

**Univerzita Karlova v Praze**  
Filozofická fakulta  
Ústav anglického jazyka a didaktiky

Bakalářská práce

Bára Mikulenková

**Corpus Based Comparison of Gendered English in Speculative  
Fiction Videogames and Cinematography**

Korpusové srovnání genderované angličtiny ve videohrách a  
kinematografii v žánru spekulativní fikce

Praha, 2024

Vedoucí práce: Mgr. Ondřej Tichý, Ph.D.

## **Poděkování**

Tímto bych ráda poděkovala vedoucímu práce Mgr. Ondřeji Tichému, Ph.D. za velkou trpělivost a vstřícnost. Také děkuji svým blízkým za podporu.

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne

podpis

Souhlasím se zapůjčením bakalářské práce ke studijním účelům. I have no objections to the BA thesis being borrowed and used for study purposes.

## Abstrakt

Tato práce je zaměřena na komparativní korpusovou analýzu genderovaného jazyka, konkrétně porovnává genderovanou angličtinu videoher oproti filmům a televizním seriálům, spadajících pod fantasy/sci-fi žánr a vydaných v rozpětí od roku 2010 po současnost. V metodologické části jsou využity již existující korpusy filmů a seriálů ve výše zmiňovaném časovém období a žánrech, kromě toho je vytvořen i korpus nový, z výběru videoher odpovídajícím daným kritériím. Tento proces vychází primárně z výzkumu *Language, Gender and Videogames* od Frazera Heritage (2021), kdy základem je získání textu z daných her, a to buď extrakcí z počítačových souborů nebo přepisem, a následného vytvoření korpusu.

Tento výzkum se zaměřuje na několik jevů, jejichž analýza je opět částečně založena na metodologii popsané v *Language, Gender and Videogames*, aplikována na objemnější množství dat a rozšířena o komparativní aspekt. Jedná se nejprve o analýzu frekvence podstatných jmen *woman* a *man* (v singulárové i plurálové formě) a zájmen *she* and *he*. Dále jsou zkoumány kolokáty slov *man* a *woman* a klíčová slova, z nichž jsou vybrána mužská a ženská jména, a mužské a ženské sociální role. Tyto jevy jsou porovnávány mezi videohrami a filmy/seriály, a jejich analýza má za cíl otestovat hypotézu, že tato média, v rámci daných kritérií, jsou obě jazykově androcentrická a to ve srovnatelné míře.

Práce je vypracována v anglickém jazyce.

Klíčová slova: korpus, porovnání korpusů, kolokáty, kvantitativní lingvistika, gender, popkultura, videohry, kinematografie

## Abstract

This work focuses on comparative corpus analysis of gendered English language, specifically comparing video games and movies/television shows that are of fantasy/sci-fi genre and were published from 2010 to the present. For movies, an already established corpus, corresponding to said criteria was chosen, while the corpus for video games is built specifically for this research. This is done by obtaining the dialogue from a set of games, either by extraction from computer files or by transcription. This process is based primarily on *Language, Gender and Videogames* by Frazer Heritage (2021).

This research focuses on several phenomena, and is again partially based on the methodology presented in *Language, Gender and Videogames*, but applied to a larger amount of data and extended by the comparative aspect. We analyse the frequency of the nouns *woman* and *man* (in both singular and plural form) and the pronouns *she* and *he*. Another focus are the collocates of *man* and *woman*, and keywords referring to male and female names, and denoting male and female social actors. These phenomena are compared between video games and movies/TV shows, and the analysis aims to test the hypothesis that these media, within the chosen criteria, are both linguistically androcentric.

The paper is written in English.

Keywords: corpora, comparing corpora, collocates, quantitative linguistics, gender, pop-culture, video games, cinematography

## **List of abbreviations**

NPC - Non-Player Character

OPUS subcorpus - Subcorpus of *OpenSubtitles corpus 2018*

RPG -Role Playing Games

SFG corpus - *Speculative Fiction Games 2010 - 2018* corpus

## List of Figures

Figure 1. A game interface in <i>Sunless Sea</i> with a storylet window.....	19
Figure 2. The storylet window in <i>Sunless Sea</i> - detail.....	20
Figure 3. An NPC Dialogue in <i>Fallout: New Vegas</i> .....	21
Figure 4. Dialogue Choices in <i>Fallout: New Vegas</i> .....	22
Figure 5. A game interface with a dialogue log in <i>Divinity: Original Sin 2</i> .....	23
Figure 6. A dialogue log in <i>Divinity: Original Sin 2</i> - detail.....	23
Figure 7. <i>Dialogue Choices in Divinity: Original Sin 2 - detail</i> .....	24
Figure 8. Excerpt from <i>Horizon Zero Dawn</i> transcription, as seen on the Game Scripts Wiki (2024).....	32
Figure 9. A cropped view of the excerpt from <i>Horizon Zero Dawn</i> transcription, after being pasted to <i>Notepad++</i> .....	33
Figure 10. View of the Replace window in <i>Notepad++</i> with a regular expression search command.....	39
Figure 11. The excerpt from <i>Figure 9</i> after running the Replace function seen in <i>Figure 10</i> .....	34
Figure 12. Screenshot of the XML file with <i>Darkest Dungeon</i> subtitles.....	36
Figure 13. View of <i>Fallout New Vegas</i> dialogue lines in <i>Lazy Voice Finder</i> .....	37
Figure 14. <i>Cyberpunk 2077</i> in-game text with ANSI Encoding.....	39
Figure 15. <i>Cyberpunk 2077</i> in-game text with UTF-8 Encoding.....	39

## List of Tables

Table 1. Games in the SFG corpus, divided by the method of obtaining the text.....	30
Table 2. Raw frequency of selected lemmas in SFG corpus.....	40
Table 3. Raw frequency of selected lemmas in OPUS subcorpus.....	40
Table 4. Lemmas <i>man</i> and <i>woman</i> as a Subjects in the SFG Corpus and the OPUS Subcorpus.....	42
Table 5. Lemmas <i>man</i> and <i>woman</i> as an Object in the SFG Corpus and OPUS Subcorpus...	44
Table 6. Modifiers of lemmas <i>man</i> and <i>woman</i> in the SFG corpus.....	46
Table 7. Modifiers of lemmas <i>man</i> and <i>woman</i> in the OPUS subcorpus.....	47
Table 8. Gendered Keywords in the SFG corpus.....	50
Table 9. Gendered Keywords in the OPUS subcorpus.....	51
Table 10. Table of Keywords from the SFG corpus.....	65
Table 11. Table of Keywords from the OPUS subcorpus.....	68



# Table of contents

<b>1 Introduction</b> .....	<b>9</b>
<b>2 Theoretical Background</b> .....	<b>11</b>
2.1 Theory of gender.....	11
2.1.1 Sex and Gender.....	11
2.1.2 Linguistic gender.....	12
2.1.3 Gender in this work.....	13
2.2 Gender in media.....	13
2.2.1. Gender in video games.....	13
2.2.2 Gender in cinematography.....	15
2.3. Video game and film genres.....	15
2.3.1 Video game genres.....	15
2.3.2 Film Genres.....	17
2.4 Corpus Data: Country of Origin and Gender Distribution of Protagonists.....	18
2.5 Dialogue in Cinema and Games.....	18
2.5.1 Dialogue Forms in Video Games.....	19
2.5.2 Previous works on dialogue in video games and cinema.....	24
<b>3 Material and Method</b> .....	<b>26</b>
3.1 Objectives.....	26
3.2 Analytic approaches.....	26
3.2.1 Frequency.....	26
3.2.2 Keywords.....	27
3.2.3 Collocates.....	28
3.3 Material.....	28
3.3.1 Building the Video Game Corpus.....	29
3.3.2 Transcription.....	30
3.3.3 Extraction from the Game Files.....	34
<b>4 Analysis</b> .....	<b>40</b>
4.1 Frequency.....	40
4.2 Collocates.....	41
4.2.1 Man and Woman as Object and Subject.....	41
4.2.2 Modifiers of Man and Woman.....	45
4.3 Keywords.....	49
<b>5 Conclusion</b> .....	<b>55</b>
<b>Bibliography</b> .....	<b>58</b>
<b>Résumé</b> .....	<b>64</b>
<b>Appendix</b> .....	<b>65</b>

# 1 Introduction

Gendered language in media is certainly not a new research focus. As the discussion about gender evolves both within academia and the mainstream, so do approaches from related fields, and linguistics is no exception. But the use of corpus linguistics in analysing gender is still an approach with a lot of novel potential. And that is the aim of this paper - to use corpus-based methods to analyse gendered English in dialogue, specifically cinema and video game dialogue.

This work was inspired by “Language, Gender and Videogames: Using Corpora to Analyse the Representation of Gender in Fantasy Videogames” by Frazer Heritage (2021). The aim of this paper is to apply some of Heritage’s methodology on two corpora of a higher word count - one of video game dialogue and one of movie and TV show dialogue - while also utilising methods from other sources. The corpora utilised will be a videogame corpus compiled specifically for this research, called Speculative Fiction Games 2010 - 2024 (SFG), and a film and television subcorpus created from OpenSubtitles 2018 parallel corpus (OPUS subcorpus). A timeframe of the year 2010 to present was chosen, as this will show tendencies in media that are relevant to present day, while also allowing for a bigger sample size. The genres concerned will be fantasy and science fiction (“speculative fiction”), as majority of video games belong under these two genres. This approach allows us to reach novel conclusions while still building upon the foundation for videogame corpus based research Heritage lays down in his work.

A part of the work will be allotted to the creation of the videogame corpus, explaining the choices made in the process and the data gathering and processing methods. In the corpus analysis itself, we will compare the frequencies of pronouns “she” and “he”, collocates of words “woman” and “man,” and analyse a keyword list generated by comparison between the specialised corpora and the OpenSubtitles 2018 corpus. The aim is to learn whether there exists a gender bias in how characters of different genders are portrayed, and whether the two media forms differ in their portrayal of linguistic gender. The initial assumption that will be challenged, based upon general public opinions and previous studies, is that both game and cinema dialogue is androcentric, placing men in roles of bigger importance.

In addition to the conclusions that may offer some new insight into English language cinema and video games, the overarching other aim of this work is to contribute to the

growing video game research, both with the specific results, and a new corpus that may serve as a basis for more future research.

## 2 Theoretical Background

### 2.1 Theory of gender

#### 2.1.1 Sex and Gender

In order to discuss the various ways in which gender and language intertwine, it needs to be established what the term “gender” means, especially in contrast to the term “sex”. Since the discussion surrounding these concepts is constantly evolving, and due to gender’s social nature, the concrete distinctions between the two terms are not - and perhaps cannot - be fully defined, but there are still certain agreed upon differences. The simplest definition one may encounter is that sex is biological and gender social. However, this statement is not sufficiently accurate, as it misses vital nuances of the relationship between sex and gender. Sex is indeed a biological attribute - but, as Penelope Eckert and Sally McConnell-Ginet point out in *Language and Gender* (2013), “there is no single objective biological criterion for male or female sex,” as “sex is based in a combination of anatomical, endocrinal, and chromosomal features” (p. 2). There are people who do possess all biological attributes denoting one of the sexes, but there are also intersex people, who possess a mix of these attributes. Anne-Fausto Sterling in *Sexing the Body: Gender Politics and the Construction of Sexuality* asserts that the act of labelling someone man or woman is a social decision, which may be informed by scientific knowledge, but not defined by it (pp. 3). Sex is thus inherently linked with social optics, but still represents something that can be, under a pre established set of knowledge and beliefs, biologically categorised.

In the present day, the mostly accepted approach towards gender is the constructionist perspective, which sees gender as a social construct (Coates, 2013, pp. 6). In their various works, prominent gender studies philosopher, Judith Butler, presents gender as something that is performed, as “an identity tenuously constituted in time - an identity instituted through a stylized repetition of acts” and “instituted through the stylization of the body,” including “bodily gestures, movements, and enactments” (1988, p. 519). This idea of gender as a performance, following a set of attributes associated with either gender, has become widely accepted within gender and feminist theory. Gender is “the social elaboration of biological sex, [carrying] biological difference into domains in which it is completely irrelevant” (Eckert & McConnell-Ginet, 2013, pp. 2). Gender encompasses one’s identity in a way sex does not, it includes not only the body in which a person was born, but how the person feels within the societal context and how the society perceives them. So when we discuss gender within corpus research, we refer to the identity people (real or fictional) identify with. In

games, this can be recognized mainly by how the characters refer to themselves and others refer to them. A way a character looks and behaves can be also taken into consideration, but it is important not to fall into the trap of easy stereotypes or personal bias. Lastly, there may be official statements by the game's creators online, in media or in supplement documentation (eg. game manuals) that may make a character's gender clear.

### 2.1.2 Linguistic gender

Eckert and McConnell-Ginet (2013) believe it was Robin Lakoff's article "Language and woman's place" published in 1973, during second wave feminism, that changed how language's relation to gender was being discussed. The 1970s' dominant structuralist perspective in social science saw society as a set of several social categories. This meant that "male" and "female" were seen as strictly defined, opposite concepts (pp. 37). Lakoff and other feminists rejected the unchanging nature of the structuralist movement that did not account for any meaningful change or disruption of social order. In her controversial paper, she argued that the "woman's language," marked by hedges, "empty adjectives" or exaggerated intonation, was weak and unassertive. This approach to language is called the *deficit* approach, and it was heavily challenged, as the implication is that the language of women is inherently defective (Coates, 2013, pp. 6). In the years following the Lakoff article, there developed two paradigmatic approaches to gender in language - the *dominance* and *difference* approaches (Eckert and McConnell-Ginet, 2013, pp. 39). The *dominance* approach sees women as an oppressed group, whose subordination under men is enacted through linguistic practice. Female oppression is perpetuated and sustained by all participants in discourse, including both men and women (Coates, 2013, pp. 6).

The *difference* approach asserts that women and men speak in different ways because they have different relations to their language (Eckert and McConnell-Ginet, 2013, pp. 39). Women and men were seen as belonging to two different subcultures (Coates, 2013, pp. 6). The positive aspect of this approach is that "it allows women's talk to be examined outside a framework of oppression or powerlessness" (Coates, 2013, p. 6).

The most recent is the *dynamic* approach that utilises a social constructionist perspective, discussed in the 2.1.1 Gender and Sex section. It is the *dynamic* approach that is most prominent in the present day, and is also the most flexible, rejecting concrete dichotomies and seeing gender as a wider concept.

### **2.1.3 Gender in this work**

This particular work will discuss gendered language through binary optics, focusing on distinctions between characters presented within the context of the works they appear in as either male or female. However, the goal is not to promote the binary approach to gender as the singular proper way, nor deny or ignore the existence of non-binary and/or gender non-conforming people, who “resist gender dichotomies altogether” (Eckert & McConnell-Ginet, 2013, p. 3). The reasons for this omission are, firstly, difficulties in distinguishing pronouns. Unless one is very familiar with the works in the corpus, it is challenging to distinguish between singular ‘they’ and plural ‘they’, even in concordance analysis - in this case, the familiarity with works present in the OPUS subcorpus was not sufficient. Secondly, as of yet, there exists no noun to complement the gendered man/woman distinction. The noun ‘person’ is semantically too wide, and the compound “non-binary” was not present in the corpora whatsoever. The similar applies to the adjective “transgender”, which only appeared once in the OPUS subcorpus and again was not present in the SFG corpus. This leads to the third reason: after analysing the corpora and based on the familiarity with the games present in the SFG corpus, it can be concluded that the representation of openly non-binary people, and transgender people in general, within the chosen media and time frame, is insufficient for any meaningful analysis. This is not an isolated issue, as a similar conclusion was drawn in the 2023 article *Gender bias in video game dialogue* (Rennick et al.). However, there already exist films, television shows and games that include transgender characters, and should this trend continue, future works focusing on gendered language within these media forms will have an opportunity for a wider and more inclusive research.

## **2.2 Gender in media**

The relationship between gender and media has been a discussed topic for years. There are two interconnected layers of this particular discussion - how gender is portrayed within the media themselves, and what role gender plays in how the people consuming or creating the media are treated.

### **2.2.1. Gender in video games**

2014 marked the beginning of Gamergate - a harassment campaign initiated as a reaction to feminist work of media critic Anita Sarkeesian, lasting into 2015 and targeting a number of women involved with game development and consumption (Mortensen &

Sihvonen, 2020). Gamergate, taking place primarily on the Internet and reaching mainstream attention, unravelled to everyone the sexism prevalent in the gaming community and shaped the landscape of feminist discourse about media in the digital age. Albeit the campaign eventually ended, the misogynistic opinions still echo into the present.

The percentage of female gamers in the United States in the year 2023 was 46% (Clement, 2023), yet many studies show that women are mistreated and ostracized within the gaming community (Rennick et al., 2023, pp. 2). Both non-male gamers and developers are at a risk of encountering gendered harassment, including sexual aggression (deWinter & Kocurek, 2017, pp. 61). A 2016 study by Allison McDaniel showed that out of 141 female players of online first-person shooter games, 107 (75.9%) experienced verbal harassment or discrimination because of their gender. Moreover, only 23 (16.9%) of the surveyed women felt that the tendency of other players present during the harassment is to actively help defend the target (pp. 29 - 31). There is a degree of normalised sexist behaviour within the community that so far does not seem to change. The situation in game development also implies systematic issues: a 2022 survey on the State of the Game Industry revealed, as paraphrased by Winter and Masters (2023), that out of 2700 game developers, only 20% identified as women and 4% as non-binary. Moreover, non-male employees have to often deal with discriminatory conditions in the workplace. Even famous developers like Riot Games or Blizzard have dealt with gender-based discrimination (Winter & Masters, 2023, p. 104). All this makes video games a media form surrounded by unprecedented deep social tensions that are important to keep in mind when discussing other aspects of this medium.

The portrayal of women in video games is closely connected with how women are seen and treated by both male developers and consumers. It was Anita Sarkeesian's work on differences between male and female characters in video games that led to her harassment, and it is arguably the entitlement of some male players to female bodies that has often led to an over-sexualized portrayal of women in games. Female characters are often victims of stereotyping - 2015 study by Xeniya Kondrat surveyed 234 game players, out of which 180 (76.9%) believed female gender to be stereotyped in video games (pp. 183). The particular stereotypes mentioned were the sexual objectification of women in games and infrequent appearance of female characters and protagonists (pp. 184). Downs and Smith (2009) found out that within 60 chosen videogames, female characters were underrepresented, women making 14.3% out of 489 characters and men making 85.7%. Women were also significantly more likely to be shown partially nude, have an unrealistic body, and wear revealing clothing (pp. 721). Rennick et al. (2023) analysed 50 roleplaying games, and out of characters labelled

male or female, 29.37% percent were female and 70.63% were male (pp. 5). A number of other studies came to similar conclusions - female characters are often underrepresented, and when they are present, they are more likely to be sexualized (Rennick et al., 2023).

## **2.2.2 Gender in cinematography**

When it comes to the film industry, the situation is not dissimilar to that of video games. A report on European films in the years 2006 to 2013 from EWA Network (2015) shows that 21% of films in a set of chosen countries were directed by women, and 84% of funding resources went to films not directed by women. Moreover, female graduates from film schools made a 44% of graduates, but female directors made only 24% of all film directors (pp. 8). Another report (Gender Equity Policy Analysis Project, 2024) found that the share of women in key creative positions was on average only 23.1% (pp. 29).

Films and television shows also demonstrate a statistical imbalance in portrayal of female and male characters. Haris et al. shows that within selected 30 movies, in years 2015 to 2019, 43.9% of characters were female. This may seem like there is a trend towards inclusion, but in 2023, within the Top Grossing U.S. Films, the percentage of female speaking roles was 35%, and the percentage of female protagonists was 28% (Lauzen, 2024, pp. 1). Moreover, when Anderson and Daniels (2016) analysed 2000 film scripts, they found out that only 22% of the films had a female lead. In 1513 films (75.65% of the 2000 films), men's lines made at least 60% out of the whole dialogue. In 307 films, at least 90% of all dialogue was men's, in comparison to only 9 films where at least 90% of dialogue was women's. Similarly to video games, there is also a degree of stereotyping - for example, studies found that "female characters occupied a more limited range of occupations, held lower-status positions, and wielded less power than men have" (Lauzen & Dozier, 2005, p. 438); there is also still a tendency to portray women as needing a strong man to help them (Kumari & Joshi, 2015, pp. 46).

## **2.3. Video game and film genres**

### **2.3.1 Video game genres**

There is still not much linguistic research focusing on video games, compared to other media forms. One reason is undeniably the relative youth of games, as the first video game made just for entertainment, Tennis for Two, was created in 1958 (Donovan, 2010, p. 8). This means that games have not yet had much opportunity to become an established part of the



sociolinguistics and corpus research tradition. Moreover, video games are still often considered to only serve as entertainment, with no linguistic value. Although their reputation has somewhat ameliorated within the last decade, the image of an asocial teenager, screaming at the screen while virtually shooting monsters, still did not fully disappear from the public subconsciousness. And yet, games as they are now offer a wide array of not only genres, but registers, writing styles, and ways the player can interact with the language. Thus, the aim of the following section is to offer a brief categorization of games, and explanation of video game dialogue with examples, in order to familiarise the reader with some basic game concepts.

Games can be single-player, where the player plays through the game alone, or multiplayer, where more players join together to play, either cooperatively or competitively. Some games can be played in both ways, some offer only one. This paper's SFG corpus only includes games with single-player mode, as these are typically more dialogue-heavy and thus the more productive option for linguistic research.

Games are a wide medium, with different settings, mechanics, and overall look. TV Tropes (2024), a popular website primarily focused on analysing different tropes and genres as seen in pop culture and individual works, divides video games into several general genres and subgenres. The most prominent ones are action, adventure, action-adventure, casual, role-playing (abbreviated as RPG), simulation and strategy. Note that the general video game genres combine relatively freely with "aesthetic" (TV Tropes, 2024) genres found in other forms of fiction, as is for example fantasy, sci-fi or horror.

Action games are "primarily about physical challenges" (TV Tropes, 2024). They typically contain less dialogue and/or written word than adventure or RPG games, but can still have a full-fledged story and dialogues, as is the case of *God of War* or *Metal Gear Revengeance* included in the SFG Corpus.

Adventure games are not focused heavily on action, instead centre exploration or puzzle solving (TV Tropes, 2024). The focus on dialogue differs widely between specific games, ranging from none (eg. *Machinarium*) to more than 100,000 words (eg. the Higurashi No Naku Koro Ni games (When They Cry Wiki, 2024)). No works were chosen for the SFG corpus, as there are not many purely adventure games of suitable word count from the year 2010 to present whose dialogue data are readily accessible.

Action-adventure games blend the two previous genres together. The focus on story and dialogue is typically bigger than in action games, but just as with adventure games, the differences between specific games are very significant. Action-Adventure games chosen for

this research are *Alice: Madness Returns* and *Portal 2*, and to some extent also *Bioshock 2* and *Bioshock Infinite* (these two games are a heavy mix of action-adventure, action and RPG).

Casual games primarily serve as a way to relax, and are relatively simple to learn and play. This genre is often utilised in mobile and website-based games, or in family-oriented console games (eg. *Wii Sports*). Strategy games focus on strategic decisions and planning, for example in games like *Civilization*, where the player tries to build a successful nation. Simulation games are then games simulating a certain experience - the most known is possibly the life simulator series *The Sims*. All three genres do not offer much direct speech and dialogues in general, which makes them unsuitable for this particular research.

Role Playing Games are one of the most popular genres. Close relatives to tabletop games such as *Dungeons and Dragons*, they are games in which “the player controls a character or party of characters in a statistically abstracted way” (TV Tropes, 2024). They generally offer substantial freedom to the player in how the game plays out, and boast a high number of dialogues, intricate, sometimes branching storylines and customisation of gameplay. They are also very popular within the mainstream, with active fan participation. Because of this, most of the games in SFG corpus belong under the RPG genre, including games with less dialogue and player control (*Dark Souls* series, *Darkest Dungeon*), dialogue heavy games with minimal player influence (*Shadow of War*) and robust tabletop-like games with an enormous degree of player freedom (*Divinity: Original Sin 1 and 2*, *Cyberpunk 2077*).

As implied earlier, the genres can overlap and mix. This can lead to hard to categorise games, like *Death Stranding*, described by TV Tropes (2024) as “post-apocalyptic sci-fi stealth-action-horror-sandbox-simulator.” It is this possibility to combine and extend beyond genres, to invent new approaches to present the stories told and engage the player, that makes video games a field worth of interest.

### **2.3.2 Film Genres**

Cinema and its approach to genres is more straightforward. The portal IMDb (2024) lists the following genres: action, adventure, animation, biography, comedy, crime, documentary, drama, family, fantasy, horror, music, musical, mystery, romance, sci-fi, short, sport, thriller, war and western genre. It also includes one film specific genre: film-noir, and television specific genres: game-show, news, reality-tv and talk-show. Out of the selection, fantasy and sci-fi were chosen because they are the prominent “aesthetic” genres within

games and still heavily present in cinema. The choice was also made to work with both television shows and films, as limiting the research to only one of the two would unnecessarily limit the resulting word count, considering there are no drawbacks for including both forms.

## **2.4 Corpus Data: Country of Origin and Gender Distribution of Protagonists**

For a more complete picture, it would be suitable to show the countries of the movie and game studios whose work compiles the corpora we will work with. The gender of the main character may also be relevant information. However, the OpenSubtitles 2018 parallel corpus has a significant handicap - it is difficult to analyse the data by its origin. The metadata of the corpus in Sketch Engine are tagged “countries the film was shot in”, “film name”, “film release date”, “full film name” and “genres of the film”, but the subcorpus cannot be divided by the country due to Sketch Engine’s technical limitations. Moreover, some movie and TV show names are missing. As such, rather than manually finding details about every movie or TV show in the subcorpus, it has been decided to omit the OPUS subcorpus in this section, and only focus on the SFG corpus, as its smaller size enables us to analyse it properly.

The countries where the developer studio is based are the following: USA (11 games), Japan (8 games). Canada (3 games), Belgium, Netherlands, Poland (2 games each) and China (1 game). The gender distribution of the game protagonist is as follows: 9 games have a male protagonist, 5 have a female protagonist and 14 allow the player to choose. A special case is *Darkest Dungeon*, which does not have a singular protagonist. None of the games had an option for the player to play as a non-binary or gender non-conforming character. This implies a slight bias towards male protagonists in our data set, which may reflect in the corpus analysis.

## **2.5 Dialogue in Cinema and Games**

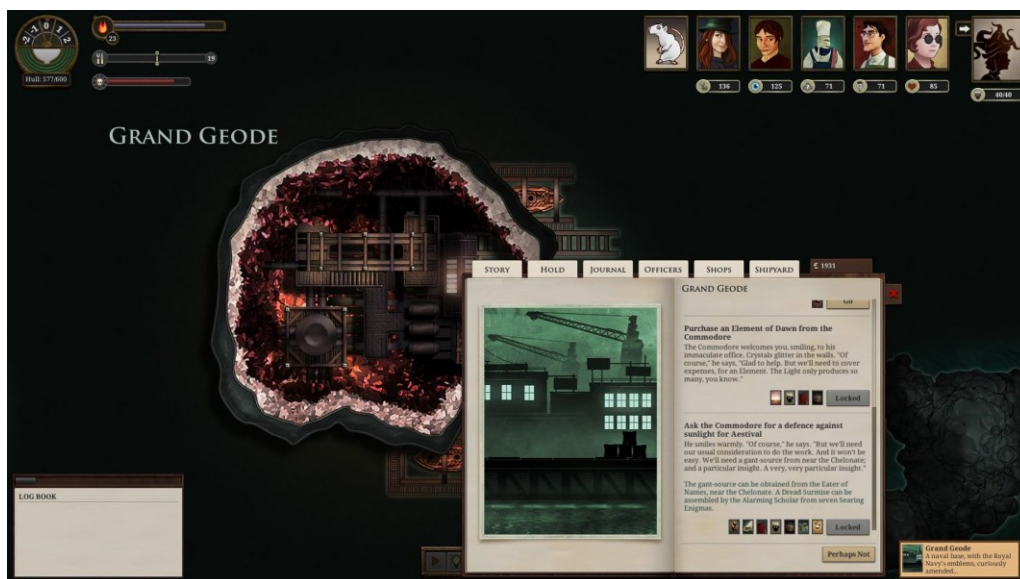
Dialogue in film is “written-to-be-spoken-as-if-not-written.” (Gregory & Carroll, 1978, p. 42). Busso and Vignozzi (2018) note that “corpus-based studies have proved that spontaneous conversation and scripted dialogues are very similar in nature” (p. 71). In cinema, the majority of text within the work will be a direct speech, with only minimal interjections (eg. by a narrator). This is why the games for the SFG corpus had to contain standart dialogue akin to film dialogue, which does not always apply.

## 2.5.1 Dialogue Forms in Video Games

A typical game will have diegetic text (primarily dialogues, but also eg. letters, signs or even books the player can read) and non-diegetic text (tutorial messages, item descriptions). But sometimes, the diegetic and non-diegetic can intertwine significantly. For example, the game *Sunless Sea* describes all in-universe events in written form, using the second person when referring to the player character and quotation marks for direct speech (see Figures 1 and 2).

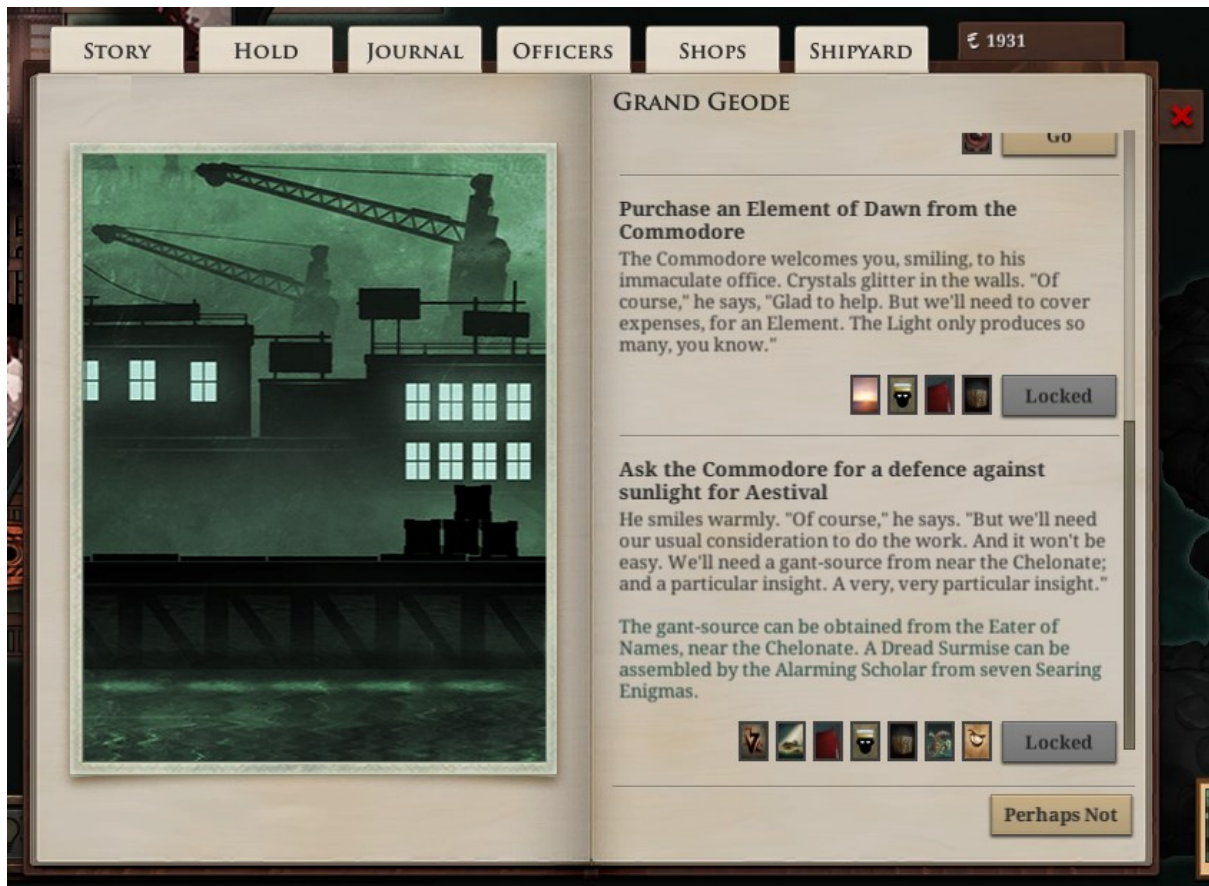
**Figure 1**

*A game interface in Sunless Sea with a storylet window*



**Figure 2**

*The storylet window in Sunless Sea - detail*

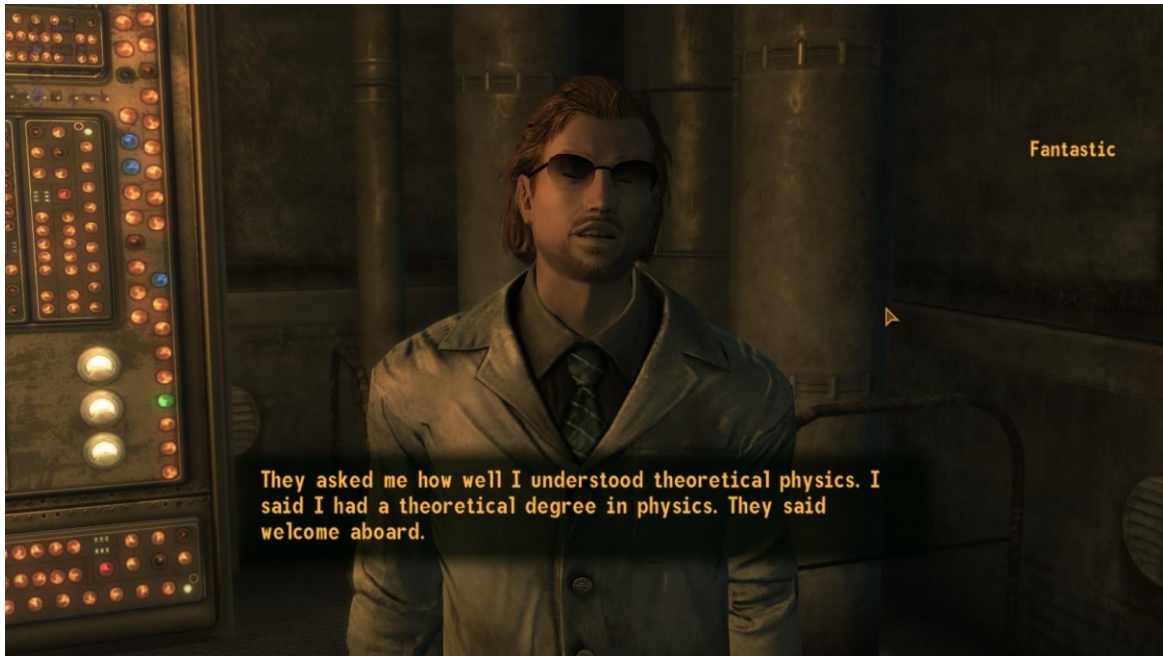


This makes experiencing the story more akin to reading a book rather than watching a film or a television show. Thus, games like *Sunless Sea* were not deemed suitable for this research.

An opposite type of games are those where all events are visually shown to the player and all dialogue is said (and subtitled) or written. That is the case of for example *Fallout: New Vegas*. Here, there are no other descriptions outside the dialogue and the communicative situations happen similarly as they would in a film, except that they can be influenced by the player's decisions.

**Figure 3**

*An NPC Dialogue in Fallout: New Vegas*



**Figure 4**

*Dialogue Choices in Fallout: New Vegas*

Note: The player influences dialogue by choosing from an offered list of options.



A combination of these modes can be found in Divinity: Original Sin 2. The standard dialogue is often interjected by a narrator, who adds context and descriptions of the immediate surrounding and actors in the communicative situation (see Figures 5 - 7) However, this did not disqualify the game for the corpus, as the non-diegetic text could be separated and deleted during the gathering of data (more detailed explanation can be found in Chapter 2).

### Figure 5

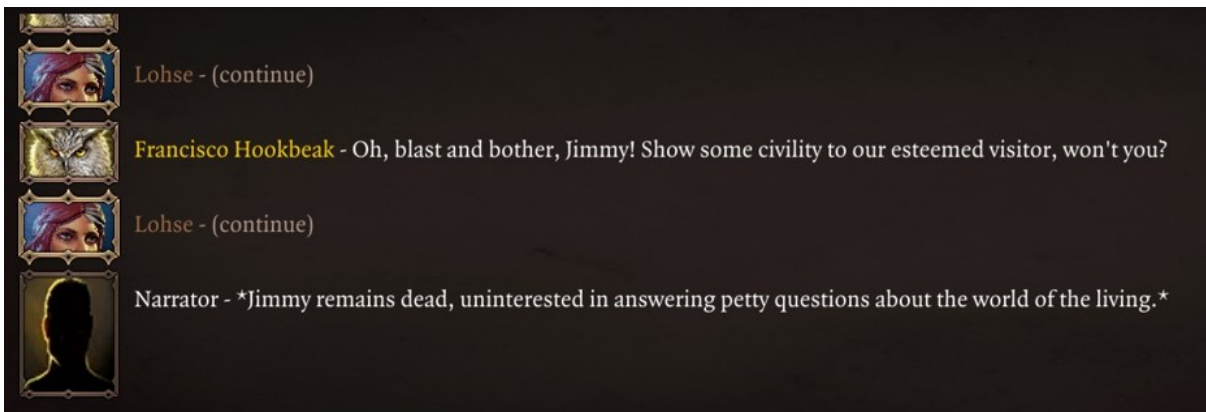
*A game interface with a dialogue log in Divinity: Original Sin 2*



**Figure 6**

*A dialogue log in Divinity: Original Sin 2 - detail*

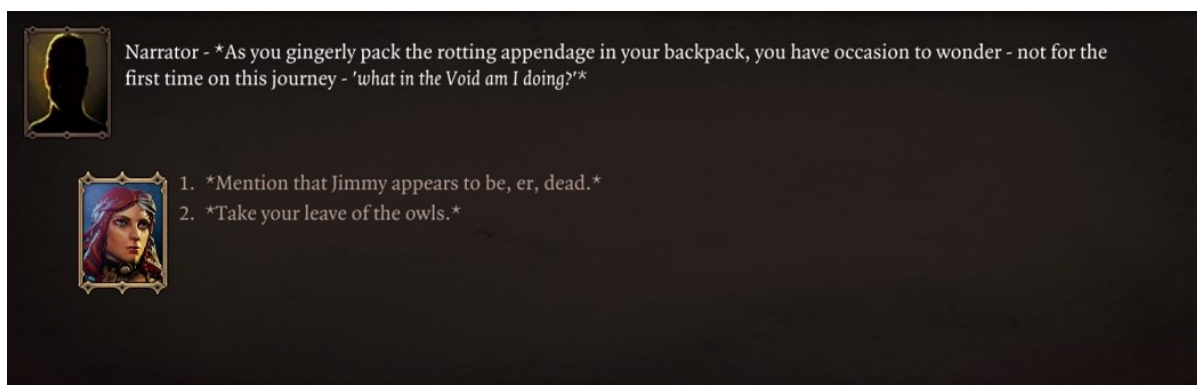
Note: dialogue log in Divinity: Original Sin enables the player to read through the dialogue they are currently in up until the last said line, including the NPC lines, narrator lines and the player characters lines.



**Figure 7**

*Dialogue Choices in Divinity: Original Sin 2 - detail*





Narrator - \*As you gingerly pack the rotting appendage in your backpack, you have occasion to wonder - not for the first time on this journey - 'what in the Void am I doing?'

1. \*Mention that Jimmy appears to be, er, dead.\*  
2. \*Take your leave of the owls.\*

## 2.5.2 Previous works on dialogue in video games and cinema

As stated in Chapter 1.0, this work takes inspiration from Heritage's *Language, Gender and Videogames*. Here he compiled 10 games into the VG2014 corpus with a word count of 327 027 (pp. 116). In a collocate analysis, he found out that the pronoun *he* was "more likely to occur with particular nouns and verbs that denoted violence" and the pronoun *she* was "more likely to occur with social actors and language denoting familial and platonic relationships." (p. 139). The collocates of *she* also exhibited a tendency to focus on the physical aesthetics of women, which did not happen with the collocates of *he*. Moreover, the analysis revealed that male characters were often the agents of an action, as they were more likely to "do verbs", while female characters were the passive receivers, more likely to have "verbs done" to them (pp.139). But Heritage also points out that "a number of concordance lines demonstrated that some women were beginning to be represented as physically strong and could enact physical violence (p. 140).

Another study using corpus to analyse video game dialogue is *Gender bias in video game dialogue* by Rennick, Clinton et al. (2023). They compiled a corpus from 50 games worth 6 280 892 words of dialogue (pp. 5). In the study, they coded all characters as either male, female, or other genders. The female dialogue ranged from 6% to 80% in individual games. As mentioned in Chapter 2.2.1, out of characters labelled male or female, 29.37% percent were female, 70.63% were male. This means that female characters did not talk less, but were simply less featured. Female characters also displayed more gratitude, used more hedging, apologised more and swore less (pp. 8). In "Gender Stereotypes in Films Language: Corpus Assisted Analysis", Busso and Vignozzi (2017) also focus on the speakers in communicative situations, but in the context of cinema. The analysed films were romantic comedies, which makes the work less relevant for this research, but it still offers some degree of insight. The women tended to speak about shopping, cleaning personal care and family,

men then discussed money, sports, work and male friendship. Women also used more polite forms and fewer swear words than men (pp. 73), which is consistent with the findings by Rennick et al. (2023).

These studies seem to indicate that there are some categorizable differences in how characters of different genres talk, are talked about and are portrayed in general. As such, the assumption of this paper is that our findings will also demonstrate a significant degree of gender difference.

## 3 Material and Method

### 3.1 Objectives

The goal of this work is to analyse and compare the SFG corpus and OPUS subcorpus by focusing at several aspects:

- 1) The raw frequency of the lemmas *man* and *woman* and pronouns *he* and *she*.
- 2) Collocative analysis of lemmas *man* and *woman*, analysing verbs with said lemmas as object and subject, and modifiers of the lemmas.
- 3) Keyword analysis, focusing on gendered keywords.

All analyses will be made via the web based tool Sketch Engine. Concordance analysis will be also used in order to further understand the context of different linguistic instances and patterns.

There are three main assumptions that will be addressed and verified during the analysis:

- 1) Words referring to the male gender (lemma *man*, lemma *he*, male names and male social actors within a set of keywords) will be more frequent in both corpora than words referring to the female gender.
- 2) Both genders will exhibit clear collocation patterns that would imply a significantly gendered and/or stereotypical view of men and women in both media forms.
- 3) The language used to refer to women will have a tendency to sexualize them, as opposed to language used about men.

### 3.2 Analytic approaches

As stated earlier, our research utilises several different corpus analysis methods: frequency, keywords, collocation and concordance. Aside from concordance analysis, these methods may differ significantly between specific tools and research, and as such, the following section aims to establish these approaches as they are employed within the context of this work.

#### 3.2.1 Frequency

Frequency is perhaps one of the most straightforward, but also “one of the most central concepts underpinning the analysis of corpora” (Baker, 2023, p. 81). It denotes how

frequent an appearance of a word is within the corpus. There are several approaches to frequencies, but perhaps the most prominent are: the raw/absolute frequency, that shows the number of occurrences of the word without any modifiers (Sketch Engine, 2024, Frequency Section); the percentage frequency, which according to Paul Baker in *Using Corpora in Discourse Analysis* (2023) denotes “the percentage contribution that each word makes towards the corpus” (p. 86); and a relative frequency, which shows how many occurrences appear per one million words. The choice of frequency calculations depends on the size of the corpus and the researcher's goal. Baker (2023) sees the percentage frequency as a more sensible way to compare data than raw frequency, especially when doing comparative analysis using datasets of different sizes (pp. 51), but unfortunately, Sketch Engine does not offer this tool. For that reason, this research will only utilise the raw frequency, which Baker (2023) deems reasonable when working with a corpora of sample size bigger than one million words (pp. 86).

### 3.2.2 Keywords

Another concept that will be important during the research is keyness. A keyword list may be similar to a word list at a first glance, but as noted by Heritage (2021), “keyword lists differ to wordlists in that they rely on statistical comparisons” (p. 71). More specifically, keyness analysis is a comparison of two wordlists against each other in order to “determine which words occur statistically more often in word list A when compared with word list B and vice versa” (Baker, 2023, p.165). The resulting list of words that are present more often than expected is the keyword list. This makes the keyword lists a possible way to show “what is statistically salient within a corpus and where the language in one corpus differs to the corpus that it is being compared to” (Heritage, 2021, p. 72).

As Heritage (2021, p. 72) points out, when choosing the reference corpus, it is preferable to compare data within the same mode. In the case of this paper, both the specialised corpora and the reference corpus are compiled from direct dialogue, as the chosen reference corpus is the OpenSubtitles 2018 parallel corpus.

In Sketch Engine documentation (2024, Simple Maths) describes the keyness score used by Sketch Engine, “simple maths.” The formula is as follows:

$$\text{Frequency per million in the focus corpus} + N$$

---

$$\text{Frequency per million in the reference corpus} + N$$

N stands for “smoothing parameter”, with a default value of 1. The user can change the value N, with higher values shifting focus to higher frequency and thus more common words, and lower values focusing on low-frequency, rarer words. For this particular research, after experimentation, the value 100 was chosen as the most suitable.

### 3.2.3 Collocates

Baker (2023) defines collocation as a phenomena when “a word regularly appears near another word, and the relationship is statistically significant in some way” (p. 135 - 136). In its Word Sketch module, Sketch Engine uses logDice to express the typicality of the collocation. LogDice score is not affected by a corpus size and is not as heavily influenced by raw frequency, which makes it preferable to the association scores T-score and MI-score, which are both based on the frequency of the collocates (Rychlý, 2008). In this research, we will work with collocates of all logDice scores, not setting a minimal limit, as the size of the SFG corpus is already quite small and limiting it further may lead to a loss of valuable data for comparison with the bigger OPUS subcorpus. Moreover, it should be noted that Word Sketch does not consider the collocation window when calculating collocates, meaning that during our analysis, logDice is the singular parameter used.

Word Sketch offers a division of collocates into several categories, according to grammatical relations. For this research, we will work with three: modifiers of the searched lemmas, verbs with the lemma as object and verbs with the lemma as subject.

## 3.3 Material

As stated in Theoretical background, in order to make a corpus based comparison, this paper uses two corpora - one for movies and TV shows and one for video games. The former is a subcorpus of OpenSubtitles 2018 parallel corpus as processed by Sketch Engine. The choice to utilise this particular corpus and not another (eg. Film Corpus from Sketch Engine) was made because the genres and years of publication are clearly labelled in the metadata, allowing for a creation of a subcorpus following the criteria established for this research, that would also be of sufficient size. This subcorpus, labelled OPUS 2010 - 2018 (Speculative Fiction), contains all text types of the whole corpus tagged with fantasy or sci-fi genre and the years from 2010 to 2018, containing 57,592,746 tokens and around 43,766,64 words, representing 3.6% of the whole Open Subtitles corpus.

It should be noted there are obvious issues with using the Open Subtitles corpus. Firstly, it is primarily a parallel corpus used for multilingual research, but fortunately, this does not pose an issue as it offers the same options and range of data as a monolingual corpus. Another visible issue is that the corpus data end with the year 2018, while the SFG corpus ends with the year 2024. This discrepancy has been chosen to ignore, as the dataset offered by OPUS subcorpus is still suitable and the missing six years have been deemed not a significant enough loss of data for the scope of this research.

The SFG corpus is significantly smaller, but since it has been built for the purposes of this paper, it offers a familiarity with the material and thus more context. It contains 6,142,076 tokens and 4,904,458 words, and is made out of 29 documents.

### 3.3.1 Building the Video Game Corpus

The selection of the games was an important part of the research. As explained in Theoretical background, we have chosen fantasy and sci-fi games, from which the specific titles have been chosen according to their availability and cultural relevance. Availability means that the game dialogue can be obtained without significant complications, while cultural relevance estimates whether the game was/is influential within the gaming landscape. This decision was based on the estimated number of sales for the individual games (the majority of selected games boasted at least one million copies sold), but also on personal exposure to the video game community, journalism and discourse.

While both factors have been considered, the availability of the games had a bigger priority. For example, there are no suitable sources that would estimate the number of sales for *Alice: Madness Returns*, and the game is not commonly discussed in online spaces in the present years, outside of dedicated fan sites and forums. Yet, the transcription for the game was readily available, making it sufficiently suitable. Similarly, the *Dragon Age* series was heavily influential in the RPG genre, and is still being actively discussed today, but due to not having the access to the game files, the dialogue from *Dragon Age 2* and *Dragon Age: Inquisition* could not be included.

The games were also chosen to represent sci-fi and fantasy genres at least semi-equally. 51.7% percent of the documents are of sci-fi genre, while 48.3% was of fantasy genre. The token distribution is slightly less balanced, with sci-fi making 55.2% percent of the overall token distribution.

There were three ways of gathering the files needed for the corpus. Manual extraction for this research, manual extraction by a 3rd party and a transcription by a 3rd party.

**Table 1**

*Games in the SFG corpus, divided by the method of obtaining the text*

Manual Extraction (researcher)	Manual Extraction (3rd party)	Transcription (3rd party)
Darkest Dungeon	Cyberpunk 2077	Alice: Madness Returns
Divinity: Original Sin	Dark Souls	Bioshock 2
Divinity: Original Sin 2	Dark Souls 2	Bioshock Infinite
Fallout New Vegas	Dark Souls 3	Bloodborne
Mass Effect 2	Elden Ring	God of War
Mass Effect 3	Fallout 4	Horizon Zero Dawn
The Witcher 3	Portal 2	Horizon Zero Dawn 2
Transistor	The Elder Scrolls IV: Skyrim	Metal Gear Rising: Revengeance
		Resident Evil 2 (remake)
		Shadow of Mordor
		Shadow of War
		The Last Of Us II

### 3.3.2 Transcription

Heritage (2021) chose to mostly avoid transcription because of concerns about their accuracy (pp. 105), but this paper, which is allotted less resources, does make use of them, as their quality was deemed sufficient and their inclusion attributes to the greater size and variety of SFG corpus. Still, there are issues unique to the transcribed texts that need to be addressed. The process of transcribing differs with the individual cases, but the standard approach is to watch a video of someone playing through the game. This means that if the person playing misses some content, the transcription will not include it. This is most prominent in the case of open world games such as *Death Stranding*, where a substantial number of dialogues is encountered outside of the main story. Moreover, transcripts may include typos and other mistakes. On the other hand, transcripts enable us to get the text from games whose dialogue have not been published online, and/or which we either do not own or are released on platforms that do not offer a simple file extraction (eg. consoles such as

PlayStation 4). The editing of the transcribed script is also very straightforward, as demonstrated in the following paragraphs.

The transcripts for *Bioshock 2*, *God of War*, *Horizon Zero Dawn*, *Horizon Zero Dawn 2*, *Metal Gear Rising: Revengeance*, *Resident Evil 2 Remake*, *Shadow of Mordor*, *Shadow of War* and *The Last of Us II* were taken from Game Scripts Wiki Blog (2018, 2019, 2020, 2022) with the permission of the admin of the blog and the creator of the transcripts, Reddit user *u/Snow\_Guard*. *Bioshock Infinite* transcript was taken from GameFAQs (Summers, 2014) with the permission from the author Patrick Summers. Dialogues for *Alice: Madness Returns* was extracted from the fan-administered site Alice Wiki (2024) and the *Bloodborne* script was compiled by the Reddit user *u/Vocazone\_2* (2020) with explicit permission to be used by other users.

All transcribed text used share similar characteristics: as with a film script, all spoken lines were preceded by the name of the character speaking. Many also had non-diegetic notes with scene descriptions, transitions and character modes of speaking. In order to get only the diegetic dialogue, every file was opened in Notepad++ (2024) as a plain text, and the Search and Replace function was used to strip the text of all non-diegetic text and markers. This was achieved with usage of both textual search and regular expressions (RegEx)<sup>1</sup> search.

---

<sup>1</sup> As defined by the site *TechTerms* (2024), regular expression “is a search pattern used for matching one or more characters within a string. It can match specific characters, wildcards, and ranges of characters.”



## Figure 8

Excerpt from Horizon Zero Dawn transcription, as seen on the Game Scripts Wiki (2024)

*[We see stunning views of wildlife and a hut in the middle of the forest. From the hut comes a strong man with a braided beard. A man has a child sitting in a shoulder bag. There is still snow around, but spring is already beginning to come into its reign. A man carries a child and speaks to her]*

**Rost:** What's that now? Don't like the cold? Can't stay in today. We have a ritual to perform, you and I. Here wear this. It belonged to my daughter. Good. Today I speak your name, girl. But - will the Goddess speak it back? Normally it would be the mother who declares... if you had one. The whole village would attend, and Matriarchs would perform the ritual. But... we are outcasts. Even so, we keep the tribe's rituals. Otherwise we might become like the faithless Old Ones, who turned their backs on the Goddess. But their wickedness doomed them. To us were left the splendors of creation. Beasts of air, water, earth... and steel. It is one thing to hunt a beast, another to hunt a machine. You must be humble and respect their power. I will teach you this, one day...

*[They climb a high mountain to a small shrine.]*

**Rost:** High Matriarch Teersa? What is she doing here? Does she mean to forbid the ritual?

**Teersa:** No-no-no, off your knees! It's nearly time. And yes, you may speak to me!

**Rost:** You came to bless the naming?

**Teersa:** Have not six months gone by since we entrusted her to you?

**Rost:** But we are outcasts.

**Teersa:** You by choice, and she, well... I'm a High Matriarch, Rost. I bless whom I choose.

**Rost:** Then... you honor us.

**Figure 9**

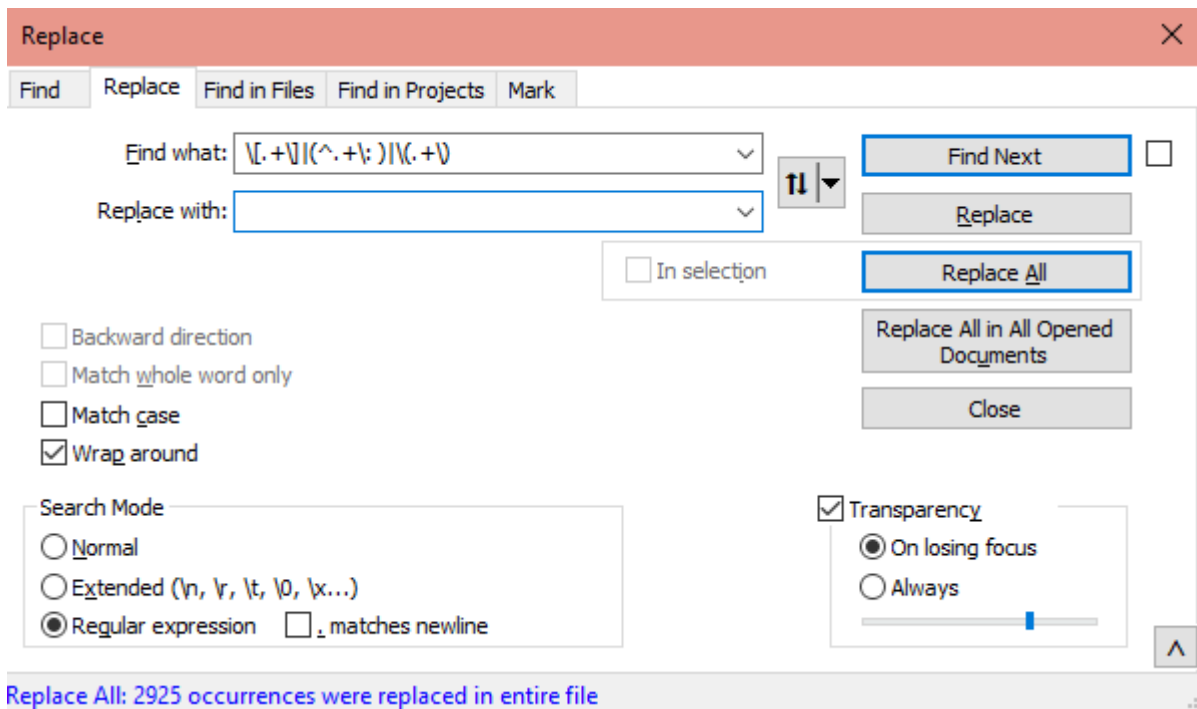
*A cropped view of the excerpt from Horizon Zero Dawn transcription, after being pasted to Notepad++*

```
1 [(We see stunning views of wildlife and a hut in the middle of the forest. From the hut cor
2
3 Rost: What's that now? Don't like the cold? Can't stay in today. We have a ritual to perf
4
5 [They climb a high mountain to a small shrine.]
6
7 Rost: High Matriarch Teersa? What is she doing here? Does she mean to forbid the ritual?
8
9 Teersa: No-no-no, off your knees! It's nearly time. And yes, you may speak to me!
10
11 Rost: You came to bless the naming?
12
13 Teersa: Have not six months gone by since we entrusted her to you?
14
15 Rost: But we are outcasts.
16
17 Teersa: You by choice, and she, well... I'm a High Matriarch, Rost. I bless whom I choose.
18
19 Rost: Then... you honor us.
```

**Figure 10**

*View of the Replace window in Notepad++ with a regular expression search command*

Note: the regular expression `\[.+\](^.+: )\(.+\)` searches for phrases starting and ending with square brackets, phrases that begin at a new line and end with a colon and a blank space, and phrases starting and ending with standart brackets. All results were replaced with empty/zero space.



**Figure 11**

The excerpt from Figure 9 after running the Replace function seen in Figure 10

```
1  
2  
3 What's that now? Don't like the cold? Can't stay in today. We have a ritual to perform, y  
4  
5  
6  
7 High Matriarch Teersa? What is she doing here? Does she mean to forbid the ritual?  
8  
9 No-no-no, off your knees! It's nearly time. And yes, you may speak to me!  
10  
11 You came to bless the naming?  
12  
13 Have not six months gone by since we entrusted her to you?  
14  
15 But we are outcasts.  
16  
17 You by choice, and she, well... I'm a High Matriarch, Rost. I bless whom I choose.  
18  
19 Then... you honor us.
```

After running several searches and replacements, clearing the text from any aforementioned non-diegetic notes, but also from the leftover ‘noise’ (such as empty spaces and lines), the text was checked manually. Afterwards, it was exported as a plain .txt file, ready to be used in the corpus.

The other way of obtaining video game texts is the extraction from the game files. The dialogue can be stored in many different formats, sometimes native to the particular development studio (eg. PAK, TLK) or stored as a standard document file (eg. JSON, CSV, XML). The text may be represented in a single file or in multiple, and the dialogue may be separate from other in-game texts or compiled in one file. Sometimes the file includes other info, like the ID of the text or the associated strings, and as such, there are several steps one has to make in order to create a file that is fit to be added to the corpus.

### 3.3.3 Extraction from the Game Files

For this research, the dialogue of eight games was extracted and processed manually specifically for this purpose, and the dialogue for another eight games was retrieved from Internet sources, as processed by other people.

For *Divinity: Original Sin* and *Divinity: Original Sin 2* (Henceforth *DOS* and *DOS2*), *GR2 Converter* was used to decode the localization files, by default encoded in a PAK format, which were found in D:\SteamLibrary\steamapps\common\Divinity Original Sin Enhanced Edition\Data\Localization and E:\SteamLibrary\steamapps\common\Divinity Original Sin 2\DefEd\Data\Localization respectively. The exported XML files were then

loaded in Microsoft Excel. The resulting table included columns “content”, with the game text, and “contentuid”, with an id of the particular line or textual unit. The cells contained a combination of diegetic and non-diegetic text, with the latter needing to be deleted. In order to do so, the table was sorted by line length (with the use of an auxiliary column in which every line was attributed a number according to its length, via the function LEN). This enabled an easier search, as the shortest lines were usually names of characters or objects, and the longest were texts of in-game books, letters or non-diegetic descriptions. Moreover, an auxiliary table was created to filter the lines by different conditions. For example, lines could not start and simultaneously end with the character ‘\*’, as this marked the narrator’s dialogue and non-direct speech. Such lines were filtered out. Some conditions needed a thorough manual check up - for instance, some lines including square brackets (‘[]’) were non-diegetic texts, but some were not, so after filtering the lines to only see ones with ‘[‘ present, it was necessarily to go through them and delete the unsuitable lines manually. After using the combination of semi-automatic and manual filtering, all lines that were left were copied to Notepad++. A number of these lines contained non-diegetic notes, some of them referring to a gameplay action, some referring to a character speaking. These have been deleted the same way as the previous cases, with the Replace function and manual proofreading.

The steps for the games *Mass Effect 2* and *Mass Effect 3* were nearly identical, with the difference that the files (both called ‘BIOGame\_INT.tlk’), found in ‘E:\Mass Effect 2\BioGame\CookedPC and E:\Mass Effect 3\BIOGame\CookedPCConsole’, were converted from a TLK format to XML by the application *Mass Effect 2 TLK Tool v.1.0.4* (2015). The resulting XML file was organised very similarly to that of DOS and DOS2, with two extra columns: ‘TLKToolVersion’ and ‘position’. Because of that, the rest of the process was the same as for DOS and DOS2.

*The Witcher 3* stores all dialogue in a .w3string format. Found in ‘E:\SteamLibrary\steamapps\common\The Witcher 3\content’ under subfolders ‘content0’ to ‘content12’, they had to be decoded using the tool *w3strings Encoder* (2015) and *w3strings Encoder GUI* (2016) for easier manipulation. The decoded files were in CSV format, with the values in a row referring to ‘id’, ‘key(hex)’, ‘key(str)’ and finally, ‘text’. After merging the files together into one CSV, the following process was identical to the previous games.

For the game *Darkest Dungeon*, all subtitled dialogue was present in the dialogue.string\_table.xml file, found in local files at the path ‘E:\SteamLibrary\steamapps\common\DarkestDungeon\localization.’ The xml file included

an ‘id’ column with the language of the subtitle line, ‘entry’ column with the text and ‘id2’ column, with strings associating the line with an in-game trigger. After uploading the file to Microsoft Excel, the contents were filtered by the ‘id’ column to show only English lines, and then the column with the in-game text was copied, pasted into Notepad++ and saved as a text file.

**Figure 12**

*Screenshot of the XML file with Darkest Dungeon subtitles*

	A	B	C
1	id	entry	id2
2	english	I'm boiling!	str_prisoner_damage_cauldron_full
3	english	My flesh... it burns!	str_prisoner_damage_cauldron_full
4	english	Get me out of this infernal cookpot!	str_prisoner_damage_cauldron_full
5	english	I don't want to be eaten!	str_prisoner_damage_cauldron_full
6	english	Aiyeeee! The pain!	str_prisoner_damage_cauldron_full
7	english	Strike the pot! Strike the pot!	str_prisoner_damage_cauldron_full
8	english	I fight for love!	str_control_before_turn_siren
9	english	For the tides!	str_control_before_turn_siren
10	english	Do not threaten my Queen!	str_control_before_turn_siren
11	english	I will live forever here, with my Love!	str_control_before_turn_siren
12	english	These waters are my home now!	str_control_before_turn_siren
13	english	My eyes clear!	str_control_return_siren
14	english	I had the most tranquil dream.	str_control_return_siren
15	english	The tides...they took me!	str_control_return_siren
16	english	Wh-where am I?	str_control_return_siren
17	english	...I am sorry my friends!	str_control_return_siren
18	english	...can't breathe!	str_prisoner_damage_drowned_anchored
19	english	The fathoms call me down!	str_prisoner_damage_drowned_anchored
20	english	Release me! I cannot swim!	str_prisoner_damage_drowned_anchored
21	english	I reject your pact!	str_prisoner_damage_drowned_anchored
22	english	The tide washes over me!	str_prisoner_damage_drowned_anchored
23	english	Crushing weight...	str_prisoner_damage_drowned_anchored
24	english	Those of faith have no tolerance for those with my condition.	str_incompatible_party_abomination_religion
25	english	I will not serve with this... creature.	str_incompatible_party_member_abomination_religion

Similarly straightforward was the extraction of text from *Transistor*. All files in the folder E:\SteamLibrary\steamapps\common\Transistor\Content\Subtitles\en were compiled into one file in Notepad++. All the files were in CSV (comma separated value) format, with the first value before comma being the subtitle line’s ID and the second the actual text. All the ID values were deleted, once again by the usage of Replace function along with RegEx, which identified all first values including the comma, making it simple to replace them with zero spaces.

The text file from *Fallout New Vegas* was the simplest to obtain. After loading the path to the game (‘E:\SteamLibrary\steamapps\common\Fallout New Vegas enplczru\Data’) to the application *LazyVoiceFinder* (2017), it automatically showed all instances of dialogue that could be then copied and pasted elsewhere.

Figure 13

View of Fallout New Vegas dialogue lines in Lazy Voice Finder

Since many of the lines repeated, as they appeared in slightly different instances within the game, the text was first pasted to *Microsoft Excel*, where the ‘Remove Duplicates’ function was used. Afterwards, the standard refining of the file in *Notepad++* followed, as there were some non-diegetic notes in parentheses ‘()’ and curly brackets ‘{}’ in certain lines, which had to be deleted by the Replace tool.

Some text files obtained from the internet were suitable to be immediately uploaded into the corpus, with only the file name being changed. This was the case of dialogues from *Elden Ring*, *Dark Souls 1*, *Dark Souls 2*, *Dark Souls 3*, *The Elder Scrolls IV: Skyrim* and *Fallout 4*.

Other files had to be further changed. *Portal 2* text, taken from the site Gamefaqs (Oblivion, 2011), contained non-diegetic notes from the user for easier orientation, and

developer commentary. This meant that the process of cleaning up the file was identical to that of transcriptions, as discussed in the previous section.

In the case of *Cyberpunk 2077*, a lot of changes had to be made in the file, as it included markers of the particular files from which the specific parts were taken. Moreover, the data were taken from JSON<sup>2</sup> type files and after being converted into a text format and opened in Notepad++, a high number of extra graphic characters were present. Before refining the text by once again using the Replace function, another issue had to be solved: the default ANSI encoding<sup>3</sup> showed certain characters incorrectly (eg. ‘á’ was shown as ‘Ăˇ’), but the UTF Encodings showed some of the extra characters as hexadecimal codes<sup>4</sup> that could not be easily searched for and replaced. After experimentation with different encoding views and conversions, the approach chosen was to search for all accented letters in the UTF-8 encoding, changing the view to ANSI encoding to see by what combination of characters were the letters represented and noting down said equation (a single character in UTF-8 had always a single corresponding character sequence in ANSI). Afterwards, the process was similar to the one used when working with transcribed dialogue - the extra characters (the ‘noise’) were found and replaced by zero spaces, and the character sequences within the text itself were replaced by the corresponding letter.

---

<sup>2</sup> JSON stands for "JavaScript Object Notation" and is a "standard text-based data interchange format" that "formats data in a way that is both human- and machine-readable." (TechTerms, 2024)

<sup>3</sup> Character encoding is a process of assigning a numeric code to a graphic character. This is because while humans see text documents as lines of text, computers view them as binary data (series of ones and zeros). Every type of encoding has a different set of codes for the graphic characters, and the most popular encoders are ASCII and Unicode. (TechTerms, 2024)

<sup>4</sup> Hexadecimal is a "base-16 number system used to represent binary data." (TechTerms, 2024)

Figure 14

Cyberpunk 2077 in-game text with ANSI Encoding

```
..
..
subtitles\media\animated_billboards\ab_ad_caliente.json..
CR2WĀ.....]c,Ā.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....$a†.....
..... N>D=Ī.
.....ŽAhoy, sea dogs!...
.....ĐD žš.
Y.....ÓSet sail for flavor with CapitĀ'n Caliente's world-famous tacos. Now in Whale Size!...
.....ĐDpžžš.
*.....AOne eurodollar for an ocean of taste!.....
..
subtitles\media\animated_billboards\ab_ad_chromanticore.json..
CR2WĀ.....'.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....*āL.....
..... N>Ī.
%..... Feeling tired? Bored? Powerless?...
.....ĐD'C"š.
J.....ĀNot anymore. Experience all of ChroManticore's 16 new flavors and.....
.....ĐDŠ"š.
.....ŠMIX IT UP!.....
..
subtitles\media\animated_billboards\ab_ad_foreign_body.json..
CR2WĀ.....-...-...%@ť.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....-ī,.....
..... Nlm?Ī.
9.....'When a phantom itch leads to a fatal system error.....
.....ĐD/G=š.
3.....@"Foreign Body" -â€" Install at your own risk.....
.....ĐDžĐ=š.
1.....Available only on braindance. In stores now!.....
..
```

Figure 15

Cyberpunk 2077 in-game text with UTF-8 Encoding

```
..
..
subtitles\media\animated_billboards\ab_ad_caliente.json..
CR2WĀ.....]c,Ā.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....$a†.....
..... N>D=Ī.
.....ŽAhoy, sea dogs!...
.....ĐD žš.
Y.....ÓSet sail for flavor with CapitĀ'n Caliente's world-famous tacos. Now in Whale Size!...
.....ĐDpžžš.
*.....AOne eurodollar for an ocean of taste!.....
..
subtitles\media\animated_billboards\ab_ad_chromanticore.json..
CR2WĀ.....'.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....*āL.....
..... N>Ī.
%..... Feeling tired? Bored? Powerless?...
.....ĐD'C"š.
J.....ĀNot anymore. Experience all of ChroManticore's 16 new flavors and.....
.....ĐDŠ"š.
.....ŠMIX IT UP!.....
..
subtitles\media\animated_billboards\ab_ad_foreign_body.json..
CR2WĀ.....-...-...%@ť.....Ó.....ČŘóYs.....'B.....ă.....UK»ěó.....-ī,.....
..... Nlm?Ī.
9.....'When a phantom itch leads to a fatal system error.....
.....ĐD/G=š.
3.....@"Foreign Body" -â€" Install at your own risk.....
.....ĐDžĐ=š.
1.....Available only on braindance. In stores now!.....
..
```



## 4 Analysis

### 4.1 Frequency

**Table 2**

*Raw frequency of selected lemmas in SFG corpus*

Woman	1,458	She	12,298
Man	5,927	He	23,091

**Table 3**

*Raw frequency of selected lemmas in OPUS subcorpus*

Woman	19,201	She	185,560
Man	76,040	He	315,976

The frequency analysis is the most straightforward, but still revealing. Looking at the raw frequencies, we see that in the SFG corpus, the lemma *man* is 4 times more frequent than *woman*, and similarly, in the OPUS subcorpus, it is 4.1 times more frequent. The reason that the difference between man and woman is so high is undoubtedly that man is often used as a gender neutral noun, as seen in following examples from SFG corpus:

- (1) “So dead men tell no tales, right?”
- (2) “Stay calm, man.”
- (3) “Any man neglecting to care for his own hut's a fool.”

This makes the pronoun comparison a more accurate way to judge the representation of the male gender. Here, the discrepancy in numbers when it comes to the lemmas *she* and *he* is not as striking, yet *he* is still used more often. In the SFG corpus, lemma *he* is 1.9 times more frequent than *she* and in the OPUS subcorpus, it is 1.7 times more frequent.

## 4.2 Collocates

In his 2008 paper, Michael Pearce compared the collocates of the lemmas *woman* and *man* in the *British National Corpus* (BNC) using Sketch Engine, and determined that there were significant asymmetries in the way men and women were presented in relation to power and dominance. The lemma *man* was more strongly associated with physical strength and endurance (Pearce, pp. 7), physical size and potency (pp. 8), and the general exercise or ownership of power (pp. 8) than the lemma *woman*. *Man* also appeared more frequently in collocations with words describing criminal activities, both as the subject and the object (pp. 10). Women were, on the other hand, more likely to be victims of crimes of sexual nature than men (pp. 10). Verbs denoting a woman's role as a beneficiary of a positive action often implied "weakness, lack or shortcoming on the part of the beneficiary" (p. 10 - 11). *Woman* was also found as an object with verbs involving the exercise of power by others. (pp. 11).

Pearce's work lays down a solid framework for analysing gendered collocates, having presented a way to categorise the data and having identified several clear asymmetries. As such, his paper will serve as the basis for this section.

It should be noted that some of the collocates recognized by Word Sketch have been omitted from the analysis, as either 1) the source of the collocate was only a singular document, or 2) the word did not semantically fit in any of the predetermined categories and did not offer any insight regarding the subject matter of the research.

### 4.2.1 *Man and Woman as Object and Subject*

Pearce (2008) suggests several categories for the lemmas *man* and *woman* in the subjective position. Actions requiring physical strength and endurance; exercise/ownership of power; criminal and/or violent acts; intense and passionate verbal/vocal expression; and emotionally intemperate verbal/vocal expression. During this research, this division has been used, except the last two categories, as there were no suitable words belonging to them. The collocates found via Word Sketch have been distributed between the rest of the groups when applicable:

**Table 4***Lemmas man and woman as a Subjects in the SFG Corpus and the OPUS Subcorpus*

	SFG CORPUS		OPUS SUBCORPUS	
	MAN	WOMAN	MAN	WOMAN
Actions requiring physical strength and endurance	come, go, stand, walk	go	breach, break, build, flee, hold, jump, run, stand, wield	endure, run, shove
Exercise/ownership of power	own	run (transitive)	enslave, kidnap	run (transitive)
Criminal and/or violent acts	attack, fight, kill	attack, fight, kill	attack, betray, burn, fight, kill, murder, rob, shoot, steal, torture	fight, murder, stab
Emotionally intemperate verbal/vocal expression	-	-	-	cry

As seen in Table 4, the bigger scope of OPUS subcorpus influences significantly the number of collocates. *Man* in both SFG corpus and OPUS subcorpus is found more frequently co-occurring with actions requiring physical abilities than *woman*. A movement connected to weakness is found only with men in the OPUS subcorpus (*flee*). *Run* (denoting a movement) appears as a collocate with *woman* in the same corpus, but only refer to ‘running away’ in two out of six instances:

- (4) “I’ve had women ran away from me before but never that fast!”
- (5) “And some of these women were running from the Everlasting.”

Meanwhile, women in both cases pattern with transitive *run*, while men do not. There is a possible argument to be made - *run* can be seen as slightly less prestigious than *own*, so

this occurrence may be an implication that women are often seen as the workers or a running force behind a business, but not as the truly powerful owners. Or, conversely, men may be seen as powerful but only as an empty symbol of prestige, while women are the ones having the real influence. While there is not enough data to support either of the two interpretations, a study with a bigger data sample might find a deeper pattern.

It should be said that men in OPUS subcorpus do not collocate with the discussed *own*, but with significantly more aggressive *enslave* and *kidnap*, an active exercising of power, while limiting other's. The lemma *man* in OPUS subcorpus is also more frequently found with verbs of violence. The lemma *woman* co-occur with *cry* in OPUS, but not in SFG corpus.

For lemmas *man* and *woman* in objective positions, Pearce's research once again serves as a basis, as the collocates were divided into several of his suggested categories:

**Table 5***Lemmas man and woman as an Object in the SFG Corpus and OPUS Subcorpus*

	SFG CORPUS		OPUS SUBCORPUS	
	MAN	WOMAN	MAN	WOMAN
Undergoing actions of legal system	-	-	arrest, execute, punish	arrest, execute
Victims of violence	fight, hang, kill, murder, slaughter	attack, hit, kidnap, kill, murder	attack, bite, burn, decapitate, fight, flay, hang, kill, murder, sacrifice, scare, shoot, slaughter, strike, torture	abduct, assault, attack, burn, drag, harm, hit, hurt, kidnap, kill, murder, rape, shoot, stab, strangle, strike, violate
Ideological and physical coercion	break	-	compel	-
'Recipients' of sexual activity	-	-	-	impregnate, screw, seduce

A pattern similar to Pearce's (2008) research emerges when looking at men and women as victims of violence. While both genders appear to collocate with violence, only women get *hit* or *kidnapped*. Men do not occur with *hit* presumably because it is a relatively weak action. That is not to say that it is not an act of violence (and in this case of gendered violence), but men seem to be victims of more brutal, non-sexual violence (*slaughter*, *flay*, *decapitate*). Meanwhile *woman* seem to occur with violence targeting one's agency (*kidnapped* or *dragged* - one is moved from one place to another against their will, *violate* or *rape* - the utmost transgressions against one's own independent wishes). Important is to note that the violence sexual in nature that was found in collocation with women in Pearce's work appears here in the OPUS subcorpus, but does not appear in the case of the SFG corpus.

Surprisingly, only men appeared with verbs denoting coercion, which may simply be connected to their overall bigger presence and activity in their respectable media.

Women serve as recipients of sexual activity only in the OPUS Corpus. One reason for this is that while the selection of video games for the SFG Corpus contained mainly ones suitable for older audiences, games are still a medium where sexual activity (or talk thereof) does not appear as often as in television. But considering that a significant portion of the games chosen (at least 10 out of the 29) do contain such themes to some capacity, another explanation is that the chosen games are not as gendered in presentation of said themes.

#### **4.2.2 Modifiers of *Man* and *Woman***

Pearce (2008) divided modifiers of *man* and *woman* into several categories: physical size and potency; power, wealth, influence; deviancy; marital/reproductive status and sexual orientation; nationality, religion, ethnicity, class; personality traits and appearance. In this section, we will use a similar schema of division, with a few changes. “Personality traits” were split into “positive” and “negative” to have a more detailed overview. “Marital/reproductive status and sexual orientation” and “nationality, religion, ethnicity, class” were omitted as once again, not many modifiers belonged to these categories. The categories of “age” and “weakness, vulnerability, misfortune” were added. The reason for the latter is that in speculative fiction media, a lot of characters find themselves in unfavourable situations, either to be saved by the heroes or to save themselves, or for their situation to serve as an exposition, atmospheric backdrop or motivation for other characters. And since a high number of modifiers seemed suitable for this category, it has been added to facilitate investigation of this aspect of sci-fi and fantasy storytelling.

**Table 6***Modifiers of lemmas man and woman in the SFG corpus*

Note: modifiers appearing only with one lemma are in boldface

	MAN	WOMAN
Physical size and potency	<b>big</b> , little, <b>smallish</b>	little, <b>strong</b>
Appearance	<b>green, metal, smelly, tin</b>	<b>ashen-haired, beautiful, pretty</b>
Age	young, old	young, old
Power, wealth, influence	<b>wealthy, rich, important</b>	-
Deviancy	dangerous, <b>evil</b>	dangerous
Weakness, vulnerability, misfortune	<b>blind</b> , dead, <b>lesser</b> , poor	dead, poor, <b>sick</b>
Personality traits - positive	<b>brave, clever</b> , decent, <b>fine</b> , good, great, happy, honest, <b>honorable, patient, powerful, reasonable</b> , smart, wise	<b>cunning</b> , decent, good, great, great, happy, honest, <b>lovely, nice, resourceful</b> , smart, wise
Personality traits - negative	<b>bad, crazy, stupid</b>	<b>angry, crazy, mad, stubborn</b>

**Table 7**

*Modifiers of lemmas man and woman in the OPUS subcorpus*

Note: modifiers appearing only with one lemma are in boldface

	MAN	WOMAN
Physical size and potency	<b>able-bodied, tall, strong, weak</b>	-
Appearance	<b>bearded, handsome, hooded</b>	<b>attractive, beautiful, blonde, pretty, sexy</b>
Age	grown, <b>middle-aged</b> , old, young	grown, old, young
Power, wealth, influence	<b>rich</b> , powerful	powerful, <b>wealthy</b>
Deviancy	<b>dangerous, evil</b>	<b>deadly, hateful, violent</b>
Weakness, vulnerability, misfortune	<b>blind</b> , dead, drunk, <b>homeless</b> , lonely, mortal, poor, <b>sorry</b> , <b>unarmed, wounded</b>	<b>crying</b> , dead, <b>helpless</b> , lonely, <b>missing</b> , poor
Personality traits - positive	brave, brilliant, <b>capable</b> , <b>charming, clever, cool, fine, funny, godly, good</b> , honest, <b>honorable</b> , intelligent, <b>loyal, merry, okay, sane</b> , smart, wise	<b>amazing</b> , brave, brilliant, <b>extraordinary, gifted, honest</b> , intelligent, <b>kind, lovely, lucky</b> , sacred, smart, strong, wise
Personality traits - negative	<b>crazy, desperate</b> , mad, <b>proud, stubborn</b>	<b>dreadful</b> , mad, <b>repellant</b>

Women in both videogames and television and cinema tend to have their appearance positively evaluated (*beautiful, pretty, attractive, sexy, pretty*), while men tend to be described neutrally (*little, smallish, big, metal, green, tin, hooded, tall, bearded*). Only the lemma *man* patterned with a negative description of appearance (*smelly*). When evaluating physical potency, men are described as *big, tall* and *strong*, as well as *little, smallish, weak*. Women have less collocates referring to bodily prowess - in the OPUS subcorpus, there are



no modifiers in this category, while in the SFG corpus, there are the adjectives *little* and *strong*. It should be noted that *strong* does not have to refer to physical strength, but may refer to mental strength or a combination of the two. In fact, when co-occurring with *women*, concordance lines that imply the strength to be physical in character are rare:

- (6) “My mother was a strong woman...she's the one who instructed me on my sword fighting techniques.”

More common are the more ambiguous meanings of strength:

- (7) “She'll handle it. She's a strong woman.”  
(8) “You can't stand the sight of a strong Nord woman?”

On the other hand, when *strong* modifies *man*, it seems to refer almost exclusively to physical strength:

- (9) “Killed some of our strongest men.”  
(10) “But a big strong man like you must be hungry, no?”  
(11) “Even the strongest man is a weakling compared with a wolf.”

In the SFG corpus, only men are in co-occurrence with words denoting power and influence (*wealthy*, *rich*, *important*). The OPUS subcorpus is more balanced - both men and women occur with *powerful* and both occur with either *rich* or *wealthy*. Interestingly, when it comes to monetary possessions, the context is often a promise of wealth, both for men and women, rather than a description of a present situation:

- (12) “You can return to the Free Cities and live as a wealthy woman for all your days.”  
(13) “I can make you a wealthy woman.”  
(14) “I'm going to make you a rich man.”  
(15) “He plans on becoming the richest man on Earth [...]”

Collocates referring to deviancy are relatively symmetrical - both *man* and *woman* appear together with *dangerous* (with logDice score of 6.8 for women and men in the SFG corpus, and 5.9 for men in the OPUS subcorpus), but only men (both in SFG corpus and

OPUS subcorpus) are described with the adjective *evil*. Women in OPUS subcorpus are also uniquely modified by *violent* and *deadly*. This result is very different to Pearce's (2008) findings in the BNC, where only *man* had a clear patterning with modifiers denoting deviancy. This can be explained by the fact that our research focuses on sci-fi and fantasy genres, which often include action, fighting scenes and an overall higher level of violence than one encounters in daily life. It is safe to assume that this extends to both genders.

In both corpora, *man* co-occur with more unique modifiers denoting weakness or vulnerability. For both genders, there is a weakness connected to economic situation (*poor*; *homeless* - uniquely with men in OPUS subcorpus), death (*dead*) and sickness (*sick*; *blind* - uniquely with men in both corpora), loneliness (*lonely*). In the OPUS subcorpus, men create patterns with violence-related (*unarmed*, *wounded*) and vice-related (*drunk*) vulnerability. Women in the same corpus then appear with modifiers signifying mental vulnerability (*crying*) and victimhood (*helpless*, *missing*).

Positive traits denoting mental capabilities (*intelligent*, *smart*, *brilliant*, *wise*) appear with both men and women, but *cunning*, a word with both negative and positive connotations, co-occur only with women in the SFG corpus.

Men and women co-occur with modifiers evaluating sanity negatively (*mad*, *crazy*), but only *man* in OPUS subcorpus patterns with *sane*. This avoids the stereotype of women being perceived as mentally unstable.

### 4.3 Keywords

Heritage (2008) uses keyword analysis to investigate the gender in *The Witcher* game series. He divided the list of keywords into seven categories: male names, female names, male social actors, female social actors, gender neutral words, non-gendered words and pronouns. In this context, male and female social actors are words that are innately gendered, eg. *queen*. Gender neutral words are those that can be associated with all genders, while non-gendered words do not imply any gender (pp. 153). Heritage's way of differentiating between the two is to test whether the word could be modified by modifiers *male* or *female*. If so, the word was categorised as gender neutral, if not, it was deemed non-gendered (pp. 153). In Heritage's research, some of the words were categorised on the basis of familiarity of the material. The same applies in this paper: *witcher* is traditionally a male profession, and so is categorised as an male social actor, and *asari* is an all-female alien race, thus being categorised as a female social actor.

Using the same approach as Heritage, the keyword list generated in Sketch Engine was divided into the established categories, with an eighth category, listing names that are used by either gender within the material (see Appendix: Table 10 and 11). Afterwards, a simplified tables with male and female names, and male and female social actors was created:

**Table 8**

*Gendered Keywords in the SFG corpus*

Female Name	Freq.	Male Name	Freq.	Female Social Actor	Freq.	Male Social Actor	Freq.
Aloy	446	Alexandar	361	asari	500	king	1894
Ciri	532	Arhu	311	goddess	539	witcher	1501
Dallis	430	Braccus	473	mum	780		
Edi	318	Caesar	589	witch	576		
Lohse	336	Dandelion	314				
Tali	410	Geralt	966				
		Kemm	309				
		Lucian	868				
		Ulfric	394				
	2472		4585		2 395		3 395

**Table 9***Gendered Keywords in the OPUS subcorpus*

Female Name	Freq.	Male Name	Freq.	Female Social Actor	Freq.	Male Social Actor	Freq.
Abby	2 179	Alaric	1 059	queen	9 397	brother	22 040
Alice	3 943	Alec	2 439	princess	3 685	duke	2 843
Alison	1 964	Arthur	3 135	witch	9 425	king	17 682
Amelia	1 394	Artie	2 622			lord	10 931
Audrey	2 605	Castiel	1 290			ser	1 454
Barbie	1 577	Crowley	2 147			sire	1 824
Belle	1 461	Damon	7 230			wizard	1 711
Beth	1 964	Dean	7 659				
Bo	3 379	Dyson	1 305				
Bonnie	4 730	Elijah	2 721				
Caroline	3 279	Enzo	1 121				
Cass	3 108	Ethan	2 459				
Cassie	2 524	Finn	1 633				
Chloe	2 416	Francis	1 918				
Clarke	1 709	Henry	8 471				
Clary	2 255	Jace	1 580				
Cosima	1 018	Jeremy	3 112				
Elena	8 729	Josh	2 756				
Emma	3 891	Killian	1 016				
Freya	1 480	Klaus	4 575				
Hayley	1 175	Lucifer	4 466				
Helena	1 487	Marcel	1 843				
Julia	2 325	Matt	3 853				
Juliette	1 239	Merlin	1 723				
Katherine	2 296	Nathan	3 086				
Kenzi	1 163	Nick	7 828				
Kiera	999	Pete	3 414				
Lucy	3 117	Peter	6 412				
Magnus	1 546	Silas	1 518				
Mary	5 948	Stefan	7 422				

Myka	1 271	Thorin	1 188				
Olivia	2 860	Tyler	3 200				
Rebekah	1 524	Walter	4 803				
Regina	2 401						
Sarah	7 246						
	92 202		111 004		22 507		58 485

The trend in the SFG corpus exhibits similarities to Heritage's (2021) findings on keywords in the *Witcher* series, where he found male names to be 2.3 times more frequent than female names and references to male social actors 3.1 more frequent than to female social actors (pp. 166). The male names in SFG corpus are also more common than female ones, but the difference is slightly smaller, with the raw frequencies of male names being 1.9 times higher than those of female names. A more significant discrepancy shows with the social actors - in SFG corpus, there were four keywords referring to female social actors and only two referring to male social actors. The raw frequency of male actors was still higher, but only 1.4 times then the frequency of the female actors.

The keywords themselves are worth looking at: the male social actor *king* is a word denoting power and authority, while *witcher* denotes dangerous, violent and socially frowned upon job, as *witcher* is a hunter of monsters. *Witch* is likewise a word implying a degree of danger and distrust, making it a reasonable equivalent of *witcher* within the semantic context. Meanwhile, the word *queen* is missing to serve as an equivalent of *king*, but the word *goddess*, which denotes the utmost authority and power, is present. The other female social actors are *mum* and *asari*. As stated earlier, *asari* is an all-female race. A point of interest is a comparison with other sci-fi and fantasy races (turian, salarian, krogan, orc, elf, dwarf) which are not gender-specific. Albeit manually comparing concordances and using the Sketch Engine Word Sketch module to compare collocates did not reveal any strong trends, it did reveal a detail worth mentioning. Only *asari* appeared as a modifier of the gendered expletives *slut* and *bitch*, while no other fantastical race did. While the collocation was not frequent enough for it to illustrate a general problem, this finding would appear to be consistent with the idea that there is underlying tendency to associate women with sexual themes.

The presumed reason for the female social actor *mum* to be included in the keywords is the form, typical for British English. In the reference OpenSubtitles 2018 corpus, the word

*mom* appears 509,242 times, while *mum* only 74,958. Meanwhile, in the SFG corpus, *mom* has a frequency of 255, while the word *mum* 780. Yet, there is a value in looking at this entry in more detail.

As stated, the raw frequency of the keyword *mum* in the SFG corpus is 780, while the frequency of the lemma *mother* is 1,507. This makes the abbreviation unexpectedly common, in comparison to *dad*, which is present only 320 times, and *father*, present 1 635 times, only slightly more often than *mother*. Concordance analysis reveals that most of the usage of the word comes from children. See the following examples:

- (16) “Mum says I'm not to play with the other children, because they're "being raised on a diet of dog-eat-dog.”
- (17) “Maybe mum and dad are stuck out there...waiting for me to come and find them.”
- (18) “I'm going out in the world, and I'm going to make my mum proud!”

This may imply that women in the games within the SFG corpus are more likely to be portrayed as beloved mothers than men as beloved fathers. One may insinuate that this ties into the notion that women are better caretakers than men.

Unlike in the SFG corpus, there were more female than male names within the selected keywords, albeit the difference was only that of two names. Moreover, frequency wise, male names were 1.2 times more frequent than female names. There were significantly less female social actors than male social actors, with only three keywords referring to female actors and seven referring to male actors. The keywords denoting male social actors were also 2.6 times more frequent.

The keywords denoting male social actors in OPUS subcorpus are mainly signifiers of authority: *king* (in this case semantically paired with *queen*, albeit *king* is 1.9 times more frequent), *duke*, *lord*, *ser*, *sire*. *Wizard* appears to be a relatively neutral term, but collocate analysis reveals the presence of collocation with modifiers like *devious*, *wretched*, *evil* and *dark*. However, not one of the negative modifiers of *wizard* reaches near the 9.8 logDice score of *wicked witch*. While the two nouns can be considered equivalent to each other, *witch* seems to have stronger negative connotations.

The female social actor *princess* may either imply a degree of authority or on the contrary, refer to the romantic image of a gentle princess, possibly akin to a damsel in distress. Another role *princess* may have is that of a child. The collocational analysis points

towards the latter two options being more common, as the term *princess* is mainly used to signify disrespect, describe the woman or girl as either weak, or refer to a child. The most frequent modifier of *princess* is *little* with the frequency of 54 (albeit with a score of only 0.8):

- (19) “I hope our little princess hasn't caused you too much trouble.”
- (20) “Run back home, little princess.”
- (21) “Well, I see a spoiled little princess who ran away from her troubles.”

Less frequent occurrence, but still significant is the modifier *warrior* (with the frequency of 19 and score of 7.2). While this may seem to show that *princess* may reference a woman in power, concordance analysis reveals this is not always the case, as the term is sometimes used either in a mocking way (example 22), or in connection to sexual or inappropriate acts (examples 23 and 24).

- (22) What's your plan, oh warrior princess?
- (23) I just want to know if I'm bedding a warrior princess.
- (24) I am a warrior princess. And I will not be giving an 11-year-old my bra.

Based on the raw frequencies, both SFG corpus and OPUS subcorpus show the male names and social actors to be more utilised. The discrepancy between the gendered social roles is especially visible with the OPUS subcorpus, and along with the number of male social actors signifying power, this may lead us to believe that the movies and TV shows in the OPUS subcorpus are more likely to have word building based upon patriarchal society.

## 5 Conclusion

In this work, we were investigating the differences between the language surrounding male and female gender across video games, TV shows and movies.

A part of the work was the methodology of creating a new corpus from scratch. This decision was made to add to the still fairly vacant space in sociolinguistics research - language and video games. While the creation process happened with limited resources, leading to certain restrictions, such as having to use fan transcriptions, possibly sacrificing parts of the data in the process, the resulting corpus was sufficient for further analysis and is suitable for use by other researchers.

The assumption that men will be talked about more frequently was proven true: in both corpora, lemmas *man* and *he* were more frequent than *woman* and *she*, with the ratio being very similar in both corpora. Similarly, keywords referring to male social roles or to individual male characters (via name) were more frequent, albeit in the case of OPUS subcorpus, the difference in frequency between male and female names was not significant.

While there were patterns of stereotyping, they were not as significant as expected. The most significant tendencies were shown via keywords denoting social roles: in the OPUS subcorpus, most of the male roles implied a position of power. Some collocates could also be interpreted as submitting to a gender stereotype, such as *man* having more collocates relating to physical activity, only *woman* collocating with *cry* or men being victims of more brutal bodily violence, while women being victims of more dehumanising nature.

Another assumption was that the language referring to women will be sexualized. Women's appearance was evaluated positively in contrast to men, creating a clear difference, but the terms used were largely respectful. Aside from the expletive collocations of *asari*, the SFG corpus did not contain any sexualizing language, while the OPUS subcorpus contained collocation of women with sexual violence and sexual acts. An assumption can be made that games, which have been widely criticised for sexualization of female characters for years, have actually shifted towards more respectful representation. We can speculate that one possible reason may be the want of developers to cater to a wider demographic. And perhaps the cinema business has no such motivations, and as such, there has not been enough effort to deconstruct the ways female characters are presented. But in order to draw conclusions relating to the whole industry, more research would be needed.

A study using a bigger video game corpus could be able to find more patterns and linguistic behaviour. As this paper showed, there already is a number of video game corpora



from different researchers - in a possible future research, the corpora could be merged (after processing the data to avoid duplicates), giving the opportunity to work with bigger numbers. This paper also used approaches from several researchers in order to conduct the analysis upon an already established framework, which made working with the data more straightforward and offered the option to compare the results with previous research. However, it should be recognized that such an approach can become limiting, as the data have to fit certain assumptions, which may deprave the analysis of certain perspectives. As such, works revisiting subjects matter of this paper might benefit from a more flexible approach.

In preparation of my bachelor's thesis, I have not used any generative AI tools.

## Bibliography

Oblivion. (2011, April 22). *Text Dump (PC)*. GameFAQs.

<https://gamefaqs.gamespot.com/pc/991073-portal-2/faqs/62236>

Alice Wiki. (n.d.). *Transcript:Alice: Madness Returns*. Retrieved April 8, 2024, from

[https://alice.fandom.com/wiki/Transcript:Alice:\\_Madness\\_Returns](https://alice.fandom.com/wiki/Transcript:Alice:_Madness_Returns)

Anderson, H., Daniels, M. (2016). *Film Dialogue*. [https://pudding.cool/2017/03/film-](https://pudding.cool/2017/03/film-dialogue/)

[dialogue/](https://pudding.cool/2017/03/film-dialogue/)

araveugnitsuga. (2022, March 13). *A Complete Dump of All of the Game's Text (Includes Dialogue and Item Descriptions)*. Reddit.

[https://www.reddit.com/r/Eldenring/comments/tdbody/a\\_complete\\_dump\\_of\\_all\\_of\\_the\\_games\\_text\\_includes/](https://www.reddit.com/r/Eldenring/comments/tdbody/a_complete_dump_of_all_of_the_games_text_includes/)

Aylett, H. (2015). *Where are the women directors? Report on gender quality for director in the European film industry (2006 - 2013)*. European Women's Audiovisual Network.

[https://www.ewawomen.com/wp-content/uploads/2018/09/Complete-report\\_compressed.pdf](https://www.ewawomen.com/wp-content/uploads/2018/09/Complete-report_compressed.pdf)

Baker, P. (2014). *Using corpora to analyse gender*. Bloomsbury.

Baker, P. (2023). *Using corpora in discourse analysis* (2nd ed.). Bloomsbury.

*Bioshock 2 Full Transcript*. (2018, July 2). Game Scripts Wiki Blog. <https://game-scripts-wiki.blogspot.com/2018/07/bioshock-2-full-transcript.html>

BowmoreLover. (2017). *Lazy Voice Finder*. Retrieved from

<https://www.nexusmods.com/skyrim/mods/82482?>

Busso L. & Vignozzi G. (2017). Gender Stereotypes in Film Language: A Corpus-Assisted Analysis. *Proceedings of the Fourth Italian Conference on Computational Linguistics*

*CLiC-it 2017*: 71 - 76. <https://doi.org/10.4000/books.aaccademia.2367>

- Butler, J. (1988). Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal*, 40(4), 519–531. <https://doi.org/10.2307/3207893>
- byronhulcher/dark-souls-markov/dialogue.txt. (2016, Apr 27). GitHub. <https://github.com/byronhulcher/dark-souls-markov/blob/master/dialogue.txt>
- Clement, J. (2023, November 6th). *Distribution of video gamers in the United States from 2006 to 2023, by gender*. Statista. <https://www.statista.com/statistics/232383/gender-split-of-us-computer-and-video-gamers/>
- Coates, J. (2013). *Women, Men and Language: A Sociolinguistic Account of Gender*
- Dark Souls 3 NPC dialogue. (2016, April 25). Pastebin. <https://pastebin.com/JTUv2CFM>
- DeWinter J., Kocurek C. A. (2017). ‘Aw fuck, I got a bitch on my team!’: women and the exclusionary cultures of the computer gamecomplex. *Gaming representation: race, gender, and sexuality in video games*, 57–73. Bloomington.
- Differences in Language* (3rd ed.). Routledge.
- Don Ho. (2024). *Notepad++ v8.6.9*. Retrieved from <https://notepad-plus-plus.org/downloads/>
- Donovan, T. (2010). *Replay: The History of Video Games*. Yellow Ant.
- Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender* (2nd ed.). Cambridge University Press.
- Gocławski, J. (2015). *Mass Effect 2 TLK Tool v.1.0.4*. Retrieved from <https://github.com/jgoclawski/me2-tlk-tool>
- God of War (2018) Full Transcript*. (2019, January 15). Game Scripts Wiki Blog. <https://game-scripts-wiki.blogspot.com/2019/01/god-of-war-2018-full-transcript.html>

- Heritage, F. (2021). *Language, Gender and Videogames: Using Corpora to Analyse the Representation of Gender in Fantasy Videogames*. Palgrave.
- Horizon Zero Dawn Full Transcript*. (2018, October 9). Game Scripts Wiki Blog.  
<https://game-scripts-wiki.blogspot.com/2018/10/horizon-zero-dawn-full-transcript.html>
- Horizon Zero Dawn II Forbidden West Full Transcript*. (2022, February 20). Game Scripts Wiki Blog. <https://game-scripts-wiki.blogspot.com/2022/02/horizon-ii-forbidden-west-transcript.html>
- IMDb. (n.d.). *Genre*. Retrieved June 4, 2024, from <https://www.imdb.com/feature/genre/>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*. 1: 7-36.
- Kondrat, X. (2015). Gender and video games: How is female gender generally represented in various genres of video games? *Journal of Comparative Research in Anthropology and Sociology*, 6(1), 171 - 193.  
<http://compaso.eu/wp-content/uploads/2015/08/Compaso2015-61-Kondrat.pdf>
- Kumari, A., Joshi, H. (2015). “Gender Stereotyped Portrayal of Women in the Media: Perception and Impact on Adolescent”. *IOSR Journal Of Humanities And Social Science*. 20(4), 44 - 52. <https://doi.org/10.9790/0837-20424452>
- Lauzen, M. (2024). *It's a Man's (Celluloid) World: Portrayals of Female Characters in the Top Grossing U.S. Films*. <https://womenintvfilm.sdsu.edu/wp-content/uploads/2024/02/2023-Its-a-Mans-Celluloid-World-Report.pdf>
- Lison P., Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Metal Gear Rising: Revengeance Full Transcript*. (2020, June 12). Game Scripts Wiki Blog.  
<https://game-scripts-wiki.blogspot.com/2020/06/metal-gear-rising-revengeance-full.html>

Microsoft Corporation. (2024). *Microsoft Excel*. Retrieved from

<https://office.microsoft.com/excel>

*Middle-earth: Shadow of Mordor Full Transcript*. (2019, January 22). Game Scripts Wiki

Blog. <https://game-scripts-wiki.blogspot.com/2019/01/middle-earth-shadow-of-mordor-full.html>

*Middle-earth: Shadow of War Full Transcript*. (2019, January 28). Game Scripts Wiki Blog.

<https://game-scripts-wiki.blogspot.com/2019/01/middle-earth-shadow-of-war-full.html>

Mortensen, T.E., Sihvonen, T. (2020). Negative Emotions Set in Motion: The Continued

Relevance of #GamerGate. In T. Holt, A. Bossler (Eds.), *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, 1353–1374. Palgrave Macmillan.

[https://doi.org/10.1007/978-3-319-78440-3\\_75](https://doi.org/10.1007/978-3-319-78440-3_75)

Nana\_Neobard. (2022, December 11). *Subtitles Resource*. Reddit.

[https://www.reddit.com/r/FF06B5/comments/ziee6v/subtitles\\_resource/](https://www.reddit.com/r/FF06B5/comments/ziee6v/subtitles_resource/)

Pak Extractor v1.11.0. Retrieved from [https://docs.larian.game/Pak\\_Extractor\\_Guide](https://docs.larian.game/Pak_Extractor_Guide)

Pearce, M. (2008). Investigating the collocational behaviour of MAN and WOMAN in the

BNC using Sketch Engine. *Corpora*. 3(1): 1–29.

<https://doi.org/10.3366/E174950320800004X>

pMarK. (2016). *W3strings Encoder GUI 1.010*. Retrieved from

<https://www.nexusmods.com/witcher3/mods/1203>

Rennick, S., Clinton, M., Ioannidou, E., Oh, L., Clooney, C., T. E., Healy, E., Roberts, S. G.

(2023). Gender bias in video game dialogue. *Royal Society Open Science*. 10(5): 221095.

<http://doi.org/10.1098/rsos.221095>

- Resident Evil 2 (2019) Full Transcript*. (2019, July 12). Game Scripts Wiki Blog.  
<https://game-scripts-wiki.blogspot.com/2019/07/resident-evil-2-2019-full-transcript.html>
- rmemr. (2015). *W3strings encoder 0.4.1*. Retrieved from  
<https://www.nexusmods.com/witcher3/mods/1055/>
- Rychlý, P. (2008). A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*. 2: 6-9.
- Sketch Engine. (n.d.). *Frequency*. Retrieved June 4, 2024, from  
[https://www.sketchengine.eu/my\\_keywords/frequency/](https://www.sketchengine.eu/my_keywords/frequency/)
- Sketch Engine. (n.d.). *Relative frequency, frequency per million*. Retrieved June 4, 2024, from [https://www.sketchengine.eu/my\\_keywords/freqmill/](https://www.sketchengine.eu/my_keywords/freqmill/)
- Sketch Engine. (n.d.). *Simple maths*. Retrieved June 4, 2024, from  
<https://www.sketchengine.eu/documentation/simple-maths/>
- Summers, P. [Shotgunnova]. (2014, April 4). *BioShock Infinite – Game Script*. GameFAQs.  
<https://gamefaqs.gamespot.com/ps3/605051-bioshock-infinite/faqs/69191>
- TechTerms. (n.d.). Encoding Definition. Retrieved June 10, 2024 from  
<https://techterms.com/definition/encoding>
- TechTerms. (n.d.). Hexadecimal Definition. Retrieved June 10, 2024 from  
<https://techterms.com/definition/hexadecimal>
- TechTerms. (n.d.). JSON Definition. Retrieved June 10, 2024 from  
<https://techterms.com/definition/json>
- TechTerms. (n.d.). Regular Expression Definition. Retrieved June 10, 2024 from  
[https://techterms.com/definition/regular\\_expression](https://techterms.com/definition/regular_expression)

*The Last of Us Part II Full Transcript*. (2020, October 7). Game Scripts Wiki Blog.  
<https://game-scripts-wiki.blogspot.com/2019/01/middle-earth-shadow-of-war-full.html>

TV Tropes. (n.d.). *Death Stranding*. Retrieved June 4, 2024, from  
<https://tvtropes.org/pmwiki/pmwiki.php/VideoGame/DeathStranding>

TV Tropes. (n.d.). *Video Game Genres*. Retrieved June 4, 2024, from  
<https://tvtropes.org/pmwiki/pmwiki.php/Main/VideoGameGenres>

Vocazone\_2. (2020, August 22). *Bloodborne Script PDF*. Reddit.  
[https://www.reddit.com/r/bloodborne/comments/ieauoo/bloodborne\\_script\\_pdf/](https://www.reddit.com/r/bloodborne/comments/ieauoo/bloodborne_script_pdf/)

When They Cry Wiki. (n.d.). *Fun Facts*. Retrieved June 4, 2024, from  
[https://wiki.whentheycry.org/wiki/When\\_They\\_Cry\\_Wiki:Fun\\_Facts#endnote\\_ArcWords](https://wiki.whentheycry.org/wiki/When_They_Cry_Wiki:Fun_Facts#endnote_ArcWords)

Winter L. & Masters S. (2023). “Better Dead than a Damsel”: Gender Representation and Player Churn. *CHI PLAY Companion '23: Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 102–107.  
<https://doi.org/10.1145/3573382.3616083>

zilav. (2015, 30 December). *Full dialogues and voice files list*. Nexus Mods.  
<https://www.nexusmods.com/fallout4/mods/7273>

## **Résumé**

Tato bakalářská práce má za cíl prozkoumat vyjadřování genderu v anglickém jazyce ve sci-fi a fantasy počítačových hrách a filmech/televizních seriálech za pomoci korpusové analýzy. Konkrétně byl výběr zúžen na hry a kinematografická díla vydaná v letech 2010 až 2024. Jako korpus reprezentující kinematografii byl vybrán *OpenSubtitles 2018* korpus, ze



kterého byl vytvořen subkorpus na základě žánru a roku publikace jednotlivých filmů a seriálů. Pro hry byl vytvořen nový *Speculative Fiction Games 2010 - 2024* korpus. Zkoumané jevy jsou frekvence, kolokáty, klíčová slova a konkordance.

V teoretické kapitole je nastíněná problematika genderu a biologického pohlaví, dále jsou popsány rozdílné přístupy k genderu v sociolinguistice. Další část poukazuje na možné tendence k diskriminaci žen v herním a filmovém průmyslu, dále jsou vysvětleny herní a filmové žánry a jak ovlivnily výběr korpusů. Poslední část je věnována dialogům ve filmech a hrách, konkrétně jsou popsány dialogy ve hrách, aby měl čtenář představu, s jakým typem textu se bude pracovat. Teoretická kapitola je pak zakončena souhrnem dosavadních prací na téma herních a filmových dialogů.

Metodologická kapitola se kromě stanovení cílů práce a vysvětlení základních analytických postupů soustředí na tvorbu herního korpusu. Jsou zde detailně popsány způsoby zisku dat a jejich zpracování.

V analytické kapitole se postupně rozebírají vybrané jevy. Je zde krátce zmíněná frekvence výskytu slov *man*, *woman*, *she* a *he*. Zde se ukazuje, že muži jsou v obou médiích zmiňováni častěji. Poté se přechází na analýzu kolokací - ty jsou rozděleny podle role lemm *man* a *woman*, které mohou být podmětem, předmětem a nebo modifikovány dalším větným členem. Výsledky v této části se občas vymykají předchozím výzkumům na dané téma, ale stejně zde nacházíme náznaky stereotypů. V analýze klíčových slov je diskutována frekvence mužských a ženských jmen, a více detailně také slova popisující ženského nebo mužského sociálního aktéra. Tato část odhaluje tendence vyobrazovat muže v pozicích moci, obzvláště ve filmovém korpusu.

Práce je zakončena zhodnocením hypotéz. Je dokázáno, že slova referující na mužský gender se vyskytují výrazně častěji v obou korpusech. Nejvíce sporná se ukázala otázka kolokátů - některé spadají do genderových stereotypů, ale nejedná se o tak výrazné rozdělení, jaké bylo předpokládáno. Nakonec je reflektováno na sexuálně orientovaný jazyk - ač se v analýze neobjevuje často, bylo vyzorovno, že je zásadně mířen na ženy, a to zvláště ve filmovém korpusu.

## Appendix

**Table 10**

*Table of Keywords from the SFG corpus*

Female Names	Aloy, Ciri, Dallis, Edi, Lohse, Tali
--------------	--------------------------------------

Male Names	Alexandar, Arhu, Braccus, Caesar, Dandelion, Geralt, Kemm, Lucian, Ulfric
Gender-Neutral Names and Surnames	Shepard, V
Female Social Actors	asari, goddess, mum, witch
Male Social Actors	king, witcher
Gender-neutral words	admiral, ai, beast, collector, commander, creature, demon, divine, dragon, elf, enemy, fiend, ghoul, godwoken, guard, human, hunter, immaculate, jarl, knight, krogan, lizard, mage, magister, magisters, minuteman, monster, mutant, nord, orc, paladin, quarian, raider, ranger, robot, rogue, salarian, seeker, soldier, sourcerer, spirit, stranger, synth, thief, troll, Turian, undead, voidwoken, warrior, wolf
Non-gendered words	access, against, alliance, among, an, ancient, Arasaka, area, armor, armour, around, Arx, as, attack, await, battle, beyond, biotic, blade, blood, brotherhood, c'mon, cap, caravan, cave, cerberus, certain, chem, choom, citadel, city, claim, clan, coin, combat, commonwealth, contract, corpse, council, craft, cure, curse, Cyseal, dam, damage, damn, dark, datum, dead, deal, death, deathfog, defeat, deserve, destroy, detect, divinity, doubt, dunno, earn, eh, elder, em, empire, end, enough, escape, eternal, expect, facility, far, fate, fear, few, fight, find, flame, flee, fleet, flesh, focus, folk, fool, force, forge, fort, free, fuckin, galaxy, genophage, geth, glad, glory, gold, guess, guild, ha, hah, heavy, heh, help, hide, hm, hope, hunt, im, imperial, increase, indeed, institute, isle, join, keep, knowledge, lead, learn, least, legion, lemme, location, lookin, luck, machine, magic, manage, master, may, mhm, might, mind, mission, more, must, NCR, near, need, normandy, nothin, novigrad, offer, once, order, outta, o', path, perhaps, place, potion, power, powerful, quite, rather, realm,

	reaper, remain, require, research, rest, return, reward, risk, rivellon, rot, sacrifice, safe, save, seek, seem, serve, shadow, shall, shame, shield, ship, sight, skellige, skill, skyrim, somethin, soon, soul, source, spare, spell, steel, stone, strength, strip, strong, supply, survive, sword, tale, target, tech, technology, temple, terminal, than, their, though, threat, tower, trade, trouble, truly, trust, ugh, uhh, upgrade, upon, use, useful, vault, void, war, waste, weapon, while, wish, within, worth, ya
Pronouns	our, someone, us, we, whatever

**Table 11***Table of Keywords from the OPUS subcorpus*

Female Names	Alice; Alison; Amelia; Audrey; Barbie; Belle; Beth; Bo; Bonnie; Caroline; Cass; Cassie; Chloe; Clary; Cosima; Elena; Emma; Freya; Hayley; Helena; Julia; Juliette; Katherine; Kenzi; Kiera; Lucy; Magnus; Mary; Monroe; Myka; Olivia; Rebekah; Regina; Sarah
Male Names	Alaric; Alec; Arthur; Artie; Castiel; Clarke; Crowley; Damon; Dean; Dyson; Elijah; Enzo; Ethan; Finn; Francis; Henry; Jace; Jeremy; Josh; Killian; Klaus; Lannister; Lucifer; Marcel; Matt; Merlin; Nathan; Nick; Pete; Peter; Sam; Silas; Simon; Stefan; Thorin; Tyler; Walter
Gender-Neutral Names and Surnames	Dunham; Lannister; Monroe; Sam; Stark; Winchester
Female Social Actors	princess; queen; witch
Male Social Actors	brother; duke; king; lord; ser; sire; wizard
Gender-neutral words	alien; angel; astronaut; beast; creature; demon; devil; dragon; elf; enemy; fae; fairy; figure; ghost; grace; grimm; grunt; guard; heavily; human; hunter; lead; monster; mortal; mystic; orc; raven; rex; scientist; sheriff; supernatural; vampire; werewolf; Wesen; wolf
Non-gendered words	aah; access; actually; alive; ancient; anyone; artifact; ash; attack; back; battle; beep; beeping; being; believe; betray; beyond; blade; blood; body; brain; breathe; build; burn; cage; camelot; castle; cell; century; choice; choose; chuckle; code; compel; connect; connection; control; council; create; cure; curse; dagger; danger; dangerous; dark; darkness; dead; death; destiny; destroy; device; die; dome; door; earth; energy; evil; exactly; exhale; exist; experiment; fail; falls; family; far; fate; fear; fight; find; forever;

	<p>future; gasp; gate; gonna; groan; growl; happen; haven; heal; heh; hell; help; hide; hmm; host; humanity; hunt; hurt; hybrid; infect; inside; kill; kingdom; lab; level; lock; magic; map; mars; maze; memory; mission; narrator; need; north; okay; original; pain; part; path; plan; planet; portal; possible; power; powerful; previously; promise; prophecy; protect; protocol; realm; research; risk; sacrifice; safe; save; science; scoff; scream; secret; shadow; shh; ship; sigh; snow; somewhere; soul; source; space; spell; stone; stop; storm; storybrooke; strong; survive; sword; system; technology; threat; throne; town; track; trap; trust; ugh; uh; um; universe; until; upon; virus; vision; wake; wall; warehouse; weapon; weird; whoa; whoosh; winterfell; wood; world</p>
Pronouns	our; someone; us; we; whatever