Do languages originate and become extinct at constant rates?

The Automated Similarity Judgment Program

Current collaborators:
- Dik Bakker
- Oleg Belyaev
- Cecil H. Brown
- Pamela Brown
- Dmitry Egorov
- Anthony Grant
- Eric W. Holman
- Hagen Jung
- Robert Mailhammer
- André Müller
- Viveka Velupillai
- Søren Wichmann
- Kofi Yakpo

The ASJP project aims at achieving a computerized lexicostatistical analysis of ideally all the world's languages.
The two main purposes are to provide a classification of all languages by a single, consistent and objective (if perhaps not ideal) method and to perform various statistical analyses regarding the historical and areal behavior of lexical items

Simple birth and death process (Yule 1925, Kendall 1948)

1. Languages split to form new languages at a constant rate over time.

2. Languages become extinct without descendants at another constant rate over time.

3. These events are independent.

Parameter-free test: imbalance of phylogenetic trees

- Kiowa Tanoan (6)
  - Kiowa-Towa (2)
    - Kiowa (1)
      - Kiowa [kio] (USA)
    - Towa (1)
      - Jemez [tow] (USA)
  - Tewa-Tiwa (4)
    - Tewa (1)
      - Tewa [tew] (USA)
    - Tiwa (3)
      - Piro [pie] (USA)
      - Tiwa, Southern [tix] (USA)
      - Tiwa, Northern [twf] (USA)

Prediction (Farris 1976)

If two coordinate branches in a phylogenetic tree have a total of N languages between them, then each possible split of the languages between the branches is equally likely:
1 vs N-1, 2 vs N-2, and so on up to N-1 vs 1.

N=6: P[1-5] = P[2-4] = P[3-3] = P[4-2] = P[5-1]

This is true for any origination and extinction rates, as long as they are constant.

Imbalance of a binary node (Fusco and Cronk 1995)

I = (observed discrepancy) / (maximum possible)
= (number of languages on larger branch – number
for most even split) / (N – 1 – number for most even
split).

N = 6: 1-5 or 5-1, I = 1 N = 7: 1-6 or 6-1, I = 1

2-4 or 4-2, I = .5 2-5 or 5-2, I = .5

3-3, I = 0 3-4 or 4-3, I = 0

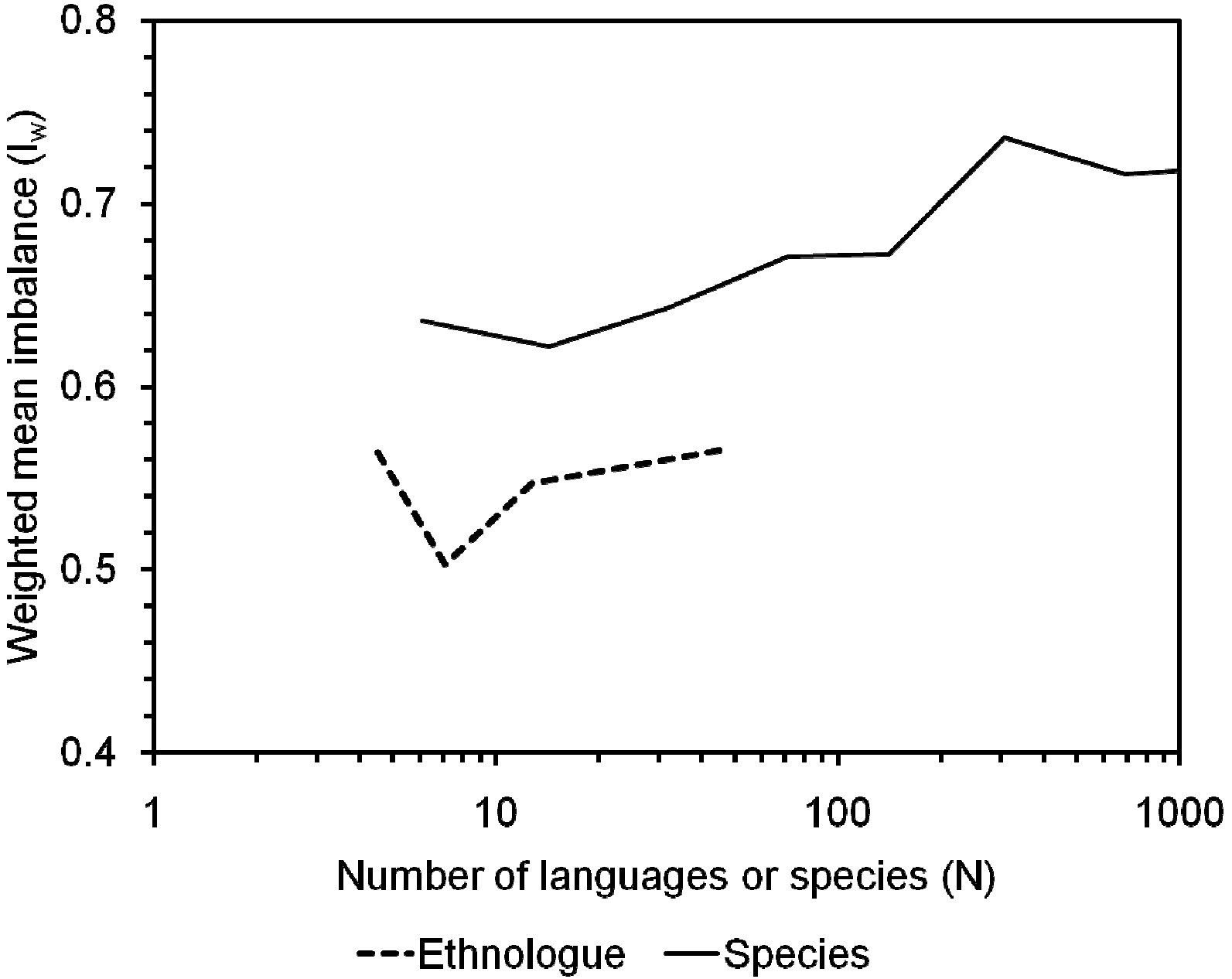
Weighted mean imbalance (Purvis et al. 2002)

If N is odd: w = 1.
If N is even and I > 0: w = (N-1)/N.
If N is even and I = 0: w = 2(N-1)/N.

N = 6: 1-5 or 5-1, I = 1 w = 5/6
2-4 or 4-2, I = .5 w = 5/6
3-3, I = 0 w = 10/6

I_w is the weighted mean of I with weights w.
This can be defined for any set of nodes, such as nodes with a particular value of N.

Prediction: I_w has expected value .5 for any N.
Test: calculate I_w as a function of N in published trees.


0.8
0.7
0.6
0.5
0.4
Weighted mean imbalance (I_w)
1
10
100
1000
Number of languages or species (N)
Ethnologue
Species

Birth and death model:
I_w = .5 for all N.

Languages in Ethnologue (Gordon 2005):
I_w = .544 (.502 - .585), little or no change with N.

Species in biological literature:
I_w at least .6, increasing with N.

Ethnologue trees are handmade.
Most nodes aren't included in test because they have more than two branches.

Species trees are now made by computers.
Most nodes have two branches.

ASJP for computerized language trees based on word lists

ASJPcode (Brown et al. 2008): standard orthography with 7 vowels, 34 consonants, and 4 modifiers

40-item list (Holman et al. 2008): 40 most stable items from Swadesh (1955) 100-item list, where stability is inferred from similarity of items in related languages relative to unrelated languages

Levenshtein distance between two languages based on word lists (Steps 1-3 from Serva and Petroni 2008)

1. For two words: LD = total number of insertions, deletions, and substitutions necessary to change one word into the other.
2. LDN = normalized LD = LD divided by length of longer word.
3. For two languages: Find average LDN between words on list for same meaning in the two languages.
4. Correct for random similarity: divide by average LDN between words for different meanings in the two languages to get LDND, which ranges from 0 to about 100%.

ASJP trees based on LDND matrix

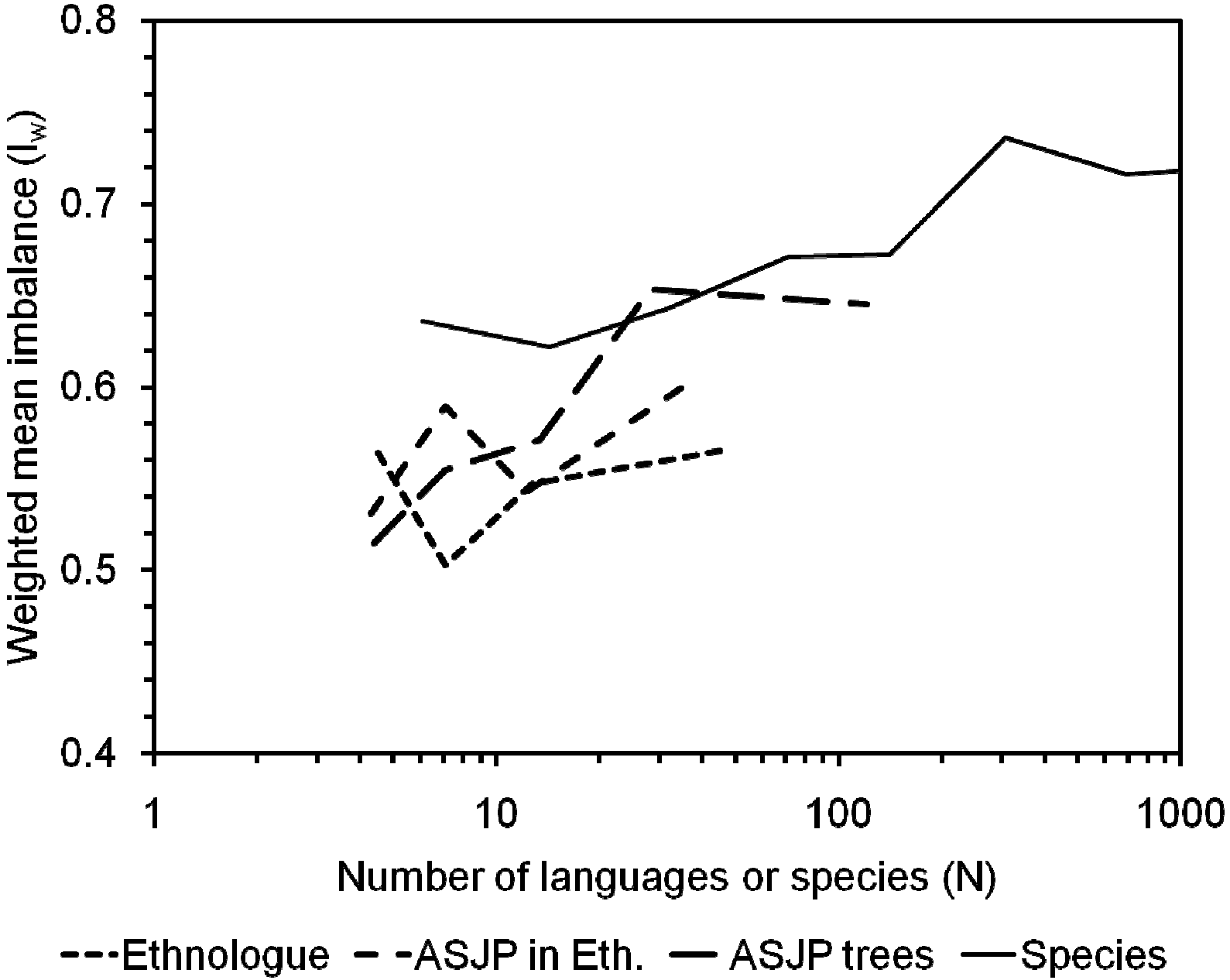
Separate tree for each family defined in WALS (Haspelmath et al. 2005).

Trees constructed by neighbor joining (Saitou and Nei 1987); all nodes have two branches.

ASJP is incomplete, with only about one-third as many languages as Ethnologue. Most large species trees are incomplete too. Does this matter to imbalance?

Theoretically: no, if birth and death model holds and sample is random.

Empirical test: imbalance of subset of Ethnologue that is also in ASJP.


0.8
0.7
0.6
0.5
0.4
Weighted mean imbalance (I_w)
1
10
100
1000
Number of languages or species (N)
Ethnologue
ASJP in Eth.
ASJP trees
Species

Birth and death model:
I_w = .5 for all N.

All Ethnologue: I_w = .544 (.502 - .585), little or no change with N.

ASJP subset of Ethnologue: I_w = .559 (.498 - .624), little or no change with N.

ASJP trees: I_w = .562 (.535 - .588), increasing with N.

Species: I_w at least .6, increasing with N.

Explanations for imbalance

1. Differences between branches in rates of origination or extinction.

2. Errors: adding random error increases imbalance of simulated trees.

3. Population size:
Larger populations could increase origination or decrease extinction.
Smaller populations could reflect oversplitting.

Proportion of nodes with larger populations on larger branch

All Ethnologue: .443 (.390 - .496)

ASJP subset of Ethnologue: .465 (.390 - .540)

ASJP trees: .458 (.426 - .490)

Species: mixed results in the literature

Test for effect of oversplitting on imbalance:
Define languages uniformly by LDND in ASJP

Set threshold value of LDND (for instance, 50%).
If average LDND between languages (or branches) is below threshold, count them as a single language.

For ASJP trees, this reduces average I_w to about .5, but I_w still increases with N.

For ASJP subset of Ethnologue, this has little effect on I_w.

Another prediction from birth and death model

If a single ancestral language has any living descendents, the expected number of descendents as a function of the origination rate λ, the extinction rate μ, and the time t:

$$E(N) = \frac{\lambda \exp[(\lambda - \mu)t] - \mu}{(\lambda - \mu)}$$

So $E(N)$ increases exponentially at rate λ for small t, and at rate $(\lambda - \mu)$ if $\lambda > \mu$ for large t.

- Kiowa Tanoan (6)
  - Kiowa-Towa (2)
    - Kiowa (1)
      - Kiowa [kio] (USA)
    - Towa (1)
      - Jemez [tow] (USA)
  - Tewa-Tiwa (4)
    - Tewa (1)
      - Tewa [tew] (USA)
    - Tiwa (3)
      - Piro [pie] (USA)
      - Tiwa, Southern [tix] (USA)
      - Tiwa, Northern [twf] (USA)

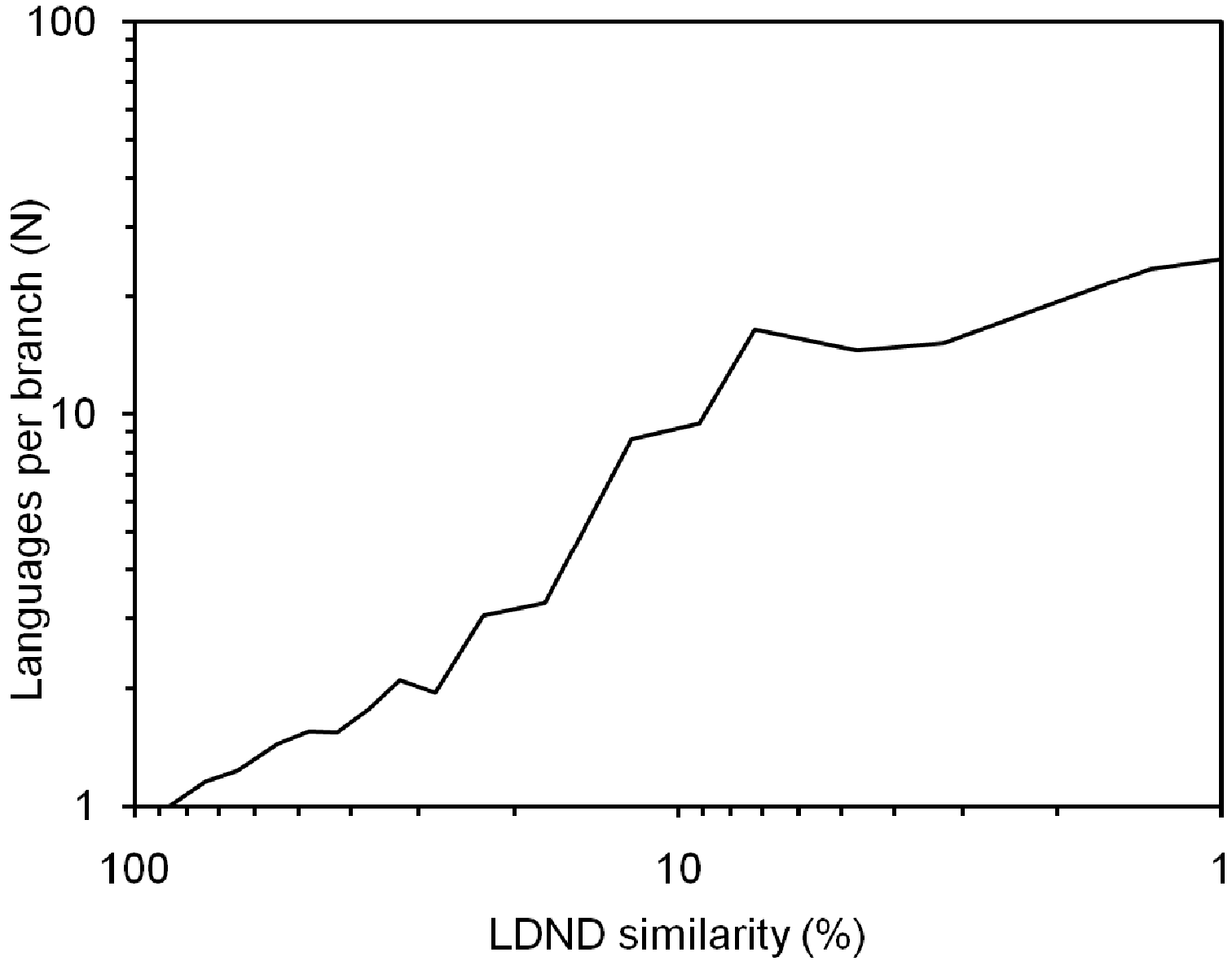
N is counted in Ethnologue on branches of trees:

Plotted on log scale, because if *N* increases exponentially, then log(*N*) increases linearly.

t is estimated from ASJP:

Origin of branch: LDND is averaged between languages on branch and languages on coordinate branches.

Plotted on reversed log scale, because by glottochronology, *t* is proportional to –log(1–LDND).


100
10
1
Languages per branch (N)
100
10
1
LDND similarity (%)

Increase in *N* is delayed and starts gradually.

Separate languages aren't recognized until some time after lineages split; this time is variable.

Similarity of different ASJP lists from same Ethnologue language: average = 65.2%, standard deviation = 16.8%.

Birth and death model can be generalized so that dialects are recognized as separate languages only after a fixed delay period.
Delay affects imbalance if and only if it's different on the two branches.

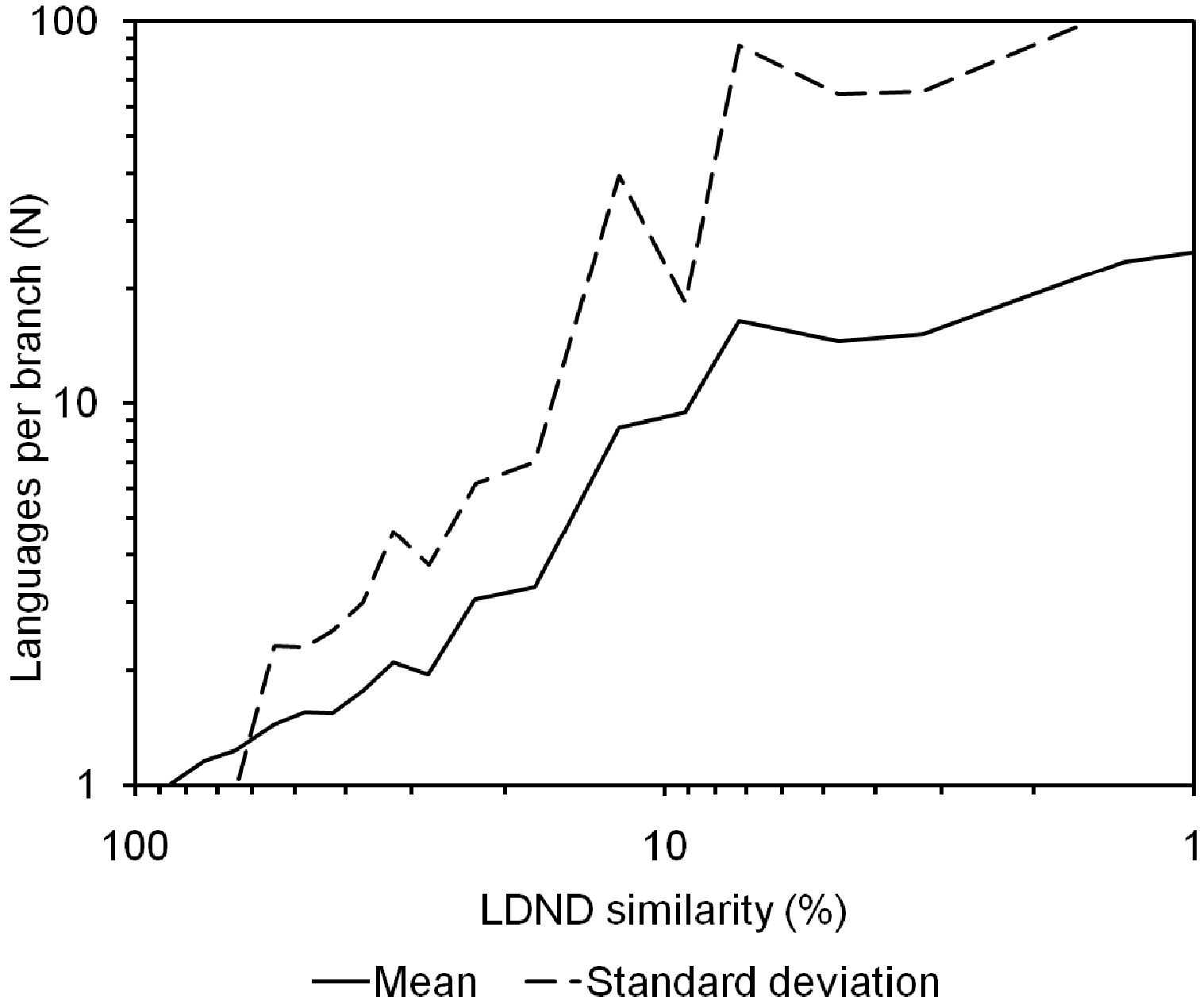
Slope of curve decreases substantially for LDND similarity below about 10%.

This implies that (λ - μ) is substantially lower than λ, so the extinction rate is almost as high as the origination rate.

This conclusion is based only on living languages, but it is consistent with the fact that the oldest recorded languages are all extinct without living descendants.

Another prediction from birth and death model

At any time *t*, the standard deviation of *N* is lower than the mean of *N* (where *N* is the number of languages per branch).


100
10
1
Languages per branch (N)
100
10
1
LDND similarity (%)
Mean
Standard deviation

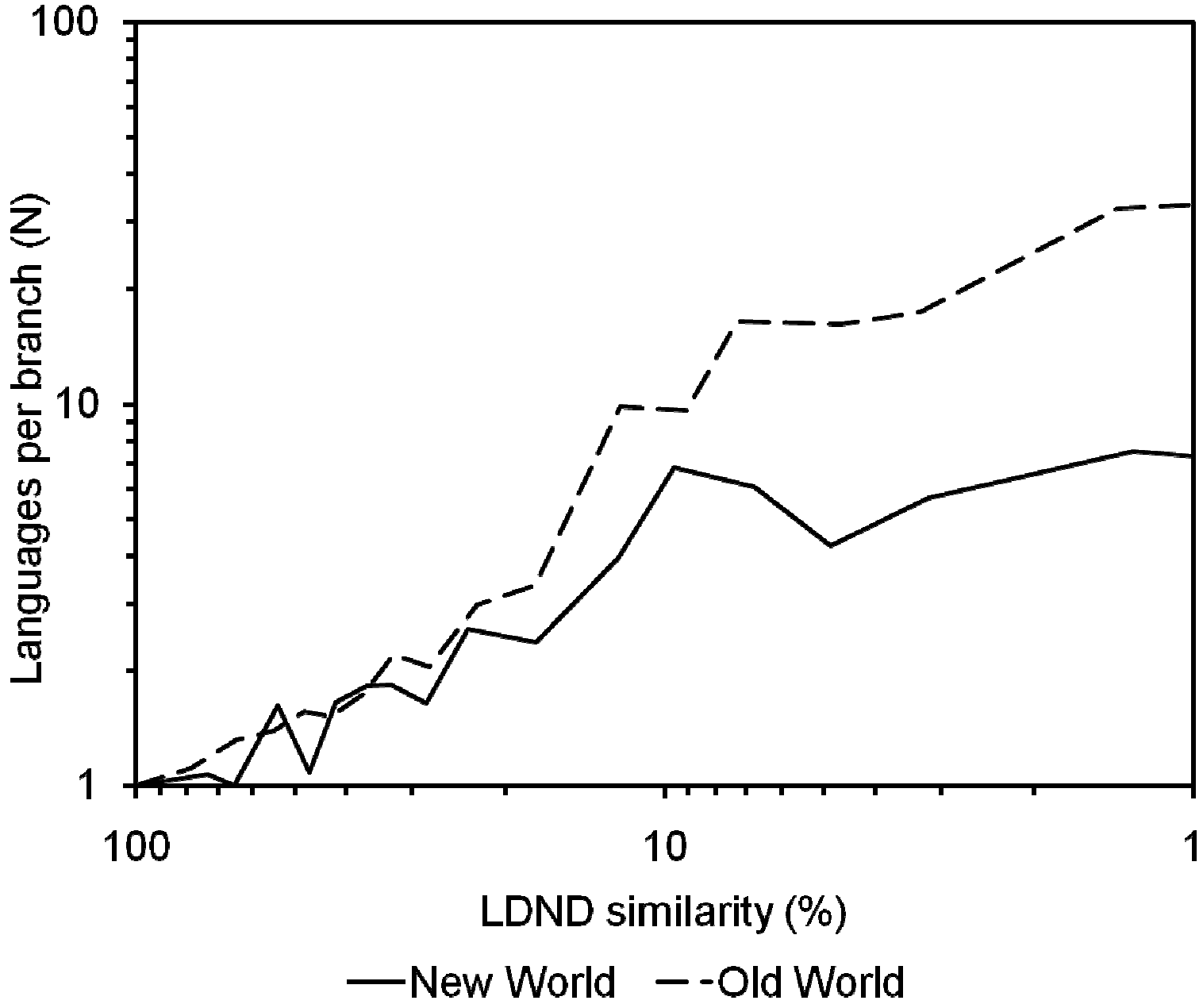
Possible reasons why *N* is too variable

1. Variability in time estimates, which inflates variability of *N* because *N* is a function of *t*.

2. Variability in boundary between languages and dialects, which inflates variability of number of languages counted.

3. Imbalance of trees, which inflates variability of *N* between branches.

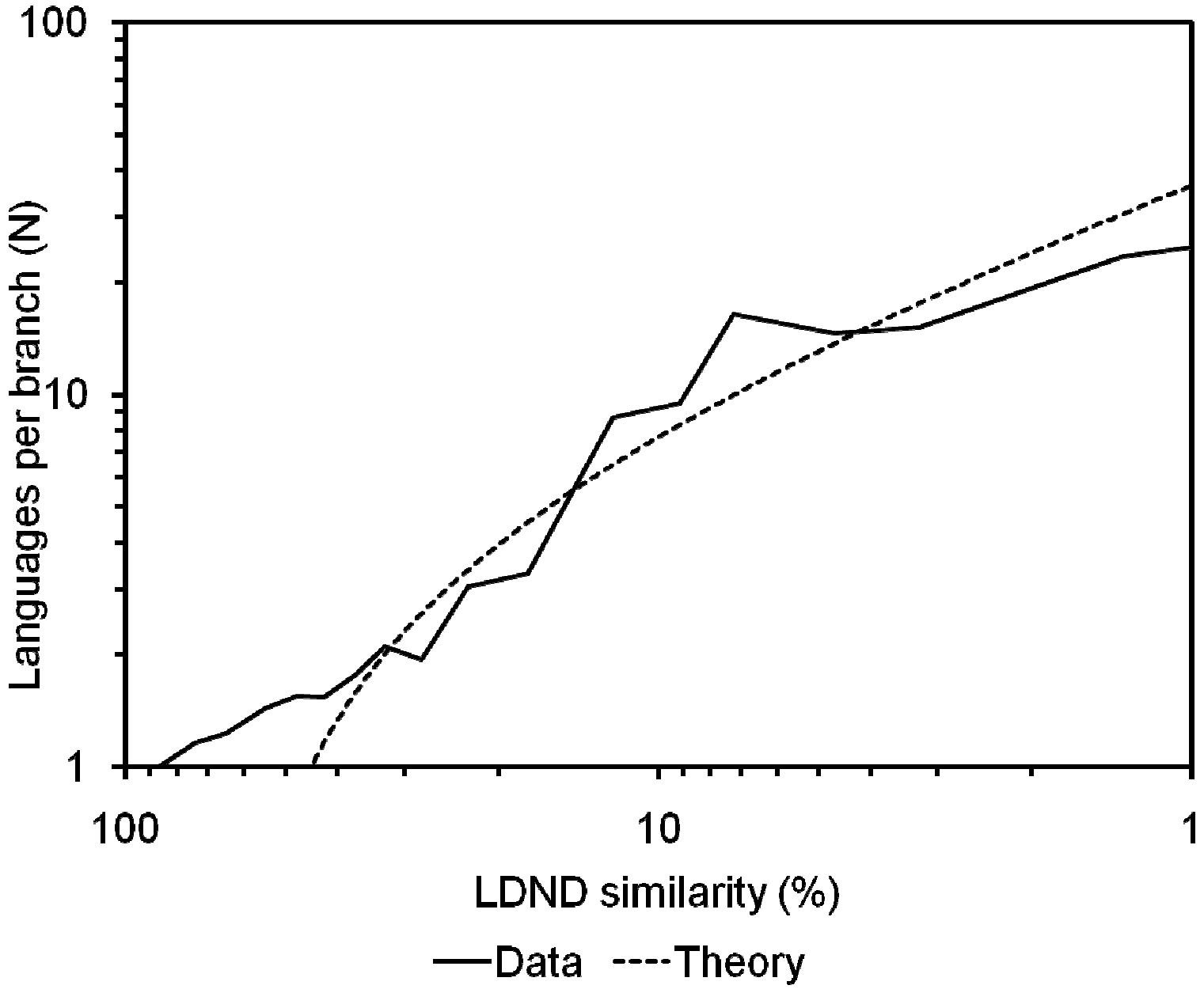
4. Differences in evolutionary rates between families or geographical regions, which inflate variability of *N* between trees.


100
10
1
Languages per branch (N)
100
10
1
LDND similarity (%)
New World
Old World

Main empirical problem with birth and death model: variability in parameters

Some variability is undoubtedly random.

Some seems to reflect patterns of historical events.


100
10
1
Languages per branch (N)
100
10
1
LDND similarity (%)
Data
Theory

Parameter values for theoretical curve

Baseline similarity within languages = 65%.
Retention rate = .79.
This makes I-E about 5500 years old.

Language-dialect boundary = 735 years.
Origination and extinction rates approach infinity with difference $\lambda - \mu$ held constant at .266 per millennium.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the World's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61: 285-308.

Farris, James S. 1976. Expected asymmetry of phylogenetic trees. *Systematic Zoology* 25:196-198.

Fusco, Giuseppe, and Quentin C. B. Cronk. 1995. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology* 175:235-243.

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue*. 15th Edition. SIL International. <www.ethnologue.com>.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42: 331-354.

Kendall, David G. 1948. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* 35:6-15.

Purvis, Andy, Aris Katzourakis, and Paul-Michael Agapow. 2002. Evaluating phylogenetic tree shape: Two modifications to Fusco and Cronk's method. *Journal of Theoretical Biology* 214:99-103.

Saitou, Naruya, and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstruction of phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Serva, M., and F. Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EuroPhysics Letters* 81: 68005.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21:121-137.

Yule, G. Udny. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London Series B* 213:21 - 87.