

# Basic Word Order Frequencies and Transition Probabilities in the Languages of the World

Harald Hammarström  
harald2@cs.chalmers.se

June 4, 2008

# The Search for Language Universals

- Possible *Universal* features of human language have interested scholars from the earliest times, more so since the 19th century and intensively since the 1970s
  - Do all languages have definite/indefinite articles?
  - Do all languages employ nasal consonants?
  - Are verb tenses found in all languages?
  - Are suffixes more common than prefixes?
  - Do all languages Obligatory plural marking on nouns?
  - Is there a universal order of noun and determiner?
  - ...

# The Value of Universal Tendencies

- As knowledge of the world's languages grew, it became apparent that there are hardly any interesting *absolute* (= exceptionless) universals
- However, strong universal *tendencies* (aka near-universals or statistical universals) are found
- Goal is to find tendencies and plausible explanations for them, e.g.,

Non-Historical:	Historical:
-Innate specification	-Common ancestry
-Processing preferences	-Areal contact
-Communicative needs	
-...	

# Traditional Approach

- Sample a large number of languages to arrive at a *frequency distribution*
- Try discount for areal and genealogical biases

Possible problem?

- Are all areal/genetic “historical accidents” detectable?
  - Dependencies that outlive the genealogical time-window?
  - Large-scale areal relationships?
- Whether or not this is an actual problem, one solution to it has already been suggested

# Dynamic Model for Lg Universals: Idea

Maslova (2000), Cysouw (2007) [or Straw Man (2008)]

NOT frequencies INSTEAD transition probabilities

- Look at languages of the same family and estimate the likelihood of change from one feature to another
- A set of transition probabilities determine a unique stationary frequency distribution!

Advantages:

- Immune to historical accidents
- Easier to gather
- Elegant diachronic-synchronic mirror

The probability of a language having a certain type at some random moment in its *history*  $\approx$  the probability of that a random language in a large language *population* has the type

# Dynamic Model for Lg Universals: Model

- For every linguistic feature there is a *constant probability of change* (CPCH) from one value to another

	Nominative-Accusative	Ergative
Nominative-Accusative	0.8	0.2
Ergative	0.8	0.2

- Languages die
- Languages split or “clone”

Birth and death processes are *independent* of feature values

# **Aim of This Talk**

---

Argue that CPCH is not valid

This is not so easy...

- Model is underspecified for transition unit time
- Model is underspecified for actual values of birth- death-rates
- For most of the languages in the world, we don't have historical data

# Data on Basic Word Order

## 1. **Ethnologue:** 1097 data points

- Sources for the data points are not indicated.
- It is not clear how the data points/languages were selected.

## 2. **WALS:** 1203 data points

- Sources for the data points are indicated.
- It is not clear how the data points/languages were selected, but it may be guessed that it is some kind of convenience sample.

## 3. **Hammarström:** 338 data points

- Sources for the data points are indicated.
- Languages were sampled *at random*, one for *every* attested language family in the world.

Total [without overlap] 2086 languages from 338 families/isolates!

	All 2086		Hammarström		Isolates		Majority	
SOV	977	<b>46.8%</b>	208	<b>61.5%</b>	121	<b>61.1%</b>	86	<b>61.4%</b>
SVO	659	<b>31.5%</b>	49	<b>14.4%</b>	28	<b>14.1%</b>	25	<b>17.8%</b>
NODOM	166	<b>7.9%</b>	30	<b>8.8%</b>	17	<b>8.5%</b>	11	<b>7.8%</b>
VSO	181	<b>8.6%</b>	21	<b>6.2%</b>	12	<b>6.0%</b>	9	<b>6.4%</b>
VOS	46	<b>2.2%</b>	9	<b>2.6%</b>	6	<b>3.0%</b>	3	<b>2.1%</b>
OVS	14	<b>0.6%</b>	6	<b>1.7%</b>	3	<b>1.5%</b>	2	<b>1.4%</b>
VSO/VOS	9	<b>0.4%</b>	7	<b>2.0%</b>	6	<b>3.0%</b>	0	<b>0.0%</b>
OSV	13	<b>0.6%</b>	1	<b>0.2%</b>	1	<b>0.5%</b>	2	<b>1.4%</b>
SVO/VSO	6	<b>0.2%</b>	2	<b>0.5%</b>	1	<b>0.5%</b>	1	<b>0.7%</b>
SOV/OVS	4	<b>0.1%</b>	2	<b>0.5%</b>	2	<b>1.0%</b>	0	<b>0.0%</b>
SVO/VOS	6	<b>0.2%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>
SOV/OSV	2	<b>0.0%</b>	2	<b>0.5%</b>	0	<b>0.0%</b>	1	<b>0.7%</b>
SOV/SVO	2	<b>0.0%</b>	1	<b>0.2%</b>	1	<b>0.5%</b>	0	<b>0.0%</b>
SOV/VOS	1	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>
	2086		338		198		140	

	Africa		Austr.		Eurasia		N Amer.		Papua		S Amer.		Total	
SOV	16	<b>45.7%</b>	10	<b>38.4%</b>	26	<b>83.8%</b>	30	<b>46.1%</b>	72	<b>78.2%</b>	56	<b>62.2%</b>	210	<b>61.9%</b>
SVO	10	<b>28.5%</b>	4	<b>15.3%</b>	3	<b>9.6%</b>	4	<b>6.1%</b>	16	<b>17.3%</b>	11	<b>12.2%</b>	48	<b>14.1%</b>
NODOM	0	<b>0.0%</b>	10	<b>38.4%</b>	1	<b>3.2%</b>	13	<b>20.0%</b>	2	<b>2.1%</b>	4	<b>4.4%</b>	30	<b>8.8%</b>
VSO	6	<b>17.1%</b>	0	<b>0.0%</b>	1	<b>3.2%</b>	8	<b>12.3%</b>	1	<b>1.0%</b>	5	<b>5.5%</b>	21	<b>6.1%</b>
VOS	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	4	<b>6.1%</b>	0	<b>0.0%</b>	5	<b>5.5%</b>	9	<b>2.6%</b>
VSO/VOS	0	<b>0.0%</b>	1	<b>3.8%</b>	0	<b>0.0%</b>	3	<b>4.6%</b>	0	<b>0.0%</b>	3	<b>3.3%</b>	7	<b>2.0%</b>
OVS	0	<b>0.0%</b>	1	<b>3.8%</b>	0	<b>0.0%</b>	1	<b>1.5%</b>	0	<b>0.0%</b>	4	<b>4.4%</b>	6	<b>1.7%</b>
SOV/OSV	1	<b>2.8%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>1.0%</b>	1	<b>1.1%</b>	3	<b>0.8%</b>
SVO/VSO	2	<b>5.7%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	2	<b>0.5%</b>
SOV/OVS	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>1.5%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>0.2%</b>
OSV	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>1.1%</b>	1	<b>0.2%</b>
SOV/SVO	0	<b>0.0%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>1.5%</b>	0	<b>0.0%</b>	0	<b>0.0%</b>	1	<b>0.2%</b>
	35 <b>(10.3%)</b>		26 <b>(7.6%)</b>		31 <b>(9.1%)</b>		65 <b>(19.1%)</b>		92 <b>(27.1%)</b>		90 <b>(26.5%)</b>		339	

SOV-majority	n	SOV	SVO	NODOM	VSO	VOS	Other
Sino-Tibetan	172	<b>91.2%</b>	<b>8.7%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Indo-European	106	<b>52.8%</b>	<b>33.9%</b>	<b>7.5%</b>	<b>5.6%</b>	<b>0.0%</b>	<b>0.0%</b>
Trans New Guinea	94	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Afro-Asiatic	85	<b>44.7%</b>	<b>40.0%</b>	<b>5.8%</b>	<b>9.4%</b>	<b>0.0%</b>	<b>0.0%</b>
Pama-Nyungan	57	<b>47.3%</b>	<b>12.2%</b>	<b>31.5%</b>	<b>0.0%</b>	<b>3.5%</b>	<b>5.2%</b>
Quechuan	42	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Uto-Aztecan	31	<b>41.9%</b>	<b>22.5%</b>	<b>19.3%</b>	<b>12.9%</b>	<b>0.0%</b>	<b>3.2%</b>
Omotic	23	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Turkic	20	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Tupi	18	<b>50.0%</b>	<b>33.3%</b>	<b>0.0%</b>	<b>5.5%</b>	<b>0.0%</b>	<b>11.1%</b>
Uralic	16	<b>50.0%</b>	<b>37.5%</b>	<b>12.5%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Mande	16	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Sepik	14	<b>92.8%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>7.1%</b>
Chibchan	14	<b>92.8%</b>	<b>0.0%</b>	<b>7.1%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Tucanoan	12	<b>66.6%</b>	<b>0.0%</b>	<b>8.3%</b>	<b>8.3%</b>	<b>0.0%</b>	<b>16.6%</b>
Panoan	12	<b>83.3%</b>	<b>0.0%</b>	<b>16.6%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Dravidian	12	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Siouan	10	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Nakh-Dagestanian	10	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Mongolian	10	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>

SVO-majority	n	SOV	SVO	NODOM	VSO	VOS	Other
Austronesian	240	<b>7.9%</b>	<b>64.1%</b>	<b>6.2%</b>	<b>12.0%</b>	<b>5.8%</b>	<b>3.7%</b>
Atlantic-Congo	201	<b>4.4%</b>	<b>91.5%</b>	<b>3.9%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Mayan	50	<b>0.0%</b>	<b>44.0%</b>	<b>4.0%</b>	<b>26.0%</b>	<b>16.0%</b>	<b>10.0%</b>
Austroasiatic	32	<b>6.2%</b>	<b>84.3%</b>	<b>6.2%</b>	<b>0.0%</b>	<b>3.1%</b>	<b>0.0%</b>
Tai-Kadai	21	<b>4.7%</b>	<b>95.2%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Central Sudanic	21	<b>0.0%</b>	<b>71.4%</b>	<b>28.5%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Torricelli	9	<b>0.0%</b>	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Miao-Yao	9	<b>11.1%</b>	<b>88.8%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Totonacan	5	<b>0.0%</b>	<b>60.0%</b>	<b>0.0%</b>	<b>40.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Songhay	5	<b>40.0%</b>	<b>40.0%</b>	<b>20.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
North Halmahera	5	<b>40.0%</b>	<b>60.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Guaicuruan	5	<b>0.0%</b>	<b>80.0%</b>	<b>20.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Zaparoan	3	<b>33.3%</b>	<b>66.6%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Koman	3	<b>0.0%</b>	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
Iwaidjan Proper	3	<b>0.0%</b>	<b>100.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
...							

# Can the CPCH be tested?

---

Observation as to Stationary Distributions:

- If CPCH, then there is a stationary distribution
- A stationary distribution may exist even if CPCH is not true

Observation as to availability of historical data:

- Historical data is unavailable (at present) for almost all the families at hand

Observation as to distorting power of birth/death effects:

- Unbounded appeal to birth/death effects can make any data consistent with CPCH

# Line of Attack: Idea

- If CPCH then each language *family* should drift towards the stationary distribution internally
- More specifically: the larger the family, the closer to the stationary distribution
- However: This drift is perturbed by birth (and death) effects
- Even allowing for birth effects: We do *not* observe such a drift

Note: A family-internal distribution depends on:

- Initial feature value of the ursprache
- The CPCH probability matrix
- The topology of the family tree

# Line of Attack: Implementation

- Assume the simplest kind of transition matrix for the stationary distribution
- Assume a very strong form of perturbation power by birth effects
- Assume the transition unit is of the same order as the birth-unit
- (With these assumptions and the real data, the initial feature of the ursprache matters little)
- Simulate families that correspond in number and size with observed world families
- Measure how much perturbation can be expected in the drift towards the stationary distribution
- The observed families show much more perturbation
- => It is highly unlikely that CPCH is true

# Distance between Distributions

---

- Define the distance between two distributions:

$$\|p_1 - p_2\| = \frac{\sum_c |p_1(c) - p_2(c)|}{2}$$

- E.g., let

	SOV	SVO	NODOM	...
$p_1$	0.8	0.2	0.0	...
$p_2$	0.2	0.8	0.0	...
$U$	0.62	0.14	0.09	...

- Then

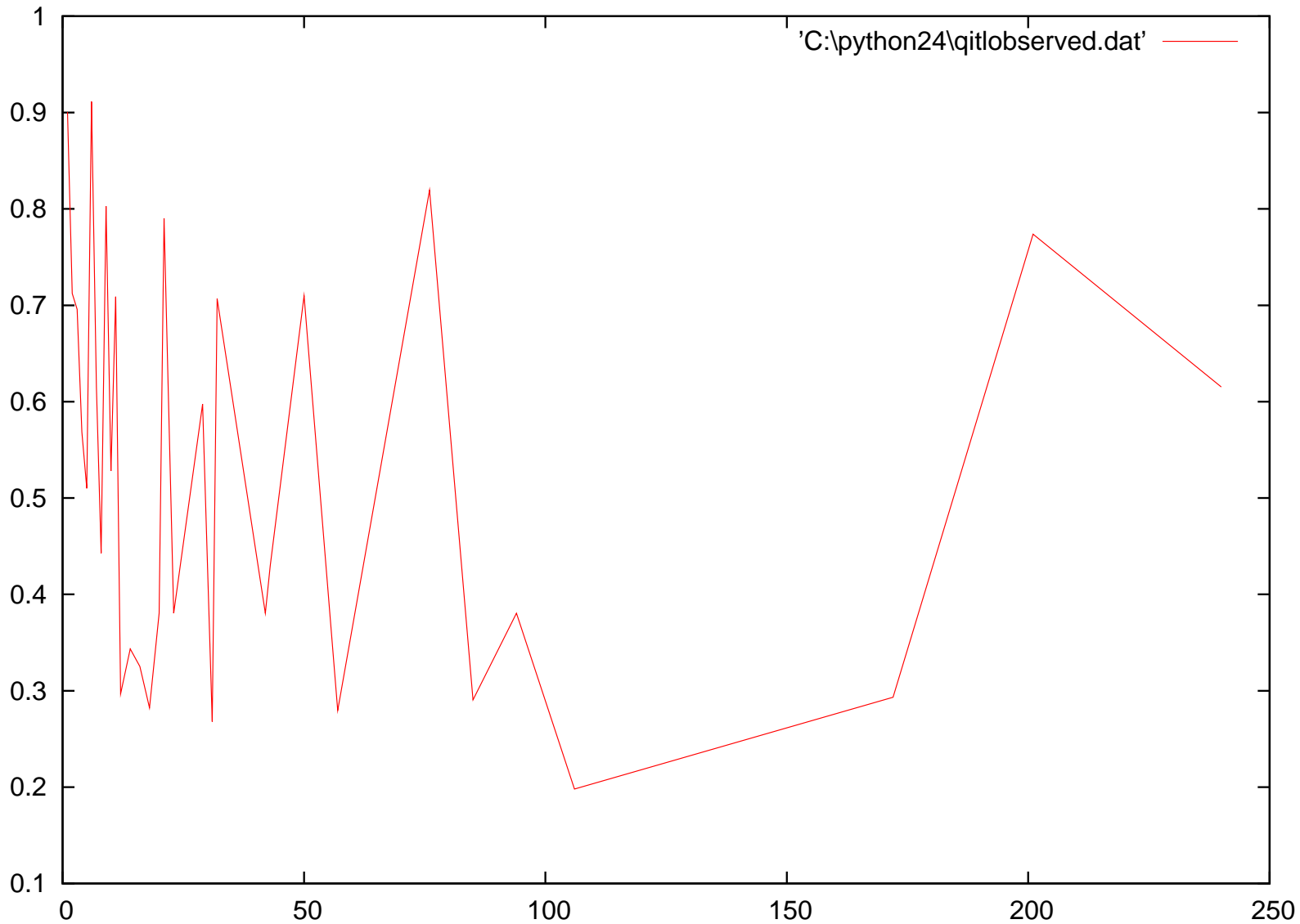
$$\|p_1 - p_2\| = 0.6$$

$$\|p_1 - U\| = 0.23$$

$$\|p_2 - U\| = 0.66$$

# Distances for Real World Families

---



# Simulation of World Families

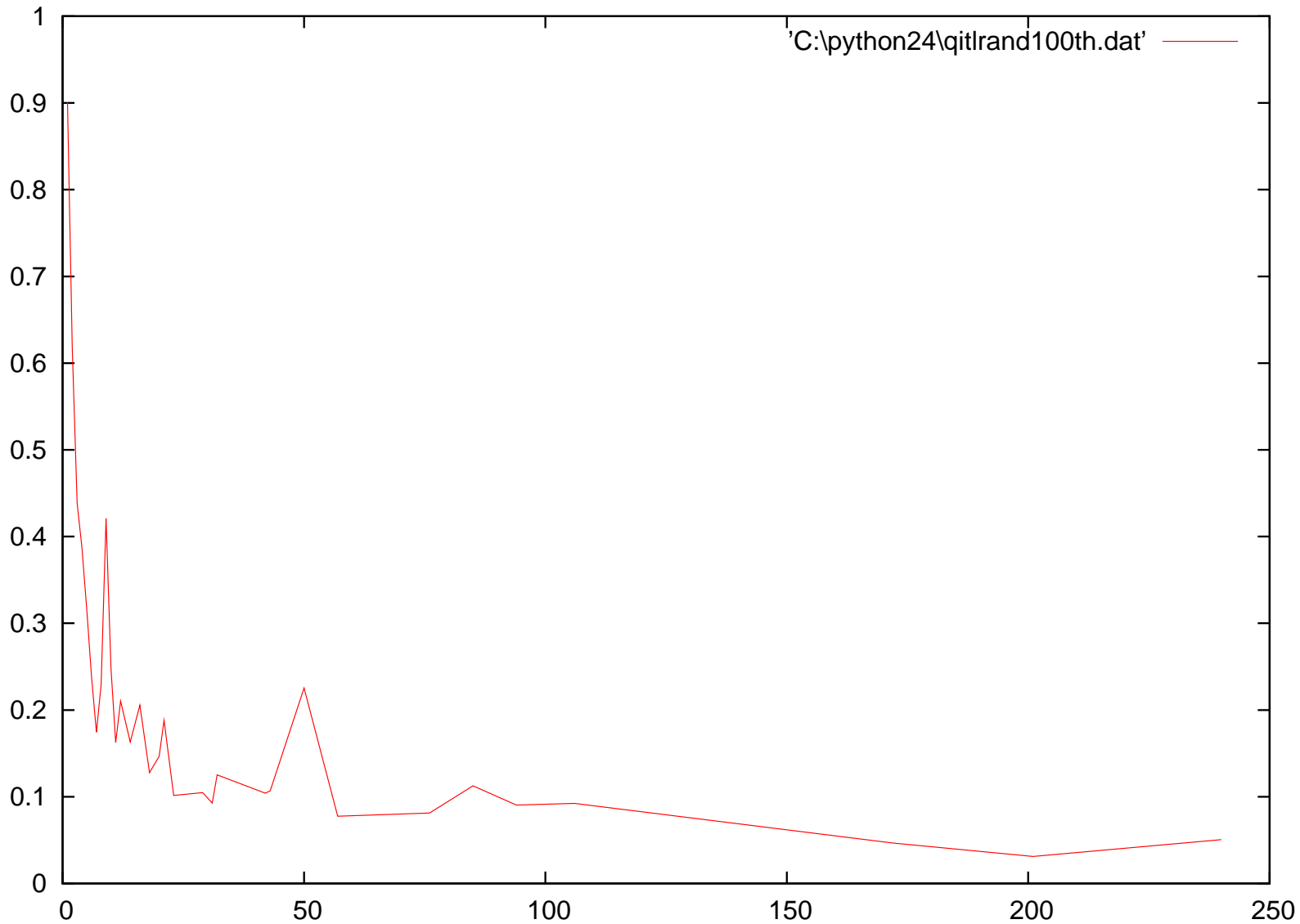
---

One Run =

1. For each real world family  $f$  of size  $n$
2. Generate a simulated family of the same size like this:
  - (a) Start with one language/feature value in the pool
  - (b) Repeat until pool size ==  $n$ :
    - i. Clone every language in the pool
    - ii. Apply the matrix transitions to each language in the pool

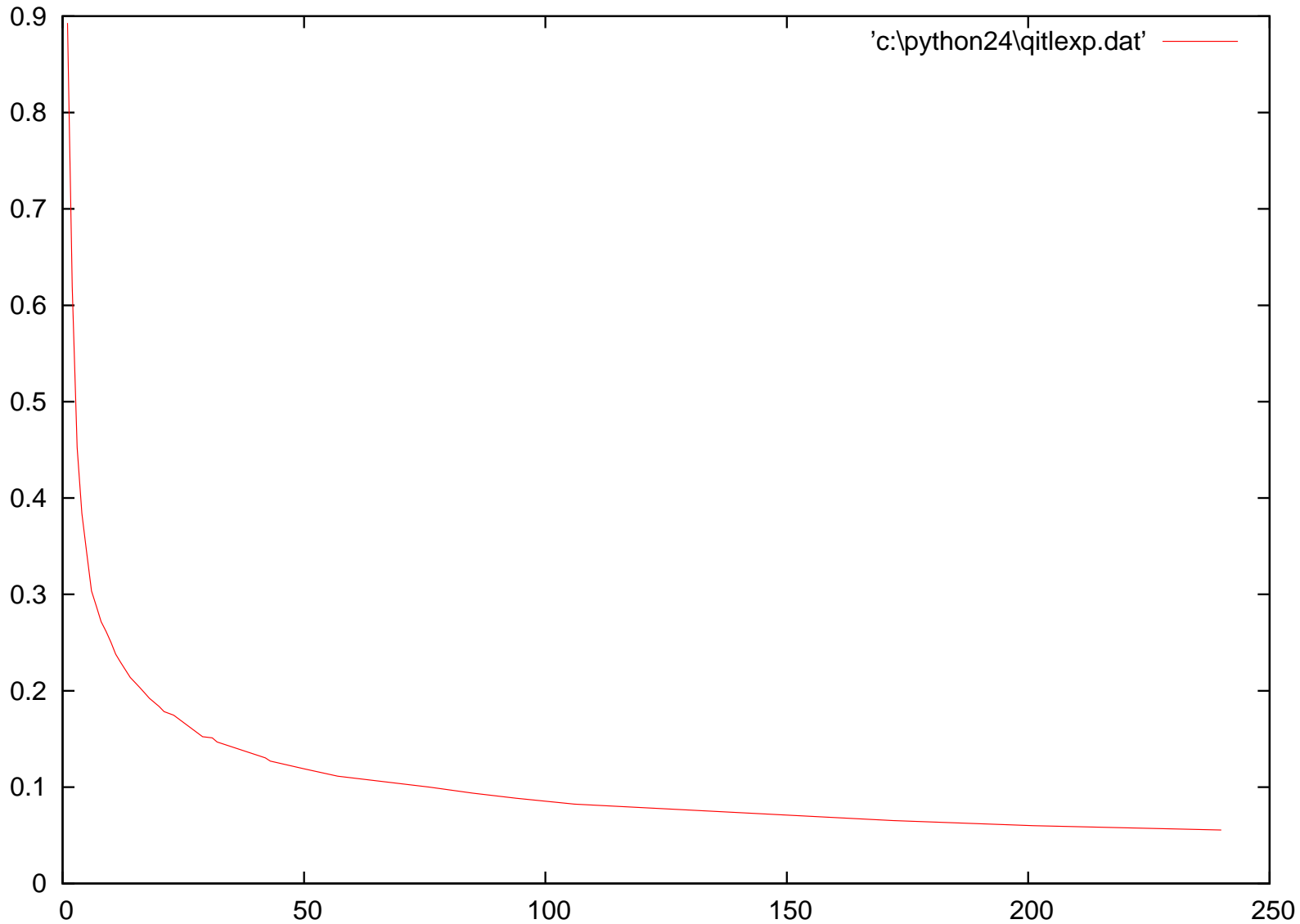
# Example Outcome of One Run

---



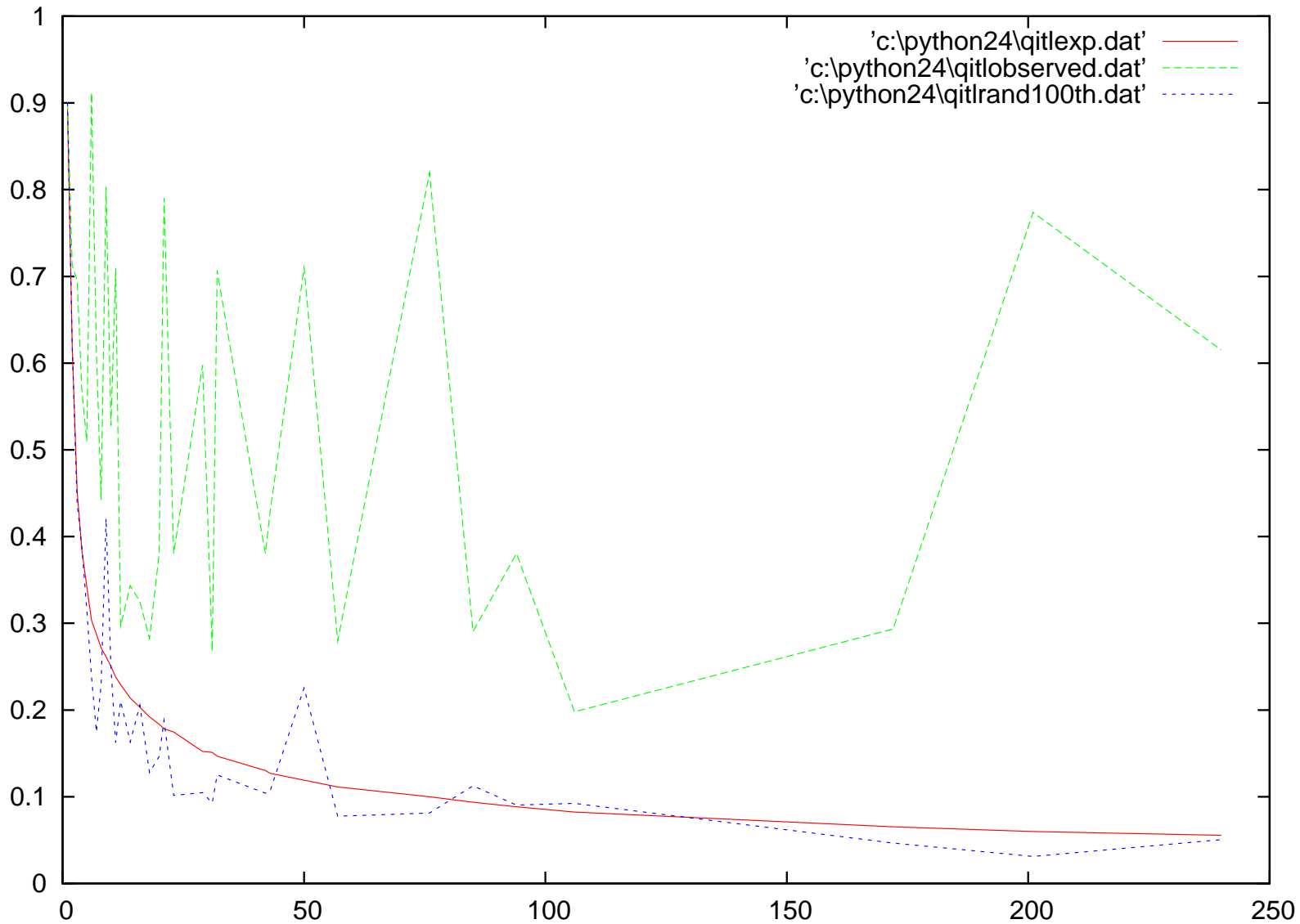
# Expectation of One Run

---



# Expectation/Example/Observed

---



# Distance from Expectation

For one run  $R(s)$ , define the distance from expectation  $E(s)$ :

- $D(R) = \sum_{s \in \text{family-sizes}} (R(s) - E(s))^2$
- E.g.,  $D(R_{100}) = 1.1810367326863338$
- Whereas for the observed real-world  $O$ ,  $D(O) = 9.7770054237386557$

Main result:  $D(O) > D(R)$  with  
 $p < .001$

# The Status of CPCH

- It is very unlikely that the real world is the outcome of the above CPCH simulations
- It stays unlikely also when we vary some parameters individually, e.g.,
  - if the transition time-unit is thrice the birth time-unit
  - if the birth-rate is slower
  - if the ursprache feature values are set in favour of the CPCH

But CPCH may still be saved if

- Real world historical data is very different from simulations
- Birth/death effects are much more dynamic than assumed here, e.g., with non-binary clonings?
- Some joint settings of birth-rates and transition matrix may (or may not) provoke the observed perturbation

# **My Own Belief**

---

- CPCH is not valid
- For one and the same feature, there are different transition probability matrices, depending on historical factors and (other) language features
- These matrices are normally distributed around centroid a matrix
- The centroid matrix determines the stationary distribution

**The End**

---

Thank You for Listening