# Linguistics 334 Final Project

**Marisa Gudiño** and **Emran Majidy**
Northwestern University

## Abstract

This paper is a partial replication of Fu et al. (2016) which used language models to quantify gender bias in sports journalism. This paper aims to replicate the first portion of the investigation which involves constructing a bigram model based on match commentary and calculating the perplexity with respect to this game language model. Like Fu et al. (2016), we find evidence of gender bias in sports journalism. We also discuss the efficacy of the KenLM language model (Heafield, 2011) in comparison to the simpler model featured in A2.

## 1 Introduction

In recent times, there has been increased conversation surrounding the issue of media bias and representation based on gener in sports media. An article in Nielsen (de Armas, 2021) notes that "sexism in sports is ingrained from the time our children are in youth sports." In fact, in 2023, only one woman earned a spot in Forbes' 50 highest paid athletes (Knight, 2023), Serena Williams at No. 49. Indeed, Fu et al. (2016) found issues of systemic bias against female tennis players in their investigation and we aim to partially reproduce this using LING 334 Assignment 2. Though Fu et al. used the KenLm language model toolkit (https://kheafield.com/code/kenlm/) to produce their results, we aim to compare the outputs of both language modeling frameworks in their efficacy of capturing bias.

## 2 Data

In order to best compare the effectiveness of the language models, the same data sets were used in each respective computation. Transcripts of live match commentary were used as a training corpus for the bigram model, and sourced from Sportsmole UK. Question data was used for the perplexity computation, and contains processed questions that players were asked in post-game interviews, covering 6467 post-match press conferences. This data was sourced from ASAP Sports. The data used was collected and converted into a .JSON format by Fu et al. (2016).

| **Sample Game Commentary** |
|---|
| "Keys seeks to push Radwanska a little harder and she wins out as her polish opponent makes a mistake trying to return on the forehand as the ball flies out of the court." |

Table 1: Game commentary acts as the training corpus. It is manipulated to contain one sentence per line, already tokenized, so it can be split it up on white-space.

### 2.1 Game Commentary

The aim of the Fu et al. (2016) investigation is to determine whether there is bias in sports journalism based on gender. This was determined by assessing the amount of sport-related material in the questions posed to male and female players in their post-game conferences. In order to create a baseline for comparison, Sportsmole UK[1] match commentary was used. 3962 pieces of live-text play-by-play commentaries, evenly split between male and female singles players were collected. Each commentary also included two other .JSON fields: gender and scoreline. Gender refers to the gender of the player ('F' for women and 'M' for men). Scoreline refers to the match's score when the the text update was posted, where '*' indicates the serving player. The commentaries are short, averaging around 40 words. There is a sample of game commentary in Table 1.

### 2.2 Question Data

This data set contains transcripts of post-match commentary for men's singles and women's singles tennis players. The data is made up of " 6467

[1]This game commentary is originally from http://www.sportsmole.co.uk/.

| Sample Interview Questions |
|---|
| "Surely you 're not a one direction fan ?" |
| "Have you played Novak ?" |
| "Has what you have achieved sunk in yet ?" |
| "What did your mom say ?" |

Table 2: Here is a sample of interview questions posed to male and female players. The above questions are merely a sample and not intended to be representative of the results.

interview transcripts and a total of 81906 question snippets posed to 167 female players and 191 male players," (Fu et al., 2016). These questions are not posed by single reporters in post-match conferences, so the data is a consolidation of questions from various sources. Duplicate questions were removed and the remaining question snippets were processed by Fu et al. (2016). The .JSON file contains five fields: gender, player, questions, ranking, and result. The 'gender' field denotes the player gender identity and 'player' indicates the name of the player being interviewed. The 'questions' field contains a list of question snippets, where each entry represents one turn from one reporter. The 'ranking' field indicates the ranking of the player at the time of the conference, and the 'result' field indicates the result of the match; '1' denotes a win and '0' denotes a loss. Full interview transcript data was available and used in (Fu et al., 2016), however, it was not made use of in this investigation. There is a sample of interview questions in Table 2.

## 3 Methods

This investigation aims to compare the methods for bias detection used in (Fu et al., 2016) and in the A2 for Linguistics 344. (Fu et al., 2016) made use of the Ken LM language modeling toolkit to train the bigram model and calculate perplexity values. Unlike the previous investigation, we decided not to analyze the effects of player ranking, and match outcome on the gender outcomes. This was because (Fu et al., 2016) did not find evidence of the statistical significance of these factors with respect to question bias. We also did not replicate the computation of question typicality, because although there existed a statistically significant change in magnitude of perplexity, the overall trend was the same between typical and atypical questions.

$$\text{perplexity}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

Figure 1: Equation describing perplexity calculation for bigram language models.

### 3.1 Ken LM Library

The Ken LM toolkit is a python library used to perform interpolated modified Kneser Ney Smoothing for estimating the n-gram probabilities[2]. It significantly reduces time and memory costs associated with performing computations on large amount of data compared to alternative methods (Heafield, 2011). (Fu et al., 2016) explains their process; "As a preliminary step, we apply a word-level analysis to understand if there appear to be differences in word usage when journalists interview male players compared to female players. We then introduce our method for quantifying the degree to which a question is game-related, which we will use to explore gender differences." That is, the investigators considered how often a word is used in a question, and then identified the words with the greatest percent difference in usage by gender. Then, these words were used to train the bigram language model in order to quantify how game-related a question is.

### 3.2 A2 Language Model

The A2 language model is much simpler in that it manually computes bigram and perplexity probabilities. First, we processed the data, tokenizing it on white-space. Then, we trained the bigram language model on the set of gender-balanced live text play-by-play game commentaries. As in (Fu et al., 2016), for an individual question q, we measured its perplexity with respect to the game commentary language model as an indication of how game-related the question was. Perplexity was calculated using the equation in *Speech and Language Processing*, (Jurafsky, 2023), Fig. 1.

## 4 Results

(Fu et al., 2016) found that when comparing perplexity values between male and female player groups, they found that the mean perplexity of questions posed to male players was significantly smaller (p-value <0.001) than that of questions posed to female players. This suggests that the questions male athletes receive are more game-

[2]https://github.com/kmario23/KenLM-training

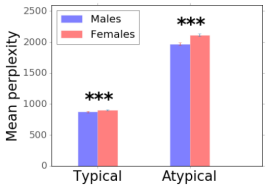


Figure 2: The mean perplexity of the bigram language model based on gender and typicality of questions. We did not replicate the typicality calculations, though (Fu et al., 2016) did find its effects to be salient.

related. Fig. 2 shows a visualization of the difference in mean perplexity.

Our computation of the same data revealed similar trends. We calculated that questions for male-identifying players had a perplexity score of 623 and questions for female-identifying players had a perplexity score of 1143, with respect to the live game commentary data set. In fact, the difference we calculated was much more dramatic compared to (Fu et al., 2016), with our computation of female-question-perplexity being nearly double that of male-question-perplexity, with respect to the game commentary.

## 5 Discussion

In this work, we aimed to compare the efficacy of the computational methods of (Fu et al., 2016) and the Linguistics 344 A2. Though we made the same conclusions overall using the same data, it is interesting to note that the differences in computational methods resulted in a different depth of results. The Ken LM toolkit used in (Fu et al., 2016) allowed the researchers to use interpolation[3] and smoothing effects when computing the bigram models for each group. The effects of interpolation, or a a method of constructing new data points based on the range of a discrete set of known data points, in addition to K-smoothing, may have worked to minimize the differences in perplexity among the two groups. While we also made use of K-smoothing when constructing the bigram language model, we did not use interpolation. This could explain the disparity in magnitude of the gender-based differences in perplexity. Furthermore, it in interesting to consider that in other instances, depending on the computational methods used, one may come to different conclusions about the significance of certain biases. For instance, in our case, it could be that the simpler method of estimation fails to adequately explain smaller data sets, as outliers may skew the results. Indeed, this work is extremely limited in scope (focused on game-related language) and used much less advanced tools than the KenLM toolkit. Still, there may be instances where the use of interpolation disguises the magnitude of certain biases. Indeed the calculations in (Fu et al., 2016) revealed that though statistically significant, the difference in perplexity was not nearly as much as calculated in the A2 method. Perhaps in the future, we will be able to distinguish the appropriate use cases for each method, whether that be in education or otherwise.

[3]https://github.com/kmario23/KenLM-training

## References

Stacie de Armas. 2021. On different playing fields: The case for gender equity in sports. *Nielsen*.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *Proceedings of the IJCAI workshop on NLP meets Journalism*.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries.

Dan Jurafsky. 2023. *Speech and Language Processing, 3rd Ed.*

Brett Knight. 2023. Why only one woman made the ranks of the world's 50 highest-paid athletes. *Forbes*.

Richard Lapchick. 2021. Sports media remains overwhelmingly white and male, study finds. *ESPN*.

Talya Minsberg. 2021. When gender equality at the olympics is not so equal. *New York Times*.