

УДК 004.415.25

## ПОДХОД К СОЗДАНИЮ КОРПУСА ТЕКСТОВ ВИДЕОИГР НА ОСНОВЕ УНИВЕРСАЛЬНОЙ СТРУКТУРЫ

Н. Р. Нурлыгаянов<sup>1</sup> [0000-0002-7969-0221], В. В. Кугуракова<sup>2</sup> [0000-0002-1552-4910]

<sup>1, 2</sup>Институт информационных технологий и интеллектуальных систем  
Казанского федерального университета, ул. Кремлёвская, 35, г. Казань, 420008

<sup>1</sup>nikita25@mail.ru, <sup>2</sup>vlada.kugurakova@gmail.com

### **Аннотация**

Рассмотрена проблема высокой и увеличивающейся стоимости разработки видеоигр, для её решения предложено применить процедурную генерацию контента, что позволит снизить затраты на разработку.

Работа является частью масштабного исследования по автоматическому созданию прототипов видеоигр и посвящена обработке игровых сценариев, то есть текстов на естественном языке. Предложено выделять из сценариев необходимые сущности и передавать их дальнейшим шагам алгоритма, который по текстовым описаниям будет генерировать игровые ресурсы.

Существует несколько публикаций, посвящённых обработке игровых текстов, в которых предложено несколько различных структур хранения выделенной информации. В настоящей статье предложен универсальный формат, который подойдёт для обработки текста любой видеоигры и позволит создать корпус текстов для использования в дальнейших исследованиях и автоматической генерации игровых прототипов.

**Ключевые слова:** PCG, NLP, разработка видеоигр.

## ВВЕДЕНИЕ

В индустрии видеоигр существует проблема, связанная с тем, что бюджеты игр<sup>1</sup> быстро увеличиваются, поэтому разработка огромных и инновационных игр становится слишком дорогой (рис. 1) [1].



Рис. 1. Быстрое повышение стоимости разработки игр [1]

В последние годы игровая индустрия претерпела значительные изменения в своих бизнес-модели и подходах к разработке. Эти изменения во многом обусловлены экспоненциальным ростом затрат на создание игр, особенно в сегменте AAA<sup>2</sup>-проектов. Отчёт Управления по конкуренции и рынкам Великобритании [2] свидетельствует о том, что затраты на разработку новой видеоигры *Call of Duty*

<sup>1</sup> Перечень видеоигр, упоминаемых в этом исследовании, приведён в разделе Лудография.

<sup>2</sup> AAA (произносится как «трипл-эй») — неофициальный термин, используемый в игровой индустрии для обозначения игр с самыми высокими бюджетами разработки и маркетинга. Происхождение термина не имеет однозначного объяснения, но предполагается, что он возник по аналогии с системой оценки облигаций, где AAA обозначает наивысший рейтинг. В контексте видеоигр AAA указывает на наивысшее качество, масштаб и ожидаемый коммерческий успех проекта. AAA-игра характеризуется крупными инвестициями, франшизами, включающими несколько названий, множеством персонажей и длинными сюжетными линиями, а также большой базой поклонников.

достигли 300 миллионов долларов, а бюджет ожидаемой в 2025 году новой версии из известной серии<sup>3</sup> видеоигр *Grand Theft Auto VI* (или *GTA6*) составляет минимум 250 миллионов долларов. Это резко контрастирует с бюджетом типичных AAA-игр пять лет назад, составлявшим от 50 до 150 миллионов долларов.

Чтобы справиться с затратами игровой разработки [1], растущими экспоненциально, игровые компании прибегают к различным стратегиям, таким как изменение моделей монетизации.

В последние годы наблюдается тенденция к фрагментации игрового контента: вместо полноценных расширений, как это было ранее (например, дополнения для *Half-Life*), сейчас чаще встречаются небольшие, короткие по игровому времени пакеты DLC<sup>4</sup>. Яркий пример такого подхода демонстрирует видеоигра *Payday 2*.

Кроме того, даже видеоигры, продающиеся по полной цене, часто включают микротранзакции [3] — операции во внутриигровом магазине для покупки цифровых товаров [4]. К числу спорных изменений последних лет относятся введение NFT<sup>5</sup>, а также повышение базовой цены на копию AAA-игры с \$60 до \$70 [5–8].

Чтобы снизить эти затраты, некоторые компании обращаются к методам машинного обучения для оптимизации процесса разработки. Например, в 2023 году компания Ubisoft<sup>6</sup> представила *Ghostwriter* — инструмент, который генерирует контекстно-зависимые фразы для персонажей, сокращая время, которое нарративные дизайнеры тратят на стандартные диалоги, позволяя им сосредоточиться на важных диалогах [9]. Интересно отметить, что один из его создателей

---

<sup>3</sup> GTA — серия видеоигр в жанре экшен-приключений от компании Rockstar Games.

<sup>4</sup> DLC, сокр. от англ. Downloadable Content — дополнительный загружаемый контент.

<sup>5</sup> NFT, сокр. от англ. Non-Fungible Token — невзаимозаменяемый токен, уникальный цифровой актив, представляющий право собственности на определённый виртуальный или физический объект, зарегистрированный в блокчейне. NFT позволяет подтвердить подлинность и уникальность цифровых предметов, таких как произведения искусства, игровые предметы или коллекционные объекты.

<sup>6</sup> Ubisoft — французская компания, занимающаяся разработкой, изданием и дистрибуцией видеоигр. Основана в 1986 году пятью братьями Гийемо. Ubisoft известна своими популярными игровыми сериями, такими как, например, *Assassin's Creed*. Компания является одной из крупнейших в игровой индустрии и имеет множество студий и офисов по всему миру, занимаясь разработкой игр для различных платформ, включая консоли, ПК и мобильные устройства.

ранее работал над *AI Dungeon* — текстовой игрой, в которой используются модели GPT, в том числе GPT-2 и GPT-3 — для создания уникальных сюжетных линий на основе действий игрока [10]. Этот опыт демонстрирует преимущество в развитии технологий искусственного интеллекта (ИИ) для игровой индустрии — переход от экспериментальных проектов, к более специализированным инструментам.

Такая эволюция подчёркивает растущую роль ИИ в автоматизации различных аспектов игрового дизайна, от генерации диалогов до создания целых сюжетных линий.

Это, в свою очередь, открывает новые возможности для процедурной генерации контента (ПГК, PCG), позволяя разработчикам создавать более разнообразный и динамичный игровой опыт при потенциальном снижении затрат на разработку. В [11] рассматривалась возможность автоматического создания прототипа видеоигры на основе набора входных данных. Сделан вывод, что использование компьютерной генерации позволит сократить время, которое необходимо потратить специалистам в предметных областях, значит, и расходы на оплату их труда.

Настоящая статья посвящена описанию первого шага алгоритма автоматической генерации игровых прототипов, а именно, сбора сценариев видеоигр и разметки их для создания корпуса, который позже может быть использован для процедурной генерации контента.

### **БЛИЗКИЕ РАБОТЫ**

Ван Стегерен и Тёне [12] предложили качественные критерии для корпусов видеоигровых текстов, описали источники игровых текстов и представили три корпуса с примерами использования. Корпус — хорошо организованный набор данных, предназначенный для того, чтобы исследовать определённый языковой признак (набор признаков) или обучать инструмент NLP [13].

Критериями качества набора данных обозначены:

- богатство — наборы данных должны содержать игровой текст и информацию о внутриигровом контексте,
- репрезентативность — тексты должны быть написаны профессионалами и получены из хорошо известных (коммерческих) игр,

- разнообразие — наборы данных должны отражать разнообразие видеоигр,
- переносимость — наборы данных должны храниться в текстовом формате, не требующем специальных инструментов для чтения и редактирования.

Мы предполагаем, что за счёт интеграции этих критериев и использования предложенной файловой структуры сценарии различных игр можно будет обрабатывать с использованием единого набора алгоритмов. Этот подход призван облегчить разработку методов процедурной генерации контента для видеоигр, если использовать методы обработки естественного языка для улучшения генерации уровней, персонажей и других игровых элементов.

В контексте создания корпусов для исследований в области видеоигр стоит отметить работу [14], в которой представлен значительный набор данных. Авторы этой статьи создали корпус, содержащий информацию о 755 играх вместе с их метаданными. Этот набор данных был разработан специально для содействия исследованиям по применению ИИ в сфере видеоигр. Важно отметить, что игры в этом корпусе были разработаны студентами, что отражено в названии набора данных. Это обстоятельство ставит под вопрос репрезентативность корпуса для всей индустрии видеоигр. Авторы статьи, однако, обосновывают ценность своего набора данных тем, что он отражает реальный опыт студентов в разработке игр. Тем не менее, для целей нашего исследования, которое сфокусировано на аспектах разработки видеоигр с точки зрения информатики (а не на психологических аспектах), вопрос репрезентативности остаётся критически важным. В рамках данной работы предпринимается попытка создать корпус, который бы более точно отражал общие тенденции и практики в индустрии видеоигр в целом.

Исследование [15] демонстрирует потенциал использования ИИ для генерации игрового контента. Авторы показали, что игровые задания со структурой **<название, задача, описание>** могут быть сгенерированы при помощи GPT-2, обученной на специально подготовленном наборе данных с заданиями. Для оценки качества сгенерированного контента был проведён онлайн-опрос. Результаты показали, что хотя сгенерированные задания не могут быть использованы в игре сразу после генерации, они представляют собой ценную основу, значи-

тельно упрощающую работу авторов игровых заданий. Важно отметить, что данное исследование было сфокусировано на одной игре — *World of Warcraft*, что ограничивает возможность прямого обобщения результатов на другие игры.

Работа [16] продолжает [15], в ней предложено генерировать базовые варианты заданий для ролевых игр в общем, а модели обучать на текстах пяти игр (*Baldur's Gate* части I и II, *The Elder Scrolls* части IV и V, *Torchlight* часть II). В [16] также предложен новый формат данных, ориентированный на задания.

Вернёмся к работе [12], рассмотрим предложенные в ней три набора данных — *Torchlight II quests*, *Knights of the Old Republic dialogue*, *The Elder Scrolls documents* (два из которых представлены в формате CSV и один в формате JSON, что не ограничивает их понимания, так как такие файлы можно открывать и редактировать с помощью любых привычных текстовых редакторов).

- *Torchlight II quests* имеет сложную иерархическую структуру (см. рис. 2), можно отметить, что имя вложенного поля в структуре описания диалога содержит его тип.

```
[
  {
    "name": "A2-after-Search_for_mana_guardian",
    "questrankmin": "21",
    "questrankmax": "35",
    "dialogpriority": "8",
    "autoassign": "true",
    "questcontrollercompletes": "true",
    "quest_guid": "3243555067944249791",
    "requirements": {
      "questscomplete": {
        "quest": "A2-SEARCH_FOR_MANA_GUARDIAN"
      }
    },
    "dialog": {
      "passive": [
```

Рис. 2. Пример *Torchlight II quests*

Этот набор данных предлагается использовать для сравнительного изучения основных и второстепенных заданий, исследования «вкусного (flavor)» текста

---

и, так как цель задания — это его резюме, для оценивания техник суммаризации в контексте видеоигр [17].

- Knights of the Old Republic dialogue состоит из следующих полей:
  - id;
  - speaker;
  - listener;
  - text;
  - animation;
  - comment;
  - next;
  - previous;
  - source\_dlg;
  - audiofile.

Этот набор данных предлагается использовать для изучения диалогов и анализа тональности.

- Структура The Elder Scrolls documents:
  - url:
    - game;
    - url;
    - description;
    - author;
    - title;
    - text.

Первую версию этого набора данных использовали для адаптации анализатора тональности англоязычных текстов для учёта характерного для данной игровой серии лексикона [18].

Каждый из этих трёх наборов данных использует свою собственную структуру и хранит тексты для применения в различных целях. Более того, многие поля избыточны или затрудняют использование собранных данных. Например, поле `animation` избыточно, так как в игре может не быть анимации.

Таким образом, в [12] описаны три набора данных, которые нельзя считать сведёнными в корпус, так как они не имеют единой структуры. Для каждого

набора данных даны примеры применений, но они не обязательно совпадают или пересекаются.

### ПРЕДЛАГАЕМЫЙ ПОДХОД

Мы предлагаем другую структуру (хранящуюся в файле формата JSON) для корпуса видеоигровых текстов, содержащую только данные, необходимые для ПГК и обобщённых исследований. Каждая запись внутриигрового текста рассматривается как часть диалога.

- id:
  - id\_dialogue;
  - id\_entry;
  - requires;
  - opens;
  - speaker;
  - text;
  - comment.

id — неотрицательное число, идентифицирующее определённую часть текста,

id\_dialogue — неотрицательное число, идентифицирующее определённый диалог (отличие этого поля от предыдущего в том, что записи с уникальными id могут иметь одинаковый id\_dialogue),

id\_entry — неотрицательное число, идентифицирующее определённую часть диалога,

requires — список id: какие части текста игрок должен открыть, чтобы разблокировать данную часть,

opens — список id: какие части текста игрок разблокирует, открыв данную часть,

speaker — неотрицательное число, идентифицирующее персонажа (игрового или неигрового),

text — строка, содержащая «реплику»,

comment — внутриигровой контекст для данной записи.

Такая структура подходит для любого типа внутриигрового текста. Например, для текста на текстовых объектах (надписи, вывески, плакаты и т. п.), размещённых в игровом окружении, большая часть полей может быть оставлена пустой, в зависимости от игры задание может являться частью невырожденного диалога. Вырожденный диалог — любой текст, записанный в предложенном формате, не являющийся репликами между персонажами.

Также эта структура позволяет объединять тексты из любых разновидностей видеоигр. Применение предложенного подхода приведёт к созданию обобщённого корпуса, который можно использовать в дальнейших исследованиях.

Поля text и comment соответствуют критерию богатства [19], критерий переносимости [20] удовлетворяется хранением обработанных данных в файлах с форматом JSON.

### **АПРОБАЦИЯ ПРЕДЛАГАЕМОГО ПОДХОДА**

Мы обработали тексты из игры *James Bond 007: From Russia with Love* (версии для PSP<sup>7</sup>), которая удовлетворяет критерию репрезентативности. Мы использовали тексты диалогов с контекстными комментариями [21] и видео прохождения [22] для разделения диалогов.

Несмотря на то, что экшен<sup>8</sup>-игры обычно содержат меньшее количество текста по сравнению с ролевыми, на этапе обработки сценария мы столкнулись с неожиданной сложностью в анализе и структурировании имеющегося текстового контента.

Анализируя игру после её прохождения и обладая полным представлением о её сюжете и персонажах, мы можем установить, что для всех реплик определённого персонажа следует использовать одно и то же значение speaker, даже

---

<sup>7</sup> PSP— PlayStation Portable, портативный аналог домашней консоли.

<sup>8</sup> Экшен (от англ. action — действие) — процесс активного взаимодействия игрока с игровым миром, включающий интенсивные действия, такие как сражения, погони или преодоление препятствий, часто являющийся ключевым элементом геймплея в играх различных жанров. (термин также используется для жанра, характеризующимся динамичным игровым процессом, требующим от игрока быстрой реакции и координации)

если в тексте этот персонаж упоминается под разными именами или псевдонимами. Разработка алгоритма для автоматического выполнения этого процесса идентификации и унификации говорящих является ключевым моментом с точки зрения применимости всего подхода к анализу игровых сценариев.

Был также проведён эксперимент со сценарием игры *Ведьмак 3* [23]. Производилось сопоставление текста происходящему в игре. Когда находилась значимая информация, она переносилась на диаграмму (рис. 3). Уже на этом, начальном, этапе эксперимента мы столкнулись с несколькими сложностями:

1. Неполная информация. Некоторые фразы, описывающие локацию и произносимые главным героем, отсутствовали в сценарии.
2. Вариации в именовании сущностей. Несколько названий могут называть одну и ту же сущность.

Ранее такое встречалось во время обработки *James Bond 007: From Russia with Love*. Дополнительные имена не обязательно могут быть краткими формами основного имени. Например, «*one of the sorceress' conjured crustacean-like creatures*» и «*crab*» — это одно и то же существо.

3. Недостаточное понимание контекста. Без знания игры некоторые части сценария могут быть истолкованы неправильно.

Например, после первого видео на локации Kaer Morhen игрок получает контроль над персонажем. Можно взаимодействовать с объектами в первой комнате или вступить в диалог с другим персонажем. В сценарии представлено краткое описание «*If Geralt inspects things in Yen's room:*». Далее идут только реплики, которыми обмениваются персонажи. Информация о названиях объектов или том, что персонаж делает в данный момент, отсутствует.

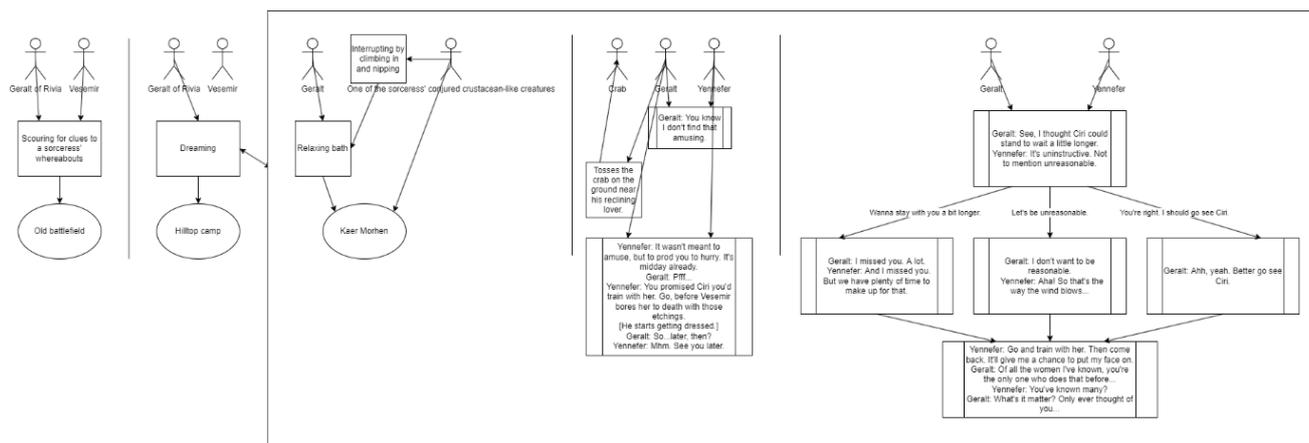


Рис. 3. Схема начала сценария игры *Ведьмак 3: Дикая Охота*

Перечисленные трудности подчёркивают важность понимания контекста и потребность в устойчивых алгоритмах, которые способны разбираться в вариациях наименований сущностей и извлекать точную информацию из сценариев игр. Несмотря на выявленные сложности, предложенный подход показал свою перспективность для создания универсального корпуса текстов видеоигр.

Однако для его успешной реализации требуется разработка дополнительных алгоритмов и методов обработки текста. Для преодоления выявленных трудностей необходимо:

- разработать алгоритмы для автоматической идентификации и унификации именования персонажей и других игровых сущностей,
- создать методы извлечения контекстной информации из игровых сценариев,
- исследовать возможности применения методов машинного обучения для улучшения понимания контекста игры.

Апробация предложенного подхода на реальных игровых сценариях позволила выявить ключевые проблемы в создании универсального корпуса текстов видеоигр. Полученные результаты демонстрируют необходимость дальнейшего совершенствования методов обработки игровых текстов и подчеркивают важность разработки специализированных алгоритмов для работы с игровым контентом. Преодоление выявленных трудностей откроет новые возможности для автоматизации процессов разработки игр и проведения исследований в области игровой индустрии.

## ЗАКЛЮЧЕНИЕ

В рамках продолжающегося исследовательского проекта (которому посвящены работы [11], [24–26] и ряд других) ведётся работа над решением общей задачи — генерацией видеоигровых прототипов на основе набора данных, в том числе текстовых (сценариев компьютерных игр). Эти прототипы должны быть такими, чтобы их можно было оценить и изменить специалистам.

В частности, в [11] было показано, что возможно генерировать раскадровки с использованием текста в качестве входных данных, поэтому некоторые из следующих логичных шагов — это создание коллекции видеоигровых текстов и обработка их для создания корпуса.

Будущие исследования будут в первую очередь сосредоточены на разработке алгоритма, который станет начальной стадией запланированного конвейера создания прототипов. Этот алгоритм должен отвечать следующим требованиям:

1. Принимать сценарии видеоигр в качестве входных данных.
2. Создавать структурированные файлы JSON, содержащие все внутриигровые тексты, отформатированные в соответствии с предложенной нами структурой файлов.
3. Включать такие методы, как распознавание именованных сущностей (Named Entity Recognition), связывание именованных сущностей (Named Entity Linking) и нормализация именованных сущностей (Concept Normalization), для решения выявленных в ходе эксперимента проблем. Однако важно отметить, что этого может оказаться недостаточно, поэтому мы признаём необходимость дальнейших исследований и поиска альтернативных решений.

После разработки алгоритма исследование будет продолжено в нескольких направлениях:

1. расширение корпуса добавлением обработанного текста из игр разных жанров для удовлетворения критерия разнообразия,
2. использование корпуса для ПГК,
3. извлечение информации из поля комментариев (comment) для получения информации о локациях, предметах и событиях.

Предложенный подход и дальнейшие исследования в этой области имеют потенциал значительно упростить и ускорить процесс разработки видеоигр, что особенно актуально в контексте растущей сложности и стоимости игровых проектов.

### Благодарности

Работа выполнена за счёт средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»).

### ЛУДОГРАФИЯ<sup>9</sup>

Bethesda Game Studios. (2007)[2006.] *The Elder Scrolls IV: Oblivion*. [видеоигра][загрузка][Windows, Xbox 360, PlayStation 3]. Bethesda Softworks, 2K, Ubisoft.

Bethesda Game Studios. (2022)[2011.] *The Elder Scrolls V: Skyrim*. [видеоигра][загрузка][Windows, Xbox 360, PlayStation 3, Xbox One, PlayStation 4, Switch, Xbox Series X|S, PlayStation 5]. Bethesda Softworks.

BioWare Edmonton. (2021) [2003.] *Star Wars: Knights of the Old Republic*. [видеоигра][загрузка][Xbox, Windows, macOS, iOS, Android, Switch]. LucasArts.

BioWare, Overhaul Games. (2019) [1998.] *Baldur's Gate*. [видеоигра][загрузка][Windows, macOS, iOS, Android, Linux, Switch, PlayStation 4, Xbox One]. Black Isle Studios, Interplay Entertainment, GraphSim Entertainment.

BioWare. (2019) [2000.] *Baldur's Gate II: Shadows of Amn*. [видеоигра][загрузка][Windows, macOS, iOS, Android, Linux, Switch, PlayStation 4, Xbox One]. Black Isle Studios, Interplay Entertainment.

---

<sup>9</sup> Лудография (от лат. *ludus* — «игра» и греч. *graphō* (γράφω) — «пишу») — список игр, упоминаемых в тексте, который отделён от библиографии. В лудографию должны быть включены все игры, упомянутые в тексте, даже если они не являются основным объектом исследования. Здесь описание каждой игры представлено в формате: <Авторский корпус. (год публикации)[год первой публикации]. *название игры* (версия).[тип игры][поддержка][платформа], издатель.> Формат выбран, исходя из рекомендаций <http://semionerd.blogspot.com/p/how-to-write-ludography.html>. Отметим, что использование термина может варьироваться в зависимости от конкретного исследования и контекста. В некоторых работах «ludography» может использоваться аналогично библиографии, но для списка игр, в то время как в других он может относиться к более глубокому анализу игрового процесса.

---

Blizzard Entertainment. (2024) [2004.] *World of Warcraft*. [видеоигра][загрузка][Windows, macOS]. Blizzard Entertainment.

CD Projekt Red. (2022) [2015.] *Ведьмак 3: Дикая Охота*. [видеоигра][загрузка][Windows, PlayStation 4, Xbox One, Switch, PlayStation 5, Xbox Series X|S]. CD Projekt.

Latitude. (2021) [2019.] *AI Dungeon*. [видеоигра][онлайн][браузер]. Latitude.

Rebellion Developments. (2006) [2006.] *James Bond 007: From Russia with Love*. [видеоигра][UMD][PlayStation Portable]. Electronic Arts, MGM Interactive.

Rockstar Games. (2025) [2025.] *Grand Theft Auto VI*. [видеоигра][загрузка][PlayStation 5, Xbox Series X|S]. Rockstar Games.

Runic Games. (2019) [2012.] *Torchlight II*. [видеоигра][загрузка][Windows, macOS, Linux, Switch, PlayStation 4, Xbox One]. Runic Games.

Starbreeze Studios, Overkill Software, Sumo Digital. (2023) [2013.] *Payday 2*. [видеоигра][загрузка][PlayStation 4, Switch, PlayStation 3, Xbox 360, Xbox One, Windows, Linux, GeForce Now]. Starbreeze Studios, 505 Games.

Treyarch, Raven Software. (2024) [2024.] *Call of Duty: Black Ops 6*. [видеоигра][загрузка][Windows, PlayStation 5, Xbox Series X|S, PlayStation 4, Xbox One]. Activision.

Valve Corporation, Gearbox Software. (2004) [1998.] *Half-Life*. [видеоигра][загрузка, CD][Windows, PlayStation 2, Linux, macOS]. Valve Corporation, Sierra Studios.

### СПИСОК ЛИТЕРАТУРЫ

1. Art Direction Bootcamp: Building a Creative Future with Artificial Intelligence. URL: <https://youtu.be/9FAXAgRrOSE?t=1737>.

2. Hu J., Wu C. Microsoft's Acquisition of Activision Blizzard: A Case Study // Highlights in Business, Economics and Management. 2023. Vol. 7. P. 363–371.

3. Behuria S. Microtransactions as a Business Model in Video Gaming Industry: Its Comparison with Traditional Model and Effect on Other Industries // Available at SSRN 4418003. 2022.

4. *Mahendra F. A., Jayadi R., Halim E.* Examining Drivers of Urge to Purchase in Video Game Microtransactions // *Journal of System and Management Sciences*. 2024. Vol. 14. No. 5. P. 131–147.

5. *Tavares R. et al.* Gamers' Reaction to the Use of NFT in AAA Video Games // *Procedia Computer Science*. 2023. Vol. 219. P. 606–613.

6. Exclusive: Ubisoft CEO speaks.

URL: <https://www.axios.com/newsletters/axios-gaming-534001ff-b32e-4a65-92c7-afb79f52dc13.html>.

7. Microsoft Raising Prices on New, First-Party Games Built for Xbox Series X|S to \$70 in 2023.

URL: [https://www.ign.com/articles/microsoft-raising-prices-new-first-party-games-xbox-series-70-2023-redfall-starfield?utm\\_source=facebook,twitter](https://www.ign.com/articles/microsoft-raising-prices-new-first-party-games-xbox-series-70-2023-redfall-starfield?utm_source=facebook,twitter).

8. 2K breaks gaming's de facto \$60 ceiling, asks \$70 for next-gen NBA 2K21.

URL: <https://arstechnica.com/gaming/2020/07/will-nba-2k21s-70-price-set-a-new-next-gen-console-standard/>.

9. The Convergence of AI and Creativity: Introducing Ghostwriter.

URL: <https://news.ubisoft.com/en-us/article/7Cm07zbBGy4Xm16WgYi25d/the-convergence-of-ai-and-creativity-introducing-ghostwriter>.

10. AI Dungeon. URL: <https://aidungeon.com/>.

11. *Sahibgareeva G.F., Bedrin O., Kugurakova V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // *CEUR Workshop Proceedings*. 2020. Vol. 2784. P. 267–282.

12. *van Stegeren J., Theune M.* Fantastic strings and where to find them: The quest for high-quality video game text corpora // *12th Intelligent Narrative Technologies Workshop, INT 2020*. — CEUR, 2021.

13. *Mitkov R. (ed.)*. The Oxford handbook of computational linguistics. Oxford university press, 2022. Vol. 1. 329 p.

14. *Anderson B.R., Smith A.M.* The Undergraduate Games Corpus: A Dataset for Machine Perception of Interactive Media // *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021. Vol. 35. No. 1. P. 3–11.

15. *van Stegeren J., Myśliwiec J.* Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation // Proceedings of the 16th International Conference on the Foundations of Digital Games. 2021. P. 1–8.

16. *Värtinen S., Hämmäläinen P., Guckelsberger C.* Generating Role-Playing Game Quests with GPT Language Models // IEEE Transactions on Games. 2022.

17. *Summerville A. et al.* Procedural Content Generation via Machine Learning (PCGML) // IEEE Transactions on Games. 2018. Vol. 10. No. 3. P. 257–270.  
<https://doi.org/10.1109/TG.2018.2846639>.

18. *Bergsma T., van Stegeren J., Theune M.* Creating a Sentiment Lexicon with Game-Specific Words for Analyzing NPC Dialogue in The Elder Scrolls V: Skyrim // Workshop on Games and Natural Language Processing. 2020. P. 1–9.

19. *Слива М.В.* Стратегии адаптивного построения в компьютерных играх // Компьютерные инструменты в образовании. 2013. № 3. P. 31–43.  
<https://doi.org/10.32603/2071-2340-2013-0-3-31-43>

20. *Суворова Т.Н., Rogozin К.И.* Подходы к оценке обучающих компьютерных игр // Вестник Оренбургского государственного педагогического университета. Электронный научный журнал. 2020. № 4. С. 380–394.  
<https://doi.org/10.32516/2303-9922.2020.36.26>

21. From Russia with Love – Game Script – PSP – By ironyisntdead.  
URL: <https://gamefaqs.gamespot.com/psp/929602-from-russia-with-love/faqs/75458>.

22. James Bond 007: From Russia with Love PSP Playthrough - With The Best Bond Ever, Sean Connery. URL: <https://youtu.be/dWckl9yd71M>.

23. Game Script (PS4) by Shotgunnova.  
URL: <https://gamefaqs.gamespot.com/ps4/702760-the-witcher-3-wild-hunt/faqs/73353>.

24. *Sahibgareeva G.F., Kugurakova V.V.* Branched Structure Component for a Video Game Scenario Prototype Generator // CEUR Workshop Proceedings. 2021. Vol. 3066. P. 101–111.

25. *Сахибгареева Г.Ф., Кугуракова В.В.* Редактор интерактивной структуры для инструмента генерации сценарных прототипов // Электронные библиотеки. 2022. Т. 24, № 6. С. 1184–1202.  
<https://doi.org/10.26907/1562-5419-2022-24-6-1184-1202>

---

26. Сахибгареева Г.Ф., Кугуракова В.В., Большаков Э.С. Инструменты балансирования игр // Электронные библиотеки. 2023. Т. 26, № 2. С. 225–251.

<https://doi.org/10.26907/1562-5419-2023-26-2-225-251>.

---

## **A NEW APPROACH TO CREATING A CORPUS OF VIDEO GAME TEXTS**

**N. R. Nurlygaianov**<sup>1</sup> [0000-0002-7969-0221], **V. V. Kugurakova**<sup>2</sup> [0000-0002-1552-4910]

<sup>1,2</sup>*Institute of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University, Kremlevskaia st., 35, Kazan, 420008*

<sup>1</sup>nikita25@mail.ru, <sup>2</sup>vlada.kugurakova@gmail.com

### **Abstract**

The problem of high and increasing cost of video game development is considered, and to solve it is proposed to apply procedural content generation, which will reduce development costs.

The work is a part of a large-scale research on automatic prototyping of video games and is devoted to the processing of game scenarios, i.e. natural language texts. It is proposed to extract the necessary entities from the scripts and pass them to further steps of the algorithm, which will generate game resources based on the textual descriptions.

There are several publications devoted to game text processing, in which several different structures for storing the extracted information are proposed. In this paper we propose a universal format that is suitable for processing the text of any video game and allows to create a corpus of texts for use in further research and automatic generation of game prototypes.

**Keywords:** *PCG, NLP, game development.*

### **REFERENCES**

1. Art Direction Bootcamp: Building a Creative Future with Artificial Intelligence. URL: <https://youtu.be/9FAXAgRrOSE?t=1737>.

2. Hu J., Wu C. Microsoft's Acquisition of Activision Blizzard: A Case Study // Highlights in Business, Economics and Management. 2023. Vol. 7. P. 363–371.

3. *Behuria S.* Microtransactions as a Business Model in Video Gaming Industry: Its Comparison with Traditional Model and Effect on Other Industries // Available at SSRN 4418003. 2022.

4. *Mahendra F. A., Jayadi R., Halim E.* Examining Drivers of Urge to Purchase in Video Game Microtransactions // Journal of System and Management Sciences. 2024. Vol. 14. No. 5. P. 131–147.

5. *Tavares R. et al.* Gamers' Reaction to the Use of NFT in AAA Video Games // Procedia Computer Science. 2023. Vol. 219. P. 606–613.

6. Exclusive: Ubisoft CEO speaks.  
URL: <https://www.axios.com/newsletters/axios-gaming-534001ff-b32e-4a65-92c7-afb79f52dc13.html>.

7. Microsoft Raising Prices on New, First-Party Games Built for Xbox Series X|S to \$70 in 2023.

URL: [https://www.ign.com/articles/microsoft-raising-prices-new-first-party-games-xbox-series-70-2023-redfall-starfield?utm\\_source=facebook,twitter](https://www.ign.com/articles/microsoft-raising-prices-new-first-party-games-xbox-series-70-2023-redfall-starfield?utm_source=facebook,twitter).

8. 2K breaks gaming's de facto \$60 ceiling, asks \$70 for next-gen NBA 2K21.  
URL: <https://arstechnica.com/gaming/2020/07/will-nba-2k21s-70-price-set-a-new-next-gen-console-standard/>.

9. The Convergence of AI and Creativity: Introducing Ghostwriter.  
URL: <https://news.ubisoft.com/en-us/article/7Cm07zbBGy4Xml6WgYi25d/the-convergence-of-ai-and-creativity-introducing-ghostwriter>.

10. AI Dungeon. URL: <https://aidungeon.com/>.

11. *Sahibgareeva G.F., Bedrin O., Kugurakova V.* Visualization Component for the Scenario Prototype Generator as a Video Game Development Tool // CEUR Workshop Proceedings. 2020. Vol. 2784. P. 267–282.

12. *van Stegeren J., Theune M.* Fantastic strings and where to find them: The quest for high-quality video game text corpora // 12th Intelligent Narrative Technologies Workshop, INT 2020. — CEUR, 2021.

13. *Mitkov R. (ed.)*. The Oxford handbook of computational linguistics. Oxford university press, 2022. Vol. 1. 329 p.

14. *Anderson B.R., Smith A.M.* The Undergraduate Games Corpus: A Dataset for Machine Perception of Interactive Media // Proceedings of the AAAI Conference on Artificial Intelligence. 2021. Vol. 35. No. 1. P. 3–11.

15. *van Stegeren J., Myśliwiec J.* Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation // Proceedings of the 16th International Conference on the Foundations of Digital Games. 2021. P. 1–8.

16. *Värtinen S., Hämäläinen P., Guckelsberger C.* Generating Role-Playing Game Quests with GPT Language Models // IEEE Transactions on Games. 2022.

17. *Summerville A. et al.* Procedural Content Generation via Machine Learning (PCGML) // IEEE Transactions on Games. 2018. Vol. 10. No. 3. P. 257–270.  
<https://doi.org/10.1109/TG.2018.2846639>.

18. *Bergsma T., van Stegeren J., Theune M.* Creating a Sentiment Lexicon with Game-Specific Words for Analyzing NPC Dialogue in The Elder Scrolls V: Skyrim // Workshop on Games and Natural Language Processing. 2020. P. 1–9.

19. *Sliva M.V.* Adaptive Construction Strategies in Computer Games [Strategii adaptivnogo postroeniya v komp'yuternykh igrakh]. Computer Tools in Education [Komp'yuternye instrumenty v obrazovanii]. 2013. № 3. S. 31–43. (In Russian).  
<https://doi.org/10.32603/2071-2340-2013-0-3-31-43>

20. *Suvorova T.N., Rogozin K.I.* Approaches to the Evaluation of Educational Computer Games [Podkhody k otsenke obuchayushchikh komp'yuternykh igr]. Bulletin of Orenburg State Pedagogical University. Electronic Scientific Journal [Vestnik Orenburgskogo gosudarstvennogo pedagogicheskogo universiteta. Elektronnyy nauchnyy zhurnal]. 2020. №. 4. S. 380–394 (In Russian).  
<https://doi.org/10.32516/2303-9922.2020.36.26>

21. From Russia with Love – Game Script – PSP – By ironyisntdead.  
URL: <https://gamefaqs.gamespot.com/psp/929602-from-russia-with-love/faqs/75458>.

22. James Bond 007: From Russia with Love PSP Playthrough – With The Best Bond Ever, Sean Connery. URL: <https://youtu.be/dWckl9yd71M>.

23. Game Script (PS4) by Shotgunnova.  
URL: <https://gamefaqs.gamespot.com/ps4/702760-the-witcher-3-wild-hunt/faqs/73353>.

24. *Sahibgareeva G.F., Kugurakova V.V.* Branched Structure Component for a Video Game Scenario Prototype Generator // CEUR Workshop Proceedings. 2021. Vol. 3066. P. 101–111.

25. *Sahibgareeva G.F., Kugurakova V.V.* Interactive Structure Editor for a Scenario Prototype Generation Tool [Redaktor interaktivnoy struktury dlya instrumenta generatsii stsenarnykh prototipov]. Electronic Libraries [Elektronnye biblioteki]. 2022. Vol. 24, No. 6. P. 1184–1202. (In Russian). DOI: 10.26907/1562-5419-2022-24-6-1184-1202

26. *Sahibgareeva G.F., Kugurakova V.V., Bolshakov E.S.* Game Balancing Tools [Instrumenty balansirovaniya igr]. Electronic Libraries [Elektronnye biblioteki]. 2023. Vol. 26, No. 2. P. 225–251 (In Russian).

<https://doi.org/10.26907/1562-5419-2023-26-2-225-251>.

---

## СВЕДЕНИЯ ОБ АВТОРАХ



**НУРЛЫГАЯНОВ Никита Рамильевич** — ассистент кафедры программной инженерии ИТИС КФУ. Сфера научных интересов — автоматизация процесса разработки компьютерных игр.

**Nikita Ramilevich NURLYGAIANOV** — Assistant of the Software Engineering Department of ITIS KFU. Research interest is automation of the computer game development process.

email: [nikita25@mail.ru](mailto:nikita25@mail.ru)

ORCID: 0000-0002-7969-0221



**КУГУРАКОВА Влада Владимировна** — доцент, к. т. н., и. о. зав. кафедрой индустрии разработки видеоигр ИТИС КФУ. Сфера научных интересов — различные аспекты проектирования игр.

**Vlada Vladimirovna KUGURAKOVA** — Associate Professor, acting head of the Video Game Development Industry Chair of ITIS KFU. Research interests include various aspects of game design.

email: [vlada.kugurakova@gmail.com](mailto:vlada.kugurakova@gmail.com)

ORCID: 0000-0002-1552-4910

*Материал поступил в редакцию 10 июня 2024 года*

---