

# The VGLC: The Video Game Level Corpus

**Adam James Summerville, Michael Mateas**

University of California, Santa Cruz  
1156 High Street  
Santa Cruz, CA 95066  
(+1) 831-459-0111,  
asummerv@ucsc.edu, michaelm@soe.ucsc.edu

**Sam Snodgrass, Santiago Ontañón**

Drexel University  
3141 Chestnut Street  
Philadelphia, PA 19104  
(+1) 215-571-4109  
sps74@drexel.edu, santi@cs.drexel.edu

## ABSTRACT

Levels are a key component of many different video games, and a large body of work has been produced on how to procedurally generate game levels. Recently, Machine Learning techniques have been applied to video game level generation towards the purpose of automatically generating levels that have the properties of the training corpus. Towards that end we have made available a corpora of video game levels in an easy to parse format ideal for different machine learning and other game AI research purposes.

## Keywords

Video Games, Level Design, Procedural Content Generation, Machine Learning, Corpus

## INTRODUCTION

For many different video games, levels are one of the critical pieces of game content. They represent the virtual space wherein the majority of player interaction occurs. As such, they represent a very attractive target for Procedural Content Generation (PCG), i.e. the creation of artefacts via an algorithm. Most PCG level creation has been accomplished via human-authored rules from early computer games such as *Rogue* up through modern games such as *No Man's Sky*. A large body of academic work has been performed in this field utilizing classical AI techniques such as constraint satisfaction (G. Smith et al. 2010), Answer Set Programming (A. M. Smith et al. 2012), Evolutionary Algorithms (Sorenson and Pasquier 2010), and others. More recently, statistical AI, i.e. Machine Learning (ML), techniques have been used such as Bayes Nets (Summerville, Behrooz, et al. 2015), Markov Chains (Snodgrass and Ontañón 2013, Summerville, Philip, et al. 2015, Dahlskog et al. 2014), clustering (Guzdial and Riedl 2015), non-negative matrix factorization (Shaker and Abou-Zleikha 2014), PCA (Summerville, Behrooz, et al. 2015), and others. While a large number of different ML techniques have been used they all have one thing in common, they require a training corpus.

Proceedings of 1<sup>st</sup> International Joint Conference of DiGRA and FDG

©2016 Authors. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

Towards that end we have assembled a corpus consisting of 428 levels from 12 games<sup>1</sup>. These levels exist as parseable text files, along with the corresponding image representation of the level, ideal for machine consumption for either ML PCG or other game AI applications. In the following sections we will first discuss the games in the corpus and the details of the levels included; we will then show a small subset of work that has been done with these levels; and finally we will discuss ways these levels could be used in the future.

## RELATED WORK

The collection of resources for the creation of a shared corpus is commonplace in fields such as Natural Language Processing (NLP). A large number of corpora exist in the NLP community, but the two most noteworthy are COCA: "the Corpus of Contemporary American English" (Davies 1990) and the Wall Street Journal Corpus (Paul and Baker 1992). Similarly, many corpora are available in the machine learning community for a variety of different tasks including image classification (Krizhevsky et al. 2016), hand-writing recognition (Lecun and Cortes 2016), and speech-recognition (Godfrey et al. 1992).

Machine learning based approaches for the procedural generation of video game levels have relied on a couple of different data sources: synthetic data, annotated images, and video. While a commonplace activity in many machine learning communities, the only known videogame level PCG to use synthetic data is the work of Shaker and Abou-Zleikha (Shaker and Abou-Zleikha 2014) which used data generated by other generation systems. A number of works have used annotated level images (either annotated by human hand or automatically via image-processing), mostly of *Super Mario Bros.* (Summerville, Philip, et al. 2015, A. Summerville and Mateas 2016, Snodgrass and Ontañón 2013, Snodgrass and Ontañón 2014, Dahlskog et al. 2014, Hoover et al. 2015, Londoño and Missura 2015) but occasionally other games such as those from *The Legend of Zelda* series (Summerville, Behrooz, et al. 2015, A. J. Summerville and Mateas 2015) or *Lode Runner* (Snodgrass and Ontañón 2014). The work of Guzdial and Riedl 2015 used Long Play videos of *Super Mario Bros.* gathered from Youtube to learn the spatial relationships of different sprite groupings.

## DATASETS

The levels from 12 games are present in the corpus as of the publication of this paper. There are three annotation formats: **Tile**, **Graph**, and **Vector**. A breakdown of the games can be seen in table 1. Most of the games in the corpus are 2D tile-based sidescrolling games which are all annotated with the **Tile** format, as are the levels from *Doom*, *Doom 2*, *Mario Kart*, and the first quest of *The Legend of Zelda*. The levels from *The Legend of Zelda* series which have a room based level structure are also annotated in the **Graph** format, and the levels from the *Doom* series can also be found in the **Vector** format.

## Annotation Formats

We will now detail the annotation formats. The **Tile** format is an intuitive format for tile based games, particularly those where reasoning over the space of the game should be done in a tile grid. These tile based levels exist as a two dimensional grid of size  $w \times h$  with  $w$  being the width and  $h$  being the height. Each entry in this grid is annotated by a single

---

1. The corpus can be found at: <https://github.com/TheVGLC/TheVGLC>

Game	Levels	Minimum Sizes	Median Sizes	Maximum Sizes
<i>Super Mario Bros.</i>	20 <b>Tile</b>	150 x 14	188 x 14	374 x 14
<i>Super Mario Bros. 2</i>	25 <b>Tile</b>	161 x 12	194 x 15	364 x 16
<i>Super Mario Land</i>	9 <b>Tile</b>	261 x 15	294 x 16	441 x 16
<i>Super Mario Kart</i>	7 <b>Tile</b>	128 x 128	128 x 128	128 x 128
<i>Kid Icarus</i>	6 <b>Tile</b>	16 x 159	16 x 205	16 x 281
<i>Lode Runner</i>	150 <b>Tile</b>	33 x 22	33 x 22	33 x 22
<i>Rainbow Islands</i>	28 <b>Tile</b>	32 x 83	33 x 165	33 x 252
<i>Doom</i>	36 <b>Tile</b>	36 x 69	121 x 117	225 x 213
<i>Doom 2</i>	32 <b>Tile</b>	54 x 60	131 x 123	231 x 274
<i>The Legend of Zelda</i>	9 <b>Tile</b>	67 x 32	89 x 80	89 x 128
<i>The Legend of Zelda</i>	18 <b>Graph</b>	V  = 12  E  = 22 $\Delta(G) = 3$	V  = 27  E  = 58 $\Delta(G) = 4$	V  = 66  E  = 161 $\Delta(G) = 6$
<i>The Legend of Zelda: A Link to the Past</i>	12 <b>Graph</b>	V  = 14  E  = 31 $\Delta(G) = 2$	V  = 34  E  = 76 $\Delta(G) = 4$	V  = 65  E  = 125 $\Delta(G) = 8$
<i>The Legend of Zelda: Link's Awakening</i>	8 <b>Graph</b>	V  = 21  E  = 43 $\Delta(G) = 3$	V  = 43  E  = 98 $\Delta(G) = 5$	V  = 59  E  = 136 $\Delta(G) = 8$
<i>Doom</i>	36 <b>Vector</b>	122 lines, 53 objects	956 lines, 251 objects	1764 lines, 463 objects
<i>Doom 2</i>	32 <b>Vector</b>	93 lines, 69 objects	774 lines, 253 objects	1690 lines, 509 objects

Table 1: The games included in the corpus as of publication

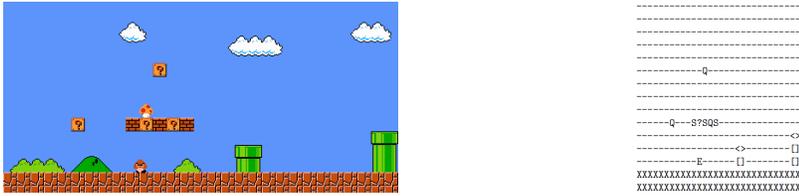


Figure 1: A section of level 1-1 (left) and the annotated text file version (right).

character. Along with the  $w \times h$  array for each level there is a JSON file that acts as a legend. Each tile character is an entry in a dictionary named *tiles* and has an associated array of possible annotation tags, e.g. for *Super Mario Bros.* the tile character for the  has the tags **[solid, ground]** and the character for the  has the tags **[solid, breakable]**. The JSON legend file for *Super Mario Bros.* is included below:

```
{
  "tiles" : {
    "X" : ["solid", "ground"],
    "S" : ["solid", "breakable"],
    "-" : ["passable", "empty"],
    "?" : ["solid", "question block", "full question block"],
    "Q" : ["solid", "question block", "empty question block"],
    "E" : ["enemy", "damaging", "hazard", "moving"],
    "<" : ["solid", "top-left pipe", "pipe"],
    ">" : ["solid", "top-right pipe", "pipe"],
    "[" : ["solid", "left pipe", "pipe"],
    "]" : ["solid", "right pipe", "pipe"],
    "o" : ["coin", "collectable", "passable"],
    "B" : ["solid", "bullet bill", "hazard", "enemy"],
    "b" : ["solid", "bullet bill"]
  }
}
```

A section of the annotated version of Level 1-1 of *Super Mario Bros.* alongside the image it is derived from can be seen in Figure 1. It should be noted that multiple different images can be mapped to the same tile character, e.g. all enemies are mapped to the same character, “E”, and all solid unbreakable tiles (ground, stairs, tree tops, giant mushroom tops) are mapped to the same character, “X”.

The **Graph** annotation format is used for games where the high-level topology should be reasoned about at differently than the low-level structure. Games with discrete room-to-room structures, e.g. *The Legend of Zelda*, or classic adventure games such as *Zork* and *King's Quest* are well represented by graphs. The graph format we chose was the DOT language used by Graphviz (Gansner and North 2000). This format was chosen for 3 reasons:

- **Easily Parseable** - The format is very easily parsed with nodes being represented by `<Node ID> [label=<Node Label>]` and edges being represented by `<Source ID> -> <Target ID> [label=<Edge Label>]`
- **Easily Visualized** - As part of Graphviz, DOT files can be consumed by the *dot* program to visualize the graphs
- **Portable** - Because it is a popular, well-documented format it is able to be used by other programs

As with the **Tile** format there is a corresponding JSON file that acts as a legend for each game. The *Legend of Zelda* legend can be seen below:

```
{
  "vertices" : {
    "e" : ["enemy" ],
    "S" : ["switch" ],
    "b" : ["boss" ],
    "k" : ["key"],
    "K" : ["boss key"],
    "I" : ["key item"],
    "p" : ["puzzle"],
    "s" : ["start"],
    "t" : ["triforce"]},
  "edges" : {
    "S" : ["switch locked" ],
    "b" : ["bombable" ],
    "k" : ["key locked"],
    "K" : ["boss key locked"],
    "I" : ["key item locked"],
    "I" : ["soft locked"],
    "S" : ["switch locked"],
    "s" : ["visible", "impassable"]}}
```

Additionally, below we have a section of the DOT file for the first dungeon from *The Legend of Zelda*, “The Eagle.” The original annotation and the output of the DOT file can be seen in Figure 2:

```
digraph {0 [label=""]
  1 [label="e,I"]
  2 [label="I"]
  ...
  17 [label="e,k"]
  18 [label="p"]
  7 -> 8 [label=""]
  8 -> 7 [label=""]
  ...
  3 -> 13 [label="b"]
  13 -> 3 [label="b"]}
```

The final annotation format is the **Vector** format. For this we chose the Scalable Vector Graphics (SVG) format as it is the most readily readable and viewable vector format. Games

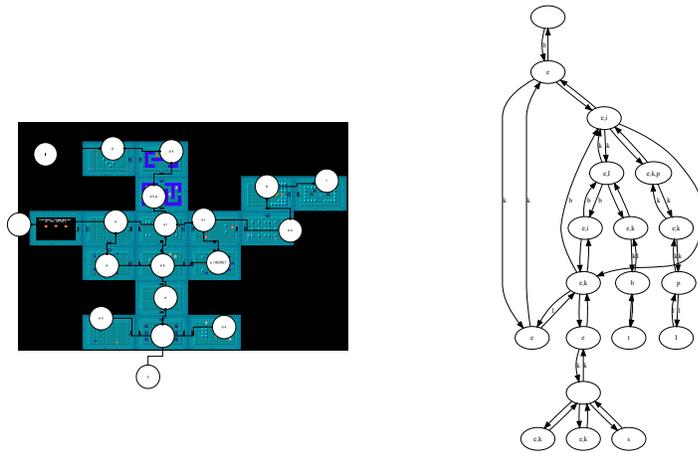


Figure 2: The annotated tilemap for Level 1 “The Eagle” of *The Legend of Zelda* (LEFT) and the corresponding svg output of the DOT file for it (RIGHT).

such as those in the *Doom* series (which is ostensibly 3D but gameplay exists solely in a 2D plane) are best thought of as line segments (walls, doors, etc.) and objects (enemies, weapons, decorations). The SVG format readily handles line segments and discrete objects can be handled as graphical primitives (circles, rectangles, ellipses). Once more, each game has a corresponding JSON legend, *Doom*’s can be seen below:

```
{ "element" : { "line" : { "stroke" : { "#333" : ["solid", "wall"],
"#999" : ["floor", "walkable", "stairs"],
"#d00" : ["door", "locked"],
"#00d" : ["teleport", "source", "activatable"],
"#666" : ["door", "walkable", "activatable"],
"#0d0" : ["exit", "activatable"], } },
"rect" : { "stroke" : { "#900" : ["enemy", "walkable"],
"#090" : ["decorative", "walkable"],
"#009" : ["teleport", "walkable", "destination"] } },
"circle" : { "stroke" : { "#900" : ["weapon", "walkable"],
"#090" : ["ammo", "walkable"],
"#009" : ["health", "armor", "walkable"] } },
"ellipse" : { "stroke" : { "#900" : ["explosive barrel", "walkable"],
"#090" : ["key", "walkable"],
"#009" : ["start", "walkable"] } } } }
```

In the legend, the color of the stroke determines the annotation for that line or object primitive. An SVG representation of *Doom*’s first level can be seen in Figure 3.

The annotations explained above are the initial set of annotations we explored. We recognize that there are limitations. For instance, the **Tile** annotation does not have a tile representation for each enemy, and some other special tiles, which results in a loss of information. In the future, we intend to explore more lossless representations, including more expressive tile sets and ontologies for those tile sets, allowing for more complete representations of the levels. Additionally, the original map images are included in the corpus to allow users to create their own annotations.

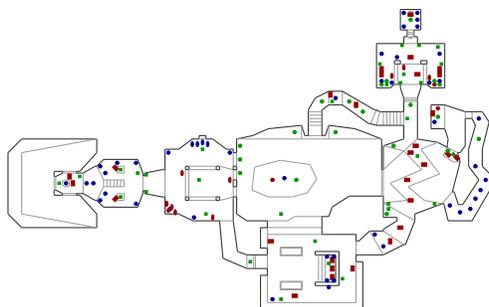


Figure 3: The first level of Doom rendered in SVG. Blue circles represent health/armor, green circles represent ammo, red squares represent enemies, red ovals represent explosive barrels, and other shapes correspond to other less common elements such as teleporters, start points, weapons, etc.

## TOOLS

Along with the levels, we have also included tools that we have used as part of processing. These include a tile based platformer A\* solver and a parser and rasterizer for Doom WAD files. The platformer solver takes in a JSON file that acts as a legend (telling it which tiles are solid) as well as the dynamics of the game (what does a jump arc look like) and produces paths through a given level. The Doom parser and rasterizer take in WAD files (the data format for Doom and Doom-likes) and produces either the SVG or tile based levels found in the corpus.

## POTENTIAL USAGE

In hope of sparking future research, below are some potential uses of the VGLC.

### Corpora Based Procedural Content Generation

This is the most obvious use, or at least the most common use for this work so far. The **Tile** format is set up in a way where nearly any text generation based approach (Markov chains, recurrent neural networks, etc.) can produce results, but given their grid based nature they are also translatable into a form that can easily be consumed for image based methods (Markov random fields, convolutional neural networks, etc.). The **Graph** could be used for graph grammar learning or other graph based approaches (spectral graph analysis, relational learning). The levels can also be processed such as in the work of Londoño and Missura (Londoño and Missura 2015) or Summerville and Mateas (A. Summerville and Mateas 2016) where simulated agents are run through levels to determine how a player could actually traverse through the levels. To our knowledge, the only games to have been used for corpora based generation are *Super Mario Bros.*, *Kid Icarus*, *The Legend of Zelda*, and *Lode Runner* meaning any of the other games are ripe for generation.

### Design Analysis

A large amount of PCG work has relied on assumptions about successful design decisions, but the levels from these games represent actual examples of successful design decisions. Dahlskog and Togelius (Dahlskog and Togelius 2012) performed an analysis on 20 of the

levels from *Super Mario Bros.* to find design patterns that could be used for PCG research, but even in the *Super Mario Bros.* domain, this corpus contains an additional 34 levels to be analyzed. Similarly, Dormans (Dormans 2010) performed an analysis on the mission and physical structure of a level from *The Legend of Zelda: Twilight Princess*, and while none of the 3D Zelda games are represented in this corpus, there are 48 levels from the 2D games of the series that could be analyzed. Beyond those series, the 3 other games could be compared and contrasted with the closer to saturated *Super Mario Bros.*

## Style Transfer

Along those lines, we do not know of any work that has successfully transferred level design style across different games. All work has focused on a single game or series, and when work has included multiple games the work has always been partitioned. Following the work of Gatys et al. (Gatys et al. 2015) there has been an interest in applying different artistic styles to images, similarly, we imagine that an interesting avenue for future work would be one that could reimagine levels of one game in the style of another, or a user could sketch the skeleton of a level and in turn generate variants based on different game styles.

## CONCLUSIONS AND FUTURE WORK

We present this corpus in hopes of helping the community. Each group of researchers that have used corpora based machine learning approaches have needed to reinvent this, admittedly, not altogether exciting wheel, which is why we expect this work to be adopted by the community so that focus can be placed on more exciting and innovative work. The VGLC is already available online (<https://github.com/TheVGLC/TheVGLC>), ready to use. Furthermore, we encourage researchers to contribute additional games and tools to this corpus, to make it become more useful to the community.

## BIBLIOGRAPHY

- Dahlskog, Steve, and Julian Togelius. 2012. "Patterns and Procedural Content Generation: Revisiting Mario in World 1 Level 1." In *Proceedings of the First Workshop on Design Patterns in Games*.
- Dahlskog, Steve, Julian Togelius, and Mark J. Nelson. 2014. "Linear levels through n-grams." In *Proceedings of the 18th International Academic MindTrek Conference*.
- Davies, Mark. 1990. *The Corpus of Contemporary American English: 520 million words*. Available online at <http://corpus.byu.edu/coca/>.
- Dormans, Joris. 2010. "Adventures in Level Design." In *Workshop on PCG in Games*.
- Gansner, Emden R., and Stephen C. North. 2000. "An open graph visualization system and its applications to software engineering." *SOFTWARE - PRACTICE AND EXPERIENCE*.
- Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2015. "A Neural Algorithm of Artistic Style." *CoRR* abs/1508.06576.
- Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. "SWITCHBOARD: telephone speech corpus for research and development." In *Acoustics, Speech, and Signal Processing*.
- Guzdial, Matthew, and Mark O. Riedl. 2015. "Toward Game Level Generation from Gameplay Videos." In *Proceedings of the FDG workshop on PCG in Games*.
- Hoover, Amy K., Julian Togelius, and Georgios N. Yannakakis. 2015. "Composing Video Game Levels with Music Metaphors through Functional Scaffolding." In *ICCC Workshop on CCG*.

- id Software. 1993. *Doom*. [PC] GT Interactive Software, Richardson Texas.
- . 1994. *Doom II: Hell on Earth*. [PC] GT Interactive Software, Richardson Texas.
- Infocom. 1987. *Zork: The Great Underground Empire - Part I*. [PC] Infocom, Cambridge Massachusetts.
- Krizhevsky, Alex, Vinod Nair, and Geoffrey Hinton. 2016. *CIFAR-10 and CIFAR-100*. Available online at <https://www.cs.toronto.edu/~kriz/cifar.html>.
- Lecun, Yann, and Corinna Cortes. 2016. *The MNIST database of handwritten digits*. Available online at <http://yann.lecun.com/exdb/mnist/>.
- Londoño, Santiago, and Olana Missura. 2015. “Graph Grammars for Super Mario Bros Levels.” In *FDG workshop on PCG in Games*.
- Nintendo EAD. 1991. *The Legend of Zelda: A Link to the Past*. [Super Famicom] Nintendo, Kyoto Japan.
- . 1992. *Super Mario Kart*. [Super Famicom] Nintendo, Kyoto Japan.
- . 1993. *The Legend of Zelda: Link’s Awakening*. [Game Boy] Nintendo, Kyoto Japan.
- Nintendo RD1. 1983. *Super Mario Land*. [Game Boy] Nintendo, Kyoto Japan.
- . 1986. *Kid Icarus*. [Famicom] Nintendo, Kyoto Japan.
- Nintendo RD4. 1983. *Super Mario Brothers*. [Famicom] Nintendo, Kyoto Japan.
- Nintendo RD4. 1986. *The Legend of Zelda*. [Famicom] Nintendo, Kyoto Japan.
- Paul, Douglas B., and Janet M. Baker. 1992. *The Design for the Wall Street Journal-based CSR Corpus*.
- Shaker, Noor, and Moahamed Abou-Zleikha. 2014. “Alone We Can Do So Little, Together We Can Do So Much.” In *AIIDE*.
- Sierra On-Line. 1987. *King’s Quest*. [PC] IBM, Oakhurst California.
- Smith, Adam M., Erik Andersen, Michael Mateas, and Zoran Popović. 2012. “A Case Study of Expressively Constrainable Level Design Automation Tools for a Puzzle Game.” In *FDG*.
- Smith, Doug. 1983. *Lode Runner*. [PC] Brøderbund, Renton Washington.
- Smith, Gillian, Jim Whitehead, and Michael Mateas. 2010. “Tanagra: a mixed-initiative level design tool.” In *FDG*.
- Snodgrass, Sam, and Santiago Ontañón. 2013. “Generating Maps Using Markov Chains.” In *AIIDE*.
- . 2014. “Experiments in Map Generation using Markov Chains.” In *FDG*.
- Sorenson, Nathan, and Philippe Pasquier. 2010. “Towards a generic framework for automated video game level creation.” In *Applications of Evolutionary Computation*.
- Summerville, Adam James, and Michael Mateas. 2015. “Sampling Hyrule: Multi-Technique Probabilistic Level Generation for Action Role Playing Games.” In *AIIDE*.
- Summerville, Adam, Morteza Behrooz, Michael Mateas, and Arnav Jhala. 2015. “The Learning of Zelda: Data-Driven Learning of Level Topology.” In *FDG*.
- Summerville, Adam, and Michael Mateas. 2016. *Super Mario as a String: Platformer Level Generation Via LSTMs*. eprint: arXiv:1603.00930.
- Summerville, Adam, Shweta Philip, and Michael Mateas. 2015. “MCMCTS PCG 4 SMB: Monte Carlo Tree Search to Guide Platformer Level Generation.” In *AIIDE*.
- Taito. 1987. *Rainbow Islands: The Story of Bubble Bobble 2*. [Arcade] Taito, Tokyo Japan.
- Toy, Michael, Glenn Wichman, and Ken Arnold. 1980. *Rogue*. [Unix], Santa Cruz, California.