INFORMED OPINIONS

GENDER GAP TRACKER

Powered by SFU

Methodology Backgrounder

The Gender Gap Tracker tool works in three stages. The first stage collects articles from online news sources and stores them in a database. This is done by scraping the listed news sources through either their RSS feed or via a media-oriented scraping tool in Python (a programming library called "newspaper").

The second stage takes the collected articles, identifies the people in the article and pulls out whether they are speaking or not, and what their gender is. Various natural language processing tools are used for this stage and the support of language researchers at Simon Fraser University is crucial for this step.

The third stage draws on the information produced from the first two and produces the website and visualizations.

Neither of the first two stages is perfect. We know that we are not scraping every article from each site, and that some are produced by newswires like CP and Reuters, which may be duplicated across media outlets. In addition, the language processing tools get some names wrong in various ways (by misidentifying or missing a person, misidentifying or missing a speaker, and misidentifying or missing a gender). Among every 100 names that we query, two will not be tagged male or female by the gender services software. We believe that the negative consequences of these problems are reduced in three ways:

1. Research and improvements are ongoing. We have been cross-checking the data and refining the process since last fall, and continue to do so. By making the tool public, we're inviting questions, knowing that these will help the development team improve it further. The tools used to produce this work are also improving independently.

2. We are dealing with data at a large scale*.* This has the effect of minimizing errors particular to individual articles, such as those due to a particular writer's style and/or the subject. (For example, if "Dr. Pat Smith" is only quoted once, there may be no associated "he" or "she" pronoun.) More general errors are distributed across the sites (a speaker misidentified as "male" will be misidentified on every site where the speaker is quoted). While we work to correct such errors, no media organization is likely to be unfairly affected. In addition, our analyses currently show that these errors are not biased in favour of one gender: the tool is as likely to misidentify men as women.

3. Revealing this research will help change how things are done. A key aim of tracking this data is that first, more people from a range of genders will be recognized as experts and their voices sought as experts. Secondly, we hope to make it easier for media organizations to discover who they are quoting as experts, possibly through additional meta-data in the article, so that they, too, are able to more easily track the diversity of their sources.

How we extract quotes:

The Gender Gap Tracker algorithm looks for quotes and their speakers. To identify the gender of the speakers, we rely on online services that provide databases of names. We also rely on the context around the quote and any pronouns referring to the speaker. An example:

(https://www.cbc.ca/news/canada/toronto/beaches-residents-bag-stations-woodbine-1.4973284)

> The Vuntut Gwitchin First Nation in Old Crow, Yukon has a new chief-elect: Dana Tizya-Tramm.
>
> The former councillor and president of Youth of the Peel won the election over incumbent chief Bruce Charlie and former MLA Darius Elias in a vote Monday.
>
> Tizya-Tramm described the campaign and election as "very humbling.
>
> "I think I reached really deep down inside myself, and found this wonderful place in my heart that, no matter what, it would have been a success for myself," he said.

In this example, our gender identification service may not know for sure whether Dana is a man or a woman. But the quote extracted from the last paragraph is immediately followed by the "he said" reference, which allows us to identify the speaker as male.

When a person is quoted several times in the same article, a subsequent quote may not be tied to a gender identifying reference. But because we count sources only once per article, this does not increase the error rate significantly.

Other information we are capturing

The algorithm analyzes and stores many different data points about each article. Currently, the dashboard is only showing *sources* (people quoted). However, we also have the capacity to measure:

- Gender breakdown for **authors** of articles
- Gender breakdown of **people mentioned** in the article (whether quoted or not)

Future work:

We're interested in exploring the distinction between **sources** and **experts**. People quoted may be witnesses, victims, politicians or newsmakers. As we develop the tool, we aim to clarify how many of those quoted provide an expert opinion, and what their gender breakdown is.

For more information about the research methodology, please contact:

John Simpson jsympson@gmail.com
Digital Humanities Specialist
University of Alberta

Maite Taboada mtaboada@sfu.ca
Director, Discourse Processing Lab
Simon Fraser University