Gender bias in Video Game dialogue: Materials and Methods

Stephanie Rennick, Seán G. Roberts, Melanie Clinton, E. I., Liana Oh, Charlotte Clooney, Edward Healy

1 Introduction

This set of supplementary materials describes each step of the creation and analysis of the Video Games Dialogue Corpus. Parts of it are written in Rmarkdown, which includes R code that implements an analysis, the output of that code, and plain commentary and explanations. This supporting materials document itself is compiled from the output of several documents. Each individual document and the original data and R code are available in the accompanying online repository.

The rest of this introduction gives a broad overview of the methods.

Sample

A sample of 50 video games was selected in the Role-Playing Game (RPG) genre where dialogue was a major game mechanic. The sample was chosen to be representative of several characteristics such as: publication date (balanced from 1986 to 2020), style ('Western' vs. 'Japanese' RPGs), and target audience (rated for 'Everybody', 'Teen' and 'Adult' by the Entertainment Software Rating Board). The games included ones developed by large companies (e.g. franchises such as *Final Fantasy, Persona, Mass Effect, Dragon Age, The Elder Scrolls, Kingdom Hearts*) and from smaller developers (e.g. *Monkey Island, Stardew Valley*). All games either individually sold, or belong to series that sold, at least 1 million copies worldwide. Dialogue for each game was located from a range of sources including data directly from the game code and public websites such as wikis and fan-made transcripts. See SI 1.2 for details.

Script parsing

For each game script, a custom python program was written which scraped and parsed the script into a common format. This parser used systematic pattern recognition, but also applied specific manual edits listed in the metadata files. There were approximately 20,000 manual edits applied to the corpus, mostly fixing mappings between character names and lines of dialogue. The scraping and parsing programs are available in an online repository alongside programs for calculating all the measures and statistics presented in this paper (`https://osf.io/b2qcg/?view_only=c194016e73544b60b57bccff453dd93a`).

The game script format represented lines of dialogue paired with the name of the character who spoke them, as well as actions and changes in location. The format used a recursive JSON structure in order to represent dialogue trees common in games.

The game scripts were validated with a systematic error-checking procedure (see SI 1.2). Transcription errors in the source were identified by finding a video of the game being played, choosing random dialogue in the video, and checking that this dialogue appears accurately in the corpus. Parsing errors from the automatic parsers were identified by manually checking random lines of dialogue. For each game, 15 lines were checked for transcription errors and 5 lines were checked for parsing errors. Any errors were raised as issues on the GitHub repository, and fixed. After this, a second round of error checking and fixing was conducted following the same steps as above.

Gender Coding

Conferred gender of characters was coded manually according to a set of defeasible indicators, as discussed above and in more detail in SI 1.3. The coding scheme did not assume binary gender. Evidence for edge cases is documented in the corpus repository. Where there was insufficient evidence for a character's gender, they were labelled as "neutral" (around 7.6% of characters).

To establish the reliability of gender coding, a sample of characters was coded by a secondary coder. For each game, 10 characters were randomly chosen with the probability of being chosen being in proportion to the amount of dialogue they spoke. Agreement between coders was 'almost perfect' (Landis & Koch, 1977, raw agreement = 96%, Cohen's kappa = 0.92 [0.89, 0.96], see SI 1.15).

Measures

The measures of dialogue length and readability were obtained using the python module *textatistic* (https://pypi.org/project/textatistic/) that was designed for looking at gender differences in large text corpora (Hengel, 2020). Length of dialogue was measured in number of words, number of lines, number of sentences and number of syllables. All of these measures were correlated with each other with $r > 0.98$ (measured at the group level), indicating that the measures are robust.

Several of the games in the sample were originally written in Japanese, so there may be differences in estimates of female dialogue in the original script versus the English translation. To test this, we analysed three versions of *Chrono Trigger*: the original Japanese script and two English translations. The measures of dialogue length were highly similar between all texts (correlation between number of English words and Japanese characters per line r = 0.93) and all estimates of the proportion of female dialogue are within 0.7 percentage points of each other (see SI 1.13). This suggests that the general gender biases are not caused by translation.

Statistical methods

The aim of the statistical measures is to assess whether there is a statistically significant bias in the distribution of dialogue by gender within a game. Standard parametric tests are inappropriate because the data is highly non-independent (words belong to lines, and lines belong to coherent characters) and highly skewed (a small number of characters say a lot while a large number of characters say little). Instead, a permutation framework was used that compares an observed measure with the range of measures that would be expected if there really was no bias. This was done as follows (for more details, see SI 1.4).

The proportion of dialogue by gender is assessed in comparison to two baselines which reflect two possible sources of bias. The first source is a 'character bias' where more male characters are included in the game than female characters, which has a knock-on effect on the proportion of dialogue for each gender. A hypothetical script was generated where the mapping between gender categories and individuals is randomly determined, with each character having an independent and equal probability of being male or female. That is, the link between gender and characters is randomised to remove any potential bias. Generating 100,000 scripts created a distribution of probable values for the proportion of female dialogue if there was no bias. This was compared to the true proportion of female dialogue. This produced

a z-score that represents the strength of the bias (in number of standard deviations away from the expected mean), and a p-value that represents how likely the baseline process results in a measure that is more extreme than the observed distribution. Lower p-values indicate that the baseline model assumptions are unlikely to hold, and suggests that the imbalance in dialogue is due to the imbalance in the proportion of characters.

Similarly, the second possible source of bias is the 'dialogue bias' where the average male character is given more dialogue than the average female character. To model a baseline to test this, a hypothetical script is generated where the mapping between gender and characters remains unchanged, but the mapping between characters and lines is randomly permuted (all of one character's lines can be swapped for all of another character's lines). This maintains the same proportion of female characters as the real data, but the proportion of dialogue for female characters will vary if they are given systematically less to say than male characters. 100,000 scripts were generated to create a distribution of values, to which the true proportion of female dialogue can be compared.

References

Hengel, E. (2020) Publishing while female, CEPR Press.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. Biometrics 33(1): 159–174.