**How to code gender in the Video Game Dialogue Corpus**

The procedure is explained in [this video](https://cardiff.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=e9506a4a-bda6-4138-abcb-ace600ec2f38). But briefly:

1. Go to the corpus github page: <https://github.com/seannyD/VideoGameDialogueCorpus>
2. Open the “data” folder.
3. Find the folder for the game you want to code.

You’ll see various files that might help:

* meta.json: Meta data about the game, source, parser, and character groups.
* data.json: The dialogue data, created by the parsing program.
* characters.txt: A simple list of all unique characters, created by the parsing program.
* stats.csv: Basic stats for the game as a whole and each group in the metadata.

To edit the meta.json file:

1. Click on meta.json to see the contents
2. Click the ‘pencil’ icon in the top right.
3. Make changes to the file (you might want to copy and paste this to a file on your computer so you can work on it offline)
4. Scroll to the bottom where it says “Commit changes”
5. Write a short description of what you changed (and optionally an explanation in the larger text box)
6. Choose “Create a new branch”
7. Click “Commit changes”

These will be reviewed by me and integrated into the corpus as soon as possible.

JSON format for the “characterGroups” section of the meta.json file:



The python scripts calculate the stats for each group. You can add any character groups that you want stats for, e.g. playable characters, characters from specific ages or backgrounds, or individuals.

**Aliases**

Sometimes, a character has multiple names in the script. This can happen if:

* The character is disguised as another character (e.g. Prince Edgar is transformed into King Otar in King’s Quest VII).
* The character speaks before revealing their name (e.g. Aerith in Final Fantasy VII)
* The name is shortened (e.g. “Red” instead of “Red XIII” in Final Fantasy VII)
* There are stage directions in the name (e.g. "Cara [to Mid]" in Final Fantasy V)
* There is variation in upper case/lower case letters (e.g. “Shinra manager” and “Shinra Manager”).
* There is a typo in the script.

These issues can be fixed by adding alias information to the metadata. This is placed after the “characterGroups”. It includes a list of ‘wrong’ names and what they should be corrected to. E.g. below all instances of “Flower girl” are converted to “Aerith”.



These changes apply during parsing, so the incorrect names won’t appear in the script. Therefore, when including aliases, the names in the “characterGroups” list should reflect the **corrected** name, not the original name.

**Gender coding scheme**

Raise an issue on GitHub if there are errors in the character name, or uncertainties about the gender of a specific character. Then the group can comment.

Code all characters that have speaking parts in the script that’s included in the corpus (data.json). This might be fewer than the total number of characters in the game. The file “characters.txt” has an automatically generated list of all unique names in the script.

**Coding**

The aim is to code the **(Western) player community’s interpretation** of the gender of each character. There may be disagreement: reasons for coding decisions in edge cases should be documented in CodingNotes.txt for each game.

Any label can be used as a gender category. Letters should be lowercase. Some conventions help the unification of the data later on. If you assign a character to the male category, use the label “male” (rather than “man”, “boy” etc.). If you assign a character to the female category use the label “female” (rather than “woman”, “lady”, “girl” etc.). Characters can be coded as having no gender (e.g. “genderless”), but only with positive evidence.

You might assign characters to the “neutral” category if:

* They are supernatural or cosmic beings, or very different species for which there are no gender norms.
* The name refers to a group (e.g. “people”) or general species (e.g. “Chocobo”) which might include more than one gender.
* They have a generic name (e.g. “Guard”, “Villager”) AND their gender isn’t readily identifiable from the indicators below. This includes where it isn’t possible to identify the specific character. They are an invisible third person narrator, especially in text-based adventures. (Although sometimes the gender of the narrator can be determined and should be coded accordingly (for example where they are a named character – such as Varric in DA2 – or where there is sufficient indication of their gender from their voice/pronouns used etc.)
* The dialogue comes from the game system not tied to a character, e.g. in tutorials.

There are various indicators used to code the gender of the game characters. Some are more authoritative than others, but any individual indicator is defeasible. Usually several indicators are required to make a decision, although the community wikis serve as ‘shorthand’ for these (as they are based on visual gender signifiers, character voice, game documentation etc.)

**1. Category defined by community wiki**

Some games have community-written wikis or discussion boards with pages for each character. Some of these define the gender of the character (e.g. <https://finalfantasy.fandom.com/wiki/Cloud_Strife> under “Gender” in the info bar on the right).

The category can be debatable, or change according to the language/release of the game.

**2. Character name**

The gender of some characters is revealed in their name, e.g. “King Graham”.

**3. Appearance**

Visual appearance may provide clear cues to gender, including visual signifiers (e.g. pink bows or makeup are often restricted to female characters) or secondary sex characteristics (e.g. breasts).[[1]](#footnote-1) Some wiki pages have higher-resolution fan art for characters (though these may not be reliable).

**4. Claims that the character themselves make in dialogue**

Characters rarely announce their gender directly, but still might bring it up in conversation. E.g. in King’s Quest Chapter 1:

*Amaya: “I'm a woman of action, if you couldn't tell.”*

In Stardew valley:

*"Abigail": "Oh, it's because I'm a girl... isn't it? Ugh...”*

**5. How other game characters refer to them**

Including pronouns, but this can be unreliable and switch between translations. E.g. <https://finalfantasy.fandom.com/wiki/Quina_Quen>

**6. Pronouns used in community wikis**

In a wiki page about a character, they may be referred to by a specific pronoun that might give you a clue.

**7. Gender of the voice actor**

If the voice actor is known, this might provide a clue to the gender, but may be quite unreliable.

**If you find an example that you think is difficult to assess, we would be very interested in discussing it. Add a comment to the Teams group (might be easier for identifying who is commenting) or an issue to the github page.**

**Error checking**

If the data source is directly from the game files, then just do the check for false positives and parsing errors. Otherwise, do both of the tests below.

After you have done the checks, place the results into the metadata after the “source” entry.

"errorChecks": {

"truePositive\_numTestsDone": "5",

"truePositive\_numParsingErrors": "0",

"truePositive\_numSourceErrors": "1",

"truePositive\_notes": "One line inaccurate transcript: [EXAMPLE]",

"falsePositive\_numTestsDone": "5",

"falsePositive\_numErrors": "1",

"falsePositive\_notes": "Parsing error: no space after full stop."

}

Raise a github issue detailing any problems that could potentially be fixed (mainly be false positive errors).

**Check for true positives and transcription errors**

1. Find a video on YouTube of someone playing the game. Try to find one that documents an entire play through the game (rather than clips), without mods, and that is not a speed run or specialist run (e.g. pacifist). Typically, “let’s play” videos will be suitable.

2. If a run is split over several videos, choose one at random.

3. Choose a random place in the video. This website will help you do that: <https://correlation-machine.com/VideoGameCorpus/randomVideoLocation.html>

4. Find the next piece of dialogue. If you reach the end of the video, loop around to the beginning. Look at up to three lines of dialogue that are spoken together.

5. Search the data.json file to answer:

* Does the dialogue in the video exist in the corpus? (ignoring small errors in punctuation, capitalisation, and also ignoring typos - the question is whether the line is represented somehow)
* Is the text of the transcript of the video accurate? Note that consecutive lines spoken by the same character are collapsed into one line in the corpus.
* Is the structure of the conversation correct? (Are options defined in “CHOICE” structures? Are all options available? Does the sequence match?). Note that the dialogue in the game may be randomised, optional or status-dependent, so all lines in the corpus may not appear in the video. The question is whether the dialogue in the video is covered by the corpus.
* If there are any errors, can we identify the source?
  + Error in parsing program.
  + Error in original transcript source.

Repeat steps 2-5 for 5 parts of the video.

**Check for false positives and parsing errors**

Pick a random line in the corpus data.json file.

Confirm that:

* The character name is plausible (not some possible parsing error like “and so”).
* There are no strange typographic characters.
* There are no obvious parsing errors (e.g. another character’s dialogue line enclosed in the dialogue string, words not separated properly).

If it is possible, find this line in the source transcript. This might involve finding the location of the source in the meta.json file, then using a google search like:

site:http://www.yinza.com/Fandom/Script/ “This is a church in the”

* Confirm that the line in the source has been correctly parsed into the corpus.

Repeat this procedure 5 times.

1. This is not to suggest that all and only females wear makeup, or have specific secondary sex characteristics. uenTheir use in a game, however, is frequently a choice by developers in order to indicate that such characters are not male. [↑](#footnote-ref-1)