# *Linguistic discrimination in scientometrics*

**Luc Boruta, Thunken Inc.**
luc@thunken.com — @thunkenizer
RefResh 2018, Cologne, 2018/12/05


THUNKEN

# Conflict of interest

http://bit.ly/2rfgwAm

# Linguistic diversity

**>7k languages** are spoken today.

23 languages account for >50% of the world population.

L1-English accounts for <5% of the world population.

https://www.ethnologue.com/guides/how-many-languages



# Linguistic diversity on the web

UNESCO report by Pimienta et al. (2009):

| | EN | SP | FR | IT | PO | RO | GE | CAT | SUM[11] | REST[12] |
|---|---|---|---|---|---|---|---|---|---|---|
| **09/98** | 75.0% | 2.53% | 2.81% | 1.50% | 0.82% | 0.15% | 3.75% | | 11.56% | 13.44% |
| **11/07** | *(45.0%)* | 3.80% | 4.41% | 2.66% | 1.39% | 0.28% | 5.90% | 0.14% | 18.46% | |

Wikipedia: **303 languages**, **49M articles**

English Wikipedia: 5.7M articles (<12%)

http://unesdoc.unesco.org/images/0018/001870/187016e.pdf
https://meta.wikimedia.org/wiki/List_of_Wikipedias



# Linguistic diversity in citation indexes

## Scopus (Elsevier)

*“The main language of the international scientific community [...] is considered to be English. Therefore, all content of the records that are available in Scopus [...] need to be in English.”*

## Web of Science (Clarivate Analytics)

*“English is the universal language of science [and] it is clear that the journals most important to the international research community are publishing [...] in English.”*



# Linguistic diversity in altmetrics BC

= before Cobaltmetrics

Not all indicators are sensitive to linguistic diversity, but...

Citation tracking in Wikipedia:

– Altmetric: 3 languages (en, fi, sv)

– PlumX Metrics: 3 languages (en, es, pt)

– ALM: 25 most popular languages



# Are your metrics alt- enough?

**NO.**



# Why should we care?

Metrics are a sampling game. Selection biases are an issue.

Materials in LOTE are not aligned translations.

**Imbalanced datasets reinforce discrimination.**

**Diversity** and **fairness** must be integral parts of how we drive our projects.



# Latent discrimination, real consequences

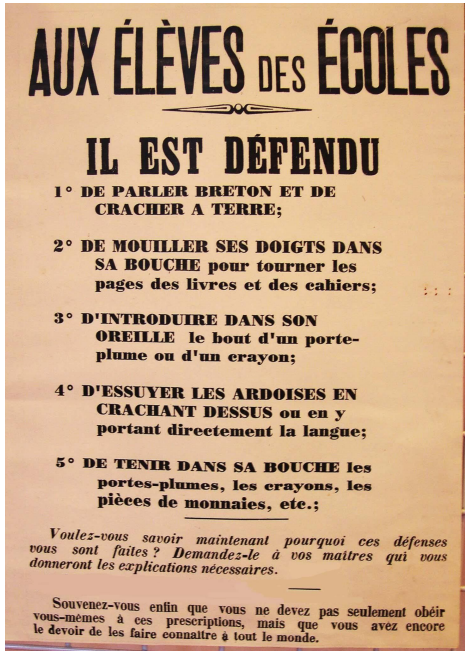
**Systematic bias** in stats on cross-linguistic citation practices.

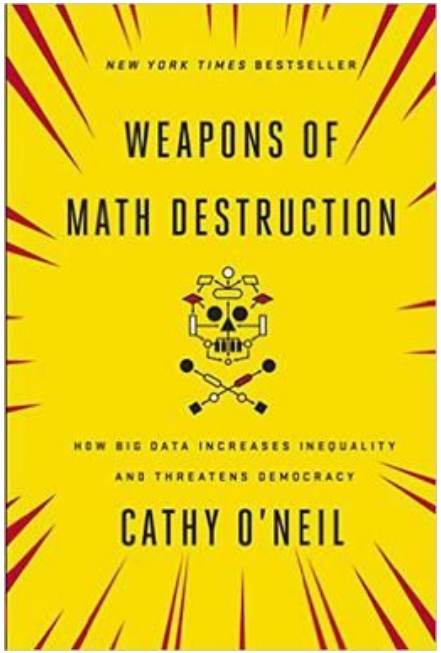
**Systematic exclusion** of some contributors from metrics-based evaluations.

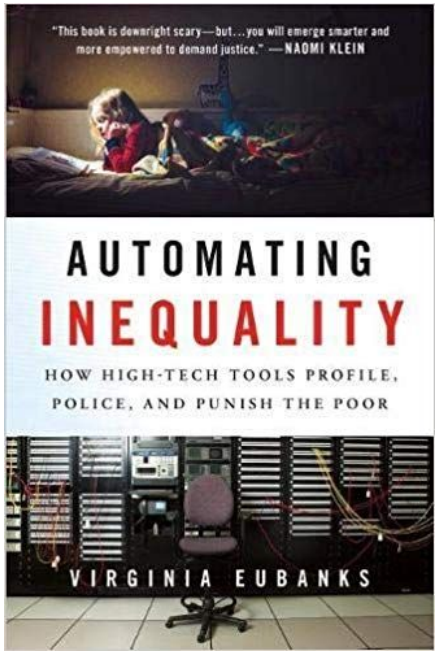
Reinforces discrimination in other parts of the community, cf. Nylenna et al. (1994) and Lazarev & Nazarovets (2018).



# From la vergonha to math destruction



# Toward fair and inclusive scientometrics

The scientific community is global and diverse.

Our corpora need to be global and diverse:
all languages, all document types, all identifiers, etc.

**It is not up to metrics providers to decide what is citable.**



# Cobaltmetrics: altmetrics for all

Altmetric, ALM, CED, Plum, et al. are great projects.

But **diversity is good**, and we think we can do even better.

With Cobaltmetrics, we include **low-frequency signals**, and we root for the underdogs.



# Natural language processing to the rescue!

Anglo-centrism is prejudicial to science.

Algorithmic complexity cannot be used as an excuse.

Citation data is mostly **machine-readable**,
and/or can be described using **local grammars**.



# Linguistic diversity in altmetrics: Wikimedia

Altmetric: 3 languages (en, fi, sv)

PlumX Metrics: 3 languages (en, es, pt)

ALM: 25 most popular languages

**Cobaltmetrics: 180+ languages!**

https://medium.com/thunken/altmetrics-for-all-4b9039249850



W **Szemantikus web** (November 20, 2018), from hu.wikipedia.org.
Matching URIs: rfc:3986 http://tools.ietf.org/html/rfc3986

W **URL** (November 10, 2018), from hu.wikipedia.org.
Matching URIs: rfc:3986

W **Kasutaja:Martin457345/Päringusõne** (November 7, 2018), from et.wikipedia.org.
Matching URIs: rfc:3986

W **شوێندۆزی سەرچاوەی جیهانی** (October 16, 2018), from ckb.wikipedia.org.
Matching URIs: rfc:3986

W **Lokator Sumber Seragam** (October 13, 2018), from id.wikipedia.org.
Matching URIs: rfc:3986 http://tools.ietf.org/html/rfc3986

W **Web resource** (October 4, 2018), from bg.wikipedia.org.
Matching URIs: https://tools.ietf.org/html/rfc3986 rfc:3986

W **ইউনিফর্ম রিসোর্স লোকেটর** (September 14, 2018), from bn.wikipedia.org.
Matching URIs: rfc:3986

# Where do we go from here?

New and upcoming features:

– New data sources, new endpoints, more filters

– 2019: web-scale altmetrics with >2.5B webpages

– 2019: cross-linguistic benchmark using the Érudit corpus

Push for more linguistic diversity in other projects!

Crossref Event Data? OpenCitations?




one size
does not
fit all