Radar de Parité: An NLP system to measure gender representation in French news stories

Valentin-Gabriel Soumah†, Prashanth Rao†, Philipp Eibl†, Maite Taboada†,*
†Discourse Processing Lab, Simon Fraser University, Canada

Abstract
We present the Radar de Parité, an automated Natural Language Processing (NLP) system that measures the proportion of women and men quoted daily in six Canadian French-language media outlets. We outline the system’s architecture and detail the challenges we overcame to address French-specific issues, in particular regarding coreference resolution, a new contribution to the NLP literature on French. We also showcase statistics covering over one year’s worth of data (282,512 news articles). Our results highlight the underrepresentation of women in news stories, while also illustrating the application of modern NLP methods to measure gender representation and address societal issues.

Keywords: Natural language processing, French, quote extraction, coreference, news, gender

1. Gender representation in news stories
Natural Language Processing (NLP) has excellent potential for applied research in various areas, such as capturing sentiment in reviews or news stories, highlighting toxicity in social media, or mining the biomedical literature. The commonality in most applied NLP research projects is the need to reliably and scalably extract information from unstructured text data. In this paper, we describe one such application: extracting quotes from news stories to quantify gender representation.

Gender representation in the media is a long debated topic. From the 1970s, there have been studies into how much women and gender-diverse people are portrayed in news stories, with the general hypothesis that they tend to be underrepresented [1, 2]. There is also research studying how they are represented, i.e., whether sexist or homophobic tropes are present when we discuss women and gender-diverse people [3, 4]. In this work, we tackle one specific aspect of representation: who is quoted and in what proportions. Our starting hypothesis is that we hear less from women than from men in news stories, that is, that men are quoted more often than is to be expected from their proportion in the general population. To fully answer this question, we formulate a quantitative approach, collecting large amounts of representative data and extracting quotes from the unstructured text. This is the goal of the Radar de Parité.

We define quotes as either direct or indirect reproductions of what a person said, and we define that person as a source in news articles. In order to extract quotes, we employ a full NLP pipeline, focusing on parsing to identify speakers, verbs, and quotes, in each news story. We then predict the gender of the speaker (or source), using external gender-prediction services. Finally, we display the results on a public-facing dashboard.1

The contributions of our paper are multifold. First, we address the difficulty of doing large-scale NLP in languages other than English. We show that, even for a relatively well-resourced language such as French, off-the-shelf language models and specialized NLP modules are not always readily available. Next, we show how we built a modular system for applied research, where the goal is not to improve a foundational algorithm or to achieve state of the art results, but rather to produce a reliable, accurate system for a specific use

1French dashboard: https://radardeparite.femmesexpertes.ca/
English dashboard: https://gendergaptracker.informedopinions.org/
Code: https://github.com/sfu-discourse-lab/GenderGapTracker

*Corresponding author: mtaboada@sfu.ca

This article is © 2023 by author(s) as listed above. The article is licensed under a Creative Commons Attribution (CC BY 4.0) International license (https://creativecommons.org/licenses/by/4.0/legalcode), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the author(s) identified above.
2. Multilingual NLP and the hegemony of English

One of the promises of NLP is the ability to reuse resources from one language and apply them to another. In principle, many innovations in tokenization, part-of-speech tagging, parsing, word sense disambiguation, and coreference resolution, if developed for English, can carry over to other languages, after adjusting for linguistic differences. In practice, however, ready-to-use algorithms suffer from a lack of annotated data to train on. This is the case even after the advent of large language models, most of them trained on English-language data [5, 6]. Chinese is perhaps the exception [7], but we welcome the recent efforts of groups such as HuggingFace in providing large open-source multilingual models [8]. We see this lack of resources for languages other than English as an obstacle in the practical application of NLP. Despite the inherent power of transformer-based language models, as well as related breakthroughs in multi-modal and transfer learning, their computing requirements are still a challenge when applied at scale [5].

As a result, we rely on spaCy, an industrial-strength NLP library (see Section 3) to perform the analysis shown in this paper. We chose spaCy because it is robust, well supported, and regularly updated, with a full pipeline. There are perhaps state-of-the-art resources for each of the components (e.g., for French NER [9]), but integration into a pipeline is at best complex and often impossible. In prior work, we had built a quote detection system for English, which is well supported by spaCy’s existing English language models [10, 11]. For French, we lacked a module for coreference resolution. spaCy, until recently, provided it via neuralcoref [12], a pre-trained and trainable neural coreference module. neuralcoref, however, works only in English, with no equivalent functionality for other languages. As we detail in the next section, we ended up adopting and expanding another library, coreferee, which supports English, French, German, and Polish.

3. NLP pipeline

We scrape news articles daily from six French-language outlets in Canada. We store the articles and their metadata to a database and pass it to our NLP pipeline. We apply a typical sequence of steps to preprocess unstructured text data, including tokenization and part-of-speech tagging, in order to extract information for downstream analysis (Figure 1).

The goal of Named Entity Recognition is to identify the names of people mentioned, a subset of which are typically quoted in a news article. Dependency parsing is used to extract syntactic quotes. Coreference resolution is a key step that allows us to cluster references to the same person (e.g., Valérie Plante, Mme Plante, she, her).

3.1. Quote extraction

To measure the gender representation gap in news media, we identify the number of men and women who are quoted in news articles; in other words, people who have not only been mentioned, but have also seen their voices reflected in the news. We consider both direct speech, surrounded by quotation marks (She stated, “...”), and indirect speech (She

---

2https://spacy.io/universe/project/neuralcoref
3https://spacy.io/universe/project/coreferee
stated that...) to be quotations. We extract all the quotes from the news article text, following which we align the quoted speakers with the unified named entities that were gender-labelled (see Section 3.5). Based on our study of the literature on reported speech [13–15], we separate quotes into three different types: direct quotes, indirect quotes, and selon quotes. Examples of each are shown in Example (1). More detailed descriptions of the quote extraction process can be found in the full version of the paper, with Appendix.4

(1) **Direct**: «N’entre pas qui veut dans le cercle de Vito Rizzuto. Importer 2000 lb de haschich à 26 ans, il n’a sûrement pas fait ça tout seul», dit M. Mansueto.

**Indirect**: Le sergent-détective a également expliqué que les criminels avaient besoin de s’impliquer dans des activités légales pour blanchir leur argent sale.

**With selon**: Selon eux, l’individu n’offre qu’une partie des services de distribution alimentaire et n’a aucun intérêt dans la Grande Roue.

### 3.2. Recognizing people

By **people** we mean all human entities or persons that are mentioned in the news text. The term **sources** refers to the subset of people who are quoted within a news article. We use Named Entity Recognition (NER), a commonly used procedure in NLP, to identify the mentioned people in each article. Current NER techniques, largely statistical in nature and based on neural network architectures, work fairly well on our French data, although we observed that they are less reliable than the techniques in our prior work on English [10]. We believe that the extensive efforts we put in to improve the recognition of people’s names may be useful for other researchers, and thus provide detailed information in an Appendix (see footnote 4).

### 3.3. Coreference resolution

A coreference resolution algorithm takes in text and returns clusters of text spans that refer to the same entity. For our use case, these text spans are either pronouns, partial or full names of people, or noun phrases that refer to someone without naming them, for example, *la comptable* (‘the [female] accountant’). To apply coreference resolution, we use the **coreferee** library, including a Rules Analyzer for French built for this project. Processing the data via **coreferee** involves several steps, detailed in the subsections below.

#### 3.3.1. Coreference clusters

The output of coreference resolution is a set of indexed clusters of mentions. A ‘mention’ is a list of tokens (typically one) that correspond to the syntactic heads of noun (or pronoun) phrases. **coreferee** produces coreference ‘chains’ for each mention, and for each chain we create a coreference ‘cluster’. The cluster is an array of mentions’ character spans where each list corresponds to one mention, and the character spans consist of the beginning and end index of a sequence of characters, as shown in Example (2).

(2) Selon nos informations, M. Legault a commencé à ressentir les premiers symptômes durant le trajet de Québec vers Montréal, jeudi, après la période de questions à l’Assemblée nationale. «Ce sont des symptômes apparentés à un rhume», affirme son directeur des relations médias, Manuel Dionne. Un test rapide s’est révélé positif et Legault a annoncé sur Twitter en début de soirée qu’il se plaçait en isolement, même s’il assure qu’il se sent «bien».

**Coreference chains:**

0: M.(4), son(46), Legault(64), il(74), se(75), se(86)

1: directeur(47), Manuel(52)

#### 3.3.2. Mention-entity alignment

Next, we align each of these clusters from the coreference algorithm to the named entities extracted in the named entity recognition step. To perform alignment, for each named entity and cluster pair, we search if the named entity has span coverage (i.e., overlaps)

4https://arxiv.org/abs/2304.09982
with all the heads of a mention belonging to the cluster. If so, we infer that the coreference cluster contains mentions of that named entity. Targeting the heads and not the whole noun phrase allows us to exclude entities that overlap with the noun phrase, but not with the head, dealing with several alignment problems. Named entities that cannot be assigned to another existing mention are considered singletons. In the example below, Christine Boyle and Jean Swanson are mentioned only once, and are hence considered singletons.

3.3.3. Entity unification

Once we align named entities to their mention clusters, we still need to merge some clusters, because the coreference resolution algorithm is not perfect—it does not directly combine all mentions of a named entity across a document, especially if the entity is mentioned in two places that are far from each other within the text. Here, we apply some domain knowledge about human names in order to find clusters pointing to the same person.

First, we extract different potential components of a name: the first name, the middle name(s), and the last name. The middle names can correspond to one of the following: (a) cases where there is no clear last name (typically with names from Middle Eastern or Latin American origin, such as Mohammed bin Salman Al Saud, or Andrés Manuel López Obrador; the last part of the first name; (b) composite middle names (e.g., Ursula von der Leyen); or (c) the spouse’s name.

Our name extraction system compares each part of the name individually between named entities representing different clusters. Several cases of matching are covered:

- Same last name and same first name: Due to computational cost, the coreference algorithm may not keep track of multiple mentions of a specific person in a long text. Sometimes, two full names that are exactly the same fall into separate clusters. We assume two full names that match exactly always refer to the same person, and combine their clusters.
- One shared name: One other common case is when some mentions in the text are only a first name or last name. For example, we can have three different entities aligned to three different clusters for one person: Justin Trudeau; Justin; Monsieur Trudeau; Trudeau. In that case it’s safe to assume they are the same person. There are possible cases of different people sharing one name but not the other. This typically happens with people of the same family, e.g., Sophie Grégoire Trudeau.

When either the first name or last name is shared but the other one is different, we infer that the two entities are different people. A possible exception is when two different people who share a last name are mentioned with only their last name. Those cases, however, are relatively rare, since journalists tend to use full names to avoid ambiguity in their writing.

When merging two clusters, we try to preserve the named entity mention that is deemed the most representative. This helps us prioritize the full-name representations for the cluster, which is useful for the gender prediction step. The priority order also uses the extracted name parts and is fairly straightforward:

- Presence of both first name and last name is more representative
- Presence of last name is more representative
- Presence of first name is more representative
- More middle names is more representative

After these steps, we come up with a unique cluster for each person containing all mentions in the text referring to that person, represented by one full name. We then move on to the next step of mapping the extracted quotes to the names of people who said them.

3.4. Mapping quote speakers to their references

To identify the name and gender of the sources (quoted speakers), we find the corresponding named entity for each extracted quote (the reference), in three successive steps.
First, we compare the speaker index field of a quote to the indices of each named-entity-mention in our unified coreference clusters. If a mention span (rather, all its heads) and a speaker span have two or more overlapping characters, we assume the mention is the speaker and attribute the quote to the unified named entity (coreference cluster) of the mention.

After trying to align all quotation speakers with potential named entities, there may still remain some quotes with speakers that could not be matched to named entities. We check if the text of the speaker matches with one of the text representatives of the clusters. If a match is found, the corresponding entity’s representative is assigned. This can happen when the named entity recognition step failed to identify a name somewhere in the text, but managed to identify the same name somewhere else (because of different contextual clues).

As a last resort, we use a custom Rules Analyzer to check if the entity is a potential introducing common noun, i.e., a noun referring to a person that is not necessarily named anywhere in the text before. This is typically the case in nouns with an indefinite article (e.g., ‘a nurse’). In those cases, the text of the speaker is assigned as the reference. Failing all the methods above, our current system ignores the speaker and leaves the quote without a reference. There are several categories of these cases, such as quotes with a pronoun speaker (e.g., she said) where the pronoun is still a singleton after all named entity and coreference cluster merging steps are finished. We provide statistics in Section 4.

3.5. Gender prediction

We do acknowledge that gender is non-binary, that there are different social, cultural, and linguistic conceptualizations of gender, and that a binary is an oversimplification of the reality of gender in the population. For this project, however, we rely on self- and other-identification of gender through names and pronouns in order to classify people mentioned and quoted as women, men, or other. This is largely due to our use of external gender services that only offer binary classification of names.

Our approach to assigning gender to a first or full name was inspired by our prior work on English [10], albeit with an extra step: title-based gender prediction, where we gain useful context from a potential title associated with the named entity (e.g., Monsieur Justin Trudeau). This is because titles are more extensively gendered in French than in English.

For gender prediction, we utilize external gender services, which can be divided into two main groups: services that use only the first names (assigning gender based on statistics for those names from historical data) and services that rely on the full name. We use the services on a daily basis, and then sometimes correct errors manually as we encounter them. Genderize and Gender-API5 are two such services.

3.6. Gender annotation

Based on all the previous steps, we produce a structured object for each news article with fields in Table 1, with people-related fields on the top half and fields for each quote on the bottom (see also the Appendix in footnote 4 for an example of the quote fields).

4. Evaluation

Evaluation was carried out separately for each component (people and source extraction, gender prediction, and quotation extraction) several times over the course of this work to test out new ideas that enhance the system. In this section, we explain our evaluation methodology and our current system’s performance vs. human-annotated data.

4.1. Manual annotation for French articles

We prepared a collection of human-annotated gold samples to compare against our system’s predicted quotes, people, and sources. We selected a sample of nine articles from each of the six newspapers’ websites we scrape (Journal de Montréal, La Presse, Le Devoir, Le Droit, Radio-Canada, TVA Nouvelles), for a total of 54 articles.

Table 1. People and quote fields produced by the NLP system

We had one annotator who was fluent in French and English do the entire annotation. We began by using the English annotation guidelines, and extended the guidelines with French examples using feedback from the annotator, especially for cases that did not have parallels in English. For each article, we have a JSON file which contains an array of extracted quotes, verbs, and speakers, together with their character span index in the text (see the Appendix, footnote 4), where the speaker, quote, and verb are identified by their position in the file. The annotation also added each person named in the articles as well as their gender.

4.2. Quote evaluation

We evaluate the output of the system by comparing it to the human annotations. First, we align the annotations with the extracted quotes. Let $q_a$ be the span of annotated quote and $q_e$ be the span of the extracted quote. The match between the two quotes is defined as:

$$\text{score} = \frac{\text{len}(q_a \cap q_e)}{\text{len}(q_a)}$$

For each annotated quote $q_a$, the best matching quote in extracted quotes is the one with the highest matching score if the score is above a certain threshold (we experimented with 0.3 and 0.8 as easy and hard thresholds, respectively). In the following text example, the human annotated and automatically extracted quote spans are highlighted using italic and underlined text, respectively: “It’s premature for us to make any sort of pronouncement about that right now, but I can tell you this thing looks and smells like a death penalty case.”

The alignment score is 0.45, which is the ratio of the length of the overlapping portion (69 characters) to the overall length of the annotated span (153 characters). Thus, this quote would be a match for the 0.3 threshold, but not for the 0.8 threshold.

We evaluate the identification of speakers and verbs independently. We compute recall, precision and F1 score for the two, considering whether they are linked to the correct quote. That way, we can assess how many of the speakers will be identified for the next steps.

Table 2 shows the result of evaluating the quotation extraction code on the manually-annotated dataset. The first three columns reflect how well the system captures the quotation content span (according to each of the set thresholds of overlap, 0.3 and 0.8) and the
last two columns show system accuracy on verb and speaker detection. We consider the verb to have been correctly detected if the verb extracted by the system has exactly the same span as the expert-annotated span for the verb of that quotation. In order to evaluate the speaker detection quality at the surface textual level, we apply a simple overlap threshold: if the system-annotated span for the speaker has at least one character overlap with the expert-annotated text span for the speaker, it will be accepted as a correct annotation. For example, if the system-annotated span was [12:25], corresponding to the string Valérie Plante, while the human-annotated span was [20:25], corresponding to the string Mme Plante, the span overlap of 5 characters would mean they were considered the same speaker. Verb and speaker evaluations are applied only to the matched quotes (the quotations that are already passed as aligned between system and expert annotations based on the content span overlap). That is why the accuracy scores for Verb and Speaker in the table were higher when we used a stricter quote matching technique (hard match threshold).

<table>
<thead>
<tr>
<th>Quotation content</th>
<th>Speakers/verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Easy match threshold (0.3)</td>
<td>84.0%</td>
</tr>
<tr>
<td>Hard match threshold (0.8)</td>
<td>74.6%</td>
</tr>
</tbody>
</table>

*Table 2. Quote extraction evaluation based on manually-annotated data*

The evaluation of speaker identification when taken independently (Table 3) shows that almost 4 speakers out of 5 are correctly identified. The missing speakers, however, will impact the next step. Although the verbs are not used anymore downstream in the pipeline, their identification is encouraging and could be useful for other applications.

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers (independently)</td>
<td>79.1%</td>
<td>78.8%</td>
</tr>
<tr>
<td>Verbs (independently)</td>
<td>84.5%</td>
<td>83.8%</td>
</tr>
</tbody>
</table>

*Table 3. Speaker and verb evaluation*

**4.3. Mapping speakers of quotes to their references**

Mapping the speakers of the quotes to their respective most representative person named entity is a crucial task that we call Quote merging. Quote merging can be constructed as an open classification task (i.e., with an unlimited number of classes). Given a set of speakers, the purpose of the quote merger is to assign to each speaker its correct corresponding reference, i.e., the most representative mention of that entity in the same document.

The evaluation of the quote merger considers a corpus of documents annotated with their speakers (speakers-GOLD) and corresponding references (references-GOLD). speakers-GOLD is given as input to the quote merger and the system will predict references-SYS. We then compare references-SYS and references-GOLD to obtain the performance scores as follows: A reference for a given speaker is considered correct (and thus counted as a true positive) if after lowercasing, the Levenshtein distance between references-GOLD and references-SYS is less than 2 (so as to take typos into account). More simply, it can be said that a reference is counted as correct if for a given speakers-GOLD:

\[
\text{referenceGOLD} = \text{referenceSYS}
\]

(4.2)
The correct references ($\text{CorrectReferences}$) are counted across all documents. In conjunction with this, we also count the total number of references given by the system ($\text{SystemReferenceCount}$) and the total number of references in the GOLD annotation ($\text{GOLDReferenceCount}$). Those two numbers may differ since the system does not necessarily assign a reference to each speaker. In the end we compute the following metrics:

$$ Recall = \frac{\text{CorrectReferences}}{\text{GOLDReferenceCount}} $$  \hfill (4.3)

$$ Precision = \frac{\text{CorrectReferences}}{\text{SystemReferenceCount}} $$ \hfill (4.4)

$$ F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} $$ \hfill (4.5)

The recall and precision are an average of all the speaker/reference pairs across documents, not an average of the speaker/reference pairs per document. Since each speaker/reference pair is weighted equally in the final score, documents containing more speakers contribute more to the final score than documents with few speakers.

It is important to note that, since the quote merger relies heavily on the output of the preceding step, the evaluation of the quote merger can also be considered as a good insight into the performance of the named entity recognition stage.

We reach good precision with our quote merger: a large majority of assigned references are correct, as seen in Table 4. The few precision errors come from coreference errors where pronouns were wrongly assigned to other people of the same gender. Since personal pronouns in French are often gender dependent, such errors do not impact the gender ratio.

<table>
<thead>
<tr>
<th>Speaker reference</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90.3%</td>
<td>66.6%</td>
<td>76.7%</td>
</tr>
</tbody>
</table>

$\text{Table 4. Speaker reference evaluation}$

The recall, although lower than the precision, is also acceptable for our purposes: Less than 25% of the speakers have a missing reference. Based on visual inspection of the results, there are several cases that explain the misses:

- A speaker pronoun is wrongly assigned a non-person reference during the coreference resolution. This can happen when pronouns have contexts that are compatible with different types of entities, for instance, between those of type ‘Person’ and ‘Organization’ (because organizations are often personified in speech).
- The speaker is further away from the preceding coreferring mention than is allowed by coreferee. In our implementation of the coreference algorithm for French, we specify a ‘distance limit’ (5 sentences for pronouns and 3 sentences for nouns), beyond which the chains will not necessarily be merged. Thus, some chains containing speakers do not contain named speakers, because the speaker was named earlier than the distance limit.
- Some rarer names are never identified as person-type named entities by spaCy’s NER model. As a consequence, they are not included in the list of potential references and can never be associated with a speaker.

4.4. Gender annotation
4.4.1. Entire pipeline

This step measures how many of the people mentioned in the text were correctly detected by our system and how many were missed. According to the human annotation instructions, the most complete name of each person in the text needs to be provided by the annotators in the annotation files.
Using these manually-annotated sets, we can calculate the number of entities our system detects and misses. We first convert all system- and expert-annotated entities in these lists to lowercase and trim the start/end space characters. We consider the human annotated sets (true_names) and the sets produced by the system (pred_names) and perform set operations to obtain:

- **True Positives**: true_names ∩ pred_names
- **False Positives**: pred_names/true_names
- **False Negatives**: true_names/pred_names

Then we sum up those over all the articles to calculate the precision, recall, and F1-score of each identification task. The evaluation of the entire pipeline helps us to assess the performance of all the previously evaluated components when put together. Evaluating all components of the pipeline independently also allows us to assess the impact of each part of the pipeline on the resulting gender annotation.

As we see in Table 5, we reach an excellent F1 for people mentioned and thus an excellent F1 for women mentioned and men mentioned (which is expected given the high performance of gender prediction). Precision is slightly better than people which reflects the performance of both the named entity recognizer and the boundary fixes performed afterwards: almost all named entities recognized as people are indeed people, and few people are missed. This also reflects the quality of entity unification: Nearly all people appear to be correctly merged. Note that there were no non-binary or unknown people in the evaluation set.

The performance for identification of sources is foreshadowed by the numbers obtained when evaluating quote merging. Precision is much better than recall and the latter is still reliable enough to detect the majority of speakers. Interestingly enough, both precision and recall for sources are higher than the multiplication of the metrics for quote merging and speaker extraction. This shows that some errors in the speaker extraction and quote merging can actually be helpful, by catching a source that was not identified by the correctly identified speakers and references. For instance, a speaker il can be incorrectly marked as a speaker and then said speaker is correctly mapped to the reference Tom Lanneau. Tom Lanneau is actually a reference to the speaker le professeur elsewhere in the text, but the coreference algorithm did not correctly resolve coreference for this occurrence. As a consequence, Tom Lanneau would be correctly included in the sources by accident.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>People mentioned</td>
<td>96.4%</td>
<td>88.7%</td>
<td>92.4%</td>
</tr>
<tr>
<td>Women mentioned</td>
<td>89.9%</td>
<td>84.9%</td>
<td>87.3%</td>
</tr>
<tr>
<td>Men mentioned</td>
<td>96.7%</td>
<td>88.1%</td>
<td>92.2%</td>
</tr>
<tr>
<td>Unknown People Sources</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Sources</td>
<td>93.9%</td>
<td>57.1%</td>
<td>71.0%</td>
</tr>
<tr>
<td>Women Sources</td>
<td>83.3</td>
<td>50.0%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Men Sources</td>
<td>96.0%</td>
<td>58.7%</td>
<td>72.8%</td>
</tr>
<tr>
<td>Unknown Sources</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 5. Evaluation of gender prediction in the entire pipeline

4.4.2. Gender ratio

We compared the gender ratio, which is the ratio of women and men sources predicted by our system to those of our human annotations, with results in Table 6. We aim for the system ratio to be the closest possible to the actual (manually-annotated) ratio. The ratios for both the people mentioned and sources quoted are very close to the actual ratios. Most of the classification errors are explained by the algorithm wrongly categorizing women
mentioned as men. This is a common problem in French, because some names are associated with a different gender in English and French, and our system uses the same lookup table for English and French. An example is the first name Jean, typically female in English, but typically male in French, as Jéan, potentially explaining why our system overshoots the human annotations for the gender ratio of women mentioned. Overall, we observe that our gender prediction algorithm is able to deliver rather accurate results.

Judging by these metrics, any improvement to this stage should come from speaker identification (part of quote extraction) and coreference resolution (part of quote merging).

<table>
<thead>
<tr>
<th>People</th>
<th>Men</th>
<th>Women</th>
<th>Unknown/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human annotation</td>
<td>73.5%</td>
<td>26.5%</td>
<td>0.0%</td>
</tr>
<tr>
<td>System</td>
<td>72.7%</td>
<td>27.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sources</th>
<th>Men</th>
<th>Women</th>
<th>Unknown/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human annotation</td>
<td>75.3%</td>
<td>24.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>System</td>
<td>75.5%</td>
<td>24.5%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 6. Evaluation of gender ratio

5. The gender representation picture in French Canadian media

We analyzed 15 months of news stories in French Canadian media, between October 1, 2021 and December 31, 2022, for a total of 282,512 news stories. The overall gender breakdown in these news articles is: 71.5% men quoted, 28.3% women quoted, and 0.2% unknown or non-binary people quoted. Table 7 shows a breakdown of these numbers per news organization.

<table>
<thead>
<tr>
<th>Organization</th>
<th>% Men</th>
<th>% Women</th>
<th>% Unknown/Other</th>
<th>Total articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Le Journal de Montréal</td>
<td>74.6%</td>
<td>25.2%</td>
<td>0.2%</td>
<td>60,300</td>
</tr>
<tr>
<td>La Presse</td>
<td>72.2%</td>
<td>27.6%</td>
<td>0.2%</td>
<td>61,674</td>
</tr>
<tr>
<td>Le Devoir</td>
<td>71.6%</td>
<td>28.2%</td>
<td>0.1%</td>
<td>35,706</td>
</tr>
<tr>
<td>Le Droit</td>
<td>71.8%</td>
<td>28.1%</td>
<td>0.1%</td>
<td>31,623</td>
</tr>
<tr>
<td>Radio Canada</td>
<td>66.9%</td>
<td>32.9%</td>
<td>0.1%</td>
<td>65,280</td>
</tr>
<tr>
<td>TVA Nouvelles</td>
<td>71.8%</td>
<td>28.0%</td>
<td>0.2%</td>
<td>27,929</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>71.5%</strong></td>
<td><strong>28.3%</strong></td>
<td><strong>0.2%</strong></td>
<td><strong>282,512</strong></td>
</tr>
</tbody>
</table>

Table 7. Percentages of sources per outlet, October 1, 2021 to December 31, 2022

We also wanted to explore in which capacity the most frequent sources are quoted, i.e., whether we hear more from politicians, celebrities, or experts. We extracted the 100 top men and women sources for each of the 15 months we studied, and manually annotated the profession of each of those sources. We arrived at a list of top professions or categories, based on existing work [16], as the most typical categories of those quoted in the news. The results are presented in Table 8.

There are some interesting observations we can derive from the results. The most quoted persons are politicians, for both men and women. Sports and Unelected government officials are second and third for men, and third and second for women. A significant difference in gender is in Health profession, where a much higher percentage of women are quoted. There
are more errors or unknowns for women, for two reasons. Many names of men end up being
categorized as women. The second reason is that names are not unique enough to help us
decide why the person was quoted (and are thus classified as Unknown). When we search
for Isabelle Côte or Anna Walker, it is unclear which individual those names refer to, since
there may be several people with those names in the news.

<table>
<thead>
<tr>
<th>Profession</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politician</td>
<td>160,581</td>
<td>44,059</td>
</tr>
<tr>
<td>Sports</td>
<td>19,071</td>
<td>2,611</td>
</tr>
<tr>
<td>Unelected govt. official</td>
<td>11,795</td>
<td>8,049</td>
</tr>
<tr>
<td>Health profession</td>
<td>13,205</td>
<td>21,553</td>
</tr>
<tr>
<td>Leader (union, school, activist)</td>
<td>3,447</td>
<td>2,207</td>
</tr>
<tr>
<td>Police</td>
<td>2,201</td>
<td>2,031</td>
</tr>
<tr>
<td>Private business</td>
<td>2,124</td>
<td>2,144</td>
</tr>
<tr>
<td>Legal profession</td>
<td>2,251</td>
<td>1,620</td>
</tr>
<tr>
<td>Creative industries</td>
<td>2,479</td>
<td>1,610</td>
</tr>
<tr>
<td>Perpetrator</td>
<td>1,273</td>
<td>360</td>
</tr>
<tr>
<td>Academic/researcher</td>
<td>1,602</td>
<td>1,716</td>
</tr>
<tr>
<td>Victim/witness</td>
<td>869</td>
<td>1,775</td>
</tr>
<tr>
<td>Media</td>
<td>628</td>
<td>427</td>
</tr>
<tr>
<td>Non-govt. organization</td>
<td>430</td>
<td>1,186</td>
</tr>
<tr>
<td>Error/unknown</td>
<td>300</td>
<td>728</td>
</tr>
<tr>
<td>Person on the street interviews</td>
<td>0</td>
<td>225</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>222,256</strong></td>
<td><strong>92,301</strong></td>
</tr>
</tbody>
</table>

*Table 8. Top 100 men and women sources, by category, in each of the 15 months between
October 1, 2021 and December 31, 2022*

6. Discussion and conclusions

We present an NLP system based on linguistic-motivated heuristics, built on top of a
robust open-source framework, *spaCy*. Our contributions can be summarized as: First, we
defend the need for NLP systems in applied contexts that rely on linguistic analyses. Second,
we contribute a robust coreference resolution methodology for French text, thoroughly tested
and deployed in a live environment.

The success of large language models and end-to-end-systems in NLP seems to have
relegated linguistic-based NLP to a second plane. In contexts where it is impossible to
annotate enough data to reach acceptable accuracy, and where a pipeline approach seems
reasonably accurate, we advocate for the latter. It is concerning, however, that recent
research advancing progress in NLP (POS tagging, parsing, coreference resolution, discourse
parsing, semantic analysis) is diminishing in its generality and that flexible, open-source
systems that generalize across languages are not widely available for languages beyond
English. Our work shows that linguistics-driven approaches can be successful, robust, and
produce reliable results at scale. As Richard Sproat suggests, [17], sometimes ‘boring’
problems require traditional solutions.

We also contributed to applied NLP by extending the *coreferee* library to support
French, a capability that did not exist before we began this work.
A final contribution is our visualization and dissemination of the statistics of gender representation in French media to the general public, through the Radar de Parité. Our results and analyses show a persistent underrepresentation of women and a practical absence of non-binary people in Canadian French language media.

Acknowledgements

The Radar de Parité and its English counterpart, the Gender Gap Tracker, are a large team effort. We thank especially members of the Research Computing Group at Simon Fraser University, present and past, who maintain the database and support data analyses: Jillian Anderson, Philip Chen, Alexandre Lopes. This research was initiated and supported by Informed Opinions, a non-profit dedicated to amplifying the voices of women and gender-diverse people. Funding support: Informed Opinions, Simon Fraser University, Social Sciences and Humanities Research Council, Natural Sciences and Engineering Research Council.

References