Re-examining the tradeoff between lexicon size and average morphosyntactic complexity in recursive numeral systems

David Yang and Terry Regier
UC Berkeley

How do you count in your language?

How do you count in your language?

Recursive numeral systems

How do you count in your language?

English

“one”, “two”, “three”, …, “ninety-nine”

1 2 3 $9 \cdot 10 + 9$

Recursive numeral systems

How do you count in your language?

English

"one", "two", "three", …, "ninety-nine"

1 2 3 $9 \cdot 10 + 9$

Mandarin

"一", "二", "三", …, "九十九"

1 2 3 $9 \cdot 10 + 9$

Recursive numeral systems

How do you count in your language?

English

"one", "two", "three", …, "ninety-nine"

1 2 3 $9 \cdot 10 + 9$

Mandarin

"一", "二", "三", …, "九十九"

1 2 3 $9 \cdot 10 + 9$

French

"un", "deux", "trois", …, "quatre vingt dix neuf"

1 2 3 $4 \cdot 20 + 10 + 9$

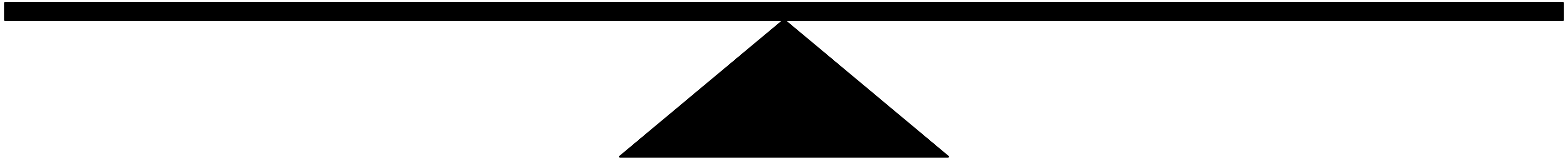
Recursive numeral systems

What explains this cross-language variation?

Efficient communication

communicatively
informative

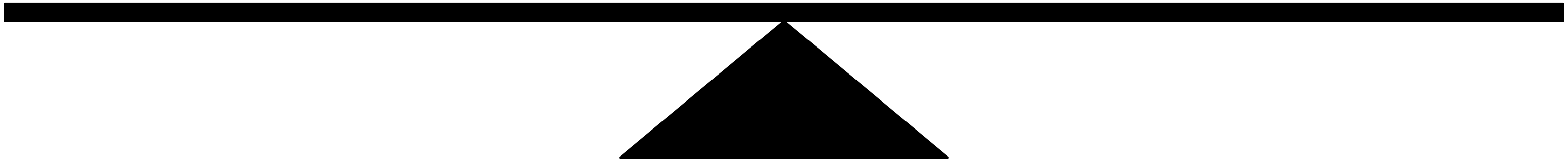
representationally
simple



Efficient communication

communicatively **informative**

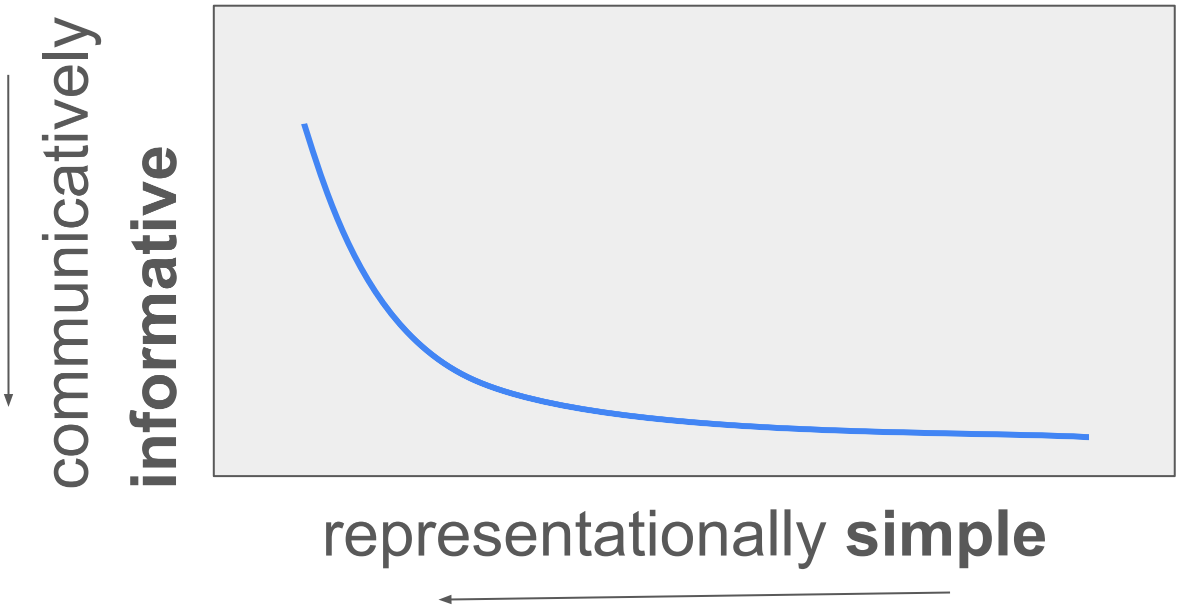
representationally **simple**



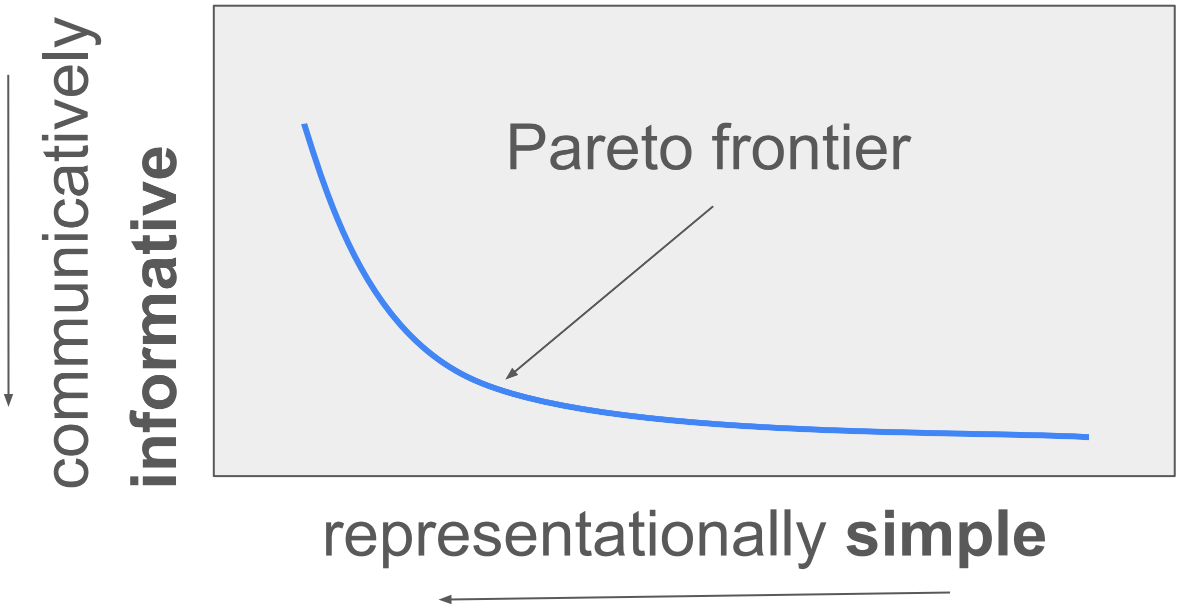
- Kinship (Kemp & Regier, 2012)
- Color (Regier et al., 2015)
- Numeral systems (Xu et al., 2020)

Efficient communication

Efficient communication



Efficient communication



Efficient communication

English

"one", "two", "three", …, "ninety-nine"

1 2 3 $9 \cdot 10 + 9$

Same informativeness

Mandarin

"一", "二", "三", …, "九十九"

1 2 3 $9 \cdot 10 + 9$

French

"un", "deux", "trois", …, "quatre vingt dix neuf"

1 2 3 $4 \cdot 20 + 10 + 9$

Efficient communication

English

"one", "two", "three", …, "ninety-nine"
1 2 3 $9 \cdot 10 + 9$

Mandarin

"一", "二", "三", …, "九十九"
1 2 3 $9 \cdot 10 + 9$

French

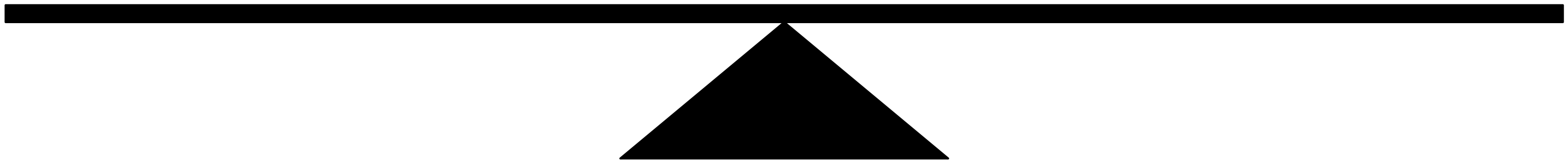
"un", "deux", "trois", …, "quatre vingt dix neuf"
1 2 3 $4 \cdot 20 + 10 + 9$

Same informativeness
Different complexities

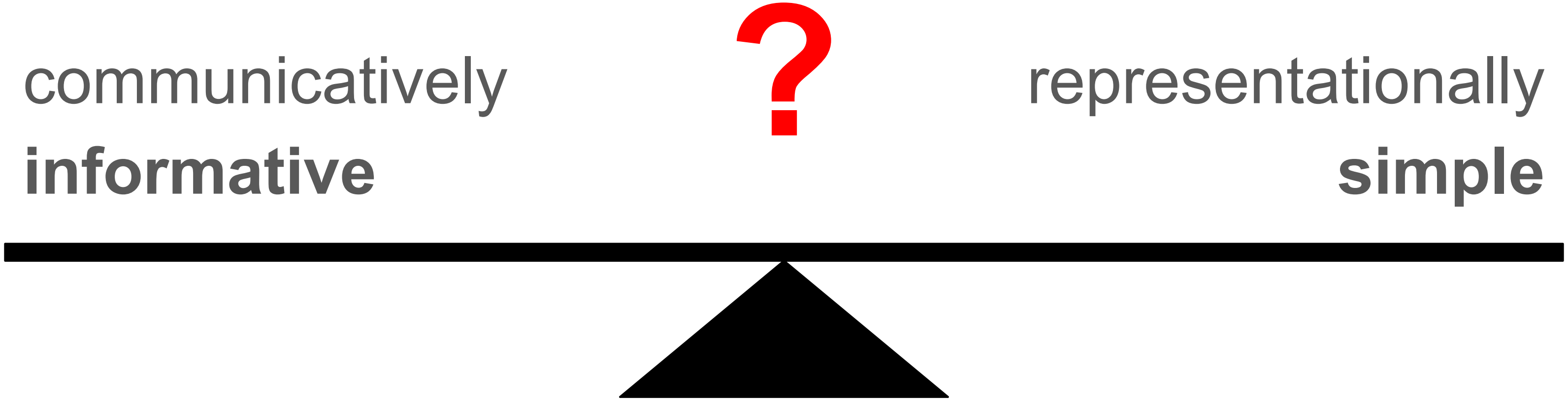
communicatively **informative**

?

representationally **simple**

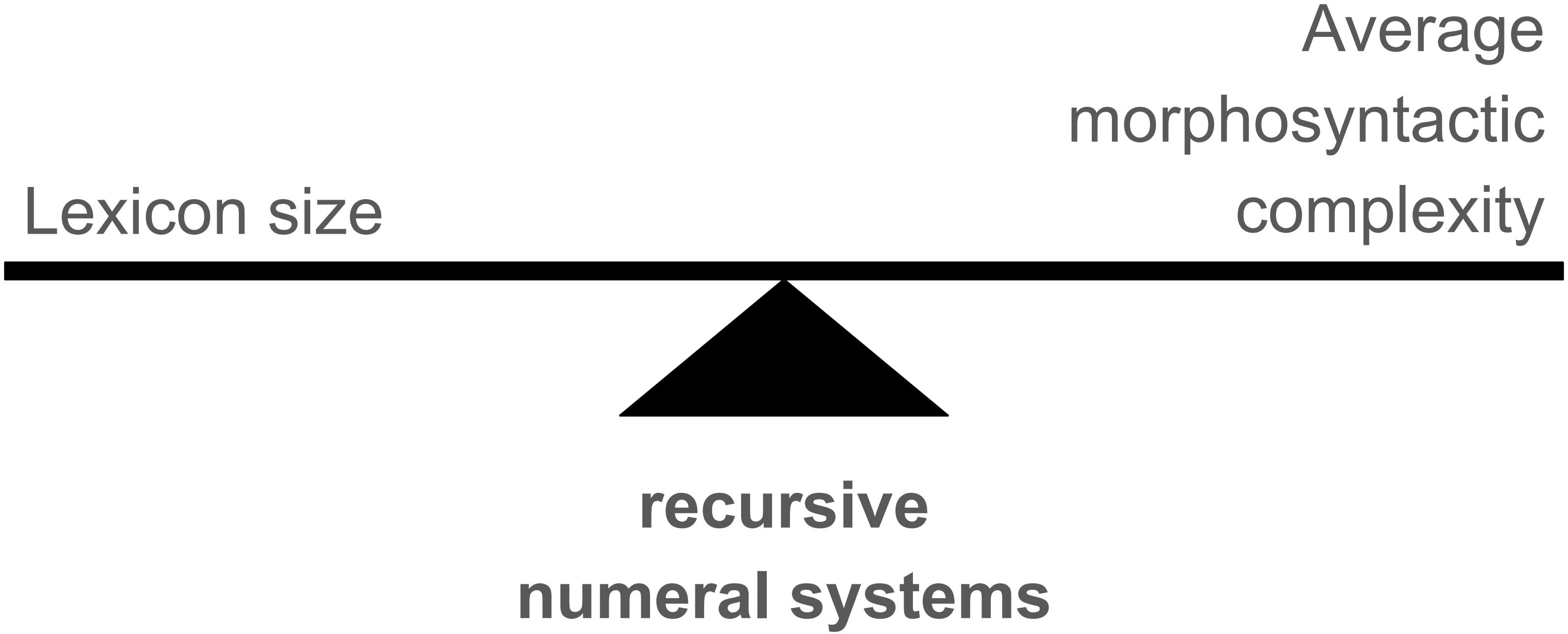


If this is not the explanation, what is?



Denić and Szymanik 2024

Denić and Szymanik 2024



Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

- Prior over numbers: $P(n) \propto n^{-2}$ (power-law prior Dehaene & Mehler, 1992)

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

Number	Numeral	Morphosyntax	Complexity
4	*four*	4	1

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

Number	**Numeral**	**Morphosyntax**	**Complexity**
4	*four*	4	1
14	*fourteen*	10 + 4	3

Denić and Szymanik 2024

Lexicon size: Number of lexicalized numerals

- E.g. English has a lexicon size of 12: {1, 2, 3, …, 11, 12}

Average morphosyntactic complexity (amsc):

Number	Numeral	Morphosyntax	Complexity
4	*four*	4	1
14	*fourteen*	10 + 4	3
49	*forty-nine*	4 · 10 + 9	5

Morphosyntax grammar

Number	Numeral	**Morphosyntax**	Complexity
4	*four*	4	1
14	*fourteen*	$10 + 4$	3
49	*forty-nine*	$4 \cdot 10 + 9$	5

Morphosyntax grammar

Number	Numeral	**Morphosyntax**	Complexity
4	*four*	4	1
14	*fourteen*	10 + 4	3
49	*forty-nine*	4 · 10 + 9	5

We need a **grammar that defines the morphosyntax** of numerals.

Grammar (Hurford 1975, 2007)

NUMBER → *D* | PHRASE | PHRASE + NUMBER | PHRASE − NUMBER

PHRASE → *M* | NUMBER · *M*

Grammar (Hurford 1975, 2007)

NUMBER → *D* | PHRASE | PHRASE + NUMBER | PHRASE − NUMBER

PHRASE → *M* | NUMBER · *M*

English

- *D* (digit): {1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12}
- *M* (multiplier): {10}

Grammar (Hurford 1975, 2007)

NUMBER $\rightarrow$ D | PHRASE | PHRASE + NUMBER | PHRASE − NUMBER

PHRASE $\rightarrow$ M | NUMBER · M

49 derivation

- NUMBER $\rightarrow$ PHRASE + NUMBER $\rightarrow$ NUMBER · M + NUMBER $\rightarrow$ D · M + D
- 49 $\rightarrow$ 40 + 9 $\rightarrow$ 4 · 10 + 9

Grammar (Hurford 1975, 2007)

NUMBER → D | PHRASE | PHRASE + NUMBER | PHRASE − NUMBER

PHRASE → M | NUMBER · M

49 derivation

- NUMBER → PHRASE + NUMBER → NUMBER · M + NUMBER → D · M + D
- 49 → 40 + 9 → 4 · 10 + 9

D&S used this grammar to create optimal artificial languages and create a Pareto frontier.

Denić and Szymanik's central result

Natural languages lie **on or near the Pareto frontier** of lexicon size and average morphosyntactic complexity

However, their artificial languages are not **entirely comparable** with natural languages

Denić and Szymanik's non-systematic artificial languages

Numeral's denotation	**Numeral's morphosyntax**
10	10
11	10 + 1
12	3 · 4
13	10 + 3
14	10 + 4
15	10 + 2 + 3
16	4 · 4

Denić and Szymanik, 2024

Our contribution

Building on Denić and Szymanik:

Our contribution

Building on Denić and Szymanik:

- Add grammar constraints on top of Hurford's grammar for a stricter comparison between artificial and natural languages

Our contribution

Building on Denić and Szymanik:

- Add grammar constraints on top of Hurford’s grammar for a stricter comparison between artificial and natural languages
- The role of the prior over numbers

Grammar constraints

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
    - English: [[{10,..., 99}, 10]] $49 = (4 \cdot \mathbf{10}) + 9$

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
    - English: [[{10,..., 99}, 9]] $49 = (4 \cdot 10) + 9$

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
- **Number subtraction constraint:** Specifies the maximum number that can be subtracted from a phrase.

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
- **Number subtraction constraint:** Specifies the maximum number that can be subtracted from a phrase.
  - English: [ ]
  - Hindi: [[{10,...,80}, 1]]

$49 = (5 \cdot 10) - 1$

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals.
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
- **Number subtraction constraint:** Specifies the maximum number that can be subtracted from a phrase.
- **Exceptions constraint:** Defines a specific (non-canonical) construction for a numeral.

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
- **Number subtraction constraint:** Specifies the maximum number that can be subtracted from a phrase.
- **Exceptions constraint:** Defines a specific (non-canonical) construction for a numeral.
  - English: [ ]
  - Gola: [20, {20}, ‘(1·20)’]

20 = **(1·20)**

Grammar constraints

- **Base constraint:** Defines the base (multiplier) used to construct numerals
- **Number addition constraint:** Specifies the maximum number that can be added to a phrase.
- **Number subtraction constraint:** Specifies the maximum number that can be subtracted from a phrase.
- **Exceptions constraint:** Defines a specific (non-canonical) construction for a numeral.

* Added a suppletives category to Hurford's grammar (e.g. English 11 & 12).

Grammar constraints

Lexicon/Grammar	English	Gola
Digits	{1,2,...,9}	{1,2,3,4}
Bases	{10}	{5,10,20}
Suppletives	{11,12}	{}
Base constraint	[[{10,..., 99},10]]	[[{5,…,9},5],[{10,…,19},10],[{20,…,99},20]]
Number addition constraint	[[{10,...,99},9]]	[[{5,…,9},4],[{10,…,19},9],[{20,…,99},19]]
Number subtraction constraint	[]	[]
Exceptions constraint	[]	[[20,{20},‘(1·20)’]]

Grammar constraints

1. All natural languages in Denić and Szymanik's study (128 languages) can be generated with our grammar constraints.

Grammar constraints

1. All natural languages in Denić and Szymanik’s study (128 languages) can be generated with our grammar constraints.

2. Artificial languages are comparable with natural languages since they are generated from the same grammatical formalism.

Artificial language evolution

Similar to Denić and Szymanik:

Artificial language evolution

Similar to Denić and Szymanik:

1. **First generation:** 300 artificial language lexicons and grammar constraints were randomly created.

Artificial language evolution

Similar to Denić and Szymanik:

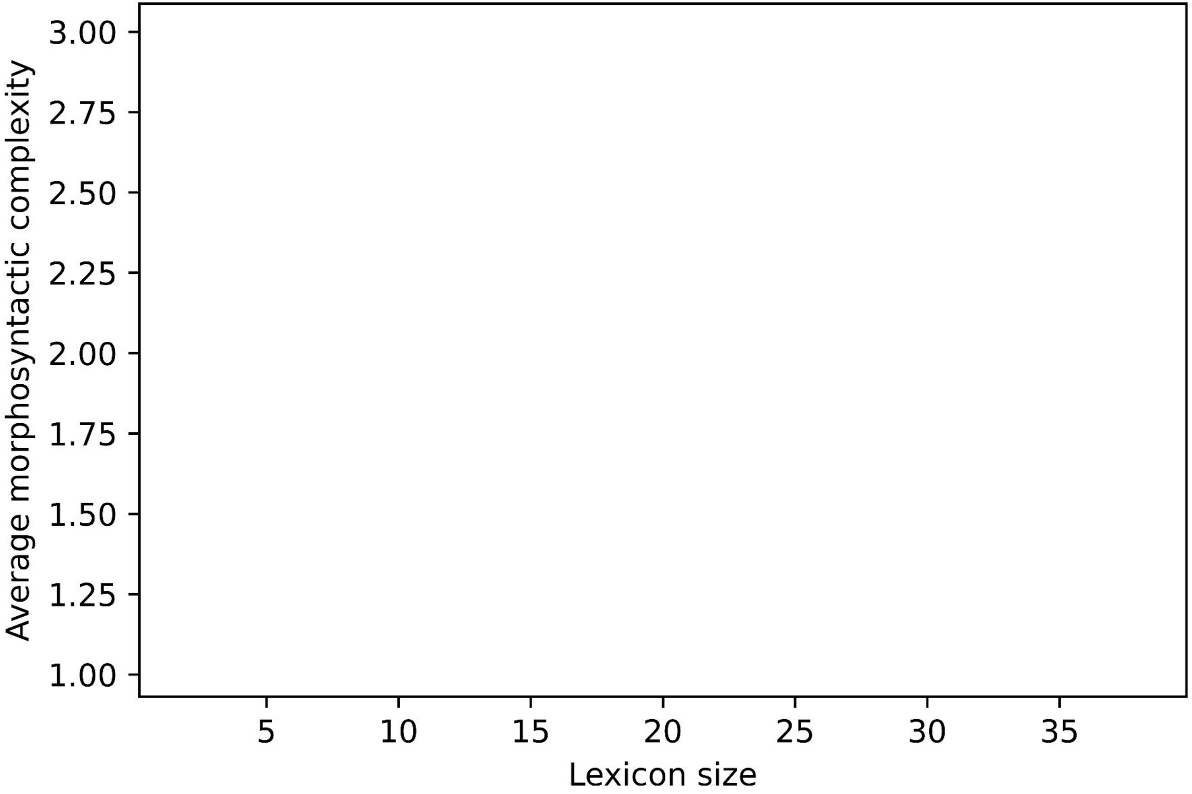
1. **First generation:** 300 artificial language lexicons and grammar constraints were randomly created.
2. **Next generation:** 50 new languages were created while mutating the optimal languages from the previous generation. This process was repeated for 100 generations.

Artificial language evolution

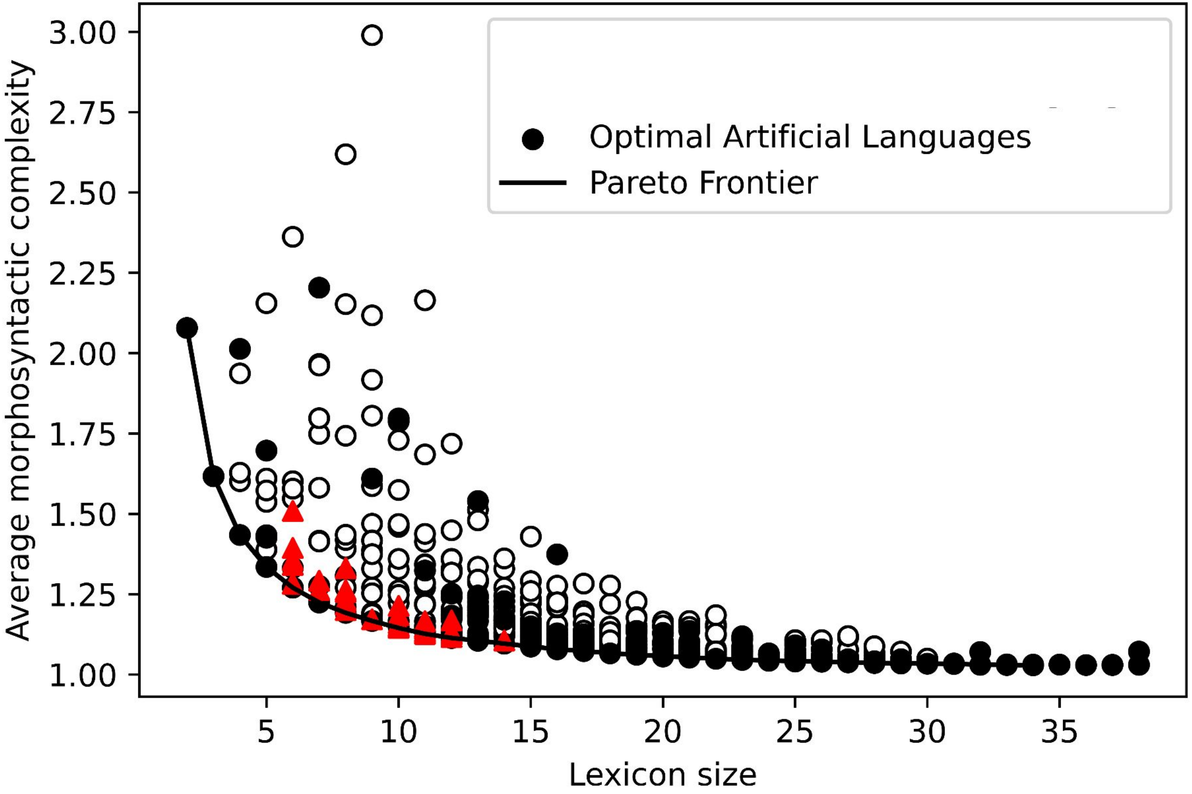
Similar to Denić and Szymanik:

1. **First generation:** 300 artificial language lexicons and grammar constraints were randomly created.
2. **Next generation:** 50 new languages were created while mutating the optimal languages from the previous generation. This process was repeated for 100 generations.
3. Finally, the last generation was combined with natural languages to create the Pareto frontier.

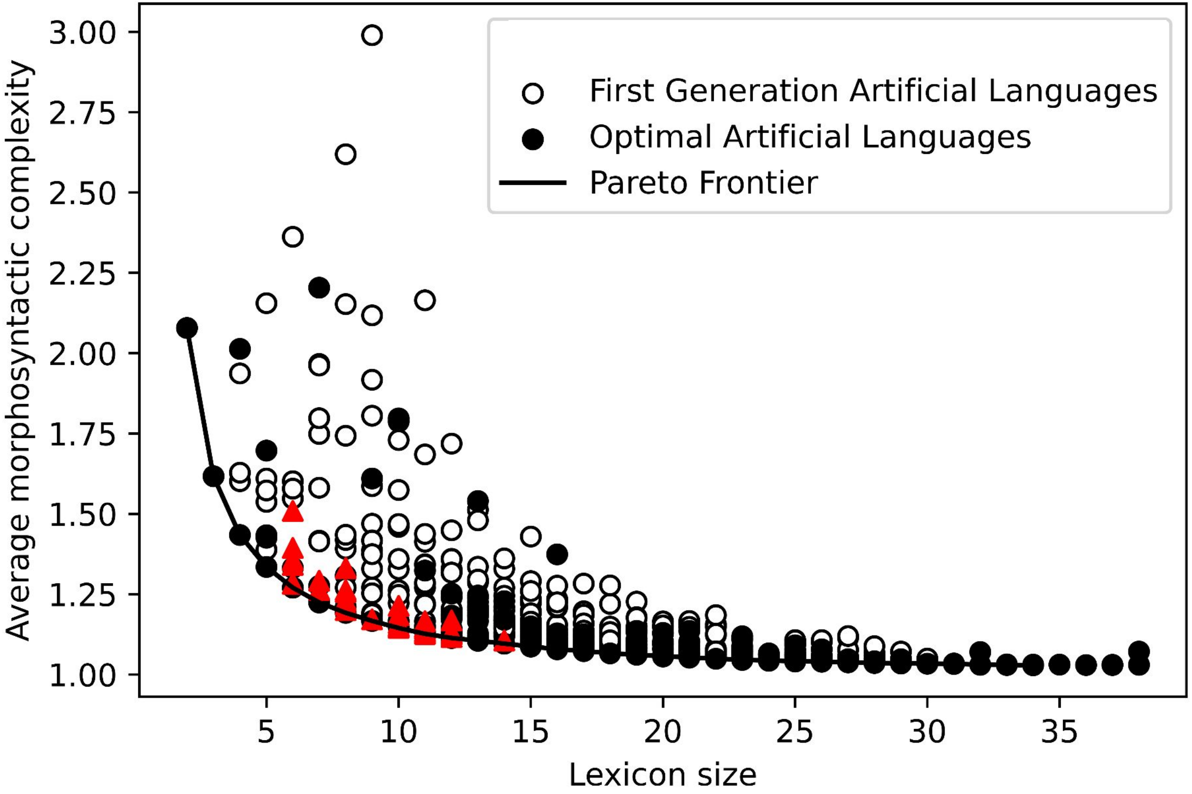
Pareto frontier with grammar constraints



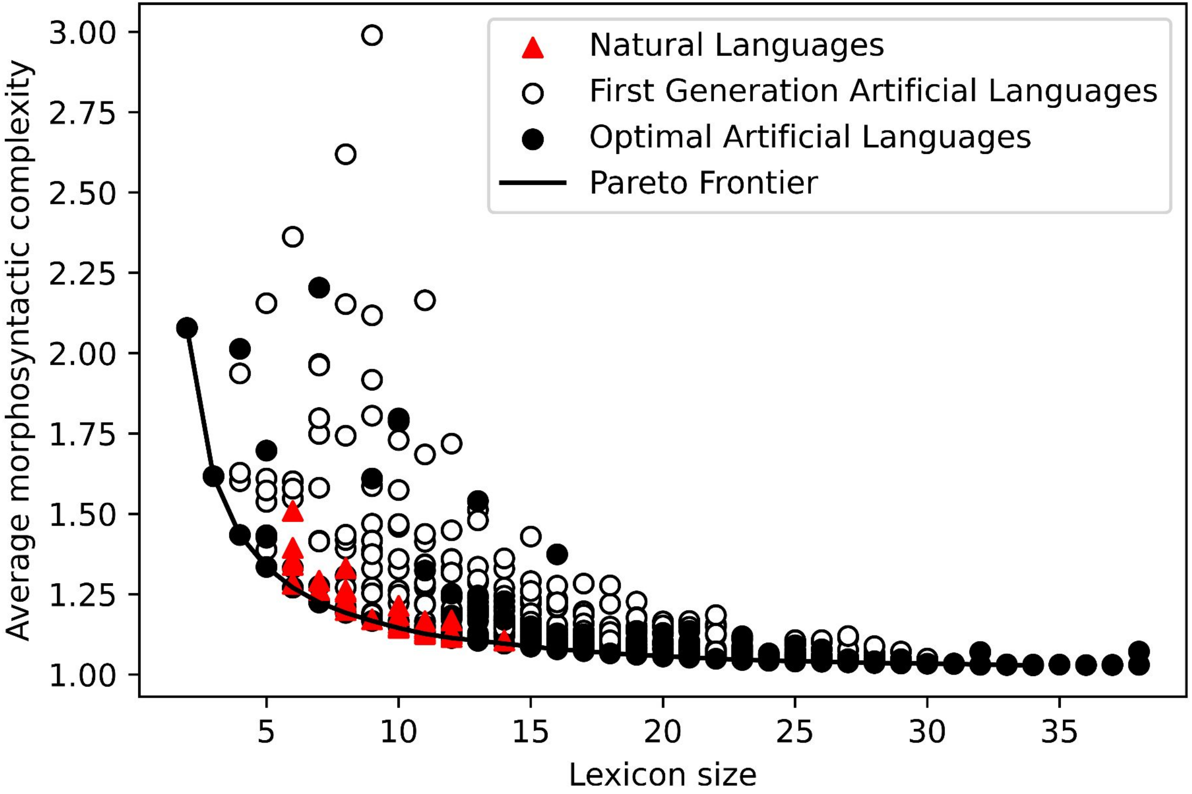
Pareto frontier with grammar constraints



Pareto frontier with grammar constraints



Pareto frontier with grammar constraints



Supports Denić and Szymanik’s main result - natural languages optimally trade off lexicon size and average morphosyntactic complexity - **even when addressing the issue of comparability** between natural and artificial languages.

The role of the prior

The role of the prior

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Power-law prior

$$P(n) \propto n^{-2}$$

The role of the prior

$$amsc(L) = \sum_{n \in [1, 99]} P(n) \cdot ms\_complexity(n, L)$$

Power-law prior

$$P(n) \propto n^{-2}$$

Represents the distribution of numeral usage frequencies actually found in natural language (Dehaene & Mehler, 1992)

Do natural numeral systems reflect the **frequency** with which people refer to specific numbers?

Xu et al., 2020
Pajot et al., 2025

Do natural numeral systems reflect the **frequency** with which people refer to specific numbers?

If so, we expect them to exhibit near-optimal tradeoff under the real prior and be **less optimal with alternate priors**.

Xu et al., 2020
Pajot et al., 2025

Alternate priors

Uniform prior

$P(n) = 1/99$

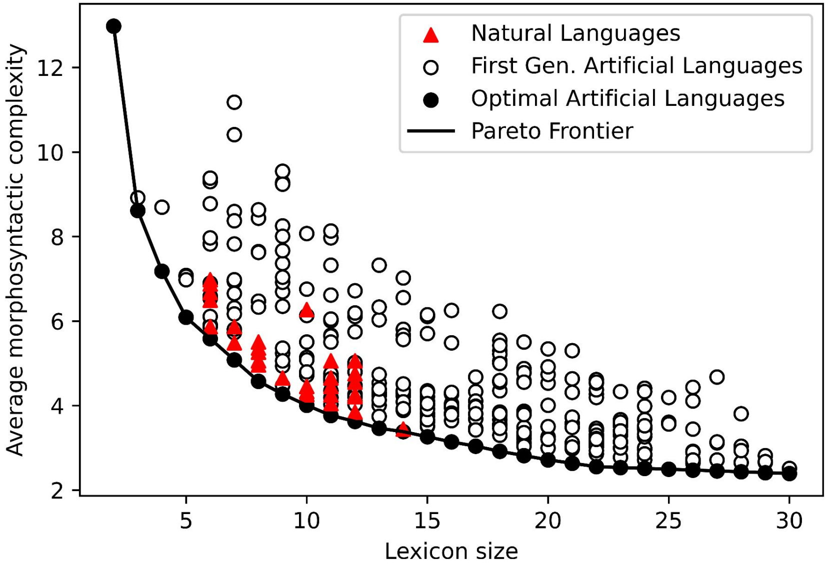
Reverse power-law prior

$P(n) \propto (100 - n)^{-2}$

Alternate priors

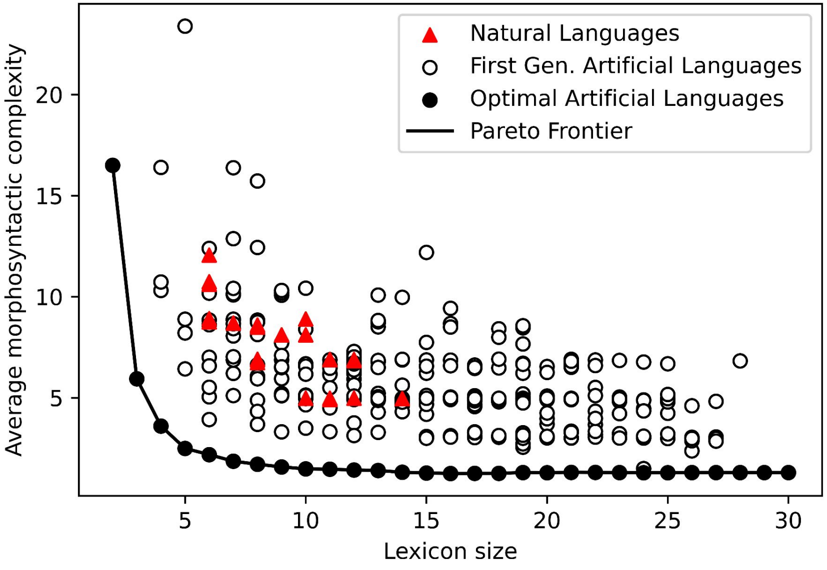
Uniform prior

$P(n) = 1/99$

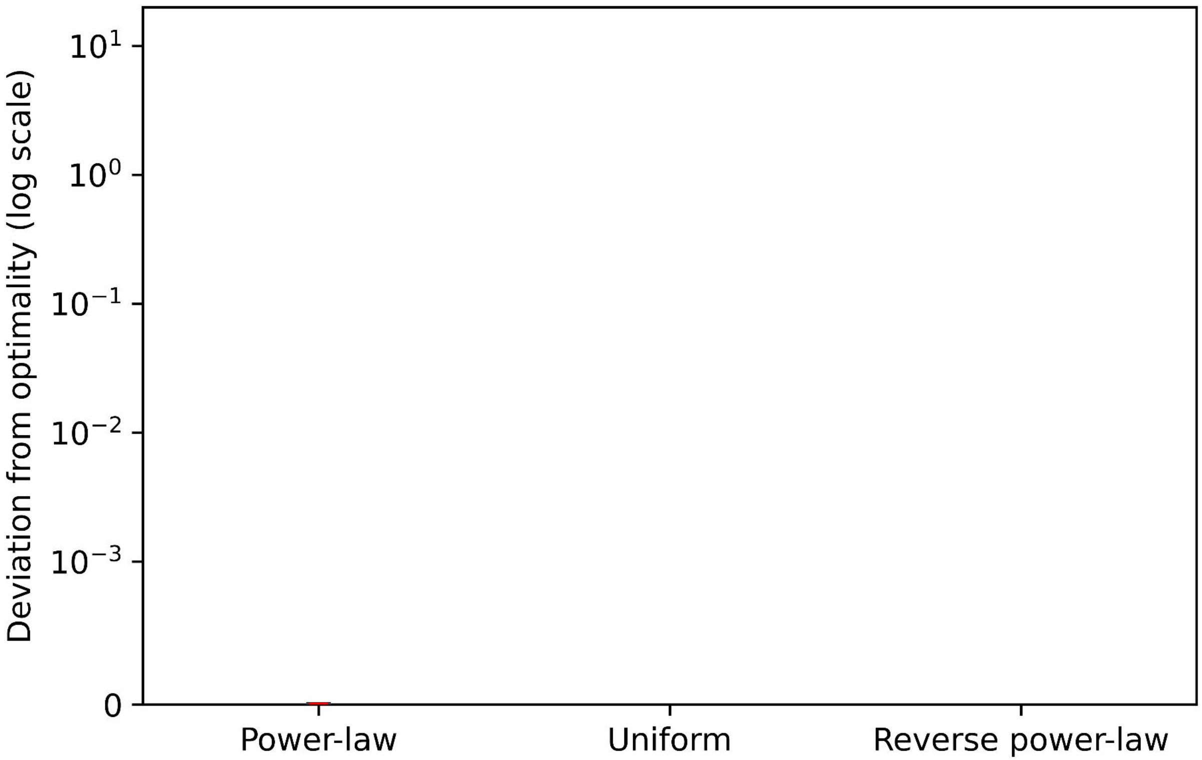


Reverse power-law prior

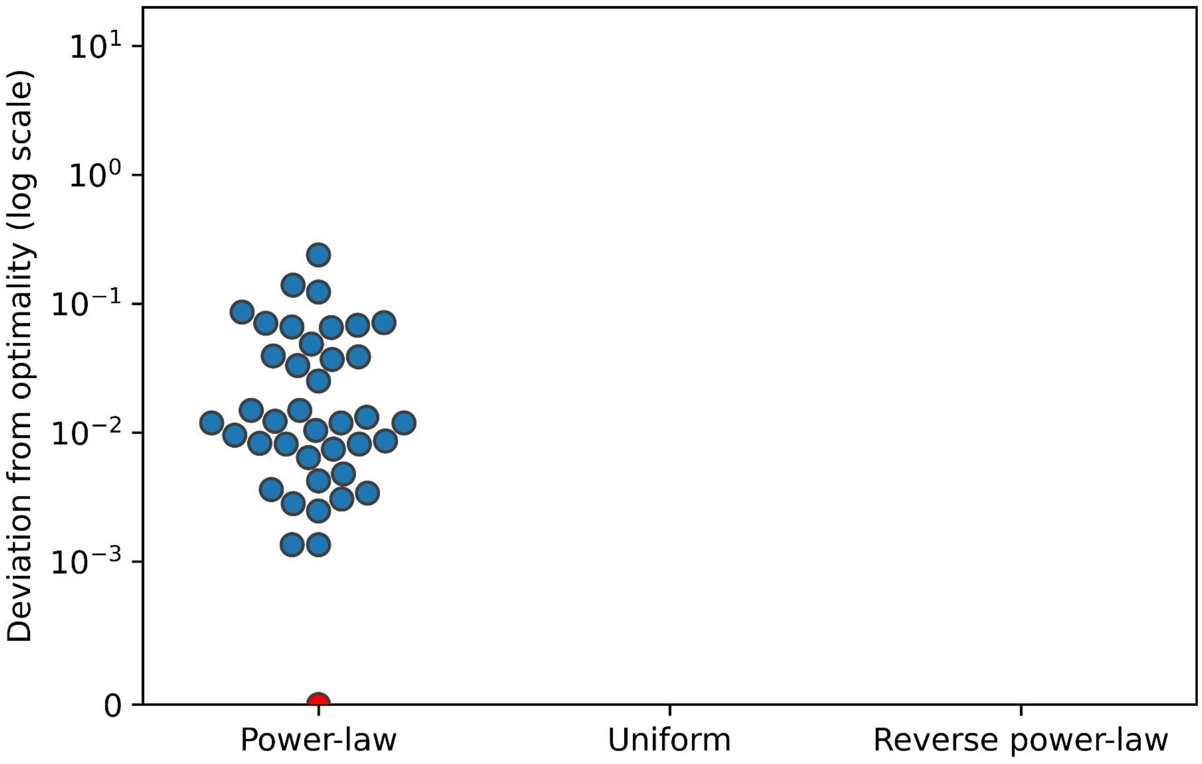
$P(n) \propto (100 - n)^{-2}$



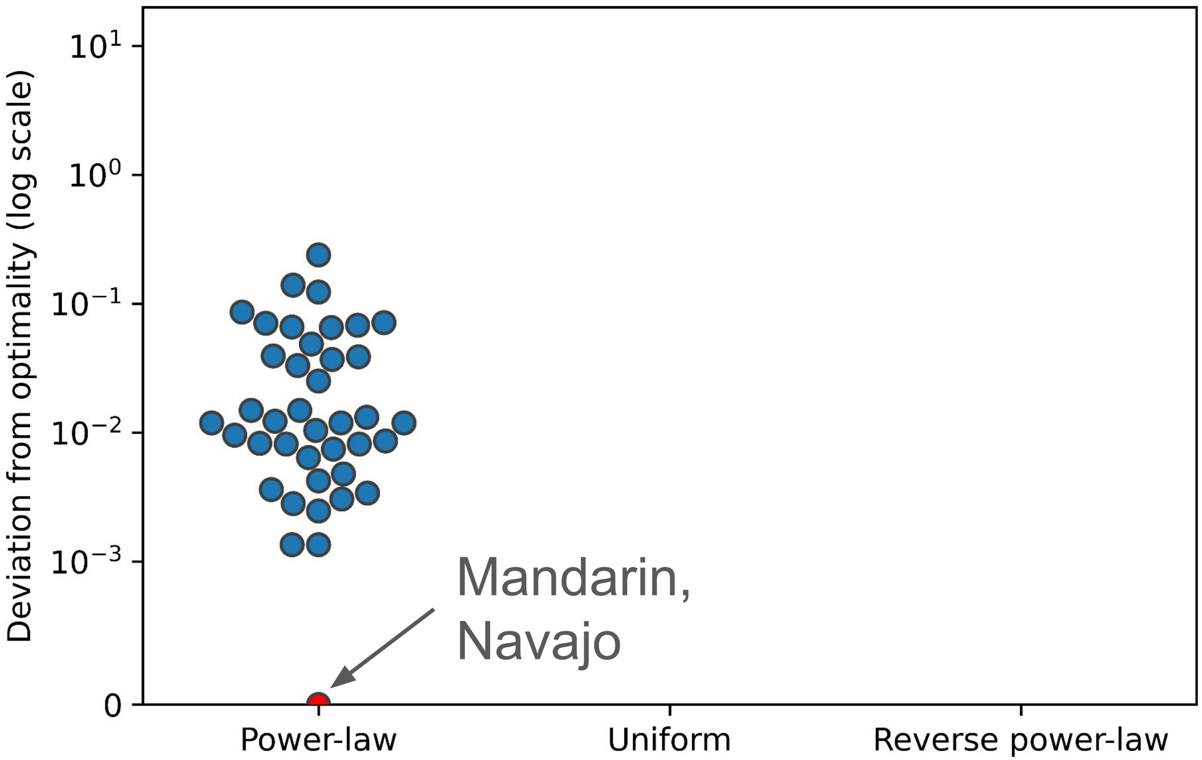
The role of the prior



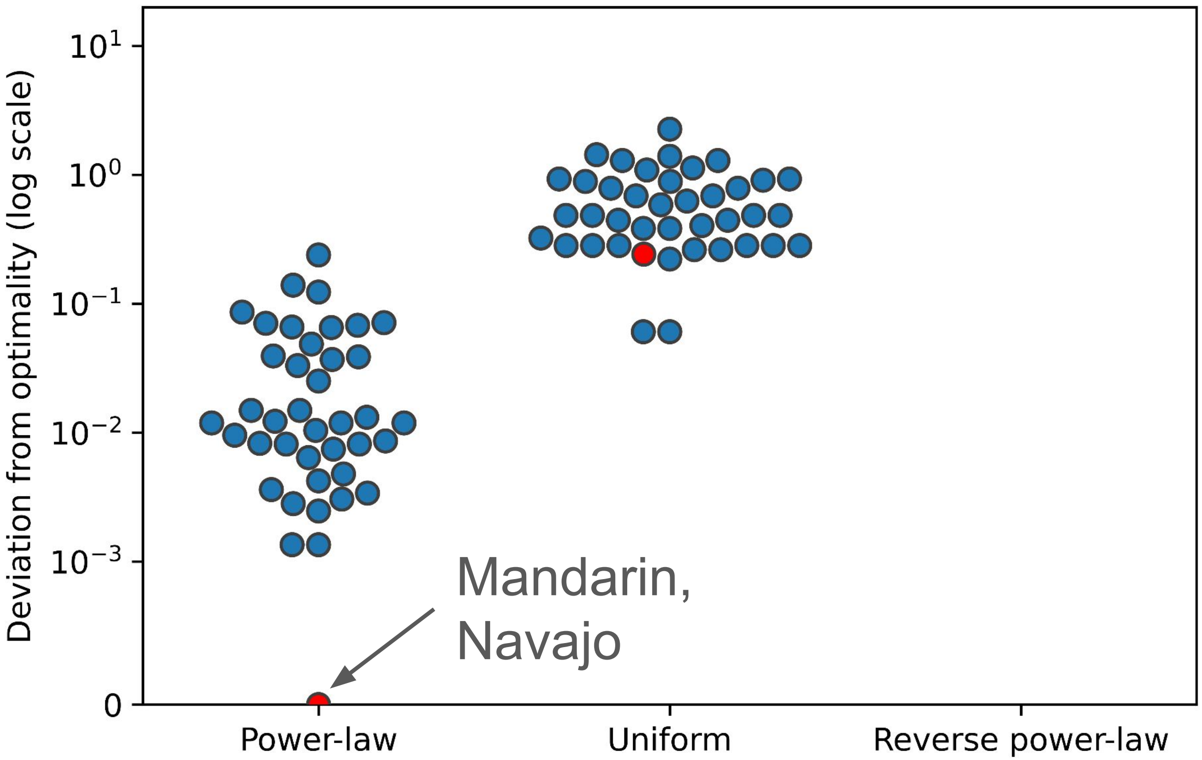
The role of the prior



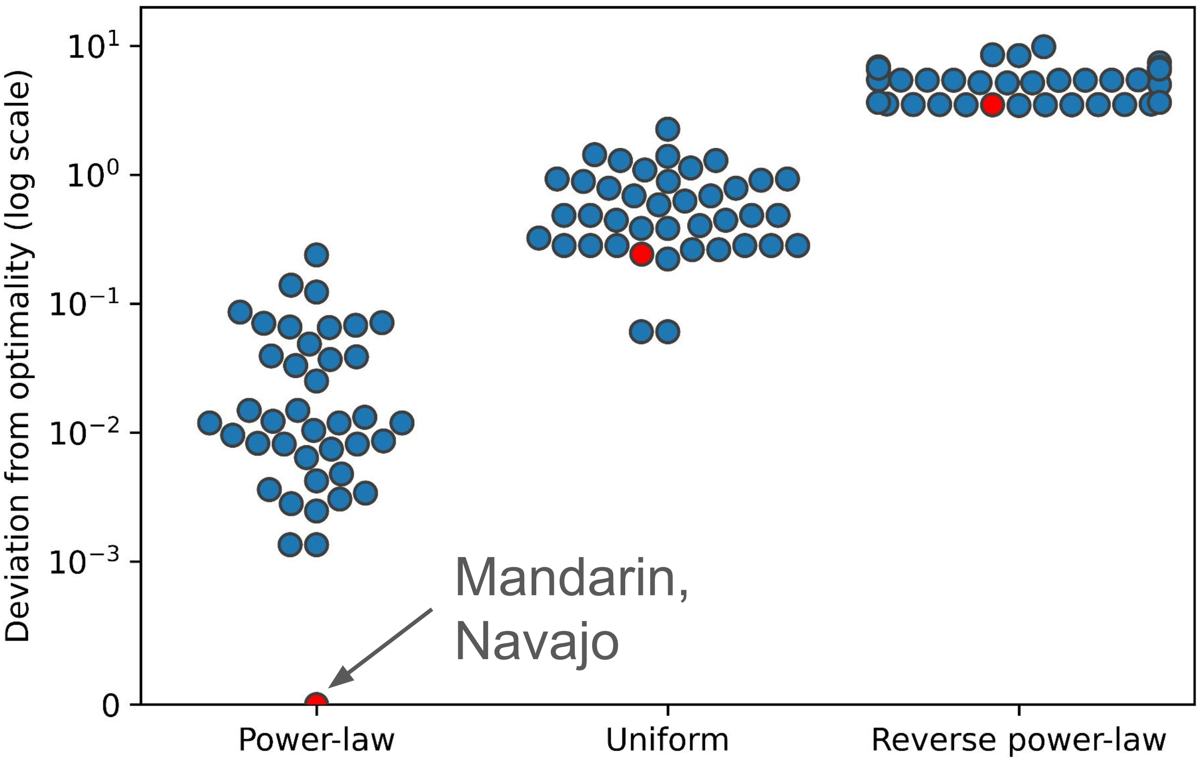
The role of the prior



The role of the prior



The role of the prior



Attested numeral systems are well-suited for communication under the power-law prior, but **are not as well-suited for communication under those alternate priors**.

Natural numeral systems **do reflect the frequency** with which people refer to specific numbers.

Natural numeral systems **do reflect the frequency** with which people refer to specific numbers.

This suggests a **possible adaptation** of numeral systems to **usage frequencies**.

Conclusions

Conclusions

1. Traditional two goals of informativeness and simplicity could not explain natural recursive numeral systems.

Conclusions

1. Traditional two goals of informativeness and simplicity could not explain natural recursive numeral systems.
2. Denić and Szymanik showed natural languages optimally trade off lexicon size and average morphosyntactic complexity.

Conclusions

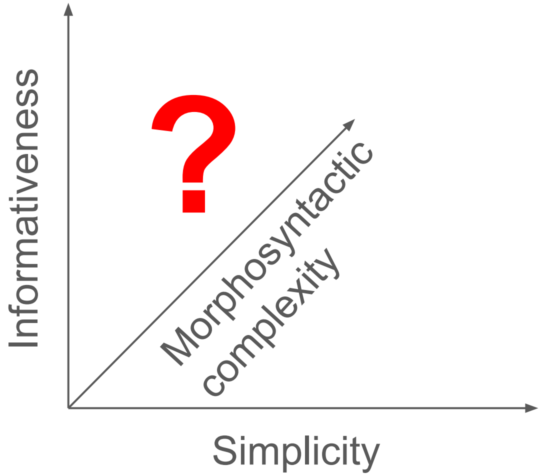
1. Traditional two goals of informativeness and simplicity could not explain natural recursive numeral systems.
2. Denić and Szymanik showed natural languages optimally trade off lexicon size and average morphosyntactic complexity.
3. Their central result **continues to hold** when natural and artificial numeral systems are made entirely comparable.

Conclusions

1. Traditional two goals of informativeness and simplicity could not explain natural recursive numeral systems.
2. Denić and Szymanik showed natural languages optimally trade off lexicon size and average morphosyntactic complexity.
3. Their central result **continues to hold** when natural and artificial numeral systems are made entirely comparable.
4. Natural numeral systems exhibit the near-optimal tradeoff when assessed under a prior that **reflects the frequencies** with which people name different numbers, but are less optimal under alternate priors.

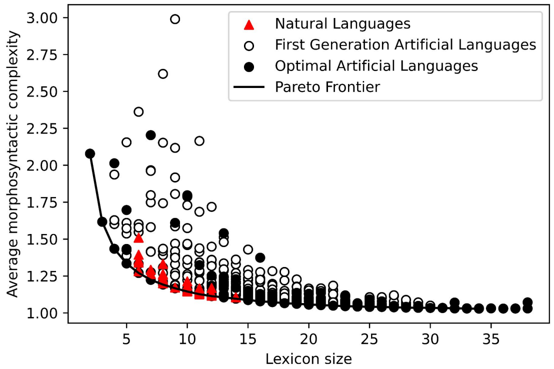
Open question

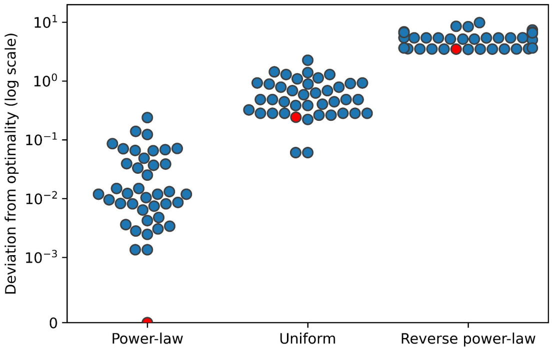
What role does **morphosyntactic complexity** play in other semantic domains and efficient communication in general?



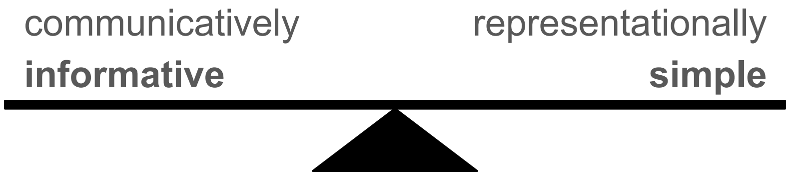
Thank you!

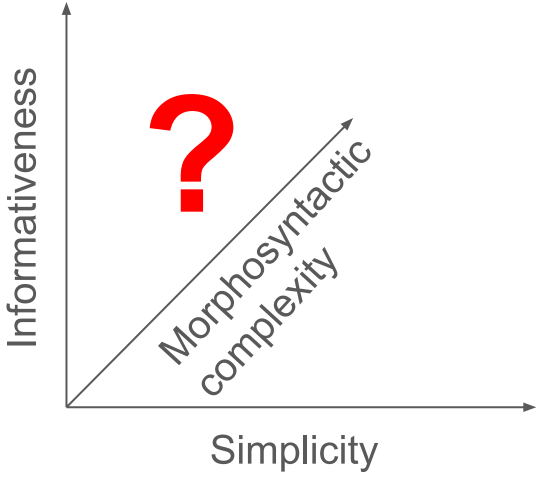
- Milica Denić and Jakub Szymanik
- Bernard Comrie


Natural Languages
First Generation Artificial Languages
Optimal Artificial Languages
Pareto Frontier
Average morphosyntactic complexity
Lexicon size
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
5
10
15
20
25
30
35


Deviation from optimality (log scale)
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
Power-law
Uniform
Reverse power-law

Questions?


communicatively
informative
representationally
simple


Informativeness
?
Morphosyntactic
complexity
Simplicity