

Applying corpus linguistics to videogame data: Exploring the representation of gender in videogames at a lexical level

Frazer Heritage

Abstract:

This paper argues for the examination of lexical patterns within videogames. In particular, it posits that researchers should examine lexical patterns across large representative samples from a variety of games in order to make more generalizable claims about how discourses around social identities are (re)produced. I thus argue for the use of a method called *corpus linguistics*. I demonstrate that corpus approaches to videogames can reveal fruitful information about how language is used within games of the same genre. This paper is designed to illustrate applications of corpora to the language of videogames. Throughout this paper, I outline the fundamental aspects of this method, primarily due to the dearth of literature which combines this method with videogame texts.

In this paper, I demonstrate the validity of corpus approaches to videogame language via an analysis of the representation of gender. I argue that male and female social actors are represented in different ways, with male characters being associated with physical violence and female characters being constructed in more multifaced ways. Non-binary characters proved representationally lacking in this corpus, and thus because they were statistical outliers in the data are not reported on here.

Keywords: Methodological approaches, corpus linguistics, ludolinguistics, language, gender, representation

1. Introduction

In 2014, feminist media critic Anita Sarkeesian faced a tirade of online backlash for suggesting that, in general, videogames were (and still are) sexist (see Sarkeesian, 2014; Massanari, 2017). One of the main criticisms lobbed against Sarkeesian was that she based claims on close qualitative readings of videogames, and that these close readings would involve a high degree of “cherry picking” data which would fit preestablished hypotheses (see, for example, Hoff Sommers, 2014). Qualitative approaches to the representation of gender and sexuality in videogames can reveal interesting and useful findings. However, I argue that selecting an eclectic range of games, and indeed cherry-picking data from those games, to ensure that findings fit a preestablished hypothesis is deeply problematic. In other words, if one selects a text because they think it will conform to a preconceived idea, the likelihood is that the data will be “pigeon-holed” and elements of the data which do not conform to a this preestablished hypothesis (or indeed reject it) might be ignored.

Various ludological scholars have investigated the representation of gender and sexuality via large scale content analyses (for example, Beasley and Collins-Stanley, 2002; Burgess et al., 2007; Miller and Summers, 2007; Kirkland, 2009; Matthews et al., 2016). A number of scholars have also examined how gender and sexuality are represented in texts around videogames, such as in magazine articles (Summers and Miller, 2014), player-to-player communication (Gray, 2018), websites (Braithwaite, 2014), and interviews with players/content creators (see Shaw, 2014; Potts, 2015) to name a few. However, these studies typically focus on visual aspects relating to gender, such as body size, the degree to which women are visually sexualized, or player’s experiences. This paper differs, in that I argue for the implementation of a linguistic method called *corpus linguistics* (which is both a

quantitative and qualitative method) to explore the representation of gender within videogames.

Although it is useful to examine how gender is represented in videogames at a visual level, I argue that we still need to examine how gender is represented at a linguistic level within videogames as texts. Theoretically speaking, if a woman is not sexualized, but referred to with derogatory terms, then the ways in which that woman is represented is equally problematic. This, therefore, problematises the reliance on a single communicative mode, namely only relying on visual elements, to understand how gender is constructed within videogames. However, as I will demonstrate in this paper, while there is a wealth of literature which has previously examined the visual representation of gender, there is a dearth of literature which considers the role of language with regard to gendered characters in videogames.

Ultimately, I argue for examining how language is used to represent gender within a representative sample of a particular genre of videogame, and argue that corpus linguistics – which is a “big data” approach to analysing language – provides a good starting point for bridging the dearth of literature in the linguistic representation of gender in videogames. Corpus linguistics is a field whereby analysts gather large representative bodies of linguistic data, then run these bodies of data through computer software. This software can reveal information which might not have been visible by manual analysis alone – such as by demonstrating what words are statistically likely to occur with what other words, or what words are statistically likely to appear in one videogame in comparison to language in general. Although this method has been applied to the study of how gender and sexuality are

represented in different forms of media (for example, Baker, 2014; Wilkinson, 2019; Heritage and Koller, 2020), it has not yet been applied to videogames as a text-type. I call for more research into how this method can be implemented in tandem with other methods.

2. Background to ludolinguistics research

This paper contributes to the ludological literature “in one of the most fledgling areas of games studies: communication and discourse” (Ensslin, 2012: 3). The literature within this fledgling subdiscipline of discourse and communication in videogames (i.e. ludolinguistics) is rare, but ever-growing.

The ludolinguistic literature appears to be divisible into four different foci: language teaching (see, for example, Gee, 2005; 2007; Vásquez and Ovalle, 2019); lexicography, localization, and variation (see, for example, Mangiron and O’Hagan, 2006; Fernández Costales, 2012; Ray, 2019); player interactions (see Potts, 2015; Rudge, 2019); and multimodal approaches to ludolinguistics (Machin and Van Leeuwen, 2016; Toh, 2019). Typically, the data which previous ludolinguistics investigations have examined can be categorised into three camps: the first is the language used in videogame paratext, such as in fora and manuals associated with a particular videogame (see Balteiro, 2019; Campos-Pardillos, 2019; Ensslin and Finnegan, 2019). The second is player-interactions, typically gathered through netno/ethnographic methods (see Graham and Dutt, 2019; Kiourti, 2019; Rudge, 2019). The third, and much less frequent, is close analyses of videogames as textual sources of data (though, see Machin and Van Leeuwen, 2016; Goorimoorthee et al., 2019; Ray, 2019).

This latter group of research is the focus of this paper. Clearly, there are a large number of foci and text types, though I would also argue that studies which examine the power

structures inherent in discourse, such as those created by hetero-cis patriarchal values, are relatively under-studied (though, see, for example, Machin and Van Leeuwen, 2016; see also Goorimoorthee et al., 2019 for a discussion on power structures caused by race). I argue that, although there are some studies which have examined the communicative and discursive strategies used within pre-scripted videogames, these studies tend to approach the data with problematic methods. For example, when examining how racialised characters are “othered” through their accents, Goorimoorthee et al., (2019: 274) claim that they: “collect[ed] and label[ed] character speech data to map the distribution of accents throughout a typical playthrough of DAO [*Dragon Age: Origins*]”. While this is certainly a laudable effort – and a considerably laborious process – one must question whether a game such as *Dragon Age: Origins* (Bioware, 2009), which is an open-world game with a multi-layered semi-linear narrative, can ever have a “typical playthrough”.

Indeed, one of the few studies from natural language processing (a field closely related to corpus linguistics) which attempts to reconcile this issue is Fekete and Porkoláb’s (2019) investigation into the linguistic features of place names in *The Elder Scrolls* series (Bethesda, 1994-2018). Fekete and Porkoláb demonstrate that by using computerised methods, it is possible to gather and examine a much more representative sample of data that removes the subjectivity of what constitutes typicality in a playthrough. Although their focus was to examine how different place names were constructed, as opposed to examining how power structures are lexicalised within videogames, their paper demonstrates the fruitful nature of using computers to assist linguistic analysis.

Although previous research has examined how gender is communicated through discourse in both auditory and visual communicative modes (see, for example, Machin and Van Leeuwen, 2016), there is a dearth of research which examines how it is represented through the words and grammatical structures within pre-scripted videogames. Furthermore, while previous ludolinguistics literature has explored lexical approaches to the representation of gender and sexuality, these investigations are typically confined to the study of language within videogame paratext (for example, Ensslin, 2012; Summers and Miller, 2014; Potts, 2015). While these studies provide an insight into how discourses around games are (re)produced, they do not necessarily reflect the discourses within the games themselves.

Previous linguistic research has utilized a variety of methods in lexical analyses. However, drawing on the work conducted analysing the representation of gender in videogame paratext (work which has implemented this method in the study of videogame paratext is discussed in more detail in Section 3.2), I propose the use of corpus linguistics as a complementary method in videogame studies. Corpora (the plural of corpus) are large collections of texts which are representative of one subset of language (or in some cases, language in general). Indeed, Ensslin (2015: 6) draws attention to the possible synergies between corpus methods and the language in videogames. In addressing the fact that these have not previously been combined, she notes: “large, corpus-based studies are needed to study specific aspects of gamer language (e.g., in the world vs. in constituted play) in greater detail than has previously been achievable”. It is this gap in the literature which this paper seeks to begin to bridge.

Having now broadly outlined the trends within the subdiscipline of videogame communication, in the following section, I review the literature which is closest to the present study. I start by reviewing the background literature on corpus linguistics. Given that this paper is aimed at ludological scholars, I introduce the foundational theories in corpus linguistics, with a focus on how these can be used in videogame corpora. Furthermore, I draw on how corpora have been used to analyze the representation of gender in comparable media (such as videogame paratext, television shows, and YouTube let's play videos). Following this, I outline how the corpus of this paper was built, and what software was used to analyze the data.

Once I have outlined the literature which has informed the study, and the data used within the study, I demonstrate how gender can be explored in a corpus of videogames. I start by analysing words which are statistically more likely to occur within the data before exploring words which co-occur with the terms “he” and “she”. Before I discuss the rest of the structure of this paper, I want to take a step back and discuss why these pronouns (which can be viewed as sustaining a problematic assumption on gender) were selected. These gendered pronouns were selected because they emerged as statistically more likely to appear within the data than in comparison to a corpus built around the language used on websites. Although this, therefore, examines fractals of the gender spectrum (see Eckert, 2014), words which denoted non-binary gender identities did not appear in the most statistically significant keywords. Further, although a good number of queer scholars, including those interesting in queer games design, have demonstrated the performed and socially constructed nature of gender and sexuality (see, for example, Butler, 1990; Gray, 2018), media outlets (including large videogame companies) often represent gender identities in typically binary ways (see Coffey-Glover, 2019 for a discussion of this in relation to magazine companies). While different types of masculinity and femininity can be performed within different forms of

media, including videogames, a number of media texts under-represent those who do not conform to the writing team's perceived binary view of gender. Thus, although I would rather avoid a dichotomous analysis of these gender fractals, pronouns such as "they" did not indicate whether the term was used in the singular form or plural form, and of the top 500 words which were statistically more likely to appear in the data than in the reference corpus, none indicated a non-binary gender identity (for example, there were no terms such as "two spirited").

The focus of this paper is primarily to demonstrate the fruitful nature of combining corpus linguistic methods and ludological studies. However, it is not the intended purpose of this paper to argue for corpus linguistics as a replacement method. On the contrary, as I note in the conclusion of this paper, corpus linguistics can be complimentary to videogame studies. While the focus of this paper remains centred on the linguistic features of videogames, it is designed to lay the foundations for combinations with other methodological approaches.

3. Background to corpus linguistics

This section is dedicated to reviewing the literature on corpus linguistics. In this section, I draw on applications of corpus linguistics to the social sciences, namely where corpora have been used to investigate the representation of gender in media which bear similarities to videogames.

3.1 What is Corpus Linguistics?

Corpus linguistics is a subdiscipline of linguistics which is formed around methodological approaches to the study of language in context and how language is used in authentic

examples. Within corpus linguistics, computer software is used to identify patterns of language in and across large data sets (McEnery and Hardie, 2011). Examination of these patterns can reveal information about language, which would be difficult to discover without such software. For example, previous research has used corpus software to reveal patterns of grammar use (Hunston, 2010), phraseological expressions (Römer, 2016), and ideologies (Baker, 2014; McEnery et al., 2015; Heritage and Koller, 2020). The data provided by corpus linguistics allows linguists to use authentic examples to highlight arguments and nuanced differences in patterns of language, which may challenge (or indeed confirm) existing intuitive assumptions of language (see McEnery and Hardie, 2011).

One of the foundational aspects of corpus linguistics is that the corpus being investigated must be a representative sample of language (Sinclair, 2005; Wynne, 2005). In some cases, this sample is representative of the language in general (for example, Love, 2016; 2020; Hawtin, 2017; McEnery and Love, 2018). In other cases, researchers compile specialized corpora, which can reveal how language is used within a specific genre or context. This latter point is particularly important for videogame scholars, who deal with a number of different genres. For example, the language used in a game like *Candy Crush Saga* (King, 2012), is likely to be different, and represent identities in a different way, to a game like *Grand Theft Auto* (DMA Design, 1997). Thus, while we might want to create a corpus of videogames, differences in genre will influence what we can do with a corpus. With regards to representativeness, the notion that corpora should be a representative sample of language within a specific context thus offers a possible way to address the kind of criticisms lobbied

against Goorimoorthee et al.'s (2019) "typical playthrough" approach discussed in Section

2 

There are two types of corpora: specialised corpora and reference corpora (see McEnery and Hardie, 2011), though it is good to view these on a cline. Specialised corpora are collections of texts which are representative samples of one specific text-type. This could be, for example, corpora built around forum posts for one specific videogame (see Ensslin, 2012). Specific corpora tend to be relatively small (usually between 50,000 words and up to about 10 million words) (see Baker, 2014). By comparison, reference corpora tend to be much larger and are built in a way to capture language across a variety of genres and registers. The point of a reference corpus is that it allows analysts to look at what kind of language is used across contexts and in a more general sense. These often tend to be much bigger and can comfortably range from 1 million words (see, for example, Baker, 2009) to in excess of multiple billion words (see, for example, Davies, 2008-onwards; Love, 2020).

Work on the representation of gender in corpus linguistics has utilized both specialized corpora and large reference corpora. For example, Moon (2014) utilized reference corpora in her analysis of terms for gendered social actors of different ages (see also Caldas-Coulter and Moon, 2010; 2016). This work revolved around examining *collocates* and then close reading of *concordance lines*, both of which are considered central methods within corpus linguistics. Although corpus linguistics in itself is a discipline of linguistics, it arguably has

¹ It should be noted that Goorimoorthee et al. (2019) are not the only scholars to implement methods which I believe take a non-representative sample of data. There are a number of other approaches, such as, for example, *close playing* (discussed in Sundén, 2010) which I still view as equally unrepresentative ways of collecting data. This is not to invalidate any one approach, rather draw attention to the fact that these approaches can be criticised from this perspective, and thus I offer corpus linguistics as a way of addressing these criticisms.


three central but different methods: *word frequency analysis* (which examines the frequencies at which words occur, though this method also could include keyword analysis, which allows scholars to examine how statistically likely words are likely to appear in one corpus in comparison to another) (see McEnery et al., 2006), *collocation/collocational analysis*, and *concordance line analysis* (Sinclair, 1991).

Word frequencies (including frequency lists based on statistical keyness) reveal what words are most frequently used in a corpus (see Brookes and Baker, 2017). When two or more corpora are contrasted, keyword lists can be generated. Typically, this tends to be through contrasting a specialised corpus with a reference corpus – though meaningful contrasts can also be made between two specialised or two reference corpora. Keyword lists use statistical measures to show which words are statistically more likely to occur in corpus “A” than corpus “B”. If the corpus used for comparison is representative of general language, or indeed of a genre in a more general sense, a keyword list should reveal what words are more likely to occur, and thus could provide a “starting point” for what should be looked at in more detail. For example, Aull and Brown (2013) examined how gender was represented in news reports about a sporting event. In their investigation, they generated a keyword list, which showed that terms denoting female social actors were more likely to appear in this corpus than a reference corpus. Similarly, Heritage and Koller (2020) generated a keyword list, which demonstrated that terms denoting gendered social actors were prevalent in a corpus of forum posts from an online community of misogynistic men. It is not uncommon once these keywords are generated, to then take a smaller number of words which are statistically key and examine them in more detail, or even examine different elements of the corpus while using the keywords to guide the analysis (see also, Baker, 2014:163-165; Heritage and Koller, 2020).

One such additional way of examining how keywords are used is through *collocation*.

Collocation has its foundations in Firth's (1957:11) notion that: "you shall know a word by the company it keeps". As Baker et al. (2013:36) note, collocates occur "frequently within the neighbourhood of another word, normally more often than we would expect the two words to appear together because of chance". Central to the idea of collocation is that words can begin to take on aspects of the meaning of the words that they collocate with. As Stubbs (2003: 13) states: "Individual words can never be more than a starting point, since it is often collocations which create connections". This is a phenomenon which is aptly illustrated by the concept of discourse prosody (Stubbs, 1994; 2003; Baker, 2016), whereby a word collocates with a set of words which belong to either a specific semantic group or appears in the vicinity of words (or phrases) which indicate positive or negative effect. For example, Baker (2008) has explored the representation of terms for unmarried men and women and argues that the term "spinster" has a negative discourse prosody, as it collocates with words such as "grumpy", while "bachelor" collocates with words such as "eligible". Applied to videogame data, and returning to arguments made at the beginning of this paper, if 80% of female characters across a genre of videogame were visually sexualised but described as "powerful" or "respectable", this would indicate that there might be a disconnect between the visual and linguistic representation of gender.

However, collocates and keywords must always be checked via a close reading of *concordance lines*. Analysts cannot rely on the lexical semantics of a word alone, especially because words might be negated or used in unexpected ways. Concordance line analysis is when a particular term is searched for in the corpus, and every use of that term is provided to

the analyst in  – this can also show all instances of a particular collocation. Checking concordance lines is important in exploring collocation and word frequencies, as it allows the analyst to explore how the words are used. For example, if the word “good” collocated with the word “man”, one may initially believe that there is a positive representation of men. However, concordance lines can reveal more information which may be counter to the analyst’s initial impression. For example, it might be that all instances where “good” collocates with “man” are negated, such as in “he was not a good man”, or it may not be an adjective which directly modifies “man”, such as in “that man was only good at falling flat on his face”. Multiple concordance lines can also be checked to examine repeated patterns around the word, such as other words in context which build up similar representations through language which is not statistically significant.

3.2 (Corpus) Linguistic Approaches to (Gender in) Videogames

One key linguistic study which examined the representation of gender and sexuality in videogame paratext was Braithwaite’s (2014) investigation into the language used by players on fora dedicated to the videogame *World of Warcraft* (Blizzard 2007-onwards) in reaction to the presence of a female character who expressed same-sex attraction. The discourses surrounding specific characters became both gendered and centred on sexuality. Braithwaite’s research offers interesting and important findings on how the community construct gender roles and view gender, sexual, and romantic minorities. Although the study reveals interesting and important findings on how players engage with the representation of sexuality, the description of the methods implemented by Braithwaite are relatively vague. Further, I would argue that the analysis could have benefited from some form of quantification in order to obtain a clearer idea of which findings were representative. This kind of quantification and clarity might have been resolved through the use of corpus

methods, or a mixed-methods approach which also implemented elements of corpus linguistics. This could still have allowed for a large-scale analysis using both quantitative and qualitative methods.

Similar work conducted by Ensslin (2012; 2015) has examined the ludolects of both people who play videogames and the developers of those games. Similar to Braithwaite, Ensslin's (2012) research focused on fora for *World of Warcraft*. However, Ensslin provides more detail about how the data were gathered and analyzed. In particular, Ensslin demonstrates a rigorous and replicable study which utilized a combination of corpus methods. Ensslin's work ultimately was able to look at metaphors and the language used around gendered terms. Similar work has also been conducted by Carillo Masso (2009), who used corpus approaches to examine the language in *Diablo* and *World of Warcraft* fora.. This kind of research demonstrates that it is possible to examine how identities are constructed via corpus methods in videogame paratext – methods which have yielded findings that could be contrasted to analyses of videogames as a text using the same methods.

Finally, outside of *World of Warcraft* fora, Potts (2015) has used corpus linguistics in the examination of homosociality within “LetsPlay” communities on YouTube. Potts argues that players of the videogames and those who watch the players are able to construct homosocial bonds. Potts argues that members of the community under examination were not homophobic but indexed their heterosexual identity while simultaneously showing appreciation for other members of the community. Pott's work utilized a variety of methods, though most pertinent to this paper is her use of building corpora from comments on the YouTube videos.

Of the studies mentioned above, one common theme can be seen: so far, corpus linguistic methods appear to have only been applied to videogame paratext. Although this kind of application can reveal interesting findings, especially about the representation of gender, it has not yet been extended to videogames themselves. If videogames are seen as textual in nature (see Aarseth, 1997; 2004; Frasca, 2001; Juul, 2005), then it stands to reason that they are complicit in the (re)production of discourses *vis-a-vis* the language contained within the text. It is therefore important to analyze this language and the discourses which are presented through the linguistic features of the text. I argue that doing this across a variety of similar genre texts from a similar period may provide a “snapshot” into the discourses which are viewed as “acceptable” within a particular genre of videogame at any given point.

Ultimately, this paper has one overarching research question, which is: “what can a corpus approach to the language used within videogames reveal about the representation of gender?”. Although this is a broad question – and one which could probably be answered over several monographs, the aim of this paper is to demonstrate the validity of taking a corpus approach to this kind of data, with the hope that other scholars will continue to examine this question across different videogames (and videogame genres).

4. Methods: Building the Corpus and Analyzing It

In this section, I outline two different methodological aspects: the first is how the corpus was built. Given that the corpus presented in this paper is unique, I dedicate a large amount of space to describing how it was built, so that other scholars who are interested in applying this method may also do so. Following this, I then outline what corpus software was used in order to analyze the data.

4.1 Building the Corpus

One of the main requirements for any corpus is that it must be a representative sample of the language being analyzed (see McEnery et al., 2006; Baker, 2008; McEnery and Baker, 2017). Furthermore, McEnery and Hardie (2011) argue that the selection of texts in the corpus should adhere to the principle of *total accountability*, whereby the analyst does not only select texts in their corpus which will confirm a presumed theory, and that samples of the language should be taken at random. There is a wealth of literature to suggest that features of genre, register, and temporality will also affect the language used in a given context (for example, Baker and Egbert, 2016; Beiber and Egbert, 2018; McEnery and Baker, 2017). Therefore, there is a need to control what samples are selected for the corpus, and the limitations of a corpus must also be acknowledged. While the sample of language from texts should be random, the videogames themselves must meet preestablished criteria.

In this paper, I take data from a variety of videogames, all of which are rooted within the fantasy genre. The games selected were first-person narratives, whereby the player followed (usually) a single avatar. All games selected were published between 2012 and 2016, in order to view the corpus as a “snapshot” for the representation of gender within videogames around this period. Furthermore, all games included were considered “AAA” at time of release, meaning that they were published on high-end consoles, rather than on mobile phones (unlike the work conducted by Machin and Van Leeuwen, 2016). One of the criteria for the games is that they had to use pre-scripted language in some capacity (so games such as *Journey* (Thatgamecompany, 2012) were excluded, as this game does not contain any language). MMORPGs were excluded primarily because a good portion of their language relies on player-to-player interaction, and I am interested in looking at pre-scripted games (although

this method could be applied to the pre-scripted language used in MMORPGs, this is considerably more sparse in comparison to offline, pre-scripted videogames (see Fizek, 2012)). Finally, all games had a PEGI rating of at least 16+. This decision was implemented so as games aimed at children were not conflated with games aimed at teens and adults – the language in games like *Barbie As The Island Princess* (Human Soft and Ivolgamus, 2007) will certainly be different compared to games like *Bayonetta 2* (PlatinumGames, 2014), owing to what can legally be included for the target audience. In other words, these are videogames which are aimed at either young adults or adults. All these selection criteria ensure that what is analyzed is not influenced by additional factors – such as variation across genre and register.

In order to select the videogames, I took Metro lists of the top 100 videogames for each year (2012-2016 inclusive) (see, for example, GameCentral, 2013), and then removed any games which did not meet the criteria. I then assigned each game a sequential number. I used an automatic number generator to pick a number within the range of these sequential numbers, which indicated which texts could be included in the corpus. Random samples of language were taken from 10 different videogames. As some games had a much smaller amount of dialogue, a limit of approximately 55,500 words was implemented for each game, in order to ensure that the data drawn from (key)word lists, collocates, and concordance lines did not favour these samples from larger games. These videogames and the number of words sampled from each are listed below in Table 1:

Videogame name (year of release)	Word count in each file
Bayonetta 2 (2014)	9,592

Bioshock Infinite (2013)	17,307
Dark Souls 3 (2016)	42,165
Dragon Age: Inquisition (2014)	52,523
Dying Light (2015)	19,580
Fallout 4 (2015)	54,529
Final Fantasy 15 (2016)	36,270
Mass Effect 3 (2012)	23,233
Metal Gear Solid V: The Phantom Pain (2015)	16,397
The Witcher 3: Wild Hunt (2015)	55,425
TOTAL	327,048

Table 1. Files Contained within the Corpus

The data for the corpora were collected in various ways. As Bednarek (2018) notes, there are various tools for collecting corpora from media such as television shows, such as by using corpora of subtitles, fan transcripts, or getting the analyst to transcribe the data². However, each method has varying degrees of accuracy. Collecting data from videogames arguably can be done using similar methods, but rather than subtitles, it is possible for analysts to use code to extract “text dump” files, which contain all the language which would appear on screen to players. For this corpus, the data were gathered using a combination of the three methods outlined above. One methodological issue became that there was no standardized procedure for collecting the data: some videogames, such as *The Witcher 3*, had files which could be

² I initially tried contacting one of the publishers of these games for the data – however, they did not respond to my request. This lack of response indicated that many (if not all) would probably also not respond.

coded out of the text files that were downloaded with the game on PC versions. Others did not afford such capabilities and had to be transcribed (such as in the case of *Bayonetta 2*). In order to adhere to the principle of total accountability, as many dialogue permutations were transcribed as possible. This, therefore, captured more data than a single playthrough. However, given the size of the data, this was not always possible. When this was not possible, different points in the videogames were taken at random and fully transcribed – including for all dialogue permutations. This allowed for samples to be taken from the “open-world” in some of the games, but also allowed for some the data from some games to be captured in their entirety.

4.2 How the Corpus was Analyzed (Software Used)

Given that this paper aims to be accessible to all ludologists who are interested in combining corpora with their research, I decided to only use two different pieces of corpus software, and reserve debate about which software is most appropriate for which task. Furthermore, I use two user-friendly pieces of software, so as if ludologists also wish to conduct a similar study, they will not be overwhelmed with the various complexities often found with specialist software.

For the analysis, I used *WordSmith 7* (Scott, 2016) when analysing keywords. *WordSmith 7* is particularly useful because it triangulates different statistical measures when exploring the keyness of words in comparison to a reference corpus. These measures are BIC score (Bayesian Information Criterion Score), LogLikelihood, and Log Ratio (see Brezina, 2018 for a full explanation of these measures).

For analysis of collocation, I used AntConc (Anthony, 2019). AntConc is useful for other ludologists, as not only does it have a user-friendly interface, but it is also free (keyword analysis can also be conducted in AntConc, but it does not triangulate the statistics like WordSmith 7). In the collocational analysis, I utilize Mutual Intelligence Score (MI Score), The Mutual Information score expresses the extent to which words co-occur compared to the number of times they appear separately. Typically, an MI of ≥ 3 is considered statistically significant (see Baker, 2014). However, others have argued that an MI of ≥ 6 is required for collocates to be “psychologically real” (see Durrant and Doherty, 2010). In this paper, I consider collocates significant and worthy of investigation if they have an MI of ≥ 3 but pay particular attention to collocates with an MI of ≥ 6 .

5. Analysis

In this section, I start by analysing the keywords of the corpus. As Baker (2008) suggests, keywords are a “way in” to the corpus. They often serve as an interesting starting point and can reveal what is salient within a particular dataset. Following this, I move on to a collocational analysis of two gendered terms which appeared in the keyword list: “he” and “she”. Throughout, I refer to concordance lines and how they were examined in order to confirm initial impressions of the quantitative data.

5.1 Keywords

A keyword analysis was conducted by generating a keyword list that compared the wordlist generated for this corpus against a word list generated by using a sample of the Corpus of Global Web-Based English (GLoWbE) (Davies, 2017). The keywords, as organized by BIC score, are presented in Table 2.

Number	Keyword	Frequency	Across “x” texts	BIC score	Log L	Log R
1	THE	17,593	10	30,031.74	30,045.37	6.43
2	TO	9,685	10	15,370.89	15,384.52	5.39
3	YOU	5,898	10	10,340.47	10,354.10	7.07
4	A	6,868	10	9,695.33	9,708.96	4.42
5	OF	7,310	10	8,186.44	8,200.07	3.34
6	AND	5,303	10	7,401.07	7,414.70	4.35
7	I	4,372	10	6,677.31	6,690.95	5.02
8	IN	3,744	10	5,029.90	5,043.53	4.14
9	IT	3,140	10	4,916.86	4,930.49	5.27
10	IS	2,892	10	4,902.72	4,916.35	6.35
11	THAT	2,948	10	4,514.96	4,528.59	5.06
12	FOR	2,555	10	3,968.37	3,982.00	5.19
13	THIS	2,270	10	3,911.28	3,924.91	6.69
14	WITH	1,935	10	3,162.90	3,176.53	5.79
15	BE	1,885	10	2,854.08	2,867.71	4.97
16	YOUR	1,573	10	2,756.92	2,770.55	7.16
17	NOT	1,730	10	2,744.55	2,758.18	5.43
18	WE	1,531	10	2,663.46	2,677.09	6.96
19	WAS	1,587	10	2,625.13	2,638.76	5.98
20	BUT	1,693	10	2,580.18	2,593.81	5.03
21	WHAT	1,465	10	2,497.08	2,510.71	6.50
22	ON	1,834	10	2,409.95	2,423.58	4.04
23	ARE	1,420	10	2,315.16	2,328.79	5.78
24	ME	1,427	10	2,291.45	2,305.08	5.59
25	HAVE	1,426	10	2,266.23	2,279.86	5.46
26	HE	1,434	10	2,136.48	2,150.11	4.85
27	HIS	1,226	10	2,000.32	2,013.95	5.80
28	MY	1,144	10	1,936.45	1,950.08	6.40
29	FROM	1,299	10	1,790.76	1,804.39	4.31
30	IF	1,129	10	1,783.28	1,796.91	5.41
31	THEY	1,083	10	1,771.89	1,785.52	5.85
32	NOCTIS	906	1	1,680.71	1,694.34	140.51
33	AS	1,417	10	1,635.16	1,648.79	3.48
34	DO	1,094	10	1,568.21	1,581.84	4.58
35	HAD	902	10	1,487.23	1,500.86	5.99
36	HER	921	10	1,437.84	1,451.47	5.31
37	AN	956	10	1,406.69	1,420.32	4.78
38	CRANE	789	3	1,406.24	1,419.87	7.93
39	HIM	839	10	1,398.67	1,412.30	6.17
40	HERE	839	10	1,398.67	1,412.30	6.17
41	CAN	957	10	1,375.24	1,388.87	4.60
42	I'M	734	9	1,359.05	1,372.68	140.20
43	VOICE	868	9	1,300.90	1,314.53	4.93
44	AT	1,173	10	1,298.42	1,312.05	3.33
45	THEIR	752	10	1,296.14	1,309.77	6.86
46	GERALT	674	1	1,246.84	1,260.47	140.08

47	JUST	815	10	1,217.17	1,230.80	4.91
48	BY	1,083	10	1,210.56	1,224.19	3.37
49	THEM	735	10	1,160.16	1,173.79	5.45
50	GET	760	10	1,155.52	1,169.15	5.07

Table 2. Keywords in the Videogame Corpus Compared to GLoWbE.

As the keyword list reveals, a fair amount of the keywords are grammatical function words, such as articles (e.g. “the”). This can be useful for other linguistics who are interested in applying corpora analyses of videogames to linguistic subdisciplines, such as using videogames to teach English as a second language (see Gee, 2005; 2007). These words which often serve a grammatical function can also often lead to interesting findings for the representation of identity, even if they do not initially demonstrate it at a lexical level (see, for example, Baker, 2014; Hunt, 2015). However, given the aims of this paper (i.e. to clearly demonstrate how we can search for gender in videogames using corpora) and the limitation of space, within this article, I only focus on content words which explicitly relate to gender.

This data reveals that gendered lexical terms may not be the most frequent within the keywords, but that they are still a salient feature, with approximately 14% of the keyword list being referencing male/female social actors. Within these gendered terms, the data reveal some information about the representation of gender at a quantitative level. The keywords suggest that male characters are more frequently represented in this data set, as three of the keywords, “Noctis”, “Crane”, and “Geralt” are all male characters, as confirmed by concordance lines and examination of pronouns which occur when discussing these characters. Interestingly, female characters within the same games do not appear in the keyword list. This could confirm other research conducted which suggests that women are

underrepresented in videogames (see, for example, Scharrer, 2004; Ivory, 2006; Burgess et al., 2007; Gestos et al., 2018).

Furthermore, the keyword list reveals **that** pronouns are prevalent within this corpus. There were six pronouns which were third person pronouns³: “they”, “their”, “he”, “his”, “him” as well as “her”. Two separate analyses of a random sample of 100 concordance lines for the words “they” and “their” revealed that these pronouns were only ever used as third person plural forms, as opposed to their person singular forms (i.e. they did not denote a non-binary gender identity, rather referred to collective groups). Thus, these games appear to sustain ideologies which reinforce a perceived gender-binary through gender-neutral pronouns only being used as plural forms and not singular forms. While some games might potentially have non-binary characters, or indeed their own terms for non-binary characters, these are much less frequent in comparison to terms for male/female characters, and thus this infrequent representation might be seen as problematic from a quantitative perspective.

With regards to gendered third person pronouns (i.e. “he”, “his”, “him”, and “her”), a T-test revealed that the difference in frequency of occurrence for pronouns denoting male characters was statistically significant ($T = 2.43406$; $p \leq .036$). Similar to the above point, this suggests that there is an unequal representation of gender in the data. However, while references to female characters appear in the top 50 keywords (unlike references to non-binary characters), the data suggests they are underrepresented in comparison to male characters (which supports findings from current content analyses, see Gestos et al., 2018). The data, in combination

³ First and second person pronouns have been excluded for this analysis, because it was not possible to tell the identity of the characters who were speaking.

with previous literature, suggests three things: 1) In the best-selling games, non-binary characters are underrepresented, 2) that women are visually underrepresented in videogames and 3) that female characters are not as frequently referred to as male characters.

However, this finding does not necessarily represent how gendered characters are spoken about on a more qualitative level. The lexical semantics of each word must be examined in order to better understand how it is used. The next section, which examines the collocates of gendered pronouns, still utilizes statistical measures but allows for a deeper understanding of how various words are used to create different prosodies around gendered words.

5.2 Collocational Analysis

As evidenced above in the keyword analysis, the most frequent key gendered pronoun was “he”, and thus I decided to examine how the gendered pronouns “he” and “she” created differences in the representation of gender within the corpus. For this section, collocates within a 5 left – 5 right window span of the node words (“he” and “she”) were examined. Furthermore, a minimum occurrence of ≥ 5 was imposed for collocates.

I start by analysing the collocates of “he” (as detailed in Table 3) before analysing the collocates of “she” (demonstrated in Table 5.) Both tables are organized by MI score.

Rank	Frequency	MI score	Word
1	6	9.57494	biases
2	5	7.98998	defect
3	7	7.79733	answered
4	9	7.35255	departed
5	5	7.3119	museum
6	8	7.28954	hassrath
7	8	7.28954	ben
8	5	7.14198	decides
9	8	6.98998	murdered

10	6	6.98998	hears
11	7	6.89044	rifle
12	8	6.82005	faced
13	5	6.72694	letho
14	11	6.69452	detainee
15	12	6.62074	agreed
16	11	6.49521	felt
17	6	6.48748	empty
18	7	6.4754	showed
19	7	6.4754	capable
20	5	6.40501	ears
21	6	6.32701	mentioned
22	42	6.3162	decided
23	17	6.24455	wants
24	5	6.22444	puts
25	6	6.11551	sniper

Table 3. Collocates of “He”

The collocates of “he” appears to suggest a semantic preference which aligns with the ideas of hegemonic masculinity (see Connell, 2005). Specifically, physical masculinity, which seeks to subordinate other forms of masculinity through physical prowess, violence, and physical domination. In this sample of 25 statistically key collocates, 14 relate to an action of some kind, while 8 relate directly to the semantic domains of war and violence (though, two of these words are part of the same lexical bundle). The lexis which can be categorized as such are reproduced in Table 4. Of note is that some lexis can be categorized in more than one category. In other words, actions can also be linked to violence and were coded as such.

Actions	Defect; answered; departed; decides; murdered; hears; faced; agreed; felt; showed; mentioned; decided; wants; puts
Semantics of war and violence	Defect; (ben) hassrath; murdered; rifle; faced; detainee; sniper

Table 4. Categorisation of Collocates for “He”

A close reading of concordance lines was undertaken in order to best categorize these collocates. For example, to someone who has not played *Dragon Age: Inquisition*, the terms “ben” and “hassrath” do not necessarily reveal anything of interest. However, by conducting a concordance line analysis, it was revealed that the collocates were used as a lexical bundle, and this lexical bundle was used in regard to a group which are comparable to modern-day police, such as in the line:

- (1) [The] [d]etainee has already confessed to **resisting arrest** when **Ben-Hassrath** came for his co-worker” (Dragon Age: Inquisition)

Given this lexical semantic preference for actions, war, and violence, one might argue that male characters in videogames are constructed in a way associated with what Connell (2005) deems as “physical” masculinity: that certain men are seen as masculine due to their physical capabilities (see also, McAllister et al., 2019 for a discussion on “military masculinity”). For example:

- (2) one day **he** awoke and **began to murder** and **destroy** (The Witcher 3)
- (3) [The} [d]etainee asked why **he murdered** Ben-Hassrath, responded that he had only defended himself. (Dragon Age: Inquisition)
- (4) When Sir Yorgh faced Sinh, the slumbering dragon, **he drew blood** with a flash of his steel, but Sinh responded by spewing forth the poison (Dark Souls 3)

Ultimately, these concordance lines demonstrate that the male characters are actively engaging in the violence to some degree, rather than being bystanders. Therefore, it is

possible to suggest that male characters within the corpus are linked to violent actions and, in general, this genre of game may normalize the ideology that masculinity is linked to acts of violence. In turn, this sustains problematic ideologies towards masculinity and (re)produces discourses about what may be expected of men.

Although it is useful to understand how male characters are represented in this corpus of videogames, an examination of the single lexeme “he” does not provide a full insight into the representation of gender. In particular, if all characters (regardless of their gender) are portrayed in an identical way, then this would suggest that there is no bias within videogames at all. Therefore, this paper now explores the representation of female characters using the gendered lexeme “she”.⁴ Similar to the analysis conducted for *man*, the top 25 collocates which occurred ≥ 5 times within a 5 left – 5 right window were examined (see Table 5).

Rank	Frequency	MI score	Word
1	5	8.59281	hits
2	6	8.1554	thinks
3	16	7.74732	herself
4	6	6.99786	tough
5	9	6.60791	wants
6	132	6.5976	her
7	5	6.50534	cipher
8	21	6.39641	woman
9	6	6.33228	says
10	8	6.27088	knows
11	8	6.14159	pretty
12	5	6.06924	fell
13	5	6.06924	died
14	9	6.04849	snake

⁴As noted previously, although I do not view gender in a binary way, the third person singular “they” shares the same orthographic form as the third person plural, and as discussed earlier an analysis of 100 random concordance lines showed that it was only ever used in the plural form. To date, there is a dearth of corpus research which examines the representation of non-binary identities (though, see Zottola, 2018). One reason for this is because most computer software is unable to differentiate between singular and plural forms with identical orthography. While some non-binary identities are lexicalized through distinctly different lexemes (such as *zer* or *zim*), these lexemes do not appear in this corpus. This does, however, open up lines for future research, which might choose to examine queer-games and examine how non-binary gender identities are lexicalised in these.

15	7	6.03384	girl
16	6	6.02295	gave
17	7	6.01214	asked
18	6	5.99786	needed
19	5	5.9781	returned
20	5	5.9781	named
21	6	5.94895	learned
22	10	5.83792	knew
23	5	5.78545	duty
24	56	5.7518	she
25	8	5.71629	wanted

Table 5. Collocates of “She”

The results suggest that there are some similarities between the representation of male and female characters: that female characters are also represented as having physical prowess, through words such as “hits” and “tough”. Indeed, a concordance line analysis revealed that these characteristics were ascribed to the women, rather than other social actors who would then use them against women. For example:

(5) she's been around a long time, she's everywhere, and **she hits hard or she hits light**, but the choosing isn't up to you” (Dragon Age: inquisition)

(6) **She’s tough**, but I’d feel better if we got this over with and got back on the road” (Final Fantasy 15)

However, there are fewer lexemes in this semantic field than there are for male characters. Indeed, a T-test conducted on the frequency at which terms denoting physical violence occurred within the top 25 collocates revealed that the pronoun “he” was statistically more likely to occur with a term relating to physical prowess ($T=2.32048$; $p \leq .024$). In turn, this

suggests that male characters are disproportionately represented as being associated with physical qualities. Nevertheless, it is important to acknowledge that there is some representation of women with physical power and that this may represent a slow (but needed) step towards equal representation within this genre of videogame.

A different way female characters appear to be represented is through the use of collocates which relate to a semantic field of knowledge. For example, “knows”, “learned” and “knew” all suggest that the pronoun “she” is associated with cognitive and mental processes, which suggests that female characters are praised more for their intelligence than physical prowess than male characters. This was confirmed when looking at the concordance lines, such as:

(7) This drive burned all the higher when **she learned from Avallac'h** that Imlerith, the general of the Hunt who murdered Vesemir, would be attending the Crones' sabbath on Bald Mountain. (The Witcher 3)

(8) A Chantry Cleric by the name Mother Giselle has asked to speak to you. **She is** not far, and **knows those involved far better than I**. Her **assistance** could be invaluable. (Dragon Age: Inquisition)

(9) She wanted new memories. **She knew her personality would be all but erased.** (Fallout 4)

Therefore, female characters seem to be presented in different ways in comparison to male characters. Although female characters are less frequently referred to for their physical prowess in comparison to male characters, they are also praised for their mental capabilities- which male characters are not.

It is also worthwhile drawing attention to the adjective “pretty”. Although just a single lexeme as a collocate is evident in the list provided, the existence of this collocate suggests that there is some bias towards female characters: that they are judged on aesthetics while male characters are not. Indeed, the first collocate which describes the physique of a male character, by the statistical measures implemented for Table 3, is “looks” (which is the 154th most frequent collocate), which is semantically neutral by itself and only preceding words give an indication of male character’s physical aesthetics. The fact that this is so much less likely to appear as a higher rank means that, statistically speaking, in comparison to male characters, pronouns for female characters are more likely to be used with terms denoting that character’s appearance. Therefore, it is possible to suggest that there is still some bias towards female characters: that their appearance is still drawn into the ways in which they are discussed, whereas this is less frequent for male characters. However, this data also suggests that the representation of women in games is different to the representation of women on websites, in that the female pronoun is more likely to collocate with words like “pretty” at $MI \geq 6.14$ (for an overview of how women are referred to by their appearance in general mass media, see Caldas-Coulthard and Moon, 2010; Moon, 2014). Thus, future research might want to consider triangulating how gendered social actors are discussed in different forms relating to videogames and contrast this to the language used within videogames via keyword analysis.

6. Discussion

Having now demonstrated that there appears to be a gender bias in this corpus, I now turn to discuss the implications of these findings.

The implementation of corpus linguistic methods to this kind of data has shown that there are repeated patterns in how gender is represented across a “snapshot” of games within one

specific genre – i.e. pre-scripted games, which are rated AAA, and aimed at 16+ year-olds. Given that these games tend to reach a wide audience, looking at how social identities is represented across them is highly important – especially given the normalisation effect of mass media such as videogames.

Although previous research has looked at the frequency at which male to female characters occur (see Paaßen et al., 2016 for an overview; see also Gestos et al., 2018), this research tends to focus on the number of female characters in games, as opposed to how frequently they are referenced (see, for example, Burgess et al., 2006). This therefore potentially gives new an interesting insight into how gender is represented in videogames – are we concerned with the number of female characters? Or are we concerned with how frequently they occur and are discussed? While we may encounter, say, 100 female characters within the tutorial of a game, if they do not occur after the tutorial, and 90% of the rest of the game is dedicated to male characters, then this would suggest a problematic representation of gender. Thus, by using a corpus method, we are able to address this problem – though, this is not to say that it could not be used in tandem with analysis of how many female characters there are in games.

The findings also suggest that characters that have non-binary gender identities are under-represented within this genre. Although the gender-neutral pronouns “they” and “their” occur as statistically significant keywords, a close reading of 100 concordance lines for occurrences of both these pronouns revealed that they were only used as third person plural forms, as opposed to the third person singular forms. This, therefore, demonstrated that these pronouns referenced collective groups, as opposed to individuals who identified as neither male nor female. This lack of representation of non-binary gender identities was compounded by the

fact that words denoting non-binary gender identities (such as “two spirited” used in Native American cultures, see Davis, 2014) did not appear in the top 500 statistically significant keywords. While these characters may exist within the games which the corpus is compiled of, they ultimately have infrequent representations.

Typically, the results tend to show that the kind of male and female characters represented conform to typical and traditional views of men and women: men are represented as enacting physical masculinity (see Connell, 2005), while women are discussed in terms of their knowledge and their physical aesthetics. While there were signs of a counter-discourse (i.e. that women were also strong), this was considerably less frequent than the occurrences at which male characters were represented in this way. Thus, by using this linguistic method, we can see that some counter-discourses are beginning to emerge within the genre, though it is still not at a point of equality just yet.

7. Conclusions and lines for future research

This paper set to explore the question: “what can a corpus approach to the language used within videogames reveal about the representation of gender?”. I have demonstrated that using corpora to analyze the language of videogames can be fruitful, especially with regards to analysing the representation of gender. In particular, I have focused on the statistically driven elements involved in corpus linguistics, and supported assumptions which statistics infer by conducting a close reading of how those words are used in context.

In this paper, I elected to explore the representation of gender, mainly due to the vast wealth of research on the topic and the fact that explicitly gendered words occurred within the

statistically significant keywords. Swann (2002) posits that language and gender researchers must answer the question “yes, but is it gender?” (Swann, 2002: 43). In other words, some sort of reference to gender identities needs to occur for the researcher to safely and accurately attribute representations to that identity. The keywords confirmed that gender was a salient aspect to these games, as male names and pronouns (gender-neutral and gendered) occurred within the top 50 keywords. One issue that occurred was that a large number of lexis within the keyword list were grammatical words, which act in a functional way. While this is common in a large number of specialized corpora, the words could be used to explore different aspects of the language of videogames. Although I elected to focus specifically on gendered words, this is not to say that future research should not examine other aspects of the corpus, as sometimes there can be connections between non-gendered keywords and gendered words which are not statistically key (see Hunt, 2015).

Within this paper, I have only focused on the kinds of words used and their semantic domains – which, although it has provided a useful starting point for explaining how gender is represented within this genre of game, does not quite account for a greater and more detailed linguistic description of the linguistic component’s that build up representations. For example, it would be interesting to contrast the types of adjectives used to pre-modify gendered terms, different kinds of verb collocate could be examined, or scholars may want to examine the grammatical and semantic agency of characters. However, while all of these different routes could give some good examples of how gender is represented at a linguistic level within videogames, given the limitations of space, I have not been able to present these here.

Although previous studies have focused on the visual representation of gender and sexuality via content analyses (see, for example, Matthews et al., 2016) or have explored the linguistic representation of gender in videogame paratext (see, for example, Potts, 2015; Gray, 2018), this paper has taken a new approach to videogames as a text. This is not to dismiss the previous work, rather draw attention to the need to investigate the language within videogames with a rigorous, statistically informed approach to the data. Indeed, I would call for more work which combines the same approaches to both paratext and videogames – as this triangulation could reveal interesting findings about how gender is represented across mediums.

While previous scholars have implemented a number of methods in their analyses for examining how identity is represented in the language used in videogames (see, for example, Goorimoorthee et al., 2019), I have argued that multi-layered and semi-linear narratives, which are often “open-world” are difficult to analyze because there can never be such a thing as a “typical playthrough”. Therefore, I argued that corpus approaches, which have their routes in representativeness and total accountability, can prove useful – as representative samples can be taken across play throughs, and in some cases, all playthrough narrative permutations, and dialogue options can be included in a corpus. This, therefore, allows the possibility to look at representation across what different people might consider “typical” in their playthrough.

This paper has presented findings which are of interest to ludologists, gender scholars, and linguists. It has provided a distilled and synthesized version of corpus linguistics for a field which does not regularly use the methods associated with this linguistic subdiscipline and has

demonstrated how the methods from the field can be implemented in research directly relevant to ludologists. However, corpus linguistic methods are not perfect. Indeed, work should still be conducted in visual analysis, player-computer interactions, audio-visual studies, etc., but this work could also be combined with findings from corpus linguistic methods. While it is important to understand how gendered characters are visually represented, it is also important to examine the language used both by and about them. This raises more lines for future research from this paper: this paper looks specifically at how gender is represented via lexemes for gendered social actors. Future studies may also choose to examine the language used by female characters and the language used by male characters to explore whether or not there are any differences in the representation of language use.

Finally, this paper has only examined the language used within one genre of game. Future research may elect to examine more genres of game and ones aimed at a variety of age ranges. This research could be the starting point of more investigations around the representation of different identities (and their intersections) within videogame corpora, such as sexuality, ethnicity, and/or religion. Thus, while I have highlighted how corpora can be used to analyze the representation of gender within a specific corpus of videogames, more work is required to fully understand how identities are represented on a broader scale within more videogames.

References:

Aarseth, E. (1997). *Cybertext: perspectives on ergodic literature*. Baltimore: John Hopkins University Press.

- Aarseth, E. (2004). Quest games as post-narrative discourse. In R. Marie-Laure (Ed.) *Narrative across media: The languages of storytelling*. (pp.361-376). Nebraska: University of Nebraska Press.
- Anthony, L. (2019). *AntConc (Version 3.5.8)* [Computer Software]. Tokyo, Japan: Waseda University. Accessed 2nd August, 2020 from Laurence Anthony.Net.
<http://www.laurenceanthony.net/>
- Aull, L.L. and Brown, D.W. (2013). Fighting words: a corpus analysis of gender representations in sports reportage. *Corpora*, 8(1), 27-52.
- Baker, P. (2008). *Sexed texts: language, gender and sexuality*. London: Equinox.
- Baker, P. (2009). 'The BE06 Corpus of British English and recent language change.' *International Journal of Corpus Linguistics* 14(3): 312-337.
- Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.
- Baker, P., Gabrielatos, C. and McEnery T. (2013). 'Sketching Muslims: A corpus-driven analysis of representations around the word "Muslim" in the British press 1998-2009'. *Applied Linguistics* 34, 3255-3278.
- Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2), 139-164.
- Baker, P. and Egbert, J. (Eds.) (2016) *Triangulating Methodological Approaches in Corpus-Linguistic Research*. London: Routledge.
- Balteiro, I. (2019). Lexical and Morphological Devices in Gamer Language in Fora. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 39-57). London: Bloomsbury.

- Beasley, B. and Collins Standley, T. (2002). Shirts vs. skins: Clothing as an indicator of gender role stereotyping in video games. *Mass Communication & Society*, 5(3), 279-293.
- Bednarek, M. (2018). *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.
- Biber, D. and Egbert, J. (2018). *Register variation online*. Cambridge: Cambridge University Press.
- Braithwaite, A. (2014). 'Seriously, get out': Feminists on the forums and the War (craft) on women. *New Media & Society*, 16(5), 703-718.
- Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.
- Brookes, G. and Baker, P. (2017). 'What does patient feedback reveal about the NHS? A mixed methods study of comments posted to the NHS Choices online service'. *BMJ Open* 7(4).
- Burgess, M., Stermer, S., and Burgess, S. (2007). Sex, lies, and video games: The portrayal of male and female characters on video game covers. *Sex roles*, 57(5-6), 419-433.
- Butler, J. (1990). *Gender Trouble*. Abington: Routledge.
- Caldas-Coulthard, C. and Moon, R. (2016). Grandmother, gran, gangsta granny semiotic representations of grandmotherhood. *Gender and Language*, 10(3), 309-339.
- Caldas-Coulthard, C.R. and Moon, R. (2010). 'Curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse and Society*, 21(2), 99-133.

- Campos-Pardillos, M. (2019). End-user Agreements in Videogames: Plain English at Work in an Ideal Setting. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 116-136). London: Bloomsbury.
- Carrillo Masso, I. (2009). Developing a methodology for corpus-based computer game studies. *Journal of Gaming & Virtual Worlds*, 1(2), 143-169.
- Coffey-Glover, L. (2019). *Men in Women's Worlds: Constructions of Masculinity in Women's Magazines*. London: Palgrave Macmillan.
- Connell, R. (2005). *Masculinities*. Sydney: Polity Press
- Davis, J. (2014). "More than just 'gay Indian'": Intersecting articulations of two-spirit gender, sexuality, and indigenesness. In L. Zimman., J. Davis., J. Raclaw (Eds.), *Queer excursions: Retheorizing binaries in language, gender, and sexuality* (pp.62-80). Oxford: Oxford University Press.
- Davies, M. (2008-onwards). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Accessed 2nd August, 2020 from Brigham Young University. <https://corpus.byu.edu/coca>
- Davies, M. (2017). *The Corpus of Global Web-based English*. Accessed 2nd August, 2020 from English-corpora.org. <https://www.english-corpora.org/glowbe/>
- Durrant, P. and Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6 (2), 125-155.
- Eckert, P. (2014). The problem with binaries: Coding for gender and sexuality. *Language and Linguistics Compass*, 8(11), 529-535.

- Egbert, J. and Baker, P. (2019). Research Synthesis. In P. Baker and J. Edgbert (Eds.), *Triangulating Corpus Methodological Approaches in Linguistic Research* (pp.183-208). London: Routledge.
- Ensslin, A. (2012). *The Language of Gaming*. Basingstoke: Palgrave
- Ensslin, A. (2015). Discourse of Games. In K. Tracy, T. Sandel, and C. Ilie (Eds.), *The International Encyclopedia of Language and Social Interaction* (pp. 406-411). New Jersey: Wiley-Blackwell.
- Ensslin, A. and Finnegan, J. (2019). Bad Language and Bro-up Cooperation in Co-sit Gaming. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 139-156). London: Bloomsbury.
- Fekete, T. and Porkoláb, Á. (2019). From Arkngthand to Wretched Squalor: Fictional place-names in The Elder Scrolls universe. *ICAME Journal*, 43(1), 23-58.
- Fernández Costales, A. (2012). Exploring translation strategies in video game localization. *MonTI. Monografías de Traducción e Interpretación* 4(1), 385-408.
- Firth, J. (1957). *A synopsis of linguistic theory. Studies in linguistic analysis*. Oxford: Blackwell.
- Fizek, S. (2012). *Pivoting the Player: A Methodological Toolkit for Player Character Research in Offline Role-Playing Games*. (Unpublished PhD thesis). Bangor University, Wales.
- Frasca, G. (2001). *Videogames of the oppressed: Videogames as a means for critical thinking and debate*. (Published Master's thesis). Georgia Institute of Technology, Georgia.
- Gee, J. (2005). *Why video games are good for your soul: Pleasure and learning*. Melbourne: Common Ground.

- Gee, J. (2007). *Good games and good learning*. New York: Peter Lang Publishing.
- Gestos, M., Smith-Merry, J., and Campbell, A. (2018). Representation of Women in Videogames: A Systematic Review of Literature in Consideration of Adult Female Wellbeing. *Cyberpsychology, Behavior, and Social Networking* 21(9). 535-541.
- Goorimoorthee, T., Csipo, A., Carleton, S. and Ensslin, A. (2019). Language Ideologies in Videogame Discourse: Forms of Sociophonetic Othering in Accented Character Speech. In A. Ensslin, and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 269-287). London: Bloomsbury
- Graham, S. and Dutt, S. (2019). “Watch the Potty Mouth”: Negotiating Impoliteness in Online Gaming. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 201-225). London: Bloomsbury.
- Gray, K. L. (2018). Gaming out online: Black lesbian identity development and community building in Xbox Live. *Journal of lesbian studies*, 22(3), 282-296.
- Hawtin, A. (2017). The British National Corpus Revisited: Developing parameters for the Written BNC2014. *Corpus Linguistics 2017 Conference*. University of Birmingham, UK. July 2017.
- Heritage, F. and Koller, V. (2020). Incels, ingroups, and ideologies: The representation of gendered social actors in a sexuality-based online community. *Language and Sexuality* 9 (2), 153-180.
- Hoff Sommers, C. (2014). *Factual Feminist*. New York: American Enterprise Institute.
- Hunston, S. (2010). How can a corpus be used to explore patterns. In A. O’Keeffe and M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp.152-166). London: Routledge

- Hunt, S. (2015). Representations of Gender and Agency in the Harry Potter Series. In: P. Baker and T. McEnery (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (pp. 266-284). London: Palgrave Macmillan.
- Ivory, J. (2006). Still a man's game: Gender representation in online reviews of video games. *Mass Communication & Society*, 9(1), 103-114.
- Juul, J. (2005). *Half-real. Video games between real rules and fictional worlds*. Massachusetts: MIT Press.
- Kirkland, E. (2009). Masculinity in video games: The gendered gameplay of silent hill. *Camera Obscura*, 24(2), 161-183.
- Kiourti, E. (2019). “Shut the Fuck up Re! Plant the Bomb Fast!”: Reconstructing Language and Identity in First-person Shooter Games. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 157-177). London: Bloomsbury.
- Love, R. (2016). “Normal with a brummy twang”: dealing with metadata in the Spoken BNC2014. *IVACS 2016 Conference*. Bath Spa University, UK. June 2016.
- Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge.
- Machin, D. and Van Leeuwen, T. (2016). Sound, music and gender in mobile games. *Gender and Language*, 10(3). 412-432
- Mangiron, C. and O’Hagan, M. (2006). Game Localisation: unleashing imagination with ‘restricted’ translation. *The Journal of Specialised Translation*, 6(1), 10-21
- Matthews, N.L., Lynch, T. and Martins, N. (2016). Real ideal: Investigating how ideal and hyper-ideal video game bodies affect men and women. *Computers in Human Behavior*, 59 (1), 155-164.

Massanari, A. (2017). #Gamergate and The Fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329-346.

McAllister, L., Callaghan, J.E. and Fellin, L.C. (2019). Masculinities and emotional expression in UK servicemen: 'Big boys don't cry'?. *Journal of Gender Studies*, 28(3), 257-270.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., McGlashan, M. and Love, R. (2015). Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. *Discourse & Communication*, 9(2), 237-259.

McEnery, T. and Love, R. (2018). Bad Language. In J. Culpeper., F. Katamba, P. Kerswill, R. Wodak, and T. McEnery (Eds.), *English Language: Description, Variation and Context* (2nd ed.) (495-507). London: Palgrave.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge: London.

McEnery, T. and Baker, H. (2017). *Corpus linguistics and 17th-century prostitution: computational linguistics and history*. London: Bloomsbury.

GameCentral. (2013). 100 best-selling video games of 2012 revealed. *Metro*. Accessed 2nd August, 2020 from: <https://metro.co.uk/2013/01/14/100-best-selling-games-of-2012-revealed-3351774/>

Miller, M. and Summers, A. (2007). Gender differences in video game characters' roles, appearances, and attire as portrayed in video game magazines. *Sex roles*, 57(9-10), 733-742.

- Moon, R. (2014). From gorgeous to grumpy: adjectives, age and gender. *Gender & Language*, 8(1), 5-41.
- Paaßen, B., Morgenroth, T., and Stratemeyer, M. (2017). "What is a true gamer? The male gamer stereotype and the marginalization of women in video game culture." *Sex Roles* 76(7), 421-435.
- 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), 163-186.
- Ray, A. (2019). Playing with the Language of the Future: The Localization of Science-fiction Terms in Videogames. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 87-115). London: Bloomsbury.
- Römer, U. (2016). Teaming up and mixing methods: collaborative and cross-disciplinary work in corpus research on phraseology. *Corpora*, 11(1), 113-129.
- Rudge, L. (2019). I cut it and I ... well now what?': (Un)collaborative Language in Timed Puzzle Games. In A. Ensslin and I. Balterio (eds.) *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp.178-200). London: Bloomsbury
- Sarkeesian, A. (2014). Tropes vs. Women. Feminist Frequency: Conversations with Pop Culture. Accessed 2nd August, 2020 from YouTube.
<https://www.youtube.com/channel/UC7Edgk9RxP7Fm7vjQ1d-cDA>
- Scharrer, E. (2004). Virtual violence: Gender and aggression in video game advertisements. *Mass Communication & Society*, 7(4), 393-412.
- Scott, M. (2016). *WordSmith Tools, Version 7*. [Computer software]. Accessed 2nd August, 2020 from Lexical Analysis Software Ltd. <https://www.lexically.net/wordsmith/>
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press

- Sinclair, J. (2005). Corpus and Text - Basic Principles. in M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books:
- Shaw, A. (2014). *Gaming at the edge: Sexuality and gender at the margins of gamer culture*. Minnesota: University of Minnesota Press.
- Stubbs, M. (1994). Grammar, text and ideology. *Applied Linguistics*, 15 (2), 201-23.
- Stubbs, M. (2003). Conrad, concordance, collocation: heart of darkness or light at the end of the tunnel?'. Presented at *The Third Sinclair Open Lecture*. University of Birmingham, UK.
- Summers, A. and Miller, M. (2014). From Damsels in Distress to Sexy Superheroes: How the portrayal of sexism in video game magazines has changed in the last twenty years. *Feminist Media Studies*, 14(6), 1028-1040.
- Sundén, J. (2010). A sense of play: affect, emotion and embodiment in "World of Warcraft". In M. Liljeström and S. Paasonen. (Eds.), *Working with Affect in Feminist Readings : Disturbing Differences*. London: Taylor & Francis.
- Swann, J. (2002). Yes, but is it gender? In L. Litosseliti and J. Sunderland (Eds.), *Gender identity and discourse analysis* (pp. 43-67). Amsterdam: Benjamins.
- Toh, W. (2019). The Player Experience of BioShock: A Theory of Ludonarrative Relationships. In A. Ensslin and I. Balteiro (Eds.), *Approaches to Videogame Discourse: Lexis, Interaction, Textuality* (pp. 247-268). London: Bloomsbury.
- Vásquez, G.C. and Ovalle, J.C. (2019). Video Games: Their Influence on English as a Foreign Language Vocabulary Acquisition. *GIST--Education and Learning Research Journal* 19(1), 172-192.

Wilkinson, M. (2019). 'Bisexual oysters': A diachronic corpus-based critical discourse analysis of bisexual representation in *The Times* between 1957 and 2017. *Discourse & Communication*, 13(2), 249-267.

Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.

Zottola, A. (2018). Transgender identity labels in the British press: A corpus-based discourse analysis. *Journal of Language and Sexuality*, 7(2),237-262.

Ludology:

Human Soft and Ivolgamus. (2007). *Barbie as the island princess* [Nintendo DS]. Digital game directed by Greg Richardson, published by Activision.

Bethesda. (1994-2018). *The Elder Scrolls* [Microsoft Windows]. Digital game directed by Todd Howard, published by Bethesda.

Bethesda. (2015). *Fallout 4* [Microsoft Windows]. Digital game directed by Todd Howard, published by Bethesda.

Bioware. (2009). *Dragon Age: Origins* [Microsoft Windows]. Digital game directed by Dan Tudge, published by Electronic Arts.

BioWare. (2012). *Mass Effect 3* [Microsoft Windows]. Digital game directed by Casey Hudson, published by Electronic Arts.

Bioware. (2014). *Dragon Age: Inquisition* [Microsoft Windows]. Digital game directed by Mike Laidlaw, published by Electronic Arts.

Blizzard Entertainment. (v. 8.3.7, 2020). [2007]. *World of Warcraft: Battle for Azeroth* [Microsoft Windows]. Digital game directed by Rob Pardo, Jeff Kaplan, and Tom Chilton, published by Blizzard Entertainment.

CD Projekt Red. (2015). *The Witcher 3: Wild Hunt* [Microsoft Windows]. Digital game directed by Konrad Tomaszkiewicz, Mateusz Kanik, and Sebastian Stępień, published by CD Projekt Red.

FromSoftware. (2016). *Dark Souls 3* [Microsoft Windows]. Digital game directed by Hidetaka Miyazaki Isamu Okano Yui Tanimura, published by Bandai Namco Entertainment

Irrational Games. (2014). *Bioshock: Infinite3* [Microsoft Windows]. Digital game directed by Ken Levine, published by 2K Games

King. (2012). *Candy Crush Saga* [Iphone and Windows phone]. Digital game directed by Riccardo Zacconi, published by King.

Kojima Productions. (2015). *Metal Gear Solid V: The Phantom Pain* [Microsoft Windows]. Digital game directed by Hideo Kojima, published by Konami.

PlatinumGames. (2014). *Bayonetta 2* [Wii U]. Digital game directed by Yusuke Hashimoto, published by Nintendo.

DMA Design. (1997). *Grand Theft Auto* [Play Station 2]. Digital game directed by Keith R. Hamilton, published by DMA Design.

Square Enix. (2016). *Final Fantasy XV* [Play Station 4]. Digital game directed by Hajime Tabata, published by Square Enix.

Techland Publishing. (2015). *Dying light* [Microsoft Windows]. Digital game directed by Paweł Marchewka and Adrian Ciszewski, published by Warner Brothers.

Thatgamecompany. (2012). *Journey* [Microsoft Windows]. Digital game directed by Jenova Chen, published by Sony.

Captions list:

Table 1. Files Contained within the Corpus

Table 2. Keywords in the Videogame Corpus Compared to GLoWbE.

Table 3. Collocates of “*He*”

Table 4. Categorisation of Collocates for “*He*”

Table 5. Collocates of “*She*”

Table 6. Frequency of Man* and Woman* as Agent and Patient in Transitive Verbs