A Quantitative Analysis of Linguistic Metadata

Kavon Hooshiar
University of Hawai'i at Mānoa

ICLDC4, 2/27/15

This talk

Show some pretty graphs

Explain lots of horrible assumptions

See if you think this is meaningful by the end if it

Inspiration

How does language size relate to endangerment?

How does endangerment relate to documentation?

Size and Endangerment

“Lots of small endangered languages”

“They don’t correlate because of outliers”

Endangerment and Documentation

“There are so many languages going extinct that aren’t documented”

“Endangerment alone can’t be a reason to document a language”

An example

Paul Newman
SOAS Department Seminar
Oct. 15, 2013
The Law of Unintended Consequences: How the Endangered Languages Movement Undermines Field Linguistics as a Scientific Enterprise.

Check the data

Each topic implies knowledge of correlations

- Number of Speakers vs Endangerment
- Endangerment vs Documentation

Modern linguistic resources claim to give us access to this data, so lets actually check...

The data

Goal: a set of metadata for every language

- -Number of speakers
- -Endangerment level
- -Metric of completed linguistic research

Completed Linguistic Research?

ELCat Documentation Index

- Raw scores given for Boasian Trilogy
  - (Grammars, Dictionaries/Lexicon, Texts/Corpora)

My goal is to follow this model

- using previously compiled data

Data -> Information

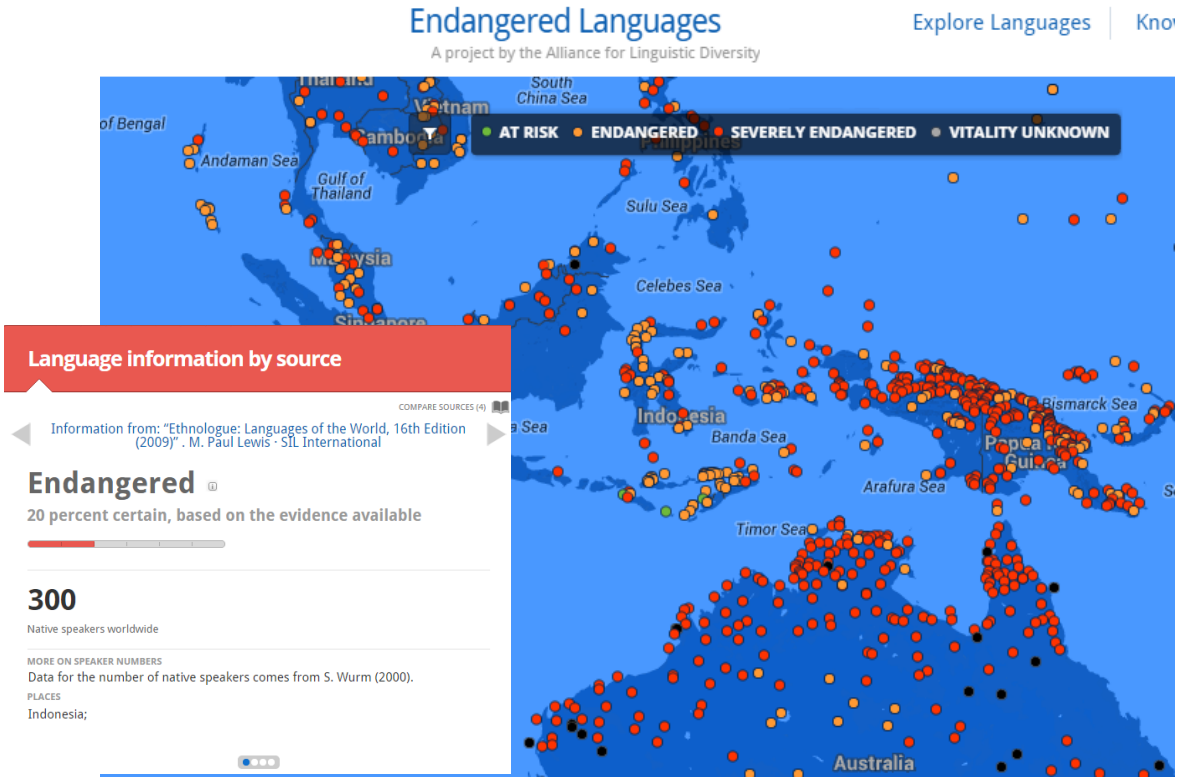
Linguists collect metadata on individual languages during their work

Other linguistics compile all of this metadata

Still other linguists (me!) process the data

Assumption: My analysis can only be as good as the work of those who came before me

Potential data sources: ELCat

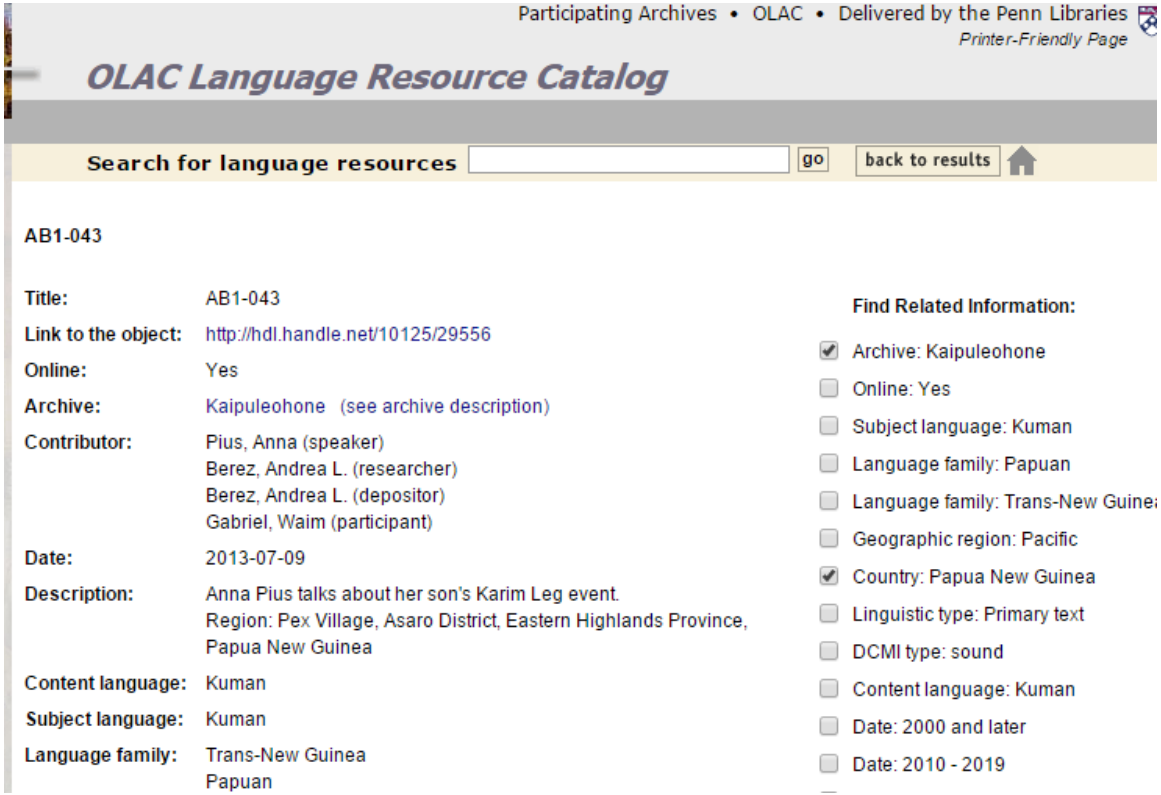


endangeredlanguages.com

Pro: Info quality

Con: endangered languages only

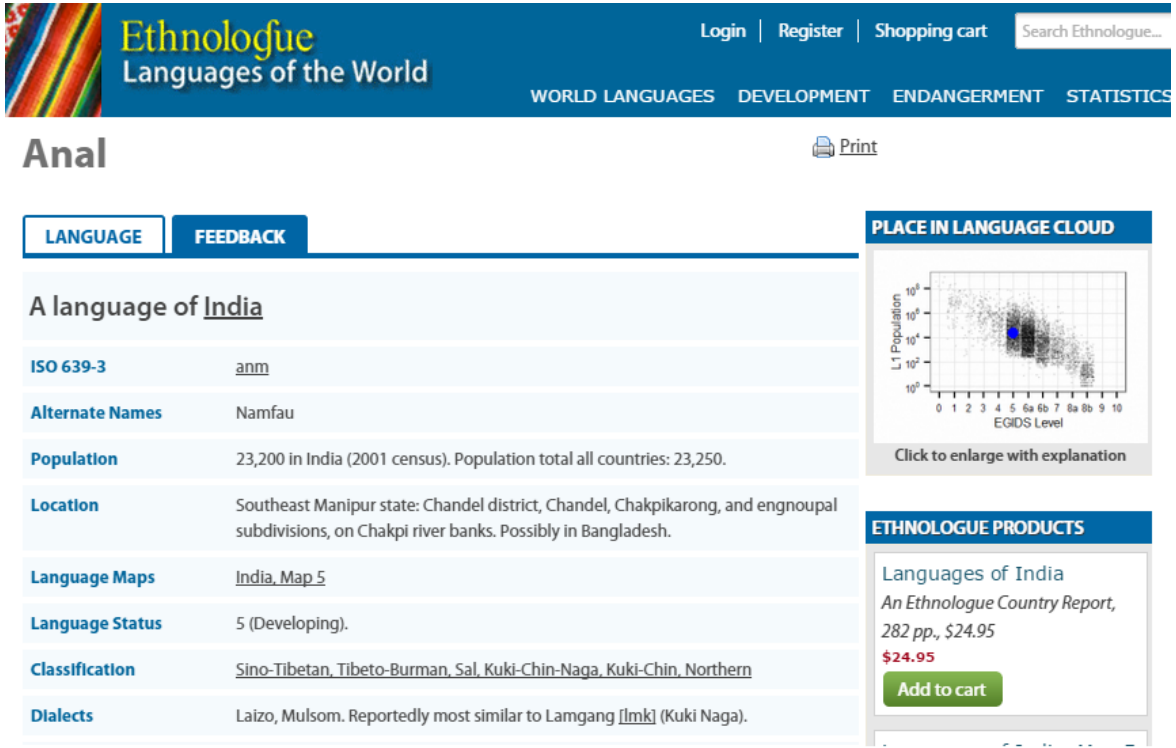
Potential data sources: OLAC



Pro: Archive info

Con: Unorganized, unmaintained

Potential data sources: Ethnologue

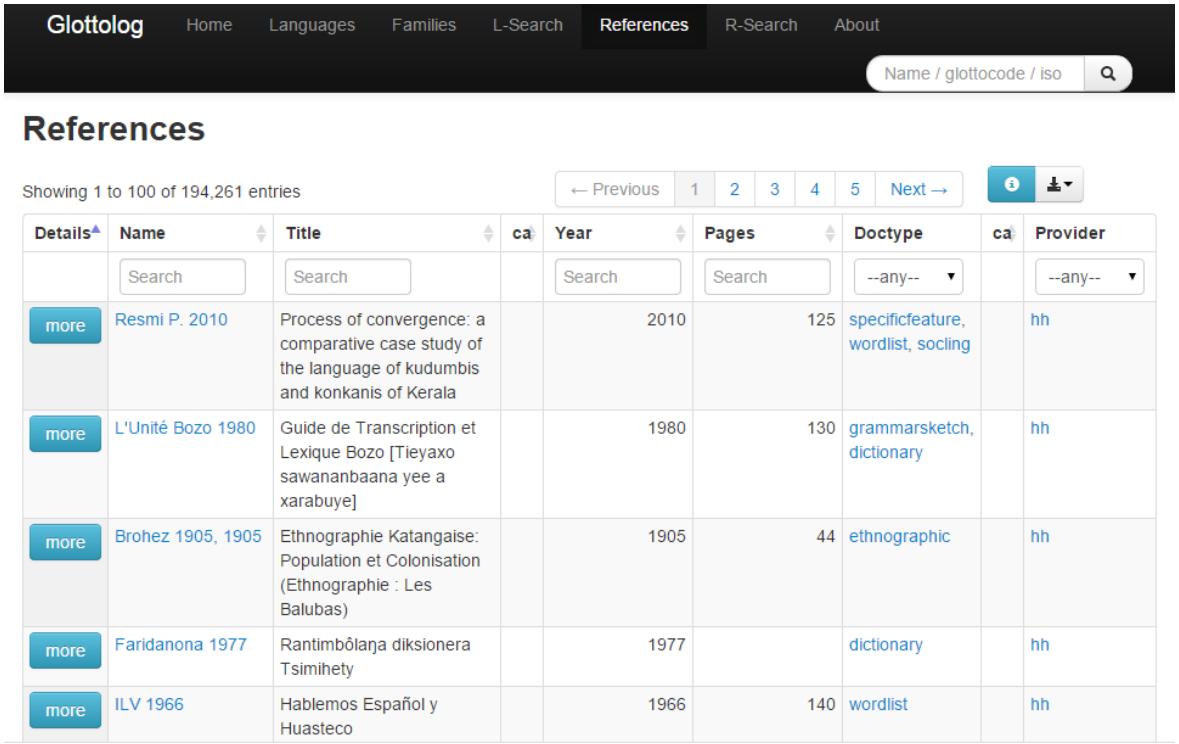


Pro: All languages

Con:

Resource data

Potential data sources: Glottolog



Glottolog Home Languages Families L-Search References R-Search About

Name / glottocode / iso

References

Showing 1 to 100 of 194,261 entries

← Previous 1 2 3 4 5 Next →

Details	Name	Title	ca	Year	Pages	Doctype	ca	Provider
	Search	Search		Search	Search	--any--		--any--
more	Resmi P. 2010	Process of convergence: a comparative case study of the language of kudumbis and konkanis of Kerala		2010	125	specificfeature, wordlist, socling		hh
more	L'Unité Bozo 1980	Guide de Transcription et Lexique Bozo [Tieyaxo sawananbaana yee a xarabuye]		1980	130	grammarsketch, dictionary		hh
more	Brohez 1905, 1905	Ethnographie Katangaise: Population et Colonisation (Ethnographie : Les Balubas)		1905	44	ethnographic		hh
more	Faridanona 1977	Rantimbôlaŋa diksionera Tsimihety		1977		dictionary		hh
more	ILV 1966	Hablemos Español y Huasteco		1966	140	wordlist		hh

Pro: All languages

Resource data

“One man job”

Con: Archive data

“One man job”

Hammarström, Harald. "Automatic annotation of bibliographical references with target language." *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*. Association for Computational Linguistics, 2008.

Gimi

Print

LANGUAGE | FEEDBACK

A language of Papua New Guinea

ISO 639-3	gim
Alternate Names	Labogai
Population	22,500 (Wurm and Hattori 1981).
Location	Eastern Highlands Province, Okapa district; small area in adjacent Chimbu Province.
Language Maps	Papua New Guinea, Map 9 Papua New Guinea, Map 10
Language Status	5 (Developing).
Classification	Trans-New Guinea, Kainantu-Goroka, Gorokan, Fore
Dialects	East Gimi, West Gimi (Gouno).
Language Development	Literacy rate in L1: 5%–15%. Literacy rate in L2: 5%–15%. NT: 1994.

References

Showing 1 to 100 of 194,261 entries

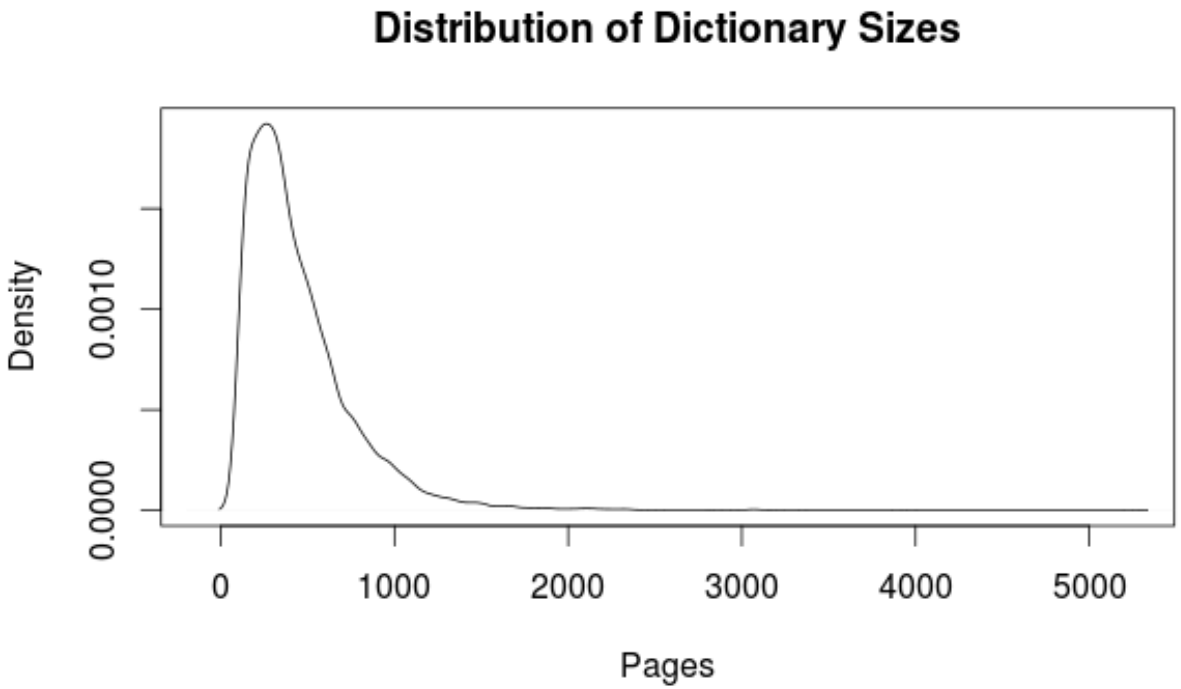
Details	Name	Title	ca	Year	Pages	Doctype	ca	Provider
	Search	Search		Search	Search	--any--		--any--
more	Resmi P. 2010	Process of convergence: a comparative case study of the language of kudumbis and konkanis of Kerala		2010	125	specificfeature, wordlist, socling		hh
more	L'Unité Bozo 1980	Guide de Transcription et Lexique Bozo [Tieyaxo sawananbaana yee a xarabuye]		1980	130	grammarsketch, dictionary		hh
more	Brohez 1905, 1905	Ethnographie Katangaise: Population et Colonisation (Ethnographie : Les Balubas)		1905	44	ethnographic		hh
more	Faridanona 1977	Rantimbôlaŋa diksionera Tsimihety		1977		dictionary		hh
more	ILV 1966	Hablemos Español y Huasteco		1966	140	wordlist		hh

My “completed research” metric

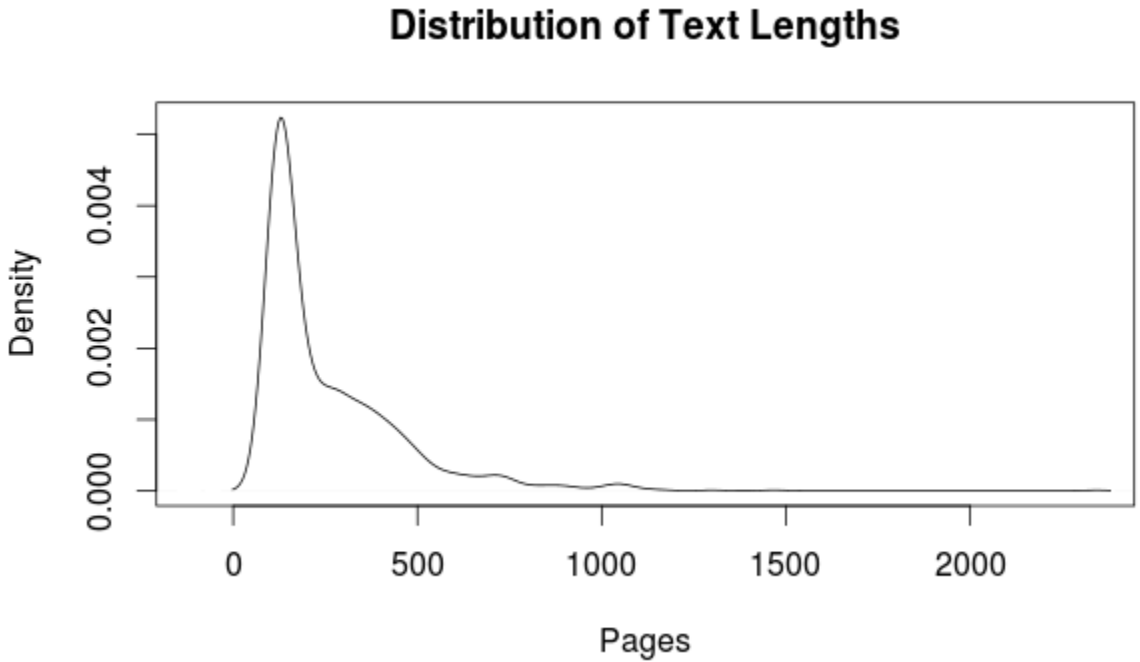
A percentage of completion for each language: 0-100%

Grammar 0-33%	Dictionary 0-33%	Texts 0-33%
Each grammar rated for page numbers and age Diminishing returns on multiple grammars	Total number of pages 250pp is 33%	Total number of pages 100pp is 33%

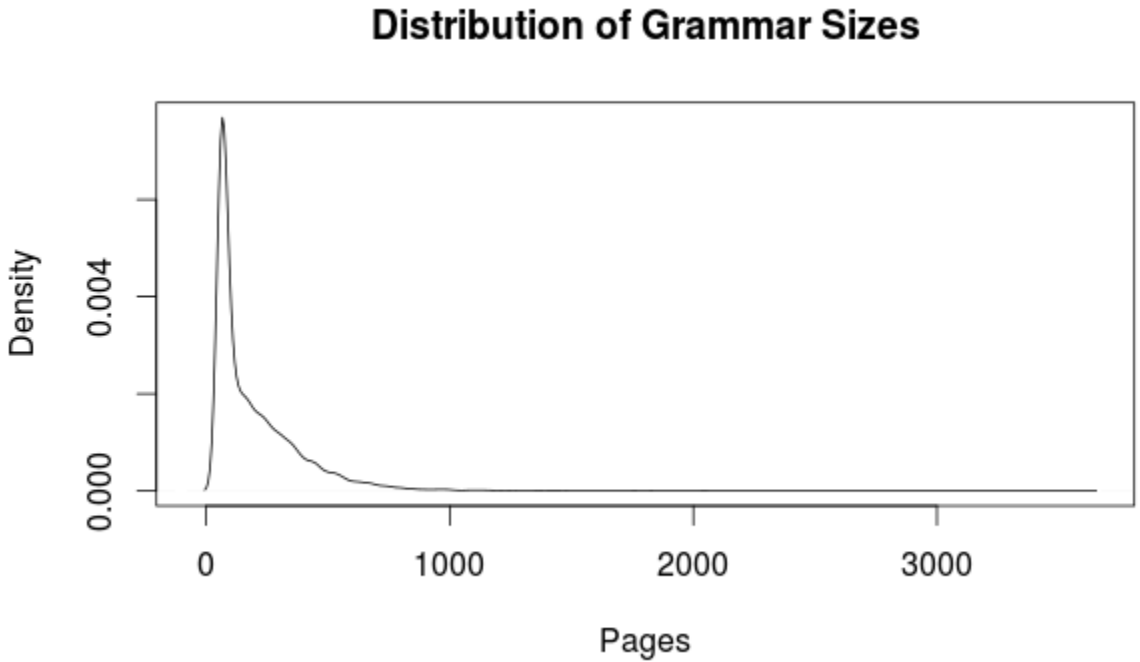
Dictionary sizes in the data



Text lengths in the data



Grammar sizes in the data



Issues

Reliance on pages numbers

Texts vs archived resources

The multiple label problem

(affecting up to 650 languages)

Other assumptions

Speaker numbers: total of L1 speakers

Only using data that has an associated ISO code

Some data created through algorithms

Data pulled in February, doesn’t include items in press

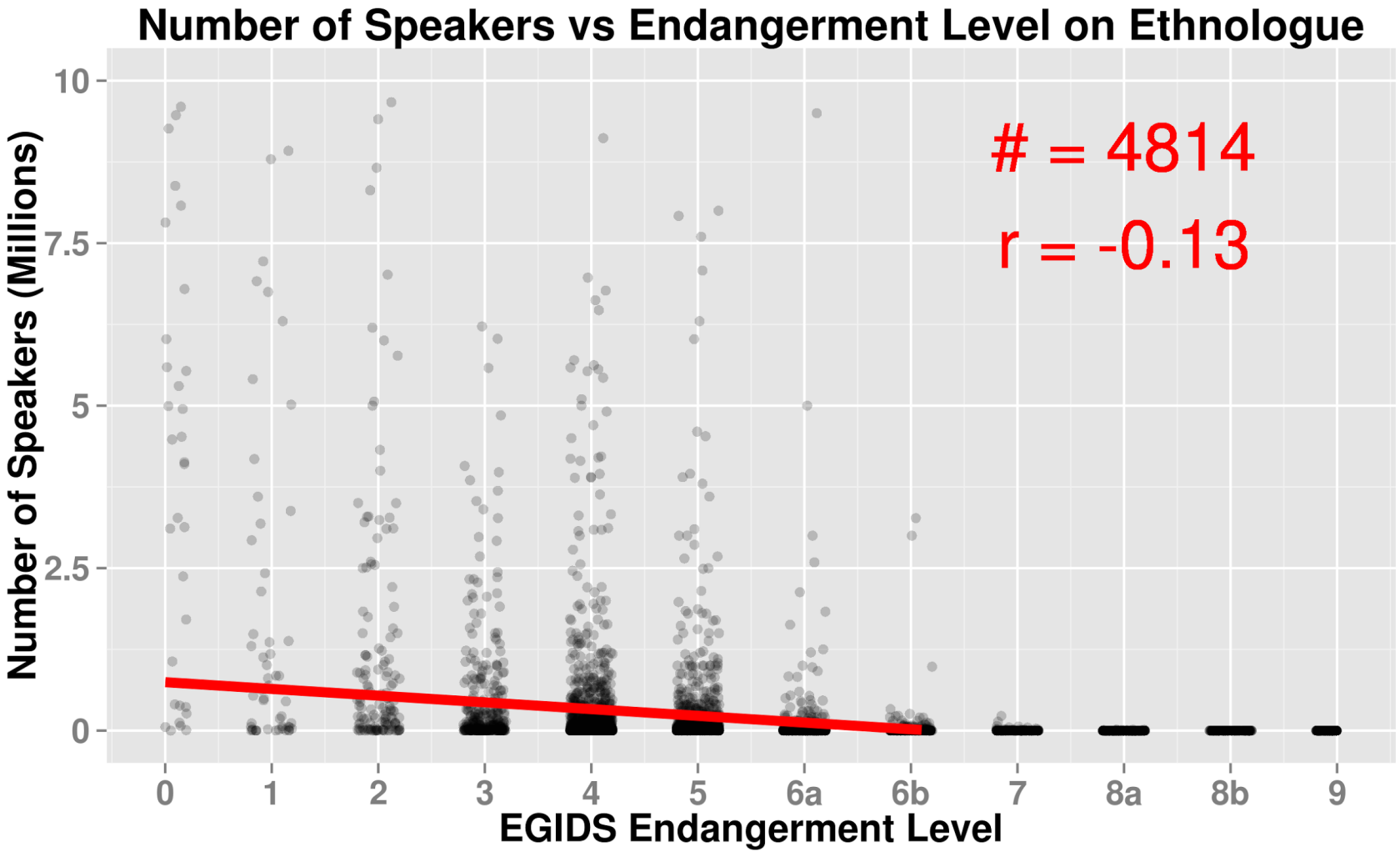
Repetition:

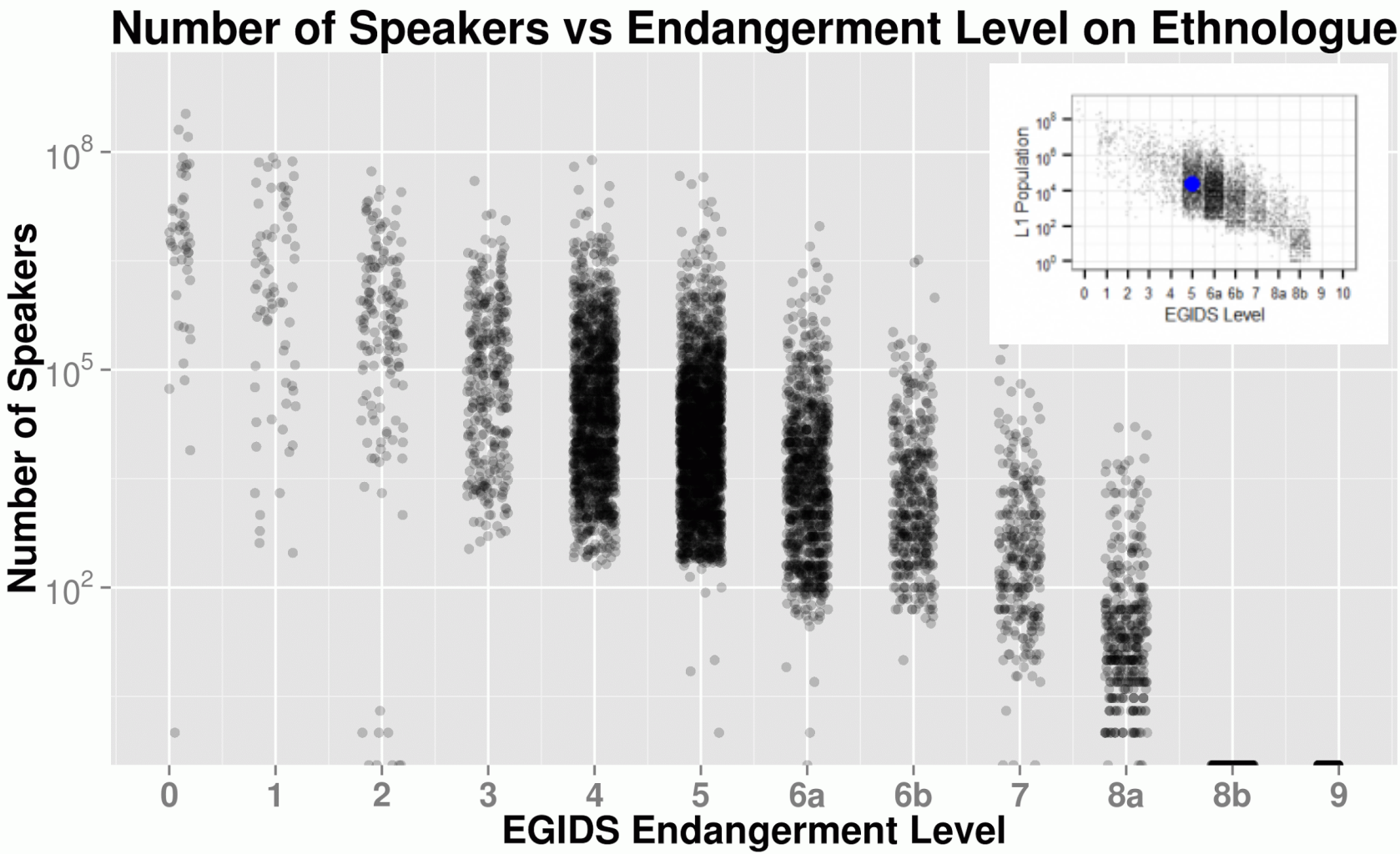
Information is only as good as the data it’s based on

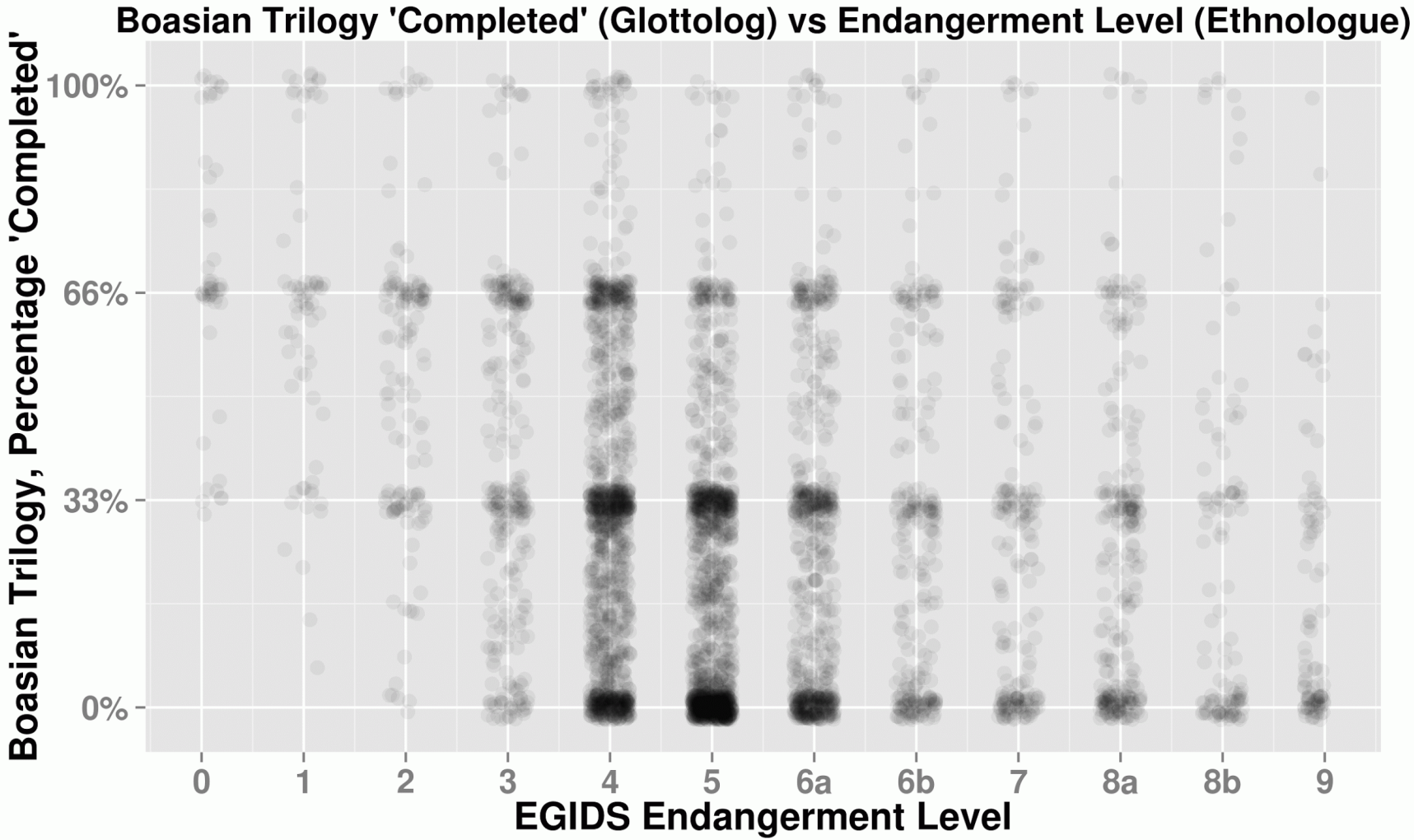
One final thought...

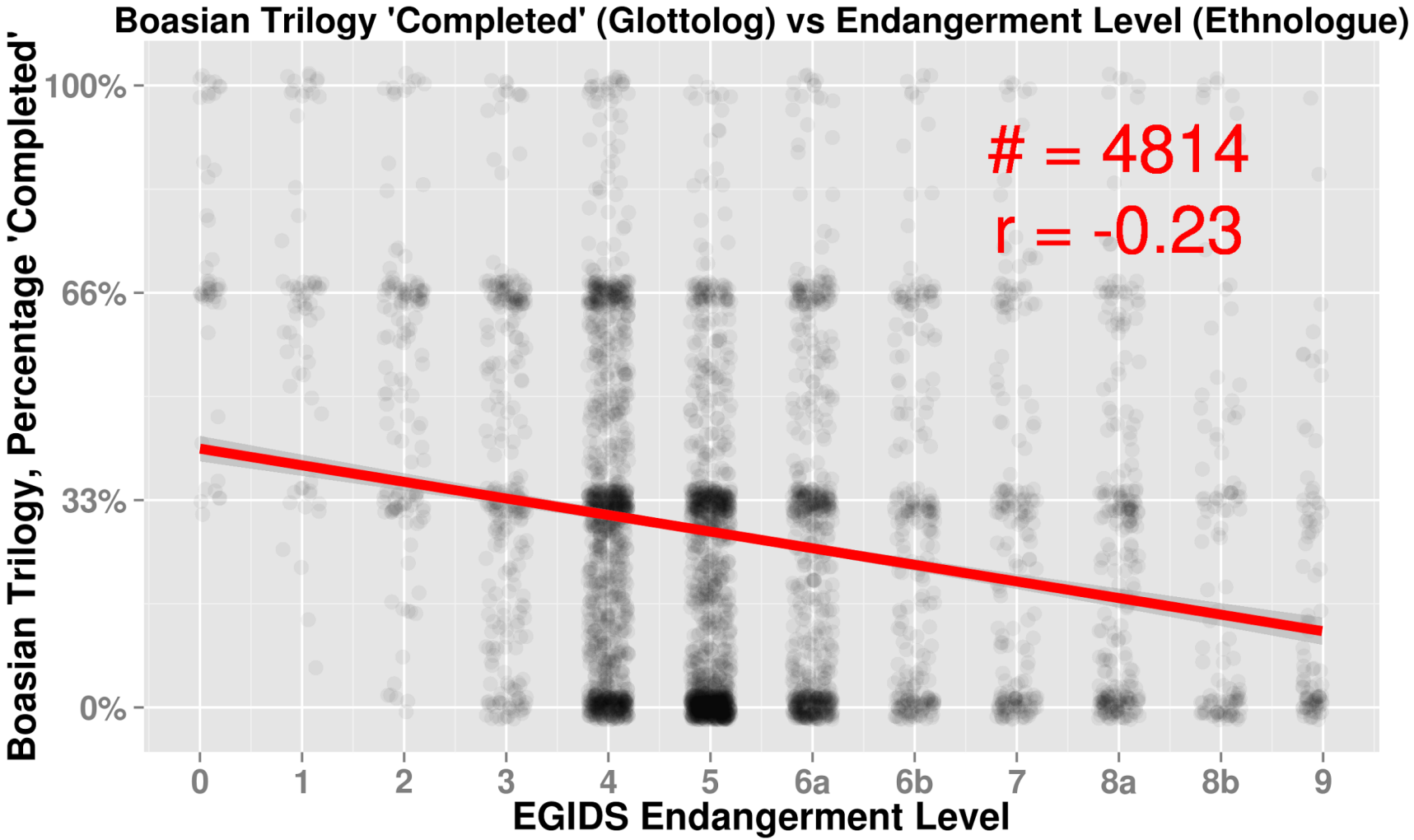
General strategy: sacrifice *some* amount of accuracy to make the analysis feasible

General goal: look for correlations and patterns among large sample sizes so that variance and outliers are acceptable


Number of Speakers vs Endangerment Level on Ethnologue
# = 4814
r = -0.13
Number of Speakers (Millions)
10
7.5
5
2.5
0
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

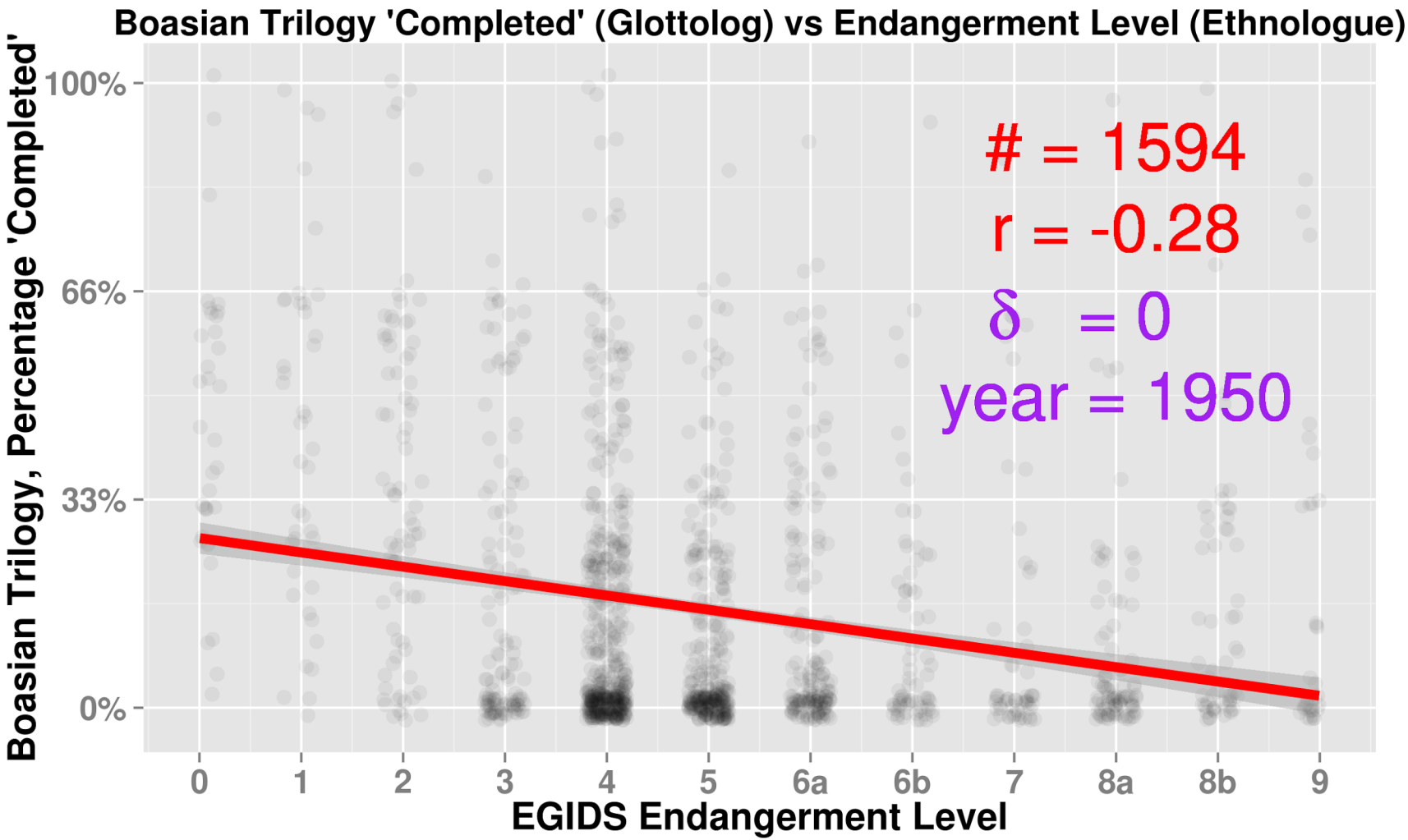

Number of Speakers vs Endangerment Level on Ethnologue
Number of Speakers
10^8
10^5
10^2
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level
L1 Population
10^8
10^6
10^4
10^2
10^0
0 1 2 3 4 5 6a 6b 7 8a 8b 9 10
EGIDS Level

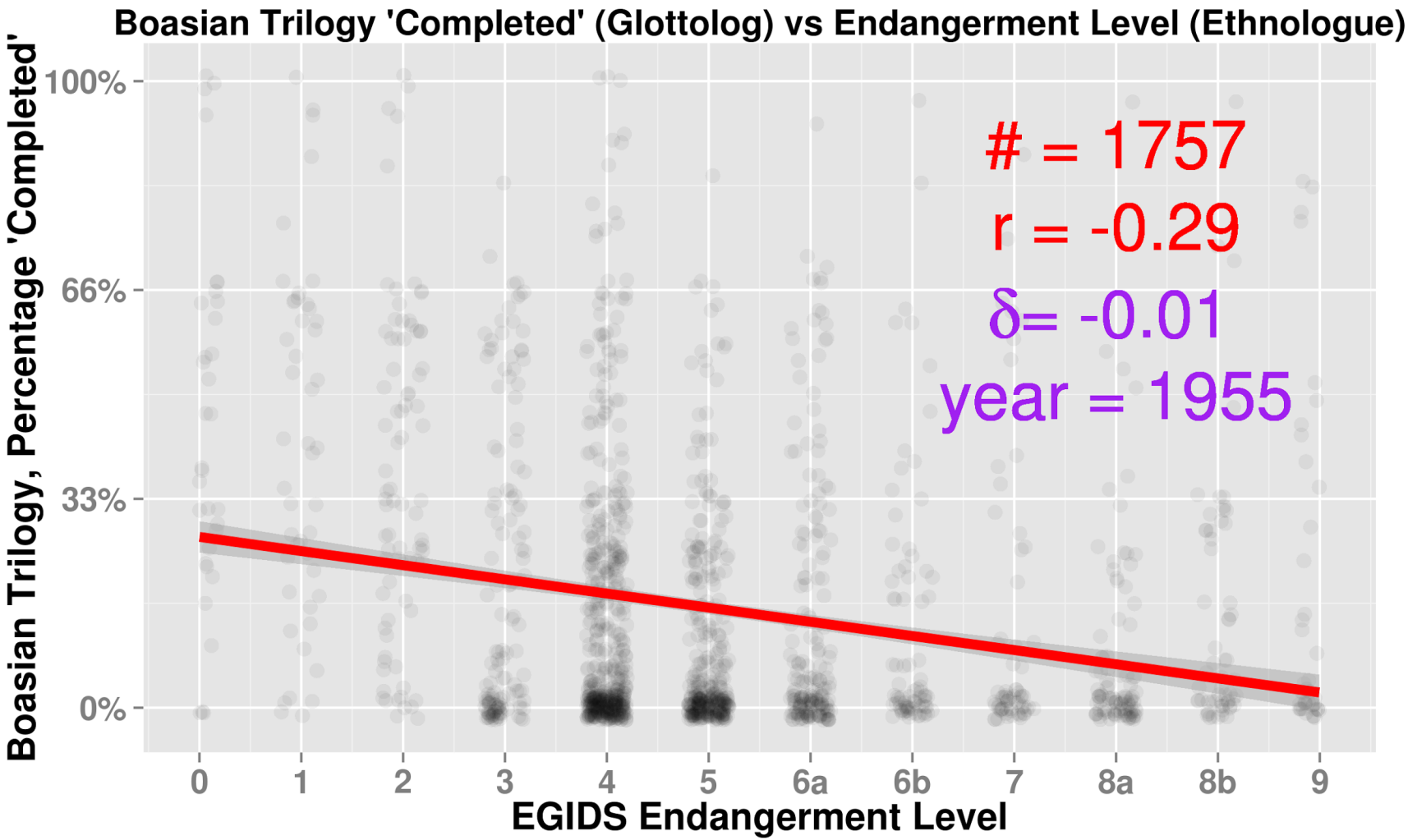

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

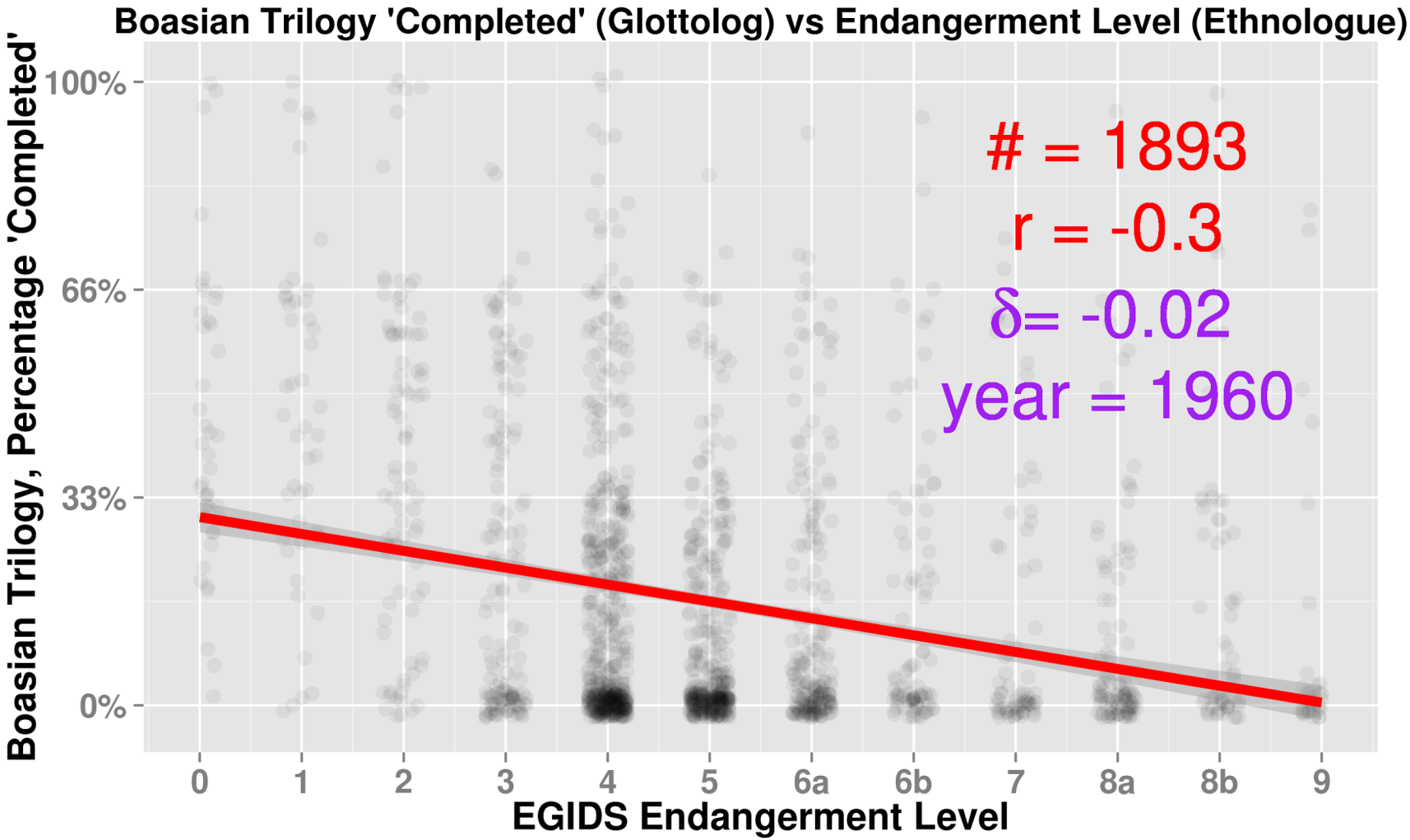

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 4814
r = -0.23
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

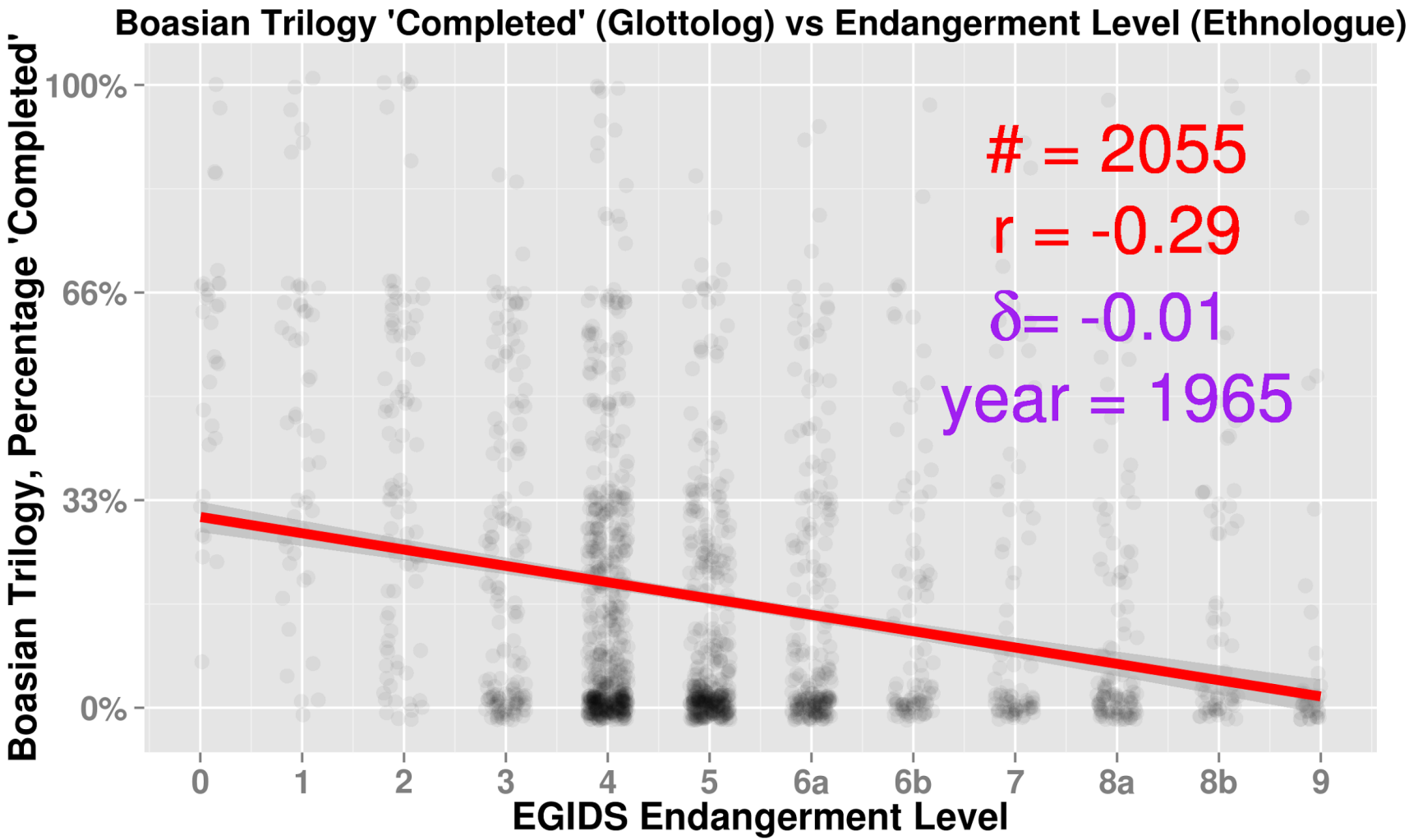
What next?

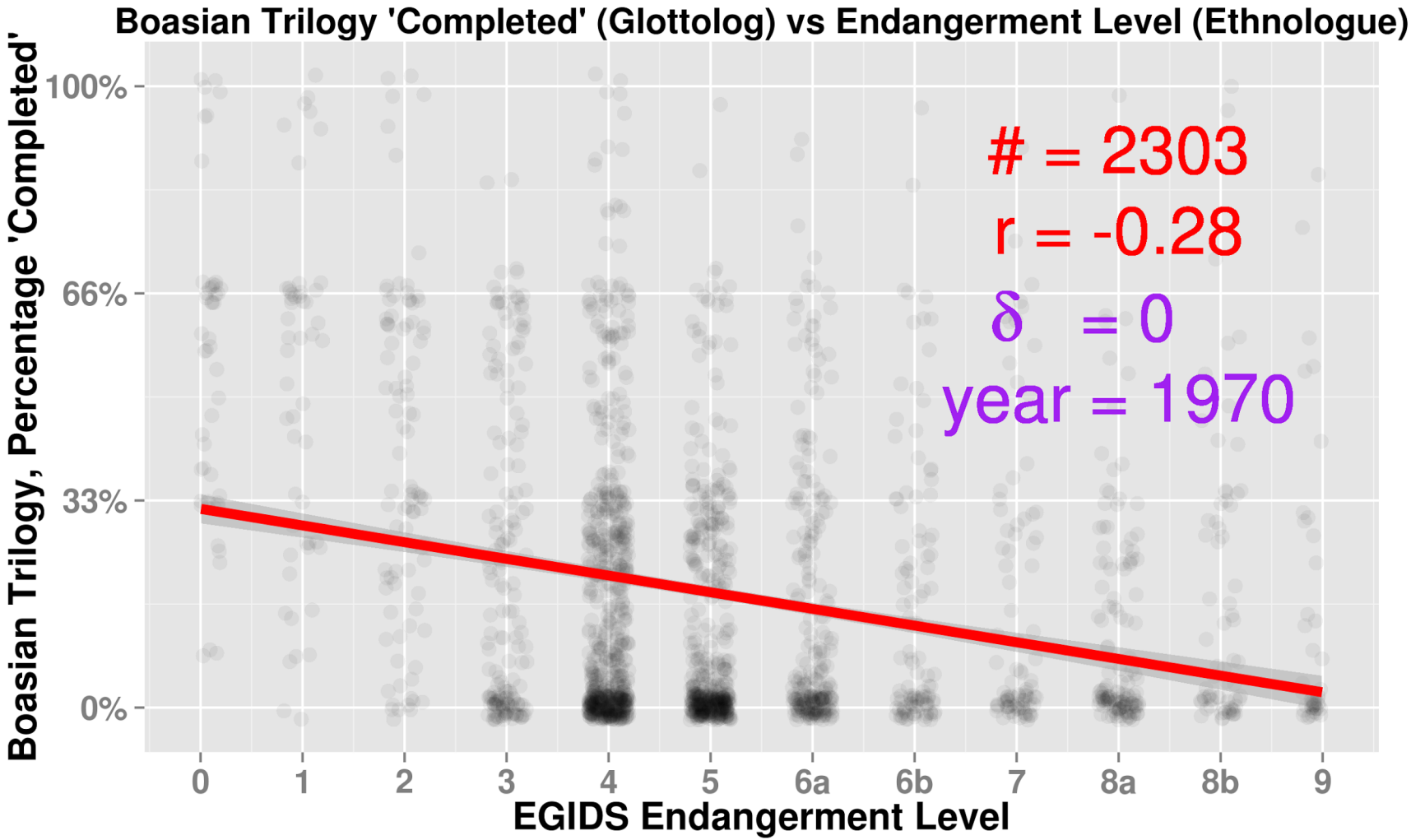
time travel, of course

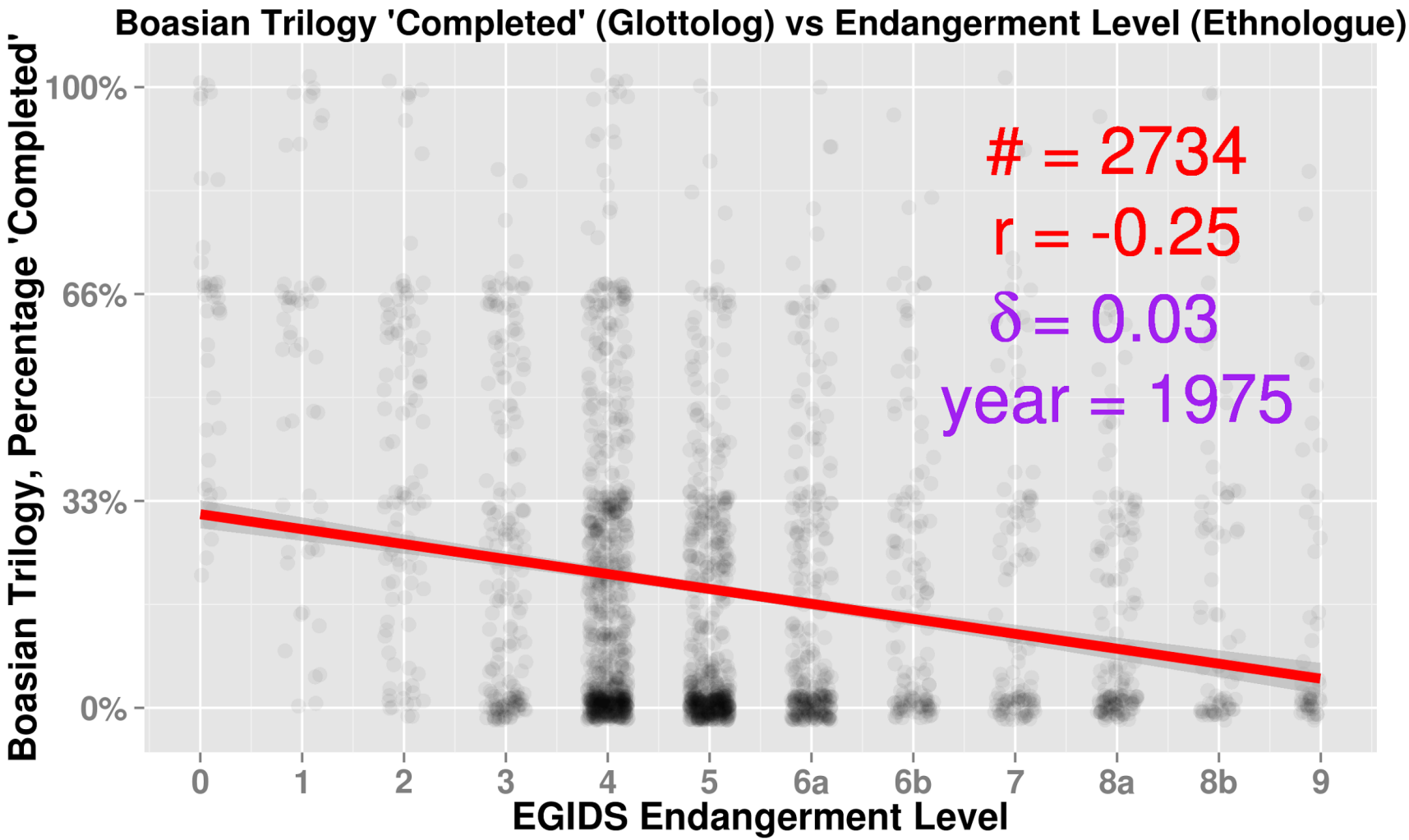

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 1594
r = -0.28
δ = 0
year = 1950
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

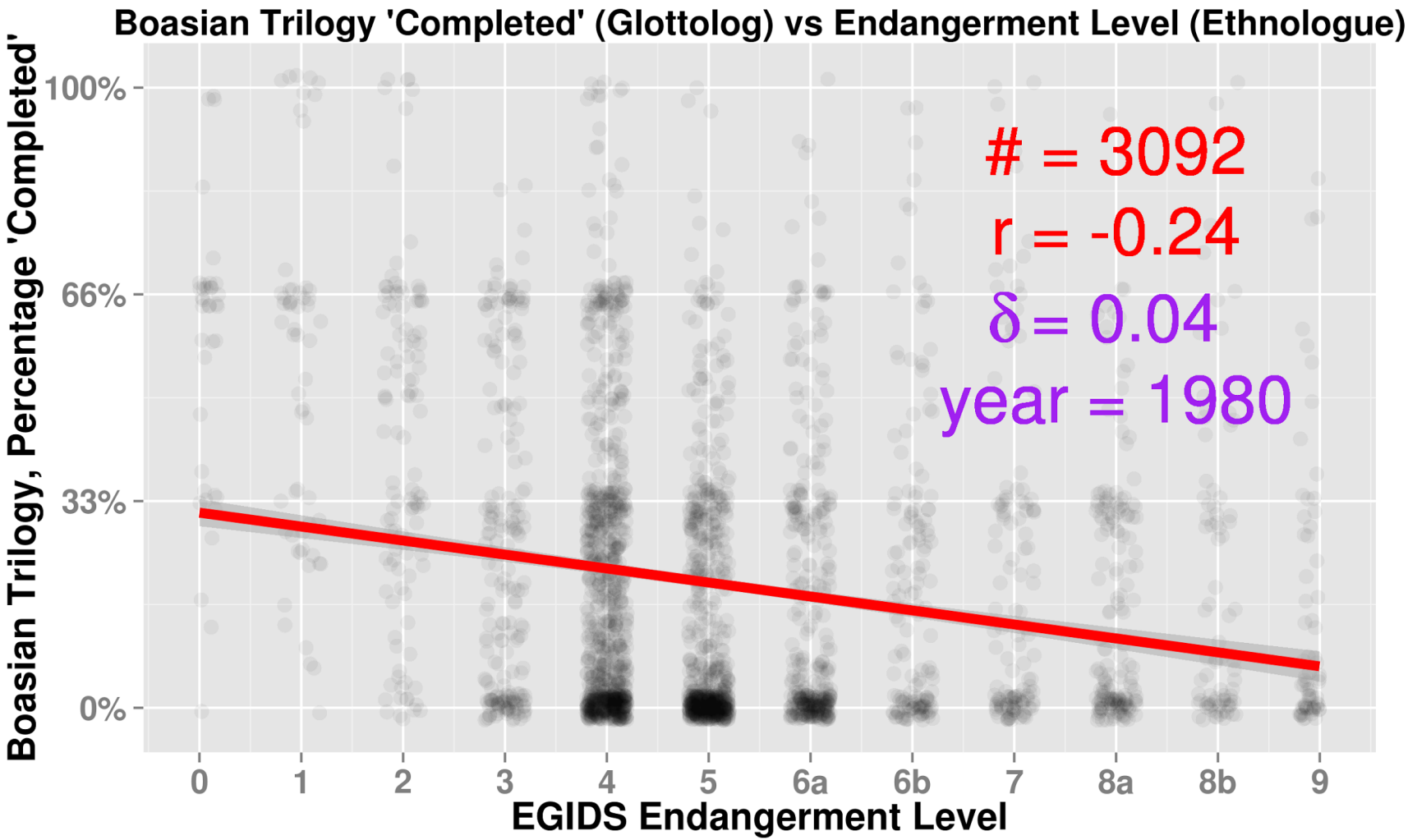

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 1757
r = -0.29
δ= -0.01
year = 1955
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

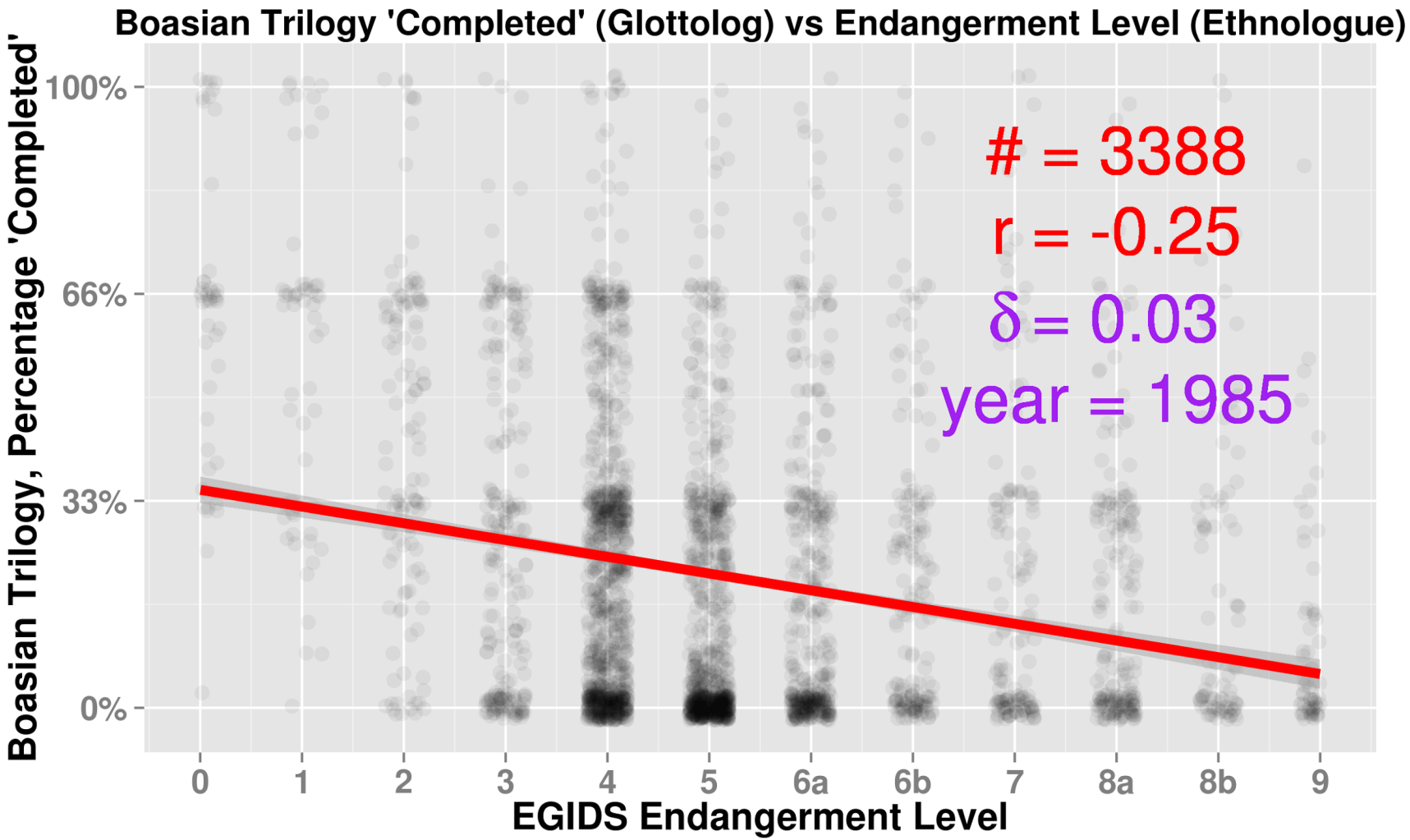

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 1893
r = -0.3
δ= -0.02
year = 1960
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

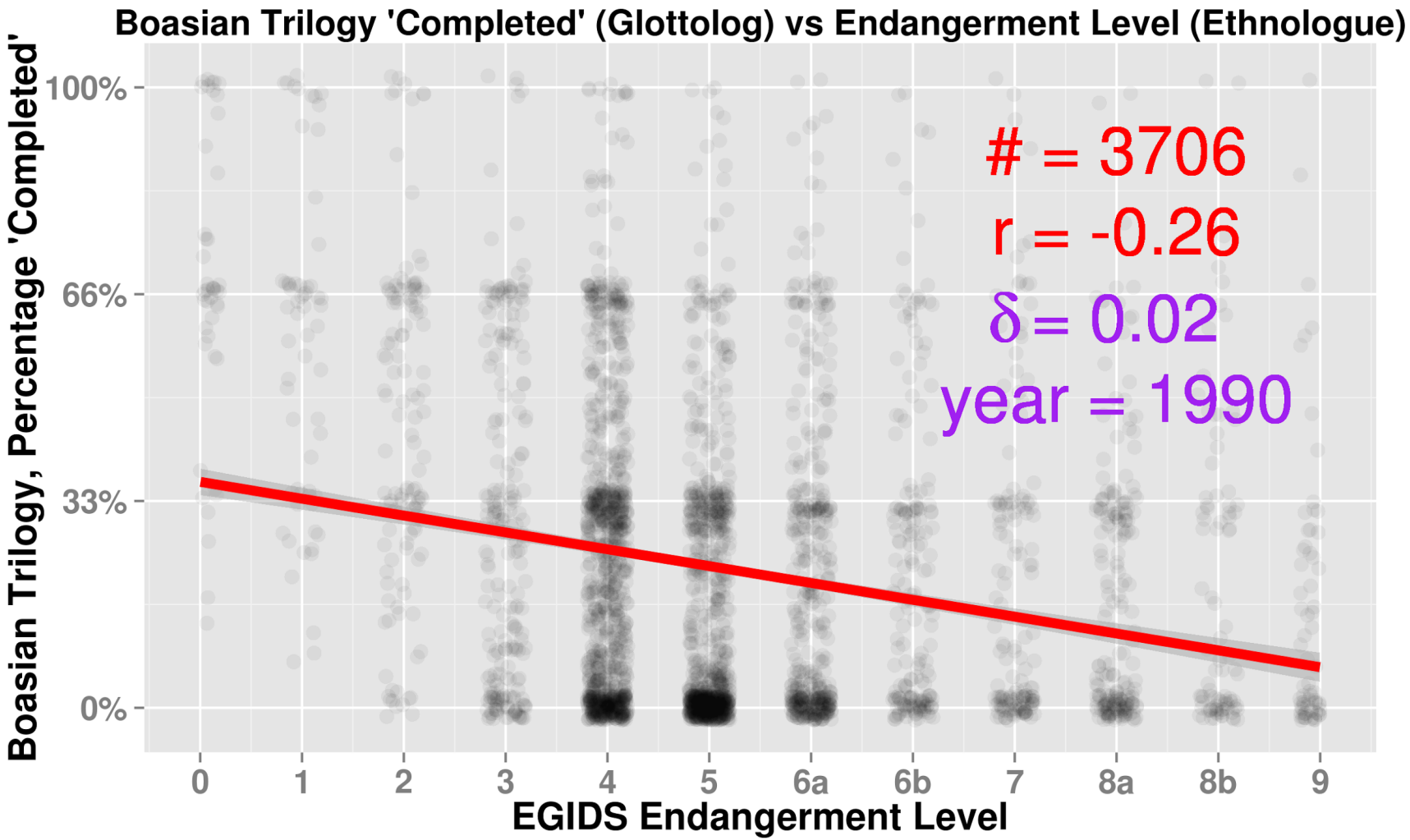

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 2055
r = -0.29
δ= -0.01
year = 1965
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

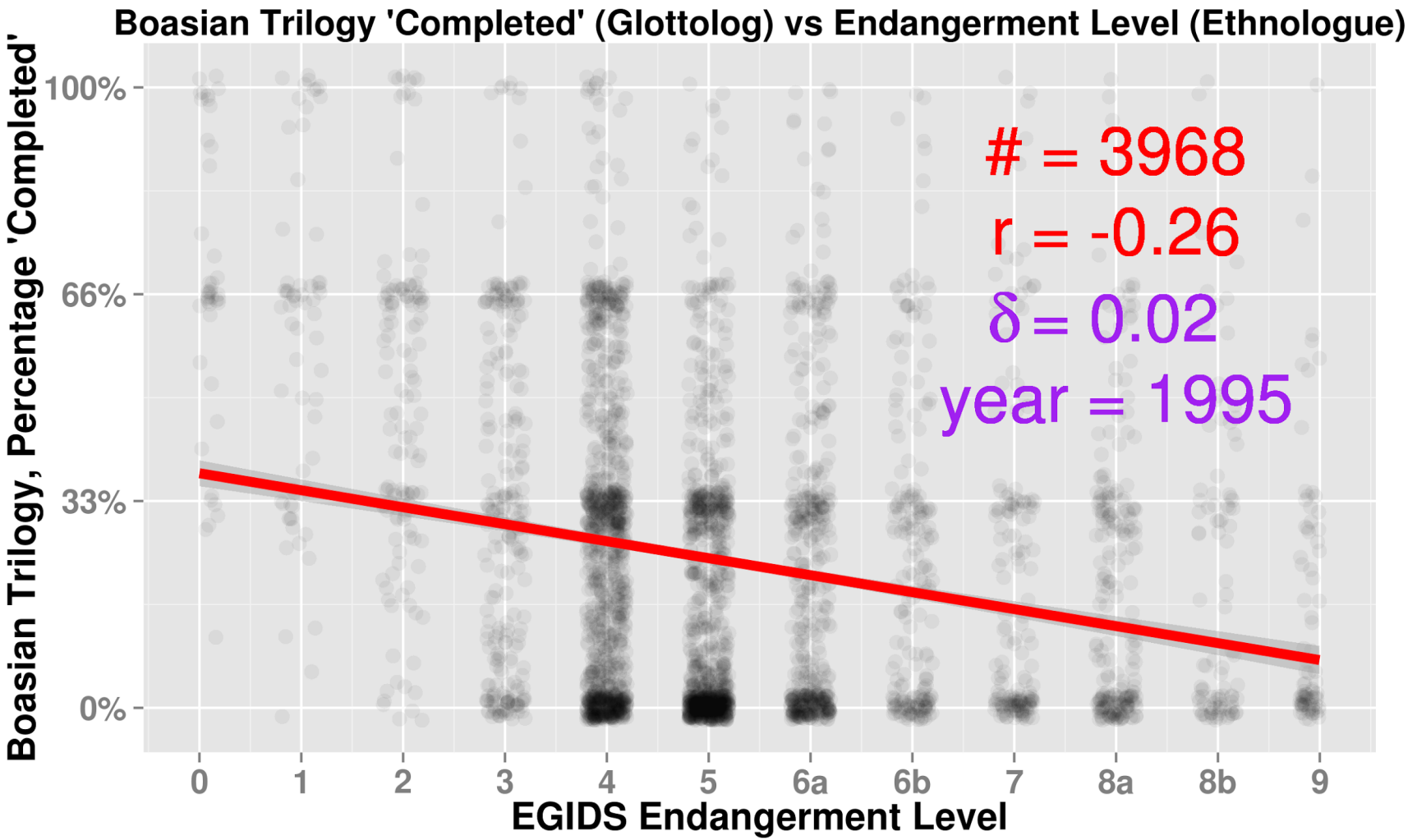

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 2303
r = -0.28
δ = 0
year = 1970
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

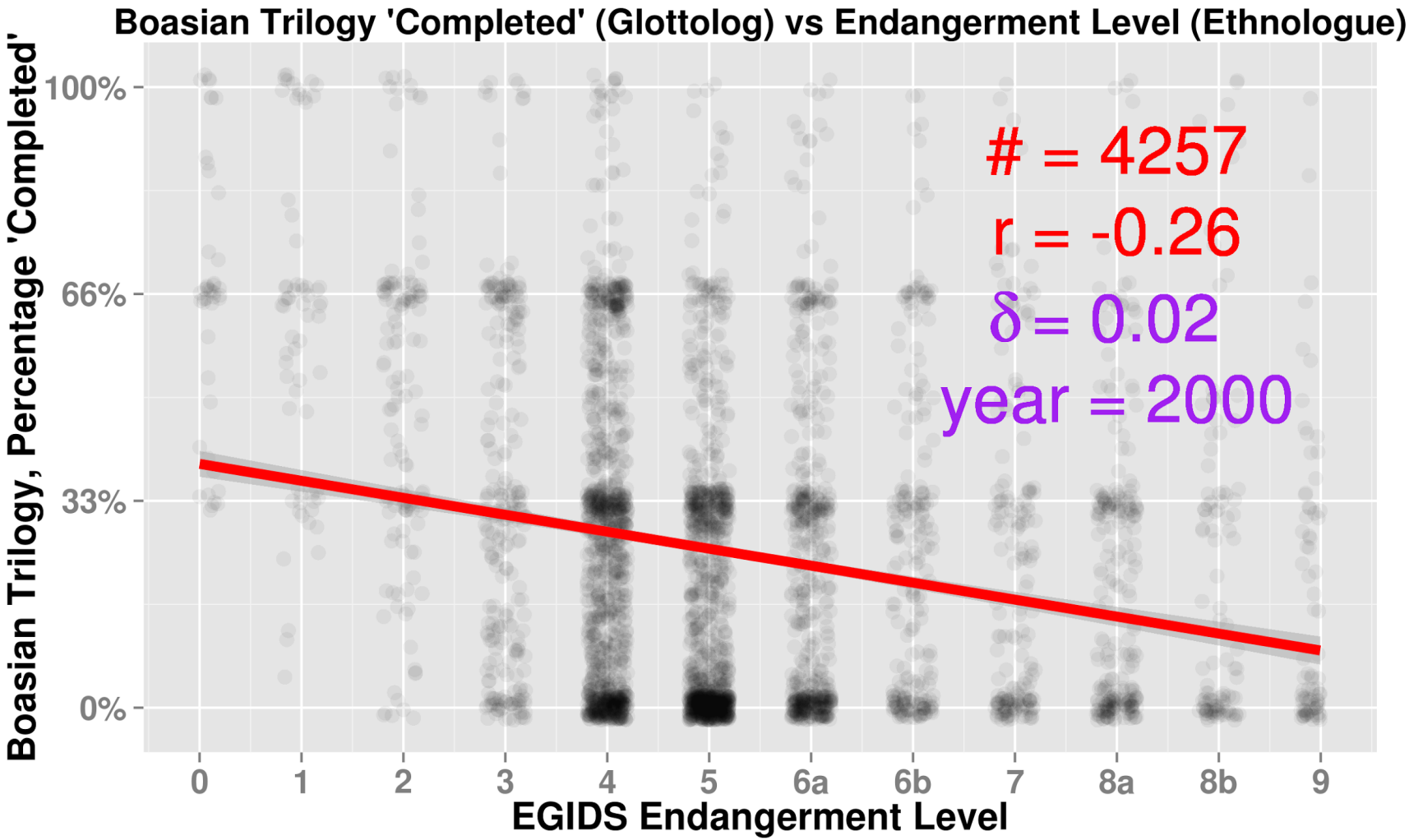

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 2734
r = -0.25
δ = 0.03
year = 1975
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

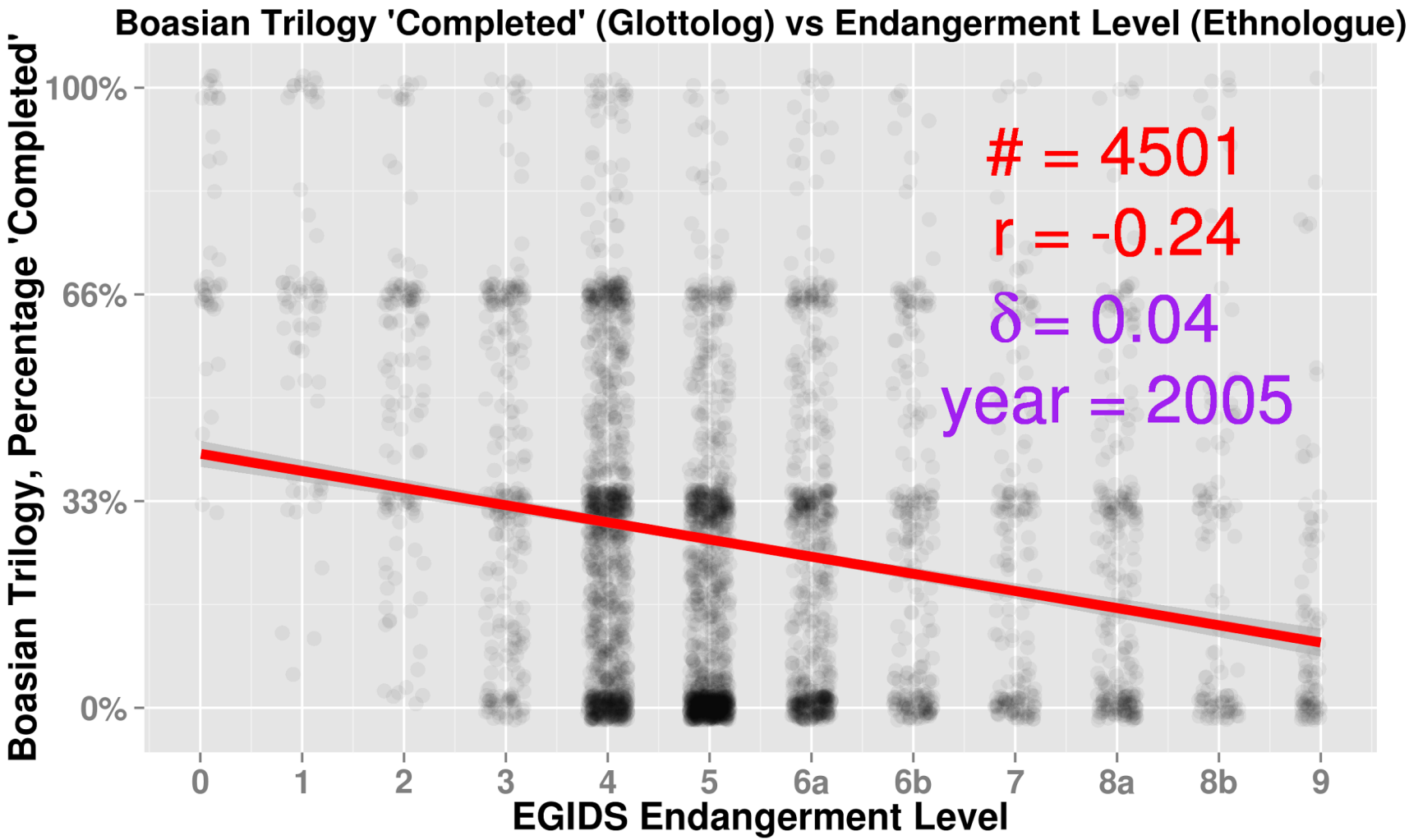

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 3092
r = -0.24
δ = 0.04
year = 1980
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

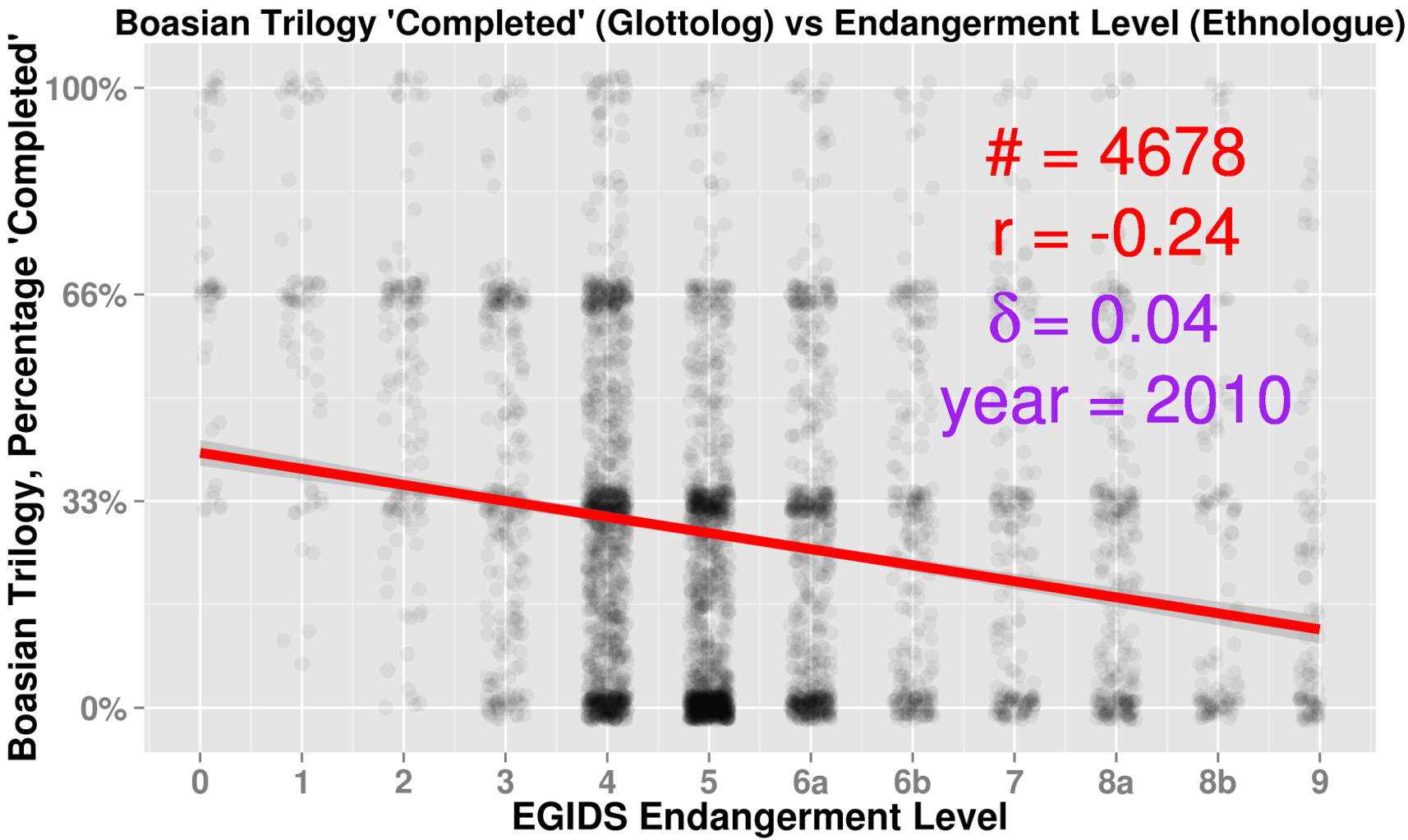

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 3388
r = -0.25
δ = 0.03
year = 1985
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

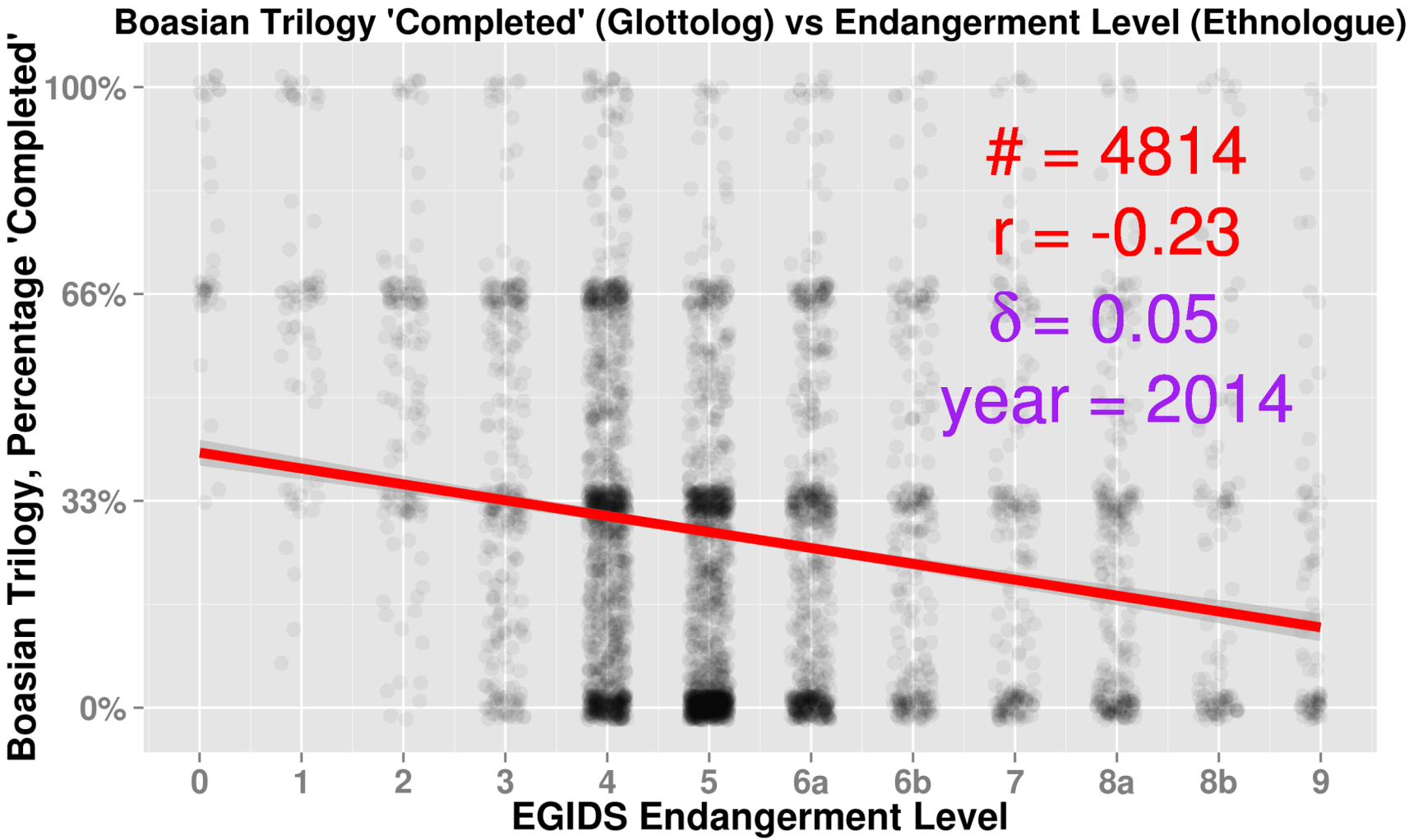

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 3706
r = -0.26
δ = 0.02
year = 1990
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level


Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 3968
r = -0.26
δ = 0.02
year = 1995
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level


Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 4257
r = -0.26
δ = 0.02
year = 2000
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

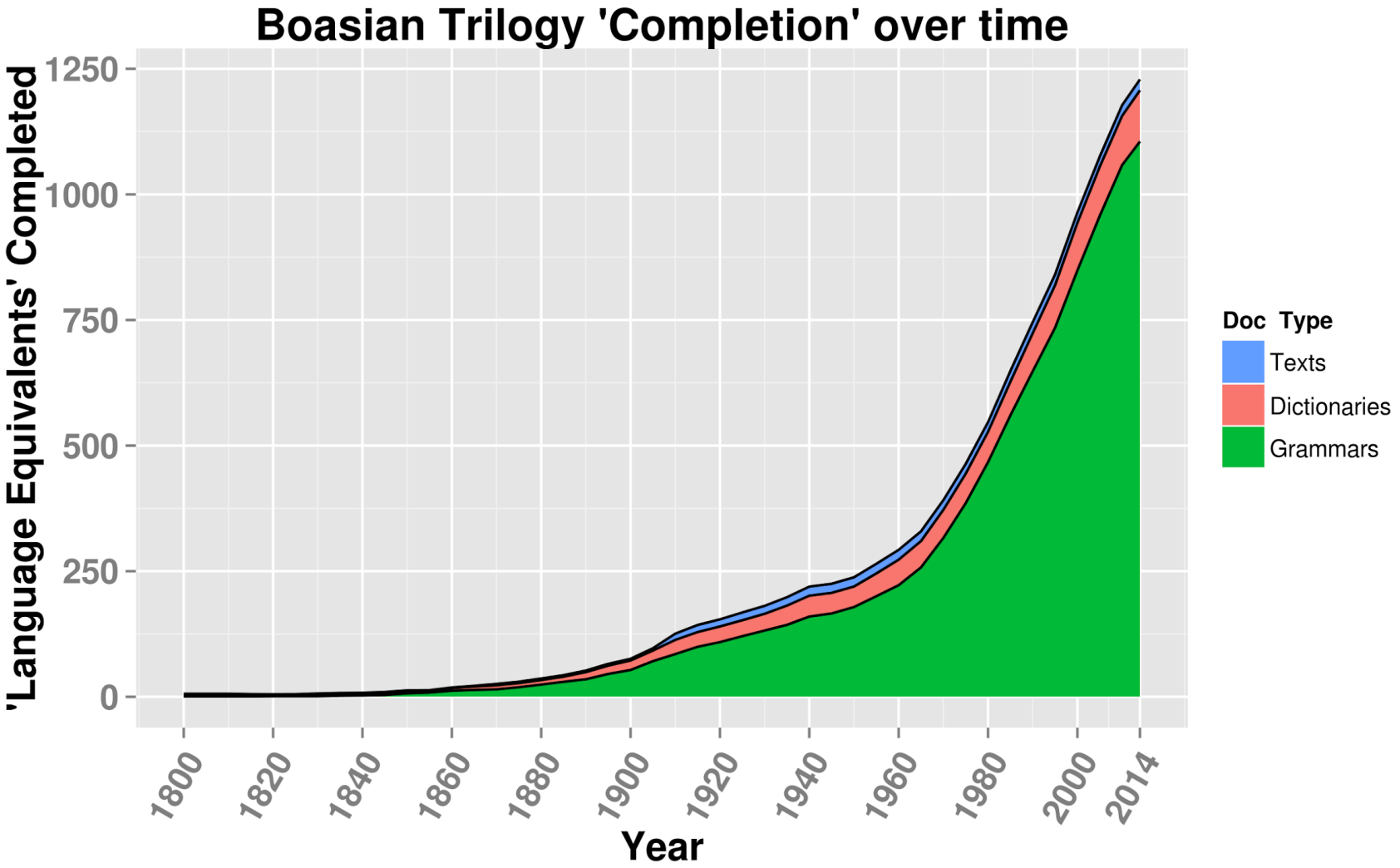

Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 4501
r = -0.24
δ = 0.04
year = 2005
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level


Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 4678
r = -0.24
δ = 0.04
year = 2010
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level


Boasian Trilogy 'Completed' (Glottolog) vs Endangerment Level (Ethnologue)
Boasian Trilogy, Percentage 'Completed'
100%
66%
33%
0%
# = 4814
r = -0.23
δ = 0.05
year = 2014
0
1
2
3
4
5
6a
6b
7
8a
8b
9
EGIDS Endangerment Level

Finally,

“Boasian completion” over time


Boasian Trilogy 'Completion' over time
'Language Equivalents' Completed
1250
1000
750
500
250
0
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
2014
Year
Doc Type
Texts
Dictionaries
Grammars

Conclusions

I didn’t assume I could answer any new questions,
my goal was to use data to verify what we qualitatively assume

Seeing the magnitudes of known phenomena is valuable

Lots of room for improving and expanding my methods

Lets move from a “one-man” to crowdsourced model

mahalo nui loa!