

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/157527/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Rennick, Stephanie and Roberts, Seán 2024. The video game dialogue corpus. *Corpora* 19 (1) , pp. 93-106. 10.3366/cor.2024.0299

Publishers page: <https://doi.org/10.3366/cor.2024.0299>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



The Video Game Dialogue Corpus

Stephanie Rennick (University of Glasgow)

Seán G. Roberts (Cardiff University)

Abstract

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

1. Introduction

Video games have become a central form of media, making more money than the film industry (BBC News, 2019) and played by nearly 3 billion people worldwide (Gilbert, 2021). The traditional view of video games as amusements for teenage boys is outdated: most gamers are adults, and half are female (e.g. ESA, 2021). They are now the focus of many social discourses including: the representation of gender, race, and sexuality (see Malkowski & Russworm, 2017; Heritage, 2021); the influence of advertising (Schmidt, 2020); the influence of violent content on crime (e.g. Markey et al., 2015); sociolinguistic ideologies (Goorimoorthee, 2019); and their use in education and second language learning (Gee, 2003; Li, 2022). Children are spending around an hour a day playing games (Qustodio, 2020), and engage in further ‘metagame’ activities (Kahila et al., 2021), making them an increasingly relevant source of language input during acquisition and language change (Eligio & Kaschak, 2020).

Accordingly, video games are increasingly studied from a linguistic perspective. Previous work has focussed on traditionally accessible texts, such as the discourse surrounding games

(Gee, 2014; Potts, 2015; Ensslin & Baltiero, 2019; Heritage, Humphreys & Roberts, 2022), paratexts such as promotional material (e.g. Scharrer, 2004) or conversations between players (e.g. Ensslin 2017; Toh, 2018). Recently, the linguistic content of games themselves has been studied (the ‘text’, compared to the paratext, see discussion in Jones, 2008; Heritage, 2022). However, these usually focus on a single game (e.g. Erdur, 2022), or a limited collection of games (Carrillo Masso, 2009; van Stegeren & Theune, 2020; Pöyhönen et al., 2022). The most extensive studies include 330,000 words from 10 games (Heritage, 2020), and 700,000 words from three games in the Witcher series (Heritage, 2022). However, in both of these studies, texts are not separated into dialogue versus non-dialogue text (item descriptions, in-game letters etc.), and dialogue is not attributed to specific characters. There is currently no large-scale, consistently coded, openly available corpus of video game dialogue. This is a major barrier to replicable, empirical investigations.

This gap is due to several challenges that video games pose for traditional corpus linguistics methods related to obtaining, representing, and sharing video game dialogue. We present the Video Game Dialogue Corpus (VGDC), which addresses each of these challenges using a “self-inflating corpus” design: an open-source pipeline of software and metadata that obtains, processes, and formats video game dialogue into a unified corpus. Section 2 provides the corpus design, including data sources and selection, data format, corpus pipeline, and error-checking procedure. Section 3 demonstrates the use of the corpus in a novel analysis of gendered language in video game dialogue, before a conclusion in section 4.

2. Corpus design

2.1 Data sources

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

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

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

Firstly, as van Stegeren & Theune discuss, accessing source files can be difficult since they are usually compiled to a proprietary format. Source files for some games have been lost (e.g. *Final Fantasy VII*, Square, 1997). Even where source files are available, obtaining dialogue can be difficult. For example, some *King's Quest* scripts are stored as context free grammars, requiring re-assembly. Furthermore, game code is often less standardised than one might expect. For example, cutscene dialogue may be stored separately or rendered into a video.

Furthermore, game code does not always map dialogue to characters straightforwardly. The code for *Dragon Age 2* (BioWare, 2011) stores a sequence of actions which must be matched by timecodes to identify who is speaking. Other games only store associations between speech and IDs for 3D models, and a character may have multiple models. Similarly, some characters may not be assigned names, or only have generic names (e.g. "Guard") which conflates lines from different individuals. These issues mean that the mapping between characters and lines requires manual editing. For example, *Star Wars: Knights of the Old Republic* (BioWare, 2003), taken from van Stegeren & Theune's curation of the game data, required over 1500 edits.

More fundamentally, games have multiple possible texts (similar to film and TV, see Bednarek, 2015), including the text that the scriptwriter wrote, voice actor recording, the dialogue implemented in the code, and dialogue the player experiences. There are larger differences between these readings for games than perhaps any other medium. For example, there are famous instances of differences between the script and voice recording for *Oblivion* (Bethesda, 2006, e.g. [here](#)), and in *Final Fantasy XV* (Square Enix, 2016) we found a small number of lines where the subtitles do not match the recorded audio (e.g. excluding interjections). Also, a player is unlikely to experience all of the dialogue in a modern game for three reasons. First, some dialogue may be considered 'optional': the game may be completable without experiencing it. Secondly, some dialogue will only appear when certain conditions are met (e.g. a combination of party members, skill level). Indeed, dialogue in the game code may be unreachable by players due to programming errors, because the scene was

cut, or if it is part of a debugging utility for developers. Finally, some dialogue depends on player choices: alternative paths in a dialogue tree are not experienced in a single playthrough. Since modern games can take a long time to complete, individual players are likely to miss large proportions of dialogue.

Therefore, we argue that the advantages of game code compared to fan transcripts may have been overstated. Fan transcripts may be more accurate and representative records of player experience (see similar arguments for film and television, Bonsignori, 2009: 187). They have been used in Heritage's (2021) corpus of video game text and to construct corpora in other media (Bednarek, 2018; Kybartas & Verbrugge, 2014). Another way to conceptualise fan transcripts are as crowdsourced transcripts, which have also been used to construct corpora (Adolphs et al., 2020). While there are concerns about the accuracy of fan transcripts (see Heritage, 2021; 104), we found this to be an infrequent issue during our error-checking procedures. For example, in the formal error checking procedures (see section 2.5), we found no transcription errors. We did find transcription errors as we were constructing the parsers, but these would account for much less than one percent of lines in the corpus. In any case, the choice of source is constrained by availability, so a mix of fan transcripts and game data was used.

2.2 Selection of data

50 games were selected for the corpus, guided by various principles of corpus design, including balance and representativeness (McEnery & Brookes, 2022). First, candidate games were identified in the Role Playing Game (RPG) genre (Zagal & Deterding, 2018), that had dialogue as a central mechanic and official dialogue in English (possibly localised from an original source). van Stegeren & Theune (2020:2) suggest that representation in a video games corpus involves sampling “popular or well-known (commercial) games that have a substantial user base”. Accordingly, candidates were only considered if they had sold, or belonged to series that sold, at least 1 million copies worldwide and featured in lists of the top-rated RPGs of all time (e.g. [IGN's top 100 RPGs of all time](#), [Game Informer's top 100 RPGs of all time](#)).

Balance was achieved using a sampling frame (Clear, 1992). The first dimension was RPG style, including ‘Japanese’ and ‘Western’ (JRPG, WRPG, see e.g. Koyama, 2022). While JRPGs originated in Japan, the distinction has come to reflect a group of game traits independent from country of origin (e.g. defined player characters vs. customisable; single main story vs. narrative plurality, see Schules, Peterson & Picard, 2018).

The second dimension was target audience, based on the official ESRB ratings: "Child" (ESRB ratings "Everyone", "Everyone 10+"), "Teen" (ESRB rating "Teen"), and "Adult" (ESRB ratings "Mature 17+" and "Adults Only 18+"). Finally, to allow diachronic comparisons, we aimed for a balance of games across time between 1985 and 2020. To facilitate comparable analyses, we collected multiple games from the same series.

For a given candidate, we searched for an accessible source of dialogue (game code, fan transcription, or wiki), prioritising accessible sources. If none was found, the remaining candidates were reprioritised to balance the sampling frame.

Tables 1 and 2 show the selected games distributed across styles, audiences, and time (see SI for further information). While the corpus is not perfectly balanced, we argue that the balance is acceptable given the constraints of multiple dimensions and the difficulty of obtaining data. Since dialogue in early games was restricted by technical limitations, achieving perfect balance is prohibitive. Instead, we present a source that can be sub-sampled according to a researcher’s requirements (see section 3).

Rating	WRPG Style	JRPG Style	Totals
Child	Stardew Valley, King’s Quest Series, Monkey Island Series	Super Mario RPG, Kingdom Hearts Series	18
Teen	Horizon Zero Dawn, Star Wars: KOTOR	Chrono-Trigger, Final Fantasy Series	20
Adult	Mass Effect Series, Elder Scrolls Series, Dragon Age Series	Persona Series	12
Totals	24	26	50

Table 1: Distribution of games over genre and age rating. The numbers indicate the total number of games in each section. Some series are represented by several games in the corpus.

Years	Games	Number of games
1985 - 1989	FFI; FFII; KQ1; KQ2; KQ3; KQ4	6
1990 - 1994	FFIV; FFV; FFVI; KQ5; KQ6; KQ7; MI 1; MI 2	8
1995 - 1999	Chrono Trigger; FFVII; FFVIII; KQ8; MI 3; Super Mario RPG; Daggerfall	7
2000 - 2004	FFIX; FFX; FFX2; KH; Star Wars: KOTOR; Morrowind	6
2005 - 2009	Dragon Age: Origins; FFXII; FFXIII; KH2; ME1; Persona3; Persona4; Oblivion	8
2010 - 2014	Dragon Age 2; FFXIV; FFXIII-2; FFXIII-LR; KH3D; ME2; ME3; Skyrim	8
2015 - 2020	FFXV; FFVII-R; Horizon Zero Dawn; KQ Chapters; KH3; Persona5; Stardew Valley	7

Table 2: distribution of games over time. FF = Final Fantasy, KQ = King's Quest, KH = Kingdom Hearts, ME = Mass Effect, MI = Monkey Island.

Version 1.0 of the corpus contains 6.2 million words of dialogue from 477,127 lines by 13,587 characters and 1.3 million words of non-dialogue text such as action descriptions, system text, and location information.

2.3 Data format

Highly structured data and metadata are key to facilitating comparative analyses (Davies, 2013; Knight & Adolphs, 2022). van Stegeren & Theune (2020:2) suggest the data format for video game dialogue should be rich (contain dialogue and in-game context), and portable (open-source format). An ideal corpus would also be machine-readable (in order to facilitate analysis with computational methods) and readable by humans (in order to facilitate qualitative analysis). Previous corpora have used a variety of formats. Heritage (2021) uses raw text files which are human- and machine-readable, but does not link dialogue to which character is speaking. Heritage also uses a hand-annotated format for documenting narrative permutations, but this is not strictly machine-readable. van Stegeren & Theune (2020) use a tabular format which lists the speaker, the line of dialogue, an ID number, and the ID numbers of lines of dialogue that can follow. This links dialogue to characters, represents recursive (and graph-like) structures, and is machine readable, but is not very human-readable.

The VGDC uses a JSON format, a plain text format that can be read with an ordinary text editor with several advantages: it can pair lines of dialogue with character names and other metadata; it can represent dialogue trees as recursive structures; is portable and open-source; is machine-readable; and looks like a screenplay script to human readers.

The script for a game is a list of dictionaries. Each dictionary has a main key which represents the name of the character who is speaking. The value associated with this key is the dialogue they speak. Reserved keys separate dialogue from non-dialogue:

- ACTION: description of the action.
- LOCATION: description of the location.
- SYSTEM: a transcription of non-diegetic text that appears to the player but is not spoken by in-game characters.
- CHOICE: a branching choice (see below).
- GOTO: script continues at another location (see below).
- STATUS: contextual status, used in branching choices (see below).

A dictionary can include minor keys, if they begin with an underscore, to convey contextual information. These are used to store a parallel corpus for *Chrono Trigger*, including the original Japanese dialogue, the official English translation and an unofficial fan re-translation (see Williams, 2014; Müller Galhardi, 2014).

For branching dialogue, the main key is labelled “CHOICE” and its value is a list of possible outcomes. Each outcome is a list of dialogue dictionaries. Any of these can itself be a choice structure, allowing recursive branching.

Figure 1 shows a recursive branching dialogue from *Final Fantasy VII* between Cloud (the player character) and Aerith. This represents the tree depicted in Figure 2.

```

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

```

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

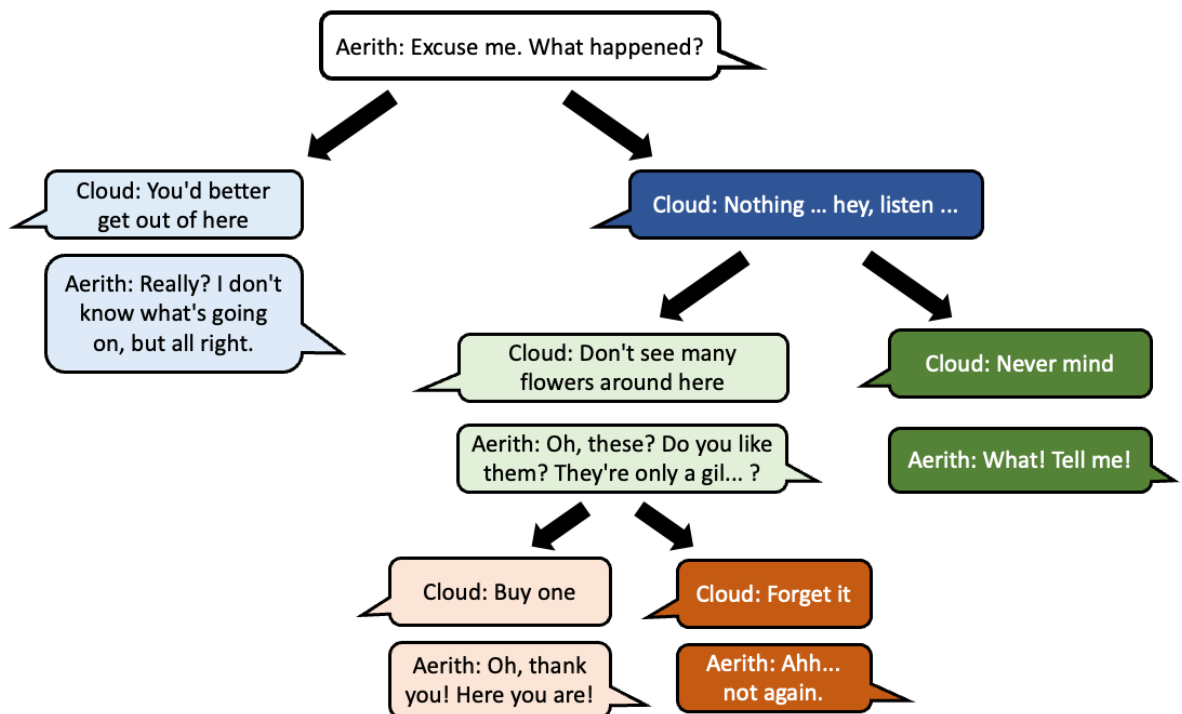


Figure 2: The branching dialogue represented in Figure 1.

Typically, each outcome is a consequence of player choice. The first entry in each outcome indicates the trigger for that outcome (e.g. choosing to buy a flower or not). Non-dialogue triggers can be represented with the main key “STATUS”, including: game conditions (e.g. a certain quest is complete); player character statuses (e.g. player character is female, represented); characters present (e.g. some games allow the player to select a subset of characters to be in their party); alternative responses (e.g. first or second time asking a question); or random choices. One outcome can be an empty list, indicating that there’s a possibility of hearing no additional dialogue.

To handle other types of transition (e.g. jumping to a position higher up the tree), each dialogue dictionary can be given a unique ID assigned to the minor key “_ID”. These can be referred to using a main key “GOTO” to indicate that the script resumes at the dictionary with that ID. This allows full graph-like traversal of dialogue. As further proof of the adequacy of the format, we show it can be used in an implementation of an RPG dialogue system (<https://correlation-machine.com/VGDC/RPG/>).

2.4 Corpus pipeline

The pipeline for creating any corpus involves obtaining data, cleaning, and formatting. VGDC uses replicable methods to create a “self-inflating corpus”. The entire pipeline process for each game is implemented in python code. This has several advantages. First, the code and metadata can be released publicly without sharing copyrighted materials directly. Other researchers can re-run the code in order to “re-inflate” the corpus. Secondly, it speeds up the iterative process of cleaning, formatting, and checking results. Thirdly, it increases transparency: every step is visible to researchers, including every manual edit. Finally, the corpus can be expanded and edited in a centralised and consistent way.

Implementing the pipeline for a game requires three parts: a scraper, a parser and metadata. The scraper is a simple python program that downloads data from the internet into temporary local files. The parser converts these into the standard JSON format. A parser can be applied to several games, though nearly every game required some customisation. The repository includes 10,000 lines of parser code.

The metadata file stores basic information (name, series, year, source), the source type ('fan transcript', 'game data', or 'wiki'), and completeness (determined either from the author statement or from the checking procedure: "complete", virtually all dialogue a player could experience; "high", most dialogue a player would experience on a typical play-through; "sample", e.g. a single play-through of the game without alternative dialogue choices). The parser to be applied is specified, and the "aliases" field stores manual changes to character names such as unifying alternative names, fixing typos or misattributions, and splitting generic names (e.g. "Guard") into individual characters based on lines of dialogue. Finally, the "characterGroups" field is a mapping from group names to a list of character names who are members of that group. Currently, VGDC identifies the gender of each character (manually coded, player-conferred gender, not assuming binary gender, see Rennick et al., under review). This required around 28,000 lines of gender coding and character name unification metadata.

2.5 Error checking

Each game underwent two types of error checking. First for true positives (ensuring that lines in the source correspond to lines in the game) and transcription errors (lines mistranscribed in the source). Three consecutive lines of dialogue were selected randomly from a YouTube video of the game being played. The checker confirmed that the dialogue existed in the corpus, that the text of the transcription was accurate, and that the structure of the dialogue was accurate. The second procedure checked for false positives (lines in the source that are not in the game) and parsing errors (e.g. wrong character assignment or dialogue structure). A random line in the parsed data was selected, and the checker confirmed that the line had been correctly parsed from the source. Both procedures were repeated 5 times per game.

Each game underwent several rounds of checking, applying fixes, and re-checking. Out of 500 tests, these procedures identified source errors in 8% of tests (mostly known limitations of the transcripts), parsing errors in less than 1%, and zero transcription errors. These errors, and many others identified informally, were filed as bug reports on the Github repository and fixed in over 900 commits.

2.6 Availability

The corpus pipeline is available from a centralised github repository (<https://github.com/seannyD/VideoGameDialogueCorpusPublic>). Anyone can suggest edits or contribute data.

3. Demonstration

To illustrate the use of the corpus, the dialogue of female and male characters was compared. Balanced subcorpora were created by selecting 1000 random lines for each gender from 25 games where there was enough data (392,138 words of male dialogue, 369,753 words of female dialogue). A keyness analysis using log likelihood was applied (female dialogue as target). A significance threshold of 0.01 identified 459 keywords. These were categorised using the UCREL system (Rayson et al., 2004), using an automatic tagger (<http://ucrel-api.lancaster.ac.uk/usas/tagger.html>) and manual editing (see SI for code and analysis).

The top key categories included personal names, pronouns, and kin, domains that are also identified in the study on gendered language in video games by Heritage (2021). For personal names, referents are more likely to be male than female (female referents = 38%, male referents = 62%, binomial test $p = 0.04$), reflecting the greater proportion of male characters. Female characters mention names about 50% more than males (Log Likelihood = 145.38, $p < 0.001$). While they mention female referents at about the same rate, female speakers are about twice as likely to mention a male referent than male speakers (Fisher's exact test $p < 0.001$).

Male characters used archaic pronouns (“thy”, “thee”, “thou”) more than female characters, reflecting gendered ideas of chivalry (see e.g. Linderoth & Öhrn, 2014). Most other pronouns were spoken more by females than males, mirroring patterns in real conversations (BNC Spoken corpus, Love et al., 2017, see SI). The reverse is true for “our”. Collocations of “our” in male dialogue included more physical entities (“homeland”, “nation”, “defences”) than the top collocations for female characters (“prayers”, “friendship”), reflecting gendered role stereotypes (see Formanowicz & Hansen, 2022).

Female characters used kin terms more than male characters (Log Likelihood = 91.81, $p < 0.001$), reflecting patterns in real conversation (see SI). The reverse is true for “son”, perhaps for several reasons. Firstly, "son of a bitch" is used more by male characters than female

characters (see SI). Secondly, "son" is used as an affectionate fictive kin term, often by an older man to refer to a younger man, while "daughter" is not necessarily used in the same way. In the *Mass Effect* series, male player characters are called "son", while female player characters are called by non-kin terms ("miss") or non-gendered terms ("child", see literature on generics e.g. Motschenbacher, 2010). Finally, many game worlds are patriarchies, so several quests involve sons and inheritance. In line with this, the frequency of "his son" (patrilineal) is higher than "her son" (see SI).

These examples demonstrate differences in the portrayal of female and male characters. While some may reflect the real world, a large body of literature has demonstrated persistent sexist biases in video games and video game culture (see SooHoo, 2022). Studies that empirically demonstrate differences in the depiction of genders are critical to understanding how these attitudes persist.

4. Conclusion

We presented a large-scale, consistently coded, self-inflating and expandable corpus of video game dialogue. We demonstrated how to apply standard corpus methodologies to VGDC and found significant patterns in gendered language. The corpus is open source and expandable. We hope VGDC will continue to grow and act as a centralising resource for the subfield.

The corpus has limitations. First, as explained in section 2.1, obtaining complete and accurate data is difficult, so the corpus is composed of sources that vary in their type and completeness. However, concerns about the quality of fan transcripts turned out to be less serious than anticipated. Secondly, although VGDC is relatively well balanced at the game level, the distribution of words is skewed towards Western RPGs (63%), adult titles (59%), and games released between 2005-2014. Despite these imbalances, we suggest that the corpus is still representative of an average gaming experience. The corpus includes as much available dialogue for each game as possible so future studies can create their own balanced samples.

We welcome contributions to the VGDC, though contributors should be aware that there are some risks involved with engaging with topics related to video games. Individuals who

comment on video games, including academics, have been the target of co-ordinated online abuse and threats by a portion of the community that see themselves as gatekeepers (Chess & Shaw, 2015). Despite these risks, or perhaps because of them, we hope that research into this medium will continue to progress.

Acknowledgements

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

References

- Adolphs, S., Knight, D., Smith, C. and Price, D., 2020. Crowdsourcing formulaic phrases: towards a new type of spoken corpus. *Corpora*, 15(2), pp.141-168.
- BBC News. 2019. Gaming worth more than video and music combined. 3/01/2019
<https://www.bbc.co.uk/news/technology-46746593>
- Bednarek, M. (2018). *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.
- Bednarek, M., 2015. Corpus-assisted multimodal discourse analysis of television and film narratives. In *Corpora and discourse studies* (pp. 63-87). Palgrave Macmillan, London.
- Bethesda (2006) *The Elder Scrolls IV: Oblivion* [Video game]. Bethesda.
- BioWare (2011) *Dragon Age 2* [Video game]. BioWare.
- BioWare (2003) *Star Wars: Knights of the Old Republic* [Video game]. BioWare.
- Bonsignori, V. (2009) 'Transcribing Film Dialogue: From Orthographic to Prosodic Transcription', in M. Freddi and M. Pavesi (eds.) *Analysing Audiovisual Dialogue. Linguistic and Translational Insights* (Bologna: Clueb), pp. 185–200.
- Carrillo Masso, I., 2009. Developing a methodology for corpus-based computer game studies. *Journal of Gaming & Virtual Worlds*, 1(2), pp.143-169.

- Chess, S. and Shaw, A., 2015. A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, 59(1), pp.208-220.
- Clear, J., 1992. Corpus sampling. *New directions in English language corpora*, pp.21-31.
- Davies, M., 2018. Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In *From data to evidence in English language research* (pp. 66-87). Brill.
- Eligio, R.B. and Kaschak, M.P., 2020. Gaming experience affects the interpretation of ambiguous words. *Plos one*, 15(12), p.e0243512.
- Ensslin, A., 2017. *The language of gaming*. Bloomsbury Publishing.
- Ensslin, A. and Balteiro, I. eds., 2019. *Approaches to videogame discourse: Lexis, interaction, textuality*. Bloomsbury Publishing USA.
- Erdur, N., 2022. Gender in Genshin Impact: A Corpus-Assisted Discourse Analysis. *Language Education and Technology*, 2(1).
- ESA. 2021. 2021 essential facts about the video game industry. Entertainment Software Association.
- Formanowicz, M. and Hansen, K., 2022. Subtle linguistic cues affecting gender in (equality). *Journal of Language and Social Psychology*, 41(2), pp.127-147.
- Fox, J. & Tang, W. Y. 2014. Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Comput. Hum. Behav.* 33, 314–320.
- Gee, J. P. 2003. *What video games have to teach us about learning and literacy*. Palgrave Macmillan.
- Gee, J.P., 2014. *Unified discourse analysis: Language, reality, virtual worlds, and video games*. Routledge.
- Gilbert, N. 2021. Number of Gamers Worldwide 2022/2023: Demographics, Statistics, and Predictions. <https://financesonline.com/number-of-gamers-worldwide/>
- Goorimoorthee, T., Csipo, A., Carleton, S. and Ensslin, A., 2019. Language ideologies in videogame discourse: Forms of sociophonetic othering in accented character speech. *Approaches to videogame discourse: Lexis, interaction, textuality*, pp.269-287.

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

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

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

Heritage, F., Humphreys, C, & Roberts, E. 2022. Are videogames violent? A corpus-assisted diachronic study of the representation of videogames in the press between 2000 and 2020. In *Proceedings of the 6th Corpora and Discourse International Conference*, Emilia-Romagna, Italy.

Jones, S.E., 2008. *The meaning of video games: Gaming and textual strategies*. Routledge.

Kahila, J., Tedre, M., Kahila, S., Vartiainen, H., Valtonen, T. and Mäkitalo, K., 2021.

Children's gaming involves much more than the gaming itself: A study of the metagame among 12-to 15-year-old children. *Convergence*, 27(3), pp.768-786.

Knight, D. and Adolphs, S., 2022. Building a spoken corpus: what are the basics?. In *The Routledge Handbook of Corpus Linguistics* (pp. 21-34). Routledge.

Korea Creative Content Agency. 2020. 7th survey report of video gamers in Korea.

<https://seoulz.com/the-korea-creative-content-agency-kocca/> (2020).

Koyama, Y. 2022. Evolution of a Genre. In R. Hutchinson & J. Pelletier-Gagnon (Eds) *Japanese Role-Playing Games: Genre, Representation, and Liminality in the JRPG*, p.19.

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

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

Linderoth, J. and Öhrn, E., 2014. Chivalry, subordination and courtship culture: being a 'woman' in online games. *Journal of Gaming & Virtual Worlds*, 6(1), pp.33-47.

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

- Malkowski, J. & Russworm, T. M. (eds) 2017. *Gaming representation: Race, gender, and sexuality in video games*. Indiana University Press.
- MacArthur, H.J., Cundiff, J.L. and Mehl, M.R., 2020. Estimating the prevalence of gender-biased language in undergraduates' everyday speech. *Sex Roles*, 82, pp.81-93.
- Markey, P.M., Markey, C.N. and French, J.E., 2015. Violent video games and real-world violence: Rhetoric versus data. *Psychology of Popular Media Culture*, 4(4), p.277.
- McEnery, T. and Brookes, G., 2022. Building a written corpus: what are the basics?. In *The Routledge handbook of corpus linguistics* (pp. 35-47). Routledge.
- Müller Galhardi, R. (2014). Video game and Fan translation: A case study of Chrono Trigger. *Fun for All: Translation and Accessibility Practices in Video Games*, 175-195.
- Potts, A., 2015. 'LOVE YOU GUYS (NO HOMO)' How gamers and fans play with sexuality, gender, and Minecraft on YouTube. *Critical Discourse Studies*, 12(2), pp.163-186.
- Pöyhönen, T., Hämäläinen, M. and Alnajjar, K., 2022. Multilingual Persuasion Detection: Video Games as an Invaluable Data Source for NLP. *arXiv preprint arXiv:2207.04453*.
- Qustodio (2020) Apps and digital natives: the new normal.
https://qweb.cdn.prismic.io/qweb/f5057b93-3d28-4fd2-be2e-d040b897f82d_ADR_en_Qustodio+2020+report.pdf
- Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of LREC 2004, Lisbon, Portugal, pp. 7-12.
- Rennick, S., Clinton, M., Ionnidou, E., Oh, L., Clooney, C., E.T., Healy, E., Roberts, S.G. (under review). Gender bias in video game dialogue.
- Scharrer, E. 2004. Virtual violence: Gender and aggression in video game advertisements. *Mass Communication and Society*, 7(4), 393–412.
- Schmidt, T., Engl, I., Herzog, J. and Judisch, L., 2020. Towards an Analysis of Gender in Video Game Culture: Exploring Gender specific Vocabulary in Video Game Magazines.
- Schules, D., Peterson, J. and Picard, M., 2018. Single-player computer role-playing games. In *Role-Playing Game Studies* (pp. 107-122). Routledge.
- SooHoo, J. 2022. A Systematic review of sexism in video games, DOI: 10.31234/osf.io/xrh36.
- Square (1997) Final Fantasy VII [Video game]. Square.
- Square Enix (2016) Final Fantasy XV [Video game]. Square Enix.

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

Toh, W., 2018. *A multimodal approach to video games and the player experience*. Routledge.

Williams, M. P. (2014) *Chrono Trigger*. Boss Fight Books.

Zagal, J.P. and Deterding, S., 2018. Definitions of “role-playing games”. In *Role-Playing Game Studies* (pp. 19-51). Routledge.