Lingue e Linguaggi
Lingue Linguaggi 66 (2024), 241-266
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v66p241
http://siba-ese.unisalento.it, © 2024 Università del Salento

VIDEO GAME IN-GAME DIALOGUE
An analysis of lexical coverage

FRANCESCA BIANCHI
UNIVERSITÀ DEL SALENTO

Abstract – The impact of extramural encounters with English in video games on L2 English learning is high. However, linguistic analysis of the language of the scripted dialogue in video games is still scarce. The current study investigates the lexical coverage of the lines of non-player characters in a small corpus of nine games from three different genres. The analyses show a lexical coverage comparable to that of British and American films. Furthermore, our data suggest that lexical coverage differences largely depend on individual games rather than genre, although setting may play some role.

Keywords: scripted language; performed language; pop-culture; non-player characters; text coverage; lexical profile.

1. Introduction

The impact of extramural encounters with English in video games on L2 English learning is high (e.g. Sundqvist, Wikström 2015). Several empirical studies have shown positive correlation between frequency and or length of engagement with video games and language skills in the foreign language and vocabulary knowledge in particular (e.g. De Wilde *et al.* 2019; De Wilde, Eyckmans 2017; Kuppens 2010; Sylvén, Sundqvist 2012; Sundqvist, Wikström 2015; Rød, Calafato 2024, in Calafato, Clausen 2024).[1]

Indeed, as "action-and-goal-directed preparations for, and simulations of, embodied experience" (Gee 2008, p. 203), video games are a powerful means for situated learning (Gee 2006, 2008). Furthermore, narrative structures provide thematically coherent linguistic as well as graphic contextualisation cues, which supports language awareness and language learning, and in particular form-meaning-function associations. Thus, vocabulary is more readily learned and retained for a longer time when encountered in narrative texts, rather than in non-narrative ones (Reinhardt 2018). This is indirectly confirmed by analyses of specific game genres. In fact, results from a few studies suggest that different game profiles may facilitate or hinder language acquisition. For instance, driving games, which lack a

[1] Studies contradicting this finding also exist (e.g. Peters 2018; Muñoz 2020).

narrative structure and involve controlling a vehicle while navigating obstacles rather than dialogic situations, have been found to negatively correlate with vocabulary learning (Calafato, Clausen, 2024). More broadly, high interactivity (e.g. actively playing the game) has been shown to impede vocabulary acquisition compared to lower levels of interactivity (e.g. watching the game; deHaan *et al*. 2010). This suggests that slower-paced games might be more conducive to vocabulary learning than fast-paced ones.

Despite all this, the attention posed by linguists to the study of the scripted dialogue in video games, which creates the narrative structure of the game (Domsch 2017), is scarce (Schmitt 2019), especially if compared to other narrative pop cultural multimedia genres, such as films or TV series. Most research on video games, in fact, has been carried out in fields other than linguistics, on the visual features of video games, or on game paratext.

The current study investigates scripted dialogue in video games, i.e. the lines of non-player characters (NPCs), in a corpus of nine titles belonging to three genres and analyses it in terms of lexical coverage. The study was inspired by a similar study by Rodgers and Heidt (2021) and other studies on text coverage. Our study, however, differs from Rodgers and Heidt's in several ways. First of all, these authors conflated in a single corpus written and spoken texts, game dialogue and paratext, which makes it difficult to compare results with other works on lexical coverage. On the other hand, our corpus exclusively comprises the spoken lines of NPCs. Furthermore, their corpus is not fully balanced in terms of game types, while ours includes three games per type. Finally, we shall also consider game setting as a potentially interesting variable for lexical coverage.

The aims of the current study are to 1) analyse the lexical coverage of the scripted language used by NPCs in the game; 2) assess whether game type and/or game setting have an impact on lexical coverage; 3) verify Rodgers and Heidt's (2021, p. 222) hypothesis that "the language that one would tend to find in video games may be expected to be more similar to the type found in movies and television (based on the aural mode of the language)".

The remainder of this paper is organised as follows: Section 2 outlines the conceptual background of this study, introducing video games (2.1) and linguistic studies of video game discourse, with a special focus on video game lexis (2.2), the notions of text-coverage and lexical coverage (2.3), and the results of previous studies on lexical coverage (2.4). Section 3 presents the materials and methods used in the current analysis. Section 4 reports the results of our analyses. Finally, Section 5 compares our results to previous ones on similar text types, adding comments and ideas for future research.

2. Conceptual background

2.1 Video games

"Video game" is an umbrella term for an enormous amount of different types of games (Winget 2011), all played electronically, to the point that scholars find it difficult to agree on a single definition, or even on what such a definition should be based on.[2] Our interest in video games resides in their being widespread multimedia pop-cultural artefacts. Like all other pop-cultural products, they "have a commercial, entertainment-related purpose, are (mass-)mediated, fall within the mainstream, and represent largely fictional and scripted content" (Werner 2022, p. 1).

An infinite number of classifications of video games exist, based on platform, age of the players, type of interaction, setting, actions performed by the players, etc., but none of them is free of problems: borders are blurred, games cross boundaries, and categories and subcategories eventually overlap (Ensslin 2012, p. 42). Despite this, in keeping with previous research (Rodgers, Heidt 2021), an attempt will be made in the current study to focus on specific genres.

For genres, we shall use Adams' (2010) classification. In particular, games from three genres will be considered: action games; action-adventure games; and role-playing games. As Adams (2010, p. 392) explains "[a]n action game is one in which the majority of challenges presented are tests of the player's physical skills and coordination. […] In such games the player doesn't have time for strategy or planning". This category includes a range of subtypes, such as shooters and fighting games.[3] In this study, we shall consider action games that include some form or narration and dialogue between NPCs – the latter presenting and explaining the fictional world of the game to the player step by step (see Domsch 2017). On the other hand, "[t]he action-adventure is a hybrid genre, combining features from both action games and adventure games. To play them well requires a fair amount of physical skill, but they also offer a storyline, numerous characters, an inventory system, dialog, and other features of adventure games" (Adams 2010, p. 398). Finally, in role-playing games "the object […] is to experience a series of adventures in an imaginary world, through an avatar character or a small group of characters whose skills and powers grow as time goes on" (Adams 2010, p. 453). As a consequence of structure and aims, these three genres differ in terms of the relevance that

[2] See for example Adams (2010, Chapter 1) for a technical definition of what a game is and what is not and for the difference between game and video game. But see also Bergonse (2017) and Arjoranta (2019) for a heated debate on the (im)possibility of outlining a definition of video game.

[3] Other researchers, including Lee *et al*. (2014), classify shooters and fighting games as separate categories.

dialogue has within the playing experience. As Domsch (2017, p. 265) points out, generally, in-game dialogue is "less prominent in first person shooters [included in this research in the adventure category], more elaborate in action adventures, prevalent in role-playing games."

Our study will also consider the setting of the video games, [4] hypothesising that the setting of the action may influence the topics of character talk. Actually, NPCs' talk performs the primary diegetic functions of describing the fictional world of the game and supporting the progression of the narrative, alongside the ludic function of giving instructions to the player (Domsch 2017). Settings are rarely classified by scholars, but amply cited by game producers, reviewers, bloggers and players (e.g. the Wikipedia page 'Category: Video games by setting' lists an enormous number of possible settings).[5] To the best of our knowledge, the only academic paper attempting a classification of settings is Lee *et al*. (2014). Based on the then existing labels from scholarly, commercial, and popular sources, they examined the subject field of video game genres and divided it into fundamental categories each representing an essential characteristic of division of the subject field, which they call 'facets'. They identify several facets, including Setting and Mood/Affect. [6] They define setting as "the surroundings or environment (spatial or temporal) in which the game takes place" (Lee *at al*. 2014, p. 134), and Mood/Affect as "the pervading atmosphere or tone of the video game which evokes or recalls a certain emotion or state of mind" (Lee *at al*. 2014, p. 135). For each facet, foci, i.e. indexing terms, are listed. Their list of foci under Setting include categories such as Cyberpunk, Futuristic, Gothic, Historic, Medieval, Modern, Renaissance, and Steampunk, while under Mood/Affect they list Horror, Humorous, Dark, and Peaceful. By the authors' admission, this scheme is incomplete and provisional and has been devised to be updated and revised upon feedback from gamers. Their considering Horror as a Mood/Affect feature rather than a type of setting is interesting, but debatable from a linguistics perspective. Although it is definitely true that horror images are meant to scare, horror products are characterised by specific types of characters, such as werewolves, vampires, zombies, demons and other monsters – physical entities inhabiting the fictional world of the game. These entities inhabit the game's spatial environment and their presence impacts on the topics of character talk as they will be named, described, talked about or event talked to by NPCs. For this reason, we prefer to consider horror as a type

[4] "Games can have identical settings and yet belong to different genres, so a medieval role-playing game belongs to a different genre than a medieval war game" (Adams 2010, p. 390).

[5] https://en.wikipedia.org/wiki/Category:Video_games_by_setting (8.7.2024).

[6] The other facets are Gameplay, Style, Purpose, Target audience, Presentation, Artistic style, Temporal aspect, Point-of-view, Theme, and Type of editing.

of setting. We are aware however that our classification, too, may be subject to future revisions and improvements.

2.2 Linguistic studies of video game discourse

Video games have been studied from a variety of perspectives, including the analysis of player-game relationship and the socio-cultural role of video games (e.g. Fernández-Vara 2014), the rhetoric of video games (Bogost 2007; Paul 2012), the study of video game characters (e.g. Fizek 2012), and video games as forms of communication (Gee 2014) to mention but a few.

Indeed, their complex nature lends itself to several forms of investigation. A few studies have analysed interaction between their various semiotic layers. Most of these studies however focus on images, written text and other graphic elements, disregarding character dialogue (e.g. Stamenković, Jaćević 2019; Stamenković 2023). An exception is represented by Stamenković *et al.* (2016), who analysed the famous 1998 action-adventure video game, *Metal Gear Solid*. In their multimodal analysis – which considers dialogue among in-game characters alongside other semiotic resources – the authors observed a network of discursive strategies aimed at persuading players to refrain from killing enemies. These include the use of both emotional and rational argumentation, expressive language, literary language (poetic imagery, symbols and metaphors) alongside referential discourse aimed at situating the game's narrative within a specific political, military and cultural paradigm.

A substantial amount of text linguistics research has focused on video game paratext and the language of players interacting online in Massive Multiplayer Online Role Playing Games (MMORPGs). These two areas have, actually, been the most productive for linguistic studies to date. This is certainly due to the indisputable importance that paratext plays in drawing a player's attention to a specific video game, constructing the player-game relationship and building and bonding the player community (Ensslin 2012). However, the fact that paratext is much easier to retrieve than in-game text may have also contributed to this focus. One of the first significant studies in this respect was carried out by Ensslin (2012), who created a corpus including several types of paratext (video game magazine articles, threads in gamer forums and chats, and live conversations during gameplay) and used it to discuss the way gamers and developers view and talk about games, gaming and game players. Her analysis considers semantic, pragmatic, metaphorical, textual and ideological aspects of gaming discourse. More recent studies on video game paratext include Álvarez-Bolado Sánchez and Álvarez de Mon (2019) on video game reviews, Balteiro (2019) on forums, Gledhill (2019) on tutorials and walkthroughs, Campos-Pardillos (2019) on end-user agreements, as well as Hancock (2019) on manuals. Studies on player chats and interactions

during gaming include Ensslin and Finnegan (2019), Rudge (2019), Graham and Dutt (2019), Kiourti (2019) and Iaia (2021).

The academic works considering the scripted dialogue of NPCs are few and they generally have specific slants: Iaia (2016, Chapter 7) analysed the scripted lines of non-English NCPs speaking in English and compared their ways of speaking to real ELF variations as described in the literature, while Heritage (2020, 2022) and Rennick *et al*. (2023) focused on gender representation.

A broader scope of analysis is pursued by Dixon, who compiled a corpus of the scripted in-game language of four highly popular titles of single-player offline role-playing games (Dixon 2022, further refined in Dixon 2024). The corpus includes both spoken and written scripted text in the game. The texts were divided into registers; in particular, the spoken texts were divided into Immersive Speech (speech where the player is not prompted to respond) and Interactive Speech (speech instances involving the player to respond). The various registers were compared to real-life discourse using Biber's (1988) Involved versus Information Production dimension (MDA Dimension 1) (Dixon 2022). Both Immersive Speech and Interactive Speech resulted towards the positive end of the involved-informational continuum, though with mean dimension scores (23.61 and 29.8, respectively) that are slightly lower than that of face-to-face conversation (35.3). The author interpreted these results declaring that "[t]he linguistic environments of these games represent language exposure that compares quite similarly to real-world language use situations" (Dixon 2022, p. 183). Dixon's results are not surprising, considering that in-game spoken text is a form of scripted text (or 'performed' text) produced for asynchronous fruition by an extradiegetic audience (the player). As Werner (2021) observes, these specific production and reception circumstances are markedly different from those of informal conversation and can only produce a linguistic outcome which is generally slightly more formal than unscripted dialogue. More specifically, although performed text may often simulate dialogue and be used to present (pseudo-)intimacy between stage participants, scripted language is a form of second-orality, as it is written-to-be-spoken. This implies that scripted texts have been preplanned and carefully edited for specific purposes and to achieve a specific effect. The result is a form of language that lacks those conversational features which are directly connected to real-time processing constraints. Furthermore, the presence of an extradiegetic audience requires performed text to be more explicit that spontaneous conversation (see Werner 2022, p. 8-9, among others).

Finally, a few studies have engaged in analyses of the vocabulary that appears in video games from an applied linguistics perspective with a view to language learning. Rodgers and Heidt (2021) analysed lexical coverage in a corpus of ten video games, discussing its distribution across game types. This

study – which inspired the current one – is more thoroughly presented in Section 2.4. Its primary finding is the indication that the lexicon in video games might be difficult for beginner or intermediate learners. This inspired Heidt *et al*. (2023) to create a pedagogical word list of gaming vocabulary that could be used by creators of game-related material or teachers to scaffold learners prior to their engaging with video games. In their study, the researchers distinguished diegetic vocabulary, i.e. words that pertain to the narrative level, from ludic vocabulary, i.e. words that are key to the manipulation of the game. Taking advantage of Rodgers and Heidt's (2021) game corpus and employing analytical methods that include keyness calculation, lexical frequency profiling (see Section 2.3), manual reading of concordance lines and manual classification of items, they produced three lists of gaming vocabulary and identified three classes of ludic vocabulary: Game System, Game Property, and Game Information. Their study demonstrates the existence of a distinctive vocabulary associated with games, shows that several game words can function in both diegetic and ludic contexts,[7] and also suggests that the kind of vocabulary found in games is genre-dependent.

2.3. Text coverage and Lexical coverage as notions

When talking about text comprehension, the terms 'text coverage' and 'lexical coverage' are frequently used as synonyms. However, they are more like two sides of the same coin.

Text coverage is the percentage of words in a text that the reader should know in order to gain some understanding of its contents. In an empirical experiment where a fictional text was manipulated by inserting varying percentages of non-existent words, Hu and Nation (2000) found a predictable correlation between the percentage of unknown words and the participants' degree of understanding. In their experiment, when knowledge of the tokens in the text was lower than 90%, no participant showed understanding. When it was 90% and 95%, a few managed to understand the text, but not many. For adequate comprehension of their narrative text, most participants needed to know as much as 98% of the words. However, while Liu and Nation (1985) suggested that 95% coverage may represent the percentage of known words necessary to correctly guess words in context, Laufer (1989) suggested that knowing 95% of text coverage is enough to understand a text. For these reasons, the 95% and 98% thresholds are generally considered as reference

[7] As the authors point out, the distinction between ludic and diegetic language is not clear cut as both the words and their context contribute to attributing an expression to one or the other category. Furthermore, while this distinction is certainly interesting form a theoretical and research perspective, its role in text comprehension and vocabulary acquisition is still unclear (Heidt *et al*. 2023, p. 156).

points for discussions of lexical coverage (e.g. Webb, Rodgers 2009a, 2009b; Rodgers, Heidt 2021).

Lexical coverage, by contrast, is a measure of the vocabulary size readers need in order to understand a text without external support. The first attempts at calculating vocabulary size date back to the end of the 19th century, but it was the advent of electronic corpora that really contributed to advances in this area.[8] As Nation and Anthony (2017, p. 360) point out, three ways of measuring lexical coverage have been devised, respectively based on the count of words (Meara, Buxton 1987; Meara, Jones 1988), lemmas (Leech *et al.* 2001) or word families (Bauer, Nation 1993). However, the word-families model is the one that has most productively been used in L2 research. For this reason, this paper will resort to Bauer and Nation's technique and related software to directly compare our results to previous studies using the same metrics and techniques.

In Bauer and Nation's (1993, p. 253) own words,

> "a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately. […] The important principle behind the idea of a word family is that once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort. Clearly, the meaning of the base in the derived word must be closely related to the meaning of the base when it stands alone or occurs in other derived forms, for example, *hard* and *hardly* would not be members of the same word family."

These authors base the creation of word families on a seven-layer ranking of affixes, which is meant to mirror ease of learning for readers at different levels of morphological awareness. The inclusion of an affix at a specific level depends on the following criteria (Bauer, Nation 1993, pp. 255-256), frequency/likelihood being calculated against a large general corpus:

- Frequency, i.e. "the number of words in which the affix occurs";
- Productivity, i.e. "the likelihood that the affix will be used to form new words";
- Predictability, i.e. "the degree of predictability of the meaning of the affix";
- Regularity of the written/spoken form of the base;
- Regularity of the spelling/pronunciation of the affix;
- Regularity of function, i.e. "the degree to which the affix attaches to a base of known form-class and produces a word of known form-class".

The border between layers – though arbitrary – is nevertheless a useful way for standardising research on vocabulary size. Indeed, their approach has been

[8] For a comprehensive review of receptive and productive vocabulary size measures see Nation and Anthony (2017).

adopted in several studies of lexical frequency profiling (reviewed in Section 2.5), also thanks to the support of two software tools for quick and easy measure of lexical coverage – *Range* (Nation, Heatley 2002) and the more recent *AntWordProfiler* (Anthony 2022) – that have facilitated testing their measure on several different types of written and spoken texts and discussing their results with reference to L2 learning.

2.4. Studies on lexical coverage

Not unexpectedly, it has been observed that lexical coverage varies with text type.[9] A study by Nation on a range of written and spoken text types (Nation 2006) showed that from newspapers the most frequent 4,000 and 8,000 word families are respectively necessary to reach 95.39% and 98.31% text coverage; with a novel like *Lady Chatterly's Lover* similar thresholds could be reached with the most frequent 4,000 and 9,000 word families, respectively; while with a graded reader 2,000 and 3,000 word families are sufficient to reach 96.75% and 98.86%. In spontaneous conversation, 96.03% text coverage is reached with the 3,000 most frequent word families, but over 9,000 word families are required to reach the 98% threshold. On the other hand, in a movie like *Shrek*, 4,000 and 7,000 word families are needed for 96.74% and 98.08% text coverage, respectively. In a wider study on 318 American and English films, Webb and Rodgers (2009a) observed 95.76% coverage with a vocabulary of the most frequent 3,000 word families and 98.15% coverage with a vocabulary of 6,000 word families. If American and British films are considered separately, then the figures change slightly, but not substantially, with 3,000 word families and 6,000 families to reach, respectively, 95.76% and 98.14% for American films, and 3,000 and 7,000 word families to reach 95.50% and 98.29% for British ones. Larger differences could be observed between film genres, in particular to reach 98% coverage. While 3,000 word families were sufficient to reach 95% coverage in eight of the eleven film genres considered – war, animated and action films requiring knowledge of 4,000 word families for the same target –, the vocabulary needed to reach 98% coverage ranged from 5,000 to 10,000 word families – the 'least demanding' types of films being horror, crime and drama (5,000 word families), and the 'most demanding' being war (9,000 word families) and animated films (10,000 word families). The same authors have also investigated both American and British television genres, namely news, drama, situation comedy, older programmes, children's programmes, and science fiction (Webb, Rodgers 2009b). Considering the whole corpus – which comprised eighty-eight television programmes – the 95% threshold was reached with the most frequent 3,000 word families while the 98% threshold with the most frequent 7,000 word

[9] Text type is not the only variable influencing lexical coverage; other factors include length and homogeneity of the text (Hu, Nation 2000, p. 409).

families. Variation in coverage was observed between genres, with TV news being the most demanding, but above all between episodes of shows of the same genre. From 2,000 to 4,000 word families were required to reach the vocabulary size necessary to gain 95% coverage, while from 5,000 to 9,000 to gain 98% coverage.

Finally, to the best of our knowledge only two studies exist on the lexical coverage of the language of video games. Chien (2019) examined YouTube videos about *Minecraft*, a famous sand-box video game. They considered videos from three channels. One of the channels features spontaneous interaction between two American players filmed while playing *Minecraft* together. The other two channels feature monologic situations where the channel owner talks while playing or creating tutorials informing listeners how to play. Their analysis, performed with *AntWordProfiler* and the British National Corpus (BNC) as reference corpus, showed that 95% coverage is reached with 4,000 word families for monologic videos and as many as 6,000 for dialogic videos, while to reach 98% coverage more than 14,000 word families are necessary.

By contrast, Rodgers and Heidt (2021) focused on language within video games. These authors created a corpus of ten video games. The corpus comprises four role-playing games, two action-adventure games, two graphic adventure games and two life simulation games. The corpus included scripted in-game dialogues between characters, but also the texts of user interfaces, menus, in-game books, narration, tutorials, etc. (Rodgers, Heidt 2021, p. 219) and was analysed with *AntWordProfiler*. Considering the whole corpus, the first 5,000 word families proved necessary to reach 95% lexical coverage and the first 10,000 word families to reach 98% coverage. Furthermore, the authors noticed individual differences between video games, but no consistent differences between game genres.

Methodologically, in all the studies above, the coverage percentage mentioned is reached by adding marginal words – i.e. exclamations, interjections and the like – and proper nouns to the percentage figures, based on the assumption that these can be recognised and learnt with very little effort (Nation 2006). Furthermore, Nation (2006) and Webb and Rodgers (2009a, 2009b) took advantage of the *Range* programme and the BNC as reference corpus, while Rodgers and Heidt (2021) used *AntWordProfiler* and both the BNC and the Corpus of Contemporary American English (COCA). *Range* and *AntWordProfiler* respectively consider 14,000 and 24,000 word families and classify unrecognised words in a separate group called *Not in the lists*. Webb and Rodgers (2009a) observed the presence of many fairly common words in the *Not in the list* category, which suggests that their results – and by extension all counts with *Range* – could perhaps be rather conservative.

3. Materials and methods

This paper analyses the scripted language used by characters of nine video games belonging to three different types. In this respect, this dataset represents a relatively balanced subset of a larger corpus of video games that is currently under construction and was conceived as a pilot corpus for testing the available analytical methods and tools. The *Video game Pilot Corpus* (ViPiCo) includes three action-adventure games (AAG), comprising the AAG sub-corpus, three action games (AG), comprising the AG sub-corpus, and three role-playing games (RPG), comprising the RPG sub-corpus. The details of the pilot corpus are illustrated in Table 1; the codes in the Type-Setting column are detailed in Table 2..

File name	Type-Setting	Title	Year	Developers	Main story length (hours)[10]	Token count
1	AA-Ho	*Resident Evil Village*[11]	2021	Capcom	10	10,324
2	AA-Pa	*The Last of Us*[12]	2013	Naughty Dog	15	27,301
3	AA-F/PA	*Devil May Cry V*[13]	2019	Capcom	11	10,385
		AAG sub-corpus			**36**	**48,010**
4	A-SF	*Dead Space 2*[14]	2011	Visceral Games	9	10,167
5	A-Hi	*Mafia Definitive Edition*[15]	2020	Hangar 13	10.5	32,928
6	A-SF	*Halo Infinite*[16]	2021	343 Industries	11.5	14,535
		AG sub-corpus			**31**	**57,630**
7	RP-F	*Bloodborne*[17]	2015	From Software	33	12,429
8	RP-F	*Chrono Trigger*[18]	1995	Square	23	15,753
9	RP-F	*Pokémon Black*[19]	2010	Game Freak	32	33,836
		RPG sub-corpus			**88**	**62,018**
		ViPiCo				**167,658**

Table 1
Details of the games comprising the corpus.

Code	Type	Setting
AA-Ho	Action-Adventure	Horror
AA-PA		Post-apocalyptic

[10] Data in this column were obtained from the *HowLongToBeat* website. https://howlongtobeat.com/ (8.7.2024).
[11] https://www.residentevil.com/village/us/ (8.7.2024); https://en.wikipedia.org/wiki/Resident_Evil_Village (8.7.2024).
[12] https://www.playstation.com/en-gb/games/the-last-of-us-part-i/?country-selector=true (8.7.2024).; https://en.wikipedia.org/wiki/The_Last_of_Us (8.7.2024).
[13] https://www.devilmaycry.com/5/ (8.7.2024); https://en.wikipedia.org/wiki/Devil_May_Cry_5 (8.7.2024).
[14] https://www.ea.com/games/dead-space/dead-space-2 (8.7.2024); https://en.wikipedia.org/wiki/Dead_Space_2 (8.7.2024).
[15] https://mafiagame.com/mafia/ (8.7.2024); https://en.wikipedia.org/wiki/Mafia:_Definitive_Edition (8.7.2024).
[16] https://www.halowaypoint.com/ (8.7.2024); https://en.wikipedia.org/wiki/Halo_Infinite (8.7.2024).
[17] https://www.fromsoftware.jp/ww/detail.html?csm=094 (8.7.2024); https://en.wikipedia.org/wiki/Bloodborne (8.7.2024).
[18] https://www.jp.square-enix.com/chronotrigger/en/ (8.7.2024); https://en.wikipedia.org/wiki/Chrono_Trigger (8.7.2024).
[19] https://www.pokemon.com/us/pokemon-video-games/pokemon-black-version-and-pokemon-white-version (8.7.2024); https://en.wikipedia.org/wiki/Pok%C3%A9mon_Black_and_White (8.7.2024).

AA-F/PA		Fantasy/Post-apocalyptic
A-SF	Action	Science fiction
A-Hi		Historical
RP-F	Role-Playing	Fantasy

Table 2
Type-setting codes.

Unfortunately, the ViPiCo could not be balanced in terms of settings, nor in terms of token counts. The year of publication was also not considered as a separate variable. These certainly represent limitations of the current study, and this will be considered in the discussion of results and in prospective research.

For each game, the scripted language was retrieved from the Internet, from fans' transcripts (a method also used by Heritage 2020). More specifically, the following search string was used in the Google search engine: *name_of_game game script*. The scripts were then copied and pasted into text files. To make sure that the corpus exclusively included the characters' lines, the scripts were manually purged from comments and descriptions of action – which luckily always appeared in either square or round brackets – and of characters' names when these were used to mark the speaker. The scripts were also checked for evident typos, and the latter were fixed. Furthermore, for comparability with previous studies based on Nation's framework and tools (see Section 2.4), the spelling of contractions, connected speech, and hyphenated words was adjusted to conform to that used in the reference lists available with the software. All typos and variants were identified through manual scanning of word lists and fixed using the search-and-replace feature in MS Word.

In keeping with Rodgers and Heidt (2021, p. 220), the analyses were carried out using *AntWordProfiler* (Anthony 2022), the most recent software tool specifically created for the identification of lexical coverage using the word-families model (Nation, Anthony 2017), and the BNC and COCA word lists were employed as terms of comparison. Furthermore, coverage was calculated cumulating percentages and including proper nouns and marginal words; items in the *Not in the list* category were manually checked for proper nouns and marginal words to add to the cumulative counts. Indeed, in all files, the off-list category included some percentage of unusual proper nouns (e.g. Gbraakon; Geryon)[20] and imaginatively-spelt marginal words (e.g. *ooof*; *whewww*). These calculations are reported in the tables under the tag *Cum. Cov. %* .

During the manual analysis of the off-list items, it was observed that in one game (the shooter *Mafia Definitive Edition*), the *Not in the list* category also included common Italian words (e.g. *mamma*, *nonna*, *impara*). A check of

[20] Gbraakon is one of several Banished warships, which fought at the Battle of Zeta Halo, in the *Halo Infinite* video game; Geryon is a character in the Devil May Cry series, named after a giant of the Greek mythology.

the full transcript of the game showed that all lines in Italian were accompanied by English subtitles – appearing in parentheses right next to the Italian line. Since the English subtitles are included in the corpus and automatically analysed in terms of word families, the Italian words were disregarded.

4. Findings

Table 3 presents percentage token counts and percentage cumulative coverage including proper nouns and marginal words (*Cum. cov. %*) of the whole corpus, as well as its three sub-corpora. Percentage token counts are reported exactly as extracted by the software. Manual calculations of the proper nouns and marginal words in the off-list category – used for the cumulative coverage % measure – are reported in the comment to the table.

Percentage token counts show analogous trends in all the (sub-)corpora: 83.27%-87.12% of the running words belong to the most frequent 1,000 families, while the second set of 1,000 word families accounts for only 3.44%-4.96% of the running words; token counts gradually decrease as the word families increase. These results are in keeping with the token-count trends observed in all previous studies.

Considering measure *Cum. Cov. %*, 95% text coverage is reached with only 3,000 word families plus proper names and marginal words considering ViPiCo as a whole, as well as in the AG and in the RPG sub-corpora, while it requires as few as 2,000 word families in the AAG sub-corpus. On the other hand, greater variation is visible across the four corpora when it comes to reaching 98% coverage, with as few as 5,000 word families plus proper names and marginal words for the AAG sub-corpus, 7,000 for the RPG one and as many as 8,000 for the AG one. In the entire corpus, 7,000 word families plus proper names and marginal words are needed to reach this threshold.

The *Not in the list* category included quite a few proper nouns and some marginal words. The percentage values of proper nouns were the following: 0.24%, 0.59%, 2.77% and 1.3% in the AAG, AG, RPG and ViPiCo (sub-)corpora, respectively. The percentage values of marginal words were as follows: 0.06%, 0.01%, 0.15% and 0.08% in the AAG, AG, RPG and ViPiCo (sub-)corpora, respectively.

	ViPiCo		AAG sub-corpus		AG sub-corpus		RPG sub-corpus	
Word families	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**

1.000	85.46	89.87	88.29	92.57	85.44	89.44	83.27	88.33
2.000	4.08	93.95	3.12	**95.69**	3.93	93.37	4.96	93.29
3.000	1.61	**95.55**	1.27	96.97	1.70	**95.06**	1.79	**95.08**
4.000	1.04	96.59	0.69	97.65	1.13	96.20	1.22	96.30
5.000	0.82	97.41	0.59	**98.24**	0.69	96.89	1.12	97.42
6.000	0.48	97.89	0.32	98.56	0.57	97.46	0.50	97.92
7.000	0.27	**98.16**	0.15	98.71	0.40	97.86	0.24	**98.17**
8.000	0.24	98.39	0.21	98.92	0.25	**98.11**	0.25	98.41
9.000	0.18	98.57	0.19	99.11	0.13	98.24	0.21	98.62
10.000	0.10	98.68	0.06	99.17	0.08	98.32	0.16	98.79
11.000	0.10	98.78	0.05	99.22	0.13	98.45	0.12	98.90
12.000	0.07	98.85	0.05	99.28	0.09	98.54	0.07	98.97
13.000	0.07	98.92	0.03	99.31	0.06	98.59	0.10	99.07
14.000	0.03	98.95	0.02	99.33	0.03	98.63	0.04	99.12
15.000	0.04	98.99	0.04	99.37	0.03	98.66	0.04	99.16
16.000	0.01	99.00	0.01	99.38	0.01	98.67	0.01	99.17
17.000	0.01	99.01	0.01	99.39	0.02	98.69	0.01	99.18
18.000	0.03	99.04	0.00	99.40	0.06	98.75	0.02	99.20
19.000	0.02	99.06	0.00	99.40	0.02	98.76	0.03	99.23
20.000	0.02	99.08	0.03	99.42	0.00	98.77	0.04	99.26
21.000	0.01	99.09	0.02	99.45	0.01	98.78	0.01	99.27
22.000	0.00	99.1	0.01	99.45	0.00	98.79	0.00	99.27
23.000	0.01	99.1	0.01	99.46	0.01	98.79	0.01	99.28
24.000	0.10	99.2	0.00	99.46	0.21	99.01	0.06	99.34
25.000	0.00	99.2	0.00	99.46	0.01	99.01	0.00	99.34
proper nouns	1.56		1.84		2.39		0.56	
marginal words	1.54		2.14		1.00		1.58	
transparent compounds	0.30		0.36		0.37		0.17	
acronyms	0.07		0.03		0.08		0.09	
ignored	0.00		0.00		0.00		0.00	
not_in_lists	1.74		0.44		1.14		3.32	
TOTAL	100.00		100.00		100.00		100.00	

Table 3
Token count (%) and cumulative coverage (%) of ViPiCo and its three sub-corpora.

	Resident Evil Village		*The Last of Us*		*Devil May Cry V*	
Word families	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**
1.000	86.32	90.45	87.78	93.75	82.20	87.35
2.000	4.26	94.72	3.16	**96.43**	4.68	92.03
3.000	1.96	**96.67**	1.27	97.45	3.27	**95.29**
4.000	0.89	97.56	0.95	97.95	1.46	96.75
5.000	0.53	**98.10**	1.05	**98.39**	0.75	97.5
6.000	0.40	98.49	0.43	98.64	0.35	97.85
7.000	0.25	98.74	0.18	98.73	0.46	**98.31**
8.000	0.21	98.96	0.11	98.97	0.20	98.51
9.000	0.07	99.03	0.22	99.21	0.26	98.77
10.000	0.06	99.08	0.16	99.23	0.12	98.88
11.000	0.05	99.13	0.06	99.28	0.11	98.99
12.000	0.08	99.21	0.09	99.31	0.07	99.06
13.000	0.03	99.24	0.04	99.34	0.10	99.16
14.000	0.00	99.24	0.02	99.36	0.01	99.17
15.000	0.06	99.30	0.03	99.40	0.02	99.19
16.000	0.03	99.33	0.02	99.41	0	99.19
17.000	0.02	99.35	0.01	99.42	0.01	99.2
18.000	0.00	99.35	0.01	99.42	0	99.2
19.000	0.00	99.35	0.00	99.42	0	99.2
20.000	0.10	99.44	0.03	99.42	0	99.2
21.000	0.00	99.44	0.11	99.42	0	99.2
22.000	0.00	99.44	0.03	99.42	0.01	99.21
23.000	0.00	99.44	0.00	99.44	0	99.21
24.000	0.01	99.45	0.00	99.44	0	99.21
25.000	0.00	99.45	0.00	99.44	0	99.21
proper names	2.35		1.26		2.73	
marginal words	1.44		1.61		1.09	
transparent compounds	0.26		0.30		0.35	
acronyms	0.01		0.05		0.08	
ignored	0.00		0.00		0	
not_in_lists	0.61		1.03		1.68	
TOTAL	100		100		100	

Table 4
Token count (%) and cumulative coverage (%) of the games in the AAG sub-corpus.

	Dead Space 2		*Mafia Definitive Edition*		*Halo Infinite*	
Word families	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**
1,000	82.20	87.35	87.3	91.75	83.5	85.67
2,000	4.68	92.03	3.26	**95.01**	4.91	90.58
3,000	3.27	**95.29**	0.75	95.76	2.73	93.31
4,000	1.46	96.75	0.77	96.53	1.73	**95.04**
5,000	0.75	97.50	0.57	97.10	0.93	95.97
6,000	0.35	97.85	0.46	97.56	0.99	96.96
7,000	0.46	**98.31**	0.23	97.79	0.72	97.68
8,000	0.20	98.51	0.25	**98.04**	0.28	97.97
9,000	0.26	98.77	0.07	98.11	0.17	**98.14**
10,000	0.12	98.88	0.03	98.14	0.17	98.31
11,000	0.11	98.99	0.11	98.25	0.20	98.51
12,000	0.07	99.06	0.06	98.31	0.16	98.67
13,000	0.10	99.16	0.04	98.35	0.08	98.74
14,000	0.01	99.17	0.06	98.41	0.00	98.74
15,000	0.02	99.19	0.05	98.46	0.01	98.75
16,000	0.00	99.19	0.01	98.47	0.01	98.76
17,000	0.01	99.20	0.01	98.48	0.05	98.81
18,000	0.00	99.20	0.04	98.52	0.12	98.94
19,000	0.00	99.20	0.03	98.55	0.01	98.94
20,000	0.00	99.20	0.01	98.56	0.00	98.94
21,000	0.00	99.20	0.01	98.56	0.03	98.98
22,000	0.01	99.21	0.00	98.57	0.00	98.98
23,000	0.00	99.21	0.01	98.57	0.02	99.00
24,000	0.00	99.21	0.37	98.95	0.00	99.00
25,000	0.00	99.21	0.01	98.96	0.00	99.00
proper names	2.73		2.93		0.92	
marginal words	1.09		1.16		0.58	
transparent compounds	0.35		0.38		0.36	
acronyms	0.08		0.09		0.07	
ignored	0.00		0.00		0.00	
not_in_lists	1.68		0.93		1.24	
TOTAL	100		100		100	

Table 5
Token count (%) and cumulative coverage (%) of the games in the AG sub-corpus.

	Bloodborne		***Chrono Trigger***		***Pokémon Black***	
Word families	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**	**Token count %**	**Cum. Cov. %**
1.000	84.61	87.98	80.51	86.76	84.06	89.17
2.000	4.01	91.99	5.41	92.17	5.10	94.28
3.000	1.49	93.47	1.85	94.02	1.86	**96.14**
4.000	2.44	**95.91**	1.34	**95.36**	0.72	96.87
5.000	0.84	96.76	1.07	96.43	1.24	**98.11**
6.000	0.54	97.30	0.60	97.03	0.45	98.55
7.000	0.43	97.72	0.39	97.43	0.10	98.66
8.000	0.31	**98.03**	0.33	97.76	0.19	98.84
9.000	0.29	98.32	0.22	97.98	0.17	99.02
10.000	0.20	98.52	0.18	**98.16**	0.14	99.16
11.000	0.36	98.88	0.05	98.21	0.06	99.22
12.000	0.12	99.00	0.09	98.30	0.04	99.26
13.000	0.05	99.05	0.30	98.59	0.03	99.29
14.000	0.04	99.09	0.07	98.66	0.03	99.32
15.000	0.05	99.14	0.05	98.71	0.04	99.36
16.000	0.01	99.15	0.03	98.74	0.01	99.37
17.000	0.02	99.16	0.00	98.74	0.01	99.38
18.000	0.02	99.18	0.02	98.76	0.02	99.40
19.000	0.01	99.19	0.08	98.84	0.01	99.41
20.000	0.08	99.27	0.08	98.91	0.00	99.41
21.000	0.01	99.27	0.01	98.92	0.00	99.41
22.000	0.01	99.28	0.00	98.92	0.00	99.41
23.000	0.00	99.28	0.02	98.94	0.01	99.42
24.000	0.00	99.28	0.23	99.17	0.01	99.43
25.000	0.00	99.28	0.00	99.17	0.00	99.43
proper names	0.23		1.20		0.38	
marginal words	2.46		1.45		1.31	
transparent compounds	0.10		0.22		0.18	
acronyms	0.08		0.11		0.09	
ignored	0.00		0.00		0.00	
not_in_lists	1.21		4.10		3.72	
TOTAL	100		100		100	

Table 6
Token count (%) and cumulative coverage (%) of the games in the RPG sub-corpus.

Within the sub-corpora ample internal variation can be observed (Tables 4-6). Considering the *Cum. Cov. %* measure, in the AAG sub-corpus, the number of word families needed to reach 95% coverage ranges from 2,000 in *The Last of Us*, to 3,000 in *Resident Evil Village* and *Devil May Cry V*, while the number needed to reach 98% coverage amounts to 5,000 in *Resident Evil Village* and *The Last of Us* and as many as 6,000 in *Devil May Cry V*. In the AG sub-corpus,

the number of word families needed to reach 95% coverage goes from 2,000 in *Mafia Definitive Edition*, to 3,000 in *Dead Space 2*, to 4,000 in *Halo Infinite*, while the number needed to reach 98% coverage amounts to 6,000 in *Dead Space 2*, 8,000 in *Mafia Definitive Edition* and as many as 9,000 in *Halo Infinite*. Finally, in the RPG sub-corpus, both *Bloodborne* and *Chrono Trigger* require 4,000 word families to reach 95% coverage; they however differ as for the 98% coverage, needing, respectively, 8,000 and 10,000 word families. On the other hand, *Pokémon Black* requires as few as 3,000 and 5,000 word families to reach the first and second thresholds, respectively.

If we group the video games by setting (Table 7), no clearer picture emerges. This is largely a consequence of the fact that only two settings out of six are represented by more than one game in the PiViCo.

Setting type	**Game title**	**95% coverage, reached at**	**98% coverage, reached at**
Historical	*Mafia Definitive Edition*	2,000	8,000
Horror	*Resident Evil Village*	3,000	5,000
Post-Apocalyptic	*The Last of Us*	2,000	5,000
Fantasy/Post-Apocalyptic	*Devil May Cry V*	3,000	7,000
Fantasy	**All F games in the ViPiCo**	3,000	7,000
	Bloodborne	4,000	8,000
	Chrono Trigger	4,000	10,000
	Pokémon Black	3,000	5,000
Science Fiction	**All SF games in the ViPiCo**	3,000	5,000
	Halo Infinite	4,000	9,000
	Dead Space 2	3,000	7,000

Table 7
Summary of lexical coverage by setting.

The only two categories including more than one video game are Fantasy and Science Fiction, and within each of them considerable variation can be observed. Unfortunately, here the other categories are all represented by a single game each. However, we could try and argue that the settings fully grounded in the world as it has really existed at some point in time – i.e. the Historical and the Post-Apocalyptic settings – require both as few as 2,000 world families plus proper nouns and marginal words to reach 95% coverage, while the games that are fully or partially set in non-existing worlds – i.e. all the other games in the corpus – require 3,000-4,000 word families plus proper nouns and marginal words for the same threshold. This is the minimum number of word families needed to be able to guess the meaning of unknown words in context. For a better understanding of the contents of the dialogues, on the other hand, 98% lexical coverage is needed, and in this respect, there seems to be little correlation between number of word families and setting.

However, Table 7 presents an interesting fact: in the two categories comprising more than one game, the number of word families plus proper nouns and marginal words needed to reach 95% and 98% coverage is

systematically lower – and in the case of the 98% threshold also considerably lower – when we look at the whole set of games sharing the same type of setting (Fantasy – All, and Science Fiction – All), compared to the individual games that comprise the group. This suggests that the games in each of these setting categories share several word families. No such clear-cut picture can be observed when grouping games by type (Table 8).

Game type	**Game title**	**95% coverage, reached at**	**98% coverage, reached at**
Action-Adventure	**AAG sub-corpus**	2,000	5,000
	Resident Evil Village	3,000	5,000
	The Last of Us	2,000	5,000
	Devil May Cry V	3,000	7,000
Action	**AG sub-corpus**	3,000	8,000
	Dead Space 2	3,000	7,000
	Mafia Definitive Edition	2,000	8,000
	Halo Infinite	4,000	9,000
Role-Playing	**RPG sub-corpus**	3,000	7,000
	Bloodborne	4,000	8,000
	Chrono Trigger	4,000	10,000
	Pokémon Black	3,000	5,000

Table 8
Summary of lexical coverage by type.

This suggests that setting could be more relevant than game type in determining the words used in the dialogues, and that it may be easier for a learner to understand game dialogues if, when choosing a new game to play, they stick to the same type of setting. However, further investigation with a much larger and more balanced dataset is needed to confirm this hypothesis.

5. Discussion and conclusions

Our first aim in this study was to analyse the lexical coverage of the scripted language used by NPCs in the game. Our analyses show that, in order to understand the NPCs lines of the games in our corpus as a whole, a player should be familiar with the most frequent 3,000 word families to be able to guess unknown words in context (95% coverage) and as many as 7,000 word families to fully understand the dialogue (98% coverage). If we consider sub-corpora divided by genre, 95% coverage is reached with 3,000 word families in the AG and RPG sub-corpora, and as few as 2,000 word families in the AAG sub-corpus, while 98% coverage with as few as 5,000 word families for the AAG sub-corpus, 7,000 for the RPG one and as many as 8,000 for the AG one. However, like Rodgers and Heidt (2021), we observed ample internal variation in each (sub)corpus.

Another aim was to assess whether game type and/or game setting have an impact on lexical coverage. Our data thus support Rodgers and Heidt's (2021) observation that lexical coverage differences largely depend on individual games rather than genre. Their findings were based on four game types, two of which also appear among the types analysed in this paper, but their corpus was not balanced for game type, and most types were represented by only two games. In our corpus, which is balanced by type, despite clear variation within each genre, dialogues in Action games as a whole would appear to be lexically slightly more challenging than in Role-Playing games, which in turn would seem to be more demanding than Action-Adventure game dialogues. This result suggests not to totally discard the possibility of genre-dependent differences and to investigate further using a much larger corpus. Indeed, the pedagogical lists of gaming vocabulary produced by Heidt *et al*. (2023) show that at least some kind of vocabulary found in games is genre-dependent.

Furthermore, our tentative analysis of games divided by setting suggests that setting does probably play a role in determining the words used in the dialogues. In particular, from a learner perspective, it suggests that it may be easier for a learner to guess unknown words when the game is grounded in reality and to understand game dialogues if, when choosing a new game to play, they continue with the same type of setting. Consequently, since many games come in series, with sequels and prequels it would be possible to hypothesise that to play titles from the same series – if they repeat the same setting, which is not always the case (e.g. the *Indiana Jones* series) – may represent an excellent way to scaffold language acquisition. This however should be verified on individual titles, as research on other types of scripted dialogue, namely telecinematic products, has observed ample lexical variation between episodes of the same series (Webb, Rodgers 2009b). It could also be interesting to investigate whether specific genre-setting configurations prove easier or more difficult for learners, as it has been observed that in films of specific genres-settings, such as horror and crime, dialogue is less lexically demanding than in others such as war (Webb, Rodgers 2009a). In any case, a corpus representing a much wider number of video games for each genre and setting and a wider variety of genres including more clear-cut ones (e.g. strategy games; simulation games; visual novels) is necessary to confirm or refine these hypotheses.

The third aim of our study was to see whether the lexical profile of in-game dialogue is somehow similar to that of other aural scripted dialogic discourse, such as movies and television. In this respect, our corpus, exclusively comprising in-game dialogue, reaches 95% and 98% lexical coverage with 3,000 and 7,000 word families respectively. These values are much lower than those identified by Rodgers and Heidt (2021) in a video game

corpus conflating both written and spoken text (5,000 and 10,000 word families for 95% and 98% coverage respectively). At the same time, our data are largely in line with those observed by Webb and Rodgers (2009a) on British and American films, where 95% coverage was reached with the most frequent 3,000 word families, and 98% coverage with 6,000-7,000 word families. This confirms the hypothesis of Rodgers and Heidt (2021, p. 222) that in-game dialogue resembles that of other spoken media. It also supports their hypothesis that the higher lexical demands observed in their analysis are likely to result from having included in the corpus user interfaces, tutorials, equipment inventories and the like. In fact, such types of text – performing a ludic function – include low frequency vocabulary such as *autosave*, *gameplay*, *overwrite*, *melee*, or *recoil* (Heidt *et al*. 2023).

The ViPiCo is definitely very small, limited in terms of number and genre of video games, and unbalanced in terms of settings, but we believe it nevertheless offered useful data. Besides the above conclusions, our data lead us to make a few more comments. First of all, the ViPiCo seems to challenge Domsch's (2017) hypothesis about the prominence of dialogue in the three genres considered (see Section 2.1). Ample variation can be observed between games within and across genres. Furthermore, if we consider main story length and calculate the average number of tokens per hour of gameplay, the genre with the greatest amount of dialogue appears to be the action-adventure one, with an average of 1,859 words uttered by NPCs per hour, immediately followed by the action genre, with 1,334 words per hour, and last comes the role-play genre, with as few as 705 words per hour of gameplay. The year when the game was published does not help in explaining this. Our best hypothesis is that the amount of dialogue (calculated in number of words) is determined by the multimodal construction of individual games. Indeed, the role of individual games has repeatedly come to the fore throughout this research, which is certainly a direct consequence of the complex and ever-evolving nature of video games. This should not however let scholars refrain from looking at groups of games sharing at least one feature in the search for common trends.

Furthermore, in order to better understand the nature of the language gamers are exposed to with in-game dialogue, it is important to go beyond lexical profiling and engage in semantic, syntactic and pragmatic analyses of NPC discourse, on its own and compared to other forms of spoken, dialogic discourse. In particular, corpus-linguistic investigations taking advantage of POS and semantic taggers could provide interesting insight into the topics game players encounter, which could then inform pedagogical uses of off-the-shelf video games. Something has been done in this direction (see Section 2.2), but several angles remain largely unexplored.

Finally, an area that deserves further research is that of video game paratext in relation to its capability to support preliminary acquisition of fundamental vocabulary prior to the actual gameplay, and to allow players with lower levels of knowledge of English to better engage with the game itself. Forums and game reviews are of course relevant, but the most interesting type of paratext could be walkthrough videos, and in particular monologic ones, as they have been observed to have simpler lexical profiles than dialogic ones (Chien 2019).

Bionote: Francesca Bianchi holds a PhD in English linguistics from Lancaster University (UK). She is associate professor of English linguistics and translation at the University of Salento (Lecce, Italy). Her research interests include corpus linguistics as an analytical method for the study of text types and forms of discourse, multimedia texts and audiovisual translation, and the use of technology in the teaching and learning of English, translation and interpreting.

Author's address: francesca.bianchi@unisalento.it

Acknowledgements: I would like to thank Tiziana Ciabattoni (University of Pavia) for her precious support, suggestions and double-checking of the quantitative data.
This work was supported by the Italian Ministry for University and Research (PRIN project grant 2020NNJTW3_001).

References

Adams E. 2010, *Fundamentals of Game Design*, 2nd ed., New Riders, Berclay, CA.

Álvarez-Bolado Sánchez C. and Álvarez de Mon I. 2019, *Video games: A Lexical Approach*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 13-38.

Anthony L. 2022, *AntWordProfiler*, version 2.0.1, [Computer Software]. Waseda University, Tokyo. http://www.laurenceanthony.net/software

Arjoranta J. 2019, *How do define games and why we need to*, in "The Computer Games Journal" 8, pp. 109-120.

Balteiro I. 2019, *Lexical and Morphological Devices in Gamer Language in Fora*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, p. 39-57.

Bauer L. and Nation P. 1993, *Word families*, in "International Journal of Lexicography" 6 [4], pp. 253-279.

Bergonse R. 2017, *Fifty Years on, What exactly is a video game? An essentialistic definitional approach*, in "The Computer Games Journal" 6 [4], pp. 239-255.

Bogost I. 2007, *Persuasive Games. The Expressive Power of Video games*, The MIT Press, Cambridge, MA.

Calafato R. and Clausen T. 2024, *Vocabulary learning strategies in extramural English gaming and their relationship with vocabulary knowledge*, in "Computer Assisted Language Learning", pp. 1-19. https://www.tandfonline.com/doi/full/10.1080/09588221.2024.2328023 (15.7.2024).

Campos-Pardillos M.A. 2019, *End-user Agreements in Video games: Plain English at Work in an Ideal Setting*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 116-135.

Chien Y.C. 2019, *The Language of Massively Multiplayer Online Gamers: A Study of Vocabulary in Minecraft Gameplay*, in "TESL-EJ" 23 [3], pp. 1-16. https://tesl-ej.org/pdf/ej91/int.pdf (8.7.2024).

De Wilde V., Brysbaert M. and Eyckmans J. 2019, *Learning English through out-of-school expo-sure. Which levels of language proficiency are attained and which types of input are important?*, in "Bilingualism" 23 [1], pp. 171-185.

De Wilde V. and Eyckmans J. 2017, *Game on! Young learners' incidental language learning of English prior to instruction*, in "Studies in Second Language Learning and Teaching" 7 [4], pp. 673-694.

deHaan J., Reed W.M. and Kuwada K. 2010, *The effect of interactivity with a music video game on second language vocabulary recall*, in "Language Learning & Technology" 14 [2], pp. 74-94.

Dixon D.H. 2022, *The Language in Digital Games: Register variation in virtual and real-world contexts*, PhD dissertation, Northern Arizona University, available from ProQuest Dissertations & Theses.

Dixon D.H. 2024, *Introducing the Single Player Offline Game Corpus (SPOC): A corpus of seven registers from digital role-playing games*, in "Corpora" 19 [1], pp. 107-122.

Domsch S. 2017, *Dialogue in video games*, in Mildorf J. and Thomas B. (Eds), *Dialogue across media*, John Benjamins, Amsterdam, pp. 251-270.

Ensslin A. 2012, *The Language of Gaming*, Palgrave Macmillan, London.

Ensslin A. and Finnegan J. 2019, *Bad Language and Bro-up Cooperation in Co-sit Gaming*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 139-156.

Fernández-Vara C. 2014, *Introduction to Game Analysis*, Routledge, London/New York.

Fizek S. 2012. *Pivoting the Player. A methodological toolkit for player character research in offline role-playing games*. ProQuest Dissertations Publishing, Bangor University, Bangor, UK. https://www.proquest.com/openview/385e0788aac018f08ac6021469e42fc1/1?pq-origsite=gscholar&cbl=51922 (12.7.2023).

Gee J.P. 2006, *Are video games good for learning?*, in "Nordic Journal of Digital Literacy" 1 [3], pp. 172-183.

Gee J.P. 2008, *Learning theory, video games and popular culture*, in Drotner K. and Livingstone S. (Eds.), *The International Handbook of Children, Media and Culture*, SAGE Publications, London, pp. 196-213.

Gee J.P. 2014, *Unified Discourse Analysis: Language, Reality, Virtual Worlds and Video Games*, Routledge, London/New York.

Gledhill C. 2019, *Phraseology and Lexico-grammatical Patterns in Two Emergent Paragame Genres: Video game Tutorials and Walkthroughs*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 58-86.

Graham S.L. and Dutt S. 2019, *"Watch the Potty Mouth": Negotiating Impoliteness in Online Gaming*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 201-223.

Hancock M. 2019, *Playing It By the Book: Instructing and Constructing the Player in the Video game Manual Paratext*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 288-303.

Heidt J., Pinchbeck G.G. and Rodgers M.P.H. 2023, *The Good Gaming (GG) list: Key vocabulary in video games*, in Reynolds B.L. (Ed.), *Vocabulary Learning in the Wild*, Springer, Singapore, pp. 143-161.

Heritage F. 2020, *Applying corpus linguistics to video game data: Exploring the representation of gender in video games at a lexical level*, in "*Game studies*" 20 [3], n.p.. https://gamestudies.org/2003/articles/heritage_frazer (19.6.2023).

Heritage F. 2022, *Magical women: Representations of female characters in the Witcher video game series*, in "Discourse, Context & Media" 49, 100627, pp. 1-9.

Hu M. and Nation P. 2000, *Vocabulary density and reading comprehension*, in "Reading in a Foreign Language" 13 [1], pp. 403-430.

Iaia P. 2016, *Analysing English as a Lingua Franca in Video Games. Linguistic Features, Experiential and Functional Dimensions of Online and Scripted Interactions*, Peter Lang, Berlin.

Iaia P. 2021, *Towards a 'COOPING' Model for the Investigation of Gamers' Online Conversations in English*, in "Iperstoria" 17, pp. 222-240.

Kiourti E. 2019, *"Shut the Fuck up Re! Plant the Bomb Fast!": Reconstructing Language and Identity in First-person Shooter Games*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 157-177.

Kuppens A.H. (2010), *Incidental foreign language acquisition from media exposure*, in "Learning, Media and Technology" 35 [1], pp. 65-85.

Laufer B. 1989, *What percentage of text lexis is essential for comprehension?*, in Lauren C., Nordman M. (Eds.), *Special language: From humans thinking to thinking machines*, Multilingual Matters Clevedon, England, pp. 316-323.

Lee J.H., Karlova N., Clarke R.I., Thornton K. and Perti A. 2014, *Facet Analysis of Video Game Genres*, Paper presented at iConference, Berlin, Germany.

Leech G., Rayson P. and Wilson A. 2001, *Word Frequencies in Written and Spoken English*, Longman, Harlow.

Liu N.A. and Nation P. 1985, *Factors affecting guessing vocabulary in context*, "RELC Journal" 16, pp. 33-42.

Meara P. and Buxton B. 1987, *An alternative to multiple choice vocabulary tests*, in "Language Testing" 4 [2], pp. 142-151.

Meara P. and Jones G. 1988, *Vocabulary size as a placement indicator*, in Grunwell P. (Ed.), *Applied Linguistics in Society*, CILT, London, pp. 80-87.

Muñoz C. 2020, *Boys like games and girls like movies. Age and gender differences in out-of-school contact with English*, in "Revista Española de Lingüística Aplicada" 33 [1], pp. 172-202.Nation P. 2006, "How large a vocabulary is needed for reading and listening?", *Canadian Modern Language Review* 63 [1], pp. 59-82.

Nation P. and Anthony L. 2017, *Measuring vocabulary size*, in Hinkel E. (Ed.), *Handbook of Research in Second Language Teaching and Learning*, vol. III, Routledge, London/New York, pp. 355-368.

Nation P. and Heatley A. 2002, *Range: A program for the analysis of vocabulary in texts*, software. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs (1.4.2023).

Paul C.A. 2012, *Wordplay and the Discourse of Video Games. Analysing Words, Design and Play*, Routledge, New York.

Peters E. 2018, *The effect of out-of-school exposure to English language media on learners' vocabulary knowledge*, in "ITL – International Journal of Applied Linguistics" 169 [1], pp. 142-167.

Reinhardt J. 2018, *Gameful second and foreign language teaching and learning: Theory, research, and practice*, Springer, New York.

Rød A.J., and Calafato R. 2024, *Exploring extramural English: Impacts, integration and future directions*, in Skulstad A.S. (Ed.), *Current issues in English teaching*, Fagbokforlaget, pp. 117-140.

Rodgers M.P.H. and Heidt J. 2021, *Levelling up comprehensible input and vocabulary learning. The lexical profile of video games*, in Werner V., Tegge F. (Eds.), *Pop Culture in Language Education. Theory, research, practice*, Routledge, London, pp. 215-227.

Rudge L.A. 2019, *"I Cut It and I … Well Now What?": (Un)Collaborative Language in Timed Puzzle Games*, in Ensslin A., Balteiro I. (Eds.), *Approaches to Video game Discourse. Lexis, interaction, textuality*, Bloomsbury, New York, pp. 178-200.

Schmitt N. 2019, *Understanding vocabulary acquisition, instruction, and assessment: A research agenda*, in "Language Teaching" 52 [2], pp. 261-274.

Stamenković D. 2023, *The stylistic journey of a video game. A diachronic approach to multimodality in the Football Manager series*, in Schubert C. and Verner W. (Eds.), *Stylistic Approaches to Pop Culture*, Routledge, New York, pp.227-246.

Stamenković D. and Jaćević M. 2019, *Video Games and Multimodality: Exploring Interfaces and Analyzing Video Game Screens Using the GeM Model*, in Wildfeuer J. Pflaeging J., Bateman J., Seizov O. and Tseng C.-I (Eds.), *Multimodality. Disciplinary Thoughts and the Challenge of Diversity*, De Gruyter, Berlin, pp.277-294.

Stamenković D., Jaćević M. and Wildfeuer J, 2016, *The persuasive aims of Metal Gear Solid: A discourse theoretical approach to the study of argumentation in video games*, in "Discourse, Context and Media" 15, pp. 11-23.

Sundqvist P. and Wikström P. 2015, *Out-of-school digital gameplay and in-school L2 English vocabulary outcomes*, in "System" 51, pp. 65-76.
Sylvén L.K. and Sundqvist P. 2012, *Gaming as extramural English L2 learning and L2 proficiency among young learners*, in "ReCALL" 24 [3], pp. 302-321.
Webb S. and Rodgers M.P.H. 2009a, *The lexical coverage of movies*, in "Applied Linguistics" 30 [4], pp. 407-427.
Webb S. and Rodgers M.P.H. 2009b, *Vocabulary demands of television programs*, in "Language Learning" 59 [2], pp. 335-366.
Werner V. 2021, *Text-linguistic analysis of performed language: Revisiting and remodeling Koch & Oesterreicher*, in "Linguistics" 59 [3], pp. 541-575.
Werner V. 2022, *Pop Cultural Linguistics*, in *Oxford Research Encyclopedia of Linguistics*, Oxford University Press, online. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-999 (8.7.2024).
Winget M.A. 2011, *Video game preservation and massively multiplayer online role-playing games: A review of the literature*, in "Journal of the American Society for Information Science and Technology" 62 [10], pp. 1869-1883.