Language and cognitive diversity and global wellbeing

# The Behemoth Language Bias

Damián Blasi

ICREA

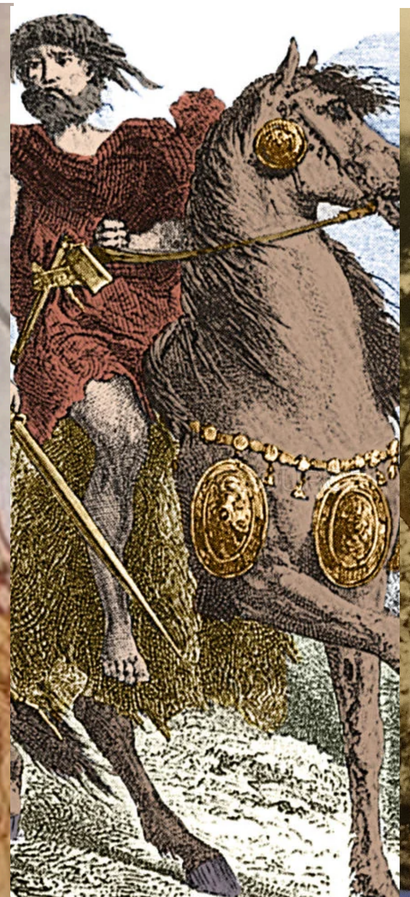
upf. Universitat Pompeu Fabra Barcelona

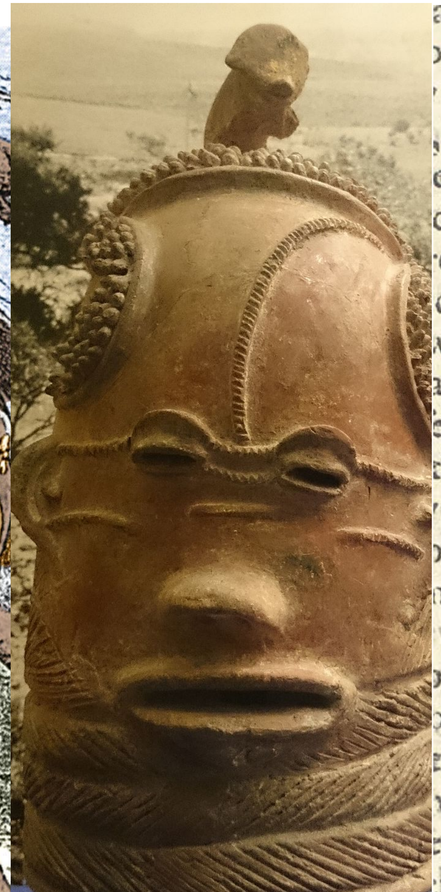
CBC Center for Brain and Cognition

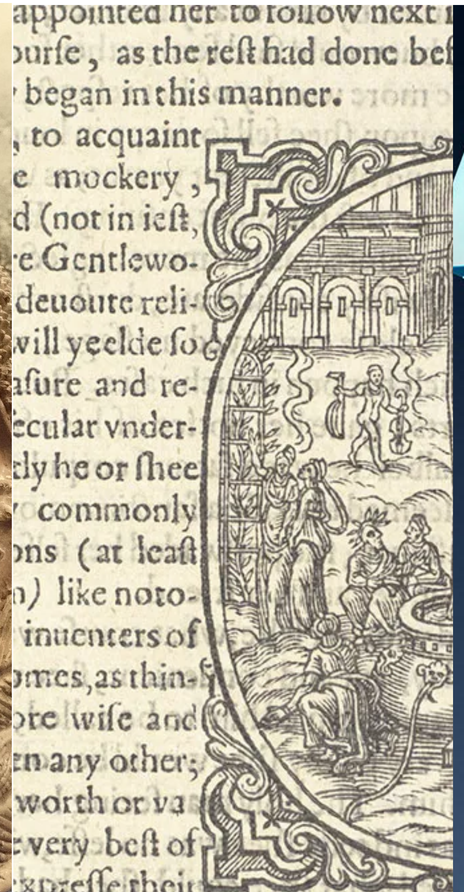
UNESCO

The Branco Weiss Fellowship Society in Science

# I research **language diversity** over broad scales of time and space through a multidisciplinary lens








*Why should anyone besides us care or invest into language diversity research?*

## Argument I

# The romantic spiel

Language diversity is a rightful object of human curiosity and awe

JUNE 1, 2023 | 15 MIN READ

# This Ancient Language Has the Only Grammar Based Entirely on the Human Body

An endangered language family suggests that early humans used their bodies as a model for reality




# 20 amazing words that don't exist in English — but really should

Shelby Slauer Updated Oct 9, 2019, 5:51 PM GMT+2

Share | Save

## "Schadenfreude" — German

**de" once in a while.** Alan Crowhurst/Getty Images

## Argument II

# The study of languaged humans

Some aspects of human behavior and cognition are influenced by the language(s) used by the individual. No true science of the languaged human without sampling across languages

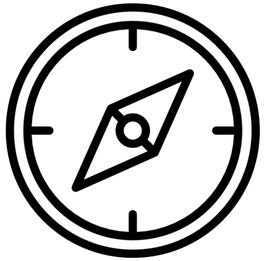
Trends in
Cognitive Sciences

CellPress
OPEN ACCESS

Feature Review

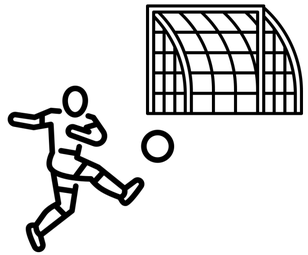
# Over-reliance on English hinders cognitive science

Damián E. Blasi, [1,2,3,*,@] Joseph Henrich, [1] Evangelia Adamou, [4] David Kemmerer, [5,6] and Asifa Majid [7,*,@]



COMPUTATIONS
AND REPRESENTATIONS

SOCIAL COGNITION



CAUSAL COGNITION

VISUAL AND
AUDITORY ATTENTION

BRAIN WIRING
PATTERNS

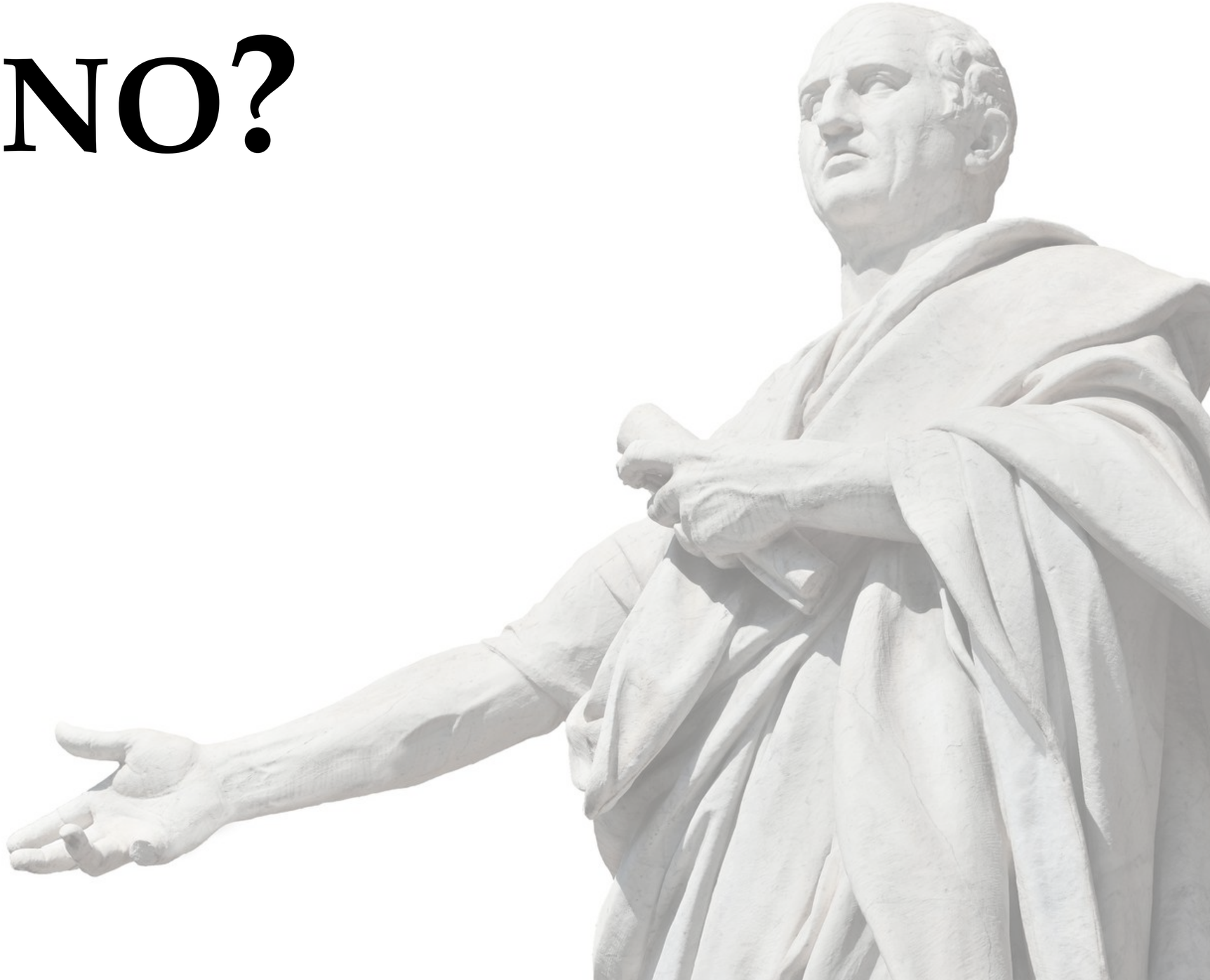
## Argument III

# The ethical imperative

Languages become dormant (i.e. lose all active users) at a concerning rate, and documenting this diversity should be at the fore of our efforts concerning language

Former rainforest environment of a Bagyeli population in Mimbosso, Centre-South province, Cameroon (June 2024)

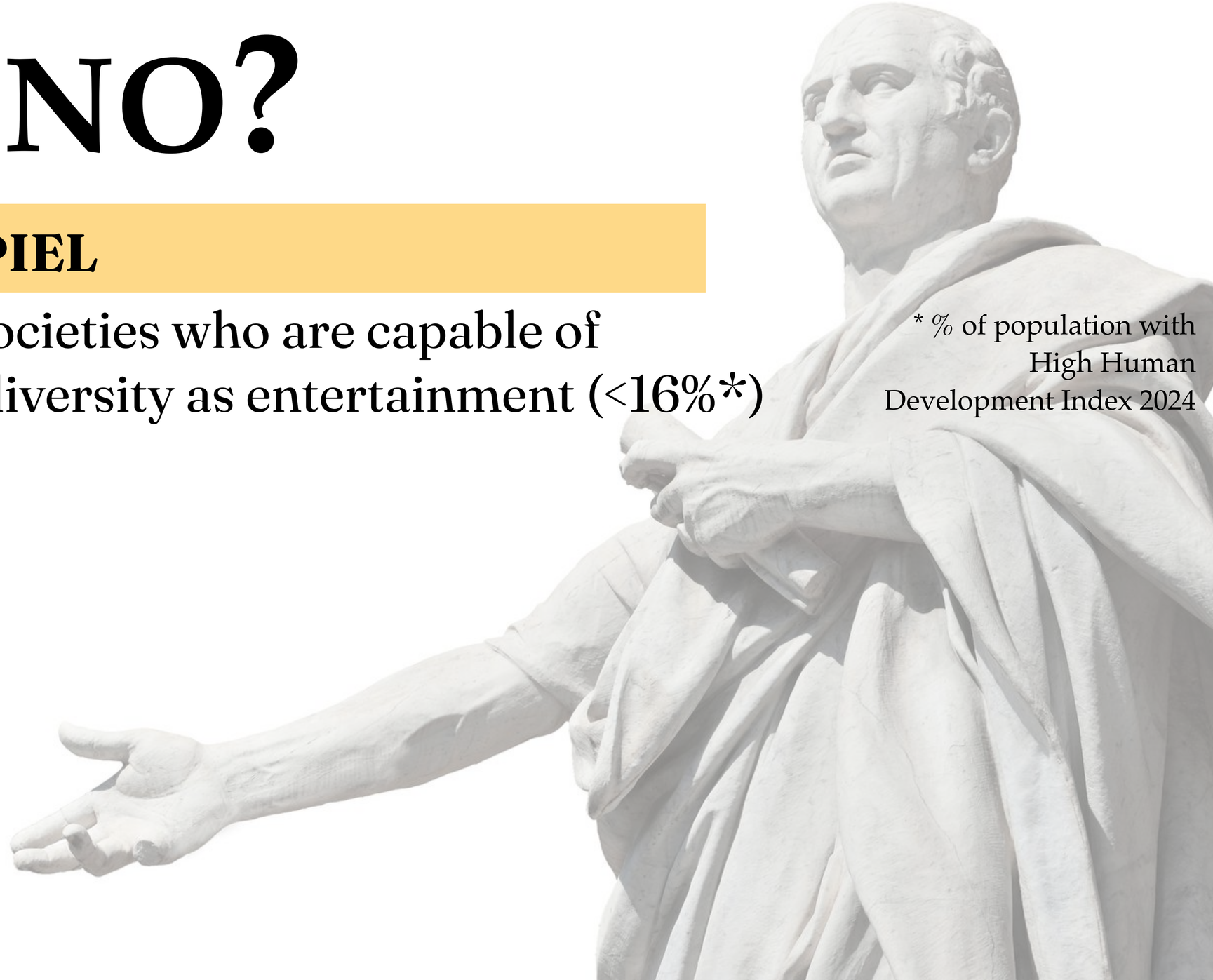
# Cui Bono?



# Cui Bono?

## The romantic spiel

Members of affluent societies who are capable of consuming language diversity as entertainment (<16%*)



* % of population with High Human Development Index 2024

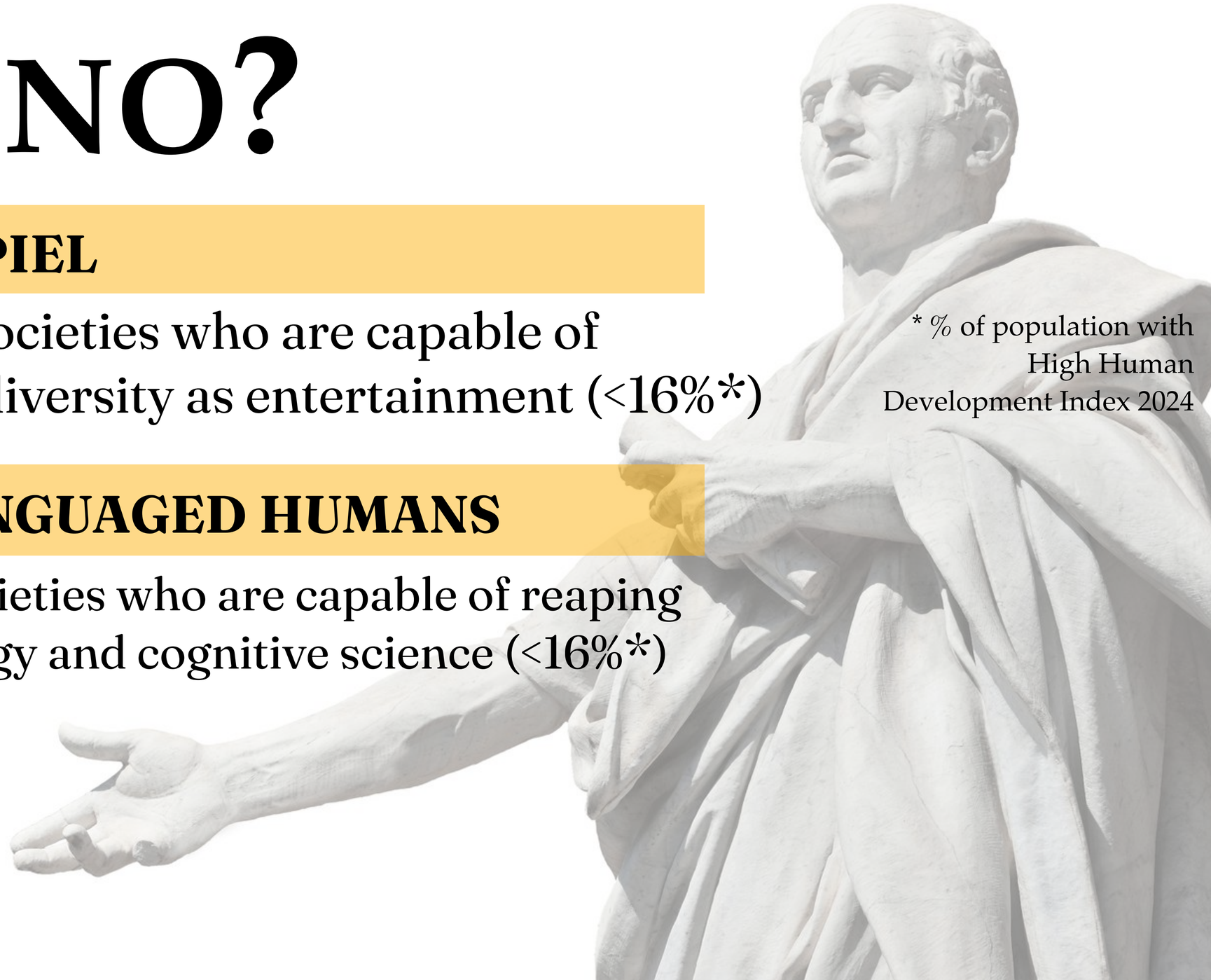
# Cui Bono?

## The romantic spiel

Members of affluent societies who are capable of consuming language diversity as entertainment (<16%*)

## The study of languaged humans

Members of affluent societies who are capable of reaping the benefits of psychology and cognitive science (<16%*)



* % of population with High Human Development Index 2024

# Cui Bono?

## The romantic spiel

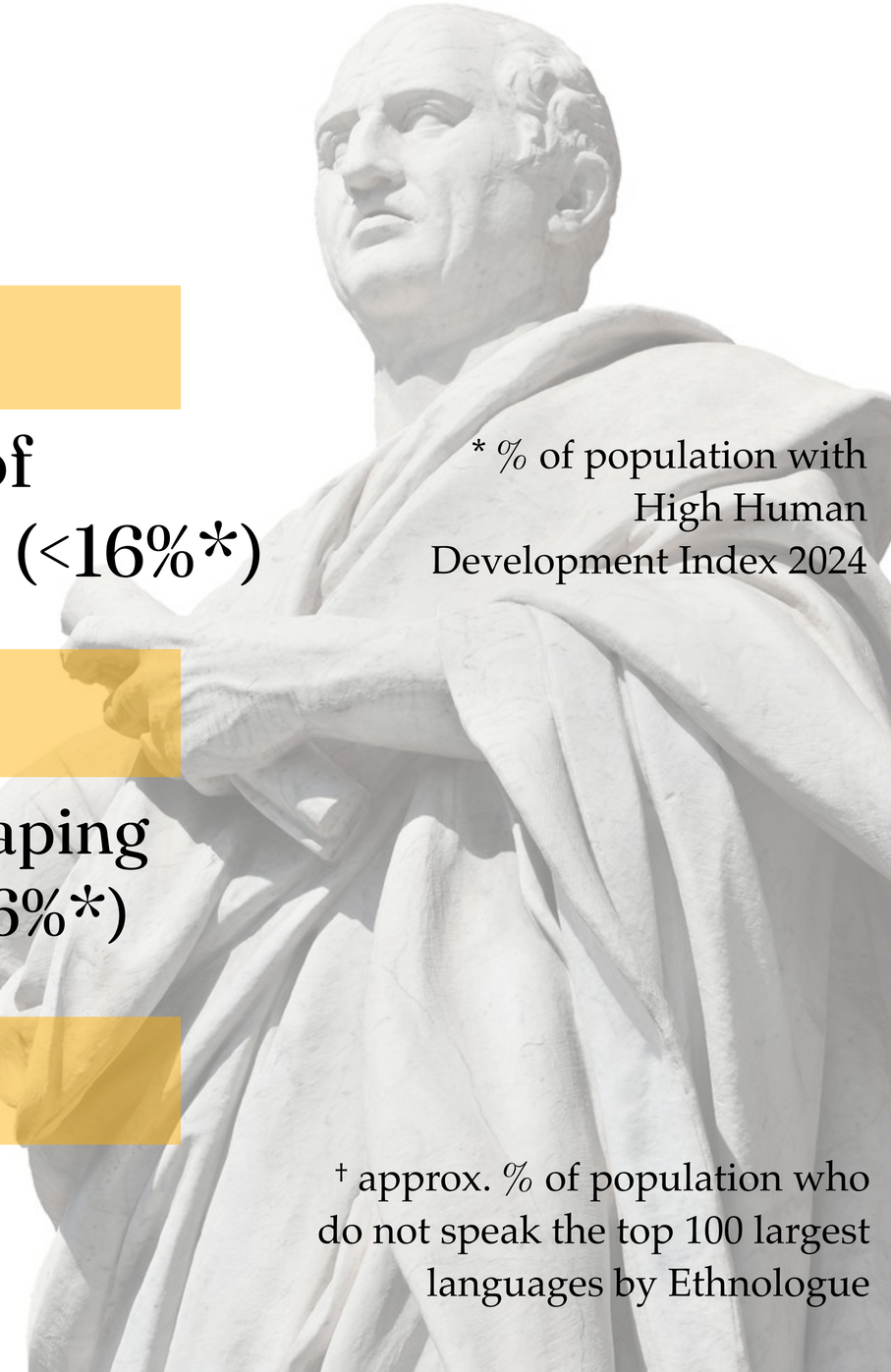
Members of affluent societies who are capable of consuming language diversity as entertainment (<16%*)

## The study of languaged humans

Members of affluent societies who are capable of reaping the benefits of psychology and cognitive science (<16%*)

## The ethical imperative

Users of endangered languages (<20%†)



* % of population with High Human Development Index 2024

† approx. % of population who do not speak the top 100 largest languages by Ethnologue

MY ARGUMENT HERE

# THE BEHEMOTH LANGUAGE BIAS

English and a few other large and influential languages ("behemoth languages") serve, in technology, medicine, education, law, etc. as prototypes models for language in general, *leading to a measurably poorer performance for users of other languages*

MY ARGUMENT HERE

# THE BEHEMOTH LANGUAGE BIAS

English and a few other large and influential languages ("behemoth languages") serve, in technology, medicine, education, law, etc. as prototypes models for language in general, *leading to a measurably poorer performance for users of other languages*

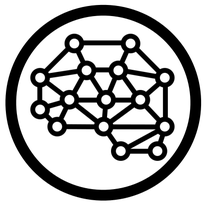
**This bias affects over 60% of the world's population, involving large vital languages with hundreds of millions of users**

Useful speech cues for detecting Parkinson's in English
lead to misdiagnose in vowel-rich and tonal languages

Pinto et al. (2017)

Direct translation of questions in the PISA evaluations is fair in most large languages
but in many minority languages leads to more complex sentences and lower scores

El Masri, Baird & Graesser (2016)



Language models are fundamental for human-like cognition in AI
but they work better in morphology-poor and fixed-word order languages like English

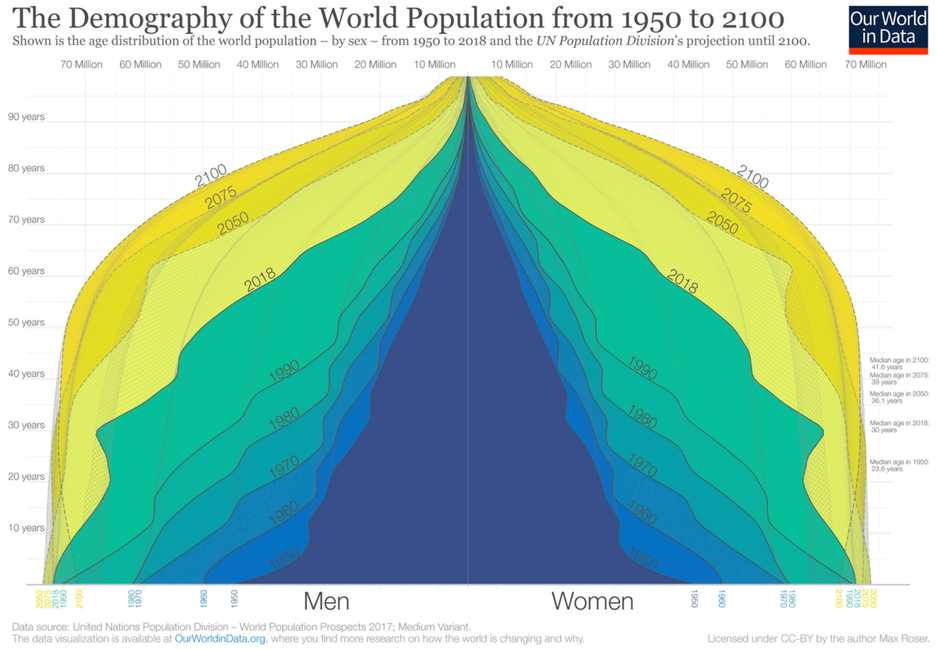
Schwartz et al. (2020), Park et al. (2020), Gerz et al. (2020)

Public health communication during COVID was clear in English
but the same guidelines applied to other languages yield difficult to understand advice

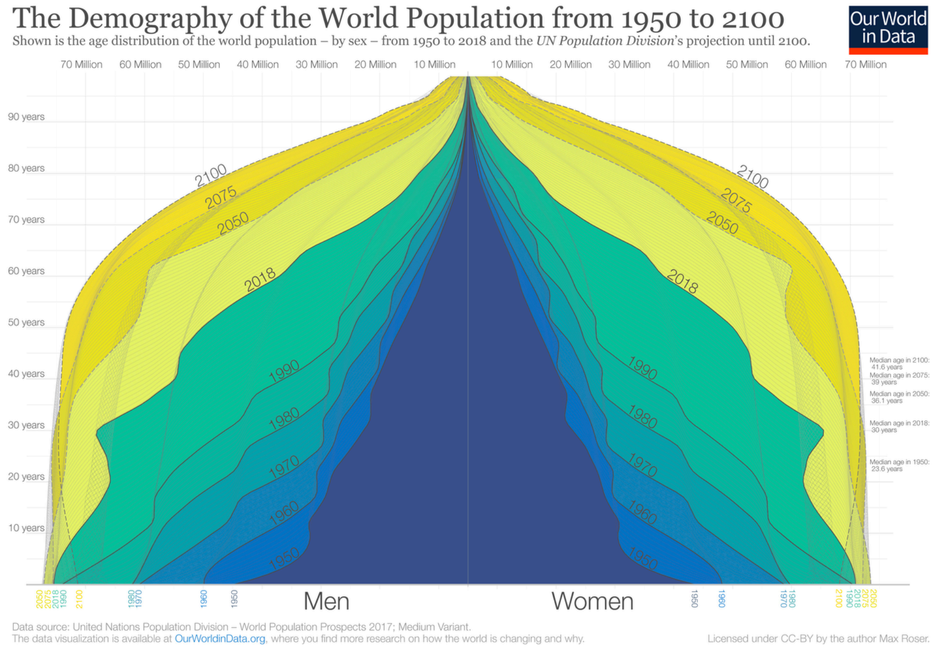
Blasi et al. (2021)

If a resource is language-based,
then very likely it's not language-independent

# The world's population is aging, which increases the burden of neurodegenerative diseases (Alzheimer's, Parkinson's, dementia)



# The world's population is aging, which increases the burden of neurodegenerative diseases (Alzheimer's, Parkinson's, dementia)

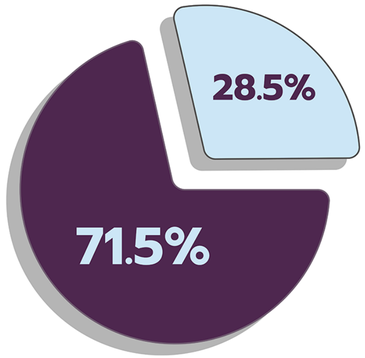



## RESEARCH

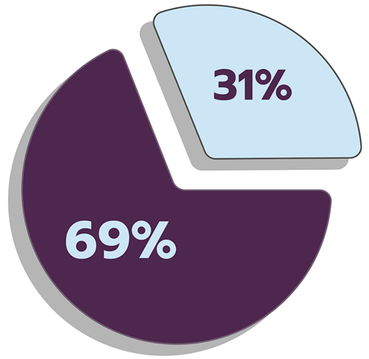
IN ENGLISH

IN OTHER LANGUAGES

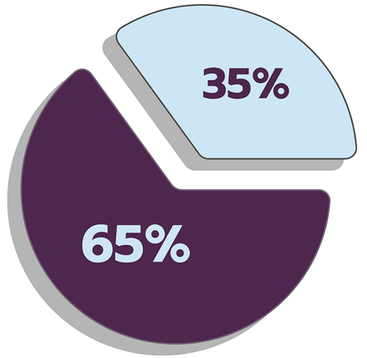
**Speech and language research on PPA (logopenic variant)**



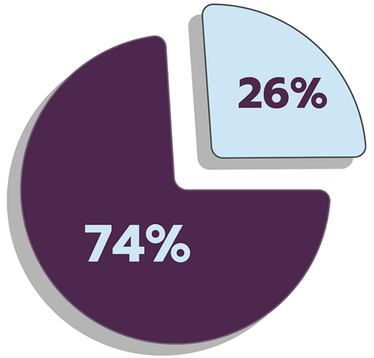

**Word retrieval research on AD**




**Verbal production research on PD**

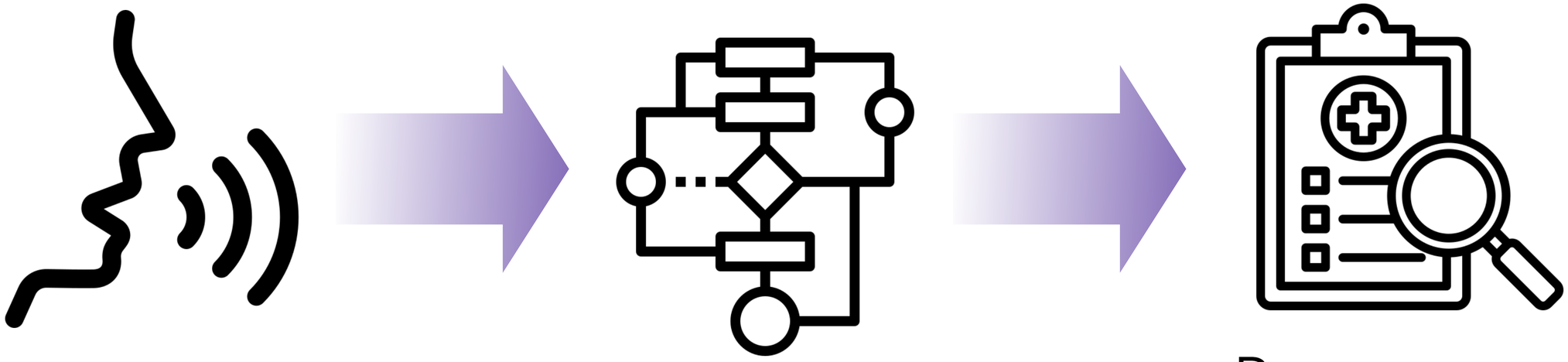



**Speech and language research on bvFTD**



There is a race for developing fast, cheap, and reliable markers for detecting these diseases early

**Linguistic biomarkers** are extremely promising and developing at a breakneck speed



LANGUAGE PRODUCTION

AUTOMATIC NLP ASSESSMENT

PROBABILISTIC DIAGNOSIS

**In English**, individuals afflicted by Alzheimer's disease (AD) use more pronouns in contrast to healthy individuals

Healthy individual:
the dog is chasing a cat

AD patient:
he is chasing it

**In English**, individuals afflicted by Alzheimer's disease (AD) use more pronouns in contrast to healthy individuals

Healthy individual:
the dog is chasing a cat

AD patient:
he is chasing it

...maybe because AD patients have problems with *perspective taking*
(Bittner et al. 2022)

**In English**, individuals afflicted by Alzheimer's disease (AD) use more pronouns in contrast to healthy individuals

Healthy individual:
the dog is chasing a cat

AD patient:
he is chasing it

...maybe because AD patients have problems with *perspective taking*
(Bittner et al. 2022)

**Convenient basis for a <u>universal biomarker</u>: pronoun frequency**

However, the **exact opposite pattern** is found in speakers of Bengali (বাংলা): they use less, no more, pronouns

However, the **exact opposite pattern** is found in speakers of Bengali (বাংলা): they use less, no more, pronouns

**English personal pronouns (NOM)**

| | | | Nominative |
|---|---|---|---|
| First-person | Singular | | I |
| | Plural | | we |
| Second-person | Singular | Standard | you |
| | | Poetic/dialectal | thou |
| | Plural | | you |
| Third-person | Singular | Masculine | he |
| | | Feminine | she |
| | | Neuter | it |
| | | Epicene | they |
| | Plural | | they |

# However, the **exact opposite pattern** is found in speakers of Bengali (বাংলা): they use less, no more, pronouns

## English personal pronouns (NOM)

| | | | Nominative |
|---|---|---|---|
| First-person | Singular | | I |
| | Plural | | we |
| Second-person | Singular | Standard | you |
| | | Poetic/dialectal | thou |
| | Plural | | you |
| Third-person | Singular | Masculine | he |
| | | Feminine | she |
| | | Neuter | it |
| | | Epicene | they |
| | Plural | | they |

## Bengali personal pronouns (NOM)

| Subject | Proximity | Honor | Singular | Plural |
|---|---|---|---|---|
| 1 | | | আমি (*ami*, I) | আমরা (*amra*, we) |
| 2 | | VF | তুই (*tui*, you) | তোরা (*tora*, you) |
| | | F | তুমি (*tumi*, you) | তোমরা (*tomra*, you) |
| | | P | আপনি (*apni*, you) | আপনারা (*apnara*, you) |
| 3 | H | F | এ (*e*, he/she) | এরা (*era*, they) |
| | | P | ইনি (*ini*, he/she) | এঁরা (*ẽra*, they) |
| | | I | এটি/এটা (*eṭi/eṭa*, it) | এগুলো (*egulo*, these) |
| | T | F | ও (*o*, he/she) | ওরা (*ora*, they) |
| | | P | উনি (*uni*, he/she) | ওঁরা (*õra*, they) |
| | | I | ওটি/ওটা (*oṭi/oṭa*, it) | ওগুলো (*ogulo*, those) |
| | E | F | সে (*she*, he/she) | তারা (*tara*, they) |
| | | P | তিনি (*tini*, he/she) | তাঁরা (*tãra*, they) |
| | | I | সেটি/সেটা (*sheṭi/sheṭa*, it) | সেগুলো (*shegulo*, those) |

However, the **exact opposite pattern** is found in speakers of Bengali (বাংলা): they use less, no more, pronouns

## English personal pronouns (NOM)

| | | | Nominative |
|---|---|---|---|
| First-person | Singular | | I |
| | Plural | | we |
| Second-person | Singular | Standard | you |
| | | Poetic/dialectal | thou |
| | Plural | | you |
| Third-person | Singular | Masculine | he |
| | | Feminine | she |
| | | Neuter | it |
| | | Epicene | they |
| | Plural | | they |

## Bengali personal pronouns (NOM)

| Subject | Proximity | Honor | Singular | Plural |
|---|---|---|---|---|
| 1 | | | আমি (*ami*, I) | আমরা (*amra*, we) |
| 2 | | VF | তুই (*tui*, you) | তোরা (*tora*, you) |
| | | F | তুমি (*tumi*, you) | তোমরা (*tomra*, you) |
| | | P | আপনি (*apni*, you) | আপনারা (*apnara*, you) |
| 3 | H | F | এ (*e*, he/she) | এরা (*era*, they) |
| | | P | ইনি (*ini*, he/she) | এঁরা (*ẽra*, they) |
| | | I | এটি/এটা (*eṭi/eṭa*, it) | এগুলো (*egulo*, these) |
| | T | F | ও (*o*, he/she) | ওরা (*ora*, they) |
| | | P | উনি (*uni*, he/she) | ওঁরা (*õra*, they) |
| | | I | ওটি/ওটা (*oṭi/oṭa*, it) | ওগুলো (*ogulo*, those) |
| | E | F | সে (*she*, he/she) | তারা (*tara*, they) |
| | | P | তিনি (*tini*, he/she) | তাঁরা (*tãra*, they) |
| | | I | সেটি/সেটা (*sheṭi/sheṭa*, it) | সেগুলো (*shegulo*, those) |

This suggests it might not be a problem of perspective taking but just AD patients using “easy” pieces of language for communication

Even with the limited number of studies on non-behemoth languages, this issue is **widespread** in the diagnose of neurodegenerative diseases

**Linguistic biomarkers rarely generalise across languages**

| | Examples of cross-linguistic differences | | |
|---|---|---|---|
| DISORDER | LANGUAGES | STRUCTURAL CONTRAST | DISTINCT MARKER |
|  nfvPPA | English | Greater phonetic and lesser morphosyntactic complexity | Phonetic distortions as most salient symptom |
| | Italian | Lesser phonetic and greater morphosyntactic complexity | Distinct syntactic alterations |
|  svPPA | English | Alphabetic script (letters represent phonemes) | High prevalence of surface dysgraphia |
| | Chinese | Logographic script (logograms convey semantic or phonological information) | Low prevalence of surface dysgraphia |
|   lvPPA | English | Less diverse morphosyntactic patterns | Frequent sentence repetition deficits |
| | German | More diverse morphosyntactic patterns | Infrequent sentence repetition deficits |
|  AD | English | Simpler pronominal system | Overuse of pronouns |
| | Bengali | More complex pronominal system | Underuse of pronouns |
|  PD | Spanish | Verb-framed language with rich verb vocabulary | Selective action-verb deficits |
| | Dutch | Satellite-framed language with fewer verbs | Non-selective action-verb deficits |

*As language diversity researchers, what can we do to tackle the Behemoth Language Bias?*

# I. Get involved with the relevant parties

“And now, a bit of feedback from our silent partner.”

Credit: Adobe Stock

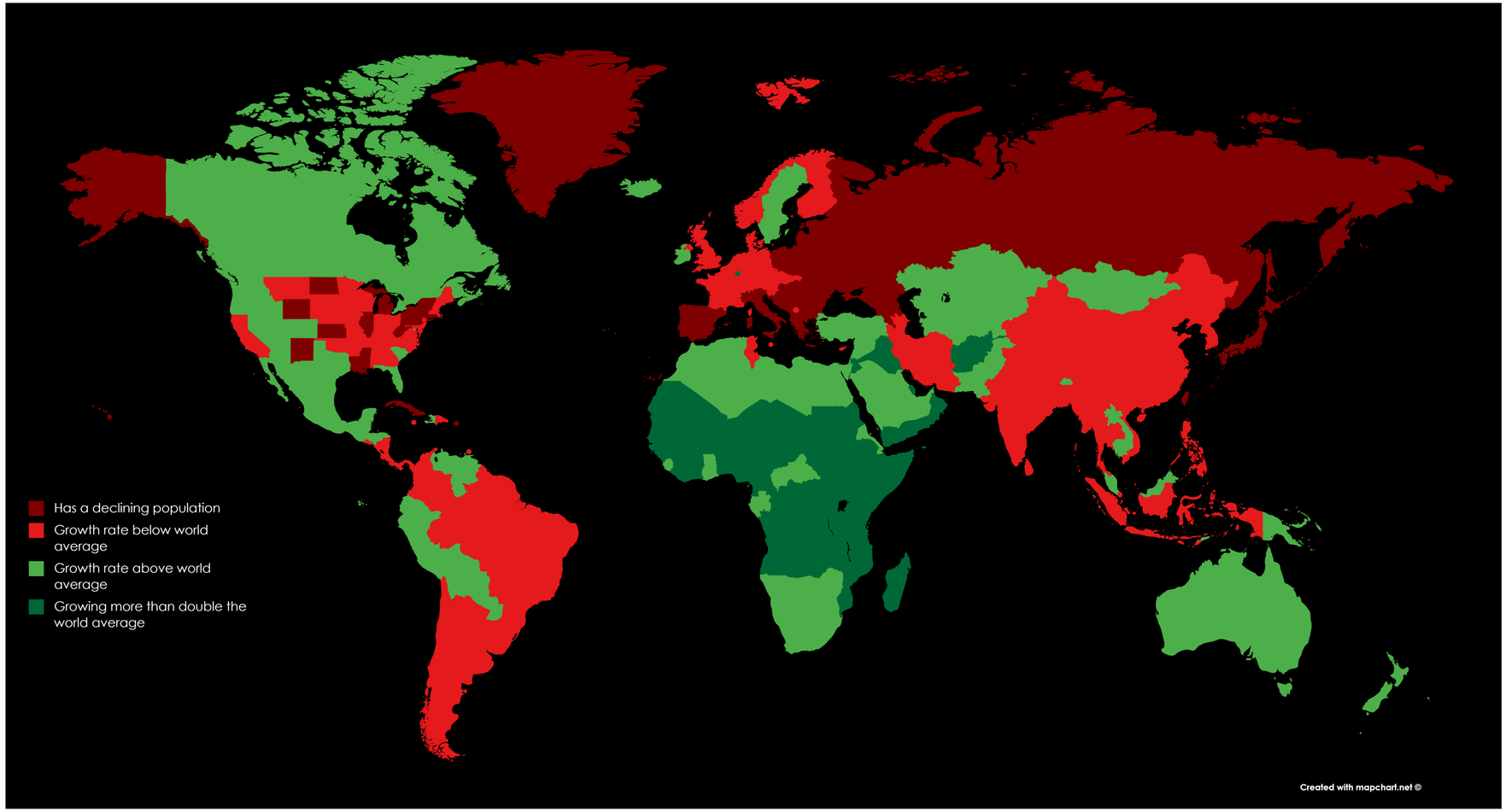
# I. Get involved with the relevant parties

“And now, a bit of feedback from our silent partner.”

Credit: Adobe Stock


Shameless staged picture of me at the UNESCO headquarters in Paris.
Disclosure: I do not represent my country in any official capacity.

# II. Be *demographically aware* of the future




**Credit: r/MapPorn user Sppoderman89 based on US Census Bureau and the World Bank**

# III. Enhance academic credit

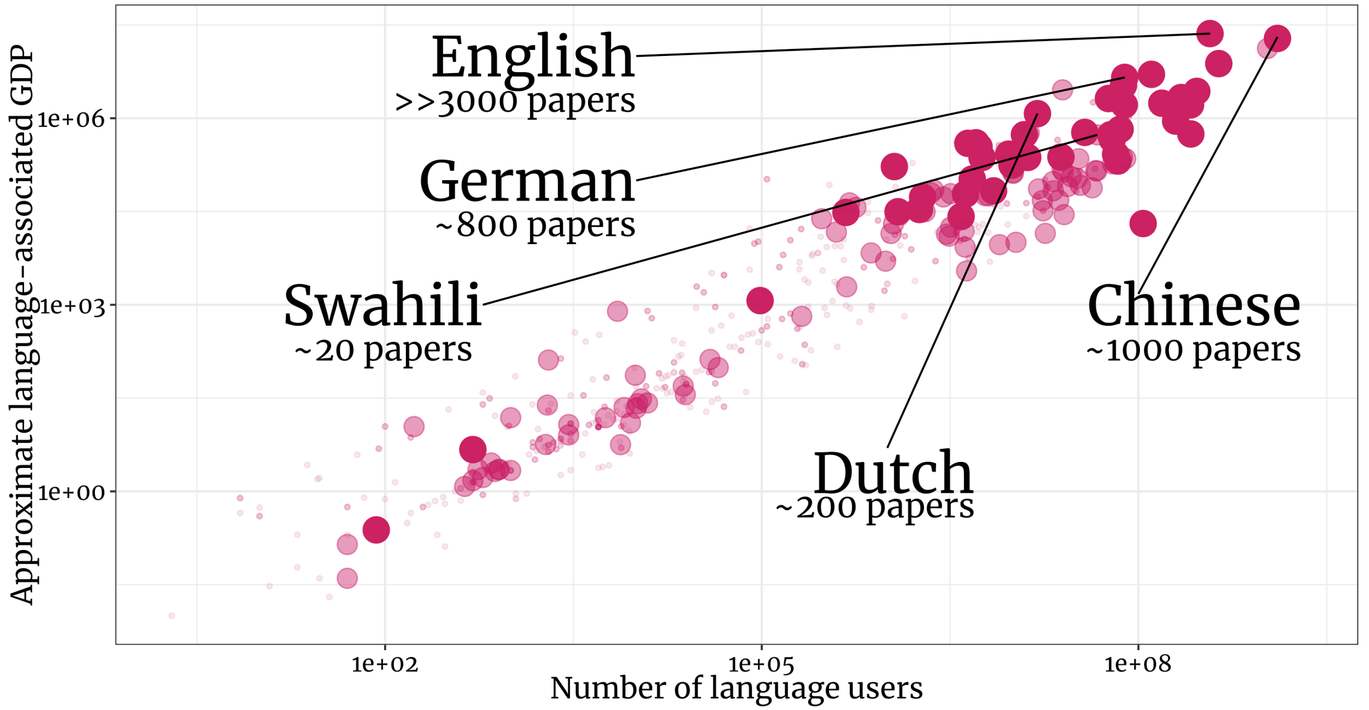
**Systematic Inequalities in Language Technology Performance across the World's Languages**


**Damián Blasi**
Harvard University
dblasi@fas.harvard.edu

**Antonios Anastasopoulos**
George Mason University
antonis@gmu.edu

**Graham Neubig**
Carnegie Mellon University
gneubig@cs.cmu.edu

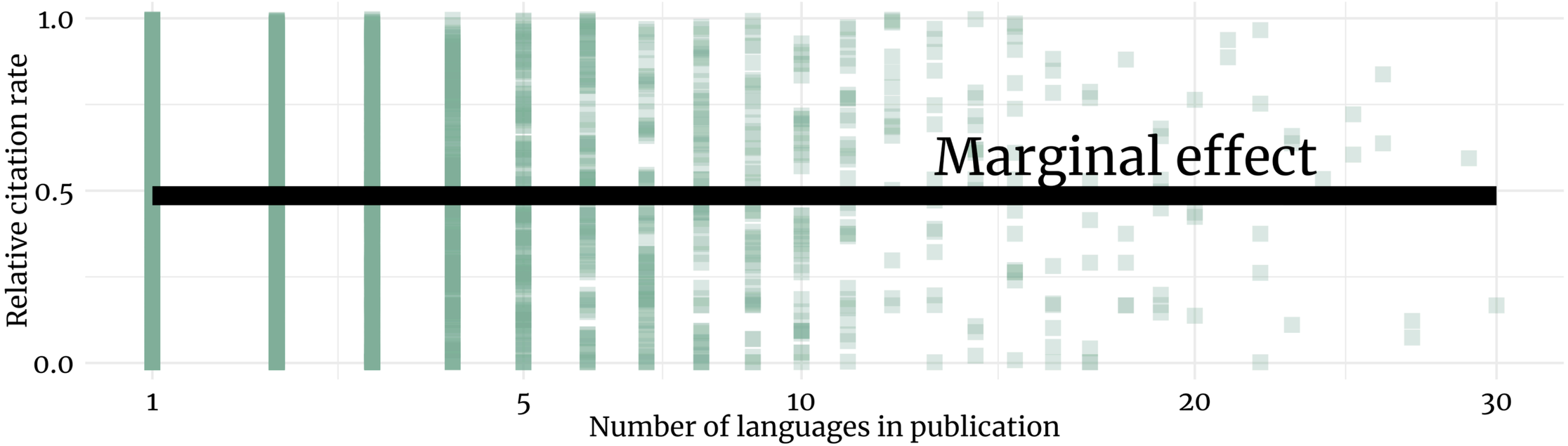

**Circle size reflects number of publications in the language, proxying magnitude of R&D in language technologies**




Contemporary language technologies, are (1) heavily centered on English and other Behemoth Languages, and (2) explained by economic centrality rather than demographic need

# III. Enhance academic credit

## The number of languages in a publication is not associated with its relative citation rate over time




→ No marginal incentive for researchers to further the language diversity of their data and models

# Fin.


ICREA


upf.
Universitat
Pompeu Fabra
Barcelona


CBC
Center for Brain and Cognition


UNESCO


The
Branco Weiss
Fellowship
Society in Science