Simulation & Gaming
http://sag.sagepub.com/

Natural Language Processing in Game Studies Research : An Overview

José P. Zagal, Noriko Tomuro and Andriy Shepitsen

Simulation Gaming 2012 43: 356 originally published online 12 October 2011
DOI: 10.1177/1046878111422560

Simulation & Gaming
43(3) 356–373

DOI: 10.1177/1046878111422560
http://sag.sagepub.com

Natural Language Processing in Game Studies Research: An Overview

José P. Zagal[1], Noriko Tomuro[1], and Andriy Shepitsen[1]

Abstract

Natural language processing (NLP) is a field of computer science and linguistics devoted to creating computer systems that use human (natural) language as input and/or output. The authors propose that NLP can also be used for game studies research. In this article, the authors provide an overview of NLP and describe some research possibilities that can be explored using NLP tools and techniques. The authors discuss these techniques by performing three different types of NLP analyses of a significant corpus of online videogame reviews: (a) By using techniques such as word and syllable counting, the authors analyze the readability of professionally written game reviews, finding that, across a variety of indicators, game reviews are written for a secondary education reading level; (b) the authors analyze hundreds of thousands of user-submitted game reviews using part-of-speech tagging, parsing, and clustering to examine how gameplay is described. The findings of this study in this area highlight the primary aesthetics elements of gameplay according to the general public of game players; and (c) the authors show how sentiment analysis, or the classification of opinions and feelings based on the words used in a text and the relationship between those words, can be used to explore the circumstances in which certain negatively charged words may be used positively and for what reasons in the domain of videogames. The authors conclude with ideas for future research, including how NLP can be used to complement other avenues of inquiry.

[1]DePaul University, USA

Corresponding Author:
José P. Zagal, College of Computing and Digital Media, 243 South Wabash Ave., Chicago, IL 60604-2300, USA
Email: jzagal@cdm.depaul.edu

Keywords

aesthetics, charged words, game reviews, linguistics, natural language processing, NLP, NLP techniques, parts-of-speech, readability, research method, review, sentiment analysis, syllable count, videogames

Natural language processing (NLP) is a field of computer science and linguistics devoted to creating computer systems that use human (natural) language as input and/or output (Jurafsky & Martin, 2008; Manning & Schütze, 1999). Broadly speaking, NLP uses a collection of tools and methods for the automated analysis of text data. NLP has most commonly been used in games for making sense of player-provided textual input. For instance, using a parser to translate player-typed text into actions that are carried out in a virtual environment (Lebling, Blank, & Anderson, 1979) or, when combined with novel artificial intelligence (AI) techniques, it has been used to create innovative gameplay experiences (Mateas & Stern, 2004).

We propose that NLP can also be used for game studies research. In this article, we provide an overview of NLP, describe a few research possibilities that can be explored using its tools and techniques, and describe some of its strengths and limitations.

Our overview of NLP is split into three sections. The first section, which we call *general techniques*, describes a few of the simple techniques that can be used such as word counting or examining the length of sentences. In the next section, we describe *NLP techniques* that are used to help identify more complex patterns and relationships between words and the texts in which they appear. Some of the techniques we cover include part-of-speech (POS) tagging, syntactic parsing, and clustering. In the third section, we explore how NLP can also be used to identify *subtleties in the use of language*. More specifically, we discuss *sentiment analysis* or the classification of opinions and feelings based on the words used in a text and the relationship between those words. In addition, in each section, we illustrate how these techniques can be used in game studies research by analyzing a significant corpus of game reviews.

Why Game Reviews?

Game reviews are undoubtedly an influence on the ways that people view, understand, and talk about games. They are one of the primary forms of videogame journalism and often overshadow other forms of journalistic discourse surrounding games such as news, investigative reporting, and commentary (D. Thomas, 2007). Videogame journalism is, in many ways, a referent regarding the popular use of words and terms for describing games. In addition, thanks to the rapid adoption of the World Wide Web, a huge number of user-submitted game reviews are available online as a resource for an analysis. Popular websites such as Gamespot[1] and IGN[2] each have multiple hundreds of thousands of game reviews submitted by their users. This supports the notion that the game review is a model of discourse that is also adopted by videogame aficionados. Zagal and Bruckman (2009) found that students taking videogame-related classes will often, when asked to describe, analyze and talk about specific games, and write in a tone and style evocative of the game reviews they are familiar with. Game reviews, especially those written by fans and nonprofessionals, are thus a relevant source of information that can help us understand how regular players, albeit those willing to write a game review, describe games, gameplay, and so on. What words do they use? How do they choose to express their feelings and emotions as they share their opinions regarding certain games? Moreover, what things do they write about when reviewing games? Might these things highlight the essential characteristics of games or help us in defining types of games based on how people describe them?

General Techniques

For practical purposes, NLP can be applied to perform the automated analysis of large amounts of textual data. Although the following section describes some of the more sophisticated tools and techniques used, it is important to note how even some of the "simpler" analyses can reveal insights that may be of interest. By simple forms of analyses, we mean things such as counting the occurrence of certain words, analyzing the types of words that appear, or examining the length of sentences. These kinds of indicators are often used to gauge the diversity of the language used in a document or to assess the overall complexity of a text. For instance, in order to be easily read and understood, sentences should be reasonably short and not too complex (Kirkman, 2005). Longer sentences, as defined by the number of words they have, are generally more complex. In fact, the Oxford Dictionaries' guide to better writing suggests an average sentence length of 15 to 20 words.[3] In addition to length, some metrics, which measure the complexity of a document, consider the ratio of passive sentences over active sentences.[4] Identifying the voice of a sentence (active/passive) requires a deeper linguistic analysis of the structure of a sentence, which NLP can provide through further syntactic parsing.

These techniques may be used in several ways for game studies research. For example, we could examine the complexity of player-written text in multiplayer virtual environments. These techniques can help us answer questions about the content of the discourse (e.g., which words are commonly used and thus presumably more important) as well as the activities being carried out. Bruckman used these simple analysis tools to gauge a sense of users' participation in a text-based children's multiplayer virtual environment. By counting the number of commands typed by each player, she was able to get a general sense of gender differences in participation in this environment (Bruckman, 2006). These tools could also be used to analyze the complexity of text in games, identify key terms used in games, and so on. In the following section, we describe the results of our analysis of the readability of professional online game reviews using simple techniques such as those we have described.

Sample Research: Game Review Readability

When talking about and discussing game reviews, one of the issues commonly raised has to do with their overall quality in terms of their use of language. Reviews have been maligned for a variety of reasons such as being inconsistent in writing and style (D. Thomas, Orland, & Steinberg, 2007), as well as being "rife with grammatical errors, historical inaccuracies, plagiarism, run on sentences, [and] clichés" (Buffa, 2006). Game reviews are supposedly written for the stereotype of the average game player—White, male, and in his teens. However, we wonder if that is really the case. We know that games are played by a broader demographic than ever (Juul, 2010) even as gaming sites' audiences are narrow (Gamespot, 2009). One way of finding out is to examine the readability of game reviews. By readability, we mean the ease of use in reading, as defined by the interaction between the reader's reading skill, prior knowledge, and the text of the game review itself. Therefore, a review written using commonly used words with few syllables and syntactically simple sentences is, generally speaking, more readable than one with longer words and sentences as well as complex sentence structures. Determining the readability of a particular document can be quite important in certain domains. For instance, legal requirements may exist for ensuring a high level of readability to ensure broad access and understanding of written materials such as in the case of research consent forms (Paasche-Orlow, Taylor, & Brancati, 2003).

Therefore, what readability do professionally written game reviews have? What level of education do they assume readers have? In order to answer this question, we analyzed 1,500 reviews posted by professional game reviewers between 2007 and 2008 on the videogame review site Gamespot. We then applied a series of commonly used readability formulas. Although many formulas exist, each with their own strengths and weaknesses, it is hard to say which is "the best" (Klare, 2000). We chose to apply a few of the more commonly used ones to get a broad sense of readability. We used the following formulas:

- Simple Measure of Gobbledygook (SMOG)
  - Estimates the years of education needed to completely understand a piece of English writing by examining the number of polysyllabic words in a sample of sentences from the text (McLaughlin, 1969).
- Coleman-Liau Index
  - Used to determine the approximate grade level (in the U.S. school system) necessary for understanding a text by examining the number of characters used per word (Coleman & Liau, 1975).
- Gunning Fog Index
  - Estimates the number of years of formal education someone should have to understand a text on a first reading based on sentence length and the percentage of words with three or more syllables (Armstrong, 1980).

Our results are perhaps surprising. Given the criticisms generally levied against game reviews, we expected their reading level to be generally low (i.e., written using

Table 1. Readability Scores for Professional Videogame Reviews

Index	Average score	Minimum	Maximum	*SD*
SMOG	10.98	8.2	15.1	1.04
Coleman-Liau	9.7	6.9	14	1.01
Gunning Fog	13.10	8.8	18.8	1.56

Note: SMOG = simple measure of gobbledygook.

simple words and sentences). In other words, we expected low scores, indicating a higher degree of readability. We found the opposite. Despite the variations between indices, online videogame reviews are not as readable as expected (see Table 1). A Gunning Fog Index of 13 is roughly equivalent to the *Wall Street Journal* and assumes a reading level of a first-year university student (Armstrong, 1980). More generally, each of the indices showed that game reviews are not written for a grade school reading level, rather they are written for secondary education reading level. The relatively low standard deviation for each of the indices is also indicative of a certain uniformity of the reviews. We might imagine that children's games are reviewed using a language that is accessible and adequate to their reading level, while mature games may be written for a higher reading level. This seems to be not the case as the readability is stable across all the reviews analyzed.

We note that the readability level measures only the ease of reading. The quality of the ideas and content in game reviews should be subject of a separate analysis. Klare (2000) notes how high readability (low scores) is desirable for better communication. It is also possible to achieve greater readability while still expressing the same content and ideas (Armstrong, 1980). In fact, anecdotally, many of today's bestselling authors write at a U.S. seventh-grade[5] reading level (Wikipedia, 2010). This is substantially lower than the readability scores we found. Whether or not this is a good thing is left for further discussion and analysis. Perhaps, one of the barriers to mainstream adoption of videogames is the inaccessibility of the writing about them, in this case game reviews?

More Complex Patterns and Relationships

Although simple metrics such as word frequency and sentence length can be considered a part of NLP at large, the heart of NLP lies in the deeper analysis of sentences based on linguistics. In general, NLP's analysis consists of the following five levels (in a hierarchy):

1. **Phonetics**, which analyzes the phonetic composition of a word (e.g., phonemes);
2. **Morphology**, which analyzes the morphological composition of a word (e.g., prefixes such as "un-" and suffixes such as "-ist");
3. **Syntax**, which identifies the phrase structure of a sentence (e.g., noun phrases, verb phrases) according to a grammar. The POS of each word

(e.g., noun and verb) is first identified, then phrases and their structural relations are discovered, and the final sentence structure, commonly represented by a (syntax) tree, is derived by the process of (syntactic) parsing;
4. **Semantics**, which derives the meaning of a sentence based on the meanings of the words/phrases (identified by parsing); and
5. **Discourse**, which derives the meaning of a paragraph or a larger unit of text consisting of several sentences.

In addition to the linguistically oriented approach, other types of approaches that use probability, statistics, and machine learning techniques have gained popularity in recent NLP research and applications. Those approaches are data driven, in that they try to extract human linguistic knowledge empirically from a large volume of corpus, in contrast to the linguistic-based approach, which uses explicit linguistic knowledge. For this reason, empirical approaches can only do a shallow analysis of texts, but are more robust especially in the presence of ungrammatical sentences or typographical errors. Because of this robustness, many of the recent NLP applications use empirical (statistical/probabilistic and machine learning) techniques in processing texts.

A number of stand-alone NLP applications are used in practice today, including machine translation (e.g., Google translate[6]), question-answering (answering a question typed in as a natural language sentence, not keywords, e.g., Ask.com[7]), and conversational agents (also called chat bots; e.g., Alice[8]). Note that some NLP applications, especially conversational/dialogue systems, involve generation as well as analysis of the language. Most of such systems (including the NLP component in façade; Mateas & Stern, 2004), however, only use manually defined, ad hoc surface-based patterns to match input sentences and generate canned responses.

In addition to traditional applications, recently, we have seen much effort in NLP to develop applications that analyze texts available on various user-generated media such as weblogs (including Twitter), discussion forums, message boards, and social networks. At the same time, component tasks and techniques used in those applications have begun to receive more attention and emphasis in the NLP research. One such task is POS tagging—the process of identifying and assigning a POS (e.g., noun, verb, and adjective) to each word in a sentence. For example, for an input sentence "The icy roads are dangerous," the POS tagging outputs "The/Det icy/Adj roads/N are/V dangerous/Adj."[9] Although the identification of the POSs is essentially a part of a syntactic analysis, POS tagging has become an NLP task of its own. It is computationally less complex than parsing (thus more feasible in processing a large amount of text), and many situations and applications exist in which the knowledge of the POS of the words in a sentence alone is sufficient (rather than the accurate syntactic structure of a sentence). Using the current state-of-the-art techniques, the accuracy of the POS tagging is about 97% for the English language.[10]

One application, in which the POS information of words is effectively used, is the automatic discovery of word associations. A word may be associated with other words by various relationships. For example, words of the same POS may be related as in the case of synonyms, hypernyms (a general-specific relationship; e.g., "vehicle"-"car"), and meronyms (a part-whole relationship; e.g., "tire"-"car"). Words may also have specific syntactic/grammatical relationships such as the adjective-noun modifier (e.g., "icy roads"). Although different relationships require varied information extracted from the texts to do the discovery, typically the POS tagging is first applied and then the candidate word pairs are extracted. Moreover, the words selected in the candidate word pairs are oftentimes clustered to obtain groups of words (instead of individual words), so that the relationship can be generalized, and the associations between the words that were not in the original data set can be discovered (i.e., bootstrapping from the initial discovery). Below, we describe the results of a study in which we clustered adjectives that modify the noun "gameplay," which were extracted from the user game reviews.

Sample Research: Gameplay Aesthetics From Game Reviews

Earlier we showed how we could analyze professionally written game reviews to gauge their readability. The sites where these reviews are posted often allow their users to post their own reviews as well. What could we learn from these reviews? One question, which we will examine in this section, is how is gameplay understood by players?

The word gameplay is frequently used when discussing and describing games. Björk and Holopainen (2005) define gameplay as "the structures of player interaction with the game system and with the other players in the game"(p. 3). They use the term to synthesize the notion of formal and structural elements of games that help shape a players' experience. Their approach is formalist and from the perspective of the designer. How do players understand, perceive, and talk about gameplay?

To answer this, we might examine the emotions, feelings, sensations, and physiological responses as reported by, or observed in, the player in a laboratory setting. This way we may achieve a deeper understanding of the connections between game design elements and emotional patterns (Ermi & Mäyrä, 2005; Nacke & Lindley, 2009). We could also examine player-generated game data, for instance, examining logs of gameplay activity (Jin Shim & Srivastava, 2010; Mahlmann, Drachen, Togelius, Canossa, & Yannakakis, 2010). These approaches do not scale to large numbers of games and players. A bottom-up approach, which examined a vast corpus of game reviews, written by game fans, aficionados, and nonprofessional writers, could also help answer this question. Game reviews are a popular form of discourse that provides a window into the thoughts and feelings on gameplay as understood in the broad sense of popular culture. It provides us with an understanding of gameplay as it is commonly understood, used, and negotiated by players (Zagal & Tomuro, 2010).

We downloaded and analyzed all the user-submitted reviews posted on Gamespot as of April 20, 2009. We found 397,759 user reviews in total, and they covered a total of 8,279 game titles. Games with the same title, but on different platform, were counted separately as we know that a game's narrative, controls, and resulting gameplay experience can vary significantly across platforms even when the game title is the same. In total, we examined all 397,759 user reviews, which were written by 111,943 unique users.

Our analysis consisted of several steps. We began by extracting all sentences in which the word "gameplay" (and variations such as "gameplays") appeared, identified the POS of the words, and parsed the sentences using a POS tagger and a parser (we used the Stanford POS Tagger[11] and the Stanford Statistical Parser[12]). Then we extracted all the adjectives that were used as a prenominal modifier to "gameplay" (e.g., "smooth gameplay") or as an adjectival complement of "gameplay" (e.g., "gameplay was smooth"). After eliminating words that only appeared once and cleaning up typographical errors, our final list had 723 adjectives.

The next step in our analysis consisted of extracting the context of each selected adjective as it appeared in all the reviews. For example, if the phrase "smooth gameplay" appeared in one sentence in a review and "smooth control" appeared in a sentence in another review, the list of context words for "smooth" would contain "gameplay" and "control." Basically, context words is the set of all the words that appeared in the context (or in close proximity) of a given adjective. Context is defined as the words in an n-word window surrounding the adjective. In this case, we chose one word preceding and one word following the adjective (thus $n = 3$, including the adjective). The list of context words extracted in this way numbered over 175,000.

We then chose the 5,000 context words that appeared most frequently and represented every original adjective by the context words.[13] Using those context words, we created a 723 × 5,000 matrix, where the rows were adjectives and the columns were context words. The value of each cell in the matrix corresponded to the number of times that a given adjective appeared together with the context word.

We then proceeded to create adjective clusters by using a clustering algorithm called Kmeans (MacQueen, 1967). By clustering adjectives, our goal was to see whether we could derive various categories of gameplay. Kmeans is a machine learning algorithm that partitions the data into k number of clusters (where k is specified a priori). The algorithm assigns each data instance to one of the clusters whose mean (called centroid; the center/average of the members assigned to a given cluster) is closest to the instance. Therefore, as the adjectives are represented by the context words in our matrix, the adjectives whose contexts are similar/close are grouped into the same cluster.[14] After some preliminary experiment, we chose $k = 30$ in this study because we observed that some meaningful clusters (i.e., clusters that correspond to our intuitions) were generated.[15]

As described, we clustered adjectives based on their contexts. This approach is based on a concept in NLP called Distributional Similarity (Lee & Pereira, 1999)—two words are similar if their distributions, in particular, the words that co-occurred with them in a context, are similar. For example, "coffee" and "juice" are considered similar because they are often used in similar phrases such as "drink coffee in the morning" (and "drink juice in the morning") or "spilled coffee on the table" (and "spilled juice on the table"). Other drinks such as "tea" and "cocoa" are used in similar

Table 2. Elements of Gameplay Aesthetics as Revealed From User Game Reviews

Element	Definition	Sample adjectives
Pacing	The perception of how often game events occur	*Fast, stressful, dull, tedious, frantic, chaotic, obnoxious, frenzied, energetic, silky, brisk*
Complexity	The measure, or sense, of the number of parts in a system and how they are interrelated	*Simple, short, complex, streamlined ingenious, flexible, uncomplicated, organized, reliable, straightforward*
Cognitive accessibility	The measure, or sense, of the opacity of a system and the challenges it poses in understanding it	*Deep, designed, unusual, twisted, uninteresting, memorable, customizable, imaginative, intricate, crafted, wacky, colorful*
Scope	The size of the possibility space afforded by a game	*Limited, unlimited, large, endless, massive, vast, tremendous, immense, minimal, maximum, moderate, infinite, extensive*
Demands	The requirements imposed upon the player by the gameplay	*Casual, sandbox, hardcore, experienced, retro, demanding, intellectual, loving*
Impact	What we feel games "do to us" when we play them, and how they make us feel	*Addictive, exciting, refreshing, exhilarating, boring, annoying, stale, monotonous, irritating, tiring, overwhelming, numbing*

contexts as well. Thus, by clustering words based on their contexts, we can derive various categories of words that have similar meanings. In this study, we first extracted adjectives that modified "gameplay" and then clustered those adjectives based on the words (nouns, verbs, and adjectives) that appeared in the contexts. Therefore, not only do the resulting clusters represent various types of gameplay, but they also indicate the (similar) aspects of a game, such as "control," "look," and "feel," which brought out the particular type or characteristics of gameplay.

Therefore, what language do we use to describe gameplay and what does this say about games as an expressive medium? Our results show that we have identified what could best be described as an aesthetic of gameplay (for additional details, refer to Zagal & Tomuro, 2010). By this, we do not mean the assessment of the "looks" of a game (i.e., the graphics and representation). This would be just one aesthetic aspect of videogames, and not necessarily one closely related to gameplay. We refer to gameplay itself, something that is intangible yet still appreciable. If music's aesthetic elements include harmony, rhythm, and mood, and film includes elements such as montage and lighting, what are the key elements for games? These elements, taken together, constitute what we call an aesthetic of gameplay. These key elements, each represented by clusters of adjectives, are pacing, complexity, cognitive accessibility, scope, demand, and impact. Table 2 lists each of these aesthetic elements together with a definition and a sampling of adjectives belonging to each of the clusters.[16]

We note that this aesthetic was generated in a bottom-up fashion from a sizable amount of discourse covering a wide variety of modern videogames. In this sense, it is a popular aesthetic that reflects how players describe gameplay. This approach complements analyses that are based on top-down reasoning and highlights how some aspects of gameplay may not be as salient or central to the player's experience as previously thought. For example, "emergent gameplay" has been described as a way to empower the player and ultimately allow for a more satisfying and interesting gameplay experience (Sweetser & Wiles, 2005). Curiously, we found no evidence to support "emergence" as a core aesthetic dimension of gameplay. Whether or not a game's gameplay is emergent may simply be less important when compared with other aesthetic elements. This discussion would be similar to focusing on the weight of a sculpture rather than its form or materials. Alternately, "emergent gameplay" may be rare and challenging enough to design, so that it simply is not talked about as much by players and thus did not register in our analysis. It may also be the case that players have difficulties describing gameplay using terms that are related to emergence. Regardless, the point is that the aesthetic elements of gameplay that are meaningful to game designers and scholars are probably different from those of players. We would have to perform a similar analysis of writings by game designers and scholars to confirm this.

The above analysis serves as an example of some of the questions that could be explored and should in no way be understood as "definitive." For instance, another thing to investigate would be to try other clustering algorithms to see whether any other notable aesthetic elements can be discovered. Clustering is an exploratory task, and different methods or parameter settings tend to produce varying partitions of the data, although usually not vastly different. Therefore, "emergent gameplay" may arise in other partitions of the concept space of gameplay. Moreover, given the sheer number of players, we could imagine a much wider range of aesthetic ideals within specific player communities, demographic groups, and so on.

Subtleties in Language Use

As we noted earlier, many NLP applications have been developed to analyze various user-generated social media, such as product reviews, weblogs, and message boards. The range of aims of those applications is quite wide, from obtaining product marketing information (Popescu & Etzioni, 2005), tracking political opinions (M. Thomas, Pang, & Lee, 2006), to searching for "buzz" (e.g., what is hot, what topics are people talking about right now) in social networks (e.g., Tweetfeel[17]). User-written texts in these applications have a distinct characteristic compared with other kinds of texts such as newspaper articles; the texts are subjective, expressing the opinions or sentiment of the users toward the thing(s) of interest.

Sentiment analysis is a task in NLP that automatically identifies the sentiment of a text, typically either positive or negative, based on the opinions or feelings written in the text (Pang & Lee, 2008). Intuitively, the sentiment of a text can be identified by the presence of words that have the undertone of the sentiment. For example, if the text contained many occurrences of words such as "great" and "wonderful," one would guess that the text is expressing positive sentiment. However, words indicative of sentiment are sometimes more subtle in their meaning depending on the context and domain of the text. Other factors also make the sentiment identification difficult, including linguistic nuances (e.g., irony, sarcasm) and cultural differences. For instance, the word “wicked” has a negative connotation in the general texts, but is used to mean “fantastic” or “excellent” by youth in Great Britain.

Therefore, the first step in the sentiment analysis is usually to discover the set of words that are strongly indicative of sentiment (for any polarity) in a given domain or context. For example, Drake, Ringger, and Ventura (2008) analyzed game reviews posted on Gamespot and used the games’ rating scores to discover words that are positively or negatively correlated with certain sentiments. Another approach, which is general (i.e., not particularly dependent on the domain of the texts), but computationally more feasible, is to focus on words of specific POSs. Typically, words that are indicative of sentiment are adjectives (e.g., “awesome”), adverbs (e.g., “beautifully”; most of them are morphological derivations of their base adjectives), and some nouns (e.g., “gem”) and verbs (e.g., “love”; Wiebe, Wilson, Bruce, Bell, & Martin, 2004). Thus, POS tagging can be effectively used to select candidate words, thereby reducing the computational cost of the analysis.

Sentiment analysis of videogame reviews can provide game designers as well as game researchers with insights into what the users/players perceive as favorable or unfavorable about a particular game. By identifying the sentiment-salient keywords of the domain and comparing them against the general use of the language, we can explore research questions such as under what circumstances are negatively charged words used to describe a game positively (and vice versa)? (e.g., obsessive, addictive). What are the commonly used sentiment-salient words and how do they compare with those used to describe other media? Similarly, we can identify and analyze the rhetoric of videogame marketing identifying such things as trends in language use. These methods can also be used in other arenas. For instance, could we, by analyzing real-time chat text in an online game, gauge the sentiment of players currently playing and make game design adjustments on-the-fly?

Sample Research: Sentiment Analysis

In the previous section, we presented our work on extracting gameplay aesthetics from game reviews using 723 adjectives that were used to qualify or modify gameplay. In this section, we look at user-submitted review scores, together with the adjectives from our earlier gameplay aesthetics study, to identify the sentiment of adjectives used in the domain of videogames. More specifically, we wanted to investigate under what circumstances are negatively charged words used to describe a game positively?

For each of the 723 adjectives, we determined its polarity in the following way. First, for each user (from a total of 111,943 unique users, who posted a total of 397,759 reviews), we computed the average rating score he or she gave to all games. Then for each review written by a user, we extracted the adjectives used in the review that were also in the list of 723 adjectives. Each adjective was then assigned a polarity value for that user. The polarity value for that user was computed as the sum of the polarity value of that adjective in each of the games reviewed by that user. The user's polarity value for an adjective for a game was calculated as the difference between the user's average rating score and the rating score he or she gave to that game. This was done to balance out the scoring criteria used by different reviewers. Lenient reviewers tend to give a relatively high score to any game, while harsh reviewers give a low score. By looking at the difference from a user's average score, we get a better sense of how positive (or negative) they are about a particular game. Finally, after we had calculated the polarity of a particular adjective for all users, we added them up to obtain the overall polarity value of that adjective.

For instance, JaneDoe's average rating score for games is 8.0. Jane also used the adjective "adventurous" in her review of GameA. Her polarity value for "adventurous" for GameA, which she rated with an 8.5, is thus 0.5 (i.e., 8.5 − 8.0). She also used "adventurous" when reviewing GameB. JaneDoe rated GameB with a 9.0, so her polarity value for "adventurous" for this game is 1.0 (i.e., 9.0 − 8.0). Finally, she used "adventurous" in her review of GameC, which she rated with a 7.0. Her polarity value in this case is −1.0 (i.e., 7.0 − 8.0). JaneDoe's polarity value for "adventurous" is calculated by adding the polarity for each of the games where she used that word. In this example, the total is 0.5 (i.e., 0.5 + 1.0 − 1.0). In order to calculate the overall polarity for "adventurous," we need to calculate its polarity for all users and add them all. Thus, a high overall (positive) polarity value for a given adjective indicates that the adjective was used by many reviewers, and it was overwhelmingly used in a positive way (vice versa for negative polarities). A value near zero would indicate that the adjective was not used much and/or its use is contested in terms of sentiment (many people use it positively, but many people also use it negatively).

As expected, the words with the highest positive and negative polarities are words commonly associated with those sentiments. In our analysis, the words with the highest polarity were "great," "new," and "amazing."[18] The words with the lowest polarity were "bad," "horrible," and "terrible."[19] The high positive sentiment associated with the word "new" is perhaps particular to modern culture.[20] It makes sense from a techno-centric perspective (newer is better), but also supports the notion that the perceived quality of a game depends on whether it provides novel experiences. Earlier research has found that professional game reviews commonly discuss and analyze games in the context of other games, highlighting the differences with earlier versions and other similar games (Zagal, Ladd, & Johnson, 2009). Thus, the practice of focusing on the new, in positive terms, seems to carry over from professional reviews to those written by regular players as well.

Other adjectives were perhaps more surprising. For example, the word "addictive"[21] is generally used negatively, referring to the persistent and compulsive use of a substance known to be harmful. In the case of games, it is used positively. The games for which this word was most commonly used in a positive sense were CALL OF DUTY 4: MODERN WARFARE (COD:MW, 2007) and GEARS OF WAR (2006). However, the term was also used in a negative light in the case of games such as TETRIS WORLDS (2001) and WII PLAY (2006). Why is that “addictiveness” is a good thing in COD:MW, but not so for WII PLAY? We imagine that certain players appreciate addictive qualities in games, while others may resent them. However, these differences should also be understood in terms of game design, player’s expectations and experience, or both. If I bought a game I intend to play for significant periods of time, I may value its addictive qualities! In another example, the word “insane”[22] is often used negatively to refer to mental disorders or capabilities of people. Its secondary use, the absurd or extreme, is used as a positive sentiment when referring to GEARS OF WAR and RESIDENT EVIL 4 (2005). However, in the case of MADDEN NFL 08 (2007) and SPIDER-MAN 3 (2007), the term is used negatively. Again, why the difference? Perhaps, players resent unrealistic features in MADDEN NFL 08 and SPIDER-MAN 3 that go against their expectations of a popular sport and the comic book world of Spiderman? Further analysis is required to fully understand these differences.

Conclusions and Future Work

We have described some of the methods and techniques in NLP and shown, via the analysis of online game reviews, how NLP could be used in game studies research. Our examples have shown how NLP can be used to explore a variety of research questions. NLP can also be used to provide baseline data to guide future inquiry or extend findings obtained using other methods. For instance, in an interview-based study, DeVane and Squire (2008) explored the meanings that players make of their experience playing GRAND THEFT AUTO: SAN ANDREAS (GTA:SA; 2004). They argue that “even though the game is a designed space, meaning is plural, multiple and situated because it is a possibility space” (DeVane & Squire, 2008) (p. 281). Their results provide the richness in detail and nuance that is characteristic of their chosen methodology. We could complement their findings by analyzing the texts of online fan sites and message boards. What other meanings for GTA:SA may we find that could then be explored more deeply? In the case of exploring gameplay, NLP could support techniques from cognitive science that explore similar questions (Lindley, Nacke, & Sennersten, 2008).

NLP is not without its limitations. For instance, NLP’s automated analysis of text data is based on linguistics and computer science, and makes no claim on the validity of the analyzed results beyond those fields. Therefore, for tasks whose analysis concerns semantics (rather than syntax), validation of the results may be necessary. Similarly, the role of the domain expert is crucial in the process of sifting data and guiding the analysis in order to achieve meaningful results. Understanding the nature of the corpus being analyzed is also crucial. For example, our analysis of all the reviews posted on a single website is most likely not representative of the broader population of game players. Regardless, we feel confident that these techniques can be applied productively in game studies research so long as special care is taken.

In conclusion, we have outlined only some of the questions that could be explored using NLP and are currently exploring some of these ourselves. Our preliminary findings are encouraging for the kinds of insights these techniques can help us obtain, and we look forward to reporting on our results as well as encouraging other researchers to make use of these techniques in their own work.

Acknowledgments

We would like to thank our reviewers Michael Nitsche at Georgia Institute of Technology (USA), Timo Lainema at Turku School of Economics (Finland), Jesper Juul at New York University (USA), and Chris Bateman (International Hobo) for their kind comments and helpful feedback on earlier versions of this manuscript.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. http://www.gamespot.com/
2. http://www.ign.com/
3. http://www.askoxford.com/betterwriting/plainenglish/sentencelength/
4. Passive Sentence Readability Score; http://rfptemplates.technologyevaluation.com/readability-scores/Passive-Sentences-Readability-Score.html
5. Equivalent to SMOG, Coleman-Liau, and Gunning Fog scores of 7.
6. http://www.google.com/language_tools?hl=en
7. http://www.ask.com/
8. http://alicebot.blogspot.com/
9. In this notation, "Det" stands for determiner, "Adj" for adjective, "N" for noun, and "V" for verb, respectively.
10. http://www.cs.umass.edu/~mccallum/courses/inlp2007/lect10-tagginghmm1.ppt.pdf
11. http://nlp.stanford.edu/software/tagger.shtml
12. http://nlp.stanford.edu/software/lex-parser.shtml
13. We used context words that occurred most frequently rather than selecting randomly in order to capture the common usage of the words/adjectives. Low-frequency context words are most likely due to idiosyncrasy or typographical errors.
14. We used Euclidean distance as the measure of similarity.
15. Before deciding on the final k, we conducted preliminary experiments using $k = 10$, 20, and 30, and inspected the derived clusters. We did this preliminary step to get a holistic sense of the data. After inspecting the derived clusters from each experiment, we determined that

the 30-way clustering (i.e., $k = 30$) generated more meaningful clusters than those derived by the 10- or 20-way clustering.

16. Each of these six elements corresponds to at least one cluster derived by the Kmeans analysis (where $k = 30$). Some elements corresponded to two or more clusters because those clusters ended up having relatively close centroids during the clustering process. Also some clusters were not meaningful—they were made of seemingly dissimilar words grouped together because of statistical coincidence.
17. http://www.tweetfeel.com/
18. Polarity values of those words were "great" (96,701.2), "new" (52,340.8), "amazing" (42,549.0). Also the mean of the 723 adjectives was 631.6 and the *SD* was 6,039.2.
19. Polarity values of those words were "bad" (−56,160.8), "horrible" (−20,805.2), and "terrible" (−18,963.2).
20. The notion that new is good is a central idea of what we call modern, so it is not surprising to note this in the context of videogames as well.
21. Polarity value was 6,983.9.
22. Polarity value was 1,684.4.

References

Armstrong, J. S. (1980). Unintelligible management research and academic prestige. *Interfaces, 10*, 80-86.

Björk, S., & Holopainen, J. (2005). *Patterns in game design*. Hingham, MA: Charles River Media.

Bruckman, A. (2006). Analysis of log file data to understand behavior and learning in an online community. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds.), *International handbook of virtual learning environments* (pp. 1449-1465). New York, NY: Springer.

Buffa, C. (2006). *Why videogame journalism sucks*. Retrieved from http://www.gamedaily.com/articles/features/opinion-why-videogame-journalism-sucks/69180/?biz=1

CALL OF DUTY 4: MODERN WARFARE. (2007). [Developed by Infinity Ward]. Activision.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*, 283-284.

DeVane, B., & Squire, K. D. (2008). The meaning of race and violence in grand theft auto. *Games and Culture, 3*, 264-285.

Drake, A., Ringger, E., & Ventura, D. (2008). Sentiment regression: Using real-valued scores to summarize overall document sentiment. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (pp. 152-157). IEEE Computer Society, Washington, DC.

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. In S. de Castell & J. Jenson (Eds.), *Changing views: Worlds in play, selected papers of DIGRA 2005* (pp. 15-27). DiGRA, Vancouver, British Columbia, Canada.

Gamespot. (2009). *GameSpot audience profile study*. Retrieved from http://www.cbsinteractive.com/brand_pdfs/GameSpot_brand-one-sheet_100209.pdf

GEARS OF WAR. (2006). [Developed by Epic Games]. Microsoft Game Studios.

GRAND THEFT AUTO: SAN ANDREAS. (2004). [Developed by Rockstar North]. Rockstar Games.

Jin Shim, K., & Srivastava, J. (2010). Behavioral profiles of character types in EverQuest II. In G. Yannakakis & J. Togelius (Eds.), *Proceedings of the 2010 IEEE Conference on Computational Intelligence in Games* (pp. 186-194). IEEE, Washington, DC.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall, Upper Saddle River, NJ.

Juul, J. (2010). *A casual revolution: Reinventing videogames and their players*. Cambridge, MA: MIT Press.

Kirkman, J. (2005). Sentence length and complexity. In *Good style: Writing for science and technology* (pp. 4-11). Routledge, London.

Klare, G. R. (2000). The measurement of readability: Useful information for communicators. *ACM Journal of Computer Documentation, 24*, 107-121.

Lebling, P. D., Blank, M. S., & Anderson, T. A. (1979). Zork: A computerized fantasy simulation game. *IEEE Computer, 12*, 51-59.

Lee, L., & Pereira, F. (1999). Distributional similarity models: Clustering vs. nearest neighbors. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 33-40). College Park, Maryland.

Lindley, C., Nacke, L., & Sennersten, C. C. (2008). *Dissecting play—Investigating the cognitive and emotional motivations and affects of computer gameplay*. *CGAMES08*. Wolverhampton, UK.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297). Berkeley: University of California Press.

MADDEN NFL 08. (2007). [Developed by EA Tiburon]. EA Sports.

Mahlmann, T., Drachen, A., Togelius, J., Canossa, A., & Yannakakis, G. (2010). Predicting player behavior in Tomb Raider: Underworld. In G. Yannakakis & J. Togelius (Eds.), *Proceedings of the 2010 IEEE Conference on Computational Intelligence in Games* (pp. 178-185). IEEE, Washington, DC.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mateas, M. & Stern, A. (2004) . Natural Language Processing in Façade: Surface Text Processing, in S. Göbel, U. Spierling, A. Hoffmann, I. Iurgel, O. Schneider, J. Dechau, and A. Feix (Eds.), *Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)* (p. 3-13). Springer: Darmstadt, Germany.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*, 639-646.

Nacke, L., & Lindley, C. (2009). Affective ludology, flow and immersion in a first- person shooter: Measurement of player experience. *Loading, 3*. Retrieved from http://journals.sfu.ca/loading/index.php/loading/article/view/72

Paasche-Orlow, M. K., Taylor, H. A., & Brancati, F. L. (2003). Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine, 348*, 721-726.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*, 1-135.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In R. Mooney (Ed.), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp. 339-346). Association for Computational Linguistics: Stroudsburg, PA.

RESIDENT EVIL 4. (2005). [Developed by Capcom Production Studio 4]. Capcom.

SPIDER-MAN 3. (2007). Activision.

Sweetser, P., & Wiles, J. (2005). Scripting versus emergence: Issues for game developers and players in game environment design. *International Journal of Intelligent Games and Simulations, 4*, 1-9.

TETRIS WORLDS. (2001). THQ.

Thomas, D. (2007). Game criticism redefined: Is this game any good. In D. Thomas, K. Orland, & S. Steinberg (Eds.), *The videogame style guide and reference manual* (pp. 97-99) Power Play Publishing.

Thomas, D., Orland, K., & Steinberg, S. (2007). *The videogame style guide*. Power Play Publishing.

Thomas, M., Pang, B., & Lee, L. (2006). Get out of the vote: Determining support or opposition from congressional floor-debate transcripts. In M. Lapata and H. Tou Ng (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 327-335). Association for Computational Linguistics: Stroudsburg, PA.

Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*, 277-308.

WII PLAY. (2006). [Developed by Nintendo EAD]. Nintendo.

Wikipedia. (2010). Readability. Retrieved from http://en.wikipedia.org/wiki/Readability.

Zagal, J. P., & Bruckman, A. (2009). Novices, gamers, and scholars: Exploring the challenges of teaching about games. *Games Studies: The International Journal for Computer Games Research, 8*. Retrieved from http://gamestudies.org/0802/articles zagal_bruckman

Zagal, J. P., Ladd, A., & Johnson, T. (2009). Characterizing and understanding game reviews. In *Proceedings of the International Conference on the Foundations of Digital Games* (pp. 215-222). ACM, New York, NY.

Zagal, J. P., & Tomuro, N. (2010). The aesthetics of gameplay: A lexical approach. In A. Lugmayr, H. Franssila, O. Sotamaa, C. Safran, & T. Aaltonen (Eds.), *Proceedings of the 14th International Academic Mindtrek Conference* (pp. 9-16). Tampere, Finland: ACM.

Bios

José P. Zagal has a PhD in computer science from Georgia Institute of Technology, an MSc in engineering sciences, and a BS in industrial engineering from Pontificia Universidad Catolica de Chile. His research interests include the use of online communities for collaborative learning and the development of frameworks for describing, analyzing, and understanding games.
Contact: jzagal@cdm.depaul.edu.

Noriko Tomuro has a PhD and an MSc in computer science from DePaul University and a BA in English language from Sophia University (Tokyo, Japan). Her research interests include natural language processing, lexical semantics, information retrieval, and machine learning.
Contact: tomuro@cs.depaul.edu.

Andriy Shepitsen is currently a PhD candidate in computer science at DePaul University and has a BS from the National Agricultural University (Kyiv, Ukraine). His research interests include natural language processing, information retrieval, and recommender systems.
Contact: tomuro@cs.depaul.edu**.**