



The Game Walkthrough Corpus (GWTC) – A Resource for the Analysis of Textual Game Descriptions

MANUEL BURGHARDT 

JOCHEN TIEPMAR

**Author affiliations can be found in the back matter of this article*

DATA PAPER

]u[ubiquity press

ABSTRACT

We present the Game Walkthrough Corpus (GWTC), which contains 12,295 unique walkthrough documents covering 6,117 games. For each game walkthrough, we provide frequencies of unigrams and bigrams, treating the walkthrough document as a Bag of Words. In addition, we provide word frequencies at the sentence level. Furthermore, the GWTC contains a number of game-related metadata, including title, publisher, developer, year, and genre. All the language statistics and metadata are stored in separate plain text files and can be referenced through uniform resource names (URN). These URNs can also be used to derive any combination of statistics and metadata. Researchers, for instance, can investigate the most frequent unigrams for games in the “Adventure” genre. This way, the GWTC can be reused for different kinds of research questions on gaming language.

CORRESPONDING AUTHOR:

Manuel Burghardt

Computational Humanities
Group, Leipzig University,
Germany

burghardt@informatik.uni-leipzig.de

KEYWORDS:

game walkthroughs; gaming
language; game philology

TO CITE THIS ARTICLE:

Burghardt, M., & Tiepmar, J.
(2021). The Game
Walkthrough Corpus (GWTC)
– A Resource for the Analysis
of Textual Game Descriptions.
*Journal of Open Humanities
Data*, 7: 14, pp. 1–7. DOI:
<https://doi.org/10.5334/johd.34>

(1) OVERVIEW

REPOSITORY LOCATION

DOI: [10.5281/zenodo.4562336](https://doi.org/10.5281/zenodo.4562336)

URL: <https://doi.org/10.5281/zenodo.4562336>

CONTEXT

The academic interest in studying games as cultural phenomena in their own right has reached a certain level of maturity. This maturity is reflected, among other things, in the existence of a number of dedicated organizations, such as DiGRA (Digital Games Research Association), and a large number of relevant publications in specific journals, such as *Game Studies*. The study of games has attracted a wide range of subject areas, including media studies, cultural studies, psychology, computer science, and many more (see Ensslin, 2012). We argue that Digital Humanities (DH) can offer yet another perspective to studying games. In DH, Moretti's (2000) notion of *distant reading* has become a central concept and metaphor for all kinds of computational and empirical approaches to text analysis. In the case of video games, however, the question arises as to how such a highly interactive and dynamic medium can be formalized and modeled in a way that allows it to be analyzed computationally. After all, game experiences are highly individual and explicitly quantifiable features are hardly available. Thus, to enable a quantitative research perspective on games, we propose focusing on their textual dimension: their language. According to Ensslin (2012, p.6), the language of gaming branches into two major aspects: "the ways in which videogames and their makers convey meanings to their audience, and the ways in which gamers and other stakeholders communicate and negotiate meanings between themselves". To study the various levels of discourse evident in the language of gaming, Ensslin created a small-scale corpus of different texts about videogames. Her *GameCorp* comprises 184 texts from videogame magazines, gamer fora and chats, and transcribed live conversations during gameplay.

In this data paper, we present a large corpus of game walkthroughs, which are textual guides for all kinds of video games that include instructions and tips that *walk* players *through* a game, so that they can complete it successfully. Walkthroughs have a specific type of gaming language that might be categorized as a mixture of languages. First, we find language as it appears in the actual games, in terms of game mechanics as well as references to the plot and its characters. Second, walkthroughs contain gaming jargon used by actual gamers, as walkthroughs are mostly written by players of the actual game (see also Krause, 2016). Walkthroughs have been shown to be a suitable document type for purposes of digital game preservation (Newman, 2011; Nylund, 2015). Newman (2011, p.111) summarizes the strengths of game walkthroughs in the following way:

"[...] player-produced walkthroughs [...] are some of the most comprehensive investigations of digital gameplay that presently exist; certainly more thorough, investigative and inventive than any professional or academic literature; [...] walkthrough texts might be better able to capture and communicate the important qualities of games, as defined and understood by their players, than the playable games themselves".

While walkthroughs have been used successfully to support the study of specific games, such as *Zelda 64* (Consalvo, 2013), they are not equally suited for all game genres. Games that involve a certain degree of creativity or that may not have a specified winning goal, such as *Minecraft*, are examples of the insufficiency of walkthrough descriptions. At the same time, complex, open gaming experiences, like *Grand Theft Auto* or current *Assassins Creed* games, and *Grand Strategy* games, like the *Europa Universalis* or *Total War* franchises, may also be rather limited in their walkthrough descriptions compared to the actual gameplay. Despite these limitations for certain genres, we agree with Newman (2011) and Nylund (2015) and believe that, in general, walkthroughs are a great source of textual game preservations that enable large-scale corpus analyses. From a quantitative perspective, walkthroughs are particularly interesting types of gaming text, as they are widely available on the Internet. Interestingly, besides one example from linguistics, where a custom German-language walkthrough corpus has been used to study imperative language (Krause, 2016), hardly any existing studies so far have utilized this type of text to study games in a quantitative fashion. With the Game Walkthrough Corpus (GWTC), we hope to promote more research in this direction in the future.

This section summarizes the main steps that were involved in creating the GWTC. It also provides an overview of the main contents of the dataset.

STEPS

- i) *Data sources and languages* – The GWTC is designed to be continuously expanded. The ultimate goal is a multilingual corpus of many different game walkthroughs. Although this first release of the GWTC focuses on English-language walkthroughs, a multilingual perspective has already been considered in the data structure by including some German-language walkthroughs. For the current corpus of 12,295 unique walkthrough documents for a total of 6,117 games, we collected English and German language walkthroughs from the following platforms:
 - *Neoseeker*: 8,729 documents were collected from Neoseeker (<https://www.neoseeker.com/>). The platform includes mostly mainstream titles like the Grand Theft Auto or Assassins Creed series, but also has a small selection of niche titles such as “Hitomi – My Stepsister”. The language of the documents is English.
 - *Jayisgames*: 2,220 documents were collected from Jayisgames (<https://jayisgames.com/>). This platform is focused on puzzle games. The language of the documents is English.
 - *Gamesetter*: 799 walkthroughs were collected from Gamesetter (<http://gamesetter.com/>). This platform also has a focus on puzzle games. The language of the documents is German.
 - *Portforward*: 318 documents were collected from Portforward (<https://portforward.com/games/walkthroughs/>). This platform is actually focused on helping with computer network-related problems but also has a side project that provides walkthrough documents for a number of popular games. The language of the documents is English.
 - *Spieletipps*: 229 documents were collected from Spieletipps (<https://www.spieletipps.de/>). The games on this platform can be considered mainstream content. The language of the documents is German.
- ii) *Data processing* – All HTML walkthrough documents were collected from the different platforms using individually implemented Scrapy (<https://scrapy.org/>) crawlers. The text content was extracted and converted into a generic uniform hierarchical TEI/XML markup. In a pre-processing step, we converted each text to lowercase and removed every character that is not a regex word or space character (`[^\w\s]`). For the sentence collocations, all punctuation was normalized to full stops, meaning that, for example, subordinate clauses are treated as sentences to break up larger sentences. All characters were further filtered, based on a lowercase whitelist of English and German letters and full stop to avoid encoding problems. We did not remove any stop words and also did not perform any lemmatization or stemming. As for potential paratextual elements (e.g., introductions, general information on the game, etc.), we kept all of those in the walkthrough documents and only removed HTML-related structural elements (e.g., navigation headers) from the documents. The normalized TEI/XML files were then used to build a Canonical Text Service (CTS), which is typically used as a citation framework in classical studies (Smith, 2009; Tiepmar, Teichmann, Heyer, Berti, & Crane, 2014). The purpose of the CTS is to have persistent URNs for each game and its structural text elements. The following is an example URN for the game “Zak McKracken and the Alien Mindbenders”:
 - urn:cts:gwtc:zak_mckracken_and_the_alien_mindbenders:
- iii) *Metadata* – Next, we added various metadata to the walkthrough documents, which we gathered from RAWG (<https://rawg.io/>) and Steam (<https://store.steampowered.com/>). For both platforms, it can be assumed that most of the metadata is subject to systematic editing. All metadata was collected using Python API packages for Steam (<https://pypi.org/project/steamfront/>) and RAWG (<https://rawgpy.readthedocs.io/>). A slight bias toward PC games is

expected, as Steam (unlike RAWG) does not include console games. While this should not impact multi-platform titles that were also published on PC, metadata for console-only games may be underrepresented.

The following metadata are available for the games, with a varying degree of coverage¹:

- game title²
- short description (booklet text)
- gameplay tags
- game genres
- publishers
- developers
- supported platforms
- supported game languages / localizations
- release dates

Aside from the game titles and short descriptions, metadata often provide multiple entries per game. For example, the gameplay tags for *Max Payne 2* are *singleplayer*, *destruction*, *drama*, *physics*, *romance*, *story*, *character*, *police* and *fall*. Gameplay tags seem to cover a wide range of topics, from platform-specific information to ludic and narrative categorization that may overlap with game genres. While it may seem strange to have multiple release dates, developers, and publishers, this is to be expected because of repeated publications and later ports to additional game platforms that are often realized by different teams.

iv) *Text statistics* – As individual copyrights protect game walkthroughs, this data set does not include the full-text documents. It rather provides various data formats that are useful for text mining and distant reading approaches while not allowing for the reconstruction of the full texts. To enable researchers to look up the original full text of specific walkthroughs, we provide the source URLs for any walkthrough document as part of the dataset. The following frequency information is available in the GWTC:

- **bagofwords**: unigram frequencies per document
- **bigrams**: bigram frequencies per document
- **sentencecollocations**: word frequencies per sentence per document
- **textlength**: number of characters per document
- **tfidf_deu**: word significance per document (German)
- **tfidf_eng**: word significance per document (English)
- **tokencount**: number of unique words per document
- **typecount**: number of words per document

(3) DATASET DESCRIPTION

OBJECT NAME

game-walkthrough-corpus.zip

The zip file contains three subfolders: data, metadata and documentation (for more details see <https://zenodo.org/record/4562336>)

¹ For an overview of the current coverage of metadata see: http://www.informatik.uni-leipzig.de/~jtiepmar/forschung/gwtc/table?data=_all&compact=true. Adding more metadata from other sources is a desideratum for future releases of the dataset.

² Note: Some videogames are published multiple times, as they may be the subject of patches or updates. Games can even be completely remade and published under the same name (e.g., “Tomb Raider”) or differ from platform to platform, because of technical differences (e.g., the PlayStation and Nintendo Wii edition of “Resident Evil 4”). As a mapping of such parallel releases cannot easily be achieved automatically, we kept the original titles as they were provided by the authors.

FORMAT NAMES AND VERSIONS

Formats: CSV, TXT

Note that there are two versions of the GWTC available for download: ver. 0.99 contains all the corpus files, plus the Git files. After downloading ver. 0.99, the Git folders may be hidden per default, depending on your operating system. Ver. 1.0 (recommended) is a cleaned-up version that comes without the Git files.

CREATION DATES

from 2020-02-12 to 2021-02-28

DATASET CREATORS

- *Jochen Tiepmar* (Computational Humanities Group, Leipzig University): Conceptualization, Data Curation, Formal Analysis, Methodology
- *Paul Starke* (Department of Computer Science, Leipzig University): Data Curation, Formal Analysis
- *Tim Karwasz* (Department of Computer Science, Leipzig University): Data Curation, Formal Analysis

LANGUAGE

English, German

LICENSE

CC BY 4.0

REPOSITORY NAME

Game Walkthrough Corpus (GWTC)

PUBLICATION DATE

2021-03-01

4 REUSE POTENTIAL

As shown in the introductory section, our proposed walkthrough corpus is intended as a textual abstraction of video games (see also Newman, 2011) that enables scholars from media studies, cultural studies, linguistics, and DH to perform large-scale game studies using the methodology of distant reading. Distant reading typically looks for overarching patterns and contexts and is primarily interested in the interplay of textual phenomena with larger macro-categories, such as genre, time, and place. In a similar vein, studies with the GWTC allow for investigating gaming language and various metadata. Some possible research questions might include the following:

- **Game mechanics:** How do game mechanics in Jump & Run Games develop over time? How similar are these mechanics when compared to other game genres?
- **Plot elements:** What are typical plot elements of Role-Playing Games (RPGs), and how do they differ with regard to different publishers?
- **Game characters:** What is the gender distribution of characters appearing in games from a particular genre, and how does the distribution develop over time? What is the sentiment toward different characters – are they viewed as positive or negative?

Furthermore, simple analyses of the extensive metadata can be very revealing in themselves. For example, the correlation of publisher and release date clearly shows that, starting in the early 1980s, large publishing houses clearly dominated the field until the 2000s. The number of releases increases continuously during this period. From the early 2000s onwards, however, releases from independent publishers increasingly appear on the scene and from then on also account for a considerable share of the total number of releases (see *Figure 1*).

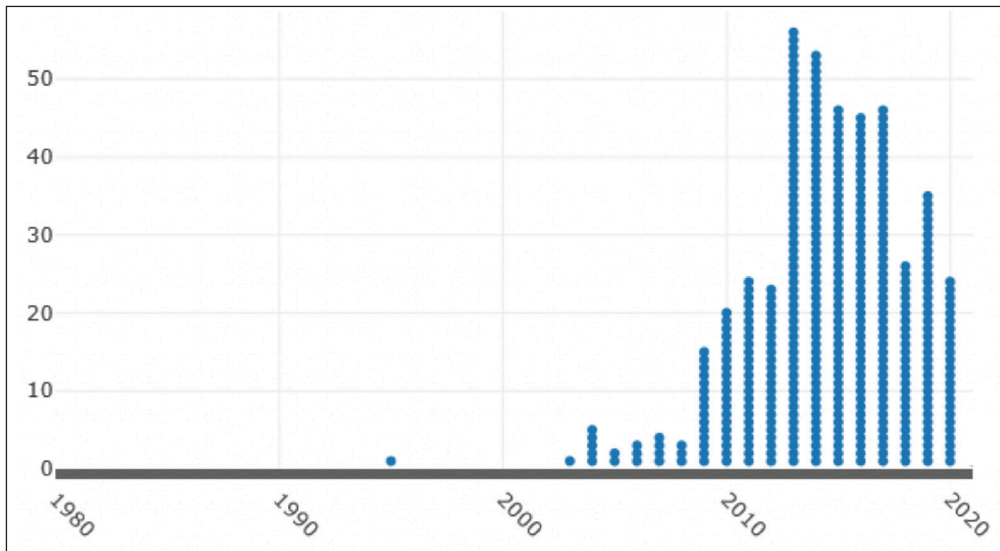


Figure 1 Overview of releases from indie publishers.

As we are only allowed to share frequency information on the document and sentence level, the analytical methods are limited to approaches that do not rely on contextual information (for instance, stylometry or keyword analyses). However, for researchers who plan to use more advanced methods, such as LDA topic modeling or Word2Vec embeddings, GWTC can still be used as a means to compose a subcorpus (according to genre, publishers, etc.), which then can be used to rebuild a full-text corpus using the available source URLs. Either way, we think walkthroughs are an excellent source for exploring games from a textual/discursive perspective, allowing for quantitative approaches that would otherwise not be possible.

COMPETING INTERESTS


The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Manuel Burghardt: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Jochen Tiepmar: Conceptualization, Data Curation, Formal Analysis, Methodology.

AUTHOR AFFILIATIONS

Manuel Burghardt  orcid.org/0000-0003-1354-9089
Computational Humanities Group, Leipzig University, Germany

Jochen Tiepmar
Computational Humanities Group, Leipzig University, Germany

REFERENCES

- Consalvo, M.** (2003). *Zelda 64 and video game fans: A walkthrough of games, intertextuality, and narrative*. *Television & New Media*, 4(3), 321–334. DOI: <https://doi.org/10.1177/1527476403253993>
- Ensslin, A.** (2012). *The language of gaming*. Basingstoke: Palgrave Macmillan. DOI: <https://doi.org/10.1007/978-0-230-35708-2>
- Krause, A.** (2016). The challenges and joys of analysing ongoing language change in web-based corpora: A case study. *Proceedings of the 10th Web as Corpus Workshop*, 27–34. DOI: <https://doi.org/10.18653/v1/W16-2604>
- Moretti, F.** (2000). Conjectures on world literature. *New left review*, 1(54), 54–68.
- Newman, J.** (2011). (Not) playing games: player-produced walkthroughs as archival documents of digital gameplay. *International Journal of Digital Curation*, 6(2), 109–127. DOI: <https://doi.org/10.2218/ijdc.v6i2.206>
- Nylund, N.** (2015). Walkthrough and let's play: Evaluating preservation methods for digital games. *Proceedings of the 19th International Academic Mindtrek Conference*, 55–62. DOI: <https://doi.org/10.1145/2818187.2818283>

Smith, D. N. (2009). Citation in classical studies. *Digital Humanities Quarterly*, 3(1).

Tiepmar, J., Teichmann, C., Heyer, G., Berti, M., & Crane, G. (2014). A new implementation for canonical text services. *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 1–8. DOI: <https://doi.org/10.3115/v1/W14-0601>

Burghardt and Tiepmar
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.34

7

TO CITE THIS ARTICLE:

Burghardt, M., & Tiepmar, J. (2021). The Game Walkthrough Corpus (GWTC) – A Resource for the Analysis of Textual Game Descriptions. *Journal of Open Humanities Data*, 7: 14, pp. 1–7. DOI: <https://doi.org/10.5334/johd.34>

Published: 14 July 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.