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―

麻子軒
関西大学 国際教育センター

概要

ゲーム内テキスト，特にキャラクターのセリフを効率的に文字起こしする方法について，4 つのゲームを対象に Google Cloud Vision の OCR 認識精度を比較した．比較は，テキストのレイアウトおよび解像度の観点から行い，誤認識パターンの分析も実施した．その結果，記号類の認識が弱い傾向が見られるものの，全体的に精度が 0.88～0.99 と，情報抽出に大きな影響はないことが判明した．ただ，極端に短い文の場合には別言語や類似文字への誤認識が発生する傾向が確認された．また，レイアウトが固定されていないゲームでは精度が相対的に低く，低解像度の古いゲームでは濁点の欠落や一部の文字が飛ばされて認識されない事例が観察された．

1 はじめに

これまで，言語資源として紙媒体や音声媒体を対象としたコーパスは数多く作成されてきたが，ゲームを対象としたものはあまり見られていない．その理由の一つとして，ゲーム内テキストを文字起こしする作業に多大な手間がかかる点が挙げられる．しかし，ゲーム内テキストは言語学的研究，ゲームデザイン研究，翻訳研究など，幅広い分野において重要な資料として活用できる可能性があり，そのテキストを効率的に抽出できる手法が求められている．

本研究では，ゲーム内テキスト，特にキャラクターのセリフを効率的に構造化し，文字起こしする方法について報告する．具体的には，筆者が構築中の『日本語ゲームコーパス（JGC）』に含まれる 4 つのゲームを対象とし，Google Cloud Vision を用いた OCR の認識結果について，テキストのレイアウトと解像度という 2 つの観点から精度を比較する．さらに，OCR による誤認識のパターンを分析し，それを通じてゲームコーパス構築における困難点と課題を明らかにする．

2 先行研究

2.1 ゲームコーパスに関する研究

これまでの言語資源は主に書籍媒体を対象としたものが多かったが，ゲームを対象としたコーパスに関しては，筆者が構築している『日本語ゲームコーパス（JGC）』[1, 2]を除いてほとんど存在していない. その理由は, 技術的な制約や作業にかかる手間, そして時間的なコストの観点から，コーパスの構築が極めて困難であるためである．特に，効率的な文字起こし手法として一般的に利用される OCR 技術が，ゲームには適用しにくいことが大きな課題となっている．

ゲームは書籍と異なり，テキストの表示形式が固定されたレイアウトでないことが多く，紙媒体のように機械的にスキャンして処理することも難しい．また，古いゲームの場合，解像度が低いという問題もあり，これらの要因が重なることで文字起こし作業の自動化が非常に難しくなっている．

2.2 OCR 技術に関する研究

OCR は，日本の文書における自動認識技術として，紙媒体やデジタル化文書の効率的な解析に広く活用されている．しかし，日本語特有の縦書きや複雑な文字構造に対応する際には，認識精度に影響を与える要因もいくつか存在する．

例えば，国立国会図書館の関連 OCR 事業では，既存の OCR サービスを使用した場合，現代の資料における文字認識精度が 0.98 に達しているのに対し，明治期の資料では 0.8 を下回る結果が得られている[3, 4]．この差は，資料の解像度や文字形状の認識に起因すると思われる．また，同事業では AI を活用した OCR 処理プログラムが開発され，学習データを用いたレイアウト抽出により，認識精度の向上が達成されたことが報告されている[5].

以上より，解像度や不規則なレイアウトの領域認識が OCR の課題であることが明らかとなった．先行研究は主に書籍や印刷物を対象としているが，ゲーム内テキストも画面上に表示される点を除けば印刷物と類似の性質を持つ．そのため，これらの研究から得られた知見は，日本語文書特有の課題に対応する OCR 技術の進展を示しており，ゲーム内テキスト抽出にも有用な示唆を与えると考えられる．

ただし，ゲームには特有のテキストに関する難しさが存在する可能性があり，本研究ではこれを検証する．検証の観点として，先行研究でも課題とされてきたレイアウトと解像度の 2 つの要因に注目する．具体的には，会話枠の位置が一定であるゲームと一定でないゲーム（レイアウトの観点），および古いゲームと新しいゲーム（解像度の観点），計 4 つのゲームを選定し，比較分析を行う．

3 ゲーム内テキストの抽出手順

筆者は，2000 年代を境に，それ以前を前期，それ以降を後期として，それぞれ 12 タイトル，合計 24 タイトルを選定し，ゲームコーパスを構築している．対象ジャンルは ACT（アクション），RPG（ロールプレイング），SLG（シミュレーション），AVG（アドベンチャー）の 4 種類である．ゲームコーパスの詳細な作業手順については，麻[1, 2]を参照されたいが，簡単にまとめると以下の流れとなる．

まずは，ゲームをプレイし，その中で表示されるテキストを含む画面をキャプチャーした．次に，画像から OCR を用いてテキスト化を行い，必要に応じて整形作業を実施した後，手作業での修正も加えた．その後，テキストを形態素解析ツールにかけることでコーパスとして整理し，最終的には語彙表などを作成する工程を経た．

以下では，本研究と関連する重要な部分，特にキャプチャーと OCR の方法について説明する．

3.1 キャプチャーの方法

キャプチャーのタイミングは，新たなテキストが表示されるたびに実施した．キャプチャー作業は，ゲームの画面を PC 経由で表示させることで行い，使用したキャプチャーボードは AVerMedia Live Gamer EXTREME 2 GC550 PLUS である．また，キャプチャーには Bandicam というソフトを使用した．画像の画質設定については，ゲーム機の本来の解像度に依存するが，例えばニンテンドースイッチの場合は 1920×1080 になる．

3.2 OCR の方法

使用した OCR ツールは Google Cloud Vision である．Python を用いて OCR 処理を実行し，その際の前処理や整形の詳細を以下に説明する．

まず，取得した画像に対して，必要に応じて会話枠の色の特定や背景の除去などの前処理を行った．具体的には，枠の位置が動的なゲームの場合，会話枠の色を特定し，それ以外の領域を黒く塗りつぶすことで，認識対象を会話枠内のみに絞る処理を実施した．これにより，不必要な情報の干渉を防ぎ，OCR の精度向上が期待できる．一方，会話枠の位置が一定であるゲームの場合，上述のような前処理が不要であるため，事前に枠の座標を特定し，OCR 処理を当該領域のみに限定することで効率化を図った．

次に，キャプチャーしたテキストデータを出力する際には，発話キャラクター名（以下，キャラ名）とセリフを分離する整形作業が必要となる．多くの場合，ゲーム内のテキストレイアウトでは，最初の行にキャラ名が表示され，その次の行以降にキャラクターのセリフが表示される，あるいはキャラ名の直後に開くかぎ括弧が付く形式が採用されていることが多い．この仕様を利用し，最初の改行または開くかぎ括弧を基準としてキャラ名とセリフを分離する処理を行った．このような規則性を活用することで，テキストデータの効率的な整形が可能となる．

4 OCR の認識結果に対する検証

4.1 選定したゲーム

本研究では，特にテキストのレイアウトと解像度が OCR 文字起こしに与える影響の検証に焦点を当てるため，『日本語ゲームコーパス（JGC）』に収録されているゲームの中から，会話枠の位置（レイアウトが固定されているものとそうでないもの）および発売時期（解像度が高いものと低いもの）の 2 つの観点に基づき，4 つのゲームを選定して比較を行う．選ばれたゲームは『がんばれゴエモン 2（以下 GM2）』『第 4 次スーパーロボット大戦（以下 SR4）』『オクトパストラベラー1（以下 OT1）』『モンスターハンターライズ（以下 MHR）』である．

GM2 と OT1 は図 1 に示したように，会話枠の位置が固定されておらず，発話キャラクターの位置に応じて会話枠の位置が変動するゲームである．

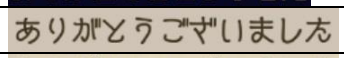
図 1　会話枠の位置が一定でないゲーム

一方，SR4 と MHR は図 2 のように，会話枠が表示される位置が常に一定しているゲームである．

図 2　会話枠の位置が一定しているゲーム

また，GM2 と SR4 は 1990 年代，OT1 と MHR は 2020 年代に発売されたものである．当然のことながら，古いゲームより新しいゲームは解像度が高い．表 1 に「ありがとうございました」で例示した．

表 1　フォント例

	フォント例
GM2	
SR4	
OT1	
MHR	

4.2 認識精度の評価基準

OCR 性能の検証方法としては，4 つのゲームから無作為に選ばれた 1 つの作業ファイルを対象に，認識精度を確認した．精度の指標には，Levenshtein の編集距離と類似度を用いた．

編集距離とは，「挿入」「削除」「置換」といった編集操作の回数を基に，2 つの文字列間の近さを数値化したものである．それを文字列の長さを考慮して標準化したものが類似度（通常は 0～1 の範囲にある）であり，1 に近いほど精度が高いことを意味する．実際の計算には，Python のライブラリである python-Levenshtein を使用した．

4.3 ゲーム別の認識結果比較

4 つのゲームすべてのレコードについて編集距離と類似度を計算し，それらの平均値を表 2 に示した．キャラ名とセリフを分けて計算したが，GM2 ではキャラ名が表示されないため，該当数値は存在しない．また，記号類，特に「…」は「．．．」「･･･」「･･･」といった異なるバリエーションとして認識されることが多いため，計算の前に正規化を実施した．

表 2　OCR 認識結果の比較

	レコード数	キャラ名		セリフ	
		編集距離	類似度	編集距離	類似度
GM2	552	-	-	1.66	0.90
SR4	1,386	0.14	0.98	0.29	0.99
OT1	957	0.76	0.88	0.92	0.90
MHR	973	0.03	0.99	0.19	0.99

Google Cloud Vision の OCR 認識精度に関する具体的な報告は確認できないが，国立国会図書館の OCR 事業で用いられた NDLOCR では，対象資料の時代にもよるものの，1930 年代以降の資料であれば認識精度が 0.90～0.97 に達している[5]．本研究では，会話枠の位置が固定されている SR4 と MHR のキャラ名とセリフの類似度は，それより高水準であることが確認された．一方で，会話枠の位置が一定でない GM2 と OT1 の場合，キャラ名とセリフの類似度が 0.9 程度まで低下することが観察された．特に GM2 のセリフでは編集距離が 1.66 に達しており，平均するとレコードごとに 1 回以上の削除や挿入といった修正が必要であることが分かった．また，低解像度の GM2 と SR4 に比べ，高解像度の OT1 と MHR の類似度が高いことも明らかになった．これらの結果は，レイアウトの特定しやすさと解像度が OCR 精度に与える影響を示していると考えられる．特に，OT1 では認識領域である会話枠を色で判定する際に，最初の改行が正しく認識されないことが頻発しているため，キャラ名とセリフの分離に失敗するケースが多く，全体的な認識精度の低下につながった．

4.4 誤認識のパターン

本節では，4 つのゲームでの誤認識のパターンを整理し，認識精度を下げる要因について探る．

表 3　GM2 に多い誤認識パターン

	正解	OCR 認識結果
別言語	いらっしゃい、いらっしゃい！	＄ｔ！«Ｊ＆Ｇｔ！＼ｔ！«Ｊａｇｔ？
	「なにぃ！？」	［ｉｚｉｚｕ！？Ｊ
濁点脱落	もどってきたぜ！	もとってきたせ！
	じごくデスよーーーー	しこくテスよーーーー

	っ！！	ッ！！
類似文字	どのしなを、かいますか？	とのしな壱、かいますか？
	とおりますか？	とありますか？
一部脱落	どのへやにしますか？	とのやにしますか？
	１かい１００両だ。	１１００両た。

GM2 のような古いゲームでは，容量節約のため漢字表記が使用されていないことが多く，その結果，文脈による判断が難しくなり，認識率の低下につながったと考えられる．特に，別言語として誤認されたケースや，濁点が正しく認識されないケースが多く見られる．また，「を」と「壱」，「お」と「あ」など，形が類似している文字の区別が困難である点も顕著である．さらに，理由は不明だが，一部の文字が認識されない場合も確認されている．

表 4　SR4 に多い誤認識パターン

	正解	OCR 認識結果
別言語	はぁーい！	ｒｉｃｈ－ｆｌｆｆ
	みんな、無事か！？	杂、無事！？
	リョウ	彐ク
類似文字	くっ､不覚をとったか。まあいい、撤退だ	〈コ、不覚在加。心、撤退尤
	うぬぅ・・・やむをえんか・・・	ラぬぅ・・・やむをえんか・・・
	了解！へへっ、面白くなってきたぜ	了解！＾＾面白くなってきたぜ
	ハ、ハーディアス将軍！？	八、ハーディアス将軍！？
小文字	ジャスティィィィィン！	ジャスティィイイイン！
	うおぉぉぉぉっ！	うおおおおおっ！
一部脱落	何だと！！	何だ！！
	敵襲です！うわっ！？	敵です！うわっ！？

SR4 では，フォントデザインの影響からか，英語だけでなく中国語として誤認識されるケースも見られた. GM2 に比べると濁点は正しく認識されているが，「〈コ」と「く」，「う」と「ラ」，「へ」と「＾」，「ハ」と「八」，など，類似文字による誤認識が発生している．また，ア行の小文字に関しては認識が弱い傾向が見られる点が特徴的である．

表 5　OT1 に多い誤認識パターン

	正解	OCR 認識結果
別言語	え…！？	Ｚ！？
	はっ…！	Ｉｔ…！
濁点付加	テリオン	デリオン
	秘書ルシア	秘書ルジア
類似文字	…つッ！	…フッ！
	アンナ夫人	アンチ美人
	プリムロゼ	ブリムロゼ

OT1 では，解像度の低いゲームで見られる濁点脱落は発生していないが，フォントデザインが独特であるためか，濁点が誤って付加される現象が確認された．また，極端に短い文とキャラ名が，「え」と「Z」，「は」と「It」，「つ」と「フ」，「ナ」と「チ」，「夫」と「美」など，別の言語または類似した文字として判別されるケースが多い傾向にある．これは，短文とキャラ名では文脈による判断材料が限定されるためと考えられる．長い文がほぼ正確に認識されていることや，全体的にキャラ名の認識精度がセリフに比べてわずかに低い傾向がある点が，その裏付けとなるであろう．

表 6　MHR に多い誤認識パターン

	正解	OCR 認識結果
類似文字	…ん？ギルドへの登録はどうなっとるのかって？	…ｈ？ギルドへの登録はどうなっとるのかって？
小文字	セイニャ～ッ！ハッ！トッ！ニャッ！	セイニャ～ッ！ハッ！トッ！ニヤツ！
	うぅ～。し、舌かんだぁ…。いてて…。	うう～。し、舌かんだぁ…いてて…。
	飴屋のコミツ	飴屋のコミッ

MHR は，最も一般的なフォントに近いデザインであるためか，認識精度が非常に高い．その誤認識の範囲は「ヤ」と「ャ」，「ツ」と「ッ」，「う」と「ぅ」，「ん」と「h」など，ごくわずかな違いに限られている．

5 考察と今後の課題

ゲーム内テキストの情報抽出において，記号類，特に「…」の認識精度が低い傾向が見られたが，記号が持つ情報量が少ないため，情報抽出に与える影響は限定的である．また，Google Cloud Vision はフォントの形だけでなく，前後の文脈も認識の判断材料としていると推測される．この特性により，文が極端に短い場合には，別言語や類似文字への誤認識といった問題が発生する傾向がある．なお，低解像度のゲームにおいて，濁点の欠落や一部の文字が飛ばされて認識されない事例も観察された．ゲームで使用されるフォントは，独自のデザインが採用されることが多いが，解像度が高く，文が十分に長い場合には, 認識に大きな問題は発生しにくい. しかし，レイアウトが一定でない場合，テキスト認識領域の特定が容易でなく，今回の分析においては精度を下げる最も大きな要因となっている．

以上の結果から，Google Cloud Vision によるゲーム内テキストの認識精度が 0.88～0.99 に達しているものの，紙媒体と同様にレイアウトの領域認識と解像度が重要な課題であることが示された．解像度については文脈を用いることである程度克服できたが，レイアウト認識が精度向上の鍵を握ると考えられる．

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）」の助成を受けたものです.

参考文献

[1] 麻子軒. 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, 2023.

[2] 麻子軒. 『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, 2024.

[3] 徳原直子・青池亨. 国立国会図書館デジタル化資料の OCR 全文テキストデータの活用可能性：NDL ラボの実験サービスにおける沖縄関連キーワードの検索結果から. デジタルアーカイブ学会誌, 6 巻 s3 号, pp.210-213, 2022.

[4] 国立国会図書館. 令和 3 年度デジタル化資料の OCR テキスト化. (オンライン) (引用日: 2025 年 1 月 1 日.) https://lab.ndl.go.jp/data_set/ocr/r3_text/

[5] 国立国会図書館. 令和 3 年度 OCR 処理プログラム研究開発. (オンライン) (引用日: 2025 年 1 月 1 日.) https://lab.ndl.go.jp/data_set/ocr/r3_software/