>Remember

By Hugo Labrande

Issue #14 : Digging through Quill games

Google N-Gram is a very interesting tool that uses Google's book digitization efforts as the raw material for textual analysis, allowing us to answer questions such as "when was sherry trendiest?" (1869, it seems). Through measuring and plotting the usage of words, one can attempt to see what are the trends, the ideas, and the worries that were permeating society at these points

This has never been done on text adventures. Until now. By building a pool of text corresponding to the strings in text adventures, then looking through it, I am hoping to be able to find some interesting things. It's not exactly science, and who knows what will turn up, but it'll be fun!

Building the corpus

For this idea to work, I need to get the text in some text adventures in a searchable form, which is to say, extract it to a text file. And, it goes without saying, without just playing a bunch of them and copying the text as I go; but also, ideally, without having to perform tool-assisted extraction by hand on each game.

In order for the extraction of text to be automated, we can look at the formats and engine, as their common structure allows them to be analyzed by a tool. Here, we will be focusing on the Quill and PAWS formats, as there are a few tools that can perform the text extraction on such games. We want this tool to be amenable to automation; the "unQuill" and "unPaws" tools, which are command-line tools that can extract all the data from a snapshot of a Quill or PAWS game, will do just that. They are both available at the IF Archive:

https://ifarchive.org/indexes/if-archiveXprogrammingXquill.html

So, yes we're going to extract the text from all the Quill and PAWS text adventures we can find.

(Note that, if you'd like to do the same thing for Inform and TADS games, look for the program "glulx-strings" :

https://github.com/jcmf/glulx-strings)

So, first, we need to download every Quill & PAWS game that we can find. I focused my efforts on the ZX Spectrum, reasoning that even though The Quill was ported to many other platforms, the Speccy had the most Quilled text adventures. This means I just have to go to a website that archives Spectrum discs and tapes, such as WorldOfSpectrum, and download everything I can. WoS has API access but, uh, I got lazy, and I downloaded the 2017 backup that can be found on archive.org:

https://archive.org/details/World_of_Spectrum_June_2017_Mirror

This is a huge folder, and I used a torrent client to download it. It didn't have many people seeding it, but there were enough that it was eventually downloaded!

Then I "just" needed to get a list of the thousands of text adventures on the website, and copy them in a separate folder. This was tricky, as I needed to massage several lists to extract what I could in minimal time. I first used WoS's CSV export and filtered text adventures (column "entry_type" : "Adventure:Text") and got a little under 1700 games. However, the names in that CSV did not exactly match the names of the disk/tape images. The approach I took was to use the list of game names to generate one-word "clues", and then copy any game that would match these into a separate folder; the result would not be a list comprised of only text adventures, but I reasoned that any non-Quill game extracted this way would just fail at the next step (the automated extraction of text). It was just enough to reduce the number of candidates from "the whole database" to "a reasonable set". So first, I removed the leading "A", "An", and "The" (these usually got shuffled to the end of the file names anyway); then I concatenated the words (removing ', ?, !, and :, but keeping the . - that's how the formatting of the names appeared to work). I then fed this to a Python algorithm that would repeatedly run the following Unix command to get a set of text adventures:

find games/ -name "thestring" -exec cp {} ./text-adventures/ \;*

(I use Linux on my computer, but that doesn't mean there wouldn't be a similar command on Windows, especially if you have WSL.)

Was this perfect? Again, no. But after 5 minutes, it did give me a list of 3776 zip files to work with! After further extraction, I got 905 TAP files, 2235 TZX files, 657 Z80 files, as well as several hundreds of DSK and MGT files. This is more than the list in

the csv file, and this would be because there are duplicates, different formats, several sides per game, re-publications, etc. - but also, perhaps, some non-text adventure files. It didn't matter though: I managed to go from over 24000 files to around 3500, and with a better probability of succeeding. This automated method guarantees a large corpus, with a lot of variety, but it also means that I am going in "blind" as to which text adventures will be included. Will there be missing games, because of copyright issues or the games being considered lost in 2017, or because my script didn't pick them up? Sure! But the fact that I probably have hundreds of Quilled adventures in one folder is promising enough.

A small but crucial hurdle is that most of the games are in ".tap" or ".txz" format, i.e. the raw data from the tape; but unQuill and unPaws require a snapshot of the game, i.e. the data after loading by a Spectrum (".sna"). There is actually a tool, "spconv", that can convert ".z80" images to ".sna" images; it was easy to build a script to do the bulk conversion, and end up with 541 ".sna" files from the ".z80" files. However, there are no tools to convert ".tap" files into ".sna" files (they are different things: one represents a tape, one represents a snapshot of the Spectrum's state). I suppose not a lot of people may have a use for that kind of feature? But after I posted on the "8-bit Text Adventures" Facebook group, Brett Norris wrote a Python program that did just that. Many thanks to him for his program! Using this program, I was able to get 902 ".sna" files. (I tried to use Brett's script to converted 2222 ".tzx" files into the ".sna" format, but it did not work; makes sense, they are two different formats!)

After creating all these snapshots, the next step is to unQuill. Here, we would like all the text that is displayed to the player: this is (if I'm not mistaken) the messages, location descriptions, and system messages. Here is the command that I used to unQuill games (wrapped in a script to automate extraction, of course):

./unquill GAME.sna -SC -SU -SN -SV -SG -SO -SF -OGAME.txt

This command failed on all non-Quill games, but outputted something on Quilled games. However, the output was initially not great: this command, applied on the 900-some files that originated from ".tap" files, gave 220 Quilled files, only 100 of which were not empty (or basically empty). Reapplying this command with the "-L" argument (for "late format Quill") gave 7 more games. Applying these commands to formerly z80 files gave 147 more unquilled files. In total, we have 247 Quilled games!

Turning now to unPAWS, you actually don't really have any control over the input, and everything is dumped. (This is fine, since we can still look through the whole text, but it makes for larger files!) The command here was

wine Unpaws211.exe GAME.SNA GAME.txt -SSNA

This took a little longer (maybe because of running Wine hundreds of times in a row...), but worked, and uncovered a significant number of games. The tap trove gave 174 unPawed games, and the z80 trove gave 309 games. However, I later realized that the tool was also able to unquill games, although with a larger output (full of things that would not be relevant to searching the full text), so this total encompasses Quilled games too!

In the end, I obtained 477 games, forming a database of 28.9MB, split almost evenly between Quilled games and Pawed games. This is a nice number, but pales in comparison to what is referenced in CASA (solutionarchive.com), which has 815 Quill games and 607 PAW games; but this is roughly a third of all games, and the method by which they were selected (i.e. by chance, "whatever my random stabs in the dark could find") means that dataset might not be that biased. In any case, I'm extremely happy that this worked, and would like to thank all the tool authors (John Elliot for unQuill; Carlos Sanchez, Jose Luis Cebrian, and Alexander Katz for unPaws; Henk de Groot for spconv; and Brett Norris for his tap2sna script) for making such great tools that perform reliably and can be automated very easily. As stated, this database is not perfect, and there is opportunity to find more games in the ".tzx" files; but I'm sure someone will eventually find a way to do this!

Text analysis

Let's dive in the text! Here I'm not doing anything fancy either; I just use standard Unix command-line tools for finding strings and counting the number of games in which they appear. This is fast, but could also lead to inaccuracies; but I hope that whichever inaccuracies there are will just balance out and not change the broad interpretation. Also, this is not scientific at all, so it doesn't really matter!

Here is the command I crafted for this:

*grep -i "WORD" --files-with-matches * | wc -l*

This is looking for every occurence of the word "WORD" in the text of all 477 games in the folder, then counting the number of games that appear in the list. Then we can draw conclusions on what Quill authors were interested in and writing about at the time. Let's dive in!

First, I wanted to look at the different languages the adventures were written in using these tools. This would be good to know in order to attempt to make statistics on English-language games (for now! But let me know if you're curious about other languages!). To determine the language, I tried to think of words that would be characteristic of a language, then tried to look if what was found was indeed in the right target language. For Spanish, I searched for " para " (for), which appeared in 77 games, and " hacer", which appeared in 73 games; so there are roughly 80 games in Spanish in my corpus. (Gareth Pitchford has stated that Quill games were rare in Spanish, while PAW was actually translated and adapted to Spanish; so most of these should be PAW games.) I also noticed a few Swedish games, which would have been produced by the Norace versions; I searched for "ligg" ("ligger", "ligga", located), and "mellan" (between), and found 7 Swedish games. Interestingly, searching for " della " and "sono " only turned up 2 Italian games; but my recollection, and hence my hypothesis, is that the Spectrum wasn't that popular in Italy, which was more of a C64 kind of country, so perhaps repeating these steps on C64 Quill games would turn up more. As for French, my search was fruitless (except for a joke message in "The Bimbles"), which makes sense - I'm pretty sure the only commercialized "Quill" localized in French was "AdventureWriter" on Atari 8-bit, which was used for precisely 1 game.

Let's look at geographical locations. What are text adventures made of? Underground caves, deserted islands, maybe castles, forests, or planets and spaceships! Let's see... 85 games with mentions of islands; 80 with castles; 45 planets and 8 spaceships; 78 ships (mostly nautical, I guess? Since there are so few spaceships?); 149 forests, and... 218 caves! Boy, adventurers love their caves and caverns! Maybe we're seeing the influence of the original *Adventure*, and maybe *Zork* too, here?

Let's look at genres! Fantasy is a popular one, and I will use "sword" and "wizard" as a proxy; science-fiction might have aliens and lasers; horror might have vampires, zombies, werewolves, and ghosts; and we should look for detectives and murders for the crime genre. I found 99 swords and 42 wizards, with 53 spells (but that could also be a figure of speech, as in "under his/her spell"). There was also 12 hobbits, 18 orcs and 41 trolls, for the Tolkien fans out there! I also counted 47 lasers, on top of the 45 planets and 31 aliens; 56 ghosts, 15 vampires, 9 zombies, and 4 werewolves; 30 murders, 16 detectives, and 9 inspectors. It seems like fantasy is the winner (with, let's say, 20% of the total? To be honest, I expected higher!), then science-fiction (maybe 12%, then?) and a small number of horror games (10%?) and detective stories (7%?). This would mean that roughly 50% of the games were fantasy, sci-fi, horror, or

crime; to be honest, this seems a bit low (what were the other 50%? Contemporary settings?), but why not!

Any indications of countries? 25 Englands, 14 Britain, 9 UKs, 9 Scotland, and 6 Wales; but somehow, 0 Irelands. (And 15 "shire", with only 1 " shire" - so, a lot of British settings!) Only 10 mentions of France, by 30 mentions of French - either because of French characters, or lots of "excuse my French"! 1 Spain but 8 Spanish; 2 Germany but 10 German; 6 Italy but 11 Italian. This pattern makes sense when you think about it: few games are about travelling to other countries or talking about geopolitics, but you might see or hear characters speaking a different language or with a different accent!

Finally - curse words! (They might actually also give us an idea of how many games written with a teenage sense of humor were written, right?) But I was actually surprised at the results of my search: 12 "shit", 7 "fuck", 1 "bollocks", 6 "wank" including 3 "wankers", which is overall not a lot. Maybe it's because most of these games were trying to be serious; or maybe because these games were distributed or sold; or maybe because British people are polite. I know that almost every French game of the era recognized curse words - but maybe that's what it is: curse words should be recognized by the parser and give a joke, but not actually be written in the text (and I removed the vocabulary tables of the Quilled games, so the curses that are recognized wouldn't show up in my output).

There's a bunch more queries that could be done this way, so I'll stop there. But this method is not just usable for textual analysis; it could also be used, I am sure, for other kinds of statistics, with a little bit more filtering. For instance, on average, how many locations did a Quill game have? All you need to do is dump a bunch of data with unQuill, then analyze it!

I am not entirely sure what would be the consequences of me publishing all these texts online; I am sure the legal risks are low, but we are still talking about copyrighted material and I wouldn't want to get into trouble. But if you'd like me to run a particular query, let me know! And since I've outlined every step, you could also probably build your own dataset in one afternoon. Be sure to let me know how it goes!