

ゲームコーパスの設計方針と構築方法

麻 子軒（関西大学）[†]

Approaches to Design and Construction of a Game Corpus

Tzu-Hsuan Ma (Kansai University)

要旨

ゲームコーパスを体系的に構築するにあたり、その目的を明確にした上で、それに合わせた形で発売年代やジャンルの諸観点から代表的なゲームを選定する必要がある。また、ゲームの場合はプレイヤーの操作によって表示されるテキストの内容と量が変化するため、テキストの認定基準をはじめとする構築方法は書籍をベースとしたコーパスとは異なるように考えられる。本稿ではゲームにおける言語的特徴の解明と日本語教育への応用との二つの目的を意識し、現在筆者が構築中のゲームコーパスの設計方針を発売年代、ジャンル、発売本数、世界観の4つの観点に分けて述べる。なお、収録環境の整備、作業人員の募集、プレイ動画と画像収録、文字化作業などの構築方法、及び構築時の問題点についても言及した。

1. 経緯

麻（2022）では、ゲームコーパスの構築理由とそれを用いた研究事例を提示された。しかし、同研究で挙げられた事例は単発的なもので一貫性に欠け、また、複数のゲームを体系的にコーパスに取り入れる構想にはまだ至っていない。ゲームコーパスを体系的に構築するにあたり、その目的を明確にした上で、それに合わせた形で発売年代やジャンルなどの諸観点からゲームを選定する必要がある。また、ゲームの場合はプレイヤーの操作によって表示されるテキストの内容と量が変化するため、テキストの認定基準をはじめとする構築方法は、書籍をベースとしたコーパスと異なるように考えられる。以上を踏まえて、本稿では、筆者が現在構築しているゲームコーパス（以下、本コーパス）の設計方針、及びその構築方法と問題点について述べる。

これより、2節では設計方針、具体的には収録ゲームの発売年代やジャンルなどの選定基準を述べ、3節では構築方法、具体的にはテキストの認定基準や電子化の方法などを述べ、最後に4節で構築時の問題点を挙げる。

2. 設計方針

コーパスを構築するにあたり、どのような目的で、どのようなデータが必要なのかを、まず明らかにしなければならない。なぜなら、調査目的によって、選定されるサンプルやその量、ないし必要なアノテーション情報も異なるからである。現在筆者が関心を持っているのは、ゲームにおける言語的特徴の解明、及びゲームを日本語教育に応用する可能性の2点であるため、以下ではこれらの目的を意識し、発売年代、ジャンル、発売本数、世界観の4つの観点からゲームの選定基準を述べる。

[†] kenji.ma@kansai-u.ac.jp

2.1 発売年代

ゲームにおける言語的特徴も時代によって変化する可能性がある。ただ、ゲームの歴史は書籍と異なり、日本で広く世に知られているのは1990年代以降のことである。2023年現在でも30年ほどの歴史しかないため、本格的な通時的調査には適していない。とはいえ、技術の進歩にともない、ゲームの表現手段も進化しており、それによる言語面での変化は確実に存在すると思われる。特にここ数年、保存媒体の容量向上により、漢字表記の多用や、キャラクターの音声付与ができるようになったため、ゲームの描写手法にも大いに影響を与えたと予想される。

本コーパスでは、据置ゲーム機の全盛期（1990～2000年、以下前期）と、直近十年間（2010～2022年、以下後期）の代表的なゲームを選定する。前期は容量制限で2Dかつ音声なしのゲームがほとんどで、後期は3Dかつ音声ありのゲームが主流となっている。これにより、技術の進歩による言語的变化が観察できると思われる。発売のプラットフォームは、据置ゲーム機に限定する。携帯ゲーム機は、後述する収録の技術的な理由により、現段階では対象としない。

2.2 ジャンル

また、ゲームはその遊び方によって、十数種類のジャンルが存在する。ゲームの言語的特徴を調査するには、なるべく全ジャンルを網羅的に取り入れるのが理想的だが、限定された時間でそれが困難であるため、主流的なジャンルを優先的に選定する。具体的に、アクションゲーム（以下ACT）、ロールプレイングゲーム（以下RPG）、シミュレーションゲーム（以下SLG）、アドベンチャーゲーム（以下AVG）の4ジャンルに限定する。それぞれのジャンルの説明と代表例は、表1に示す。

表1 本コーパスで収録するゲームジャンル

ジャンル	説明	代表例
アクションゲーム	キャラクターの行動を操作して、ストーリーを進めていくゲーム	スーパーマリオ、ロックマンX
ロールプレイングゲーム	キャラクターが他のキャラクターから情報を聞き出し、ストーリーを進めていくゲーム	ドラゴンクエスト、ファイナルファンタジー
シミュレーションゲーム	プレイヤーが戦略性を考慮しキャラクターを操作して、ストーリーを進めていくゲーム	ファイアーエムブレム、スーパーロボット大戦
アドベンチャーゲーム	プレイヤーが画面に表示されたテキストを読んで、選択肢を選んでストーリーを進めていくゲーム	ときめきメモリアル、逆転裁判

ただし、「ゼルダの伝説」のように、ACTとRPGの性質が両方揃っており、分類が複数のジャンルにまたがるゲームもある。なお、言語研究であるため、ストーリー性が薄く、最初からテキストがあまりないと分かるジャンル、例えば格闘ゲームは除外している。

2.3 世界観

書籍同様、ゲーム内に現れた言語的表現は、何を描写するかによって大きく左右される。あまり良い例えではないが、書籍の場合、哲学・歴史・芸術・文学などの分類があり、それぞれに現れた言語的特徴も異なる。ゲームの場合は、書籍と同じ分類は難しいが、ゲーム内の世界観である程度分けることができる。暫定的に、中世王道風、近現代風、未来 SF 風の3つに分ける。本コーパスは、それぞれの世界観の作品をなるべくバランスよく収録したいのだが、ゲームの性質上、結果的に中世王道風が多めに入っていることになった。なお、「スターオーシャン」のように、宇宙の探索ができる時代に未発達惑星での冒険ができるなど、分類が難しいゲームもある。表2はゲームの世界観をまとめたものである。

表2 本コーパスで収録するゲーム世界観

世界観	説明	代表例
中世王道風	剣と魔法で作られた中世を時代背景とし、現実的な世界とかけ離れている幻想的な世界	ドラゴンクエスト、テイルズオブファンタジア
近現代風	物理的法則や施設が現実の世界に近く、実在している場所をベースとしているゲームもある	龍が如く、ペルソナ
未来 SF 風	発達した技術によって構築された世界を背景とし、主に機械やロボットが登場している	ロックマンX、スーパーロボット大戦

2.4 発売本数

本コーパスは均衡コーパスではないため、母集団のすべての性格を反映させるよりも、代表性のあるものを反映させるのが目的である。そのため、収録するゲームも人気のあるものを想定している。また、ゲームの日本語教育への応用も筆者が関心のあることであるため、多くの人が興味を持つゲームを考察して得られた知見のほうに教育現場で役に立つと思われる。知名度を定義するには非常に困難であるが、原則的に日本国内での発売本数が10万本以上の作品に限定する。ただし、ゲームジャンルによって、どうしても作品数が揃わない場合は、例外的に発売本数が10万本未満の作品を対象とすることがある。

以上の4観点で選定する予定のゲームと現在の進捗状況は、表3に示す。

表3 選定される予定のゲームと進捗状況

	ジャンル	ゲーム名	発売年代	世界観	状態
前期	ACT	スーパーマリオワールド	1990	中世王道風	未着手
		ゼルダの伝説 神々のトライフォース	1991	中世王道風	未着手
		ロックマン X3	1995	未来 SF 風	済
	RPG	ドラゴンクエスト3 (リメイク)	1996	中世王道風	済
		クロノ・トリガー	1995	中世王道風	作成中
		ファイナルファンタジー7	1997	中世王道風	作成中
		テイルズオブファンタジア	1995	中世王道風	作成中
		マザー2	1994	近現代風	未着手
		スターオーシャン1	1996	未来 SF 風	未着手
		SLG	ファイアーエムブレム 紋章の謎	1994	中世王道風
		第4次スーパーロボット大戦	1995	未来 SF 風	未着手
	AVG	ときめきメモリアル1	1994	近現代風	未着手

後期	ACT	モンスターハンター：ワールド	2018	中世王道風	未着手
		ゼルダの伝説 ブレスオブワイルド	2017	中世王道風	作成中
		龍が如く 7	2020	近現代風	未着手
	RPG	ドラゴンクエスト 11	2017	中世王道風	未着手
		キングダムハーツ 3	2020	中世王道風	未着手
		オクトパストラベラー1	2018	中世王道風	作成中
		ポケットモンスター バイオレット	2022	近現代風	未着手
		ペルソナ 5	2016	近現代風	未着手
		ゼノブレイド 2	2017	未来 SF 風	未着手
	SLG	ファイアーエムブレム 風花雪月	2019	中世王道風	未着手
		FRONT MISSION 1ST (リメイク)	2022	未来 SF 風	未着手
AVG	大逆転裁判 2	2017	近現代風	未着手	

前期と後期のゲームがなるべく同数になるように選定した。ジャンルと世界観に関しては、これまで発売されたゲームの全体的な内訳を見て、偏りが出たのはある程度妥協しなければならないと思われる。

3. 構築方法

本節では、どのようにゲーム内のテキストをコーパスにするかの手順を説明する。直接ゲーム媒体の内部データにアクセスして、文字を抽出する方法が最も作業の手間が省けるが、基本的にゲームの文字コードが PC と異なり暗号化されており、一々解読するのは現実的ではない。また、許可なしに内部データにアクセスすると、法律に違反するおそれもある。そのため、代わりに地味な方法を採用するしかない。具体的には、ゲームをプレイし、テキスト情報を画像に保存しておき、後に一気に文字化する方法である。そのためには、収録環境の整備、作業人員の募集、プレイ動画収録と画像キャプチャー、文字化作業、最終確認、以上の5つのステップが必要となる。

3.1 収録環境の整備

まず、必要なゲーム機器とゲームソフトを用意する必要がある。前期のゲームをプレイするための機器であるスーパーファミコンとプレイステーション 2 は 20 年以上前に発売されたもので、すでに生産終了となっており、中古品を購入した。後期のゲームをプレイするためのニンテンドースイッチとプレイステーション 4 は新品で購入した。ゲームソフトに関しても同様で、基本的に前期のゲームは中古で購入する以外方法はない。

また、コーパスを作成するため、PC も必要である。文字化するだけであれば、スペックの低い PC でも問題ないが、動画を収録するために、ある程度高性能の PC が必要である。収録に使用したのは ASUS Zenbook 14 OLED UX3402ZA (CPU インテル® Core™ i7-1260P、メモリ 16G、ストレージ 512GB) である。なお、録画するために、ゲームの画面を PC 経由で表示させる必要がある。キャプチャーボードを繋げば、PC をテレビ代わりにできるため、録画ソフトで画面の録画と画像キャプチャーが可能である。使用したキャプチャーボードは、AVerMedia Live Gamer EXTREME 2 GC550 PLUS (HDMI 端子対応) である。前期のゲーム機には HDMI 端子が付いていないため、AV 端子を HDMI 端子に変換するコンバーターも必要である。図 1 と図 2 は、HDMI 端子に対応するゲーム機とそうでないゲーム機の収録環境をイメージしたものである。

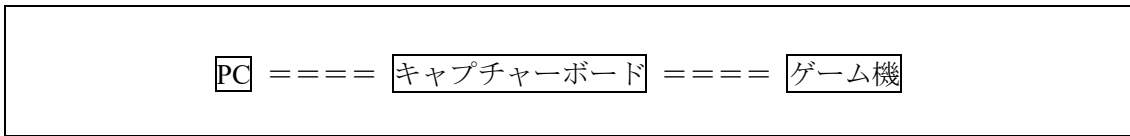


図1 ゲームプレイ動画の収録環境 (HDMI 端子対応ゲーム機の場合)

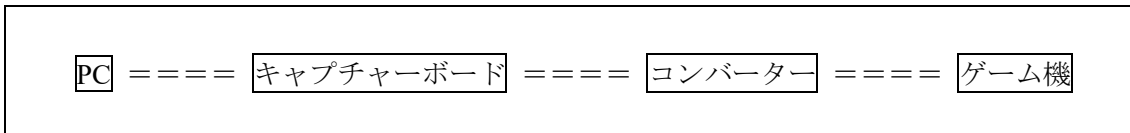


図2 ゲームプレイ動画の収録環境 (HDMI 端子非対応ゲーム機の場合)

使用した録画ソフトは **Bandicam** である。録画の画質設定は **FPS30**、画質 **640*480** (1 時間 **600M** 程度) にした。画像の画質設定は、ゲーム機もともとの解像度に依存するが、ニンテンドースイッチとプレイステーション 4 の場合は **1920*1080** になる (ゲームのシーンにもよるが、平均して 1 枚 **600~800K** 程度)。これで収録環境の整備は完了である。

3.2 作業人員の募集

本コーパスを構築するためにメインとなる作業は、ゲームのプレイ動画収録と画像キャプチャー、及び文字化 (アノテーション情報付与とデータ整形を含む) である。作業の効率性を考慮し、最も時間がかかるプレイ動画収録と文字化作業を中心に、作業人員に依頼することにした。作業者はゲーム経験者であり、かつ日本語の入力ができることが条件となる。実際募集の際に出した条件は、①ゲーム経験者 (特に **RPG**、**SLG**)、②PC での日本語入力、③EXCEL の基本操作、④国籍不問 (ただし非母語話者の場合は日本語能力試験 **N1** 必須) である。現在 2 名体制で作業を進めている。

3.3 動画収録と画像キャプチャー

このステップの作業は、プレイしながら画面のテキスト情報を保存することが目的である。必要な情報はテキスト情報だけであるため、それ以外の画面情報は不要である。この前提であれば、画像キャプチャーだけでもよいのだが、それでも同時に動画を収録させたのは、画像のキャプチャーに失敗した際の保険と、後に文字化する際に文脈 (場面) を確認するためである。時間を節約するため、作業者が当該ゲームの経験者であることが望ましいが、やむを得ずプレイしたことがないゲームを収録・文字化させる場合もある。

テキスト認定は、方針未定の部分があるが、現段階では作業者に下の指示を出している。

- (1) テキストが画面上に出てきたらキャプチャーする。
- (2) できればすべてのルートのすべての文字をテキストする。(攻略サイトを参照)
- (3) 同じ人物には少なくとも 2 回話しかける。(内容が変わる可能性があるから)
- (4) ストーリーが進むと、会話が変わる可能性のある人物に再度話しかける。
- (5) 「はい」「いいえ」のような選択肢は、どちらも選んで、収録する。
- (6) ミッションがあれば、失敗するバージョンも収録する。
- (7) 特定の行動を取る (特定の人物に話しかける、または特定の場所に移動する) とフラグが立ち、他の人物と話せなくなったり、会話内容が変わったりするので、話しかける順番に注意する。

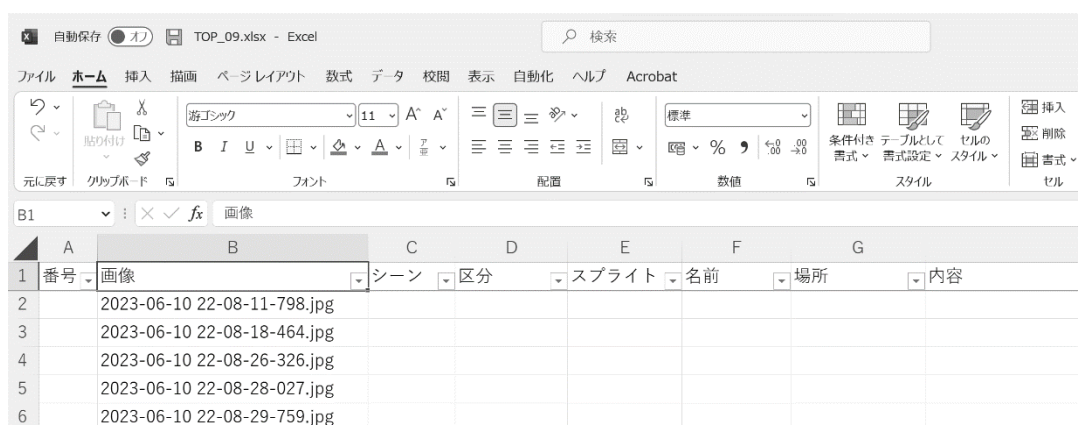
- (8) 新しい町に到着したら、行動をする前に、まずセーブを別データに保存する。(やり直しに備える)

4節で詳述するが、ゲームはプレイヤーの操作によって現れるテキストも異なるため、どこまでこだわるかによってテキスト量と作業の所要時間が変わる。現在は、プレイ動画はなるべくすべての要素を網羅できるように収録してもらい、あとで筆者が文字化データを最終確認の際に必要なものだけを残すことにしている。

3.4 文字化作業

文字化作業に関して、OCR という方法も試したが、ゲームのフォントは書籍と異なり独自のものを使用するものがほとんどで、認識精度が著しく悪かったため、手作業で入力するしかなかった。

文字化作業に使用する EXCEL の記録フォーマットは図 3 の通りである。「シーン」欄は今後、やり取りのある会話を分析するために、筆者が後に付与する情報である。なお、「シーン」欄以外は麻 (2022) で説明されたため、そちらを参照されたい。



番号	画像	シーン	区分	スプライト	名前	場所	内容
2	2023-06-10 22-08-11-798.jpg						
3	2023-06-10 22-08-18-464.jpg						
4	2023-06-10 22-08-26-326.jpg						
5	2023-06-10 22-08-28-027.jpg						
6	2023-06-10 22-08-29-759.jpg						

図 3 文字化作業の EXCEL 記録フォーマット

具体的な手順は、キャプチャーした画像のファイル名をプログラムで「画像」の列に書き込ませた後、作業者に「区分」「スプライト」「名前」「内容」の列を入力させる。「番号」の列は、次の最終確認のステップで、自動で付与する通し番号である。実際に作業者に出示した指示は以下の通りである。

- (1) 一枚の画像を、一行のレコードに入れる。入力が必要な欄は「区分」「スプライト」「名前」「内容」の 4 か所である。
- (2) 「区分」のセルについて、キャラクターの発話は「セリフ」、システムメッセージは「メッセージ」、魔法・アイテム欄の選択候補は「メニュー」と入れる。
- (3) 区分が「セリフ」の場合、発話者は「名前」、発話内容は「内容」のセルに入力する。発話者名が表示されない場合は*を入れる。
- (4) 原則、表記も含め、元テキストを忠実に再現する（会話の最初の“ ”は入力しない）。ただ、ひらがな・カタカナ・英数字・記号はすべて「全角」で入れる。

(5) 改行は「半角スペース」で入力する。

3.5 最終確認

入力が終了したファイルを、筆者が最終確認する。具体的には、①入力ミスの確認、②「シーン」「場所」など必要なアノテーションの追加、③不要な行の削除、以上の3つの作業を行なう。

すべての手順の所要時間に関して、最も時間がかかる RPG というジャンルは、ゲームによってクリア時間が変わるが、平均的に30~40時間で1本クリアできる。ACTの場合は攻略方法さえ分かれば3時間程度でクリアできるゲームもある。文字化作業は約プレイ時間の2~3倍時間かかる。延べ週12時間体制で作成させているため、理論的に文字化作業も含めて2か月にRPGが1本コーパス化できる計算になっている。

4. 今後の課題

本節では、文字化作業を行う際に、実際に遭遇した問題点を述べる。

4.1 対象とするテキストの認定

書籍は、文字が物理的に紙に印刷されているため、意図的に読み飛ばさない限り、すべてのテキストを目にすることができる。一方、ゲームの場合はストーリーの分岐や任意のサブイベントがあるため、プレイヤーの操作次第で、表示されないテキストがある。ここで、ゲーム内にあるテキストをすべて収録する方法（以下、やり込み方式）と、一部のルートのみ収録する方法（以下、一周クリア方式）との2つの選択肢がある。2つの方法で同ゲームを収録するイメージを図4と図5に示す。

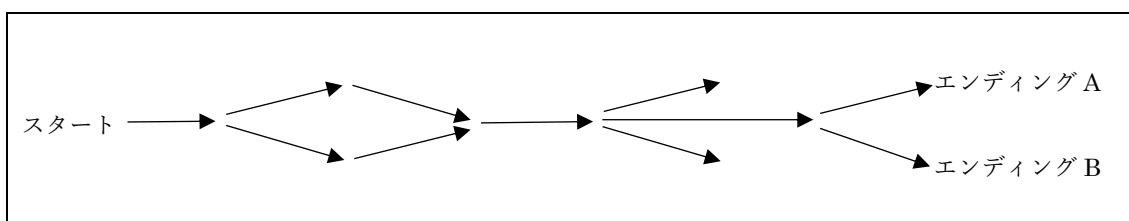


図4 やり込み方式（実線すべて収録）

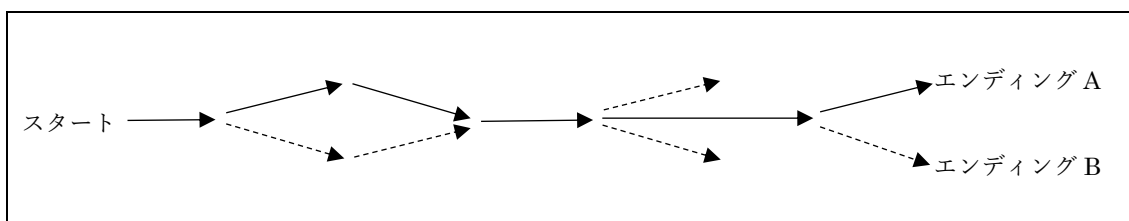


図5 一周クリア方式（実線のみ収録、点線は収録しない）

後期のRPGはやり込み方式にこだわると100時間以上かかるため、本コーパスでは基本的に一周クリア方式を採用する。ただ、前期のゲームは容易にやり込み方式を達成できるため、ゲームによってやり込み方式と一周クリア方式に分けて収録することも考えられる。ただ、一周クリア方式で収録した場合は、客観性と再現性を担保するために、選択肢の選び方やサブイベントの定義について厳密に規定する必要がある。

4.2 キャラクターの特徴の記述

ゲームでは、主人公などの重要人物は名前が付いている一方、一度しか登場しないキャラクターは名前がないことが多い。コーパスでは、後に発話者を特定できるように、当該人物の特徴を記述する（図3の「スプライト」の欄）必要があるが、前期のゲームは解像度が低く、キャラクターの特徴を記述することが困難な場合がある。記述がなくてもできる研究であれば問題ないが、役割語のような社会言語学の研究では、話者の属性が重要であるため、研究に使うには限界がある。

4.3 テキストの分類

現在、テキストの分類として、キャラクターによる発話の「セリフ」と、ナレーションかシステム説明の「メッセージ」と、魔法・道具欄の選択候補の「メニュー」の3種類を定義したが、これらの分類に当てはまらないテキストがある。例えば、キャラクターが心の中で思っている内容がテキストとして表示される場合は現在「セリフ」に入れているが、厳密に言うと「セリフ」ではないため、別項目を立てたほうがよいかもしれない。

もう一つの例は、セリフかどうか判定しにくいパターンである。例えば、キャラクターが本棚を調べたときに、書籍の内容を読み上げることがあるが、この場合、表示されたテキストは確かにそのキャラクターのセリフではあるものの、本質的には書籍の内容であるため、別扱いにする必要があるように思われる。これと類似したパターンは、あるキャラクターが別のキャラクターに憑依した場合の発話である。

4.4 テキスト以外の情報の記録

本コーパスの目的はテキストを収録することであるため、テキスト以外の情報、例えばフォントのサイズ、フォントの色、キャラクターの表情などは収録できていない。この点は書籍のコーパスの場合も同じだが、ゲームではフォントの大きさや色はパラ言語的に使われることがあるため、研究する価値はあると思われる。

4.5 アノテーション情報欄の設定

最後に、ゲームによって必要なアノテーション情報が異なる点が挙げられる。例えば、前期のRPGはスプライトの情報が重要であるが、ACTの場合はあまり意味を成さないため、記録フォーマットに当該欄をあえて設定する必要はない。如何にすべてのゲームに適用する一貫性のある記録フォーマットを設計することかが重要な課題となる。

以上挙げた5つの点は、一部未解決のものもあるため、今後の課題としたい。

謝 辞

本研究はJSPS 科研費若手研究「テレビゲームの日本語教育における可能性の探索とテレビゲームコーパスの構築（課題番号：23K12220）」の助成を受けている。

文 献

麻子軒（2022）「テレビゲームコーパスの構築とその利活用」『言語資源ワークショップ発表論文集』2022, pp.117-126