rot	0	0	0	0	⋯	1	0	0	0
grün	0	0	0	0	⋯	0	0	0	0
gelb	0	0	0	1	⋯	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
rouge	0	0	0	0	⋯	0	0	0	0
vert	0	0	0	0	⋯	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- further processing:
    - divide each row by the number n of test persons using the corresponding term
    - duplicate each row n times

Statistical feature extraction: PCA

- technique to reduce dimensionality of data
- input: set of vectors in an n-dimensional space



first step:

- rotate the coordinate system, such that
  - the new n coordinates are orthogonal to each other
  - the variations of the data along the new coordinates are stochastically independent

second step:

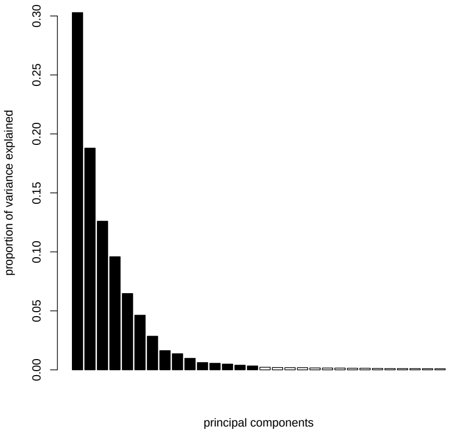
- choose a suitable $m < n$
- project the data on those m new coordinates where the data have the highest variance

Statistical feature extraction: PCA

- alternative formulation:
  - choose an m-dimensional linear sub-manifold of your n-dimensional space
  - project your data onto this manifold
  - when doing so, pick your sub-manifold such that the average squared distance of the data points from the sub-manifold is minimized
- intuition behind this formulation:
  - data are "actually" generated in an m-dimensional space
  - observations are disturbed by n-dimensional noise
  - PCA is a way to reconstruct the underlying data distribution
- applications: picture recognition, latent semantic analysis, statistical data analysis in general, data visualization, ...

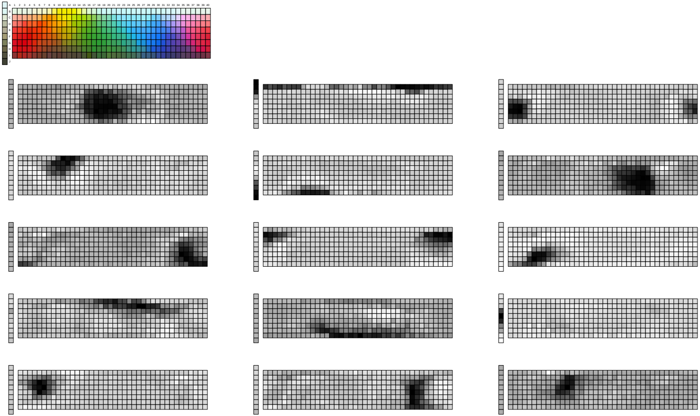
Statistical feature extraction: PCA

- first 15 principal components jointly explain 91.6% of the total variance
- choice of $m = 15$ is determined by using "Kaiser's stopping rule"



Statistical feature extraction: PCA

after some post-processing ("varimax" algorithm):



Implicative universals

- first six features correspond nicely to the six primary colors *white, black, red, green, blue, yellow*
- according to Kay et al. (1997) (and many other authors) simple system of **implicative universals** regarding possible partitions of the primary colors

Implicative universals

I	II	III	IV	V
		white red/yellow green/blue black	white red yellow green/blue black	
white/red/yellow black/green/blue	white red/yellow black/green/blue	white red/yellow green black/blue		white red yellow green blue black
		white red yellow black/green/blue	white red yellow green black/blue	
		white red yellow/green/blue black	white red yellow/green blue black	
		white red yellow/green black/blue		

source: Kay et al. (1997)

Eberhard Karls Universität Tübingen

seminar für sprachwissenschaft

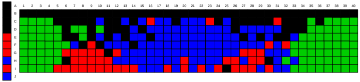
Partition of the primary colors

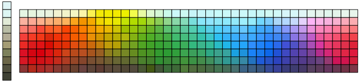
- each speaker/term pair can be projected to a 15-dimensional vector
- primary colors correspond to first 6 entries
- each primary color is assigned to the term for which it has the highest value
- defines for each speaker a partition over the primary colors

Partition of the primary colors

- for instance: sample speaker from Piraha (see above):
- extracted partition:

$$\begin{bmatrix} \text{white/yellow} \\ \text{red} \\ \text{green/blue} \\ \text{black} \end{bmatrix}$$





- supposedly impossible, but occurs 61 times in the database

Partition of primary colors

- most frequent partition types:
  1. {white}, {red}, {yellow}, {green, blue}, {black} (41.9%)
  2. {white}, {red}, {yellow}, {green}, {blue}, {black} (25.2%)
  3. {white}, {red, yellow}, {green, blue, black} (6.3%)
  4. {white}, {red}, {yellow}, {green}, {black, blue} (4.2%)
  5. {white, yellow}, {red}, {green, blue}, {black} (3.4%)
  6. {white}, {red}, {yellow}, {green, blue, black} (3.2%)
  7. {white}, {red, yellow}, {green, blue}, {black} (2.6%)
  8. {white, yellow}, {red}, {green, blue, black} (2.0%)
  9. {white}, {red}, {yellow}, {green, blue, black} (1.6%)
  10. {white}, {red}, {green, yellow}, {blue, black} (1.2%)

Partition of primay colors

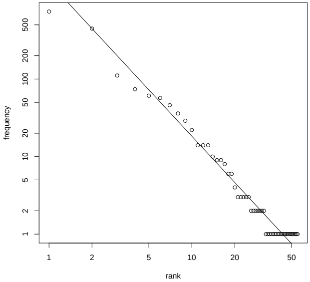
- 87.1% of all speaker partitions obey Kay et al.'s universals
- the ten partitions that confirm to the universals occupy ranks 1, 2, 3, 4, 6, 7, 9, 10, 16, 18
- decision what counts as an exception seems somewhat arbitrary on the basis of these counts

Partition of primary colors

- more fundamental problem:
  - partition frequencies are distributed according to **power law**

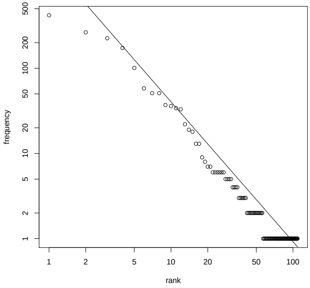
$$frequency \sim rank^{-1.99}$$

- no natural cutoff point to distinguish regular from exceptional partitions



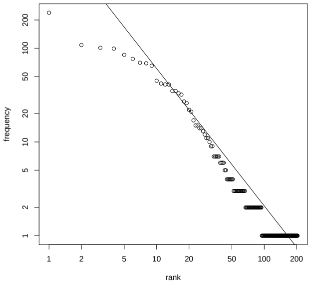
Partition of seven most important colors

$$frequency \sim rank^{-1.64}$$

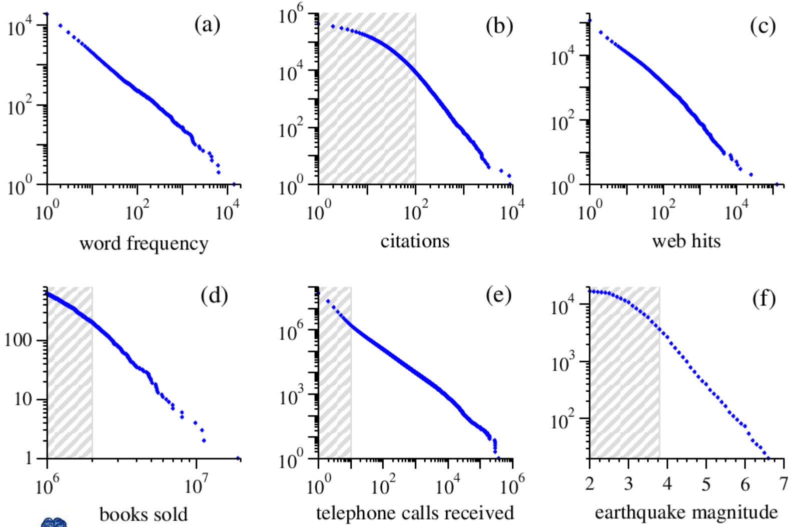


Partition of eight most important colors

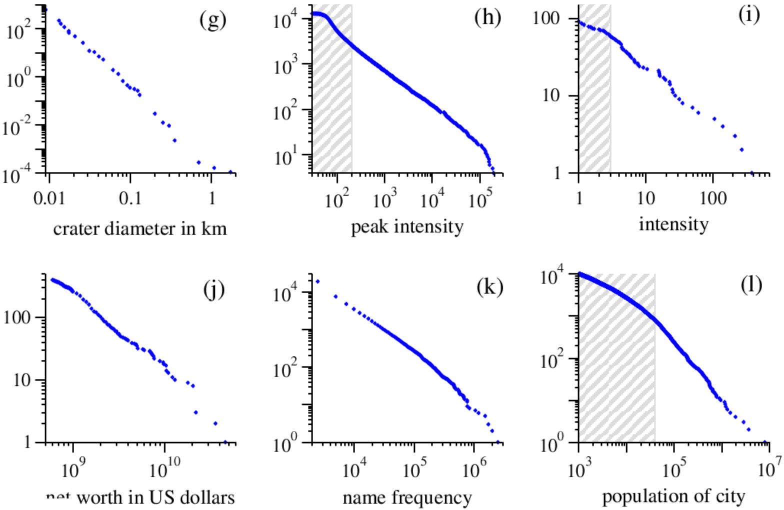
$$frequency \sim rank^{-1.46}$$



Power laws


(a)
(b)
(c)
(d)
(e)
(f)
word frequency
citations
web hits
books sold
telephone calls received
earthquake magnitude

Power laws



(g)
10^2
10^0
10^{-2}
10^{-4}
0.01
0.1
1
crater diameter in km
(h)
10^4
10^3
10^2
10^1
10^2
10^3
10^4
10^5
peak intensity
(i)
100
10
1
1
10
100
intensity
(j)
100
10
1
10^9
10^{10}
net worth in US dollars
(k)
10^4
10^2
10^0
10^4
10^5
10^6
name frequency
(l)
10^4
10^2
10^0
10^3
10^5
10^7
population of city

Power laws

FIG. 4 Cumulative distributions or "rank/frequency plots" of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel *Moby Dick* by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. (c) Numbers of hits on web sites by 60 000 users of the America Online Internet service for the day of 1 December 1997. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10 000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000.

from Newman 2006

Power laws are **not** everywhere

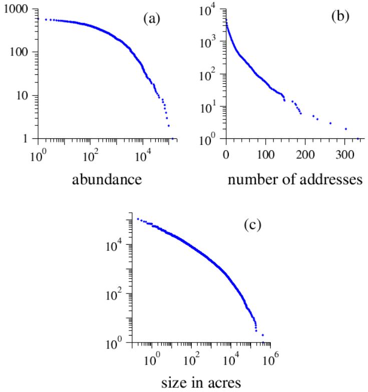


FIG. 5 Cumulative distributions of some quantities whose distributions span several orders of magnitude but that nonetheless do not follow power laws. (a) The number of sightings of 591 species of birds in the North American Breeding Bird Survey 2003. (b) The number of addresses in the email address books of 16 881 users of a large university computer system [33]. (c) The size in acres of all wildfires occurring on US federal land between 1986 and 1996 (National Fire Occurrence Database, USDA Forest Service and Department of the Interior). Note that the horizontal axis is logarithmic in frames (a) and (c) but linear in frame (b).

Are color naming systems power law distributed?

- free software by Aaron Clauset, based on Clauset et al. 2009
- performs Kolmogorov-Smirnov test
- result:
  - power law hypothesis cannot be rejected
  - jury is still out whether power law is a better fit than alternative distributions like log-normal distribution
  - anybody here who *really* knows how to do these things?

Other linguistic power law distributions

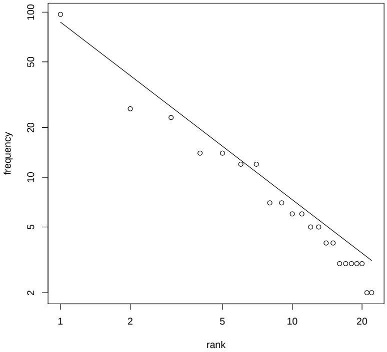
number of vowels	vowel systems and their frequency of occurrence				
3	14				
4	14	5	4	2	
5	97	3			
6	26	12	12		
7	23	6	5	4	3
8	6	3	3	2	
9	7	7	3		

(from Schwartz et al. 1997,

based on the UCLA Phonetic Segment Inventory Database)

Other linguistic power law distributions

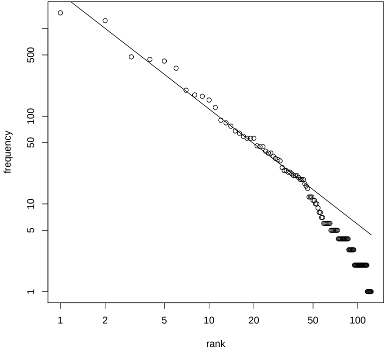
$$frequency \sim rank^{-1.06}$$



Other linguistic power law distributions

- size of language families
- source: Ethnologue

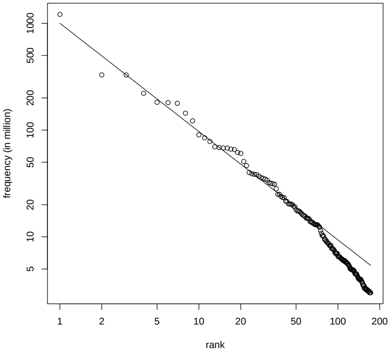
$$frequency \sim rank^{-1.32}$$



Other linguistic power law distributions

- number of speakers per language
- source: Ethnologue

$$frequency \sim rank^{-1.01}$$

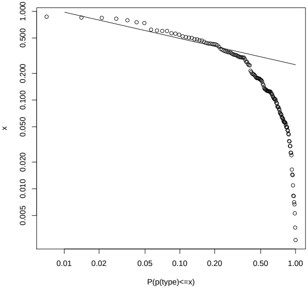


The World Atlas of Language Structures

- large scale typological database, conducted mainly by the MPI EVA, Leipzig
- 2,650 languages in total are used
- 142 features, with between 120 and 1,370 languages per feature
- available online

The World Atlas of Language Structures

- Maslova 2008, "Meta-typological distributions"
- hypothesis:
  - pick a random value for each feature
  - estimate the probability that a random language has this value
  - the likelihood that an arbitrarily chosen feature value has a probability x is proportional to a power of x



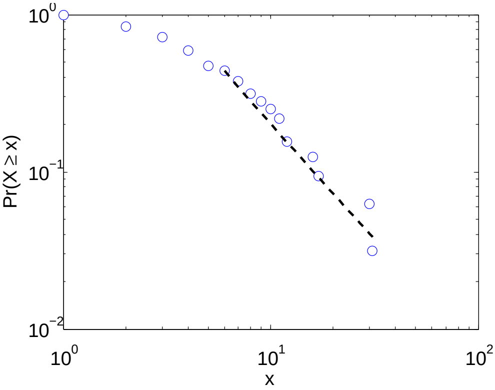
- only holds for the most frequent 30% of all types
  - for the entire range of type frequencies, the hypothesis can be rejected

The World Atlas of Language Structures

- however, Maslova is perhaps right in the assumption that languages are power-law distributed across WALS types
- worth to test it within features rather than across features
- problem: number of feature values usually too small for statistic evaluation
- solution:
  - cross-classification of two (randomly chosen) features
  - only such feature pairs are considered that lead to at least 30 non-empty feature value combinations
- pilot study with 10 such feature pairs

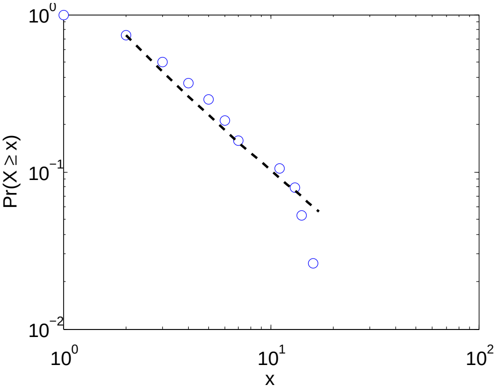
The World Atlas of Language Structures

- Feature 1: Consonant-Vowel Ratio
- Feature 2: Subtypes of Asymmetric Standard Negation
- Kolmogorov-Smirnov test: positive



The World Atlas of Language Structures

- Feature 1: Weight Factors in Weight-Sensitive Stress Systems
- Feature 2: Ordinal Numerals
- Kolmogorov-Smirnov test: positive

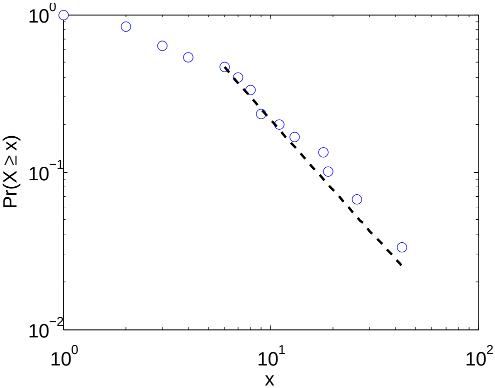


EBERHARD KARLS UNIVERSITÄT TÜBINGEN

sfs seminar für sprachwissenschaft

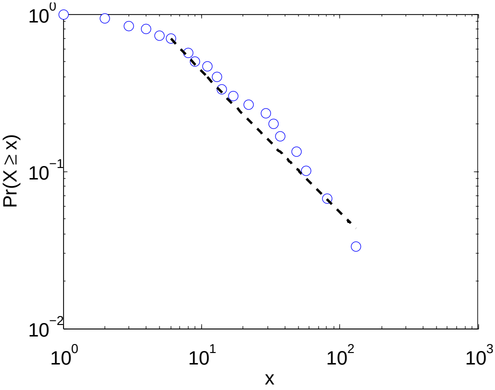
The World Atlas of Language Structures

- Feature 1: Third Person Zero of Verbal Person Marking
- Feature 2: Subtypes of Asymmetric Standard Negation
- Kolmogorov-Smirnov test: positive



The World Atlas of Language Structures

- Feature 1: Relationship between the Order of Object and Verb and the Order of Adjective and Noun
- Feature 2: Expression of Pronominal Subjects
- Kolmogorov-Smirnov test: positive

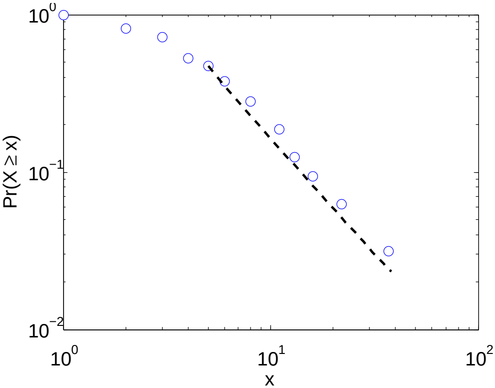


Eberhard Karls Universität Tübingen

sfs seminar für sprachwissenschaft

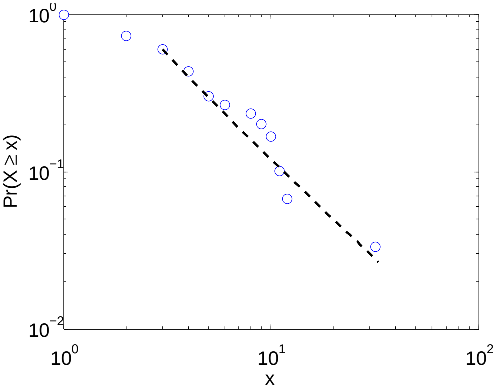
The World Atlas of Language Structures

- Feature 1: Plurality in Independent Personal Pronouns
- Feature 2: Asymmetrical Case-Marking
- Kolmogorov-Smirnov test: positive



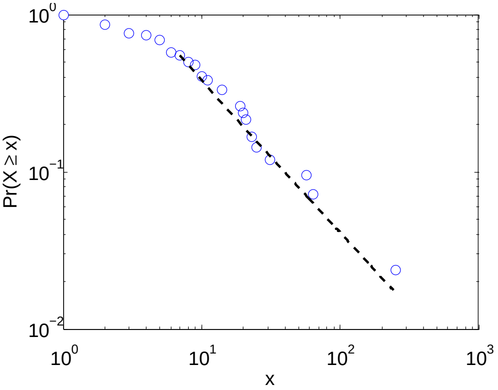
The World Atlas of Language Structures

- Feature 1: Locus of Marking: Whole-language Typology
- Feature 2: Number of Cases
- Kolmogorov-Smirnov test: positive



The World Atlas of Language Structures

- Feature 1: Prefixing vs. Suffixing in Inflectional Morphology
- Feature 2: Coding of Nominal Plurality
- Kolmogorov-Smirnov test: positive

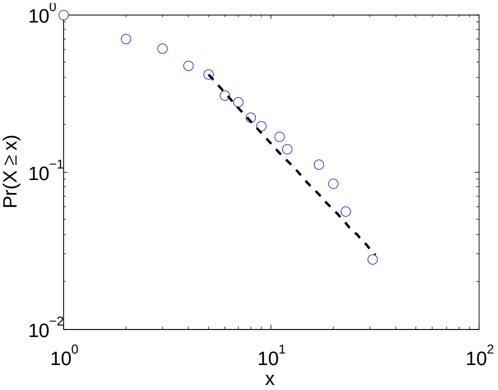


Eberhard Karls Universität Tübingen

sfs seminar für sprachwissenschaft

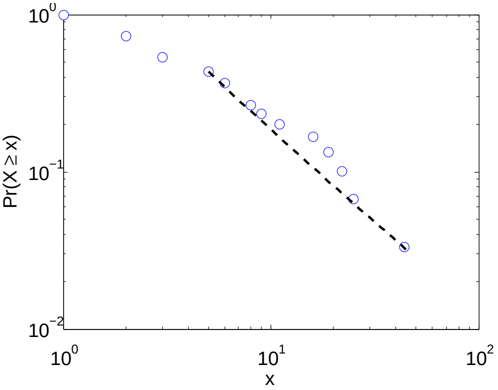
The World Atlas of Language Structures

- Feature 1: Prefixing vs. Suffixing in Inflectional Morphology
- Feature 2: Ordinal Numerals
- Kolmogorov-Smirnov test: positive



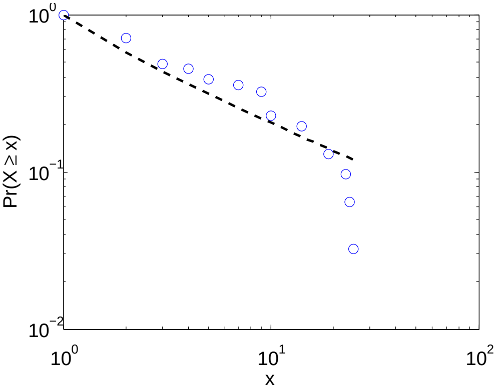
The World Atlas of Language Structures

- Feature 1: Coding of Nominal Plurality
- Feature 2: Asymmetrical Case-Marking
- Kolmogorov-Smirnov test: positive



The World Atlas of Language Structures

- Feature 1: Position of Case Affixes
- Feature 2: Ordinal Numerals
- Kolmogorov-Smirnov test: negative



Why power laws?

- critical states
- self-organized criticality
- preferential attachment
- random walks
- ...

Preferential attachment

- items are stochastically added to bins
- probability to end up in bin n is linear in number of items that are already in bin n

(Wide) Open questions

- Preferential attachment explains power law distribution *if there are no a priori biases for particular types*
- first simulations suggest that preferential attachment + biased type assignment does **not** lead to power law
- negative message: uneven typological frequency distribution does not prove that frequent types are inherently preferred linguistically/cognitively/socially
- unsettling questions:
  - Are there linguistic/cognitive/social biases in favor of certain types?
  - If yes, can statistical typology supply information about this?
  - If power law distributions are the norm, is their any content to the notion of *statistical universal* in a Greenbergian sense?