Frane Malenica

Department of English Studies, University of Zadar, Croatia
fmalenica@unizd.hr

Picking up the Scraps—Analyzing Video Game Reviews Using Web-Scraping Tools

Abstract

The methods for creating corpora from websites have been in use for almost two decades (Baroni and Ueyama 2006; Baroni et al. 2009), and numerous tools for extracting textual data and metadata from websites have been developed since either as standalone programs, browser extensions, or as packages and libraries in programming languages such as Python and R (cf. Bradley and James 2019; Diouf et al. 2019; Kumar and Roy 2023). The widespread availability of these tools has allowed scholars to create custom corpora on a wide array of very specific topics, such as song lyrics (Kreyer and Mukherjee 2009; Werner 2012; Motschenbacher 2016), comics (Dunst et al. 2017; Unser-Schutz 2011), video games (Heritage 2020), and video game reviews (Guzsvinecz 2022; Arik 2022; HaCohen Kerner et. al. 2020). Previous research in this domain, conducted by Cho et al. (2020), has also demonstrated the effectiveness of NLP methods in extracting and identifying the main themes of video games. In this paper, I will present the results of research conducted on a corpus of video game reviews collected from the GameSpot website (www.gamespot.com) using the rvest package (Wickham 2021) for web scraping in R, and analysed using a combination of traditional corpus linguistic (CL) methods and Natural Language Processing (NLP) methods available in the quanteda package (Benoit et al. 2018). The main aims of this paper are to: i) identify words and phrases typical for different genre of video game reviews; ii) test the applicability of web scraping and NLP methods for linguistic research. While frequency-based analysis is good for a cursory glance at words and phrases typical for this register, the keyword analysis offers more useful results. The results of the sentiment analysis show statistically significant correlation between polarity and ratings, further highlighting the usefulness of these methods.

Keywords: corpus collection, keyword analysis, n-grams, sentiment analysis, specialized corpora, video game reviews, web scraping

1. Introduction

The increased popularity of online discourse and communication and the availability of digital linguistic materials in the past several decades have paved the way for the creation of corpora based on texts collected from the internet. The first corpora collected under this web-as-corpus paradigm (first envisaged in Kilgarriff 2001)

were based on some of the most studied languages in the world, such as English, German, Italian, and Japanese (Baroni and Ueyama 2006; Baroni et al. 2009), but the same methodology was soon applied to a whole variety of other languages and registers (Brezina 2018: 18; Biber and Reppen 2015: 37), thus confirming Kilgarriff's (2001: 345) prophetic claim: "The corpus of the new millennium is the web". While certain aspects of usage of linguistic data collected online have been questioned in terms of several criteria, such as authenticity, representativeness, size, topics, etc. (cf. Gatto 2011, 2014), the pervasiveness of digital communication has made them an indispensable source of information two decades after the initial idea.

This is reflected in the number of various tools for extracting textual data and metadata from websites that have been developed since as standalone programs, browser extensions, or as packages and libraries in programming languages such as Python and R (cf. Bradley and James 2019; Diouf et al. 2019; Suchomel 2020; Kumar and Roy 2023). Diouf et al. (2019) provide a comprehensive overview of different approaches, tools, and areas of application for web scraping, and their list of ready-made tools, such as browser extensions, different software packages and platforms, and libraries in programming languages includes more than 20 different tools that were available at the time the article was published.[1] Bradley and James (2019) provide a detailed tutorial for using the *rvest* package in R to extract and store data from webpages and entire websites, supplementing their tutorial with example R scripts made available via the Open Science Framework system, while Kumar and Roy (2023) describe a similar technique using the Python programming language.

This paper provides a brief demonstration of contemporary methods for creating and analysing corpora using specialized libraries in programming languages like R and their usefulness for analysing specialized registers, such as video games and video game reviews. The main two aims of the research component of this paper are to see whether the methods in question can be used to identify the words and phrases typical for the genre in question and whether the NLP methods, such as sentiment analysis, can be used productively in linguistic research. The paper is outlined as follows—in Section 2, I provide a brief overview of previous research on custom corpora that utilized tools similar to those described in the paper; in Section 3, I discuss the importance of video games and video game reviews in recent linguistic research; in Section 4, I describe the research questions and the methodology of the paper; in Section 5, I present the results of the research, and in Section 6, I provide conclusions, an outlook concerning the potential use of the methods in question for future studies.

1 The list of available tools has probably increased since, but the number of tools available at that point in time should certainly suffice for the purposes of linguistic research.

2. Studies on Custom Corpora

The widespread availability of tools for extracting textual data and converting them into corpora has coincided with the tendency of scholars to create custom corpora on a wide array of very specific topics, such as song lyrics (Kreyer and Mukherjee 2009; Werner 2012; Motschenbacher 2016), literary texts (Fischer-Starcke 2009; Moustafa 2022), football commentaries (Merullo et al. 2022), comics (Dunst et al. 2017; Unser-Schutz 2011), and video games (Heritage 2020).

In his study of American and British pop songs, Werner (2012) analyses the lexico-grammatical and morphosyntactic features of commercially successful songs, and the diachronic development of pop song lyrics. The corpus for this study was collected using the Songtext website and annotated using the CLAWS tagger. While relatively small in size (less than 400,000 tokens in total), Werner's corpus reveals some interesting patterns of pop song lyrics, such as low lexical density, high number of contractions, and first and second person pronouns, as well as some different tendencies between the two subcorpora, e.g., the UK corpus leaning towards a more "Americanized" flavour. On a somewhat more traditional note, Fischer-Starcke (2009) analyses the features of Jane Austen's Pride and Prejudice (P&P), by looking at the keywords and the most frequent phrases via the Wordsmith tool. By comparing the P&P corpus with the two reference corpora (the corpus containing Jane Austen's work minus P&P, and the corpus containing 30 novels published from 1740 to 1859 by different authors), she identifies the keywords patterns for every pairwise comparison. By comparing the P&P corpus to the rest of Jane Austen's works, she identifies five main patterns—family and family relationships, women, men, personal pronouns, military, while the keyword comparison with other novels of the same period reveals six keyword patterns—mental concepts and emotions, women, love, courtship and marriage, family and family relationships, communication, men (ibid: 498). According to Fischer-Starcke, the use of corpus linguistic methods allows the research to uncover the patterns which cannot be perceived intuitively and as such provides a useful tool for supplementing traditional literary critical analyses.

The usefulness of computational tools for revealing the hidden patterns in texts is even more prominent in research dealing with more contemporary genres of texts. Merullo et al. (2019) use the *spaCy* library in Python to analyse the data from the corpus of American football commentaries from over 1400 games and over 6 decades and investigate potential racial biases in player description. While their research, according to their own admission, is faced with statistical and linguistic confounds (e.g., different racial distribution across different player positions), it does illustrate some indicative patterns in sports commentaries, such as increased first-name reference for non-white players and different positive terms used to describe white and non-white players in different positions (Merullo et al. 2019: 6358-6359). Even more importantly, it highlights the capability of corpora and NLP methods to identify various biases in language use. In a similar vein, Heritage

(2020) uses the corpus tools WordSmith and AntConc to investigate representation of gender in video games. Specifically, he uses the keyword analysis to look at the most representative words of the video game corpus and looks at the collocations of masculine and feminine pronouns (he and she) to see whether they co-occur with different sets of lexemes. His results indicate that the female characters tend to be described differently in the sense of their physical capabilities being less pronounced and more emphasis being placed on their mental abilities, which is not the case with the male characters. A similar corpus based on multimodal sources is presented by Dunst and his associates who present "the first digital corpus of graphic novels, memoirs, and non-fiction written in English" (Dunst et al. 2017: 15). Their corpus is annotated using the Graphic Narrative Markup Language (GNML) to include the interrelations between textual and visual information in the corpus and is supplemented with eye-tracker data of several readers for the first chapter of each graphic narrative. The data collected using the eye-tracker show a degree of consistency in terms of attention that the readers give to a certain part of the page, such as text in captions, faces, hands, and objects relevant for the story. According to Dunst et al. (2017: 19), the tools for the creation of this corpus can be of use to multiple disciplines in the humanities and social sciences, such as linguistics, media and literary studies, and psychology.[2]

What this brief and certainly non-exhaustive overview of custom corpora-based research shows is that the advances in technology for collecting, annotating, and analysing texts has allowed us to explore such diverse genres and modalities of language use that the limits seem virtually non-existent. Thus, the main concern for every researcher should not be whether there are any tools to do the job, but to find the right domain in which to use those tools.

3. Video Games as Sources of Linguistic Data

Parallel with the development of corpora based on digital texts described in Section 2, the past two decades have also witnessed increased interest in video games as valuable sources of linguistic data, primarily focusing on their effect on language learning (inter alia Gee 2003, 2013; deHaan 2005; Sylvén and Sundqvist 2012; Zhonggen 2018; Camacho Vásquez and Ovalle 2019). In one of the first comprehensive studies on significance of video games on culture and learning, Gee argues: "Video games are a new form of art. They will not replace books; they will sit beside them, interact with them, and change them and their role in society" (2003: 204). Although one of the first empirical studies on the correlation between language acquisition and video games showed a facilitation effect of video games (deHaan 2005), multiple drawbacks of the study did not allow for any broad-sweeping

2 A more comprehensive list of papers from this project can be found on the web page of The Hybrid Narrativity Project: https://groups.uni-paderborn.de/graphic-literature/wp/?lang=en.

generalizations to be made. As deHaan (2005: 282) himself notes, the study was limited to a single participant, it took the participant too much time to learn how to play the game, the majority of the data comes from self-report questionnaires, and the same test was used to test the acquisition before and after playing the video game. However, subsequent studies have remedied these issues by taking a larger sample. For instance, Sylvén and Sundqvist (2012) look at the data collected from 86 students aged 11-12 using a questionnaire, a language diary, and a three-part vocabulary test. Their results show correlation between the amount of time spent gaming and the results of the vocabulary test. A similar result is reported by Chen and Yang (2013), whose two studies show that playing adventure games has a positive effect on EFL students learning new L2 vocabulary items and that language learners have a positive opinion on the benefits of video games in language learning, despite some potential issues with the games themselves.[3] The significance of video games for language acquisition and learning in general is perhaps best reflected in the establishment of the Game-Based Learning paradigm (Burmester et al. 2006; Santos 2017; Kasemap 2017).

3.1. Studies on Video game Reviews

Beyond the scholarly interest in video games from the perspective of language acquisition and gender representation (Heritage 2020, 2022a, 2022b), video game reviews and similar texts have also become a valuable object of inquiry in the past several years. For instance, Guzsvinecz (2022) looked at the corpus of 993,932 Steam reviews of 21 games belonging to the so-called "Souls-like" genre[4] collected using the *steam_reviews* library in Python[5] and found a slight-to-moderate positive correlation between time spent playing and positive reviews and identified some of the most liked aspects of video games mentioned in positive reviews (e.g., medieval setting, drawn graphics, 2D graphics) and some of the least liked aspects (pixel graphics and futuristic setting). Using a similar methodology with the *PRAW* (Python Reddit API Wrapper) library, Arik (2022) created a corpus of comments from the social media platform Reddit and conducted a sentiment analysis of the comments, along with frequency analysis of the words used. Meanwhile, HaCohen Kerner et. al. (2020) carried out a sentiment analysis of Steam comments in Brazilian Portuguese using the *Steam API* (Application Programming Interface). In both cases, the combination of textual data scraped from social networks (Reddit) or videogame distribution service (Steam) and sentiment analysis enables the identification of the most positive aspects of video games, such as graphics, gameplay, soundtrack, and storytelling, and the most negative aspects, such as online

3 A more comprehensive overview of similar studies is provided by Yudintseva (2015).

4 Video games resembling the Souls video game franchise.

5 A similar package is also available for R (Fox et al. 2023).

gaming-related issues, DLCs, and various bugs, as reported by users themselves (HaCohen Kerner et. al. 2020: 401).

In addition to sentiment analysis, topic modelling is another potentially productive area of applying NLP methods to corpora based on video game reviews. Wang and Goh (2020) use text analysis methods to automatically generate topics from online user reviews and compare their effect on user satisfaction, identifying narrative and achievement as having the strongest correlation with satisfaction. Cho et al. (2020) compare the feasibility of qualitative human-based analysis of video game reviews and automated text mining analysis for identifying topics in video game reviews. While the human annotations showcase better understanding of the plot and narrative, the results provided by the machine-based methods show they can also be successfully used for identifying the main topics of games, especially when dealing with large databases.

4. Towards the Present Study—Research Questions, Methodology

As can be seen from the brief overview of previous research in Section 3, video games and video game reviews represent an interesting and a highly relevant object of linguistic inquiry. Thus, the main goal of this paper is to showcase how the contemporary methods for creating and analysing corpora using the specialized libraries in programming languages like R can be applied to this fast-evolving register. In order to build on the previous research and extend the analysis to other types of video game reviews, the main aims of this paper are: i) to identify words and phrases typical for the gaming genre and the individual subgenres; and ii) to test the applicability of web scraping and NLP methods for linguistic research. In order to achieve these aims, the paper will address the following research questions:

1) What words and phrases are typical for different genres of video game reviews?
2) Are typical words and phrases more accurately captured by frequency lists or keyword analysis?
3) Is there a correlation between the value obtained by Sentiment Analysis of video game reviews and the ratings provided by writers of reviews?

For this purpose, a corpus of video game reviews was collected from the GameSpot website (www.gamespot.com) using the *rvest* package (Wickham 2021) for web scraping in R.[6] The corpus was then analysed using the combination of traditional corpus linguistic (CL) methods, such as frequency lists, n-grams, and keywords and NLP methods, such as sentiment analysis, available in the *quanteda*

6 The method used in the paper was based on the tutorial on web scraping in R by John Little from Duke University, while a similar methodology is also described in Bradley and James (2019).

package (Benoit et al. 2018). Specifically, frequency lists for individual words and n-grams and keyword analysis were used to address RQ1 and RQ2, while the Augmented General Inquirer Positiv and Negativ dictionary in the quanteda.sentiment package was used to answer RQ3.

5. Results and Analysis

The data collected for the purpose of this paper were gathered from July 19 to 26, 2022 and include a total of 5243 reviews written by 213 different authors, distributed across three video game genre—Adventure, First Person Shooter (FPS), and Strategy. The oldest reviews were from May 1996 for games such as Star Trek: The Next Generation—A Final Unity and Crusader: No Remorse, while the most recent review was from July 2022 for the game Stray. For every review, five variables were automatically scraped from the web (title, text, author, date, rating), and two variables were manually added/derived (genre, sentiment). After the relevant data had been scraped from the web, they were cleaned to remove all the parts of the texts that were not part of the review (e.g., references to other games, galleries, warnings like "You need a javascript-enabled browser", etc.) and turned into a corpus using the corpus() function in the quanteda package. The corpus with the starting size of 6.68 million words was then tokenized, and all punctuation symbols and stopwords (function words like 'and', 'or', 'but', as well as the word 'game'[7]) were deleted from it, resulting in the final corpus size of 3.63 million words.

The most frequent words and n-grams from the resulting corpus were then extracted using the *dfm*[7] function, the results of which are shown in Table 1 and Table 2 below. As we can see from Table 1, raw frequency count is not particularly helpful in identifying the main topics of video game reviews at first glance, as a vast majority of the top 20 most frequent single words are gaming-nonspecific words (e.g., can, one, also, just) with sporadic lexical items relevant for the subcorpus in question (e.g., units and battle for the Strategy subcorpus). While this observation is particularly true for single words in Table 1, the 2-grams in Table 2 offer a bit more substance in terms of the identified review topics (e.g., sound effects, artificial intelligence, single-player campaign, turn-based strategy for Strategy, frame rate, team deathmatch, level design, multiplayer mode, capture flag for FPS), along with the names of certain video games (e.g., command conquer, rainbow six, tomb raider, metal gear). For the sake of brevity, only frequency lists for single words and 2-grams are provided here. The lists for 3-grams is provided in the Figure 3 in the Appendix, while an illustration of frequency lists for 4-grams can be seen in Figure 1.

7 Document feature matrix.

Figure 1. The most frequent 4-grams in the entire corpus

Table 1. The most frequent words in the corpus and the three subcorpora

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
can	39848	can	13394	can	8694	can	17760
one	20820	one	5521	one	4855	one	10444
like	19612	units	5419	like	4809	like	9564
also	15845	like	5239	also	3895	just	8035
just	15254	new	4568	just	3578	time	7997
time	15208	also	4566	even	3481	even	7423
even	15050	even	4146	get	3297	also	7384
get	13812	time	3999	time	3212	get	6985
new	13408	games	3692	new	3194	much	6329
much	12491	just	3641	though	2966	though	6085
though	11775	get	3530	enemies	2859	story	5834
make	11130	strategy	3463	play	2793	way	5818

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
games	10991	much	3409	much	2753	make	5731
way	10660	play	3303	weapons	2698	new	5646
play	10361	make	2948	multiplayer	2641	game's	5490
game's	10216	well	2757	make	2451	Characters	5235
well	10009	though	2724	way	2450	around	5232
good	9580	battle	2700	good	2433	games	5187
enemies	9129	good	2549	well	2405	enemies	4914
around	9113	enemy	2533	game's	2334	well	4847

Table 2. The most frequent 2-grams in the corpus and the three subcorpora

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
can also	1867	real-time strategy	1158	first-person shooter	775	resident evil	866
feel like	1426	strategy game	1070	call duty	580	can also	797
sound effects	1333	strategy games	692	can also	483	feel like	762
real-time strategy	1226	can also	586	single-player campaign	441	voice acting	730
voice acting	1159	sound effects	373	feel like	404	adventure game	728
frame rate	1110	original game	339	sound effects	350	frame rate	644
feels like	1056	game can	331	first-person shooters	338	sound effects	610
even though	1049	can get	304	rainbow six	329	feels like	599
can get	945	artificial intelligence	291	frame rate	327	tomb raider	554
can use	933	even though	278	world war	312	action game	535
first-person shooter	886	can take	278	team deathmatch	296	even though	534
resident evil	876	world war	270	pc version	296	playstation 2	519
takes place	803	feel like	260	xbox 360	289	splinter cell	505
xbox 360	797	can play	259	far cry	282	can use	497
single-player campaign	758	can use	248	feels like	281	metal gear	462
playstation 2	758	units can	248	2 s	276	star wars	457

All 3 genre		**Strategy**		**FPS**		**Adventure**	
Term	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*	*Term*	*Frequency*
pretty much	753	single-player campaign	234	level design	270	can get	433
can take	749	command conquer	232	multiplayer mode	251	xbox 360	431
pc version	743	turn-based strategy	226	capture flag	240	action adventure	431
strategy games	724	pretty much	214	artificial intelligence	238	first time	425

The next step in the analysis is to look at the keywords for every subcorpus and compare its successfulness in identifying relevant topics to those obtained by the frequency-based analysis. As one of the aims of the paper is to examine the usefulness of the different methods for identifying topics specific for a particular genre, the method which allows us to identify more terms specific for the genre (or subgenre) in question is in principle the one better suited for the task. Specifically, this means identifying terms which refer to various elements of the game like graphics, units, items, sounds, gameplay, etc.

The typical procedure for conducting the keyword analysis includes comparing a particular corpus (or subcorpus) of interest with a larger and more general reference corpus (Brezina 2018: 80). However, as the results in Fischer-Starcke (2009) indicate (cf. Section 2), it is also feasible to conduct this analysis by comparing the focus subcorpus against the whole corpus (i.e., conduct the whole analysis by using a single corpus). This option is utilized in the *quanteda_textstats* package, which uses the log-likelihood ratio as a measure of keyness. As we can see from Figures 2-4, the keyword analysis yields significantly better results as it identifies far more key terms (blue bars) specific for all three subcorpora. This is particularly true for the Strategy subcorpus where the top 10 keywords are the key concepts in the games in question (e.g., units, strategy, build), followed by names of various videogame franchises (e.g., worms, tycoon,[8] etc.). The same is true, although to lesser extent, for the remaining two subcorpora where the top 10 keywords also include genre-specific topics, such as shooter, deathmatch, and weapons for the FPS subcorpus, and puzzles, story, character for the Adventure subcorpus.[9]

8 The first one refers to the Worms game series (*Worms Battlegrounds, Worms Armageddon, Worms 3D*), the second to the Tycoon series (*RollerCoaster Tycoon, Zoo Tycoon, Railroad Tycoon*).

9 A more detailed list of keywords is available in Table 5 in the Appendix.

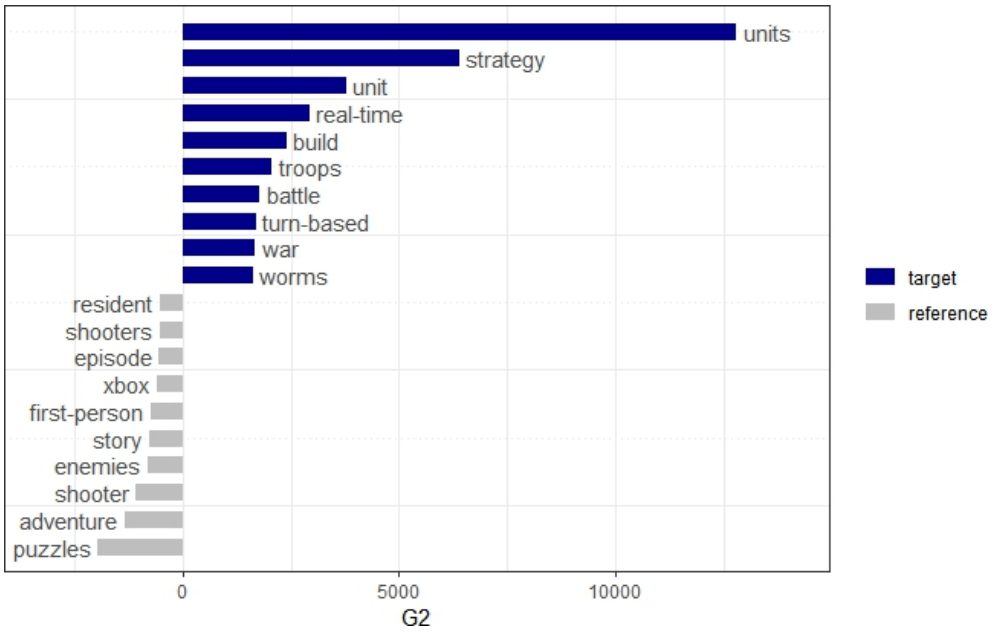

Figure 2. Keywords for the Strategy subcorpus

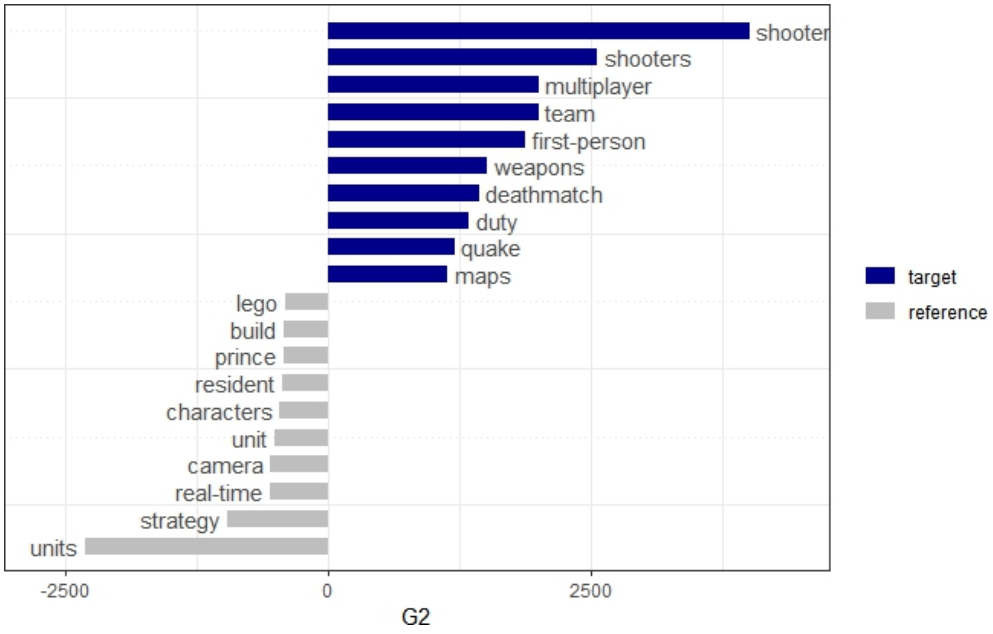

Figure 3. Keywords for the FPS subcorpus

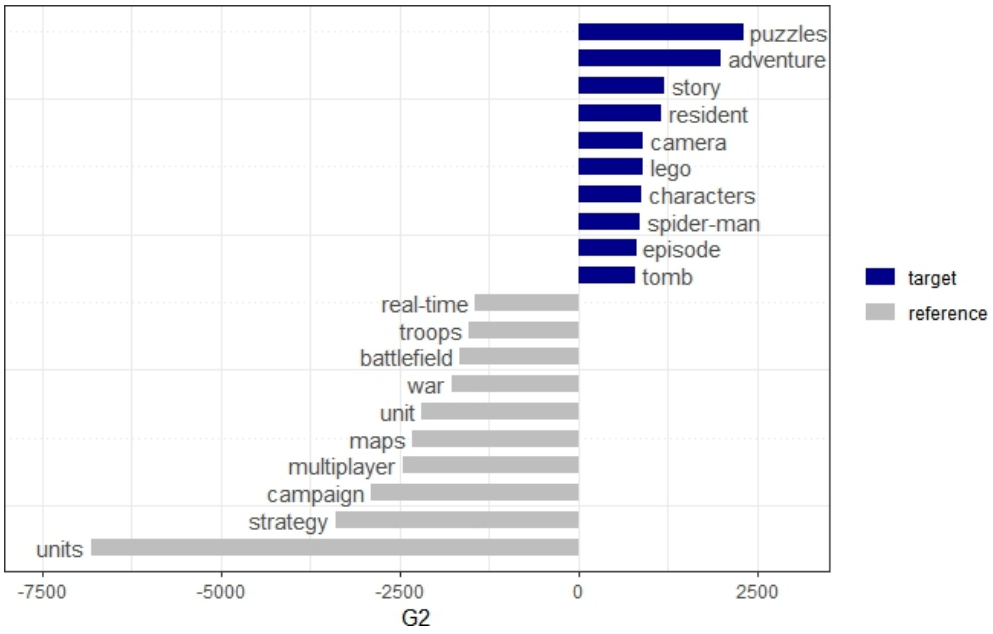

Figure 4. Keywords for the Adventure subcorpus

The last analysis, conducted to answer RQ3, is Sentiment Analysis (SA) for which the Augmented General Inquirer Positiv and Negativ dictionary in the *quanteda.sentiment* package was utilized. This SA tool is based on the presence of terms with either positive (1653 potential lexemes) or negative polarity (2010 lexemes) and assigns either a positive or a negative value. Polarity of the reviews in the corpus ranged from 1.946 (highest) to -0.765 (lowest). The obtained polarity values and ratings scraped from the website were averaged across author (N=213) and were tested for potential correlation. Pearson's correlation revealed a statistically significant (albeit relatively weak) positive correlation ($r= 0.23$, CI: 0.098 – 0.353, $p<.001$), as can be seen in Figure 5.[10] This shows that sentiment analysis can be used as a reliable tool for automatic identification and quantification of positive or negative polarity of a particular text.

10 An even stronger correlation was obtained using the *Lexicoder Sentiment Dictionary (2015)* in the same package, but is not reported here for the sake of brevity.

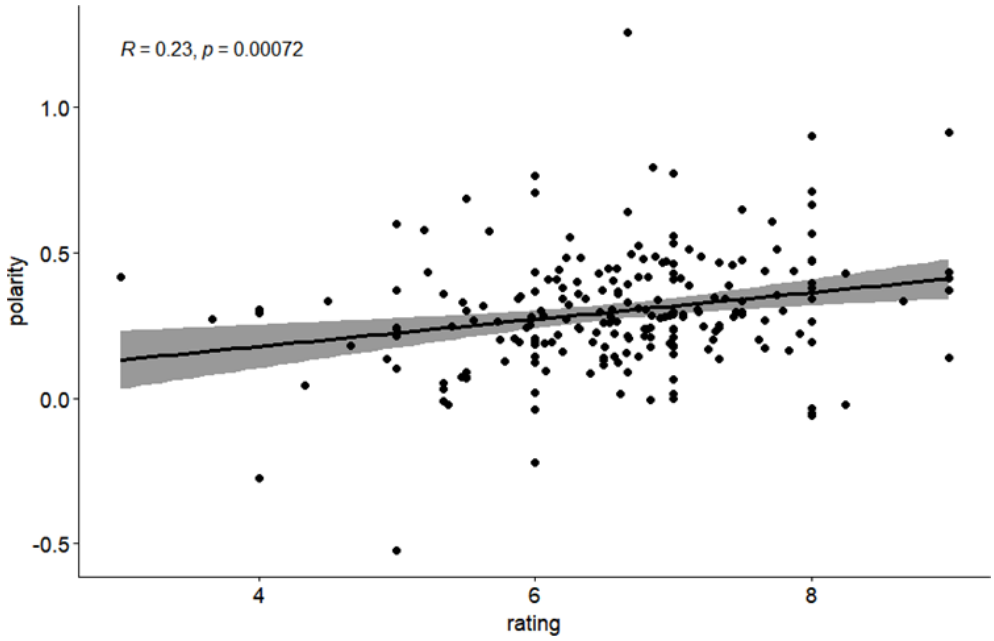

Figure 5. Correlation between polarity and ratings of the reviews

6. Conclusion

The main two aims of this paper were identifying words and phrases typical for the register of video game reviews and testing the applicability of web-scraping and NLP methods for linguistic research. As we can see from the results in Section 5, the conclusions are mixed—the most basic lists of the most frequent words and 2-grams provide only a handful of genre-specific units, while including a lot of noise in terms of general-purpose words (e.g., can, one, like in Table 1). That being said, the keyword analysis does much better since it provides a more substantial list of genre-specific lexical items, as we can see in Figures 2–4 and Table 5. Undoubtedly, this method is not without flaws, as there are again some items that do not serve the purpose of identifying the main aspects of video games, such as names of individual characters and videogame franchises. However, the benefits of such a method still outnumber the downsides shown here, which makes it quite a useful tool for identification of key topics in a particular text or register. Obviously, comparable if not even better results for identification of key topics could also be obtained by the topic modelling methodology using the *stm* package in R, but these were not employed here for the sake of brevity. Sentiment analysis, the last method used in this paper, proved to be a valuable tool as it relatively consistently recognized whether the review leaned more towards the positive or the negative end of the spectrum. The practical application of the sentiment analysis tools in linguistic research is manifold—for instance, they could be used to detect whether newspapers or politicians of different political orientation depict different social groups or minorities in more

positive or negative terms, or to examine whether there is a difference in positivity/negativity between descriptions of the same event by different individuals, that is, whether different groups perceive the same event in the same manner.

One might argue that the first two methods of the analysis used in this paper (frequency lists and keyword analysis) are available in most (if not all) software solutions for corpus management, such as WordSmith or Sketch Engine. However, what those solutions do not offer is the seamless integration of the module for collecting data (like the *rvest* package used here), the module for the basic data analysis, and the module for the more advanced NLP methods, which are all conveniently available in R. Another potential drawback of the methods used in the paper is the fact that the corpus was not lemmatized or POS-annotated. For instance, one can notice that the list of keywords for the Strategy and the FPS subcorpora both include singular and plural forms (unit and units, shooter and shooters, respectively). Integrating lemmatization and POS-tagging into the analysis using packages like *udpipe* or *spacyr* would have remedied these issues, but the idea behind this paper was to see how far one can get by using the most easily available tools, which is why this avenue was not pursued.

Ultimately, as Kilgarriff's (2001) claim about the importance of the internet for corpus linguistics remains uncontested, it is up to the linguists to embrace the new technological and analytical capabilities of the emerging tools to harness the power of the data available in online texts. Hopefully, this paper will motivate future use of tools such as those described here to incorporate new types of data into linguistic research.

References

Arik, Kaan. 2022. "Social Media Content Review of MMORPG Games: Reddit Comment Scraping and Sentiment Analysis". *Journal of Emerging Computer Technologies* 2(1). 13–21.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros. 2009. "The WaCky wide web: A collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43(3). 209–226. doi: https://doi.org/10.1007/s10579-009-9081-4

Baroni, Marco; Ueyama, Motoko. 2006. "Building general- and special-purpose corpora by Web crawling". *Proceedings of the NIJL International Workshop on Language Corpora*. Tokyo: NIJL.

Benoit, Kenneth; Watanabe, Kohei; Wang, Haiyan; Nulty, Paul; Obeng, Adam; Müller, Stefan; Matsuo, Akitaka. 2018. "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software* 3(30). 774. doi: https://doi.org/10.21105/joss.00774

Biber, Douglas; Reppen, Randi. 2015. *The Cambridge handbook of English corpus linguistics*. Cambridge University Press.

Bradley, Alex; James, Richard J. E. 2019. "Web Scraping Using R". *Advances in Methods and Practices in Psychological Science* 2(3). 264–270. doi: https://doi.org/10.1177/2515245919859535

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press. doi: https://doi.org/10.1017/9781316410899

HaCohen Kerner, Yaakov; Miller, Daniel; Yigal, Yair. 2020. "The influence of preprocessing on text classification using a bag-of-words representation". *SBC { Proceedings of SBGames*. https://dx.plos.org/10.1371/journal.pone.0232525

Burmester, Michael; Gerhard, Daniela; Thissen, Frank (Eds.). 2006. *Digital game-based learning: Proceedings of the 4th International Symposium for Information Design*. Stuttgart Media University, Universitätsverlag.

Camacho Vásquez, Gonzalo; Ovalle, Joan Camillo. 2019. "The Influence of Video Games on Vocabulary Acquisition in a Group of Students from the BA in English Teaching". *Gist Education and Learning Research Journal* 19. 172–192.

Chen, Howard; Yang, Ting-Yu Christine. 2013. "The impact of adventure video games on foreign language learning and the perceptions of learners". *Interactive Learning Environments* 21(2). 129–141. doi: https://doi.org/10.1080/10494820.2012.705851

DeHaan, Jonathan W. 2005. "Acquisition of Japanese as a Foreign Language Through a Baseball Video Game". *Foreign Language Annals* 38(2). 278–282. doi: https://doi.org/10.1111/j.1944-9720.2005.tb02492.x

Diouf, Rabiyatou; Sarr, Edouard Ngor; Sall, Ousmane; Birregah, Babiga; Bousso, Mamadou; Mbaye, Sény Ndiaye. 2019. "Web Scraping: State-of-the-Art and Areas of Application". *IEEE International Conference on Big Data (Big Data)*. 6040–6042. doi: https://doi.org/10.1109/BigData47090.2019.9005594

Dunst, Alexander; Hartel, Rita; Laubrock, Jochen. 2017. "The Graphic Narrative Corpus (GNC): Design, Annotation, and Analysis for the Digital Humanities". *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 15–20. doi: https://doi.org/10.1109/ICDAR.2017.286

Fischer-Starcke, Bettina. 2009. "Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistic analysis". *International Journal of Corpus Linguistics* 14(4). 492–523. doi: https://doi.org/10.1075/ijcl.14.4.03fis

Fox, Nathan; Van Berkel, Derek; Verge, Ramiro Serrano; Lindquist, Mark. 2023. "vGameReviews: An R package for harnessing video game reviews for scientific research". *SoftwareX* 23, 101423. doi: https://doi.org/10.1016/j.softx.2023.101423

Gatto, Maristella. 2011. "The 'body' and the 'web': The web as corpus ten years on". *ICAME Journal* 35. 35–88.

Gatto, Maristella. 2014. *The web as corpus: Theory and practice*. Bloomsbury Publishing.

Gee, James Paul. 2004. *What video games have to teach us about learning and literacy*. Palgrave Macmillan.

Gee, James Paul. 2013. *Good video games and good learning: Collected essays on video*

games, learning and literacy. Peter Lang.

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

Heritage, Frazer. 2022a. "Magical women: Representations of female characters in the Witcher video game series". *Discourse, Context & Media* 49, 100627. doi: https://doi.org/10.1016/j.dcm.2022.100627

Heritage, Frazer. 2022b. "Politics, pronouns and the players: Examining how videogame players react to the inclusion of a transgender character in World of Warcraft". *Gender and Language* 16(1). doi: https://doi.org/10.1558/genl.20250

Kasemap, K. 2017. "The Fundamentals of Game-Based Learning". In Kidd, T. & Morris, L. R. Jr. (Eds.) *Handbook of Research on Instructional Systems and Educational Technology*. IGI Global. 174–185. doi: https://doi.org/10.4018/978-1-5225-2399-4

Kilgarriff, Adam. (2001). *Web as corpus*. In Rayson, P., Wilson, A., McEnery, T., Hardie, A. & Khoja, S. (Eds.) *Proceedings of the Corpus Linguistics*. 342–344.

Kreyer, Rolf; Mukherjee, Joybrato. 2007. "The Style of Pop Song Lyrics: A Corpus-linguistic Pilot Study". *Anglia - Zeitschrift Für Englische Philologie* 125(1). doi: https://doi.org/10.1515/ANGL.2007.31

Kumar, Sumit; Roy, Uponika Barman. 2023. "A technique of data collection". In Goswami, T. & Sinha, G. R. (Eds.) *Statistical Modelling in Machine Learning*. Elsevier. 23–36. doi: https://doi.org/10.1016/B978-0-323-91776-6.00011-7

Merullo, Jack; Yeh, Luke; Handler, Abram; Grissom Ii, Alvin; O'Connor, Brendan; Iyyer, Mohit. 2019. "Investigating Sports Commentator Bias within a Large Corpus of American Football Broadcasts". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP). 6354–6360. doi: https://doi.org/10.18653/v1/D19-1666

Motschenbacher, Heiko. 2016. "A corpus linguistic study of the situatedness of English pop song lyrics". *Corpora* 11(1). 1–28. doi: https://doi.org/10.3366/cor.2016.0083

Moustafa, Basant S. M. (2022). "A comparative corpus stylistic analysis of thematization and characterization in Gordimer's My Son's Story and Coetzee's Disgrace". *Open Linguistics* 8(1). 46–64. doi: https://doi.org/10.1515/opli-2020-0183

Santos, Antonio. 2017. "Instructional Strategies for Game-Based Learning". In Kidd, T. & Morris, L. R. Jr. (Eds.) *Handbook of Research on Instructional Systems and Educational Technology*. IGI Global. 164–173. doi: https://doi.org/10.4018/978-1-5225-2399-4

Suchomel, Vít. 2020. *Better Web Corpora for Corpus Linguistics And NLP*. Doctoral Thesis. Masaryk University. Brno.

Sylvén, Liss Kerstin; Sundqvist, Pia. 2012. "Gaming as extramural English L2 learning and L2 proficiency among young learners". *ReCALL* 24(3). 302–321. doi: https://doi.org/10.1017/S095834401200016X

Unser-Schutz, Gianclarla. 2011. "Developing a text-based corpus of the language of Japanese comics (manga)". In Newman, J., Baayen, H. & Rice, S. (Eds.) *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. BRILL. 213–238. doi: https://doi.org/10.1163/9789401206884

Wang, Xiaohui; Goh, Dion Hoe Lian. 2020. "Components of game experience: An automatic text analysis of online reviews". *Entertainment Computing* 33. 100338. doi: https://doi.org/10.1016/j.entcom.2019.100338

Werner, Valentin. (2012). "Love is all around: A corpus-based study of pop lyrics". *Corpora* 7(1). 19–50. doi: https://doi.org/10.3366/cor.2012.0016

Wickham, Hadley. 2021. *rvest: Easily Harvest (Scrape) Web Pages*. R Package Version 1.0.2.

Yudintseva, Anastassiya. 2015. "Game-Enhanced Second Language Vocabulary Acquisition Strategies: A Systematic Review". *Open Journal of Social Sciences* 3(10). 101–109. doi: https://doi.org/10.4236/jss.2015.310015

Appendix

Table 3. The most frequent 3-grams in the corpus and the three subcorpora

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
world war ii	403	real-time strategy game	444	world war ii	191	metal gear solid	360
metal gear solid	372	real-time strategy games	352	deathmatch team deathmatch	176	grand theft auto	274
real-time strategy games	369	world war ii	187	rainbow six 3	94	lego star wars	181
grand theft auto	299	heroes might magic	99	deathmatch capture flag	91	devil may cry	179
xbox 360 version	227	turn-based strategy game	83	far cry 3	90	action adventure game	174
deathmatch team deathmatch	202	full spectrum warrior	82	team deathmatch capture	82	resident evil 4	142
lego star wars	181	final fantasy tactics	76	left 4 dead	75	xbox 360 version	132
devil may cry	180	romance three kingdoms	72	xbox 360 version	72	gear solid 2	131
playstation 2 version	153	age empires ii	68	call duty 3	72	right analog stick	108
make feel like	147	turn-based strategy games	56	rainbow six vegas	70	blood omen 2	108
resident evil 4	143	red alert 2	54	red faction ii	66	game takes place	101
spend lot time	136	railroad tycoon ii	43	bad company 2	63	playstation 2 version	98

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
right analog stick	134	game boy advance	41	call duty 4	61	game boy advance	92
gear solid 2	131	jagged alliance 2	40	quake iii arena	59	new york city	91
new york city	123	steep learning curve	39	far cry 2	57	star wars ii	80
spend much time	119	game takes place	37	duke nukem 3d	53	spend lot time	79
point point b	114	traditional real-time strategy	37	enemy artificial intelligence	49	resident evil 2	75
can also use	110	red alert 3	37	make feel like	46	third-person action game	72
enemy artificial intelligence	110	real-time strategy genre	35	call duty 2	46	spend much time	67
blood omen 2	108	sid meier's pirates	34	big red one	43	tomb raider legend	65

Table 4. The most frequent 4-grams in the corpus and the three subcorpora

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
metal gear solid 2	131	many real-time strategy games	25	team deathmatch capture flag	73	metal gear solid 2	131
lego star wars ii	80	starfleet command volume ii	21	deathmatch team deathmatch capture	71	lego star wars ii	80
team deathmatch capture flag	76	heroes might magic iii	20	call duty black ops	39	tom clancy's splinter cell	50

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
deathmatch team deathmatch capture	74	real-time strategy game set	16	standard deathmatch team deathmatch	35	grand theft auto iii	50
xbox 360 playstation 3	72	command conquer red alert	15	battlefield bad company 2	27	xbox 360 playstation 3	45
tom clancy's splinter cell	58	final fantasy tactics advance	15	far cry 2 s	27	devil may cry 3	42
grand theft auto iii	54	robin hood defender crown	15	medal honor allied assault	26	devil may cry 4	37
goes long way toward	47	war ii real-time strategy	14	call duty 4 modern	24	prince persia sands time	36
devil may cry 3	42	3d real-time strategy games	13	duty 4 modern warfare	24	metal gear solid 3	33
call duty black ops	39	typical real-time strategy game	12	xbox 360 playstation 3	24	resident evil code veronica	31
devil may cry 4	37	world war ii real-time	12	submachine guns assault rifles	23	goes long way toward	30
standard deathmatch team deathmatch	37	traditional real-time strategy games	12	ghost recon advanced warfighter	23	tomb raider angel darkness	30
prince persia sands time	36	3d real-time strategy game	12	tom clancy's rainbow six	22	grand theft auto games	29
metal gear solid 3	33	might magic heroes vi	11	left 4 dead 2	21	grand theft auto series	28
get point point b	31	red alert 3 s	11	far cry 3 s	19	harry potter sorcerer's stone	26

All games		**Strategy**		**FPS**		**Adventure**	
Term	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*	*Term*	*Freq.*
resident evil code veronica	31	harvest moon friends mineral	11	world war ii combat	19	like metal gear solid	24
playstation 3 xbox 360	30	moon friends mineral town	11	call duty world war	19	lego harry potter years	23
long way toward making	30	traditional real-time strategy game	10	call duty modern warfare	17	60 frames per second	22
grand theft auto games	30	age empires ii age	10	modes deathmatch team deathmatch	17	grand theft auto iv	22
tomb raider angel darkness	30	heroes might magic series	10	tom clancy's ghost recon	17	xbox playstation 2 versions	22

Table 5. Top 50 keywords for all three subcorpora

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
units	12791.63	5419	157	shooter	4014.768	2117	458	puzzles	2306.533	3489	603
strategy	6382.148	3463	486	shooters	2568.759	1154	137	adventure	1994.823	3264	633
unit	3780.281	1877	180	multiplayer	2008.116	2641	2397	story	1204.104	5834	2738
real-time	2933.624	1523	179	team	2004.8	1819	1115	resident	1168.203	946	19
build	2407.141	1769	553	first-person	1886.801	1468	725	camera	913.6884	2243	678
troops	2072.831	1215	220	weapons	1523.494	2698	3079	lego	903.9804	821	36
battle	1778.051	2700	2106	deathmatch	1441.937	795	195	characters	890.7015	5235	2678
turn-based	1688.705	759	39	duty	1346.864	624	85	spider-man	861.5379	644	4
war	1680.435	2182	1479	quake	1206.947	471	22	episode	814.7044	1046	135

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
worms	1626.313	690	20	maps	1145.582	1501	1357	tomb	799.1362	687	22
battles	1562.245	2066	1425	modes	1049.383	1164	898	film	777.4658	1123	180
rts	1493.901	625	15	single-player	1048.418	1346	1195	prince	756.8535	747	47
resources	1405.063	1000	293	halo	1005.054	484	77	raider	690.3584	558	11
building	1333.066	1392	745	teammates	975.9379	511	108	evil	667.7697	1594	470
ships	1273.965	873	236	rifle	919.7942	579	195	batman	620.9776	525	15
map	1270.758	2114	1773	online	861.2664	1283	1290	lara	620.0901	492	8
strategic	1220.25	828	219	half-life	843.5262	320	11	solve	588.5413	909	162
armies	1201.858	612	66	sniper	843.4046	568	219	splinter	550.6774	572	43
command	1170.965	972	377	doom	789.4749	487	157	harry	535.476	453	13
tycoon	1130.417	451	4	rainbow	771.0926	379	65	boss	514.4711	1325	417
interface	1125.971	1059	495	call	768.5167	890	717	moves	506.795	1507	530
heroes	1117.458	905	337	gun	720.2128	1019	983	zelda	505.9009	388	4
empire	1081.608	750	208	guns	695.3923	880	771	movie	479.4775	1239	391
scenarios	1076.796	902	355	shooting	693.5467	1007	992	button	478.4417	1955	840
resource	1067.193	597	93	ops	682.4686	376	92	puzzle	463.5554	1124	336
campaign	1016.328	2460	2641	unreal	679.5289	305	36	platforming	456.2773	531	55
campaigns	990.7091	676	181	players	677.9108	2016	3072	sequences	451.063	1459	544
factions	935.3527	643	175	f.e.a.r	661.134	234	2	creed	447.2823	379	11
army	911.6773	880	426	weapon	647.8509	1245	1498	character	446.0561	3738	2178
civilization	909.6756	520	87	battlefield	634.895	849	781	assassin's	438.9373	408	20
empires	895.2129	353	2	assault	616.9774	635	452	stealth	427.3437	1211	411
tactical	825.9579	961	582	levels	591.976	1939	3084	horror	411.4611	668	128

Strategy				**FPS**				**Adventure**			
Keyword	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*	*Keyword*	*G2*	*n_target*	*n_reference*
conquer	809.142	401	38	turok	578.3071	211	4	cell	392.9027	611	110
can	794.9819	13394	26454	crysis	568.2212	210	5	fisher	380.6659	321	9
buildings	782.1325	980	641	grenades	560.7733	499	298	potter	362.2752	284	4
skirmish	778.7961	424	60	medal	557.9524	261	37	kain	341.3627	243	0
terrain	717.5223	617	253	campaign	549.3755	1933	3168	crime	340.234	544	102
infantry	714.1497	571	208	bots	539.8228	291	67	objects	337.1887	1319	553
age	703.5445	703	356	duke	531.9984	274	55	combos	336.0768	398	43
cities	702.2771	565	208	firefights	518.7839	270	56	solving	324.3613	544	109
victory	685.059	576	228	mode	504.4413	1994	3421	ninja	316.6469	452	71
research	683.68	527	180	3	503.184	980	1189	dialogue	299.1989	1395	642
tactics	682.8009	705	372	cover	497.4774	838	923	arkham	299.0238	221	1
expansion	624.9209	655	352	shoot	489.3064	836	930	legend	293.7579	429	70
structures	624.7087	578	264	counter-strike	484.9129	176	3	sword	288.2717	708	214
wargame	623.3512	248	2	vehicles	474.9474	785	853	hitman	281.9814	259	12
historical	614.7302	434	125	grenade	470.9992	365	179	kong	279.9012	275	17
forces	601.1434	972	797	soldier	467.043	457	307	dead	273.3179	1307	609
scenario	591.8282	530	231	pc	461.3926	1227	1772	persia	272.1509	244	10
management	568.418	428	140	wolfenstein	455.2862	163	2	detective	271.3996	346	44