The State of Wikimedia Research: 2014-2015

Benjamin Mako Hill
Tilman Bayer
Aaron Shaw

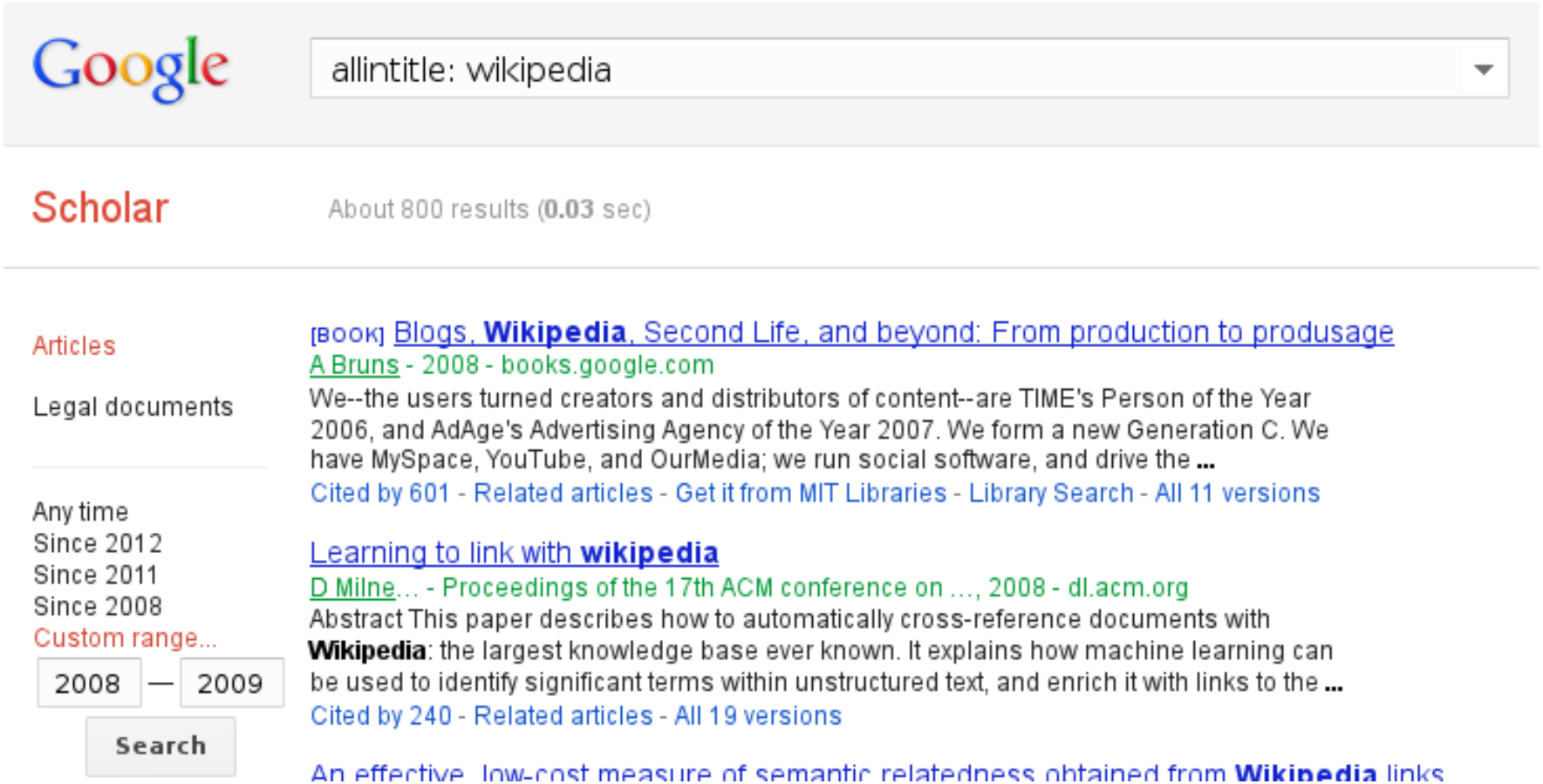
Wikimania 2015, Mexico City
July 17, 2015

“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From my Wikimania 2008 Submission

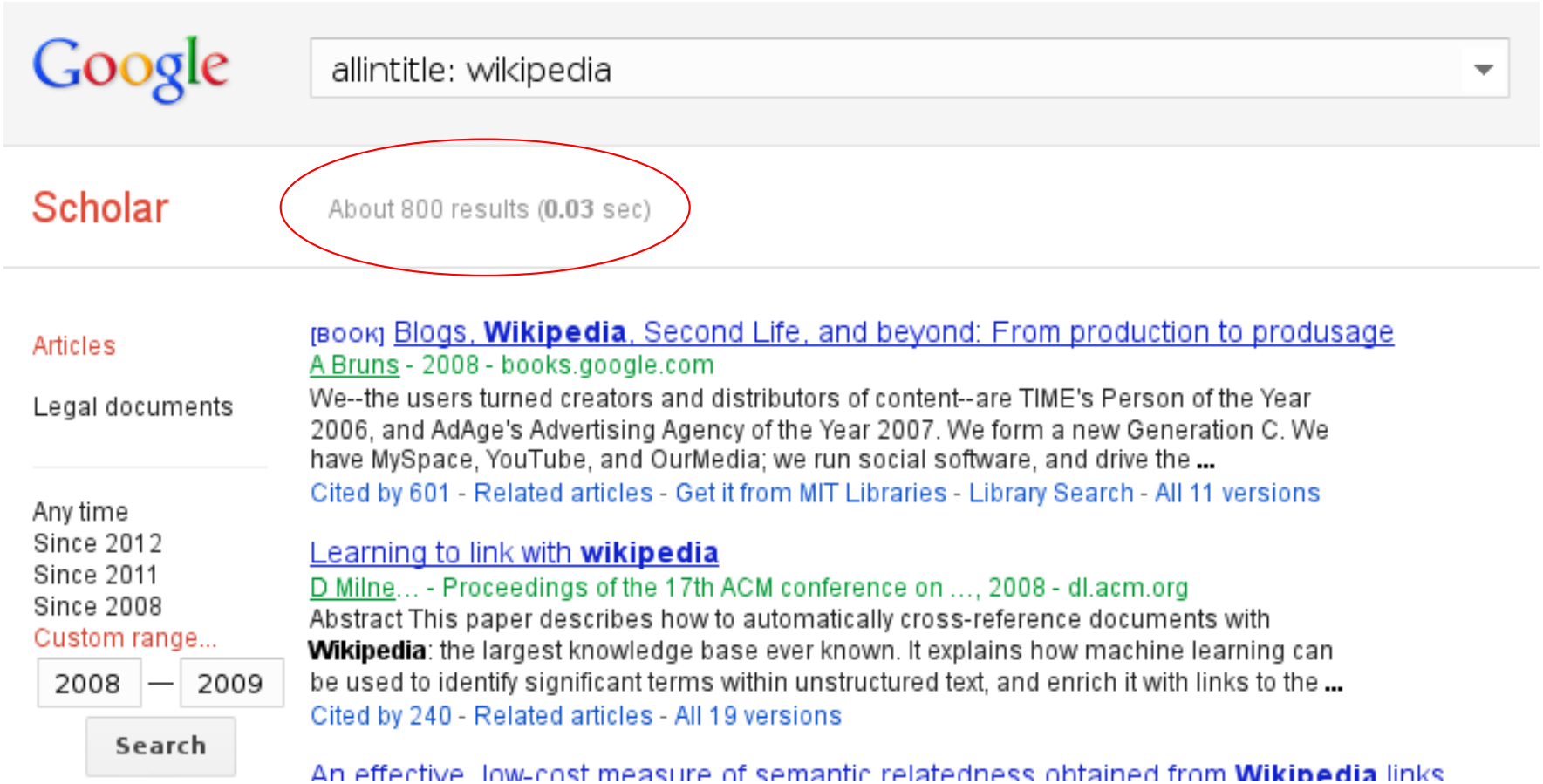
“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

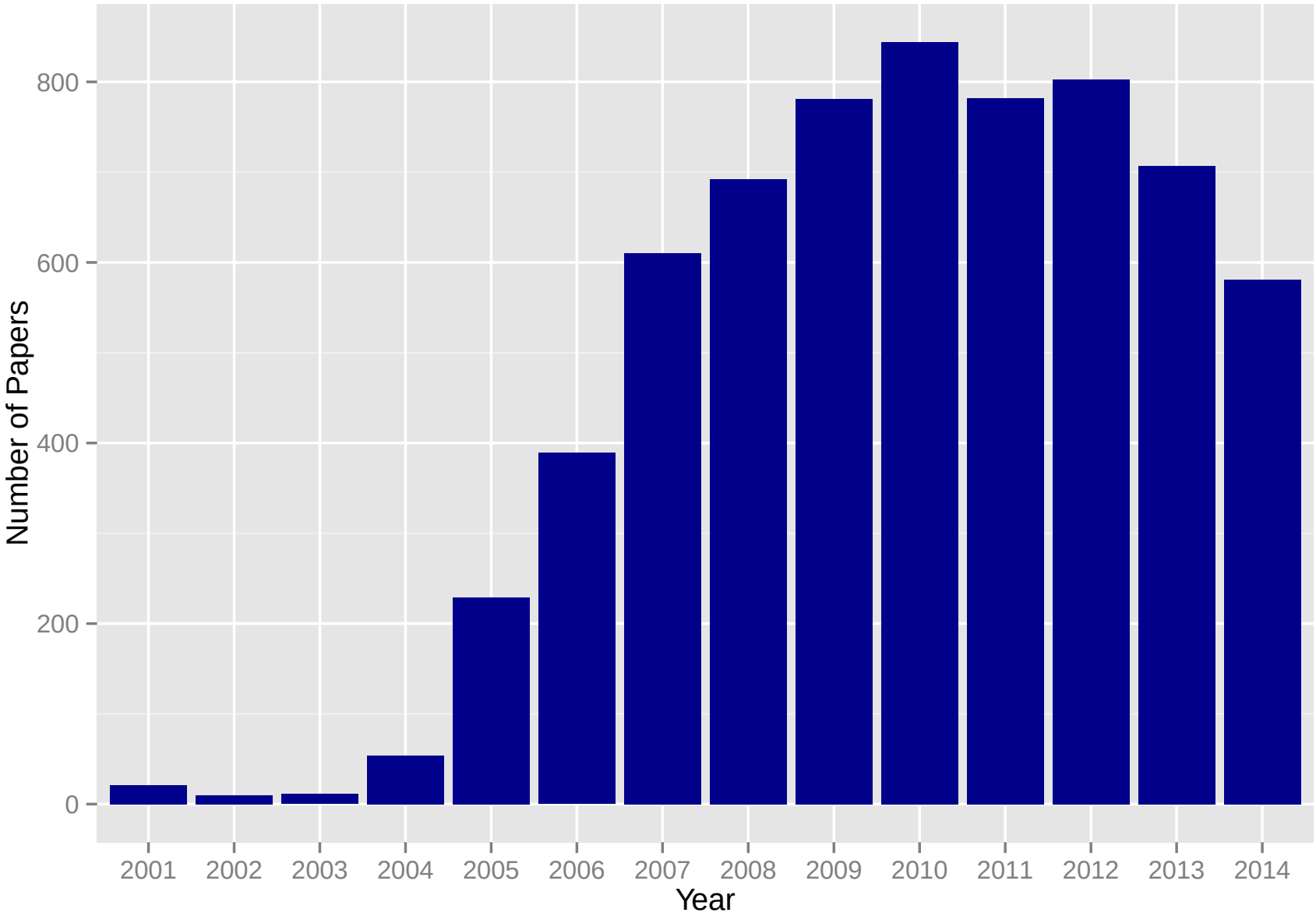
– From my Wikimania 2008 Submission



“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From my Wikimania 2008 Submission

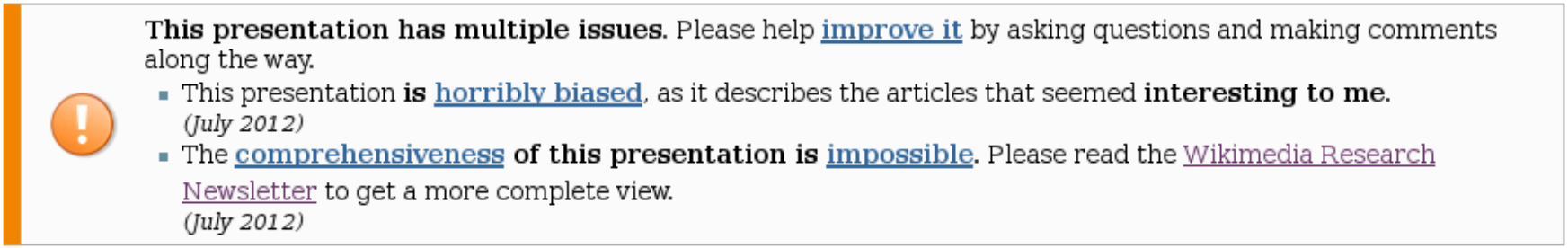




Number of citation, per year, with the term "wikipedia" in the title.

(Source: Google scholar results. Accessed: 2013-08-06)

- **2968** Wikipedia-related publications in the Scopus database as of November 2013
- **191** recent publications reviewed or mentioned in the 12 issues of the Wikimedia Research Newsletter from July 2014 to June 2015.



In selecting papers for this session, the goal is always to choose examples of work that:

- Represent **important themes** from Wikipedia in the last year.
- Research that is likely to be of **interest** to Wikimedians.
- Research by people who are **not at Wikimania**.
- . . . with a bias towards **peer-reviewed** publications

Wikipedia as a Source of Data

Wikipedia as a source of data

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). **Links that speak: The global language network and its association with global fame**. Proceedings of the National Academy of Sciences, 111(52), E5616—E5622. doi:10.1073/pnas.1410931111

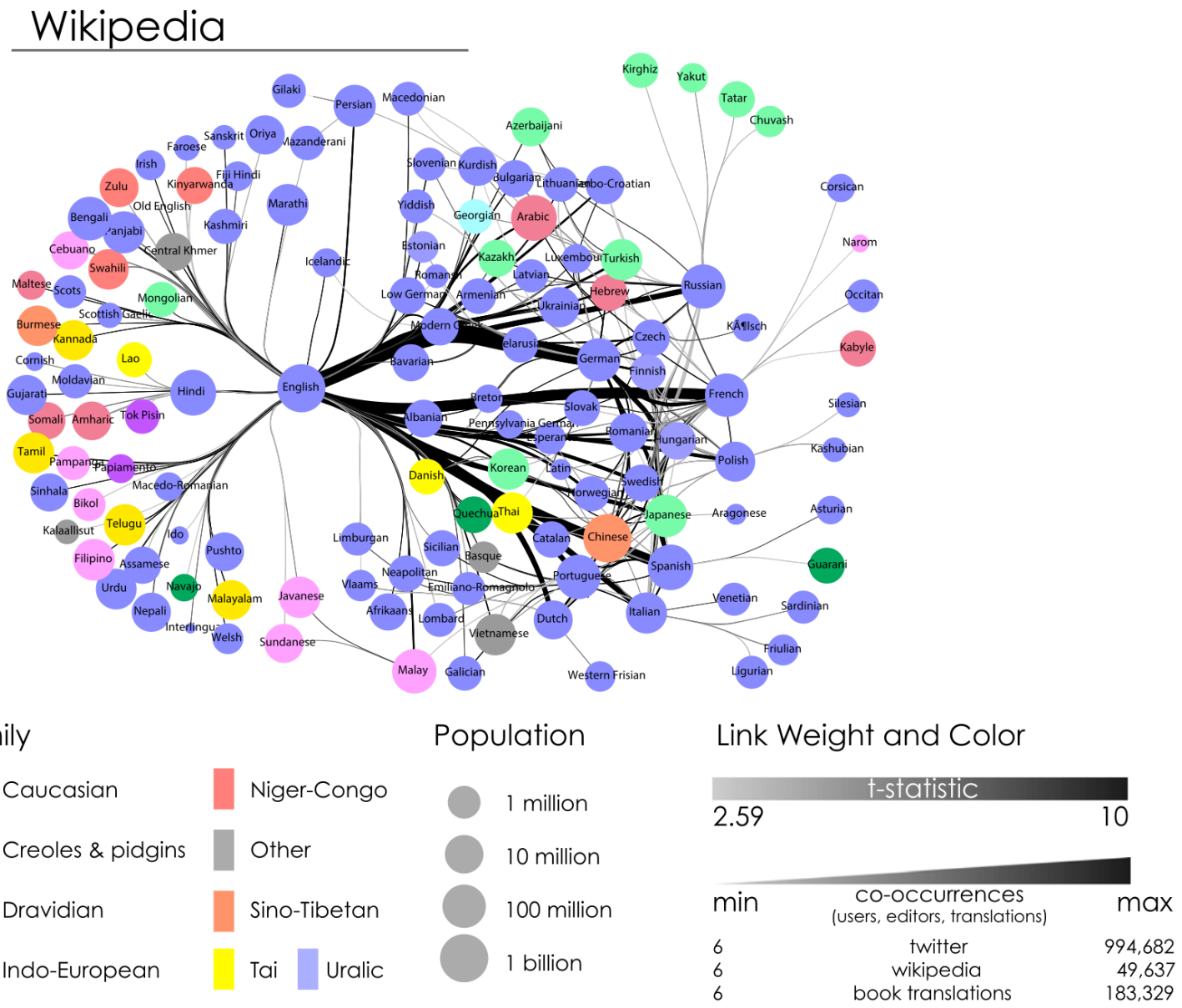
How to measure the global influence of languages?

Traditional methods rely on:

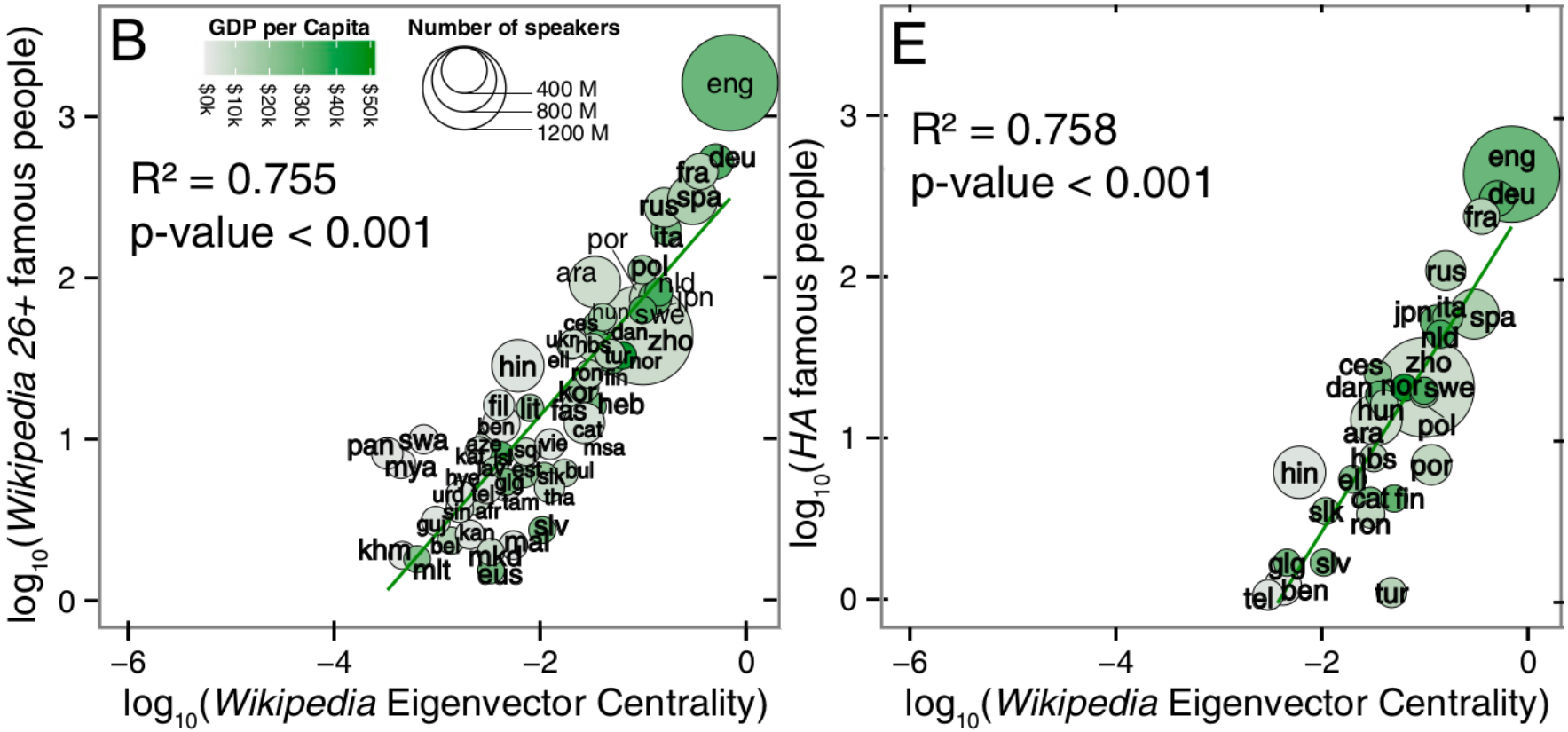
- **Population** of speakers
- **Income** or political power of speakers

Paper presents **new network method** based on measuring **co-speakers** of languages in several data sources including Wikipedia.

Wikipedia as a source of data: Ronen et al.



Wikipedia as a source of data: Ronen et al.



Community and Organization

Community and organization

Warncke-Wang, M., Ranjan, V., Terveen, L., & Hecht, B. (2015). **Misalignment Between Supply and Demand of Quality Content in Peer Production Communities**. In Ninth International AAAI Conference on Web and Social Media (ICWSM).

Perfect Alignment Hypothesis (PAH): There is an exact match between the supply of high-quality content and the demand for it.

	PAH_1	PAH_2	PAH_3	PAH_4	PAH_5	PAH_6	PAH_7
Q_1	1,710,819	477,687	30,701	6,647	657	16	64
Q_2	454,270	477,547	92,585	37,148	6,130	190	852
Q_3	43,255	71,012	26,749	19,056	6,259	232	1,344
Q_4	14,408	30,669	13,707	12,102	5,447	262	1,351
Q_5	3,649	9,416	3,192	2,136	953	62	506
Q_6	132	398	128	92	31	0	12
Q_7	59	1,994	846	766	438	32	218

Community and organization: Warncke-Wang et al.

Measure of the degree of misalignment can be used to build lists of categories that are relatively **"overproduced"** and **"underproduced"**:

Rank	Topic	N	Rel. Risk
1	Countries	144	506.9
2	Pop music	97	38.9
3	Internet	84	37.6
4	Comedy	134	21.9
5	Technology	58	15.8
6	Religion	121	15.8
7	Science Fiction	70	15.5
8	Rock music	84	11.4
9	Psychology	60	11.1
10	LGBT studies	136	9.1

Table 8: Topics most strongly over-represented in the Needs Improvement (NI) dataset, limited to topics w/at least 50 NI articles. "N" columns lists number of NI articles.

Rank	Topic	N	Rel. Risk
1	Cricket	65	159.0
2	Tropical cyclones	112	99.3
3	Middle Ages	87	13.4
4	Politics	147	12.0
5	Fungi	53	9.1
6	Birds	78	8.2
7	Military history	404	5.3
8	Ships	88	5.0
9	England	72	4.9
10	Australia	258	4.3

Table 9: Topics most strongly over-represented in the Effort (SE) dataset, limited to topics w/at least 50 SE cles. "N" column lists number of SE articles.

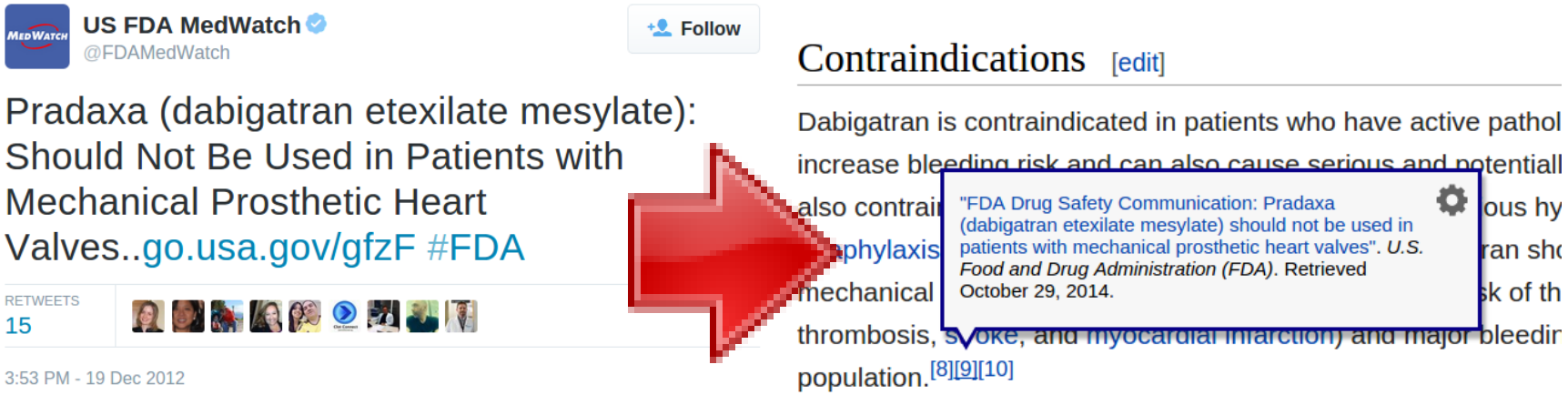
Content quality

Quality of drug articles

Hwang et al., “**Drug Safety in the Digital Age**.” N Engl J Med 2014; 370:2460-2462 June 26, 2014 doi: 10.1056/NEJMp1401767.

Kräenbring et al., **Accuracy and completeness of drug information in Wikipedia: a comparison with standard textbooks of pharmacology**. PLoS One 9 (9): e106930. doi:10.1371/journal.pone.0106930

Quality of drug articles: NEJM



- The US Food and Drug Administration (**FDA**) frequently issues safety warnings about prescription drugs. How long does it take until these are reflected on English Wikipedia?
- 41% updated within two weeks (58% for high-prevalent diseases), but 36% still unchanged after more than a year.

Quality of drug articles: PLoS One

- Selected 100 drugs from German undergrad curriculum in pharmacology
- Extracted information from two standard textbooks
- "Accuracy of drug information in [German] Wikipedia was 99.7%±0.2% when compared to the textbook data." Similar results for English Wikipedia

Quality of drug articles: PLoS One

- Completeness (as compared to the textbooks):
  - 83.8% (of 224 statements) for German WP
  - 87.2% for English WP
- Completeness of contraindications information was 100% in the En WP sample.
- English WP cited academic publications more often than German WP.
- Quality "significantly improved" in drug articles assessed in a 2010 study.

Automation in Wikipedia

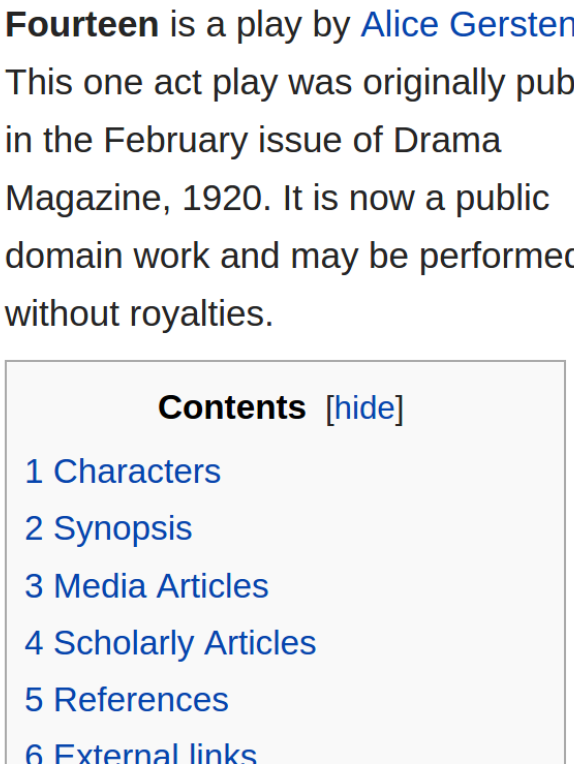
Banerjee et al., **Playscript Classification and Automatic Wikipedia Play Articles Generation**. 2014 22nd International Conference on Pattern Recognition (ICPR). pp. 3630–3635. DOI:10.1109/ICPR.2014.624

Author's copy

Automation in Wikipedia: Bot-written theatre play articles

- Bot searches for playscripts and related documents on the web
- Extract key information from them, e.g.
  - The play's main characters
  - Relevant sentences from online synopses of the play
  - Mentions in Google Books and Google News (as evidence that the play satisfies Wikipedia's notability criteria)
- Some heuristics to exclude non-encyclopedic sentences, e.g. first person statements

Automation in Wikipedia: Bot-written theatre play articles



Fourteen is a play by Alice Gerstenberg. This one act play was originally published in the February issue of Drama Magazine, 1920. It is now a public domain work and may be performed without royalties.

Contents [hide]

1 Characters
2 Synopsis
3 Media Articles
4 Scholarly Articles
5 References
6 External links

- 15 articles submitted at Articles for Creation. Two accepted by Wikipedia editors. One of them without major changes.

Gender Beyond the Gap

It's a Man's Wikipedia?

Wagner, Claudia; David Garcia; Mohsen Jadidi; and Markus Strohmaier. 2015. "**It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia**." Ninth International AAAI Conference on Web and Social Media (ICWSM).

It's a Man's Wikipedia: Motivation

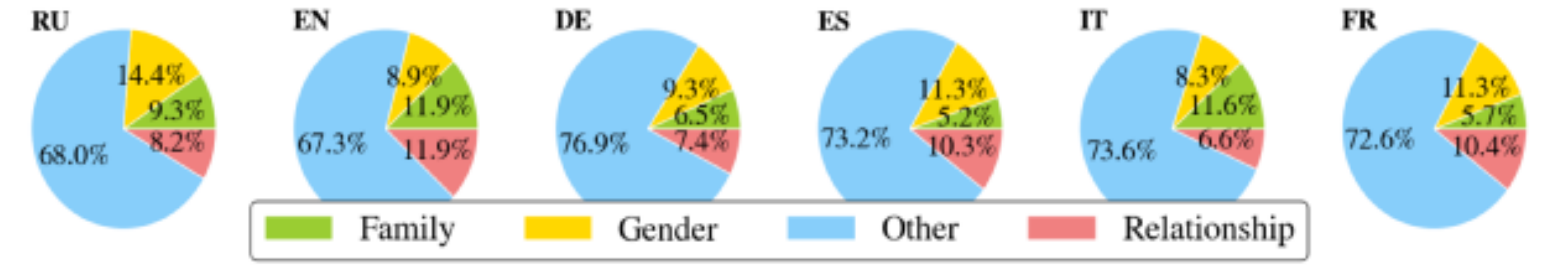
- We know there's a gender gap.
- Need for more multidimensional analysis of **how gender is represented in content of articles across Wikipedias**.

It's a Man's Wikipedia: Methods

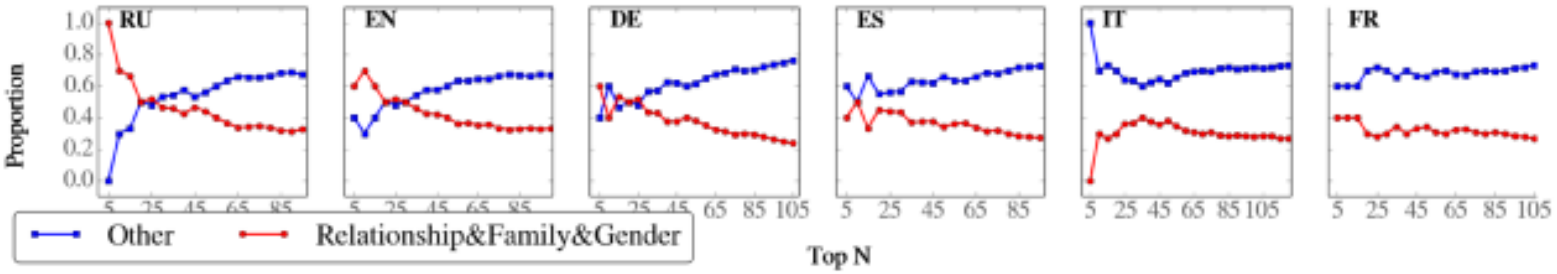
- Use data from three sources (Freebase, "Human Accomplishment," and Pantheon) as baselines for comparison with six Wikipedias (EN, ES, DE, FR, IT, RU).
- Examine multiple potential forms of bias: coverage, structure, lexical characteristics, visibility.

It's a Man's Wikipedia: Results

It's a Man's Wikipedia: Results



(a) Proportion of the 150 most discriminative words for women per category



(b) Proportion of the N most discriminative words for women per category

Figure 8: **Lexical Bias:** The proportion of the 150 most discriminative words of articles about women that belong to different categories. In all language editions between 32% and 23% of the 150 most indicative words for women belong to one of the three categories, while only between 0% and 4% of the most discriminative words for men belong to one of these categories. In some language edition, like the Russian (RU), the English (EN) and the German (DE) one, the proportion of the most discriminative words that belong to one of these three categories is especially high among the top words.

Adopting Wikipedia as a Teaching Tool

WP and the Wisdom of Crowds

Barnhisel, Greg and Marcia Rapchak. 2014. “**Wikipedia and the Wisdom of Crowds: A Student Project**.” Communications in Information Literacy 8(1): 145-159. doi:10.7548/cil.v8i1.249.

WP and the Wisdom of Crowds: Motivation

- Students use Wikipedia uncritically. Don't understand how low quality much of the information may be or how it may be manipulated.
- Professor (author) believes that WP is full of dubious information. Wants to unmask that for his students.
- Through more in-depth exposure, students may understand the limitations of collaborative, open systems of knowledge production.

WP and the Wisdom of Crowds: Methods

- Require a Senior (college) composition class to work on editing WP articles (together and individually) throughout the semester.
- Incorporate assignments to help students learn about the history of WP as well as how to use it.
- Require students to reflect on their experiences in writing.
- Require students to analyze the pros/cons of open collaborative writing in their final projects.

WP and the Wisdom of Crowds: Results

Both sources [crowds and experts] have different merits... My life experience since class pulls me in favor of the wisdom of the crowd. In my recent studies, I have found that I can learn much more from a group of my peers than from a single expert.

— Student 1

Meta-Analyses

- Mesgari, Mostafa and Okoli, Chitu and Mehdi, Mohamad and Nielsen, Finn Årup and Lanamäki, Arto. 2014. “The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia”. Journal of the Association for Information Science and Technology.
- Miquel-Ribé, Marc. 2015. “User Engagement on Wikipedia, A Review of Studies of Readers and Editors.” Ninth International AAAI Conference on Web and Social Media (ICWSM).

More Resources

- **Wikimedia Research Newsletter**
  [[:meta:Research:Newsletter]] / @WikiResearch
- **WikiSym/OpenSym** (This August in San Francisco!)
- **WikiPapers Repository** [http://wikipapers.referata.com]
- **Much More**