Altmetric

# Evaluating the diversity of scientific discourse on twenty-one multilingual Wikipedias using citation analysis

Mike Taylor, Head of Data Insights

Carlos Areia, Data Scientist

Roisi Proven, VP Product, Altmetric

Dimensions

Part of DIGITAL science

# Your presenter today

- Mike Taylor, Head of Data Insights at Digital Science
- Mostly working on Altmetric and Dimensions
- At Digital Science for over seven years
- Elsevier for 20 years (Research and Development)
- Allegedly working towards a PhD at University of Wolverhampton

# What is this research?

- The study of citations from ‘grey literature’ is fairly unrepresented in academic literature (>30 for “Wikipedia citations”)

- Wikipedia is one of these most important sources of information on the web, and one of the most visited websites (but it’s importance is far more than this: Wikipedia data is used *everywhere* – including in large language models

- There’s an assumption (sometimes expressed, sometimes not…) that the English Wikipedia is so much larger and active than the others, that it’s the only one that ‘counts’

# The inspiration!

- Over the years, I'd had two conversations with people about the role of Wikipedia, and – in one, been taken to task by a Russian art historian

- The other, a Mexican health researcher was complaining about the hegemony of English / American research / representation of knowledge

# Why now?

- A couple of years ago, we at Altmetrics decided to go an extra mile to improve diversity
- We also needed to revamp the Wikipedia citation collector (which was shoddy)
- The new pipeline is much better at handling multiple languages, without risking ire at Wikipedia…

# New pipeline is *much* more effective

- So, we added a lot of new languages! Some we did not… over 30 so far, and we're adding in a set of African languages this summer (these are small, but this is part of an initiative that we're supporting)

Earlier in the week, a Reddit user named Ultach detailed a discovery they made about the Scots language version of Wikipedia (via *The Guardian*). Alongside Gaelic, Scots is one of the indigenous languages of Scotland. The thousands of Wikipedia entries written in it make up one of the largest collections of the Scots language you can access online for free. The problem is an American teenager from North Carolina – who can't speak the language – wrote 49 percent of all the entries.

Before Ultach discovered the teen, who had gone by the username AmaryllisGardner, they had been prolific. By 2018, the 19-year-old had written more than 20,000 entries and committed approximately 200,000 edits. They were able to write so much by starting at the age of 12. The majority of entries AmaryllisGardner penned feature the occasional Scots word, often misspelled, and they include no Scots grammatical constructions. It seems AmaryllisGardner used an online translator to graft Scots words onto sentences written in American English.

Part of the reason no one noticed or stopped the vandalism is that there wasn't much interest in the Scots Wikipedia before this week. "Nobody cared about maintaining [the Scots Wikipedia]," said Wikipedian MJL, one of the website's administrators. "Someone stepped up because no one else did. That person was never given any guidance. Articles ended up being very poorly mistranslated."

# What research questions were we addressing?

- First of all, we wanted to quantify the contribution that non-English languages were making to the whole body of “Wikipedia citations”
- Second, we wanted to understand whether / to what extent different languages were representing research (through this curious lens of citation analysis…) – which I’m calling “uniqueness of voice”
- Third, we wanted to look to see if there was any evidence that there were subject discipline differences
- (There were other RQs, I’m not very disciplined :D – but Carlos is )

# The methodology…

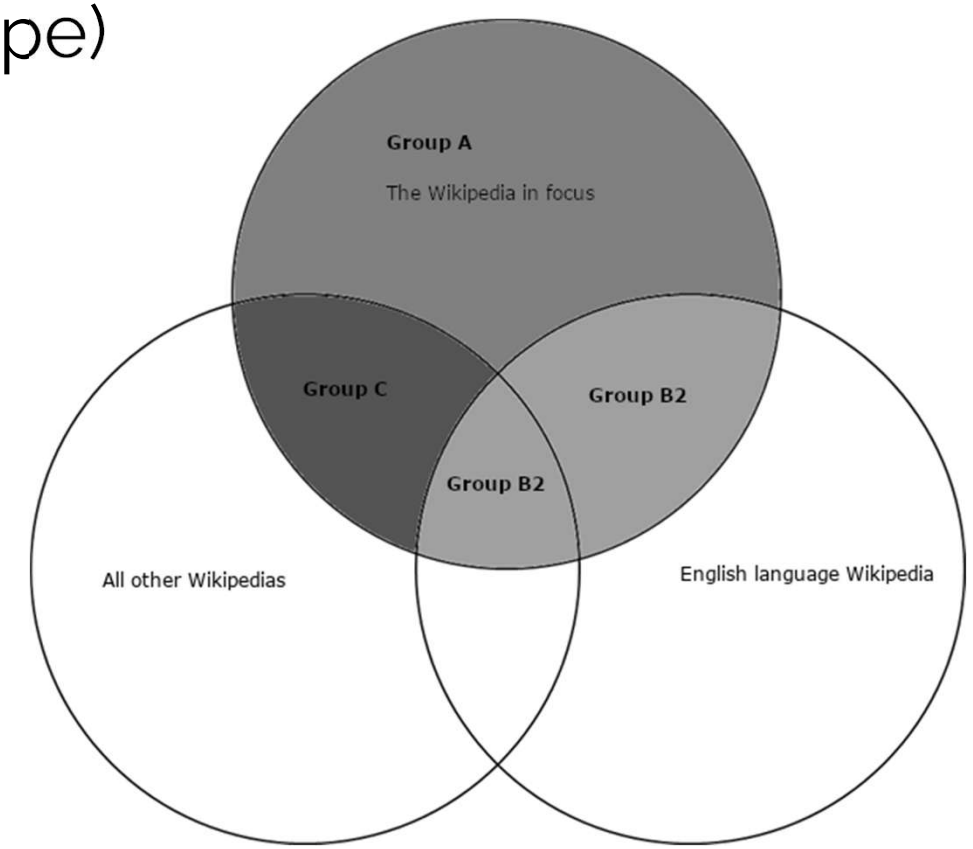
- We identified twenty languages (other than English), representing many parts of the world, but importantly are large, and have over 1,000 editors
  - French, Spanish, Catalan, Portuguese, Italian, German, Greek, Turkish
  - Chinese/Mandarin, Persian/Farzi, Indonesian, Arabic, Vietnamese, Korean, Japanese
  - Russian, Ukrainian, Polish, Czech, Serbian

*Excluded:*
Bengali
Dutch
Finnish
Hebrew
Hindi
Hungarian
Norwegian
Romanian
Swedish
Thai

All citations identified by Altmetrics were mapped to the relevant cited publications (articles, books, etc) in Dimensions, and mapped to one of six subject areas.

# The clever bit (we hope)

- For each language, we determined the number of articles that are uniquely cited by this Wikipedia (and ONLY this Wikipedia)
- And the number of articles cited by this Wikipedia and ONLY co-cited with English
- Finally, the number of articles cited by this Wikipedia and by at least one other non-English Wikipedia (whether or not English was involved)




Altmetric

Dimensions

# Some results...

Number of research outputs cited
by our 21 Wikipedias

# 3,485,474

Altmetric

Dimensions

Number of research outputs cited by EN Wikipedia

# 2,035,466 (58.4%)

The English Language Wikipedia cites the largest number of outputs

Altmetric

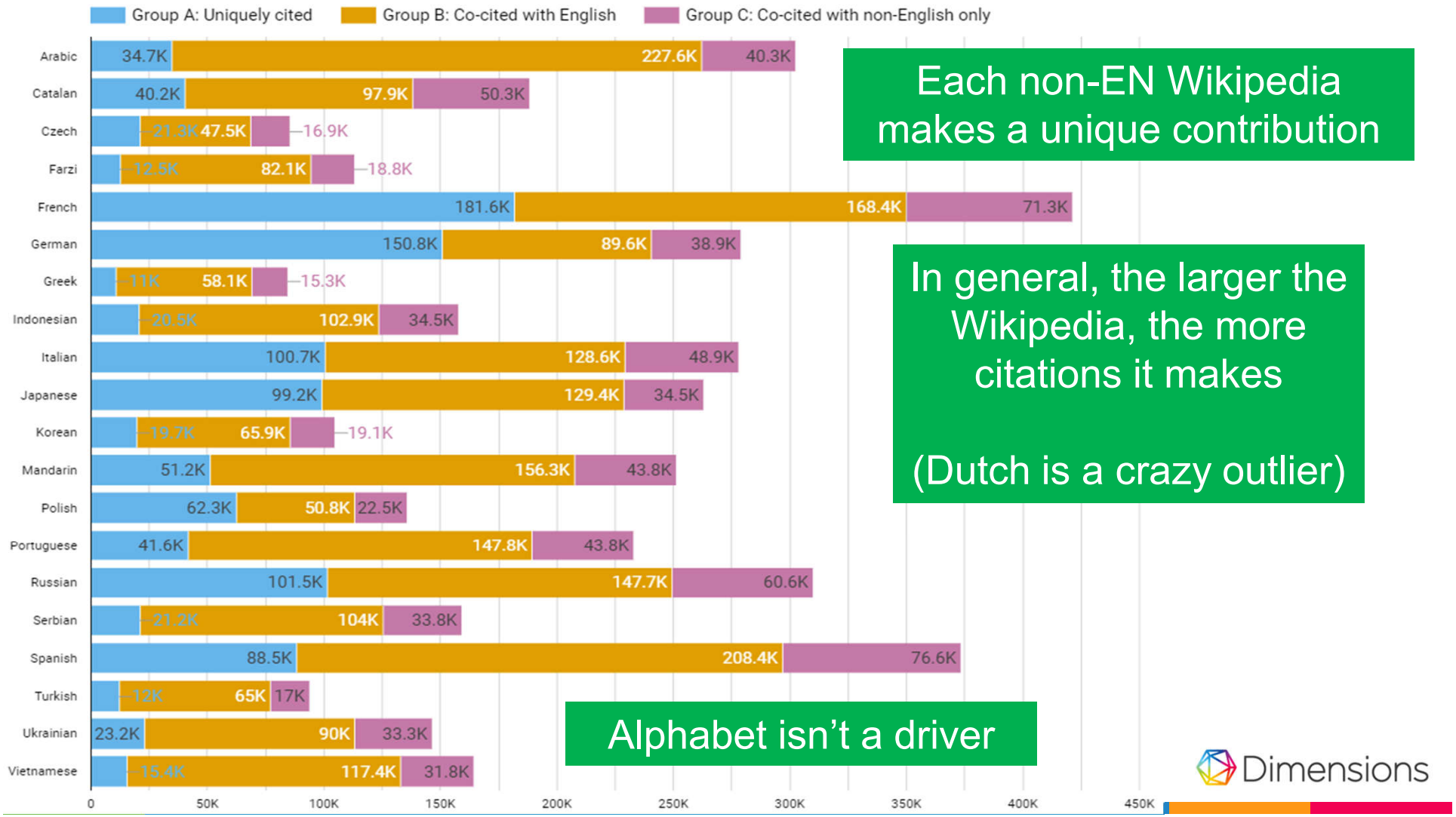
Dimensions

| Number of articles cited by EN Wikipedia *and not cited by others* | Number of articles cited by non-EN Wikipedias *and not cited by EN* |
| --- | --- |
| 1,140,160 | 1,457,138 |
| 32.7% | 41.8% |

The unique contribution of non-English Wikipedias is larger than the English

Altmetric

Dimensions



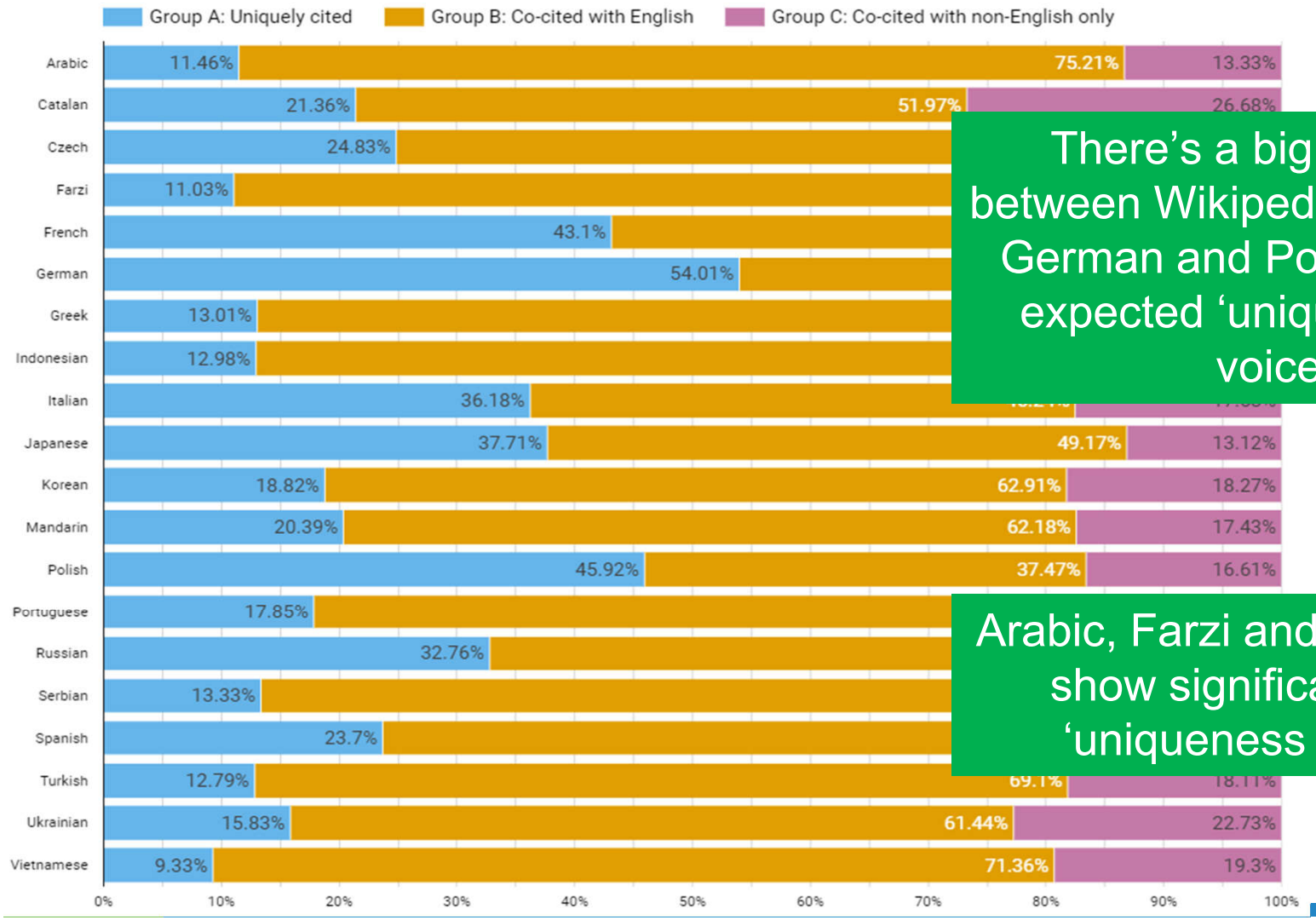

Each non-EN Wikipedia makes a unique contribution

In general, the larger the Wikipedia, the more citations it makes

(Dutch is a crazy outlier)

Alphabet isn't a driver

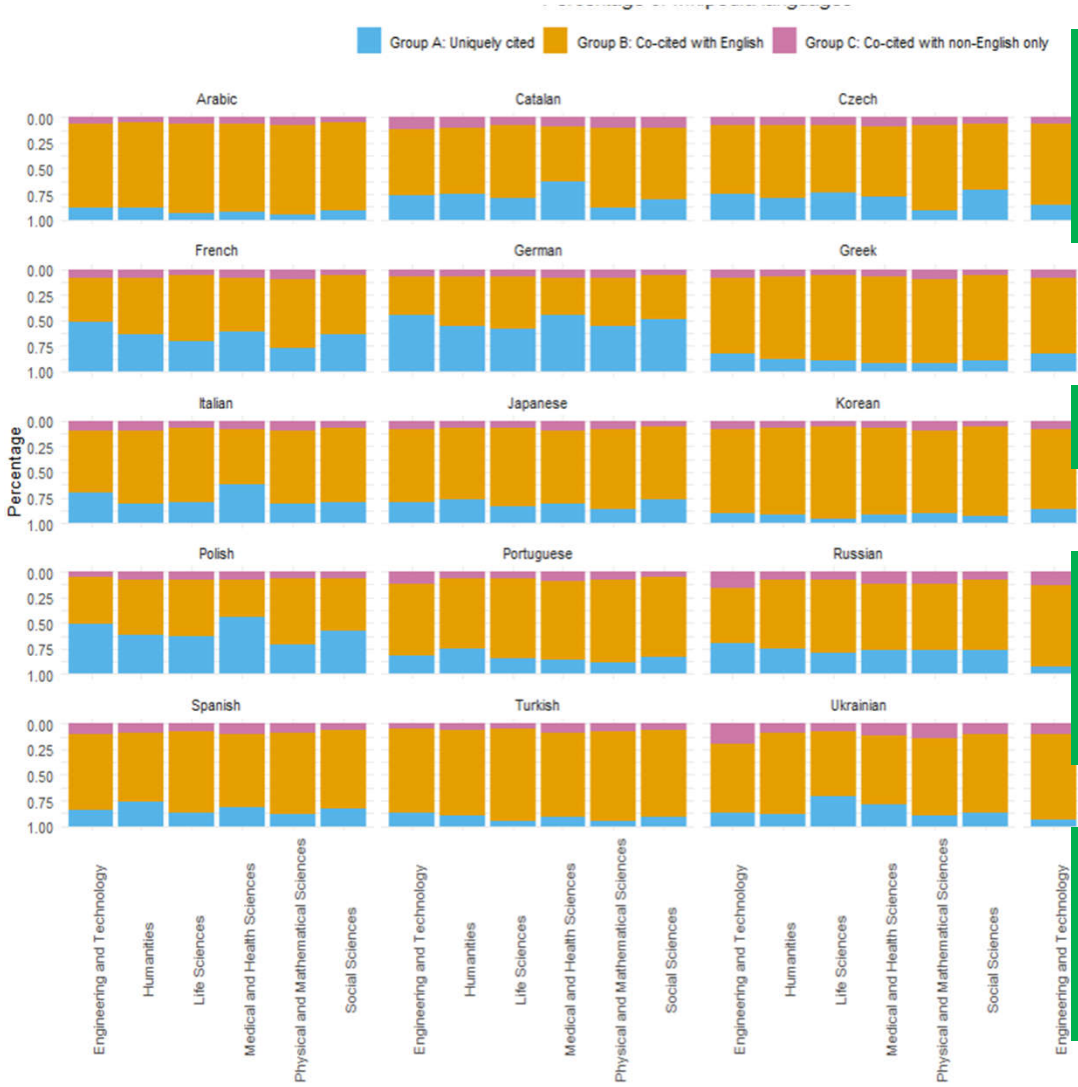
Dimensions




There's a big variation between Wikipedias – French, German and Polish exceed expected 'uniqueness-of-voice'

Arabic, Farzi and Vietnamese show significantly less 'uniqueness of voice'

Dimensions




Generally, we can say that *some* Wikipedia have pronounced uniqueness in MHS

E&T is probably the next

The less unique outliers (Arabic, Farzi, Vietnamese) have less variability

Others – Ukrainian, Polish, Spanish, Catalan, German, French, Italian – show higher

# We don't know where the variability is occuring

- Representations of Émilie du Châtelet…
- The EN personal page has more academic citations than the FR
- The FR version has much more 'grey' citations
- However, the FR Wikipedia has pages about her work, with many more academic citations

Altmetric

Dimensions

# Some suggestions of exogenous factors

- Wikipedias in languages experiencing cultural / social / political / economic stress (e.g. Catalan, Ukrainian) may be developed as a reaction
- There may be a relationship between the language of scholarship, and the language of Wikipedia (e.g. German, Polish vs. Dutch) – and hence importance of focus?
- Although books are in the dataset, we've not broken them out – previous research of mine pointed to Arts / Humans / Social Sciences showing different behaviour in non-English languages
- Can we consider the focus / expertise / interest of either groups of editors or individuals? What does it mean now that publishers and institutions are 'taking an interest' in contributing to Wikipedias

# Conclusions

The English Language Wikipedia cites the largest number of outputs

The unique contribution of non-En Wikipedias is larger than the English

Each non-EN Wikipedia makes a unique contribution

In general, the larger the Wikipedia, the more citations it makes
(Dutch is a crazy outlier)

There's a big variation between Wikipedias – French, German and Polish exceed expected 'uniqueness-of-voice'. Arabic, Farzi and Vietnamese show significantly less 'uniqueness'

Subject area differences seem tied to other factors

We cannot treat non-EN Wikipedias as subsets or translations or samples of English: they make a unique contribution to representations of scholarship and deserve more study