

『日本語ゲームコーパス』前期データの概要と出現語彙リストの試行公開

麻 子軒（京都外国語大学）[†]

Overview of Early-Period Data in the Japanese Game Corpus and Trial Release of Vocabulary Occurrence Lists

Tzuhsuan Ma (Kyoto University of Foreign Studies)

要旨

筆者が構築を進めている『日本語ゲームコーパス（JGC: Japanese Game Corpus）』のうち、1990年代発売のゲームを中心とする「前期データ」について、4ジャンル（ACT, RPG, SLG, AVG）計12タイトルの概要と量的特徴を報告する。ストーリーに関するテキストを対象に集計した結果、全体で異なり13,702短単位、延べ408,302短単位が得られた。本稿では、タイトル別の延べ・異なり語数、Guiraud's R、品詞構成比、語種構成比を示し、さらに、全12タイトルを対象として対数尤度比による特徴語分析を行い、作品ごとの世界観・物語内容・言語表現の特徴を確認する。これらを通じて、ゲームテキストの語彙的・文体的特徴を概観するとともに、他のレジスターとの比較に向けた基礎資料を提示する。また、試行的なデータ公開として、セリフ・心内発話・ナレーション・テキストの出現環境別に、語彙素・語彙素読み・品詞を付与した出現語彙リストを公開する。

1. はじめに

言語は、媒体、ジャンル、話題など、その出現環境によって語彙的・文体的特徴が異なる。これらの違いを検証するために、書籍・雑誌などを対象とする書き言葉コーパスや、講演・日常会話などを対象とする話し言葉コーパスが構築されてきた。また、近年では、LINEやXなどのSNSに見られる打ち言葉も研究対象として注目され、言語資料としての整備が進んでいる（落合 2026）。

一方、テレビゲームには物語を担うセリフの他、ゲーム内の掲示物、システム通知など多様なテキストが含まれており、言語資源として大きな可能性を有しているにもかかわらず、収録・文字化に伴う技術的・作業上の制約や著作権上の制約から体系的な言語資源の整備が遅れてきた。筆者は2022年から『日本語ゲームコーパス（JGC: Japanese Game Corpus）』の構築を進め、その設計方針と構築方法を報告してきた（麻 2022, 2023）。また、麻（2024）では前期ACT3タイトルを対象に中間報告と量的分析を行った。本稿では対象を前期12タイトル全体に拡大し、データ規模、品詞・語種構成等を概観するとともに、研究利用可能な出現語彙リストの試行公開について報告する。

2. 『日本語ゲームコーパス』について

2.1 概要

『日本語ゲームコーパス』（以下、JGC）は、ゲームという環境に現れる日本語の特徴を明らかにし、他のレジスターとの比較や日本語教育への応用につなげることを目的とする

[†] m_tzuhsu@kufs.ac.jp

コーパスである。収録作品は発売年代を二つの時期に分け、各時期 12 タイトル、計 24 タイトルを対象とし、ジャンル、作品の代表性、テキスト量などを考慮して選定している（麻 2023, 2024）。ジャンルは、アクションゲーム（ACT）、ロールプレイングゲーム（RPG）、シミュレーションゲーム（SLG）、アドベンチャーゲーム（AVG）の 4 種である。

構築作業は、ゲームを一周クリアするルートを収録し、画面内に現れたテキストを文字化した上で、原文との照合・校正とメタ情報付与を行う手順で進めている。テキスト量の多い作品では OCR も利用する（麻 2025）が、最終的には人手で確認する。さらに、テキストの出現環境を分類し、形態論情報を付与して短単位・長単位の集計データを作成する。

進捗状況としては、現在、前期 12 タイトルの構築作業を終え、後期は残り 3 タイトルの作業を進めている。本稿では、すでに完成している前期データを扱う。前期データを一括して示すことで、麻（2024）の ACT3 タイトルを対象とした分析から一歩進め、作品間・ジャンル間の比較に必要な基礎的なデータ規模と構成を確認する。

2.2 前期の収録ゲーム

本稿で扱う「前期データ」は、表 1 に示す 1991～1997 年発売の 4 ジャンル 12 タイトルである。内訳は、ACT3 タイトル、RPG6 タイトル、SLG2 タイトル、AVG1 タイトルである。表では、今後本稿で言及する際のゲーム略称も付記した。

表 1 前期の収録ゲーム

ジャンル	収録ゲーム	略称	発売年
ACT	がんばれゴエモン 2	GM2	1993
	ゼルダの伝説 神々のトライフォース	ZLD1	1991
	ロックマン X3	X3	1995
RPG	ドラゴンクエスト 3（リメイク）	DQ3	1996
	クロノ・トリガー	CT	1995
	ファイナルファンタジー7	FF7	1997
	テイルズオブファンタジア	TOP	1995
	マザー2	MT2	1994
	スターオーシャン 1	SO1	1996
SLG	ファイアーエムブレム 紋章の謎	FE	1994
	第 4 次スーパーロボット大戦	SRW4	1995
AVG	ときめきメモリアル 1	TK1	1994

2.3 テキストの分類基準

ゲームのテキストには多くの種類が存在するが、本稿におけるテキストの分類は麻（2024）で見直した基準を用いる（表 2）。大分類は「ストーリーに関する内容」「システム通知に関する内容」「メニュー表示に関する内容」の 3 種であり、さらにセリフ、心内発話、ナレーション、テキスト、メッセージ、要素名、コマンド、オプション、説明文に細分する。

本稿の分析では、このうち、物語の進行と直接関係する「ストーリーに関する内容」の 4 種（セリフ・心内発話・ナレーション・テキスト）のみを対象とする。その理由は、これらはゲームの中核的な言語データであり、一周クリアの進行に沿って収集範囲を比較的明確に定められるためである。

表 2 テキスト分類基準

	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. 前期データに見られる量的特徴

量的特徴を分析するための前処理として、前期 12 タイトルのストーリー関連テキストに形態論情報を付与し、空白・記号・補助記号を除いた上で、短単位・長単位で出現頻度を集計した。形態論解析データと集計データの例をそれぞれ図 1、図 2 に示す。

ゲーム	区分	文境界	短単位出現形	短単位発音形	短単位語彙表	短単位品詞	短単位活用型	短単位活用形	語種	語彙表番号	長単位品詞	長単位活用型	長単位活用形	長単位語彙表	長単位出現形
FF7	セリフ	B	待つ	マッ	マツ	待つ	動詞-一般	五段-タ行	適用形-促音便	和	35830 動詞-一般	五段-タ行	適用形-促音便	マツ	待つ
FF7	セリフ		て	テ	テ	て	助詞-接続助詞			和	24874 助詞-接続助詞	*	*	テ	て
FF7	セリフ		、				補助記号-読点			記号	24 補助記号-読点	*	*	、	、
FF7	セリフ		クラウド	クラウド	クラウド	クラウド-crowd	名詞-普通名詞-一般			外	115399 名詞-普通名詞-一般	*	*	クラウド	クラウド
FF7	セリフ		!			!	補助記号-句点			記号	45 補助記号-句点	*	*	!	!
FF7	セリフ	EOS													
FF7	セリフ	B	クラウド	クラウド	クラウド	クラウド-crowd	名詞-普通名詞-一般			外	115399 名詞-普通名詞-一般	*	*	クラウド	クラウド
FF7	セリフ		!			!	補助記号-句点			記号	45 補助記号-句点	*	*	!	!
FF7	セリフ	EOS													
FF7	セリフ	B	なに	ナニ	ナニ	何	代名詞			和	27920 代名詞	*	*	ナニ	何
FF7	セリフ		を	オ	ヲ	を	助詞-格助詞			和	41407 助詞-格助詞	*	*	ヲ	を
FF7	セリフ		し	シ	スル	ある	動詞-非自立可能サ行変格	適用形-一般	適用形-一般	和	19537 動詞-一般	サ行変格	適用形-一般	スル	する
FF7	セリフ		て	テ	テ	て	助詞-接続助詞			和	24874 助詞-一般	上一段-ア行	連体形-一般	テイル	ている
FF7	セリフ		いる	イル	イル	居る	動詞-非自立可能上一段-ア行	連体形-一般	連体形-一般	和	2585 *	*	*	*	*
FF7	セリフ		の	ノ	ノ	の	助詞-準体助詞			和	28990 助動詞	助動詞-ダ	終止形-一般	ノダ	のだ
FF7	セリフ		だ	ダ	ダ	だ	助動詞	助動詞-ダ	終止形-一般	和	22916 *	*	*	*	*
FF7	セリフ		?			?	補助記号-句点			記号	57 補助記号-句点	*	*	?	?
FF7	セリフ	EOS													

図 1 形態素解析データの一部

語彙素	語彙素読み	品詞	語種	GM2	ZLD1	X3	DQ3	CT	FF7	TOP	MT2	SO1	FE	SRW4	TK1	総計
男	オトコ	名詞-普通名詞-一般	和		0	3	0	32	25	62	20	42	10	6	16	4
男の子	オトコノコ	名詞-普通名詞-一般	和		0	0	0	3	1	1	0	1	0	0	0	8
男物	オトコモノ	名詞-普通名詞-一般	和		0	0	0	0	0	1	0	0	0	0	0	1
脅し	オドシ	名詞-普通名詞-一般	和		0	0	0	0	0	0	0	0	0	0	3	3
落とし	オトシ	名詞-普通名詞-一般	和		0	0	0	0	0	1	0	0	0	0	4	5
落とし穴	オトシアナ	名詞-普通名詞-一般	和		0	0	0	0	0	0	0	0	1	0	0	1
陥れる	オトシイレル	動詞-一般	和		0	0	0	0	0	0	0	0	1	1	0	2
脱める	オトシメル	動詞-一般	和		0	0	0	0	0	1	0	0	0	0	0	1
脅す	オドス	動詞-一般	和		0	0	0	0	0	1	0	0	0	0	0	1
落とす	オトス	動詞-一般	和		0	1	0	11	4	6	6	7	4	1	3	43
訪れ	オトズレ	名詞-普通名詞-一般	和		0	0	0	0	0	0	0	1	0	0	0	1
訪れる	オトズレル	動詞-一般	和		0	0	0	6	2	2	13	1	6	1	0	31

図 2 集計データの一部

以下では、語彙規模、品詞構成比、語種構成比の順に全 12 タイトルを概観し、最後に各ゲームの特徴語を示す。なお、本稿では、分析目的と外部データとの比較可能性を考慮し、短単位・長単位、および異なり語数・延べ語数を分析ごとに使い分ける。

3.1 語彙規模

語彙規模の把握には短単位を用い、異なり語数と延べ語数の双方を示す。短単位を採用する理由は、UniDic に基づく基本的な形態論単位であり、今後比較対象となりうる CEJC や BCCWJ を用いた多くの先行研究でも短単位で語数が示されているためである。

短単位の異なり語数を表 3、延べ語数を表 4 に示す。全 12 タイトルを合わせると、ストーリー関連テキストは異なり 13,702 語、延べ 408,302 語である。CEJC は異なり 29,534 語・延べ 2,419,171 語（『日本語日常会話コーパス』短単位語彙表・語数表 ver.2022.09）、BCCWJ1 の小説会話文は異なり 48,150 語・延べ 3,097,209 語（山崎 2022）である。JGC 前期データは CEJC の約 6 分の 1 の延べ語数で、異なり語数は CEJC の約 46% である。参考までに、語数差の影響をある程度抑える Guiraud's R（異なり語数／延べ語数の平方根）を算出すると、JGC は 21.44 であり、CEJC の 18.99 と BCCWJ1 小説会話文の 27.36 の中間に位置する。

作品別に見ると、延べ語数では『ファイナルファンタジー7』が 89,079 語で最も多い一方、『ロックマン X3』は 1,960 語にとどまり、作品差が大きい。全体として、登場人物の対話を通じて展開するストーリーを主要な魅力とする RPG は語数が多く、プレイヤーの操作テクニックを重視する ACT は語数が少ない傾向が見られる。ただし、『ファイアーエムブレム 紋章の謎』と『第 4 次スーパーロボット大戦』のように、同じ SLG でも語数に大きな差があり、ジャンルだけでなく物語の提示方法や作品設計の影響も大きい。

表 3 短単位出現度数（異なり語数）

	セリフ	心内発話	ナレーション	テキスト	全体
GM2	953	0	11	0	957
ZLD1	974	0	107	153	1,034
X3	456	0	118	0	510
DQ3	2,798	0	69	131	2,819
CT	3,661	0	7	315	3,702
FF7	5,145	203	0	598	5,273
TOP	3,904	142	140	466	4,038
MT2	3,858	561	219	766	4,191
SO1	3,001	88	38	187	3,045
FE	1,396	5	715	0	1,698
SRW4	4,590	151	255	0	4,652
TK1	1,358	486	1	0	1,497
総計	13,250	1,132	1,209	1,824	13,702

表 4 短単位出現度数（延べ語数）

	セリフ	心内発話	ナレーション	テキスト	全体
GM2	3,557	0	18	0	3,575
ZLD1	6,066	0	200	289	6,555
X3	1,738	0	222	0	1,960
DQ3	40,932	0	128	228	41,288
CT	40,072	0	7	681	40,760
FF7	86,948	432	0	1,699	89,079
TOP	50,786	288	277	1,695	53,046
MT2	39,137	1,990	568	2,256	43,951
SO1	27,730	146	50	332	28,258
FE	9,828	7	2,504	0	12,339
SRW4	58,528	279	550	0	59,357
TK1	21,870	6,253	11	0	28,134
総計	387,192	9,395	4,535	7,180	408,302

テキスト種の構成に関して、全体において延べ語数の 95% (387,192 語) をセリフが占め、書籍の地の文と最も機能的に近いナレーションは全体として少ない。ただし、『ファイアーエムブレム 紋章の謎』ではナレーションが全体の約 20%, 『ときめきメモリアル 1』では心内発話が約 22%を占めるなど、物語の提示方法そのものにも作品差が表れている (図 3)。

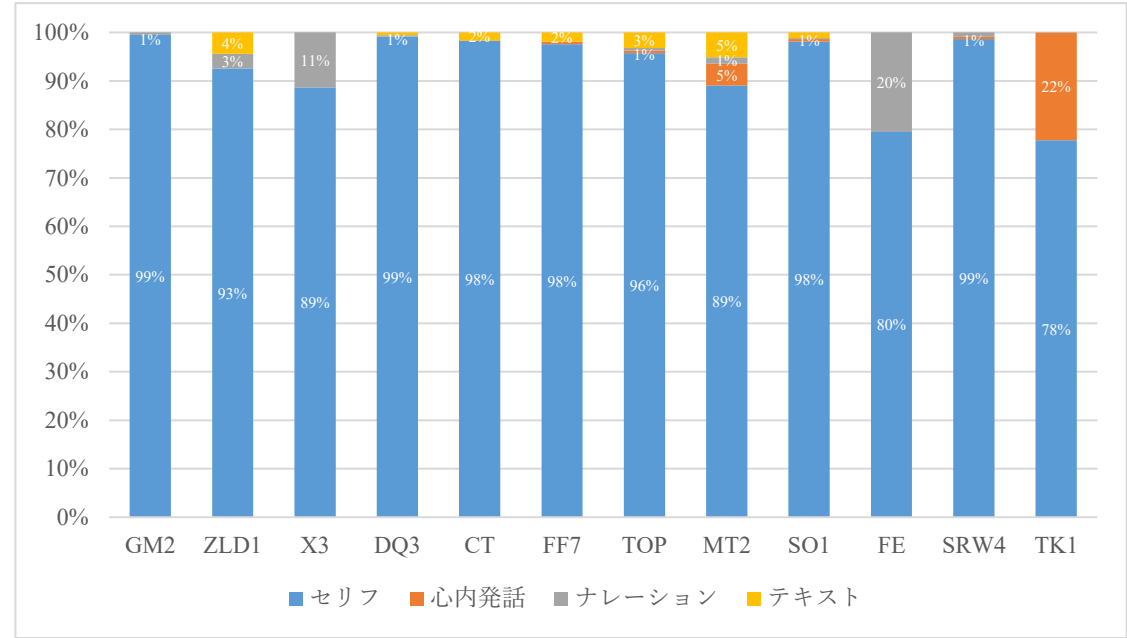


図 3 テキスト別・短単位出現比率 (延べ語数)

語彙豊富度に関して、ゲーム間の標本サイズに大きな差があるため、語数差の影響を受けない Guiraud's R (異なり語数/延べ語数の平方根) を採用する。その結果は表 5 である。ACT の 3 作品はいずれも 11.52~16.01 と比較的低いのに対し、RPG では DQ3 (13.87) を除く 5 作品が 17.53~19.99 に位置しており、RPG のほうが多様な語彙を使用する傾向がうかがえる。なお、TK1 はデートに誘う電話など類似した会話が繰り返し現れるため、同じ語彙が反復して使用され、Guiraud's R が低くなったと考えられる。

表 5 短単位 Guiraud's R

	セリフ	心内発話	ナレーション	テキスト	全体
GM2	15.98	-	2.59	-	16.01
ZLD1	12.51	-	7.57	9.00	12.77
X3	10.94	-	7.92	-	11.52
DQ3	13.83	-	6.10	8.68	13.87
CT	18.29	-	2.65	12.07	18.34
FF7	17.45	9.77	-	14.51	17.67
TOP	17.32	8.37	8.41	11.32	17.53
MT2	19.50	12.58	9.19	16.13	19.99
SO1	18.02	7.28	5.37	10.26	18.11
FE	14.08	1.89	14.29	-	15.29
SRW4	18.97	9.04	10.87	-	19.09
TK1	9.18	6.15	0.30	-	8.92
総計	21.29	11.68	17.95	21.53	21.44

3.2 品詞構成比

図4は、各タイトルについて短単位の延べ語数に基づく品詞構成比を示したものである。短単位を用いるのは、山崎（2022）によるCEJCとBCCWJ小説会話文の分析と比較できるようにするためである。また、異なり語数ではなく延べ語数を用いるのは、語彙の種類としてどの品詞が存在するかではなく、実際のテキストに各品詞がどの程度の割合で現れるかという文体的特徴を捉えるためである。

主要な品詞を見ると、名詞は19～26%、動詞は12～16%、助詞は29～34%の範囲にあり、山崎（2022）では、CEJCが名詞17%、動詞11%、助詞29%、BCCWJ1小説会話文が名詞22%、動詞14%、助詞34%である。JGCの値は若干BCCWJ1小説会話文に近く、少なくともこれらの主要品詞については、自然な日常会話よりも書かれたフィクション会話に近い側面を持つことがうかがえる。

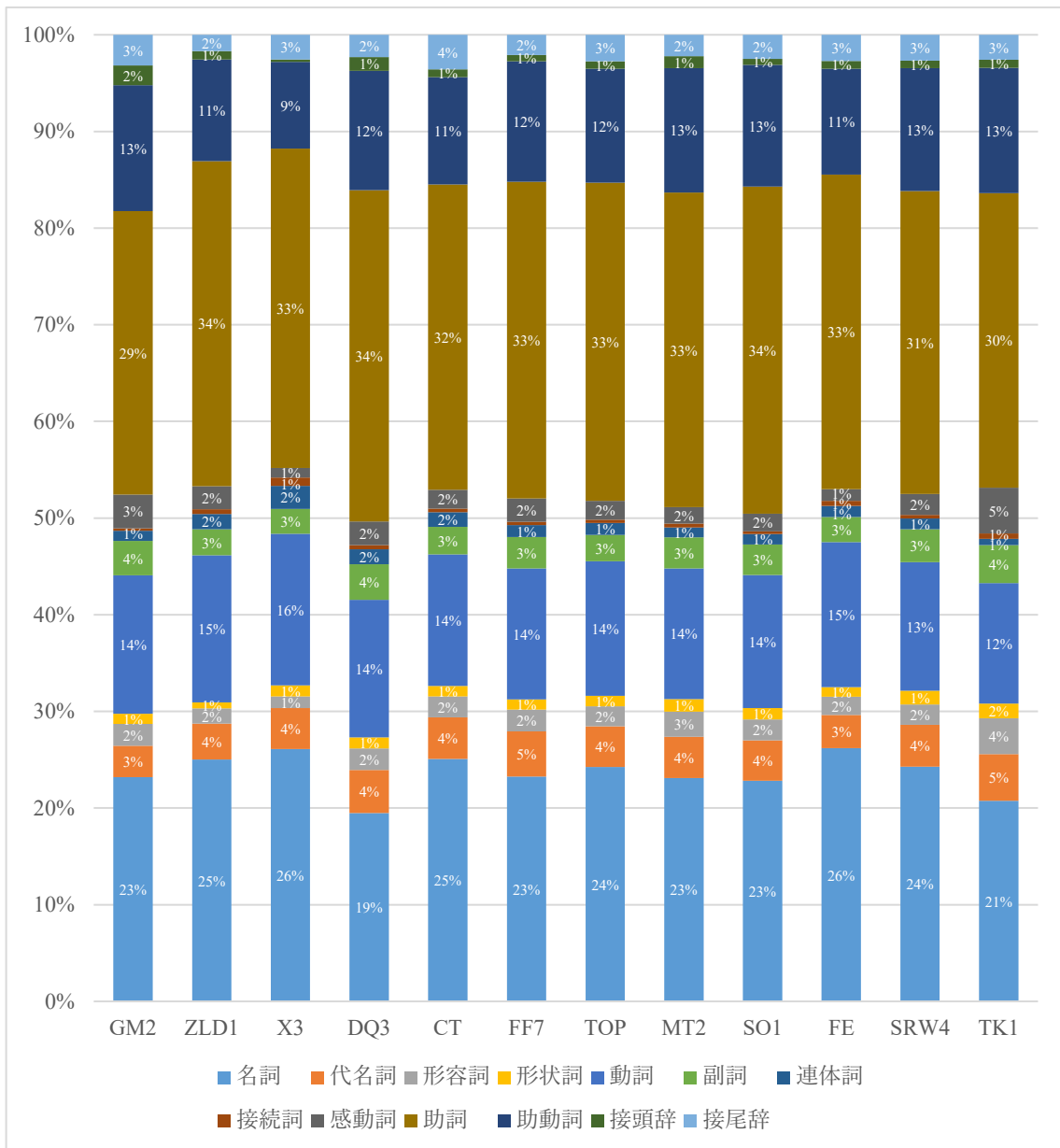


図4 短単位品詞構成比（延べ語数）

3.3 語種構成比

語種構成の計算については、UniDic の語種情報が短単位に付与されているため、短単位の延べ語数を用いる。図 5 は、語種情報のない項目と記号を除き、和語・漢語・外来語・固有名・混種語の 5 区分で構成比を計算した結果である。

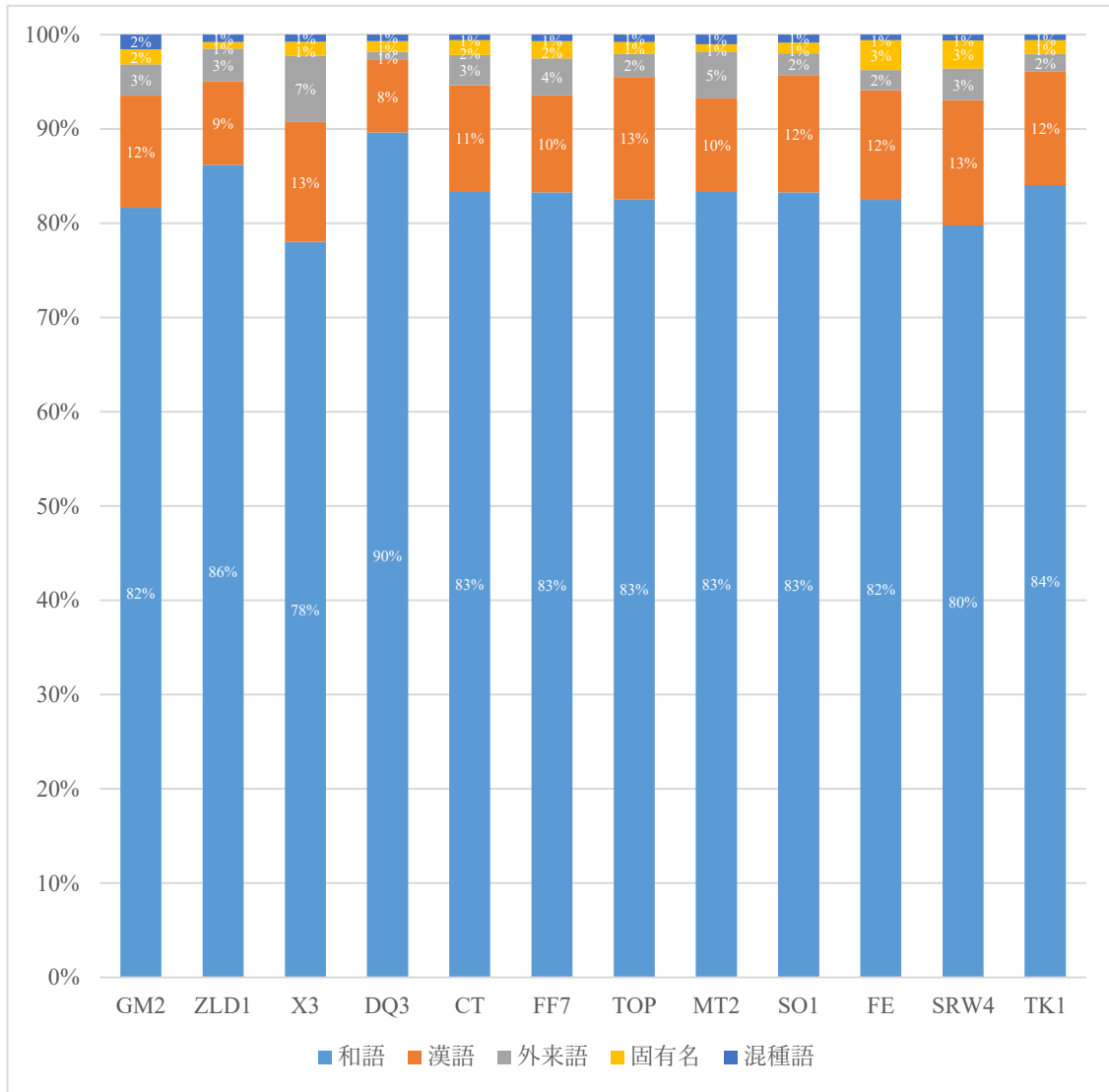


図 5 短単位語種構成比（延べ語数）

全タイトルで和語が最も多く、78～90%を占める。漢語は 8～13%，外来語は 1～7%，固有名は 1～3%，混種語は 1～2%である。山崎（2022）では、CEJC が和語 85%，漢語 9%，外来語 2%，固有名 1%，混種語 1%，BCCWJ1 小説会話文がそれぞれ 84%，12%，1%，2%，1%であり、JGC の多くの作品も大枠ではこれらに近い構成を示す。ただし、作品別には偏りも見られ、SF・機械的要素の強い『ロックマン X3』や現代的な都市生活を描く『マザー 2』で外来語が多い一方、人物・勢力・組織を多数扱う SLG である『ファイアーエムブレム 紋章の謎』『第 4 次スーパーロボット大戦』では固有名が多い。このことから、ゲームテキストの語種構成は、一般的な会話・小説会話と大きく乖離するものではない一方、各作品の世界観による差を比較的反映すると考えられる。

3.4 特徴語

特徴語の抽出には長単位の延べ語数を用いる。長単位を採用するのは、「パワーアップする」「契約する」のような複合的な内容語を意味的にまとまりのある単位として捉えやすいためである。また、対数尤度比（LLR: Log-Likelihood Ratio）は対象コーパスと参照コーパスにおける出現頻度の偏りを評価する指標であるため、異なり語の有無ではなく延べ語数の頻度を用いる。各ゲームを対象コーパス、残る 11 タイトルをまとめたものを参照コーパスとして LLR を算出し、0.1%水準に相当する 10.83 以上、かつすべてのゲームの総出現頻度が 5 以上の語を抽出した。表 6 には、名詞と動詞について LLR 上位 5 語を示す。なお、人名・地名などの固有名詞は除外した。

表 6 各ゲーム特徴語 5 位まで（名詞・動詞）

	名詞	動詞
GM2	大江戸城／祭り／メカ／地獄／温泉	御座る／勝負する／飛ぶ／いらっしゃる／然る
ZLD1	トライフォース／七賢者／司祭／力／願い	掲げる／逃れる／導く／叶える／聞こえる
X3	カプセル／ボディ／イレギュラーハンター ／空中／ダメージ	授ける／入る／受ける／パワーアップする／再生する
DQ3	王様／魔王／大魔王／国／旅	記録する／旅立つ／書き留める／倒す／行く
CT	魔王軍／魔神器／海底神殿／女王／時代	飲む／踊る／現われる／巣くう／逆らう
FF7	マテリア／星／古代種／ソルジャー／魔晄 炉	選ぶ／上る／分かる／説明する／死ぬ
TOP	精霊／指輪／契約／戦争／魔術	書く／いらっしゃる／契約する／願い奉る／操る
MT2	ドル／ゾンビ／看板／病院／バス	御買い求める／喋る／直す／御泊まりに成る／奪い返す
SO1	瞳／スキル／真実／魔界／艦長	覚える／いらっしゃる／謁見する／入手する／逸れる
FE	同盟軍／砦／王子／反乱軍／光	捕らわれる／占領する／加担する／守る／率いる
SRW4	戦力／異星人／地上／地球／通信	撤退する／出撃する／向かう／戦う／合流する
TK1	今日／待ち合わせ／電話／デート／誕生日	待つ／返る／誘う／プレゼントする／連絡する

名詞の特徴語には、各作品の舞台・世界観・物語上の中心概念が強く表れている。ACT では、『がんばれゴエモン 2』の「大江戸城」「祭り」「温泉」が和風の舞台を、『ゼルダの伝説 神々のトライフォース』の「トライフォース」「七賢者」「司祭」が神話的な物語世界を、『ロックマン X3』の「カプセル」「ボディ」「イレギュラーハンター」「ダメージ」が機械・戦闘を中心とする SF 的世界をそれぞれ示す。RPG でも、『ドラゴンクエスト 3』『クロノ・トリガー』『テイルズオブファンタジア』には「魔王」「魔王軍」「精霊」「魔術」など幻想的な世界に関わる語が現れる一方、『ファイナルファンタジー 7』では「マテリア」「古代種」「ソルジャー」「魔晄炉」といった作品固有の概念が上位に現れる。『マザー 2』の「ドル」「病院」「バス」や『ときめきメモリアル 1』の「待ち合わせ」「電話」「デート」「誕生日」は、日常生活を基盤とする作品世界を反映している。さらに、SLG の『ファイアーエムブレム 紋章の謎』と『第 4 次スーパーロボット大戦』では、「同盟軍」「反乱軍」「戦力」「異星人」など、軍事・勢力関係に関わる語が特徴的である。

動詞では、名詞が示す「何が描かれているか」に加えて、「そこで何が行われるか」やキ

キャラクターの話し方が特徴として現れる。『ファイアーエムブレム 紋章の謎』の「占領する」「守る」「率いる」、『第4次スーパーロボット大戦』の「撤退する」「出撃する」「合流する」は、軍事行動を中心とするSLGの展開をよく表す。『ときめきメモリアル1』の「待つ」「誘う」「プレゼントする」「連絡する」は対人関係の形成、『ゼルダの伝説 神々のトライフォース』の「導く」「叶える」「聞こえる」は物語上の導きや願い、『ロックマンX3』の「授ける」「パワーアップする」「再生する」は強化・機械的处理と結び付いている。また、『がんばれゴエモン2』の「御座る」「いらっしゃる」や、『マザー2』の「御買い求める」「御泊まりに成る」のように、行為内容だけでなく、キャラクターや場面に特有の待遇表現・役割語的な特徴を反映する例も見られる。

ただし、ここでの特徴語はJGC内部で12タイトルを相互比較して得た「作品間の相対的特徴」であり、一般的な日本語と比べた「ゲーム特有の語」を直接示すものではない。今後、同じ方法でCEJCやBCCWJ小説会話文等を参照コーパスとすれば、ゲームというレジスタに特徴的な語彙をより明確に抽出できると考えられる。

4. 出現語彙リストの試行公開

著作権の都合上、コーパス本文そのものを公開することはできないが、出現語彙リストを筆者のホームページ (<https://kenjima.net/>) で公開する。公開対象は、本稿で扱った4ジャンル12タイトルのストーリー関連テキスト(セリフ・心内発話・ナレーション・テキスト)を解析した短単位データと長単位データである。

データの構成に関して、1枚目のシートは「マニュアル」で、収録ゲームのジャンル、略称、概要やデータの利用規約などを示す。2枚目以降は語彙リストで、短単位(SUW)と長単位(LUW)に分けている。短単位は緑色系、長単位は青色系のシート見出しとし、それぞれに全テキスト種をまとめた「全体」と、セリフ・心内発話・ナレーション・テキストを個別に集計したシートを用意した(図6)。



図6 シート見出しの色と名称

シートの中身も色で解析単位が分かるようにしてある。緑色系のシートが短単位、青色系のシートが長単位であり、さらに左上には「全体」「セリフ」「心内発話」「ナレーション」「テキスト」など、現在閲覧しているテキスト区分を明示した(図7, 図8)。形態論情報として語彙素、語彙素読み、品詞を付与し、ゲーム名は略称で示してジャンル別にグルーピングした。Excel形式であるため、フィルタ機能と並べ替え機能も利用できる。

短単位形態論情報 (全体)				ACT					RPG					SLG			AVG	
語彙素	語彙素読み	品詞		GM2	ZLD1	X3	DQ3	CT	FF7	TOP	MT2	SO1	FE	SRW4	TK1			
少尉	ショウイ	名詞-普通名詞-一般		0	0	0	0	0	0	0	0	0	0	1	0	0		
上位	ショウイ	名詞-普通名詞-一般		0	0	0	0	0	1	1	0	0	0	0	0	0		
乗員	ショウイン	名詞-普通名詞-一般		0	0	0	0	0	0	1	0	0	0	0	0	0		
上級	ショウエイ	名詞-普通名詞-ラ変可能		0	0	0	0	0	0	0	0	0	0	0	0	1		
上級	ショウエン	名詞-普通名詞-ラ変可能		0	0	0	0	0	0	0	0	0	0	0	0	1		
罪人	ショウカ	名詞-普通名詞-ラ変可能		0	0	0	0	0	0	0	0	1	0	0	0	0		
城下	ショウカ	名詞-普通名詞-一般		0	0	0	1	1	0	0	0	1	0	0	0	0		
浄化	ショウカ	名詞-普通名詞-ラ変可能		0	0	0	0	0	0	0	0	0	0	1	0	0		
紹介	ショウカイ	名詞-普通名詞-ラ変可能		1	1	0	0	0	1	1	1	0	0	1	1	1		
障害	ショウガイ	名詞-普通名詞-ラ変可能		0	0	0	0	0	0	1	0	0	0	0	0	0		
小字	ショウガク	名詞-普通名詞-一般		0	0	0	0	0	0	0	0	0	0	0	0	1		
罪人	ショウガク	名詞-普通名詞-ラ変可能		0	0	0	0	0	1	0	0	0	0	0	0	0		
召喚	ショウカン	名詞-普通名詞-ラ変可能		0	0	0	0	0	1	1	0	0	0	0	0	0		
上巻	ショウカン	名詞-普通名詞-一般		0	0	0	0	0	1	0	0	0	0	0	0	0		
上巻	ショウカン	名詞-普通名詞-一般		0	0	0	0	0	0	0	0	1	0	1	0	0		

図7 短単位語彙リスト(全体)の一部

長単位形態論情報（セリフ）				ACT				RPG				SLG		AVG	
語彙素	語彙素読み	品詞		GM2	ZLD1	X3	DQ3	CT	FF7	TOP	MT2	SO1	FE	SRW4	TK1
天井	テンジョウ	名詞・普通名詞一般		0	0	0	1	0	0	0	0	0	0	0	0
天上	テンジョウ	名詞・普通名詞一般		0	0	0	0	1	0	0	0	0	0	0	0
天承	テンショウ	名詞・普通名詞一般		0	0	0	0	0	0	0	0	1	0	0	0
天上界	テンジョウカイ	名詞・普通名詞一般		0	0	0	0	1	0	0	0	0	0	0	0
伝承書	デンショウショ	名詞・普通名詞一般		0	0	0	0	0	0	1	0	0	0	0	0
人類	アンニョク	名詞・普通名詞一般		0	0	0	1	0	0	0	0	0	0	0	0
天宮	テンシヨク	名詞・普通名詞一般		0	0	0	1	0	0	0	0	0	0	0	0
転機	テンシヨク	名詞・普通名詞一般		0	0	0	1	0	0	0	0	0	0	0	0
転換する	アンシヨクメル	動詞一般		0	0	0	1	0	0	0	0	0	0	0	0
伝書箱	デンシヨネコ	名詞・普通名詞一般		0	0	0	0	0	0	0	0	1	0	0	0
伝書鳩	デンシヨハト	名詞・普通名詞一般		0	0	0	0	0	0	0	0	1	0	0	0
伝説	デンセツ	名詞・普通名詞一般		0	0	0	0	0	1	0	0	0	0	0	0
天降	テンキイ	名詞・普通名詞一般		0	0	0	0	0	0	0	0	1	0	0	0
伝説	デンセツ	名詞・普通名詞一般		0	1	0	1	1	1	1	1	1	1	0	1
伝説上	デンセツジョウ	名詞・普通名詞一般		0	0	0	0	0	0	1	0	1	0	0	0

図 8 長単位語彙リスト（セリフ）の一部

語彙リスト中の数字は 1 か 0 の二値データであり、1 は当該ゲームにその語が出現したこと、0 は出現しなかったことを示す。数字は出現頻度ではないことに注意が必要である。現段階では、特にセリフにおける同一発話の重複認定について最終的な確認作業が残っているため、頻度表ではなく出現リストとして公開することにした。

5. おわりに

本稿では、JGC の前期 12 タイトルについて、ストーリー関連テキストを対象としたデータ概要と量的特徴を示し、出現語彙リストの試行公開を報告した。前期データは異なり 13,702 短単位、延べ 408,302 短単位からなり、約 95% をセリフが占める一方、ナレーションや心内発話の比率には作品ごとの大きな違いが見られた。品詞構成では名詞・動詞・助詞の比率は比較的共通するが、作品ごとに若干差があり、語種構成でも外来語や固有名詞の比率に作品固有の世界観を反映すると考えられる差が確認された。また、特徴語では舞台設定、キャラクターの話し方、物語上の行為を反映する語が LLR 上位に現れることを確認した。

今後の課題は、全 12 タイトルの校正済み頻度データを用いたジャンル別の文体論的特徴の分析、CEJC や BCCWJ などの外部データとの比較、さらに後期データとの通時的比較を進めたい。

謝 辞

本研究は JSPS 科研費「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

- 落合哉人（2026）「BCCWJ2 における SNS データの扱いについて」BCCWJ2 シンポジウム：中間成果の公開と展望 発表資料。
- 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 2022 発表論文集, 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 2023 発表論文集, pp.151-158.
- 麻子軒（2024）『『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』言語資源ワークショップ 2024 発表論文集, pp. 279-287.
- 麻子軒（2025）「ゲーム内テキスト抽出における OCR の性能評価—レイアウトと解像度の影響に着目して—」言語処理学会 第 31 回年次大会 発表論文集, pp.397-401.

山崎誠（2022）『『日本語日常会話コーパス』の会話と小説の会話の比較』シンポジウム「日常会話コーパス」VII 発表資料.

関連 URL

国立国語研究所（2022）『日本語日常会話コーパス』短単位語彙表・語数表 ver.2022.09, <https://www2.ninjal.ac.jp/conversation/>（2026 年 8 月 8 日閲覧）.

麻子軒（2026）『日本語ゲームコーパス』前期データ出現語彙リスト, <https://kenjima.net/>（2026 年 8 月 8 日閲覧）.