

ScholarWorks@GSU

Introducing the Single Player Offline Game Corpus (SPOC): A corpus of seven registers from digital role-playing games

Authors	Dixon, Daniel
Citation	Dixon, D. H. 2024. Introducing the Single Player Offline Game Corpus (SPOC): A corpus of seven registers from digital role-playing games. <i>Corpora</i> 19(1).
Download date	2026-03-16 18:37:45
Link to Item	https://hdl.handle.net/20.500.14694/457

Paper accepted for publication in *Corpora*. Please cite as follows:

Dixon, D. H. 2024. Introducing the Single Player Offline Game Corpus (SPOC): A corpus of seven registers from digital role-playing games. *Corpora* 19(1).

Introducing the Single Player Offline Game Corpus (SPOC): A corpus of seven registers from digital role-playing games

Daniel H. Dixon

Assistant Professor, Applied Linguistics & ESL
Georgia State University

Abstract

This paper describes the compilation and design of the Single Player Offline Game Corpus (SPOC), which is being made freely available for research and educational purposes. The SPOC was compiled by extracting the localization files from the digital directories of four popular commercial digital role-playing games: *Divinity: Original Sin II*, *Fallout 4*, *the Elder Scrolls V: Skyrim*, and *the Witcher 3: Wild Hunt*. The 3.7-million-word corpus contains more than 30,000 texts and is unique from other game corpora in that it has the following three characteristics: (1) the texts are categorized into seven registers using Biber and Conrad's (2019) register framework, (2) texts are systematically parsed into the smallest meaningful units of observation, and (3) all texts were compiled from the data files of the games themselves. Nearly all language use in the four games is accounted for and parsed into register categories based on their underlying situational characteristics, in particular the communicative purposes and the associated contexts in which the texts appear in the games.

Keywords: game corpus, digital games, video games, register analysis, NLP data

1. Introduction

This paper describes the process of compiling and designing the Single Player Offline Game Corpus (SPOC), a 3.7-million-word corpus parsed into 30,182 texts extracted from the digital localization files of four popular commercial digital role-playing games. The SPOC will be freely available to researchers and educators by signing a waiver available at <https://sites.google.com/view/danielhdixon/spoc> and sending a signed copy to ddixon49@gsu.edu. The games included are *Divinity: Original Sin II* (Larian Studios, 2017), *Fallout 4* (Bethesda Game Studios, 2015), *the Elder Scrolls V: Skyrim* (Bethesda Game Studios, 2011), and *the Witcher 3: Wild Hunt* (CD Projekt Red, 2015). Although some data in the SPOC overlaps with data in the game corpora literature discussed below, the SPOC is unique from other game corpora in that it has all three of the following characteristics in its compilation and design: texts from the four games are (1) categorized into seven register categories that were identified through situational analyses following Biber and Conrad's (2019) register framework, (2) systematically parsed into meaningful and contextual units of observation, and (3) sourced solely from the actual localization files of the games themselves. Further, the four games in the SPOC aim to represent a broader population of digital games, games often referred to as open-world role-playing games (RPGs). Games that share these types of designs and mechanics are linguistically interesting because they tend to include hundreds of hours of recorded speech to give voice to the many automated non-player characters (NPCs) and include thousands of written texts that serve several communicative purposes like describing the games' mechanics and the many complex systems that operationalize gameplay.

Researchers have not compiled large corpora of language use from digital games to nearly the same extent as other forms of media like films and television (see Bednarek & Zago, 2019 for an overview), music (e.g., Norgaard & Römer, 2022), and online domains (e.g., Biber et al., 2015). Just in the past few years, however, interest in game corpora has been quickly gaining steam in applied linguistics and related research domains like natural language processing (NLP). For example, van Stegeren and Theune (2020) describe the use of high-quality game corpora for NLP, training AI-systems, as well as procedural content generation (PCG) used in applications like *ChatGPT* while Hämäläinen et al. (2022) praise game corpora as "an extremely useful data source for NLP [that] should be more widely studied" (p. 1). Although rare, until very recently, researchers have compiled corpora from in-game language use (Rodgers and Heidt, 2020), discussions of digital games (Ensslin, 2012) as well as corpora compiled from scraping gaming websites (Thorne, et al., 2012). This broad interest has led to a healthy debate regarding the most robust source of data for compiling a game corpus. The two most common sources for compiling game corpora include (1) extracting texts from the digital directories (i.e., localizations files) of the games themselves and (2) using web scraping tools to collect online transcriptions from community gaming websites. Extracting data from the actual game files, van Stegeren and Theune (2020) argue, will provide the highest quality data since it contains the texts that will be seen by the player during gameplay (p. 3). Rennick and Roberts (2023), however, suggest that the advantages of data extraction from games over gathering online transcriptions "may have been overstated" because online transcriptions better represent the language that a player experiences in a single playthrough. Certainly, the source of data should be driven by the goals of the research, and the advantages of one source over another should be

carefully considered in corpus design and in discussing its representativeness to the broader domain (see Egbert et al., 2022).

In fact, many researchers have drawn on both sources in compiling game corpora. For example, Masso (2009) sourced data from popular gaming community websites, player questionnaires, as well as recording in-game interaction among players and non-player characters (NPCs) to investigate discourse of gender differences in two popular RPGs. Similarly, Rennick and Roberts (2023) targeted both online game transcriptions and data extracted from the games in designing the Video Game Dialogue Corpus (VGDC), a corpus of 6.2 million words of character dialogue from 50 RPGs published between 1985 and 2020. Other researchers have chosen a single source for compilation of game corpora. For instance, Heritage (2022) used software tools to extract language use from the localization files of three games in *The Witcher* series rather than mixing game data with online transcriptions. Heritage notes that although the data is highly representative since it is the same data used in displaying in-game text, the data was "decontextualized" in that sentences from various aspects of the games were mixed together and appear in an order that would not be experienced by players. Nevertheless, Heritage argues that contextualized data was not critical to the goal of the research: to investigate gender representation in *The Witcher* series. Hämäläinen et al. (2022) also exclusively used extracted game data and targeted dialogue in *Fallout New Vegas*. They chose this source of data because many of the lines of dialogue include sentiment annotations from the game developers like *anger*, *disgust*, and *happy*, among others. Their goal was to test the efficacy of NLP tools in correctly predicting sentiment in the dialogue lines. Thus, extracting the game's dialogue as the data source was a requirement for the goals of their research. In a similar research domain, natural language generation (NLG), Juraska et al. (2019) compiled a data set they refer to as *ViGGO*, which consisted of discussions of video games posted from users of two online game databases. Their goal in using this data was to build and test a language generator that "can support a natural multi-turn exchange on the topic of video games" (p. 168), which motivated the use of data from online websites rather than actual language extracted from the games themselves. These studies reiterate the importance of aligning research goals with the source data that best represents the target domain of language use.

In the next section, the goals and methods in compiling the SPOC are discussed, followed by an outline of the situational analyses of in-game language use that identified the seven game registers. Next, the characteristics that make each register unique are detailed along with screenshots illustrating their appearance in the games. Following details of the register characteristics, the units of observation for texts in each register are defined, and text structures are detailed. The paper concludes with direction for using the SPOC in future research. To give an overview of the SPOC, Table 1 reports the number of texts and words in each of the seven registers across the four games.

Table 1. Distribution of words and texts across registers/games

	Divinity	Fallout	Skyrim	Witcher	REGISTER TOTALS
Spoken					
<i>Immersive Speech</i>					
words	110,731	484,585	511,589	212,656	1,319,561
texts	2,567	2,507	3,846	4,507	13,427
<i>Interactive Speech</i>					
words)	609,372	566,386	89,928	431,991	1,697,677
texts	1,267	1,443	518	1,373	4,601
Written					
<i>Character Text</i>					
words	12,525	23,752	6,606	2,806	45,689
texts	1,163	3,792	519	204	5,678
<i>Quest Objectives</i>					
words	4,898	4,886	6,420	18,385	34,589
texts	153	256	394	356	1,159
<i>Quest Stages</i>					
words	37,124	21,671	23,101	57,555	139,451
texts	158	146	203	356	863
<i>Lore</i>					
words	18,748	14,813	361,722	60,960	456,243
texts	350	202	788	383	1,723
<i>Tutorial Text</i>					
words	5,260	42,904	19,177	1,298	68,639
texts	247	1,602	837	45	2,731
GAME TOTALS	Divinity	Fallout	Skyrim	Witcher	
words	792,154	857,584	848,801	785,651	3,761,849
texts	5,905	9,948	7,105	7,224	30,182

2. SPOC compilation and design

The goal in compiling the SPOC was to represent all the language (both spoken and written) that a player could encounter across all contexts in the four games. Thus, the decision was made to exclusively use data extracted from the games' localization files to best represent all possible language that a player could experience in any given playthrough rather than using web scraping tools to gather online transcriptions from community gaming websites. These localization files are used to display text and play audio in the context designed by the game developers. To extract the localization files, modification software or *mod tools* were used, and each game required a unique tool (see Table 2 for a list). Customized Python scripts were written to parse the raw extracted data into meaningful units of observation. Due to space limitations, instructions

on using these mod tools are not discussed in this paper, but a discussion on their use is provided by Dixon (2022).

Table 2. Mod tools used in SPOC compilation

Game title	Mod tool title	Developer
Divinity: Original Sin II	The Divinity Engine 2	Larian Studios (2017)
Fallout 4	Fallout 4: Creation Kit	Bethesda Game Studios (2016)
The Elder Scrolls V: Skyrim	Skyrim: Creation Kit	Bethesda Game Studios (2012)
The Witcher 3: Wild Hunt	Modkit	CD Projekt Red (2015)

The seven registers within the SPOC provide a categorical framework for the situational contexts of language use common in the broader population of modern commercial digital role-playing games (RPGs). These registers were identified through a series of *situational analyses* (see Biber and Conrad, Chapter 2, 2019) that compared language use across the various contexts in the games, drawing on hundreds of screenshots to document and categorize language use. The seven registers that were identified include two spoken and five written registers. The operational definitions and communicative purposes of the registers are outlined in Table 3.

Table 3. Operational definitions and communicative purposes of the seven game registers

Register	Operational definition	Communicative purposes
Spoken Registers		
<i>Interactive Speech</i>	Speech during which the Player is prompted to provide input or a response to an action or utterance of a non-player character (NPC)	Provide a sense of conversational agency allowing the player to feel that they are an active participant in the scripted dialogue Provide direction for completing quests
<i>Immersive Speech</i>	Speech that does <i>not</i> prompt the Player to provide input	Communicate NPCs' feelings and stances Develop the characters by adding depth to the storylines and fictional settings Provide direction for completing quests
Written Registers		
<i>Character Text</i>	Any unit of text that describes any one of the following aspects related to the Player Character(s): A. Inventory (e.g., armor, weapons, usable items) B. Attributes (e.g., strength, intelligence, dexterity) C. Skills (e.g., spells, abilities)	Describe the effects and uses of inventory, attributes, and skills as well as the context in which they can be used
<i>Quest Objectives</i>	Text accessed through the game's user interface that lists the objectives of a quest	List the actions that need to be taken to complete a quest
<i>Quest Stages</i>	Text accessed through the game's user interface that provides narrative about the quest objectives	Provide the context and purpose for the actions needed to complete a quest as they relate to broader storylines and characters

<i>Lore</i>	Text that is connected to an in-game object that cannot be equipped by the Player's Character(s). Generally, it can only be read by the Player and sold for in-game currency. Less commonly, Lore items may be required to progress further in a quest or provide attribute boosts to the player character.	Describe locations, characters, and relationships within the fictional settings Provide aesthetic information that is typically not critical to progression
<i>Tutorial Text</i>	Text that pops up on screen or is accessed through the game's user interface that addresses the real-life Player (and not one of the fictional game characters) describing the mechanics or rules of the game	Teach the mechanics and rules of the games, explaining the actions that the player can take and their resulting effects on characters and game environment

2.1 Interactive Speech and Immersive Speech

Spoken language texts (i.e., texts with recorded speech audio) were separated into two registers: Interactive and Immersive Speech. The key difference between these two registers is that during Interactive Speech instances, the player is prompted to make a dialogue choice during conversations with NPCs whereas Immersive Speech instances do not prompt the player to make dialogue choices. Thus, if the text includes one or more dialogue choices, it is categorized as Interactive Speech. The absence of choice places that text in the Immersive Speech category. For most dialogue choices, the conversation travels along different scripted paths to give the player a sense of conversational agency. For example, Figure 1 illustrates an Interactive Speech instance from *Fallout 4* where the player character (i.e., the main character controlled by the real-life player) is speaking with an NPC named Piper. During the conversation, the game pauses, and the player is prompted to select one of the four dialogue options displayed at the top-right of Figure 1. Depending on the selection, the conversation will travel along one of four branches, which allows the player to decide whether to help Piper or request more information. Such choices often affect the path that relationships with characters can take and the development of the broader storyline. Some Interactive Speech texts contain several points at which a dialogue choice can be made, and how these were noted and accounted for in the SPOC texts are discussed in Section 3.

In contrast to Interactive Speech, Immersive Speech instances do not prompt the player to make dialogue choices. Rather, these spoken lines of dialogue serve the communicative purpose of providing a sense of immersion and depth to the fictional setting and the characters within it. In Figure 2 from *Fallout*, a security guard is thanking the player character for helping the town with a paint project that was an objective in an earlier quest. As the player walks by the NPC, recorded audio automatically plays and subtitles transcribing the character's speech are displayed.



Figure 1. Interactive Speech (Fallout 4)



Figure 2. Immersive Speech (Fallout 4)

2.2 Character Text

Character Text, one of the five written registers, provides labels and descriptions of the items, skills, and attributes of the player character. These texts often appear in a menu in which the player selects and equips their character with gear like weapons and armor or other usable items from their inventory. Also included in Character Text are descriptions of player character skills

and attributes like strength, intelligence, or endurance, among others. As the player progresses, they gain experience points which are used to *level up* their character. For each new level, the player can choose to invest in one attribute over another. Figure 3 illustrates the *endurance* attribute from *Fallout 4*. As the text in Figure 3 explains, investments in endurance provide the character with more *hit points*, making the character more resilient and also allowing them to sprint for a longer period. In short, the shared communicative purpose of Character Text is to label and describe these various items, skills, and attributes, and explain the contexts in which they have an effect.

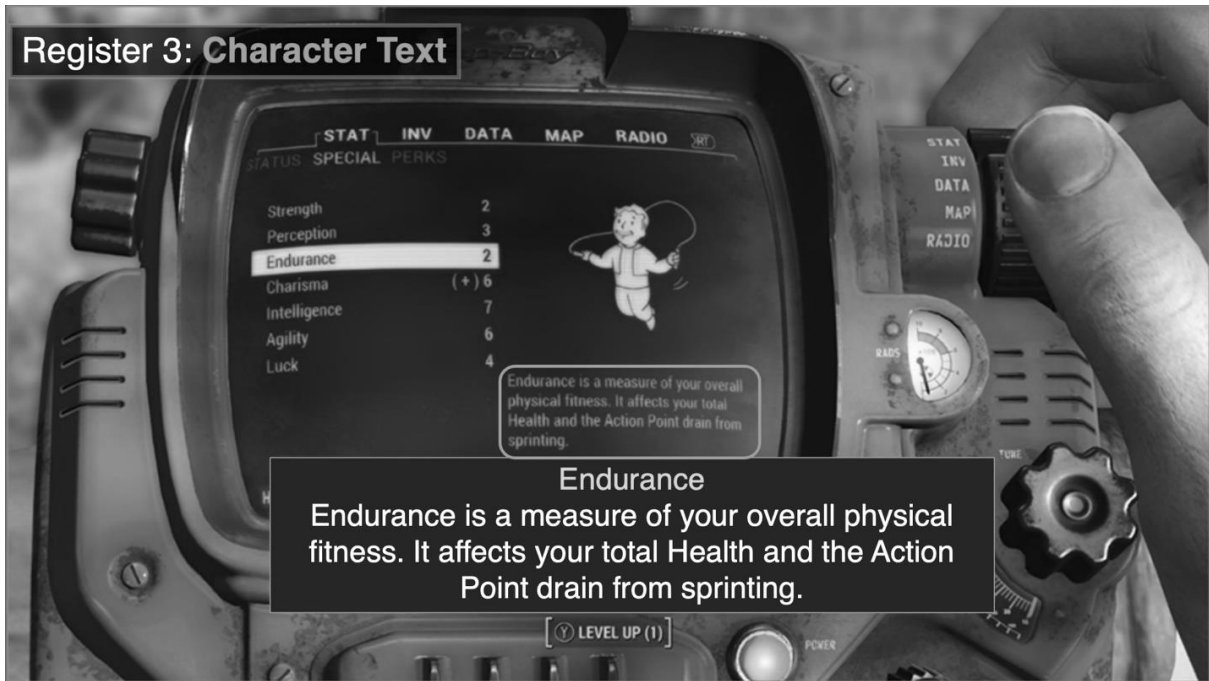


Figure 3. Character Text (*Fallout 4*)

2.3 Quest Text: Quest Stages and Quest Objectives

Although Quest Stages and Quest Objectives both relate to progressing through the games' quests, they were split into two registers due to their differences in communicative purposes and unique linguistic characteristics. Quest Objectives consist of a few short directives communicating the actions (i.e., objectives) that the player needs to take to progress in a particular quest. Quest Stages provide a narrative explaining those objectives in the context of the broader storyline of the games. To illustrate, Figure 4 shows part of the *Skyrim* in-game menu where these texts often appear. The text in the center of the screen lists the objectives, directing the player to find equipment, escape the town of Helgen, and get to *the Keep*. Above this list (i.e., quest objectives) is a short narrative describing the relationship of those objectives with the game's story (i.e., quest stages).

2.4 Lore

The Lore register consists of texts that are usually not critical to progressing through the primary quests and objectives. Rather, these texts aim to add a sense of immersion or *flavor* to the world

in which the games are set. Often, Lore texts describe the history, politics, or character relationships within the settings. Like Character Text, Lore texts are represented as in-game items, but a key difference is that Lore texts cannot usually be used or equipped by the player character and typically appear in the form of a book or a message. For example, Figure 5 displays a Lore text from Skyrim, which describes the relationship of two historic characters. The information contained in the text is not central to completing quests. Rather, the purpose is to add scale to the immersive narrative and engage players in its story.

2.5 Tutorial Text

Finally, Tutorial Text provides the player with information related to the mechanics of the game. RPGs can be quite complex, so rather than introducing all the rules and mechanics at one time, information is often scaffolded, and new mechanics are introduced as the player progresses through the game. These texts often appear in a window when a new mechanic is first introduced, but often these explanations can often be reread in the user interfaces or game menus. Figure 6 displays an example of Tutorial Text from Fallout 4. Once the player progresses through the early stages of the game, they are given the ability to build settlements in various areas of the game. The text displayed explains where structures can be placed and the material that is required to construct them.

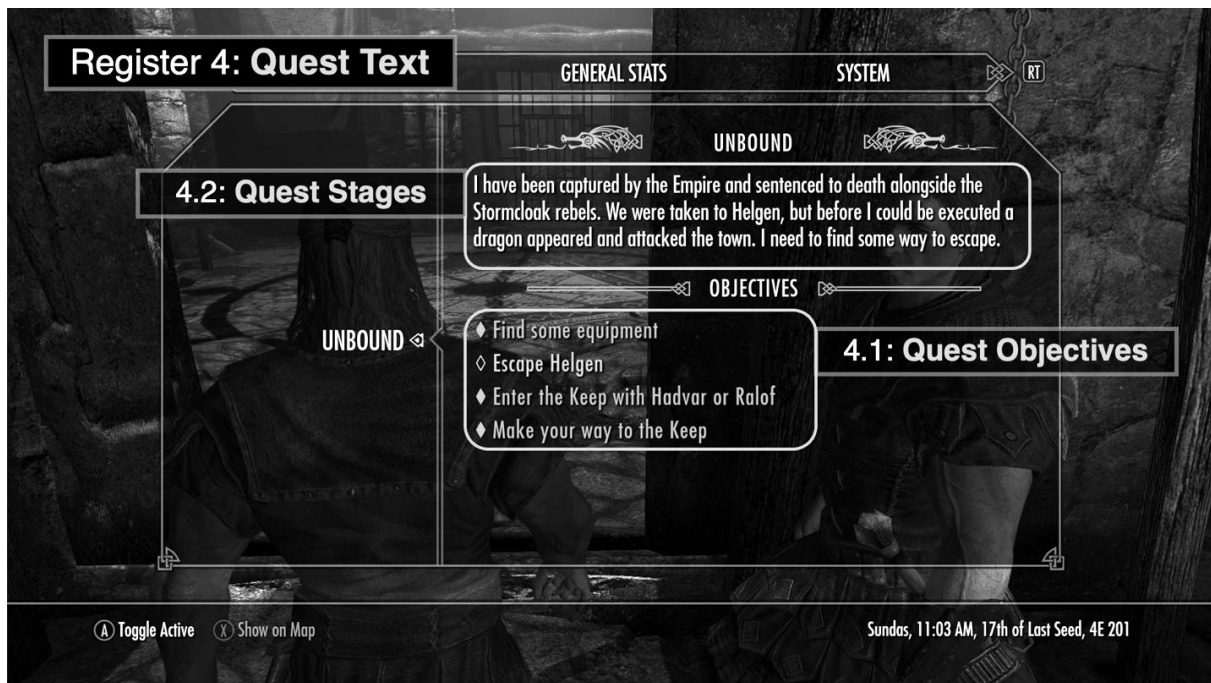


Figure 4. Quest Stages and Quest Objectives (Skyrim)

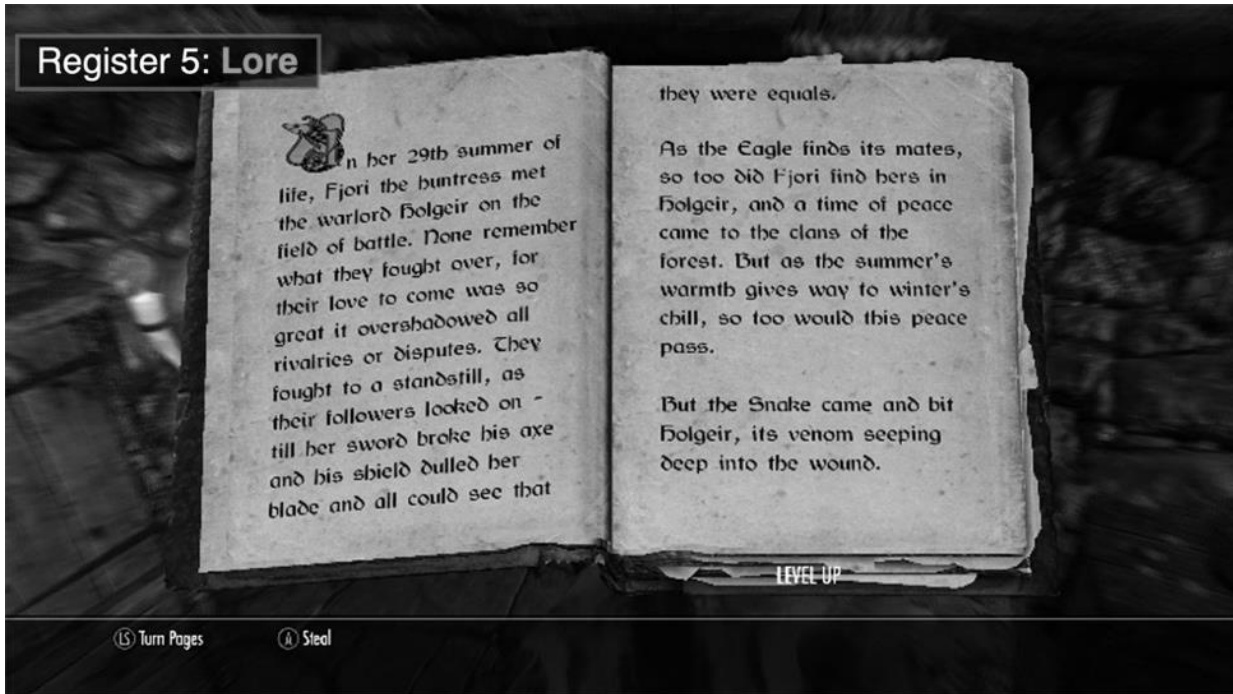


Figure 5. Lore (Skyrim)

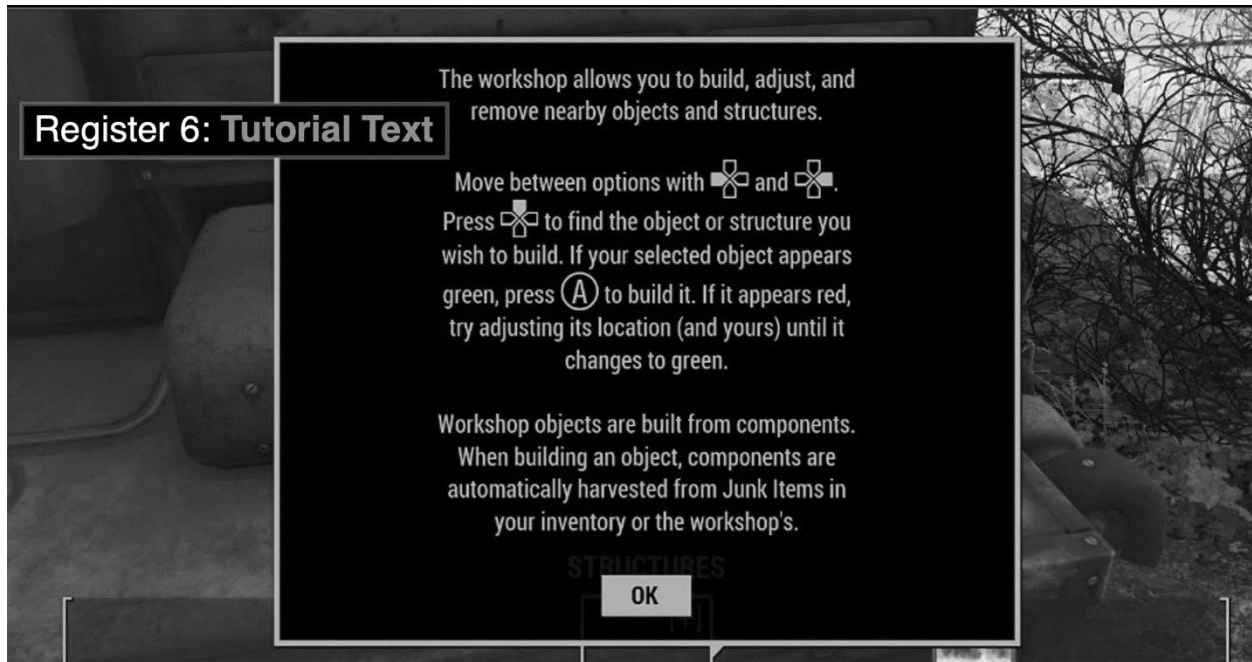


Figure 6. Tutorial Text (Fallout 4)

3. Corpus make-up and units of observation of texts

The next step in compiling the SPOC was to parse language use from the registers into units of observation (i.e., texts). A major goal for compiling the SPOC was to maintain coherent, comparable, and consistent units of language within a register across the four games. Although the seven registers are found in all four games and share common communicative purposes, the

localization files that make up the games' raw language data are not uniformly organized. Just as each game required a unique mod tool to extract and decrypt the game's localization files, each game also required customized Python scripts to be developed to convert these localization files into texts that represent the same or very similar units of language in a register. These Python scripts will be made available at <https://sites.google.com/view/danielhdixon/spoc>. In the section that follows, the parameters, structures, and contextual metadata of a single text in each register are outlined.

3.1 Spoken registers text units

Determining the unit of observation (i.e., a text) in a corpus of conversations is challenging, especially because, among other reasons, "many conversations do not have well-defined openings or closings" (Crowdy, 1995, p. 277). Despite this challenge, Egbert et al. (2021) stress that fundamentally texts must represent functional units for a corpus to be "useful for addressing register-related research questions" (p. 717), and they provide a framework based on four parameters for determining a functional "discourse unit": (1) coherent for its overarching communicative goal, (2) characterized by one or more communicative purposes, (3) recognizably self-contained, and (4) consist of a minimum of five utterances or 100 words. Drawing on Egbert et al.'s (2021) parameters, the discourse unit for the two spoken SPOC registers were compiled into the smallest coherent unit that could be reliably identified. Figure 7 illustrates an Interactive Speech text, and the discussion that follows details the structure of spoken texts.

For Divinity, the raw unparsed game data noted an audio file for each line of dialogue. Similarly, the Witcher's data had a *scene* file noted for a group of lines predetermined by the developers. For these two games, a single text consisted of all lines of dialogue within a single file. Fallout's raw data listed all lines of dialogue in a single spreadsheet with each row containing one line of dialogue and columns listed additional contextual information for each line. For example, one column listed the quest during which the line is heard, another listed a scene within that quest, and another listed a category. Skyrim's raw data was nearly identical except that instead of a *scene*, a column listed a *branch* within a quest in which the line is heard. Thus, to parse dialogue into the smallest coherent unit of observation, a single spoken text was parsed by the associated subcategory within a quest, which is a *scene* in Fallout and a *branch* in Skyrim. If no scene or branch was listed, then the associated *category* was used. In short, the spoken texts are meaningful in that each text represents dialogue from a specific time and place within the larger context of a specific quest. At the top of each text, metadata is included in angled brackets which includes text name, register, the related quest, the category used by the developers (if listed), and the associated branch or scene (if listed).

After the metadata, the spoken lines of dialogue appear. The speaker's name (as it appears in the raw data) is noted in angled brackets followed by the line of dialogue that is heard. Some lines have script notes or other contextual information that is not part of the recorded audio. For example, this often includes notes to the voice actors like *surprised* or *impressed*. This type of information is listed beneath the line in angled brackets and is followed by a double line break to indicate where the next spoken line of dialogue begins. In short, only the actual spoken lines (i.e., recorded audio) is outside of angled brackets so that researchers can automate linguistic

analysis by writing a script that ignores everything within angled brackets. Finally, the word count is noted at the end of the text, which includes all words that have recorded audio but excludes metadata and other information in angled brackets.

```
<TEXT NAME: f4_sp_int_7_BoS101.txt>
<REGISTER: interactive_speech>
<RELATED QUEST: BoS101>
<SCENE: BoS101SceneStage030>
<DEVELOPER CATEGORY: Miscellaneous>

<PLAYER CHOICES: Understood - Not my first rodeo - Hurry this up - Mission recap>

<PlayerVoiceFemale01> Understood.
<PLAYER CHOICE: Understood>
<SCRIPT NOTES: Confident>

<PlayerVoiceFemale01> This isn't my first rodeo.
<PLAYER CHOICE: Not my first rodeo>
<SCRIPT NOTES: Irritated>

<PlayerVoiceFemale01> Let's hurry this up.
<PLAYER CHOICE: Hurry this up>
<SCRIPT NOTES: Irritated>

<PlayerVoiceFemale01> Why are we at ArcJet again?
<PLAYER CHOICE: Mission recap>
<SCRIPT NOTES: Question>

<NPCMPaladinDance> Outstanding.
<SCRIPT NOTES: Impressed>

<NPCMPaladinDance> I understand that. I'm simply offering valuable tactical advice. You'd do well to listen.
<SCRIPT NOTES: Irritated>

<NPCMPaladinDance> I don't plan on rushing through the facility without extreme caution.
<SCRIPT NOTES: Irritated>

<NPCMPaladinDance> We're here to retrieve a device that will boost the signal from the radio tower at the Cambridge Police Station.
<SCRIPT NOTES: Confident>

<NPCMPaladinDance> Contrary to what you might believe, I'd like you to remain alive during the course of our mission.
<SCRIPT NOTES: Irritated>

<NPCMPaladinDance> Without it, we can't make contact with our superiors.
<SCRIPT NOTES: RE-RECORD PLEASE / Irritated>

<NPCMPaladinDance> Remember, our primary target is the Deep Range Transmitter.
<SCRIPT NOTES: Stern>

<NPCMPaladinDance> Stay focused and check your fire. I don't want to be hit by stray bullets.
<SCRIPT NOTES: Stern>

<NPCMPaladinDance> Listen up. We do this clean and quiet. No heroics and by the book. Understood?
<SCRIPT NOTES: Question>

<WORD COUNT: 128>

<DUPLICATE LINES BY AN ALTERNATE SPEAKER>

<PlayerVoiceMale01 ---> Understood.>
<PlayerVoiceMale01 ---> This isn't my first rodeo.>
<PlayerVoiceMale01 ---> Let's hurry this up.>
<PlayerVoiceMale01 ---> Why are we at ArcJet again?>

<REPEATED LINES WORD COUNT: 16>

<TOTAL WORD COUNT: 144>
```

Figure 7. An Interactive Speech Text (Fallout 4)

Two other important aspects of the spoken texts are worth noting here. First, the Interactive Speech texts have dialogue choices, and the text seen by the player is listed between the metadata and the first line of dialogue in angled brackets beginning with the words PLAYER CHOICES. This is followed by a list of the choices separated by a hyphen. After a choice is made, a specific line of dialogue is heard, and this is noted under the line by the words PLAYER CHOICE: followed by the text that was selected (in angled brackets). Unfortunately, the raw data in Divinity did not mark which choice triggered which line, so that information is not given in

the Divinity Interactive Speech texts. Second, sometimes the same line dialogue may be spoken by an alternative speaker. These duplicate lines of dialogue and their speakers are listed in angled brackets at the end of the text after the word count. To allow for automated analysis, these duplicate lines have an arrow (--->) so that researchers can choose whether to include these lines in analysis. After the duplicate lines list, the number of words in the duplicate lines is listed. Then another word count notes the total word count including non-duplicate and duplicate lines. In sum, each text includes all lines of dialogue that *could* be heard in a particular context although it is very unlikely that *all* lines will be heard in a single playthrough of the games.

3.2 Written registers text units

For the written registers, determining the unit of observation of a text was fairly straightforward compared to the spoken registers. For example, a single text in the Character Text register generally consists of the name of a single item, skill, or attribute its associated description (when a description was given). One Lore text includes the prose seen on the screen for a single Lore item. Quest Objectives and Quest Stages include all the listed objectives or all the prose that is associated with a single quest, respectively. Finally, one text in the Tutorial Text register contains the prose related to a single explanation as it was parsed in the games' raw data files by the developers. In addition, Tutorial Text includes the texts that appears in the games' loading screens. While many of these fit well into the Tutorial Text category since they often explain game mechanics; at times, these texts provide information on the games' characters and settings and would likely better fit into the Lore register. Without manually reading and sorting all 2,731 Lore texts based on their description, this determination would not be possible. This is noted as a limitation of the current version of the SPOC.

4. Conclusion

Although most language use is accounted for in the games in the SPOC, future research can take the register analysis approach and identify additional registers in games with different designs and mechanics. For example, massively multiplayer online games (MMOs) are played online with hundreds of other players synchronously, which will likely reveal additional register categories not captured in the population of games targeted in the SPOC. Further, future research can use the SPOC to compare language use in games to real-world registers or to evaluate and analyze the frequent linguistic features to inform development of L2 learning tasks in the domain of digital game-based language learning (DGBLL). Additionally, digital educational games have been found to be less effective for L2 learning than games designed for entertainment (Dixon et al., 2022). To design more engaging tasks in apps or platforms, developers of L2 educational technology can draw on analyses of linguistic features in the SPOC, potentially increasing their effectiveness for L2 learning. Future research in natural language processing and machine learning could also use the linguistic data in the SPOC for language model training, sentiment analysis, and dialogue generation (see van Stegeren and Theune, 2020). Finally, to accommodate a range of research aims, both a plain text version and a spaCy (Honnibal and Johnson, 2015) tagged version of the SPOC is available by going to <https://sites.google.com/view/danielhdixon/spoc>, signing a waiver stating its use is for research or educational purposes, and emailing the signed copy to ddixon49@gsu.edu.

References

- Bednarek, M. and R. Zago. 2019. *Bibliography of Linguistic Research on Fictional (Narrative, Scripted) Television Series and Films/Movies*. Version 3 (May 2019). Accessed at: <http://unipv.academia.edu/RaffaeleZago>
- Bethesda Game Studios. 2015. *Fallout 4* [Digital game]. Bethesda Softworks.
- Bethesda Game Studios. 2016. *Fallout 4: Creation Kit*. [Digital game]. Bethesda Softworks. <https://www.creationkit.com>
- Bethesda Game Studios. 2011. *Skyrim* [Digital game]. Bethesda Softworks.
- Bethesda Game Studios. 2012. *Skyrim: Creation Kit*. [Digital game]. Bethesda Softworks. <https://www.creationkit.com>
- Biber, D. and Conrad, S. 2019. *Register, genre, and style*. Cambridge University Press.
- Biber, D., Egbert, J., and Davies, M. 2015. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, 10 (1), pp. 11-45.
- Carrillo Masso, I. 2009. Developing a methodology for corpus-based computer game studies. *Journal of Gaming & Virtual Worlds*, 1(2), 143-169. <https://doi.org/10.1386/jgvw.1.2.143/7>
- CD Projekt Red. 2015. *ModKit*. [mod tool software]. CD Projekt Red. <https://www.thewitcher.com/us/en/news/1094/the-witcher-3-wild-hunt-now-with-the-mod-support>
- CD Projekt Red. 2015. *The Witcher 3: Wild Hunt*. [Digital game]. CD Projekt.
- Crowdy, S. 1995. The BNC spoken corpus. In Geoffrey Leech, Greg Myers & Thomas Jenny (eds.), *Spoken English on computer: Transcription, mark-up and application*, 224–235. Harlow: Longman.
- Dixon, D. H. 2022. *The language in digital games: Register variation in virtual and real-world contexts*. [Doctoral dissertation, Northern Arizona University]. ProQuest. Available from <https://www.proquest.com/docview/2703973035/A13FCC138E6A4861PQ/1>
- Dixon, D. H., Dixon, T., & Jordan, E. 2022. Second language (L2) gains through digital game-based language learning (DGBLL): A meta-analysis. *Language Learning & Technology*, 26(1), 1–25. <http://hdl.handle.net/10125/73464>

Egbert, J., Biber, D., & Gray, B. 2022. Domain considerations. In J. Egbert, D. Biber, & B. Gray (Eds.), *Designing and evaluating language corpora* (pp. 68-121). Cambridge. <https://doi.org/10.1017/9781316584880.004>

Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T., & Baker, P. 2021. Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41 (5-6), 715-737. <https://doi.org/10.1515/text-2020-0053>

Ensslin, A. 2012. *The language of gaming*. Palgrave Macmillan.

Hämäläinen, M., Alnajjar, K., & Poibeau, T. 2022. Video Games as a Corpus: Sentiment Analysis using Fallout New Vegas Dialog. *arXiv*, preprint: <https://doi.org/10.48550/arXiv.2212.02168>

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

Honnibal, M. and Johnson, M. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378: Association for Computational Linguistics.

Juraska, J., Bowden, K. K., & Walker, M. 2019. Viggo: A video game corpus for data-to-text generation in open-domain conversation. *arXiv*. <https://doi.org/10.48550/arXiv.1910.12129>

Larian Studios. 2017. *Divinity: Original Sin II*. [Digital game]. Larian Studios.

Larian Studios. 2017. *The Divinity Engine 2*. [mod tool software]. Larian Studios. https://docs.larian.game/The_Divinity_Engine_2

Norgaard, M. and Römer, U. 2022. Patterns in music: How linguistic corpus analysis tools can be used to illuminate central aspects of jazz improvisation. *Jazz Education in Research and Practice* (3) 1, pp. 3-26.

Reinhardt, J. 2019. *Gameful second and foreign language teaching and learning: Theory, research, and practice*. Palgrave Macmillan. <https://doi.org/10.1007/978-3-030-04729-0>

Rennick, S. and Roberts, S.G., 2023. The video game dialogue corpus. *Corpora*, 19(1). preprint: <https://eprints.gla.ac.uk/291596/1/291596.pdf>

Rodgers, M. and Heidt, J. 2020. Levelling up comprehensible input and vocabulary learning in V. Werner and F. Tegge (eds.) *Pop Culture in Language Education: Theory, Research, Practice*, pp 215-227. Taylor and Francis.

Thorne, S. L., Fischer, I. and Lu, X. 2012. The semiotic ecology and linguistic complexity of an online game world. *ReCALL* 24 (3), 279-301. <https://doi.org/10.1017/S0958344012000134>

van Stegeren, J. and Theune, M. 2020. Fantastic strings and where to find them: The quest for high-quality video game text corpora. In *2020 Workshop in Intelligent Narrative Technologies (INT'20)*. Retrieved from <https://judithvanstegeren.com/assets/vanstegeren2020fantastic.pdf>