『日本語ゲームコーパス（JGC）』の構築に関する中間報告：前期のアクションゲームに見られる量的特徴

『日本語ゲームコーパス（JGC)』の構築に関する中間報告 —前期のアクションゲームに見られる量的特徴—

麻 子軒（関西大学）†

Interim Report on the Construction of a Japanese Game Corpus: Quantitative Features Observed in Early Period Action Games

Tzu-Hsuan Ma (Kansai University)

要旨

紙媒体や音声媒体を対象としたコーパスが多く存在する中、ゲームコーパスは技術的な理由によりその構築が困難とされ、長らく言語資源において欠けている存在であった。筆者は、日本の代表的なゲームを前期と後期に分け、それぞれ 12 タイトルを選定し、ゲームコーパスの構築を進めている。本稿では、その構築作業の中間報告として、進捗状況及びテキストの出現環境に対する分類基準を中心に述べた上、既に完成した 3 種のアクションゲームのデータを用いた分析を報告する。具体的には、各ゲームに見られる量的特徴（異なり語数、延べ語数、文長、漢字含有率、特徴語）を明らかにする。また、語彙表も試行的に公開する。

１．はじめに

言語、特に日本語においては、その現れる環境によって語彙や文法だけでなく、様々なレベルで異なる性格を持つ。これらの特徴を観察するために、これまでに書籍をはじめとする多くの媒体に基づく言語資源が作成されてきたが、ゲームコーパスの構築は技術的な理由から遅れている。筆者は、ゲームという環境に現れる日本語の特徴を明らかにするため、2022 年から『日本語ゲームコーパス（Japanese Game Corpus、略称 JGC）』の構築を開始し、その設計方針、方法論、そして研究事例を提示した（麻 2022，2023）。

本稿では、筆者が構築している『日本語ゲームコーパス』の中間報告と、すでに完成した前期の 3 つのアクションゲームにおける統計量から観察できる量的特徴を明らかにする。まず、第 2 節では中間報告として、収録したゲームのタイトルや進捗状況を述べる。次に、第 3 節では前期に対象とした 3 つのアクションゲームについて、異なり語数、延べ語数、文長、漢字含有率、特徴語などの量的特徴を報告する。最後に、語彙表の公開と今後の課題について触れる。

２．中間報告

２.１　これまでの概要

麻（2023）で報告したように、『日本語ゲームコーパス』の構築においては、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を念頭に、発売年代、ジャンル、世界観、発売本数の 4 つの観点から 24 タイトルを選定している。詳細については麻（2023）を参照されたいが、進捗状況を説明する前提として、その概要を以下に簡単にま

† kenji.ma@kansai-u.ac.jp

とめる。

発売年代については、据置ゲーム機の全盛期である 1990～2000 年（以下、前期）と、直近の十年間である 2010～2022 年（以下、後期）の 2 つに分けて収録する。その理由は、2000 年を境にハードの性能や記録媒体の容量が飛躍的に向上し、言語的特徴にも変化が生じることが予測されるため、これらの通時的調査を可能にするからである。ジャンルは、ストーリー性やテキスト量の確保などの観点から、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 つを選定した。世界観については、当初、中世王道風、近現代風、未来 SF 風の 3 種類に分けていたが、複数の世界観が混在しているゲームが多く、完全な分類が困難であるため、現在は特定の世界観に偏らないように選定を行っている。発売本数については、代表性を考慮し、原則として 10 万本以上を目安としている。

２.２　進捗状況

初期段階のみ必要となる環境の整備や作業人員の募集を除けば、構築手順には収録、文字化、校正（テキスト種や話者などのメタ情報付与を含む）の 3 段階があり、それぞれの進捗状況は表 1 に示す通りである。

表 1　収録ゲームと進捗状況

	ジャンル	ゲーム名	発売年代	収録	文字化	校正
前期	ACT	がんばれゴエモン 2	1993	■	■	■
		ゼルダの伝説 神々のトライフォース	1991	■	■	■
		ロックマン X3	1995	■	■	■
	RPG	ドラゴンクエスト 3（リメイク）	1996	■	■	■
		クロノ・トリガー	1995	■	■	
		ファイナルファンタジー7	1997	■	■	□
		テイルズオブファンタジア	1995	■	■	■
		マザー2	1994	■	■	□
		スターオーシャン 1	1996	■	■	
	SLG	ファイアーエムブレム 紋章の謎	1994	■	■	□
		第 4 次スーパーロボット大戦	1995	■	■	
	AVG	ときめきメモリアル 1	1994	■	■	
後期	ACT	モンスターハンター：ライズ	2021	■	■	□
		バイオハザード 4	2005	■	■	■
		龍が如く 1	2005	■	■	□
	RPG	キングダムハーツ 2	2005	■	■	
		大神	2006	■	■	□
		オクトパストラベラー1	2018	■	■	□
		ポケットモンスター バイオレット	2022	■	■	□
		ペルソナ 5	2016	■	■	□
		ゼノブレイド 2	2017			
	SLG	信長の野望 革新	2006			
		トライアングルストラテジー	2022			
	AVG	逆転裁判 3	2004	■	■	□

※ ■は作業済、□は作業中、空欄は未着手

作業量や全体のバランスを考慮し、麻（2023）と比較して、選定したゲームに若干の見直しを行った。当初の計画では、2023～2024 年にゲームの収録を終え、2025～2026 年に文字化及び校正作業（タグ付けを含む）を完了させる予定であった。しかし、特に後期のゲームのテキスト量が膨大で、作業時間が想定以上に必要であり、現時点では 3 タイトルがまだ未収録の状態である。進捗状況によって、未収録の 3 タイトルも今後変更する可能性がある。文字化作業については、当初は手作業で入力していたが、途中から Google の OCR API を利用することで効率が大幅に向上し、収録済みのゲームもほぼ文字化が完了している。ただし、OCR 認識結果の校正及びタグ付けの作業がまだ残っている。この作業は現在、筆者とアルバイトの学生が進めており、2026 年までにはすべての作業を完了する予定である。

２.３　方針の見直し

実際の構築作業において、作業時間の制約や分類の体系性を考慮し、麻（2023）で述べられた方針からいくつか変更を加えた。主な変更点は以下の 3 点である。

1 点目は、２．１節で述べたように、世界観を選定基準の必須条件から外し、全体のバランスが偏らないようにするための参考条件としたことである。これは、ゲームによっては世界観の分類が困難であるためである。

2 点目は、テキストの認定（プレイルート）についてである。当初は、ゲームごとにゲーム内のテキストをすべて収録する「やりこみ方式」と、一部のルートのみを収録する「一周クリア方式」のいずれかを採用していた。しかし、２．２節で述べたように、後期のゲームで作業量が飛躍的に増加することから、すべてのゲームで「一周クリア方式」を採用する方針に変更した。

3 点目は、テキスト種の分類方法についてである。麻（2023）では、テキストを「キャラクターのセリフ」「ナレーションまたはシステム説明のメッセージ」「魔法・道具欄の選択肢としてのメニュー」の 3 種類に分類していた。しかし、これらの分類に当てはまらないテキストが存在するため、現在は表 2 に示すように、まずは「ストーリー」「システム通知」「メニュー表示」の 3 つに大分類し、その後に詳細な分類を行っている。

表 2　テキスト分類の新基準

	分類名	定義
ストーリーに関する内容	セリフ	キャラクターによる声に出す発話
	心内発話	キャラクターによる声に出ない思考内容
	ナレーション	画面外の人物による物語の説明
	テキスト	ゲーム内の書籍や掲示物の内容
システム通知に関する内容	メッセージ	システムから通知された状況確認・説明・指示
メニュー表示に関する内容	要素名	人物名、道具名、魔法名、特技名、能力値の名称など
	コマンド	「話す」「調べる」など、アクションを実行するための指令
	オプション	プレイヤーに提示された選択肢となる項目
	説明文	要素名・コマンド・オプションに対する説明内容

３．前期 3 アクションゲームに見られる量的特徴

本節では、すでに構築が完成した前期の 3 つのアクションゲーム、『がんばれゴエモン 2（以下、「ゴエモン 2」とする）』『ロックマン X3』『ゼルダの伝説 神々のトライフォース（以下、「ゼルダの伝説」とする）』について、異なり語数、延べ語数、文長、漢字含有率、特徴語などの量的特徴を述べる。なお、『ゴエモン 2』と『ロックマン X3』は漢字表記が少なく、形態素解析の誤解析が多かったため、３．３節の考察を除き漢字に変換したデータを用いた。使用した解析器は MeCab 0.996、辞書は UniDic-MeCab 2.1.2 である。

３.１　収録語数

表 3 では、句読点や括弧などの補助記号を除いた実質語と機能語の合計収録語数を、短単位と長単位に分けて、延べ語数と異なり語数で示している。

表 3　ゲーム別の収録語数（実質語＋機能語）

	ゴエモン 2				ロックマン X3				ゼルダの伝説			
	短単位		長単位		短単位		長単位		短単位		長単位	
	延	異	延	異	延	異	延	異	延	異	延	異
セリフ	3564	946	3023	992	1784	457	1526	474	6086	973	5405	1081
ナレーション	18	11	16	9	230	121	176	103	200	106	181	103
テキスト	-	-	-	-	-	-	-	-	292	156	253	140
メッセージ	170	104	92	74	69	49	18	18	909	319	820	303
要素名	-	-	-	-	74	19	20	10	54	39	45	32
コマンド	-	-	-	-	-	-	-	-	7	7	6	6
オプション	298	96	162	86	21	20	13	12	166	73	141	73
合計	**4050**	**1157**	**3293**	**1161**	**2178**	**666**	**1753**	**617**	**7714**	**1673**	**6851**	**1738**

参考までに、表 3 から助詞・助動詞を除外した実質語のみの語数を表 4 に示す。

表 4　ゲーム別の収録語数（実質語のみ）

	ゴエモン 2				ロックマン X3				ゼルダの伝説			
	短単位		長単位		短単位		長単位		短単位		長単位	
	延	異	延	異	延	異	延	異	延	異	延	異
セリフ	2053	863	1565	887	1050	412	802	411	3376	896	2791	977
ナレーション	15	9	13	7	140	100	92	77	108	86	90	78
テキスト	-	-	-	-	-	-	-	-	200	126	161	108
メッセージ	154	96	76	66	69	49	18	18	523	284	442	262
要素名	-	-	-	-	74	19	20	10	44	37	35	30
コマンド	-	-	-	-	-	-	-	-	7	7	6	6
オプション	273	87	137	77	21	20	13	12	130	62	105	57
合計	**2495**	**1055**	**1791**	**1037**	**1354**	**600**	**945**	**528**	**4388**	**1498**	**3630**	**1518**

表 4 の全体の傾向は表 3 とそれほど変わっていないため、以下では表 3 を中心に説明する。表 3 の長単位の延べ語数では、『ロックマン X3』が 1,753 語と最も少なく、『ゼルダの伝説』が 6,851 語と最も多い。後者の文量は、小説に例えればおおよそ短編小説に相当する。前期のゲームは容量の制約からテキスト量が少ない傾向があるが、現在の作業の感触では、後期のゲームになるとその量が数倍に増加している。また、今回取り上げたゲームがアクションゲームであることも、テキスト量が少ない要因の一つである。アクションゲームはストーリーの描写よりもプレイヤーの操作で楽しむことに重きを置いており、特にアクション性の強い『ロックマン X3』の語数が少ないのはその証拠と言える。今後、ストーリー性の強いロールプレイングゲームと比較することで、この特徴はさらに顕著になるだろう。なお、3 つのゲームに共通して観察されたのは、セリフ、ナレーション、オプションである。『ゼルダの伝説』は、ロールプレイングゲームの要素も少し含んでおり、テキストやコマンドといった分類が見られる。

ここでは、それぞれのテキスト出現環境における語彙多様性も観察する。実質語のみを示した表 4 の短単位における延べ語数と異なり語数をもとに計算された TTR（Type-Token Ratio）が図 1 である。

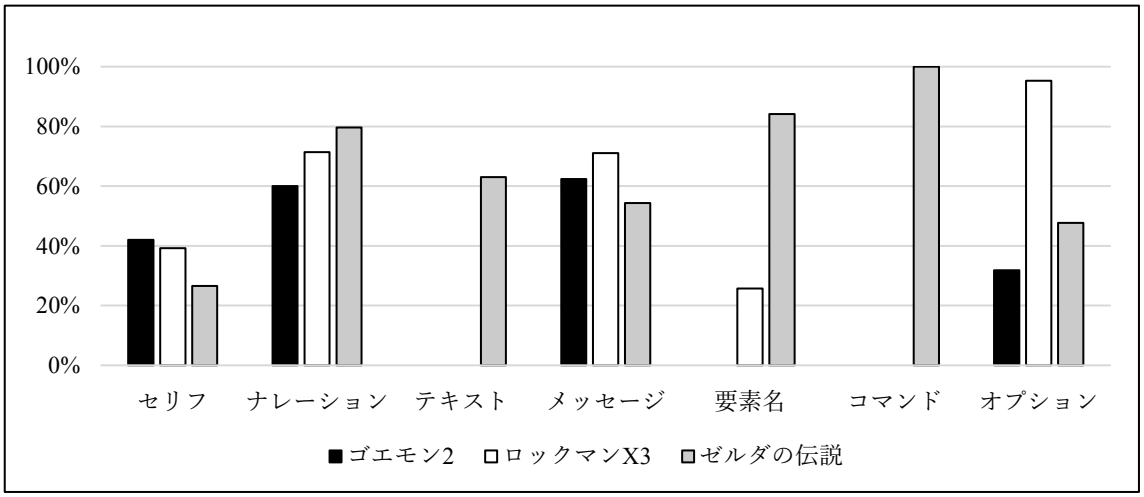

図 1　テキスト出現環境別の TTR

セリフは、物語や世界観を表す語彙が頻繁に繰り返されるため、TTR はどのゲームでも 30～40%程度と最も低い数値となっている。オプションについても、語数が少ない『ロックマン X3』を除き、「はい」「いいえ」のように、バリエーションがそれほど多くないが、繰り返しが多いため、TTR は低い数値となっている。一方、要素名やコマンドはメニューの性質上、同じ表現が繰り返し使われることが少なく、TTR は 80～100%と高い数値である。ナレーションとメッセージは、それらの間に位置している。

３.２　平均文長

平均文長は、全角と半角の区別をせず、一文あたりの文字数で計算した。ただし、スペースは含めていない。文の区切りは便宜上「。」「！」「？」「！？」「！！」を基準としたが、厳密に行ったわけではない。以上の条件で全ての文を分割し、その平均文字数を計算した結果が表 5 である。

表 5　ゲーム別の平均文長

	ゴエモン 2	ロックマン X3	ゼルダの伝説
セリフ	14.47	21.41	20.62
ナレーション	9.5	33	31.27
テキスト	-	-	18.48
メッセージ	6.85	16.61	15.59
要素名	-	6.25	5.72
コマンド	-	-	3
オプション	5.45	6.55	5.08

小説の平均文長は一般的に 20〜30 字とされている（工藤他 2010）。また、複数の話者がやり取りするセリフよりも、単独の話者が一方的に語るナレーションのほうが長くなることは容易に想像できる。表 5 では、『ロックマン X3』と『ゼルダの伝説』は小説と類似した傾向を示している。一方で、『ゴエモン 2』はセリフの文長が比較的短いが、これはゲームの特性上、画面に一度に表示できるテキストの量が限られているため、文の途中での改ページを極力避けるように意識している可能性がある。ただし、同じことは他の 2 つのゲームにも言えるため、これはゲーム制作者の方針の違いによるものかもしれない。なお、『ゴエモン 2』のナレーションの文長が極端に短いのは、文量が少ないことが影響していると考えられる（表 3 も合わせて参照されたい）。

表 5 の結果を視覚的に比較しやすくするため、図 2 ではそれを棒グラフとして示すことにする。

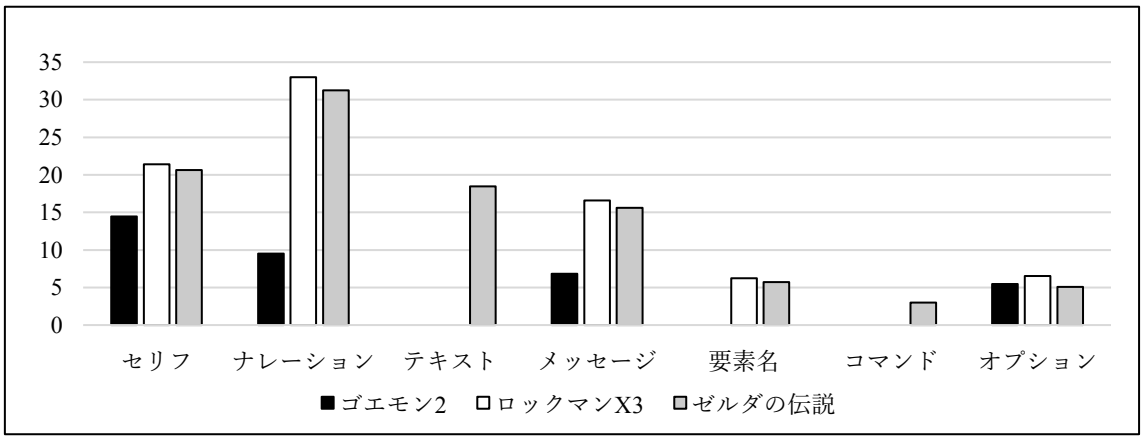

図 2　ゲーム別の平均文長を表した棒グラフ

全体的に見ると、セリフとナレーションの文長が長いことが観察できる。他の分類については、オプションは提示された選択肢であるため基本的に短くなっている。また、コマンドや要素名は通常、文ではなく、道具名や魔法名の 1 語のみで構成されているため、平均文長が短くなるのは予想通りである。

３.３　漢字含有率

本節では、文章中に占める漢字の割合について考察する。集計時には、句読点などの補助

記号を除外し、ひらがな・カタカナ・漢字・英数字の合計を分母とした。また、第 3 節の冒頭で述べたように、同じ条件で形態素解析を行うために、『ゴエモン 2』と『ロックマン X3』については漢字変換データを使用したが、本節の考察に使用したのは元の表記のオリジナルデータである。

表 6　ゲーム別の漢字含有率

	ゴエモン 2	ロックマン X3	ゼルダの伝説
セリフ	1%（51/7395）	0%（0/3879）	15%（1566/10614）
ナレーション	5%（2/42）	0%（0/543）	23%（75/329）
テキスト	-	-	18%（97/529）
メッセージ	8%（36/447）	0%（0/265）	15%（247/1649）
要素名	-	0%（0/134）	0%（0/143）
コマンド	-	-	0%（0/18）
オプション	9%（46/504）	0%（0/67）	5%（20/376）

※ 括弧内は「漢字数/文字数」

前期のゲームでは、記録媒体の容量が不足していたため、極力漢字を使用せず、文字種が少ないひらがなやカタカナのみで表記されることが多かった。特に『ドラゴンクエスト 1』のように、容量を節約するためカタカナの使用まで制限されているゲームもあった。表 6 でも同様の傾向が見られる。さらに観察しやすいように、図 3 に棒グラフも示す。

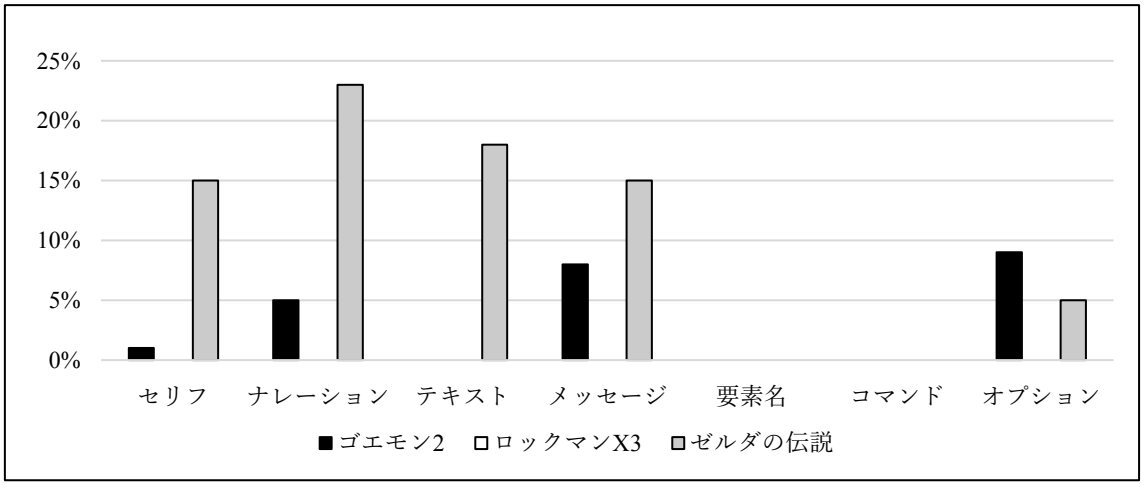

図 3　ゲーム別の漢字含有率を表した棒グラフ

図 3 を見ると、『ロックマン X3』には一切漢字が使用されていないことが分かる。一方、『ゼルダの伝説』では、セリフとナレーションにおいて漢字の含有率が 15～25%である。菅野（2017）の調査によると、1990 年代の小説における漢字比率は 25～30%程度とされている。今回の調査は、前期のゲームにおける漢字の使用率が、同時代の小説よりも低い可能性を示唆している。その理由は、調査したゲームの対象年齢が小説よりも低いことに関連しており、難しい漢字をあえて使わないようにしているためと考えられる。

もう一点注目すべきは、他のテキスト種で漢字が使用されている『ゼルダの伝説』においても、コマンドと要素名の漢字含有率だけが 0%であることである。これは『ゼルダの伝説』のみならず、一部の前期のゲームに共通して見られる特徴である。その理由は、コマンドと要素名の表示される場所が、セリフやナレーションといった「テキストフレーム」ではなく、別に開かれた「メニュー」にあることと関連していると思われる。メニューはテキストフレームとは異なるデザインであることが多かったため、別途フォントを用意しなければならず、その結果、字形データの容量や制作の手間が 2 倍になることがある。これが原因で、前期のゲームでは、セリフやナレーションに漢字が使用されていたとしても、メニュー内のコマンドや要素名はひらがなやカタカナで表記されることが多かったと推測される。

３.４　特徴語

最後に、それぞれのゲームに見られる特徴語について考察する。抽出に関しては、ゲームの世界観を最もよく反映するセリフとナレーションを対象とし、3 つのゲームを互いに対照データとして計算した対数尤度比が最も高い上位 5 語を名詞と動詞ごとに列挙する方法を取る。まず、名詞の特徴語を表 7 に示す。なお、人名は除外し、代名詞や接辞、記号類は名詞に含めている。

表 7　ゲーム別の特徴語（名詞）

順位	ゴエモン 2	ロックマン X3	ゼルダの伝説
1	大江戸城	カプセル	貴方
2	町	俺	力
3	メカ	ボディ	勇者
4	祭り	貴様	トライフォース
5	両	イレギュラーハンター	剣

抽出された特徴語は、各ゲームの世界観をよく反映していると言える。『ゴエモン 2』は江戸時代を背景にした和風ゲームであり、「大江戸城」「祭り」「両（通貨の単位）」など、日本の特徴を表す語が多く含まれている。『ロックマン X3』は未来を舞台にしたゲームで、「カプセル」「ボディ」「イレギュラーハンター」といったロボットに関連する語が上位にランクインしている。『ゼルダの伝説』は中世の王道風ゲームで、「勇者」「剣」などが特徴語となっている。

次に、動詞の特徴語を表 8 に示す。

表 8　ゲーム別の特徴語（動詞）

順位	ゴエモン 2	ロックマン X3	ゼルダの伝説
1	御座る	授ける	聞こえる
2	いらっしゃる	受ける	開く
3	飛ぶ	追加する	叶える
4	来る	入る	導く
5	勝負する	戦う	引く

『ゴエモン 2』の「御座る」は、登場キャラクターである忍者が使う語尾であり、「いらっしゃる」は店に入る際の「いらっしゃいませ」、「勝負する」は博打屋でのセリフに由来している。これらの表現は、対照データとして用いた他の 2 つのゲームには見られない要素である。『ロックマン X3』においては、「授ける」「追加する」がいずれもロボットを強化する際の表現であり、「戦う」はゲームの中核をなす概念であるため、頻繁に登場している。『ゼルダの伝説』では、「聞こえる」「叶える」「導く」は賢者が勇者にヒントを与える際のセリフに使用されている。全体的に見て、「飛ぶ」「来る」「入る」「開く」以外の動詞は、それぞれのゲームの世界観をよく反映した言葉であると言える。

４．終わりに

４.１　語彙表の公開

著作権の都合上、コーパスの全文を公開することはできないが、語彙表は筆者のホームページ（https://kenjima.net/）で公開している。公開している語彙表はテキスト種がセリフとナレーションを対象として解析された短単位データと長単位データの 2 種類があり、現段階では本稿で取り上げた 3 つのアクションゲームの語彙表のみとなっているが、今後、他のゲームの文字化と校正が完了次第、順次追加公開する予定である。

４.２　今後の課題

今回の考察は主に 3 つのゲーム間の比較に焦点を当てており、外部資料との比較はそれほど行っていない。今後は、小説や漫画、日常会話コーパスなど、他の言語資源を参照することで、ゲームに特有の言語的表現が明らかになると考えられる。また、品詞構成、語種構成や語彙難易度など、他にも様々な計量的指標が存在するが、それらについての考察は今後の課題とする。

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

菅野倫匡（2017）「漢字は無くなるのか―再び「漢字の将来」を問い直す―」計量国語学, 30:8, pp.481-498.

工藤彰・村井源・徃住彰文（2010）「計量分析による村上春樹文学の語彙構成と歴史的変遷」情報知識学会誌, 20:2, pp.135-140.

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.

麻子軒（2023）「ゲームコーパスの設計方針と構築方法」言語資源ワークショップ発表論文集, 2023, pp.151-158.