

CMPU 366 · Natural Language Processing

Ethics and Bias in Natural Language Processing

13 October 2026



Why are we talking about ethics now?

NLP systems are currently very pervasive,

Why are we talking about ethics now?

NLP systems are currently very pervasive,

We're about to dig into advanced deep learning and LLMs after break, and

Why are we talking about ethics now?

NLP systems are currently very pervasive,

We're about to dig into advanced deep learning and LLMs after break, and

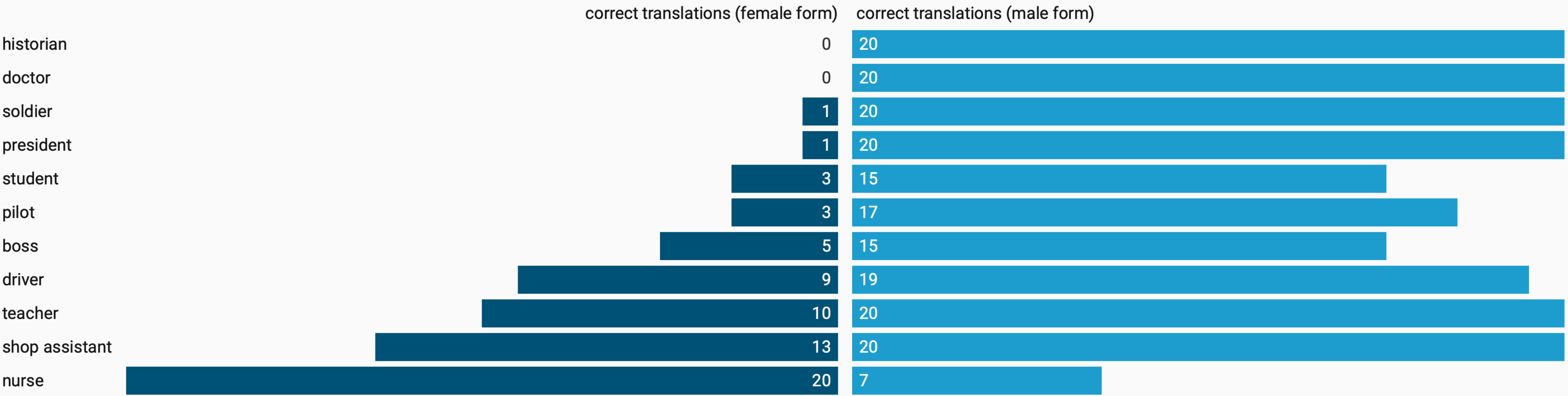


Bias has seemingly become the main topic of ethics in NLP.

For example, we've discussed bias in word embedding models, but this also creeps into downstream tasks like machine translation.

Female doctors don't exist, says Google Translate

Correct translations for 20 translation pairs to and from French, German, Spanish, Italian and Polish.



How to read the chart: Out of 20 translations of a female doctor, none were correct (e.g. "die Doktorin" become "le docteur", "la dottoressa" becomes "der Doktor" etc.)

Source: AlgorithmWatch • [Get the data](#) • Created with [Datawrapper](#)

Before



After



Kuczmariski, 2018

translate.google.com/?sl=en&tl=fr&text=developer&op=translate

Google Translate

Text Images Documents Websites

Detect language German Spanish **English**

developer

də'veləpər

Noun [See dictionary](#)

9 Advanced

Translations are gender-specific. [Learn more](#)

développeuse (feminine)

[See dictionary](#)

développeur (masculine)

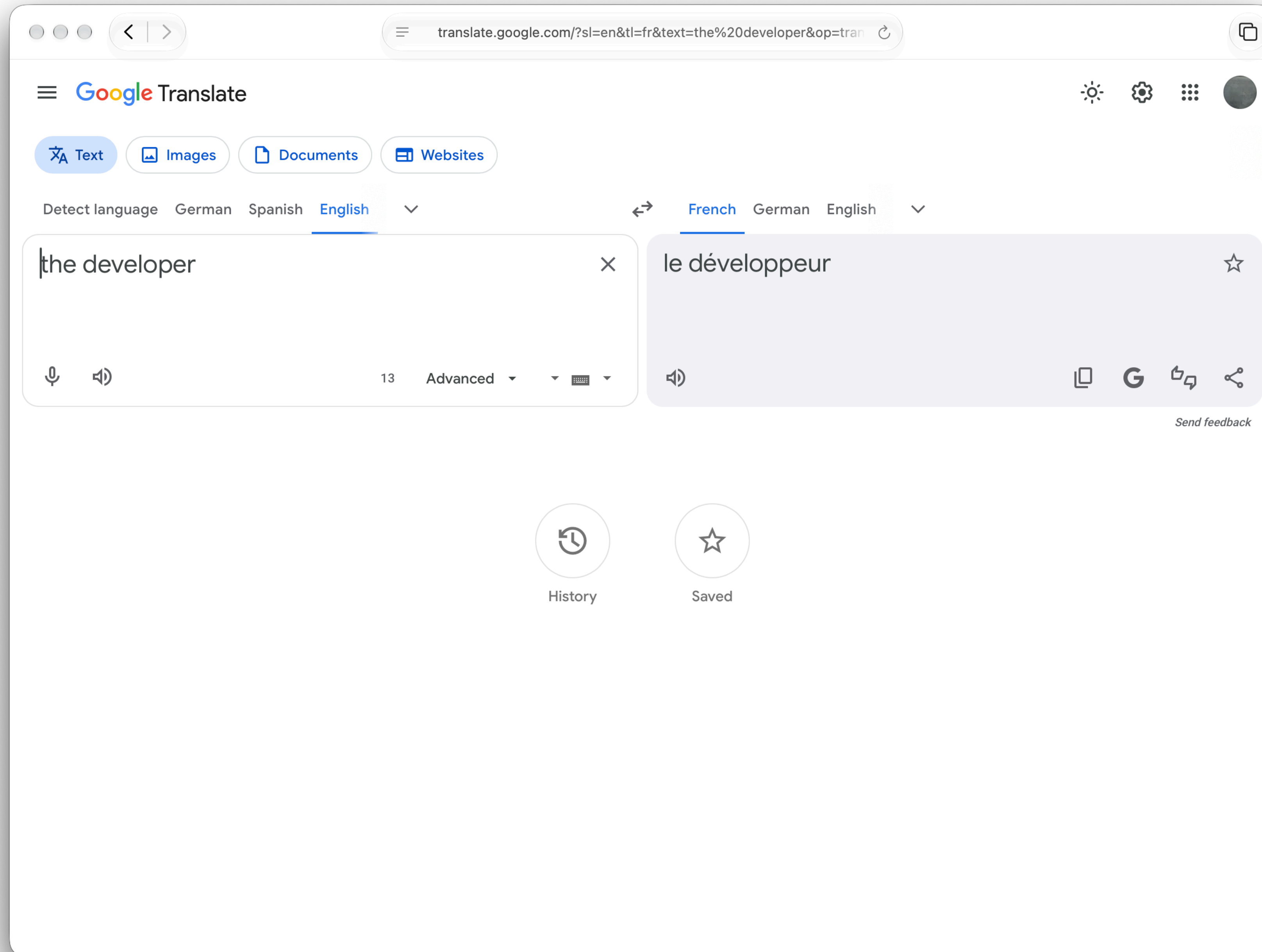
[See dictionary](#)

More translations Expand all

/e promoteur	Noun	sponsor	developer	instigator	matchmaker	▼
/e révélateur	Noun	developer	revealer	tattletale		▼
/e développeur	Noun	developer				▼

Send feedback

Cf. Kayser-Bril, 2020



Cf. Kayser-Bril, 2020

Bias can arise

- from the data itself,
- from the annotations we add,
- from the representations we use for the data,
- from the models we learn, or
- from the very design of our research.

It's almost impossible to have an unbiased system.

But bias isn't *inherently* bad.

Think of it like a prior probability. In the absence of more specific knowledge, it can help us make decisions.

However, if bias *overwhelms the evidence*, or if it *produces systematically worse outcomes for some groups*, it becomes problematic.

<

|

>

≡

scholar.google.com/scholar?hl=en&as_sdt=0,33&q=word2vec+resu

↺

📄

≡

Google Scholar

word2vec resume

🔍

👤

🏠 Articles

About 5,230 results (0.19 sec)

🎓 My profile

★ My library

Any time

Since 2026

Since 2025

Since 2022

Custom range...

Sort by relevance

Sort by date

Any type

Review articles

☐ include patents

☒ include citations

📧 Create alert

Scoring of resume and job description using Word2vec and matching them using Gale–Shapley algorithm

S Pudasaini, S Shakya, S Lamichhane... - Expert Clouds and ..., 2021 - Springer

... In this paper, the word2vec algorithm using CBOW model is implemented ... Word2Vec model trained on our custom data produced highly relevant results. Google pretrained Word2Vec ...

☆ Save 📄 Cite Cited by 22 Related articles All 4 versions 🔗

[HTML] academia.edu

🔖 Quick read

Uniqueness meets semantics: A novel semantically meaningful bag-of-words approach for matching resumes to job profiles

S Susan, M Sharma, G Choudhary - Inteligencia Artificial, 2024 - journal.iberamia.org

... Furthermore, we propose the use of GloVe and Word2Vec word embeddings to capture the semantics of the unique elite keywords extracted from each resume document. Word ...

☆ Save 📄 Cite Cited by 8 Related articles All 7 versions 🔗

[PDF] iberamia.org

🔖 Quick read

ViewIt @ Vassar

Research on the application of Word2Vec-based job-person matching methods in corporate recruitment

R Pang - Proceedings of the 2024 16th International Conference ..., 2024 - dl.acm.org

... In practice, this article first vectorizes resumes and job descriptions for recruitment positions in the form of Word2Vec embeddings. Through unsupervised learning, it identifies resumes ...

☆ Save 📄 Cite Cited by 4 Related articles 🔗

[PDF] acm.org

Full View

Resume parsing based on multi-label classification using neural network models

J Liu, Y Shen, Y Zhang, S Krishnamoorthy - Proceedings of the 6th ..., 2021 - dl.acm.org

... For the word2vec, the amount of data needed is tremendous and has sparse tagging ... As a word2vec training set, we caught teachers' resumes from 12 provinces and cities, such as ...

☆ Save 📄 Cite Cited by 23 Related articles All 2 versions 🔗

[PDF] acm.org

Full View

Words worth: Verbal content and hirability impressions in YouTube video resumes

S Muralidhar, L Nguyen... - Proceedings of the 9th ..., 2018 - aclanthology.org

... Word2Vec, GloVe) to understand verbal behavior in this context.In particular, using a dataset of 292 YouTube video resumes... improvement could be due to Word2Vec, being a predictive ...

☆ Save 📄 Cite Cited by 22 Related articles All 9 versions 🔗

[PDF] aclanthology.org

🔖 Quick read

Enhancing Candidate Selection with NLP-Driven Resume Analysis for Industry

<

|

>

www.fastcompany.com/90249309/amazons-hiring-ai-may-have-we

FASTCOMPANY

SUBSCRIBE

NEWS TECH DESIGN LEADERSHIP WORK LIFE VIDEO PODCASTS GAMES | EVENTS PARTNER STORIES

10-10-2018 | NEWS

Amazon's hiring AI may have weeded out women: Report

Bias in artificial intelligence is real.

SHARE

ADD ON GOOGLE



[Photo: Flickr user [Tony Webster](#)]

BY MELISSA LOCKER

Artificial intelligence is slowly becoming the first round of human resources departments everywhere. To increase efficiency in the

A useful concept to distinguish moral gradation:

what we want the world to be, vs

what it is

A coreference system that cannot resolve female pronouns with the noun *doctor* is

normatively wrong – we want women to be doctors

descriptively wrong – when the sentence is actually referring to a female doctor.

Racially or gender-biased word embeddings are

normatively wrong – we don't want systems to proliferate stereotypes
but they might be *descriptively* correct – they reflect how societies *talk about* gender and ethnicity.

*That doesn't mean
those attitudes are
good or are founded
in reality!*

How does NLP affect *social justice* – the equal opportunities for individuals and groups within society to

access resources,

get their voice heard, and

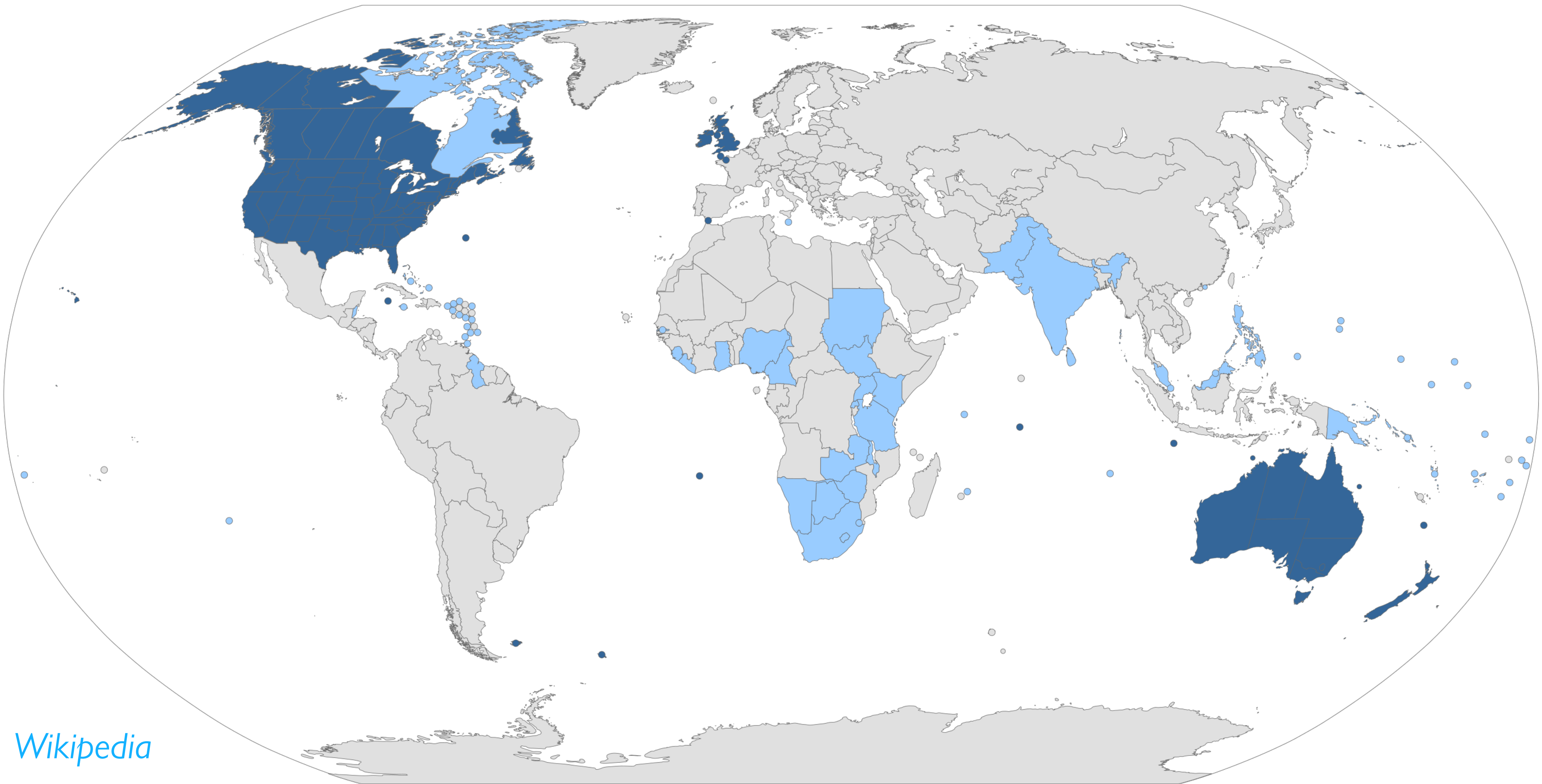
be represented in society?

African-American English (AAE) isn't well represented in training data for language identification, parsing, etc., so technologies like translation and intelligent assistants aren't as usable for its speakers.

Result: Users are forced to choose between avoiding their dialect or missing out on technological benefits.

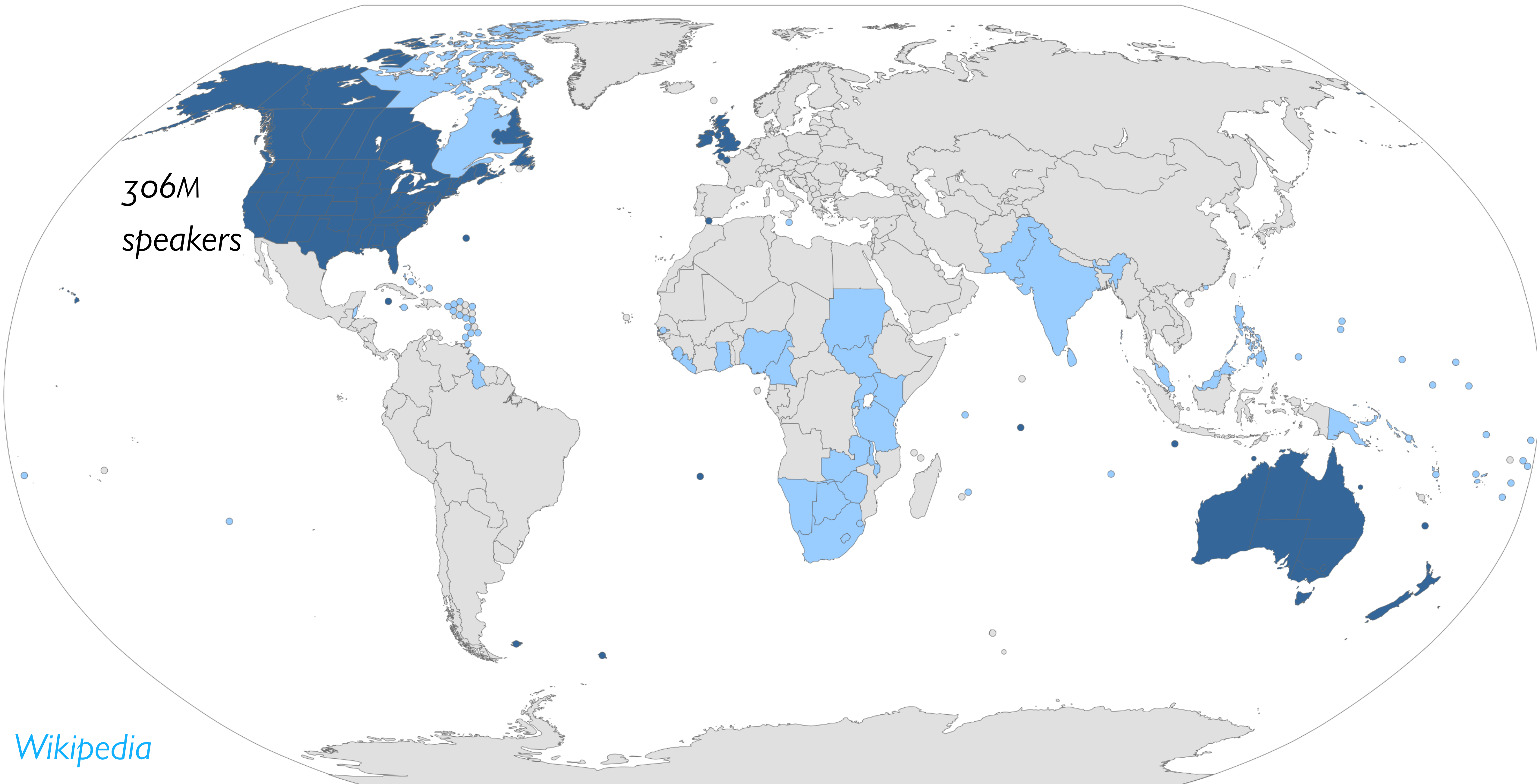
See *Blodgett & O'Connor, 2017*.

There's a similar situation for English varieties in India, Nigeria, the Philippines, Singapore, the Caribbean, etc. – and for non-standard or stigmatized *language varieties* in general.

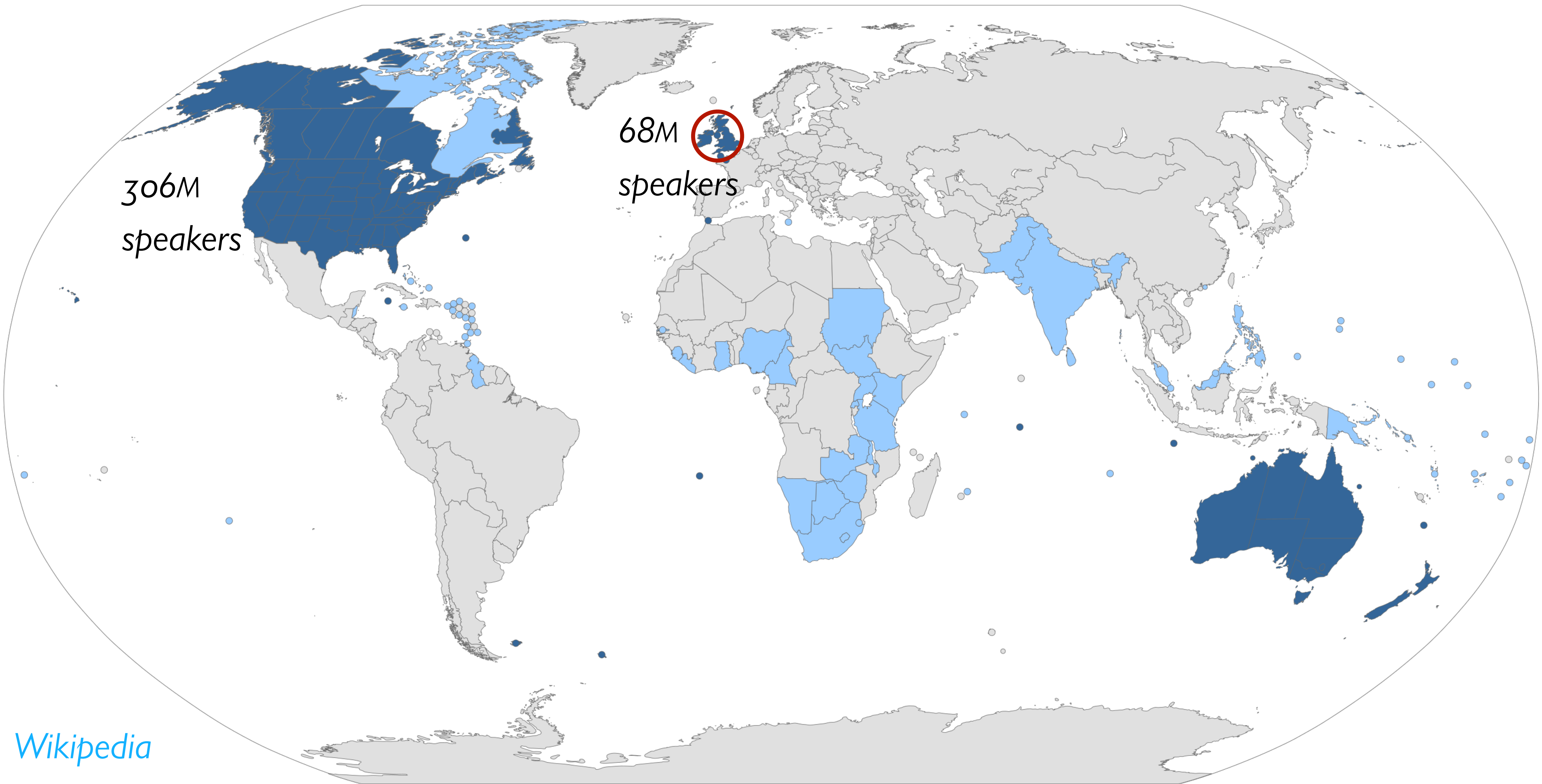


[Wikipedia](#)

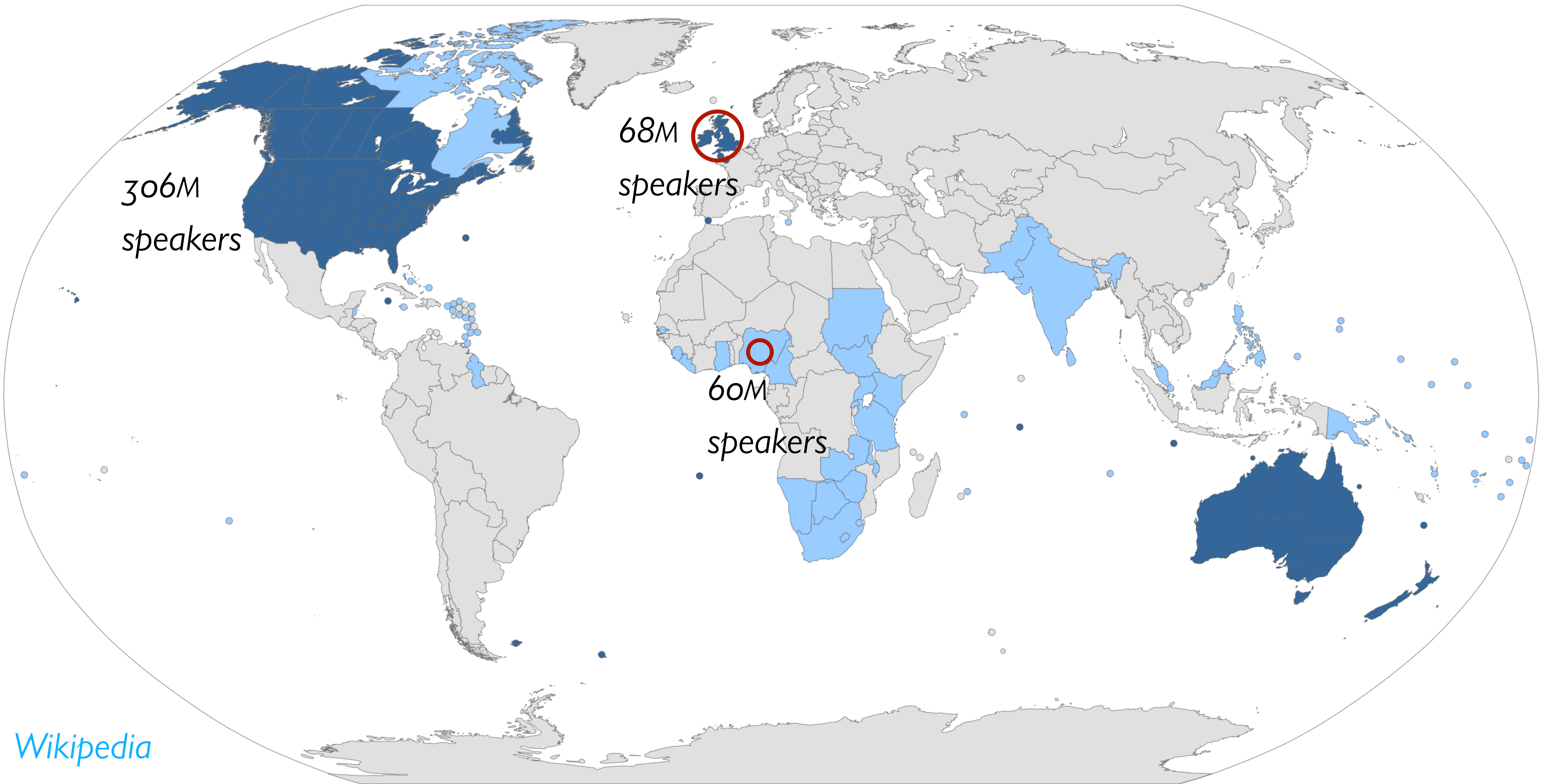
Note: This mixes different sources for the counts



Note: This mixes different sources for the counts

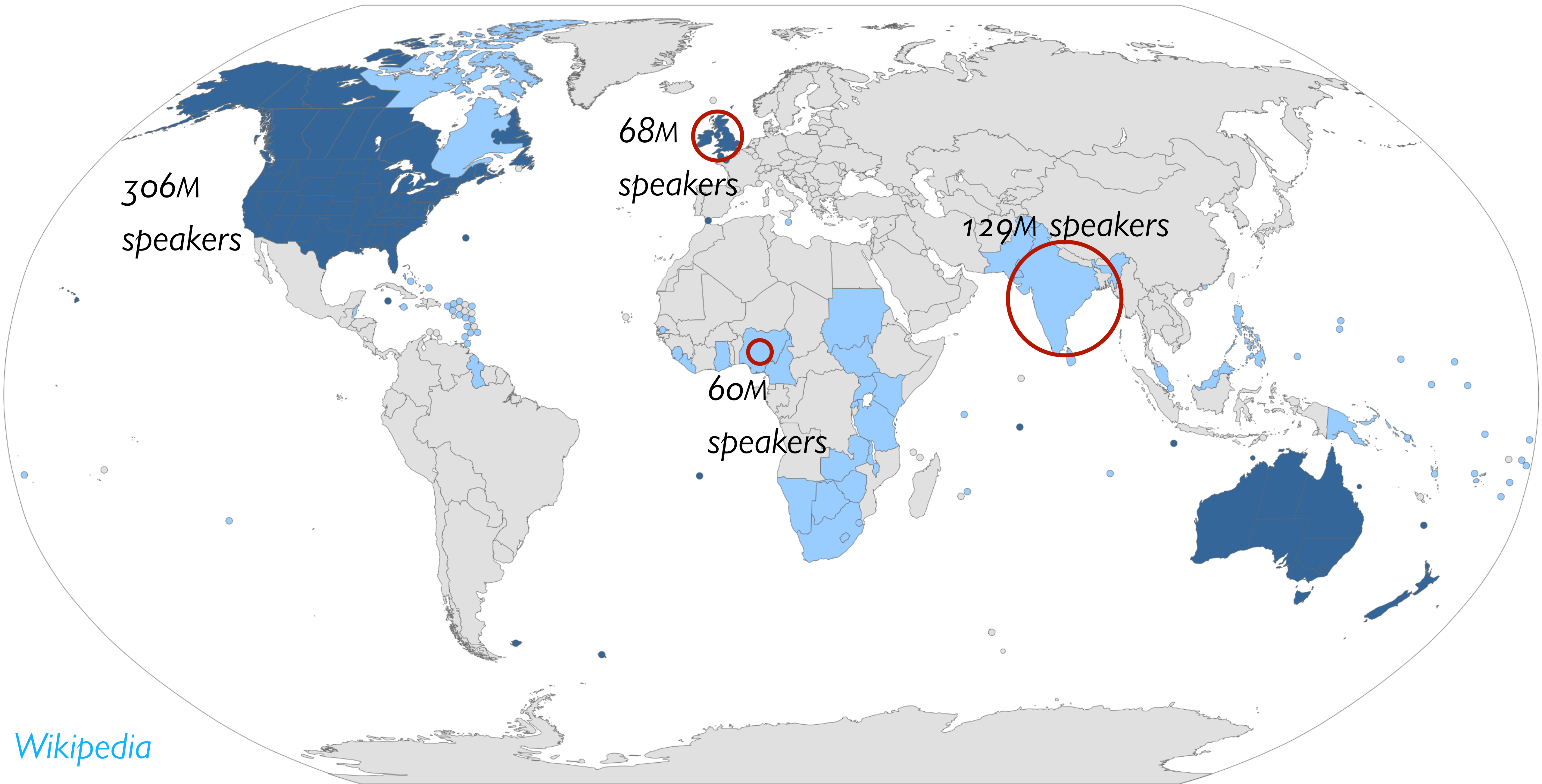


Note: This mixes different sources for the counts



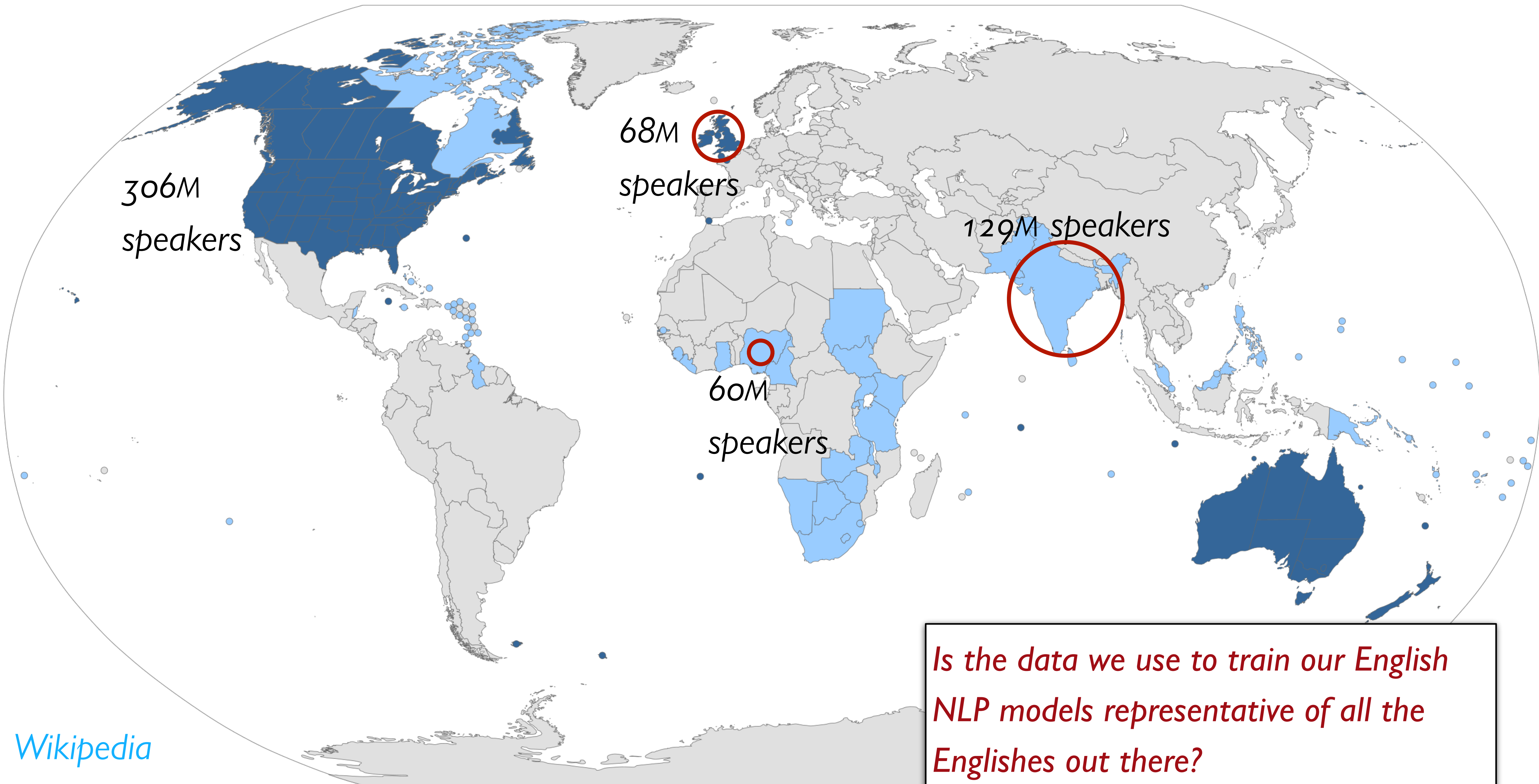
[Wikipedia](#)

Note: This mixes different sources for the counts



Wikipedia

Note: This mixes different sources for the counts



Is the data we use to train our English NLP models representative of all the Englishes out there?

Wikipedia

Note: This mixes different sources for the counts

Easy steps that help reduce problems stemming from biased or unrepresentative data:

When writing up NLP research, be clear about:

What your data looks like, why it was collected, and what kind of information your system learns from it.

Who (country, region, gender, native language, etc.) produced the text and labels in your data set(s).

Any known biases in your data set(s) – including the obvious ones!

This is *especially* important when writing for nontechnical potential users or clients.

See *Bender & Friedman (2018)* and *Gebru et al. (2018)*.

Another major concern in NLP is with preserving privacy.

Privacy means nobody knows I am doing *something*.

Anonymity means everyone knows *what* I am doing,
but not that it is *me*.

NLP researchers started out (willfully?) ignorant of the ramifications of how they were using online data.

The usefulness of being able to access so much text online led to an attitude that anything “public” was fair game.

We’re living through a moment of re-evaluation of that attitude.

The ethics of the data sets we gather and what we do with them is moving to the forefront.

Why is text from social media different from traditional corpora like newspaper articles or pre-20th Century novels?

Historically, “human participants” are viewed as people the researcher *interacts* or *intervenes* with.

Professional and academic codes haven’t required ethical scrutiny to collect, store, or study *publicly available data*.

However, with social media, many researchers are questioning these old assumptions, partly due to studies of users’ own attitudes.

What do users think?

Do Twitter users know their tweets might be used for research?

How do they feel about potential research on their tweets? E.g.,

- Is it okay at all?

- Do they want to be asked permission?

- Is it okay to publish their tweets in papers?

What do users think?

Fiesler & Proferes (2018) found:

Most (62%) of respondents didn't know that tweets are sometimes used for research.

Only 20% were uncomfortable with the idea in general, but this rose to 48% if the study includes their entire Twitter history.

Other factors matter, such as:

- Whether they explicitly give permission

- What the study is about and who is doing it

- Whether their tweets are part of a much larger dataset

- Whether profile information is also used

- Whether tweets are analyzed by humans or by computers

What do users think?

Williams et al. (2017) found:

Most respondents were at least slightly concerned about use of their tweets in research.

More concerned for commercial research than academic

Asked about their tweets being published in research papers:

80% of respondents expected to be asked for consent before

90% of respondents expected their tweets would be anonymized

User expectations vs Twitter guidelines

Users seem to want anonymity if tweets quoted in research publications.

However, Twitter guidelines explicitly stated that published tweets should be reproduced verbatim, including the user's name and Twitter handle.

Twitter also said users should be able to delete their tweets, which is effectively impossible if published non-anonymously!

Facebook used NLP techniques to identify the emotional content of posts and then reduced either the amount of positive or negative content in a user's feed.

Why? To check whether doing so would affect the sentiment of what you post.

New technology like NLP can be conceived of as a social experiment ([Van de Poel, 2016](#)).

If we assume we are all participating in a large experiment, we need to make sure it meets certain criteria of responsible experimentation:

Beneficence – no harm to subjects; maximize benefits and minimize risk

Respect for subjects' autonomy – informed consent

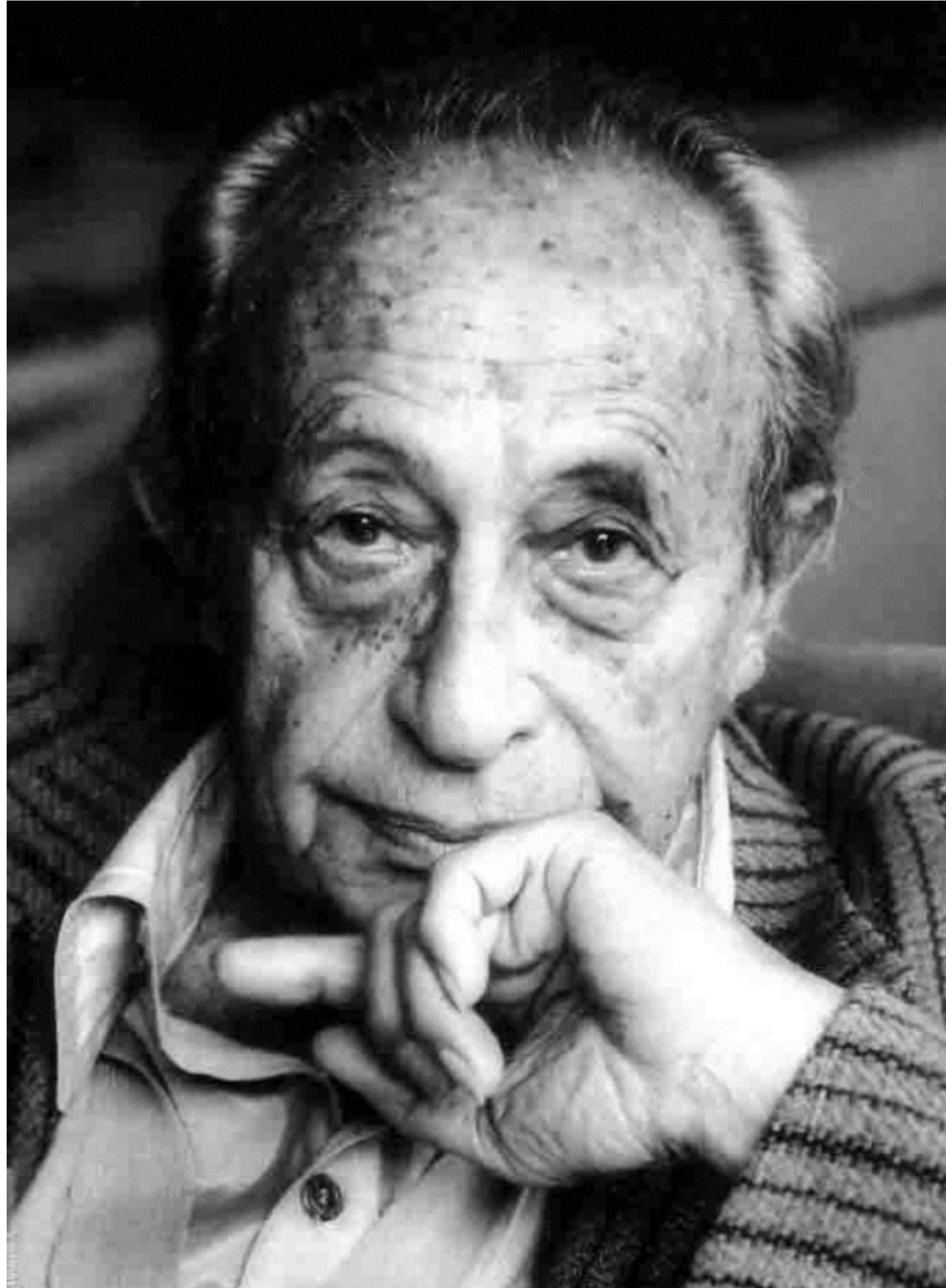
Justice – benefits vs harms, protection of vulnerable subjects

Refer to the Belmont Report, 1979

Every technology has an intended use and unintended consequences.

Nuclear power, knives, electricity can all be abused for things they weren't originally designed to do.

Since we don't know how people will use something, we need to be aware of this duality.



Hans Jonas, 1903–1993



Real Genius, 1985

Consider this scenario:

One of your classmates suggests an idea to explore for their final project – studying gendered language in the LGBTQ community.

They're very engaged in the community themselves and have access to naturally occurring data (e.g., Reddit posts).

Their plan is to write a text classification tool that distinguishes LGBTQ language.

What do you say?

You develop a system to automatically respond to spammers, trying to engage them in email conversation for as long as possible.

Is this ethical? Does this research require IRB approval?

Menu

About the ACL

News

Journals

Conferences

Events

ACL Fellows

SIGs

Anthology

Wiki

Education

Policies

Archives

Association for
Computational Linguistics

Search this site..

ACL Code of Ethics

March 13, 2020 | BY webmaster

ACL adopts the ACM Code of Ethics (<https://www.acm.org/code-of-ethics>) in the version adopted June 22nd, 2018, by the ACM Council. In its application to ACL, it is to be read in the contextually appropriate interpretation, e.g., "ACM member" is to be read as "ACL member". Sec 4.2 should be read as follows:

"4.2 Treat violations of the Code as inconsistent with membership in the ACL. Each ACL member should encourage and support adherence by all members of the CL/NLP community regardless of ACL membership. ACL members who recognize a breach of the Code should consider reporting the violation to the ACL, which may result in remedial action."

Hinrich Schütze, ACL President

User login

Username *

Password *

Create New Member Account

Request New Password

Request Username Reminder

Log in

Latest Events

NAACL Board Election: Call for
Community Nomination

.....

Call for Participation: SemEval-
2027 Task 6: MMCultureQA

.....

Reminder: Call for Papers -

aclweb.org/portal/content/acl-code-ethics

<

|

>

acrollingreview.org/responsibleNLPresearch/#guidelines-for-answe

A. For every submission:

A1. Did you describe the *limitations* of your work?

- Point out any strong assumptions and how robust your results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only held locally). Reflect on how these assumptions might be violated in practice and what the implications would be.
- Reflect on the scope of your claims, e.g., if you only tested your approach on a few datasets, languages, or did a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated. Reflect on the factors that influence the performance of your approach. For example, a speech-to-text system might not be able to be reliably used to provide closed captions for online lectures because it fails to handle technical jargon.
- If you analyze model biases: which definition of bias are you using? Did you state the motivation and definition explicitly? See the discussion in [Blodgett et al. \(2020\)](#).
- We understand that authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection. It is worth keeping in mind that a worse outcome might be if reviewers discover limitations that aren't acknowledged in the paper. In general, we advise authors to use their best judgement and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

A2. Did you discuss any potential *risks* of your work?

- Examples of risks include potential malicious or unintended harmful effects and uses (e.g., disinformation, generating fake profiles, surveillance), environmental impact (e.g., training huge models), fairness considerations (e.g., deployment of technologies that could further disadvantage or exclude historically disadvantaged groups), privacy considerations (e.g., a paper on model/data stealing), and security considerations (e.g., adversarial attacks). See discussion in [Leins et. al. \(2020\)](#) as examples.
- Does the research contribute to overgeneralization, bias confirmation, under or overexposure of specific languages, topics, or applications at the expense of others? See [Hovy and Spruit \(2016\)](#) for examples.
- We expect many papers to be foundational research and not tied to particular applications, let alone deployments.

This isn't the last time we'll talk about ethics in NLP;
this topic gets even more important – and
complicated – after we introduce LLMs!

