The Video Games Dialogue Corpus

Stephanie Rennick, University of Glasgow,
stephanie.rennick@glasgow.ac.uk

Seán G. Roberts, Cardiff University
RobertsS55@cardiff.ac.uk

Abstract

This paper presents the Video Games Dialogue Corpus, the first large-scale, consistently coded, open source corpus of dialogue from video games. It contains over 6.2 million words of dialogue from 50 video games in the Role Playing Game (RPG) genre. This includes: games produced between 1985 and 2020; rated for children, teenagers, and adults; and in both "Western" and "Japanese" subgenres. The corpus design is described, including custom data formats for representing branching dialogue. We demonstrate the use of the corpus by comparing the dialogue of female and male characters, where we find reflections of gendered language in other media as well as patterns that seem specific to video games. We provide the source code for a "self-inflating corpus": a pipeline that obtains the data then processes and parses it into a standard format. This makes the corpus available for teaching and research purposes, providing the first such resource for empirical analysis of video game dialogue.

[The corpus repository will be released on publication. If you would like to use the corpus in your research, please contact the authors]

1. Introduction

Video games have become a central form of media, now making more money than the film industry (BBC News, 2019) and played by nearly 3 billion people worldwide (Gilbert, 2021). The traditional view of video games as amusements for teenage boys is no longer accurate, with most gamers being adults, and around half being female (ESA, 2021; ISFE, 2021; Nico Partners, 2021; The China Gamers Report, 2021; Newzoo, 2019; Pandurov, 2021). Games have also become increasingly sophisticated, with Role Playing Games (RPGs) now typically featuring character animation, voice acting and production quality on par with TV and film. As they have become more central in people's lives, they have become the focus of many social discourses including: the representation of gender, race, and sexuality (see Malkowski & Russworm, 2017; Heritage, 2021); the influence of advertising (Schmidt, 2020); the influence of violent content on crime (e.g. Markey et al., 2015, Behnke, 2021); sociolinguistic ideologies (Goorimoorthee, 2019); and the use of video games in education and second language learning (Gee, 2003; Li, 2022). Children are spending increasing amounts of time playing games, around an hour a day (Qustodio, 2020), and engaged in further ‘metagame’ activities surrounding video games (Kahila et al., 2021), making it an increasingly relevant source of language input during acquisition and language change (Eligio & Kaschak, 2020).

Despite the increasing importance of video games in society, there is currently no large-scale, consistently coded corpus of video game dialogue. Many current studies either focus on a single game (e.g. Erdur, 2022, Heritage, 2022), or a limited collection of games (Carrillo Masso, 2009; van Stegeren & Theune, 2020). Heritage (2020) presents a corpus analysis of 330,000 words of text from 10 RPG video games, but the texts are not separated into dialogue versus non-dialogue, and dialogue is not attributed to specific characters. More importantly, there is no large-scale, openly available corpus. This is a major barrier to replicable, empirical investigations.

This gap in linguistic resources is due to several challenges that video games pose for traditional corpus linguistics methods. First, there is a technical challenge to obtaining the data. Secondly, there is a legal challenge to sharing the data. Thirdly, there is the question of how to represent dialogue in an interactive medium. We discuss these in detail below.

We present the Video Games Dialogue Corpus (VGDC), which addresses each of these challenges using a “self-inflating corpus” design: an open-source pipeline of software and metadata that encodes instructions for obtaining, processing, and formatting video game scripts into a consistently coded corpus. The current corpus includes over 6.2 million words of dialogue from 50 games. It also contains around 1.3 million words of non-dialogue text providing contextual information such as action descriptions, system text, and location information. Section 2 provides the corpus design, including data sources and selection, data format, corpus pipeline, and error-checking procedure. Section 3 demonstrates the use of the corpus in a novel analysis of gendered language in video game dialogue, before a conclusion in section 4.

2. Corpus design

2.1 Data sources

Heritage (2021, p.98) lists four ways of obtaining dialogue from video games:

1. Extracting dialogue directly from game data.
2. Manual transcription of narrative permutations.
3. Using fan transcripts.
4. Using wikis and community websites

Manual transcription is very time consuming, and we have found that wikis tend to be either fan transcripts or curations of game data, so the main choice is between game data and existing fan transcripts. van Stegeren & Theune (2020) suggest that the highest quality data comes directly from game data. However, we argue that this is not always the case.

Firstly, as van Stegeren & Theune discuss, it is not always easy to access the source files since they are usually compiled to a proprietary format. For some games, like *Final Fantasy VII* (Square, 1997), the source files have been lost. However, even if the source files are available there may be considerable work to obtain the dialogue. For example, some of the scripts for the *King's Quest* series are plain text, but chunked into a context free grammar to save memory, requiring re-assembly. Furthermore, due to the practical requirements of video game production, game code is often less standardised than one might expect. For example, cutscene dialogue may be saved in a separate format from normal conversation or rendered into a video rather than game code.

Furthermore, the game code will often not have a simple mapping between the character assets and character names. For *Dragon Age 2* (BioWare, 1997), there is a sense in which there is no overt link between dialogue and speaker anywhere in the code. Instead, there is just a sequence of actions which must be matched in time to identify who is speaking (e.g. 3D model *A* moves their lips, sound file *X* plays, subtitles *Y* appear). In other games, there is only an association between speech and an ID for a 3D model. So if a character changes their appearance (e.g. changes their clothes) then their ID might change. Similarly, some characters may not be assigned names, or only have generic names (e.g. "Guard") which can conflate lines assigned to different individuals. These issues often means that the mapping between characters and lines requires manual editing. For example, for Star Wars: Knights of the Old Republic (BioWare, 2003), which was taken from van Stegeren & Theune's curation of the game data, we needed to make over 1500 manual edits to the data to unify the mapping between dialogue and character names.

More fundamentally, there is a question of how to identify the text of a video game. Like any work of art, a video game might have multiple possible readings (see Bednarek, 2015), including the text that the scriptwriter wrote, the dialogue that the voice actor recorded, the dialogue implemented in the code, and the dialogue that the player experiences. There are larger differences between these readings for video games than perhaps any other medium. For example, there may be differences between the original script and what the voice actor produces simply because the volume of dialogue means that not everything can be checked. There are famous examples of recording errors in *Oblivion* (Bethesda, 2006, e.g. here), and in *Final Fantasy XV* (Square Enix, 2016) we found dozens of lines where the subtitles do not match the recorded audio. In addition, a player is unlikely to experience all of the dialogue in a modern game for three reasons. First, some dialogue may be considered 'optional': the game may be completable without experiencing some dialogue (e.g. that associated with 'side quests' or optional areas). Secondly, some dialogue will only appear when certain conditions are met (e.g. a particular combination of party members being present, particular choices being made at character creation, or a character or player having a particular level of skill). Indeed, some conditions are impossible to meet through standard gameplay; dialogue may exist in the game code, but be unreachable by the player either due to programming errors, because the scene was cut, or as debugging utilities for the developers (e.g. in *Dragon Age Origins*, BioWare, 2009). Finally, some dialogue is conditional on player choices: the nature of branching dialogues mean that some alternative dialogue will not be experienced by a player in a single playthrough. Given that modern games can take a large amount of time to complete (e.g. according to https://howlongtobeat.com, several of the games in the corpus take over 100 hours to complete to a high level), it's likely that individual players will not experience large proportions of dialogue.

For these reasons, we argue that the advantages of game code compared to fan transcripts may not be so large as previously suggested. There are respects in which the fan transcripts can be more accurate and representative records of player experience than the game code (see similar arguments for film and television, Bonsignori, 2009: 187). Fan transcripts and wikis have been used to construct corpora in other media (e.g. Bednarek, 2018; Kybartas & Verbrugge, 2014). While there are concerns about the accuracy of transcripts, we found this to be an infrequent issue during our error-checking procedures (see section 2.7). In any case, the choice of source is usually constrained by availability, so we use a mix of fan transcripts and game data to construct our corpus.

2.2 Selection of data

van Stegeren & Theune (2020:2) suggest a video games corpus should be representative (contain "popular or well-known (commercial) games that have a substantial user base"), and diverse (games should represent the diversity within the genre). 50 games were selected for the corpus (see table 1). All games were Role Playing Games (RPGs) in which dialogue was a central mechanic. They all have high sales figures: all games either individually sold, or belong to series that sold, at least 1 million copies worldwide. Every game or series frequently features in lists of the top RPGs of all time. For example, IGN's top 100 RPGs of all time and Game Informer's top 100 RPGs of all time contain all the games in the corpus or at least one game from each series. The *King's Quest* and *Monkey Island* games sold relatively poorly on release (compared to modern games), however, both of these series are frequently discussed as being highly influential on the medium (e.g. here, here, here, here). These series are sometimes categorised as "adventure games" or "point-and-click" games, although they exhibit various features typical of RPGs, including player choices and dialogue trees (and thus they are also commonly described as 'adventure RPGs'). The final requirement was that the games needed to have an accessible source of dialogue available.

Game	Words	Game	Words	Game	Words
Final Fantasy	2763	Horizon Zero Dawn	50348	Persona 3	44447
Final Fantasy II	8689	King's Quest I: Quest for the Crown	1396	Persona 4	156753
Final Fantasy IV	17822	King's Quest II: Romancing the Throne	402	Persona 5	385817
Final Fantasy V	12408	King's Quest III: To Heir Is Human	3151	Star Wars: Knights of the Old Republic	439667
Final Fantasy VI	14750	King's Quest IV: The Perils of Rosella	1772	Stardew Valley	53870
Final Fantasy VII	100584	King's Quest VI	72074	Super Mario RPG	13755
Final Fantasy VIII	51368	King's Quest V	10833	The Elder Scrolls II: Daggerfall	43855
Final Fantasy IX	99245	King's Quest VII: The Princeless Bride	21358	The Elder Scrolls III: Morrowind	173666
Final Fantasy X	84290	King's Quest VIII	15880	The Elder Scrolls IV: Oblivion	777177
Final Fantasy X-2	31635	King's Quest Chapters	113053	The Elder Scrolls V: Skyrim	150041
Final Fantasy XII	128742	Kingdom Hearts	19847	Chrono Trigger	37982
Final Fantasy XIII	12674	Kingdom Hearts II	47843	Dragon Age: Origins	701258
Final Fantasy XIII-2	68922	Kingdom Hearts 3D: Dream Drop Distance	18484	Dragon Age 2	281116
Lightning Returns: Final Fantasy XIII	47916	Kingdom Hearts III	40808	The Secret of Monkey Island	14619
Final Fantasy XIV	689360	Mass Effect	385744	Monkey Island 2: LeChuck's Revenge	9285
Final Fantasy XV	74792	Mass Effect 2	270209	The Curse of Monkey Island	30577
Final Fantasy VII Remake	86487	Mass Effect 3	361358	Total	6280892

Table 1: The selection of games in the corpus and the number of words of dialogue in each game. See the supporting document for sources.

We wanted to provide researchers with the ability to conduct comparisons between major styles of RPG as well as diachronic analyses. Therefore, we aimed to collect a balance of games according to three factors. The first was RPG style, including Western RPGs and

Japanese-style RPGs, since those are the dominant types. Secondly, the target audience age. For the latter we used three categories based on the official ESRB ratings: "Child" (ESRB ratings "Everyone", "Everyone 10+"), "Teen" (ESRB rating "Teen"), and "Adult" (ESRB ratings "Mature 17+" and "Adults Only 18+"). Table 2 shows the distribution of game series with the number of games in brackets. Finally, we also aimed for a balance of games across time between 1985 and 2020 (table 3).

RATING	WESTERN	JRPG	TOTALS
CHILD	Stardew Valley, King's Quest Series, Monkey Island Series (13)	Super Mario RPG, Kingdom Hearts Series (5)	18
TEEN	Horizon Zero Dawn, Star Wars: KOTOR (2)	Chrono-Trigger, Final Fantasy Series (19)	20
ADULT	Mass Effect Series, Elder Scrolls Series, Dragon Age Series (9)	Persona Series (3)	12
TOTALS	24	26	50

Table 2: Distribution of games by ESRB rating and style. Numbers in brackets show the total number of games for the given category.

YEARS	GAMES	NUMBER OF GAMES	PERCENTAGE OF DIALOGUE
1985 - 1989	FFI; FFII; KQ1; KQ2; KQ3; KQ4	6	0.3%
1990 - 1994	FFIV; FFV; FFVI; KQ5; KQ6; KQ7; Monkey Island 1; Monkey Island 2	8	2.8%
1995 - 1999	Chrono Trigger; FFVII; FFVIII; KQ8; Monkey Island 3; Super Mario RPG; Daggerfall	7	4.7%
2000 - 2004	FFIX; FFX; FFX2; KH; Star Wars: KOTOR; Morrowind	6	13.5%
2005 - 2009	Dragon Age: Origins; FFXII; FFXIII; KH2; ME1; Persona3; Persona4; Oblivion	8	35.9%
2010 - 2014	Dragon Age 2; FFXIV; FFXIII-2; FFXIII-LR; KH3D; ME2; ME3; Skyrim	8	30.0%
2015 - 2020	FFXV; FFVII-R; Horizon Zero Dawn; KQ Chapters; KH3; Persona5; Stardew Valley	7	12.8%

Table 2: Distribution of games by year. FF = Final Fantasy; KQ = King's Quest; KH = Kingdom Hearts; ME = Mass Effect.

Given the difficulties of balancing the three criteria above, and the constraint on being able to find dialogue sources, the corpus is relatively well balanced at the game level. However, the corpus is not balanced in terms of number of words per game. There is a difference of two orders of magnitude between the game with most dialogue and the game with least dialogue. This is mainly an effect of the change in technology, storage capacity, and development budget for video games over the last three decades. The distribution of words by RPG type is skewed towards Western RPGs (63%) compared to JRPGs (37%), and the distribution of ratings is skewed towards adult titles (8% Child, 33% Teen, 59% Adult). The distribution is also biased towards games released between 2005-2014 (see table 2), mainly due to early games being limited by memory constraints.

Despite these imbalances, we suggest that the corpus is still representative of an average gaming experience. In any case, deciding how to sub-sample each game to create a balanced corpus would depend on the particular research question and type of balance aimed for (balancing characters, character groups, certain stages of the game, or depth of dialogue tree). The corpus includes as much available dialogue for each game as possible and future studies can make their own decisions about what to sample.

2.3 Data format

van Stegeren & Theune (2020:2) suggest two key properties for the data format of a corpus of video game data: richness (datasets should contain both game text and information about their in-game context), and portability (the data should be open-source format). The ideal format for data would need to capture various features:

- Lines of dialogue, paired with the name of the character that is speaking them.
- Distinguishing dialogue that every player will experience compared to optional dialogue.
- Capturing the structure of the dialogue 'tree' when there are multiple alternative sequences of dialogue.
- Linking dialogue with other contextual information.
- All games represented with a common format.

The corpus should also be both machine-readable (in order to facilitate analysis with computational methods), and readable by humans (in order to facilitate qualitative analysis).

Previous corpora have used a variety of formats. Heritage (2021) uses raw text files. This is human- and machine-readable, but does not link dialogue to which character is speaking. Heritage also uses a hand-annotated format for documenting narrative permutations, but this is not strictly machine-readable. van Stegeren & Theune (2020) use a simple tabular format which lists the character, the dialogue, an ID number, and the ID numbers of lines of dialogue that can follow this. This links dialogue to characters, represents recursive (and graph-like) structures, and is machine readable. However, it is not easy for humans to follow the structure.

The canonical form of the data in the current corpus is a JSON format. This is a plain text format that can be read with an ordinary text editor or any open-source JSON tool. This format was chosen for several reasons: it can capture pairings of data and recursive structures; is portable and open-source; it is machine-readable (e.g. it is compatible with python dictionaries); and it looks like a screenplay script to human readers. A YAML format might look more readable, since it doesn't include parentheses, but levels of embedding quickly become difficult to distinguish. In any case, JSON can be converted to many other formats.

The script for a game is a list of dictionaries. Each dictionary has a main key which represents the name of the character who is speaking. The value associated with this key is the dialogue they speak. Below is an example of three lines of dialogue from *Final Fantasy VII* between the characters Barret and Cloud:

```
[
{"Barret": "The planet's full of Mako energy. People here use it
every day."},
{"ACTION": "Cloud shrugs."},
{"Barret": "It's the life blood of this planet."},
{"Cloud": "I'm not here for a lecture. Let's just hurry."}
]
```

According to the JSON format, square brackets enclose a list of items separated by commas, and curly brackets enclose a set of key:value pairs.

There are a number of reserved main keys that indicate particular kinds of information. This helps separate dialogue from non-dialogue. These include:

- ACTION: the value is a description of the action (see the example above).
- LOCATION: the value is a description of the location.
- SYSTEM: the value is a transcription of non-diegetic text that appears to the player, but is not spoken by in-game characters. This includes menu text and tutorial text.
- CHOICE: a branching choice (see below).
- GOTO: The script continues at another location (see below).
- STATUS: A description of some contextual status, used to interpret branching choices (see below).

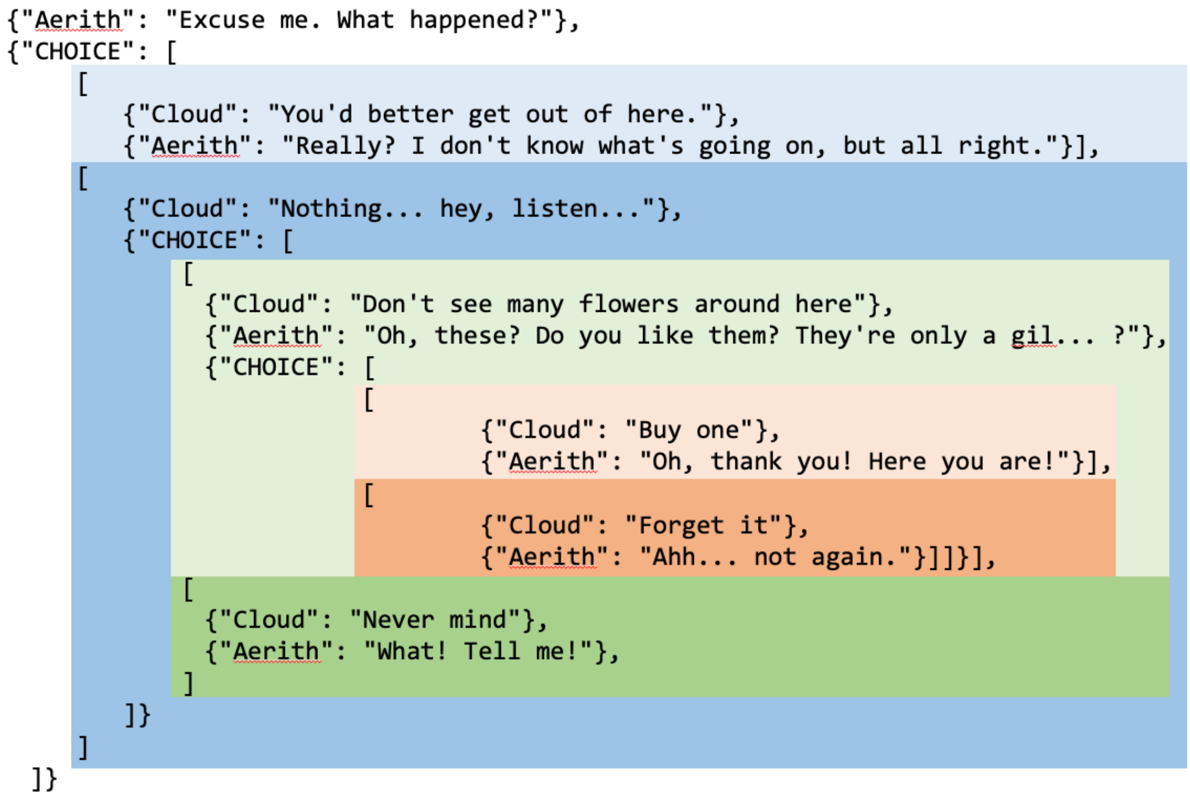
The dictionary can optionally include other keys, as long as they begin with an underscore, which might convey contextual information. In the example below from *Oblivion*, the extra data includes the race of the character, the emotion information assigned to the face animator, and the quest ID that this dialogue appears within.

```
{"Velwyn Benirus": "I got the door open. The rest is up to you.",
      "_Race": "Imperial",
      "_Emotion": "Fear 90",
      "_Quest": "0000BCD7"}
```

Similar structures are included in the data for Chrono Trigger to include a parallel corpus of lines in the original Japanese and a more recent fan “retranslation”. This allows research into translation (e.g. Müller Galhardi, 2014).

Branching dialogue is handled using a recursive structure. The main key is labelled “CHOICE” and its value is a list of possible outcomes. Each outcome is a list of dialogue dictionaries, like the normal format. Any of the dialogue dictionaries can itself be a choice structure, allowing recursive branching.

Figure 1 shows a recursive branching dialogue from *Final Fantasy VII* between Cloud (the player character) and Aerith. There are several possible binary choices, indicated by different shades of the same colour:



```
{"Aerith": "Excuse me. What happened?"},
{"CHOICE": [
        [
          {"Cloud": "You'd better get out of here."},
          {"Aerith": "Really? I don't know what's going on, but all right."}],
        [
          {"Cloud": "Nothing... hey, listen..."},
          {"CHOICE": [
               [
                 {"Cloud": "Don't see many flowers around here"},
                 {"Aerith": "Oh, these? Do you like them? They're only a gil... ?"},
                 {"CHOICE": [
                           [
                                 {"Cloud": "Buy one"},
                                 {"Aerith": "Oh, thank you! Here you are!"}],
                           [
                                 {"Cloud": "Forget it"},
                                 {"Aerith": "Ahh... not again."}]]}],
               [
                 {"Cloud": "Never mind"},
                 {"Aerith": "What! Tell me!"},
               ]
          ]}
        ]
  ]}
```

Figure 1: Representation of a branching dialogue structure, with colours representing different embedded levels.

The structure above represents the branching structure in figure 2 below:

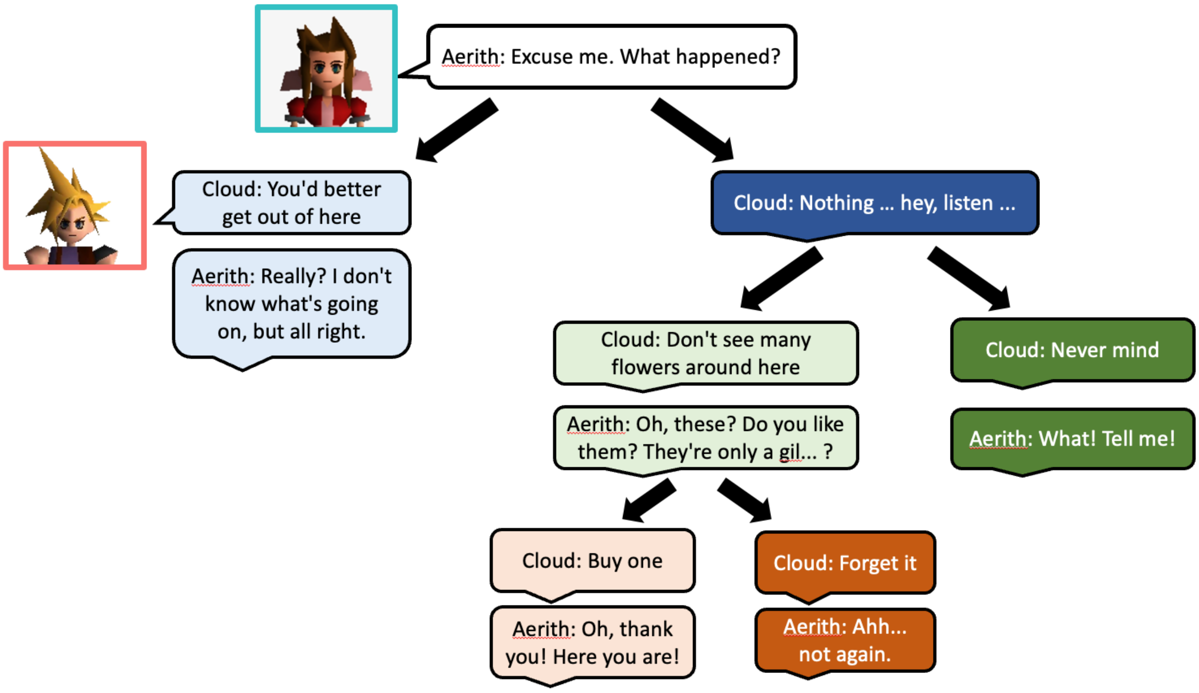

Aerith: Excuse me. What happened?
Cloud: You'd better get out of here
Aerith: Really? I don't know what's going on, but all right.
Cloud: Nothing ... hey, listen ...
Cloud: Don't see many flowers around here
Aerith: Oh, these? Do you like them? They're only a gil... ?
Cloud: Never mind
Aerith: What! Tell me!
Cloud: Buy one
Aerith: Oh, thank you! Here you are!
Cloud: Forget it
Aerith: Ahh... not again.

Figure 2: The branching dialogue represented in Figure 1.

Typically, each outcome will represent the consequences of a player choice. The first entry in each outcome will indicate the trigger for that outcome (e.g. choosing to buy a flower or not). However, the format is designed to capture any type of outcome, including those triggered by: game conditions (e.g. a certain quest is complete); player character statuses (e.g. player character is female); characters present (e.g. party composition); alternative responses (e.g. first or second time asking a question); or random choices. One outcome can be empty, indicating that there's a possibility of hearing no additional dialogue.

In the example below from *Skyrim* (2011), Ralof responds differently based on the player character's class. The first entry of each option stats the condition for that option relies on a game "status" related to the player's race or questline:

```
{“CHOICE”: [
   [
     {“STATUS”: “Player race/questline”},
     {“Ralof”: “Mage, eh? Well, to each his own.”}],
   [
     {“STATUS”: “Player race/questline”},
     {“Ralof”: “Warrior, good! Those stars will guide you to honor and glory.”}],
   [
     {“STATUS”: “Player race/questline”},
     {“Ralof”: “Thief, eh? It’s never too late to take charge of your own fate.”}
   ]
]}
```

Although a branching structure can technically represent all types of outcome, some games are more succinctly represented using a mix of branching structures and links between parts of this structure. To handle this, each dialogue dictionary can be given a unique ID (assigned to the key "_ID"). Then there is a reserved main key "GOTO" paired with an ID. This is an instruction that the script resumes at the dictionary with the given ID.

In the example below from *Mass Effect 3* (BioWare, 2012), Din Korlack's line changes depending on the gender of the player character Shepard. If Shepard is female, then the first outcome is experienced. If not, then the second outcome is experienced, but then there's a GOTO line that rejoins the script at Zaeed's first line. That is, the text dialogue experienced by the player is identical except for the pronouns.

```
{“CHOICE”: [
  [
      {“STATUS”: “Shepard is female (12/0)”},
      {“Din Korlack”: “Shepard’s investigating. She’s... a recent acquaintance.”,
       “_ID”: “689365”},
      {“Zaeed Massani”: “How recent?”, “_ID”: “689367”},
      {“Din Korlack”: “Very.”, “_ID”: “689368”},
      {“Zaeed Massani”: “Shit. All right, I’m listening.”, “_ID”: “689369”}],
  [
      {“STATUS”: “ (-1/0)”},
      {“Din Korlack”: “Shepard’s investigating. He’s... a recent acquaintance.”,
       “_ID”: “689366”},
      {“GOTO”: “689367”}]]}
```

2.4 Corpus pipeline

The pipeline for creating any corpus involves several steps, including obtaining data, cleaning and formatting it, and producing summary statistics. We approach these tasks using replicatable methods in order to create a "self-inflating corpus". The entire pipeline process for each game is implemented in python code. This has several advantages. First, it means that we can release the code and metadata publicly without needing to share copyrighted materials directly. Other researchers can re-run the code in order to “re-inflate” the corpus. Secondly, the process of cleaning and formatting involves many iterations of running a process, checking results and re-running the process. A software pipeline speeds up this process considerably. Thirdly, it increases transparency: every step of processing and cleaning is visible to researchers, including every manual edit. Finally, the corpus can continue to be expanded and edited in a centralised and consistent way.

Implementing this pipeline for a game requires three parts: a scraper, a parser and metadata. The scraper is a simple python program that downloads data from the internet into temporary local files. The parser takes this data and parses it into a consistent format, using information in the metadata. These aspects are expanded on below. The corpus repository also includes scripts for extracting text, dividing it into specific groups, and applying quantitative methods.

2.5 Parsing

A parser is a python program that take some raw format, typically a html webpage, and converts it into the standard JSON format described above. The pipeline is implemented so that a single parser can be applied to several games. The hope was that code could be re-used across games. In reality, nearly every game needed its own special parser because of the specifics of the source format or game mechanics. As a result, there are around 10,000 lines of code for the parsers in the repository.

2.6 Metadata

Specific steps for cleaning and processing the data for each game are stored in a metadata file called ‘meta.json’. The metadata file is a JSON format file with fields that store basic information such as name of the game, the series it belongs to, the year of publication, and the source of the data. The field “parserParameters” stores details of which parser should be applied to the raw data, and additional arguments to be passed from a specific game to a generic parser. The “sourceFeatures” field specifies what the source contains, including the type (‘fan transcript’, ‘game data’, or ‘wiki’), completeness (“complete”, virtually all dialogue that a player could experience; “high”, most dialogue a player would experience on a typical play-through of a game; and “sample”, e.g. a single play-through of the game without alternative dialogue choices, or only portions of the game), “dialogueOrder” (whether the lines are in the order in which they appear to players) and “choices” (“NA”: game has no choices, “not included”, “partial”, “complete”).

The “characterGroups” field in the metadata is a mapping from group names to a list of character names who are members of that group. In the current implementation, this is used to code each character’s gender, but could be used to code any arbitrary group. The group labels can be any string, and there can be as many groups as is necessary to capture the diversity in the character groupings. For example:

```
“characterGroups”: {
      “male”: [“Cloud”, “Barret”],
      “female”: [“Tifa”, “Aerith”],
      “neutral”: [“Chocobo”, “Jenova”]
      }
```

Finally, the “aliases” field stores details of manual changes to character names. This can deal with unifying alternative names, fixing typos or misattributions, and splitting generic character names (e.g. “Guard”) into individual characters based on lines of dialogue (see the repository for further details).

The current corpus identifies the gender of each character. This required around 28,000 lines of gender coding and character name unification metadata.

2.7 Error checking

The data for each game underwent two types of error checking procedures. The first was for true positives (ensuring that lines in the source correspond to lines in the game) and transcription errors (identifying lines that may have ben mistranscribed in the source). A video of the game being played was identified on YouTube and a random sequence of three lines of dialogue was selected. The checker confirmed that the dialogue existed in the corpus, that the text of the transcription was accurate, and that the structure of the dialogue was accurate (e.g. choice structure). This was done for 5 random sequences in the video.

The second procedure checked for false positives (lines in the source that are not really in the game) and parsing errors (lines that are in the data, but parsed incorrectly in terms of character assignment or dialogue structure). A random line in the parsed data was selected, and the checker confirmed that the line had been correctly parsed from the source, including checking the character name and the formatting. This was repeated 5 times.

Each game underwent several rounds of checking, applying fixes, and re-checking. Out of 500 tests, these procedures identified source errors in 8% of tests, transcription errors in 0% of tests, and parsing errors in less than 1% of tests. These errors, and many others which were identified informally, were filed as bug reports on the Github repository and all were fixed, with the development repository having over 900 commits.

The most frequent type of errors were source errors, which were observed in 17 games. For example, in the play-through video of Skyrim that was chosen, conversations with some NPCs were observed that were not in the source transcript (hence the "completeness" of this game being 'high' instead of 'complete'). Or for *The Secret of Monkey Island* (LucasArts, 1990), the source only records one path through a dialogue tree, so there was dialogue that was not captured. Both of these cases were known limitations of the source.

2.8 Availability

The corpus pipeline is available from a github repository (made public on publication). Anyone can suggest edits to the current data or contribute additional games.

3. Demonstration

To illustrate the use of the corpus, we conducted a study comparing the dialogue of female and male characters. The gender of characters was coded (see Rennick et al., under review). Balanced subcorpora were created for male and female dialogue by selecting 1000 random lines from 25 games were there was enough data. This created subcorpora of 392,138 words of male dialogue and 369,753 words of female dialogue. A keyness analysis using log likelihood was applied, using the female dialogue as the target corpus and male dialogue as the reference corpus. A significance threshold of 0.01 was applied to identify 459 most significant keywords. To aid analysis, these words were classified into categories using the

UCREL system (Rayson et al., 2004), using the online semantic tagger (http://ucrel-api.lancaster.ac.uk/usas/tagger.html) and manual coding for words that could not be matched. All methods and code for analysis are detailed in the supporting materials.

The top key categories used more by female speakers than male speakers included personal names, pronouns, and kin. While, in general, female characters mention names about 50% more than males (Log Likelihood = 145.38 , p < 0.001), some individual names were used more by male characters than female, and some were used more by female characters than by male characters. We coded the referents for gender, and found that male referents are mentioned more than female (female referents = 38%, male referents = 62%, binomial test p = 0.04), which may reflect the greater number of male characters in the general corpus. Male and female speakers mention female referents at about the same rate, but female speakers are about twice as likely to mention a male referent than male speakers (Fisher's exact test p < 0.001).

For pronouns, males used archaic forms ("thy", "thee", "thou") more often than females, probably reflecting gendered ideas of chivalry. Most other pronouns were spoken more by females than males, reflecting patterns in real conversations (estimated from the BNC Spoken corpus, Love et al., 2017, see SI). The exception is "our", which is used more by male characters than female characters in games, while in real conversation the opposite is true (see SI). We note that collocations of "our" in male dialogue included more physical entities ("home", "homeland", "nation", "defences") than the top collocations for female characters ("prayers", "friendship"), perhaps reflecting gendered role stereotypes.

Female characters used kin terms more than male characters (Log Likelihood = 91.81, p < 0.001), reflecting patterns in real conversation (see SI). The exception is "son", which is used more by male characters than female characters in games, while in real conversation the opposite is true. This may be explained by several factors. Firstly, "son" appears in the swearing phrase "son of a bitch", which is used more by male characters than female characters (see SI). Secondly, "son" is used as an affectionate fictive kin term, often by an older man to refer to a younger man, while "daughter" is not necessarily used in the same way. In the Mass Effect series, when a male player character is called "son", the alternative lines for a female player character use non-kin terms like "miss" or non-gendered terms like "child". Finally, many game worlds exhibit patriarchal systems and thus there are several quests involving sons and inheritance of positions of power. In line with this, the frequency of "his son" (patrilineal) is higher than "her son" (see SI).

Each of these examples demonstrate a difference in the way that female and male characters are portrayed. While some may reflect empirical differences in the real world, a large body of literature has demonstrated persistent sexist biases in video games and in video game culture (e.g. Stermer & Burkley, 2015; Ruvalcaba et al., 2018; Mortensen, 2018; see SooHoo, 2022). Studies that empirically demonstrate differences in the depiction of genders are critical to understanding how these attitudes persist.

4. Conclusion

We presented a large-scale, consistently coded, self-inflating and expandable corpus of video game dialogue. This allows researchers to investigate video game dialogue as a language variety, or to compare games against each other. We demonstrated it is possible to apply standard corpus methodologies to VGDC, demonstrating significant patterns in gendered language.

There are a number of limitations to the corpus. First, as explained in section 2.1, obtaining complete and error-free data is difficult, and the corpus is composed of a range of source types and completeness. We initially worried about the quality of fan transcriptions, but this turned out not to be a frequent problem, and secondary compared to the more fundamental problem of sourcing the data. Secondly, some games have more dialogue than others, due to factors such as increasing memory capacity over the last four decades. This means that specific investigations will need to sample the corpus in different ways according to their needs. Thirdly, the corpus is stored in a custom format that cannot be simply processed by standard corpus software. However, the corpus is open source and expandable, meaning that each of these limitations can be addressed incrementally. We hope the corpus will continue to grow and act as a centralising resource for the subfield.

We welcome contributions to the corpus, though contributors should be aware that there are some risks involved with engaging with topics related to video games. Individuals who comment on video games, including academics, have been the target of co-ordinated online abuse and threats by a portion of the community that see themselves as gatekeepers (Mortensen, 2018; Chess & Shaw, 2015). Despite these risks, or perhaps because of them, we hope that research into this new medium will continue to progress.

Acknowledgements

We thank the fans who have organised the sources that this project is based on. We also thank Melanie Clinton, Elena Ioannidou, Liana Oh, Charlotte Clooney, E. T., and Edward Healy for helping with corpus construction. S.R. was supported by a Swiss National Science Foundation grant (182847).

References

BBC News. 2019. Gaming worth more than video and music combined. 3/01/2019 https://www.bbc.co.uk/news/technology-46746593

Bednarek, M. (2018). *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.

Bednarek, M., 2015. Corpus-assisted multimodal discourse analysis of television and film narratives. In Corpora and discourse studies (pp. 63-87). Palgrave Macmillan, London.

Behnke, M., Chwiłkowska, P. and Kaczmarek, L.D., 2021. What makes male gamers angry, sad, amused, and enthusiastic while playing violent video games?. *Entertainment Computing*, *37*, p.100397.

Bethesda (2006) The Elder Scrolls IV: Oblivion [Video game]. Bethesda.

Bethesda (2011) The Elder Scrolls V: Skyrim [Video game]. Bethesda.

BioWare (1997) Dragon Age 2 [Video game]. BioWare.

BioWare (2003) Star Wars: Knights of the Old Republic [Video game]. BioWare.

BioWare (2012) Mass Effect 3 [Video game]. BioWare.

BioWare (2009) Dragon Age: Origins [Video game]. BioWare.

Bonsignori, V. (2009) ‘Transcribing Film Dialogue: From Orthographic to Prosodic Transcription’, in M. Freddi and M. Pavesi (eds.) Analysing Audiovisual Dialogue.

Linguistic and Translational Insights (Bologna: Clueb), pp. 185–200.

Chess, S. and Shaw, A., 2015. A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, *59*(1), pp.208-220.

Carrillo Masso, I., 2009. Developing a methodology for corpus-based computer game studies. *Journal of Gaming & Virtual Worlds*, *1*(2), pp.143-169.

Eligio, R.B. and Kaschak, M.P., 2020. Gaming experience affects the interpretation of ambiguous words. *Plos one*, *15*(12), p.e0243512.

Erdur, N., 2022. Gender in Genshin Impact: A Corpus-Assisted Discourse Analysis. *Language Education and Technology*, *2*(1).

ESA. 2021. 2021 essential facts about the video game industry. Entertain. Softw. Assoc. (2021).

Fox, J. & Tang, W. Y. 2014. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. Comput. Hum. Behav. 33, 314–320.

Gee, J. P. 2003. *What video games have to teach us about learning and literacy*. Palgrave Macmillan.

Gilbert, N. 2021. Number of Gamers Worldwide 2022/2023: Demographics, Statistics, and Predictions. https://financesonline.com/number-of-gamers-worldwide/

Goorimoorthee, T., Csipo, A., Carleton, S. and Ensslin, A., 2019. Language ideologies in videogame discourse: Forms of sociophonetic othering in accented character speech. *Approaches to videogame discourse: Lexis, interaction, textuality*, pp.269-287.

Heritage, F. 2022. "Magical women: Representations of female characters in the Witcher video game series." *Discourse, Context & Media* 49: 100627.

Heritage, F., 2020. Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level. *Game studies*, *20*(3), p.20.

Heritage, F., 2021. Language, Gender, and Videogames. In *Language, Gender and Videogames* (pp. 27-61). Palgrave Macmillan, Cham.

ISFE. 2021. 2021 key facts about the European video games sector. Interact. Softw. Fed. Eur. (2021).

Kahila, J., Tedre, M., Kahila, S., Vartiainen, H., Valtonen, T. and Mäkitalo, K., 2021. Children's gaming involves much more than the gaming itself: A study of the metagame among 12-to 15-year-old children. *Convergence*, *27*(3), pp.768-786.

Korea Creative Content Agency. 2020. 7th survey report of video gamers in Korea. https://seoulz.com/the-korea-creative-content- agency-kocca/ (2020).

Kybartas, B., and Verbrugge, C. 2014. Analysis of Re-GEN as a graph-rewriting system for quest generation. IEEE Transactions on Computational Intelligence and AI in Games 6(2):228–242

Li, J., 2022. A systematic review of video games for second language acquisition. *Research Anthology on Developments in Gamification and Game-Based Learning*, pp.1345-1371.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The spoken BNC2014. International Journal of Corpus Linguistics, 22(3), 319-344.

LucasArts (1990) The Secret of Monkey Island [Video game]. LucasArts

Malkowski, J. & Russworm, T. M. (eds) 2017. *Gaming representation: Race, gender, and sexuality in video games*. Indiana University Press.

Markey, P.M., Markey, C.N. and French, J.E., 2015. Violent video games and real-world violence: Rhetoric versus data. *Psychology of Popular Media Culture*, *4*(4), p.277.

Mortensen, T. E. 2018. Anger, fear, and games: The long event of# gamergate. Games Cult. 13, 787–806.

Müller Galhardi, R. (2014). Video game and Fan translation: A case study of Chrono Trigger. In Mangiron et al. (Eds.) *Fun for All: Translation and Accessibility Practices in Video Games*, 175-195.

Newzoo. 2019. Newzoo and gamma data report. https://newzoo.com/insights/articles/newzoo-and-gamma-data-report-female- mobile-gamers-in-japan-play-more-per-week-than-men (2019).

Nico Partners. 2021. The China Gamers Report. https://nikopartners.com/china-gamers-report/ (2021).

Pandurov, M. 2021. 15 amazing video game statistics for Canada in 2021 (2021).

Qustodio (2020) Apps and digital natives: the new normal. https://qweb.cdn.prismic.io/qweb/f5057b93-3d28-4fd2-be2e-d040b897f82d_ADR_en_Qustodio+2020+report.pdf

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Rennick, S., Roberts, S.G., Clinton, M., Ionnidou, E., Oh, L., Clooney, C., E.T., Healy, E. (under review). Gender bias in video game dialogue.

Ruvalcaba, O., Shulze, J., Kim, A., Berzenski, S. R. & Otten, M. P. 2018. Women's experiences in esports: Gendered differences in peer and spectator feedback during competitive video game play. J. Sport Soc. Issues 42, 295–311.

Schmidt, T., Engl, I., Herzog, J. and Judisch, L., 2020. Towards an Analysis of Gender in Video Game Culture: Exploring Gender specific Vocabulary in Video Game Magazines.

SooHoo, J. 2022. A Systematic review of sexism in video games, DOI: 10.31234/osf.io/xrh36.

Square (1997) Final Fantasy VII [Video game]. Square.

Square Enix (2016) Final Fantasy XV [Video game]. Square Enix.

van Stegeren, J., Theune, M. (2020). Fantastic Strings and Where to Find Them: The Quest for High-Quality Video Game Text Corpora. 12th edition of the Intelligent Narrative Technologies workshop. October 19-20, 2020.

DeWinter, J., Kocurek, C.A. 2017. "aw fuck, i got a bitch on my team!": Women and the exclusionary cultures of the computer game complex. *Gaming representation: Race, gender, and sexuality in video games*, pp. 57–73.

Keyness study of differences between male and female speech in video games

Contents

Introduction

This is the supporting document for 'The Video Game Dialogue Corpus'. It lists the sources for the corpus data, demonstrates how to load dialogue data from the corpus using R, and runs a keyword analysis comparing the language of male and female characters.

The document includes commentary which is formatted like this.

```
And R code which is formatted like this.
```

```
## [1] "And output of running the R code which is formatted like this"
```

Sources

The dialogue scripts were sourced from a variety of source types, including public fan transcripts, public wikis and directly from game data. We are very grateful for the work that fans put into organising these sources.

Table of games and sources:

Game	Year	Source	Type	Completeness
Chrono Trigger	1995	https://www.chronocompendium.com/Term/Retranslation.html	game data	complete
Dragon Age: Origins	2009	Game Data	game data	complete
Dragon Age 2	2011	Game Data	game data	complete
Final Fantasy	1987	https://archive.rpgamer.com/games/ff/ff1/info/ff1_script.txt	fan transcript	complete
Final Fantasy II	1988	https://gamefaqs.gamespot.com/ps/916670-final-fantasy-ii/faqs/61436	fan transcript	complete
Final Fantasy IV	1991	https://gamefaqs.gamespot.com/ds/939425-final-fantasy-iv/faqs/53978	fan transcript	high
Final Fantasy V	1992	http://www.finalfantasyquotes.com/ff5/script	fan transcript	high
Final Fantasy VI	1994	https://gamefaqs.gamespot.com/snes/554041-final-fantasy-iii/faqs/70118	fan transcript	high
Final Fantasy VII	1997	http://www.yinza.com/Fandom/Script/	fan transcript	high
Final Fantasy VIII	1999	https://www.neoseeker.com/finalfantasy8/faqs/136092-final-fantasy-viii-script-a.html	fan transcript	high
Final Fantasy IX	2000	https://gamefaqs.gamespot.com/ps/197338-final-fantasy-ix/faqs/42207	fan transcript	high
Final Fantasy X	2001	http://auronlu.istad.org/ffx-script/	fan transcript	high
Final Fantasy X-2	2003	https://www.ffcompendium.com/h/faqs/ffx2scriptaschthehated.txt	fan transcript	high
Final Fantasy XII	2006	http://ffnerdery.blogspot.com/p/final-fantasy-xii.html	fan transcript	high
Final Fantasy XIII	2009	https://en.wikiquote.org/wiki/Final_Fantasy_XIII#Dialogue	wiki	sample
Final Fantasy XIV	2010	https://ffxiv.gamerescape.com	wiki	sample
Final Fantasy XIII-2	2011	https://gamefaqs.gamespot.com/pc/846193-final-fantasy-xiii-2/faqs/64861	fan transcript	high
Lightning Returns: Final Fantasy XIII	2014	https://www.youtube.com/watch?v=Kl_EgMs2V4A	fan transcript	sample
Final Fantasy XV	2016	https://thelifestream.net/final-fantasy-xv-lore/final-fantasy-xv-chapter-by-chapter-lore-exposition-and-development/	fan transcript	high
Final Fantasy VII Remake	2020	https://finalfantasy.fandom.com/wiki/Final_Fantasy_VII_Remake_script	wiki	high
Horizon Zero Dawn	2017	https://game-scripts.fandom.com/wiki/Horizon_Zero_Dawn	fan transcript	sample
King's Quest III: To Heir Is Human	1986	https://kingsquest.fandom.com/wiki/KQ3_transcript	game data	complete
King's Quest I: Quest for the Crown	1987	https://kingsquest.fandom.com/wiki/KQ1SCI_transcript	game data	complete
King's Quest II: Romancing the Throne	1987	https://kingsquest.fandom.com/wiki/KQ2_transcript	game data	complete
King's Quest IV: The Perils of Rosella	1988	https://kingsquest.fandom.com/wiki/KQ4SCI_transcript	game data	complete
King's Quest V	1992	https://kingsquest.fandom.com/wiki/KQ5NES_transcript	game data	high
King's Quest VI	1992	https://kingsquest.fandom.com/wiki/KQ6_transcript	game data	complete
King's Quest VII: The Princeless Bride	1994	https://kingsquest.fandom.com/wiki/KQ7_transcript	game data	complete
King's Quest VIII	1998	https://kingsquest.fandom.com/wiki/KQ8_transcript#Connor	game data	complete
King's Quest Chapters	2015	https://kingsquest.fandom.com/wiki/KQC1_transcript	game data	high
Kingdom Hearts	2002	https://transcripts.fandom.com/wiki/Kingdom_Hearts	fan transcript	high
Kingdom Hearts II	2005	https://transcripts.fandom.com/wiki/Kingdom_Hearts_II	fan transcript	high
Kingdom Hearts 3D: Dream Drop Distance	2012	https://gamefaqs.gamespot.com/3ds/997779-kingdom-hearts-3d-dream-drop-distance/faqs/65008	fan transcript	high
Kingdom Hearts III	2019	https://gamefaqs.gamespot.com/ps4/718920-kingdom-hearts-iii/faqs/78466	fan transcript	high
Mass Effect	2007	See Readme	game data	complete
Mass Effect 2	2010	Dump from custom branch of Legendary Explorer: https://github.com/ME3Tweaks/LegendaryExplorer	game data	complete
Mass Effect 3	2012	Dump from custom branch of Legendary Explorer: https://github.com/ME3Tweaks/LegendaryExplorer	game data	complete
The Secret of Monkey Island	1990	https://gamefaqs.gamespot.com/pc/562681-the-secret-of-monkey-island/faqs/23891	fan transcript	high
Monkey Island 2: LeChuck's Revenge	1991	https://gamefaqs.gamespot.com/pc/562680-monkey-island-2-lechucks-revenge/faqs/79490	fan transcript	sample
The Curse of Monkey Island	1997	https://gamefaqs.gamespot.com/pc/29083-the-curse-of-monkey-island/faqs/60819	fan transcript	high
Persona 3	2006	https://lparchive.org/Persona-3/	fan transcript	sample
Persona 4	2008	https://lparchive.org/Persona-4/	fan transcript	sample
Persona 5	2016	https://lparchive.org/Persona-5/	fan transcript	sample
Star Wars: Knights of the Old Republic	2003	https://github.com/hmi-utwente/video-game-text-corpora/blob/master/Star%20Wars:%20Knights%20of%20the%20Old%20Republic/data/dataset_20200716.csv	game data	complete
Stardew Valley	2016	https://drive.google.com/drive/folders/0BwyXuxAqGS7ueVdFX2dQSUVzcUk	game data	high
Super Mario RPG: Legend of the Seven Stars	1996	https://gamefaqs.gamespot.com/snes/588739-super-mario-rpg-legend-of-the-seven-stars/faqs/30431	fan transcript	high
The Elder Scrolls II: Daggerfall	1996	https://github.com/Interkarma/daggerfall-unity	game data	high
The Elder Scrolls III: Morrowind	2002	https://elderscrolls.fandom.com/wiki/	wiki	sample
The Elder Scrolls IV: Oblivion	2006	https://www.mediafire.com/file/bhkqiqjhfib0waa/dialogueExport.txt/file	game data	complete
The Elder Scrolls V: Skyrim	2011	https://gamefaqs.gamespot.com/pc/615805-the-elder-scrolls-v-skyrim/faqs/69918	fan transcript	high

Corpus analysis of gendered language

This section loads data from the corpus, creates a balanced sample of male and female speech, then performs a keyword analysis.

Load libraries:

```
library(quanteda)
library(quanteda.textstats)
library(rjson)
library(lattice)
```

Functions for calculating log likelihood and collocations:

```
logLikelihood.G2 = function(a,b,c,d){
  E1 = c*((a+b) / (c+d))
  E2 = d*((a+b) / (c+d))
  G2a = (a*log(a/E1))
  G2b = (b*log(b/E2))
  if(is.nan(G2a)){G2a=0}
  if(is.nan(G2b)){G2b=0}
  G2 = 2*(G2a + G2b)
  return(G2)
}

logLikelihood.test =
  function(freqInCorpus1, freqInCorpus2,
           sizeOfCorpus1, sizeOfCorpus2,
           silent=F){
  G2 = logLikelihood.G2(freqInCorpus1,
          freqInCorpus2,
          sizeOfCorpus1,
          sizeOfCorpus2)
  c1Rel = freqInCorpus1/sizeOfCorpus1
  c2Rel = freqInCorpus2/sizeOfCorpus2
  p.value = pchisq(G2, df=1, lower.tail=FALSE)
  if(c2Rel>c1Rel){G2=-G2}
  pres = paste("= ",round(p.value,3))
  if(p.value<0.0001){pres = "< 0.001"}
  if(!silent){
    print(paste("Log Likelihood =",
                round(G2,2), ", p ", pres))
  }
  return(c(G2, p.value))
  }

collocation_mutual_information = function(myTokens, target_word, window=c(4,4),minFreq=4){
  # Keep only words that appear in a 'window' around
  # the target (e.g. within 3 words before or 3 words after).
  toks_target <- tokens_keep(myTokens, pattern = target_word, window = window)
  # Get the frequency of each word in the windowed data
  colloc = textstat_frequency(dfm(toks_target))

  # Work out frequency of all words in whole subcorpus
  # (not just appearing next to target)
  totalFreq = textstat_frequency(dfm(myTokens))
  # Match to colloc
  colloc$freqInWholeCorpus = totalFreq[match(colloc$feature,totalFreq$feature),]$frequency
  # Frequency of the target word
  targetFrequency = totalFreq[totalFreq$feature==target_word,]$frequency
```

```r
  # Total size of corpus
  numTokensTotal = sum(ntoken(myTokens))
  # Work out pointwise mutual information
  colloc$mutualInformation = log((colloc$frequency/numTokensTotal) /
          ((colloc$freqInWholeCorpus/numTokensTotal) * (targetFrequency/numTokensTotal)))
  # print results, ordered by mutual information
  colloc = colloc[order(colloc$mutualInformation,decreasing = T),]
  colloc = colloc[,!names(colloc) %in% c("rank",'group','docfreq')]
  colloc = colloc[colloc$frequency>=minFreq,]
  colloc = colloc[colloc$feature!=target_word,]
  return(colloc)
}
```

Create a balanced corpus

Choose 1000 random lines from male and female dialogue for all games with enough data.

```r
set.seed(15318)

sampleNumLines = 1000

textF = c()
textM = c()
textFMini = c()
textMMini = c()
stats = read.csv("../../../project/results/generalStats.csv",stringsAsFactors = F)
# Remove alternative measures
stats = stats[stats$alternativeMeasure!="True",]
stats = stats[!is.na(stats$words),]
games = c()
femaleNames = c("yunie","vici","askin","morag","nihon","qun","lis")
maleNames = c("minato-san", "minato-kun", "prometheus",
              "crumbler","tong","horus","zat","yamagishi")
folders = unique(stats$folder)
for(folder in folders){
  dx = fromJSON(file =
    paste0("../../../project/data/",folder,"data.json"))["text"]
  dx = unlist(dx)
  names(dx) = gsub("CHOICE\\.","",names(dx))
  names(dx) = gsub("text\\.","",names(dx))
  js = fromJSON(file = paste0("../../../project/data/",
                              folder,"meta.json"))

  femaleNames = unique(c(femaleNames,
    unlist(strsplit(js$characterGroups[["female"]]," "))))
  maleNames = unique(c(maleNames,
    unlist(strsplit(js$characterGroups[["male"]]," "))))

  flines = dx[names(dx) %in% js$characterGroups[["female"]]]
  mlines = dx[names(dx) %in% js$characterGroups[["male"]]]

  textF = c(textF, flines)
  textM = c(textM, mlines)

  if(length(flines)>=sampleNumLines & length(mlines)>=sampleNumLines){
    flinesMini = sample(flines,sampleNumLines)
    mlinesMini = sample(mlines,sampleNumLines)
    textFMini = c(textFMini, flinesMini)
    textMMini = c(textMMini, mlinesMini)
    games = c(games,rep(js$game,sampleNumLines))
  }
}

# Tidy up possessives and pronouns for later
textMMini = gsub("'s"," 's", textMMini)
textMMini = gsub("I'm","I 'm", textMMini, textFMini,ignore.case = T)
textMMini = gsub("you're","you 're", textMMini,ignore.case = T)
textFMini = gsub("'s"," 's", textFMini)
textFMini = gsub("I'm","I 'm", textFMini,ignore.case = T)
textFMini = gsub("you're","you 're", textFMini,ignore.case = T)

dM = data.frame(
  text =textMMini,
  group="male",
  game = games,
  stringsAsFactors = F
)
corpM = corpus(dM)
tokensM = tokens(corpM, remove_punct = TRUE)
maleTotal = sum(ntoken(tokensM))


dF = data.frame(
  text = textFMini,
  group="female",
  game = games,
  stringsAsFactors = F
)
corpF = corpus(dF)
tokensF = tokens(corpF, remove_punct = TRUE)
femaleTotal = sum(ntoken(tokensF))

maleNames = maleNames[!maleNames %in% c("mog")]
femaleNames = femaleNames[!maleNames %in% c("noel","ralen","mog")]
```

This corpus contains 392139 words of male speech and 369766 words of female speech from `length(unique(dM$game))` games.

Keyword analysis

Make a single corpus containin both male and female data, then calculate keywords.

```
corp = corpus(rbind(dF,dM))
tok = tokens(corp,remove_punct = T)
corpDFM = dfm(tok)

key = textstat_keyness(corpDFM,
          target = corp$group=="female",
          measure = "lr")
```

Write keywords to file in order to be tagged by the UCREL tagger website.

```
key2 = key[key$p < 0.01,]
key2$femaleRelFreq = 1000*key2$n_target/femaleTotal
key2$maleRelFreq = 1000*key2$n_reference/maleTotal
cat(key2$feature,file = "~/Downloads/key.txt", sep="\n")
```

The semantic tags are now collated using a custom matching website, and saved to a file UCRELTaggedKeywords.tab. Load this data, fix some elements, create a data frame with the keywords, scores, and semantic tags:

```
taggedKeywords = read.table("../data/UCRELTaggedKeywords.tab",
                            sep="\t", quote = "",header=T)
key2$tag = taggedKeywords[match(
  gsub("'[smtr]e?$","",key2$feature),
  taggedKeywords$word),]$tag
key2[key2$feature=="isn't",]$tag = "Negative"
key2[key2$feature=="ain't",]$tag = "Negative"

key2 = key2[!is.na(key2$tag),]
write.csv(key2, "../../corpusDescription/results/keywords.csv",row.names = F)
```

We an now analyse the data. Below are categories used more by females than males (numbers show number of unique words in the category):

```
moreByF = key2[key2$G2>0,]
knitr::kable(head(sort(table(moreByF$tag),decreasing = T),10))
```

Var1	Freq
Personal names	45
Living creatures generally	11
Discourse Bin	10
Grammatical bin	10
Geographical names	8
Kin	7
People	7
Crime, law and order: Law and order	6
Pronouns etc.	6
Judgement of appearance (pretty etc.)	5

Categories used more by males than females:

```
moreByM = key2[key2$G2<0,]
knitr::kable(head(sort(table(moreByM$tag),decreasing = T),10))
```

Var1	Freq
Personal names	35
Discourse Bin	26
People	12
Religion and the supernatural	12
People:- Male	8
Geographical names	7
Grammatical bin	6
Pronouns etc.	6
Ability:- Success and failure	5
Getting and giving; possession	5

Run log likelihood tests by total frequency for the category:

```
catFreq = data.frame(
  word = tapply(key2$tag, key2$tag, head,n=1),
  femaleFreq = tapply(key2$n_target, key2$tag, sum),
  maleFreq = tapply(key2$n_reference, key2$tag, sum))
```

```
res = sapply(1:nrow(catFreq),function(i){
  logLikelihood.test(
    catFreq[i,]$femaleFreq,
    catFreq[i,]$maleFreq,
    femaleTotal,
    maleTotal,silent = T)
})
catFreq$G2 = res[1,]
catFreq$p = res[2,]
```

Top categories mentioned more by female characters than male characters:

```
topFemale =
  catFreq[order(catFreq$G2,decreasing = T),
          c("word","G2")]
knitr::kable(head(topFemale,10),
              row.names = F,digits = 2)
```

word	G2
Personal names	145.38
Living creatures generally	102.44
Pronouns etc.	93.97
Kin	91.81
Thought, belief	63.04
Judgement of appearance (pretty etc.)	48.71
Politeness	37.62
Architecture and kinds of houses and buildings	37.46
Speech etc:- Communicative	34.49
Degree: Boosters	29.69

Top categories mentioned more by male characters than female characters:

```
topMale =
  catFreq[order(catFreq$G2,decreasing = F),
          c("word","G2")]
knitr::kable(head(topMale,10),
              row.names = F,digits = 2)
```

word	G2
People:- Male	-85.01
Discourse Bin	-51.68
Getting and giving; possession	-48.97
Ability:- Success and failure	-37.39
Being	-34.86
Money generally	-33.83
Measurement	-33.35
Science and technology in general	-32.48
Location and direction	-25.81
Movement/transportation: air	-25.39

These suggest several categories of interest which are investigated further below:

Personal names

Names are significant keywords for both male and female characters:

```
pn = key2[key2$tag=="Personal names",]
knitr::kable(
  pn[,c("feature","femaleRelFreq","maleRelFreq","G2")],
        row.names = F, digits = 2)
```

feature	femaleRelFreq	maleRelFreq	G2
cloud	0.57	0.15	99.23
ajira	0.14	0.00	76.64
crono	0.39	0.10	74.17
squall	0.38	0.13	50.18
addhiranirr	0.08	0.00	44.83
zidane	0.23	0.06	43.41
ayla	0.18	0.04	38.78
yunie	0.07	0.00	37.60
habasi	0.06	0.00	34.70
mjrn	0.06	0.00	27.64
mog	0.13	0.03	26.23
noel	0.20	0.07	25.12
minato-san	0.04	0.00	23.14
azura	0.08	0.01	22.46
carth	0.04	0.00	18.80
griff	0.03	0.00	17.35
halric	0.03	0.00	15.91
malak	0.10	0.03	15.52
rost	0.04	0.00	14.22
illusive	0.05	0.01	12.79
akihiko	0.04	0.01	10.55
uthid	0.04	0.01	10.55
jote	0.03	0.00	10.43
prometheus	0.03	0.00	10.43
flemeth	0.03	0.00	10.35
hien	0.04	0.01	9.68
shen	0.04	0.01	9.68
barthello	0.02	0.00	9.09
galbedir	0.02	0.00	9.09
ralen	0.03	0.00	9.09
brynjolf	0.03	0.01	8.53
liara	0.04	0.01	8.43
barret	0.08	0.03	8.30
alphinaud	0.13	0.06	8.19
dania	0.02	0.00	7.77
fratley	0.02	0.00	7.77
vici	0.02	0.00	7.77
steiner	0.06	0.02	7.64
serana	0.05	0.01	7.64
minato-kun	0.03	0.01	7.40
ulath-pal	0.03	0.01	7.40
vincent	0.04	0.01	7.34
yamagishi	0.04	0.01	7.34
marle	0.04	0.01	6.73
olin	0.04	0.01	6.73
askin	0.00	0.02	-6.95
ayde	0.00	0.02	-6.95
serah	0.12	0.20	-6.99
bhelen	0.02	0.05	-7.15

feature	femaleRelFreq	maleRelFreq	G2
morag	0.01	0.03	-7.35
dagoth	0.05	0.10	-7.91
olfie	0.01	0.04	-7.92
tong	0.02	0.05	-7.97
ozzie	0.00	0.03	-8.03
loghain	0.02	0.06	-8.08
fullname	0.00	0.02	-8.15
maesa	0.00	0.02	-8.15
nanako	0.00	0.02	-8.15
nibani	0.00	0.02	-8.15
nihon	0.00	0.02	-8.15
edea	0.01	0.04	-9.02
qun	0.00	0.03	-9.18
lian	0.00	0.03	-9.37
mishima	0.00	0.03	-9.37
varvur	0.00	0.03	-9.37
yusuke	0.01	0.04	-10.50
tifa	0.04	0.10	-11.58
penelo	0.01	0.04	-11.59
mitsuru	0.01	0.04	-12.69
ellone	0.03	0.10	-14.20
jobasha	0.00	0.03	-14.61
souji	0.01	0.05	-14.92
horus	0.00	0.03	-15.94
zat	0.00	0.03	-15.94
kimahri	0.04	0.11	-16.04
garik	0.01	0.05	-16.06
alduin	0.02	0.09	-16.75
o'aka	0.00	0.04	-18.60
lis	0.00	0.06	-30.56
aloy	0.05	0.20	-37.91

We can test whether females mention names more than males:

```
logLikelihood.test(sum(pn$n_target),sum(pn$n_reference),
                   femaleTotal,maleTotal)
```

```
## [1] "Log Likelihood = 145.38 , p  < 0.001"
```

```
## [1] 1.453823e+02 1.771648e-33
```

This test is significant, indicating that female characters mention names about 50% more than males.

We can also test whether male and female speakers differ in the gender of the referent they mention. We construct a table of the frequency of times a speaker of a given gender mentions a referent of a given gender:

```
pn = key2[key2$tag=="Personal names",]
pn$gender = NA
pn$gender[pn$feature %in% tolower(femaleNames)] = "Female reference"
pn$gender[pn$feature %in% tolower(maleNames)] = "Male reference"
pn = pn[!is.na(pn$gender),]
x = rbind(
  "Female speaker" = tapply(pn$n_target,pn$gender,sum),
  "Male speaker" = tapply(pn$n_reference,pn$gender,sum))
knitr::kable(x)
```

	Female reference	Male reference
Female speaker	448	1165
Male speaker	410	650

```
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 4.79e-09
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5153169 0.7214613
## sample estimates:
## odds ratio
##  0.6097486
```

Male referents are mentioned more than female, which may reflect the greater number of male characters. However, there is a significant imbalance in the table: Male and female speakers mention female referents at about the same rate. But female speakers are about twice as likely to mention a male referent than male speakers.

Pronouns

```
knitr::kable(key2[key2$tag=="Pronouns etc.",
        c("feature","femaleRelFreq","maleRelFreq","G2")],
        row.names = F, digits = 2)
```

feature	femaleRelFreq	maleRelFreq	G2
i	33.93	30.93	54.41
he	5.51	4.64	28.95
you	32.72	30.93	19.75
she	2.66	2.24	13.56
it	16.98	15.97	11.92
me	7.56	7.00	8.46
thy	0.02	0.05	-7.97
y'all	0.00	0.03	-10.34
our	2.69	3.10	-10.82
somethin	0.03	0.09	-12.58
thee	0.01	0.07	-14.38
thou	0.02	0.10	-26.19

Interestingly, "we" doesn't appear on this list, because the p value of the log likelihood test is relatively high (p = 0.02).

Does this reflect speech in real life? The table below shows the frequencies of pronouns spoken by males and females from the BNC Spoken corpus.

Word	Feq (F)	Feq (M)	Freq per 1000 words (F)	Freq per 1000 words (M)	Log Likelihood	p
I	282,477	152,417	588.99	205.47	116737.2	< 0.0001
you	189,401	120,589	394.92	162.56	60038.2	< 0.0001
she	49,779	18,149	103.79	24.47	32312.1	< 0.0001
he	61,860	35,291	128.98	47.58	23530.3	< 0.0001
we	55,341	35,938	115.39	48.45	16924.6	< 0.0001
me	22,167	10,470	46.22	14.11	10928.0	< 0.0001
our	4099	2629	8.55	3.54	1282.3	< 0.0001

This shows that female speakers also use pronouns more than male speakers in real conversations.

"Our" is different since it's used by male characters more than female characters (and the effect size in real conversations is weaker, too). We tested whether male and female characters used inclusive and exclusive 'our' to different extents. We sampled 100 random lines including "our" from the male and female corpus, and manually coded them for whether they included or excluded the main character. However, there was no significant difference:

```
# Make random selection of lines
set.seed(3456)
ourF = kwic(tokensF, "our", window = 9)
ourF = ourF[sample(1:nrow(ourF), 100),]
ourF$gender = "female"
ourM = kwic(tokensM, "our", window = 9)
ourM = ourM[sample(1:nrow(ourM), 100),]
ourM$gender = "male"

ourOut = rbind(ourF,ourM)
ourOut = ourOut[sample(1:nrow(ourOut)),]
# Write out for manual editing
write.csv(ourOut, file="../results/our.csv",row.names = F)
```

Results:

```
ourType = read.csv("../results/our_edited.csv",stringsAsFactors = F)
fisher.test(table(ourType$type, ourType$gender))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(ourType$type, ourType$gender)
## p-value = 0.4788
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7028261 2.3060375
## sample estimates:
## odds ratio
##   1.271177
```

Collocations for “our” in the male corpus:

```
colloc_ourM = collocation_mutual_information(tokensM, "our", minFreq=4)
head(colloc_ourM,n=15)
```

```
##          feature frequency freqInWholeCorpus mutualInformation
## 317    gatehouse         4                 4          5.777696
## 270       quarry         5                 6          5.595374
## 318        doman         4                 5          5.554552
## 237        homes         5                 8          5.307692
## 345   scientists         4                 8          5.084548
## 147         goal         8                17          5.023924
## 208     homeland         6                13          5.004506
## 111        minds        10                22          4.989238
## 247  destination         5                11          4.989238
## 217       nation         6                15          4.861405
## 301     defenses         4                11          4.766095
## 340 relationship         4                11          4.766095
## 255        dance         5                14          4.748076
## 223    war-chief         6                17          4.736242
## 262       guests         5                16          4.614545
```

Collocations for “our” in the female corpus:

```
colloc_ourF = collocation_mutual_information(tokensF, "our", minFreq=4)
head(colloc_ourF,n=15)
```

```
##          feature frequency freqInWholeCorpus mutualInformation
## 258      neglect         4                 4          5.917883
## 280    remainder         4                 4          5.917883
## 242      prayers         4                 6          5.512418
## 278       voyage         4                 6          5.512418
## 284   friendship         4                 9          5.106953
## 252       allies         4                10          5.001592
## 206         laws         5                13          4.962371
## 219       client         4                12          4.819271
## 228 relationship         4                12          4.819271
## 170    ancestors         6                19          4.765203
## 173     troubles         6                21          4.665120
## 238        names         4                14          4.665120
## 203        guest         5                18          4.636949
## 159      efforts         6                22          4.618600
## 229        minds         4                15          4.596127
```

The top collocations for male characters contain more physical entities than the top collocations for female characters.

Kin

The category of kin is a frequent occurrence in female speech:

```
knitr::kable(key2[key2$tag=="Kin",
        c("feature","n_target","n_reference","femaleRelFreq","maleRelFreq","G2")],
        row.names = F, digits = 2)
```

feature	n_target	n_reference	femaleRelFreq	maleRelFreq	G2
husband	78	7	0.21	0.02	73.73
father	227	120	0.61	0.31	40.14
mother	186	103	0.50	0.26	29.31
dad	68	32	0.18	0.08	15.46
sister	90	55	0.24	0.14	10.72
grampa	10	0	0.03	0.00	10.43
marry	17	5	0.05	0.01	7.64
clans	2	13	0.01	0.03	-8.38
son	58	108	0.16	0.28	-12.51

Female characters talk more about these categories of kin than male characters. The exceptions are “clans” and “son”.

Does this reflect speech in real life? The table below shows the frequencies of pronouns spoken by males and females from the BNC Spoken corpus.

Word	Feq (F)	Feq (M)	Freq per 1000 words (F)	Freq per 1000 words (M)	G2	p
sister	960	310	2.00	0.42	692.4	< 0.0001
mother	847	440	1.77	0.59	369.2	< 0.0001
brother	897	511	1.87	0.69	342.0	< 0.0001
husband	384	133	0.80	0.18	261.0	< 0.0001
daughter	406	218	0.85	0.29	169.0	< 0.0001
son	372	244	0.78	0.33	111.7	< 0.0001
father	457	332	0.95	0.45	111.6	< 0.0001
wife	392	292	0.82	0.39	90.6	< 0.0001

In real conversations, female speakers mention kin names more than male speakers. So the pattern for “son” in video games stands out. There may be several explanations for this. Let’s look at the collocations of the word “son”:

```
knitr::kable(head(collocation_mutual_information(tokensM,"son"),6),
             row.names = F, digits = 2)
```

feature	frequency	freqInWholeCorpus	mutualInformation
bitch	10	20	7.50
forever	4	43	5.82
count	4	43	5.82
his	15	811	4.21
my	38	2565	3.99
am	4	520	3.33

“Son” appears in the swearing phrase “son of a bitch”, which is used more by male characters (10 occurances, frequency per 1000 words = 0.026) than female characters (2 occurances, frequency per 1000 words = 0.005).

“Forever” relates to a poem from King’s Quest Chapters where each stanza ends in “I’ll love you forever my son”.

Furthermore, in real conversations, “son” is used as an affectionate fictive kin term, often by an older man to refer to a younger man. However, “daughter” is not necessarily used in the same way. In video games, non-kin terms like “miss” or non-gendered terms like “child” are used to refer to a female player character in place of “son”. Two examples below are from the Mass Effect series.

```
{"CHOICE": [
    [
        {"STATUS": " (144/0)"},
        {"Simon Atwell": "Ease down, miss.", "_ID": "179420"},
        ],
    [
        {"STATUS": " (-1/0)"},
        {"Simon Atwell": "Ease down, son.", "_ID": "179421"},
]]},
```

```
{"CHOICE": [
 [
    {"STATUS": "[ME3] Shepard is female (17662/0)"},
    {"David Anderson": "You did good, child. You did good. I'm proud of you.", "_ID": "680757"}],
 [
    {"STATUS": " (-1/0)"},
    {"David Anderson": "You did good, son. You did good. I'm proud of you.", "_ID": "680758"}]]}
```

However, each of the cases above are relatively rare, and cannot account for the imbalance in usage between male and female characters. Instead, we guess that the increased frequency of talk of sons by male characters is due to the patriarchal system that many game worlds exhibit. There are several quests involving sons and inheritance of positions of power. One prediction that this hypothesis makes is that there should be more cases of “his son” (patrilineal) rather than “her son”, which the table below shows is the case (Fisher’s exact test p = 0.017).

	... son	... daughter
his ...	13	5
her ...	0	4

The same pattern exists when looking at the entire corpus (Fisher’s exact test p = 0.007):

	... son	... daughter
his ...	84	47
her ...	21	30

Discourse items

Overall, discourse items are used more by male than female characters. There are a range of discourse markers:

```
knitr::kable(key2[key2$tag=="Discourse Bin",
     c("feature", "femaleRelFreq","maleRelFreq","G2")],
     row.names = F)
```

feature	femaleRelFreq	maleRelFreq	G2
oh	2.0715804	1.4459159	42.736040
um	0.3488693	0.1377063	36.235640
sorry	0.9925196	0.6579300	26.072472
whoo-hoo	0.0351574	0.0000000	18.798498
heeey	0.0459750	0.0051002	14.452176
please	1.0357902	0.8160372	9.971548
umm	0.0784280	0.0280513	9.492213
right	2.8152940	2.4787129	8.178717
hee	0.0811324	0.0331515	7.945712
tee-hee	0.0216353	0.0000000	7.769404
hoooo	0.0000000	0.0204009	-6.946951
fer	0.0108177	0.0408019	-7.023786
ho	0.0459750	0.0969044	-7.041139
tch	0.0081132	0.0357016	-7.093157
svagatam	0.0135221	0.0484522	-7.907557
hmph	0.0703147	0.1402564	-8.985446
hic	0.0000000	0.0255012	-9.374175
whoa	0.0486794	0.1122051	-9.776038
anyways	0.0081132	0.0433520	-10.014093
uh	0.2028310	0.3315151	-11.888916
dammit	0.0216353	0.0765035	-12.308011
nah	0.0135221	0.0612028	-12.451141
hey	0.8086195	1.0786991	-14.756096
em	0.1568560	0.2907132	-15.432700
ha	0.2515104	0.4207692	-16.362141
heh	0.1162898	0.2499114	-18.929578
yo	0.0027044	0.0561026	-22.446146
aye	0.0622015	0.1861585	-24.519181
yeah	0.5733356	0.8823402	-25.124968
ah	0.3380516	0.5967272	-27.543559
damn	0.1487427	0.3340652	-27.685156
eh	0.0892456	0.2524615	-30.777606
hah	0.0189309	0.1249557	-33.018866
hup	0.0000000	0.0637529	-33.212573
ya	0.1487427	0.4768717	-68.554959
kupo	0.0730192	0.8874404	-307.367827

Females use some discourse markers more than males, for example “oh” (reciept), “sorry” (apologies) and feminine laughter (“hee”, “tee-hee”). Male characters use more expletives.

Acton (2011) also finds that, in real speech, females use “um” more than males and use “uh” less than males.

References

Acton, E.K., 2011. On gender differences in the distribution of um and uh. University of Pennsylvania working papers in linguistics, 17(2), p.2.