1. Context

1.1. The emergence of social media and participatory web structures has exerted a tremendous impact on communications between members of play communities, as well as the methods available to scholars to study these social interactions. Gamers have invested most of the web 2.0 platforms which also find themselves integrated to the interface of home consoles, enabling users to produce and broadcast content with the push of a button. Reddit, Facebook, Twitter, and many other websites are now the major discursive spaces that both host these exchanges and, simultaneously, influence their content and form. As they extend the social space (Nitsche 15) of the game, they also contribute in maintaining an active promotional voice online through the user-led indexation of content. It is on these spaces, prone to frequent and impulsive uses, that social conventions and practices are being defined and negotiated.

1.2. Beyond isolated studies that mobilized a range of different text mining techniques to specific game-related corpora (Liang; Mohri et al.; Pelletier-Gagnon; Suomela), the field of game studies has yet to develop a methodological framework that fully integrate this type of data to its frame of analysis. Accounting for this new situation poses challenges to traditional approaches to communication studies scholar and calls for the adaptation of pre-existing methodologies and epistemological frameworks to the analysis of born-digital textual productions of players in online communication. As it has proven in other fields such as sociology and literature (Moretti Distant Reading; Moretti Graphs, Maps, Trees; Öztürk and Ayvaz), exploiting these types of data would offer better insights on the formation of videogames as cultural text coproduced by its software and their users, as well as the impact of marketing strategies on player communication practises by highlighting significant trends and outlier phenomena.
2. Objectives

2.1. This paper proposes a basic framework for the study of player textual data conceived to address the specificity of the object of videogames and the field of game studies through the examination of posts published on Twitter about the 2019 Nintendo Switch title *Pokémon Sword and Shield* (PokémonSaS). This project is part of a broader effort on the assessment and categorization of text mining tools and methodologies conducted by researchers at the Canada Research Chair on Gaming Communities and Big Data to provide scholars and students of game studies with the means to supplement their work with text mining-based analysis. The release of a game as anticipated as *PokémonSaS* was provided an ideal case study through which tools and methodologies could be assessed and defined, as well as identify challenges and opportunities specific to the nature of the data gathered. In addition, the data gathered was used in the analysis of the evolution of player communication over time and in different linguistic regions, leading to the identification of discursive trends in the *PokémonSaS* Twitter community in relation to cultural capital among players.

2.2. In sum, the objectives of this project are situated on both levels of methodology and content analysis. The exploration of each aspect was motivated by the following questions:

1. What tools and methodology would provide the best ratio balancing the quality of analytical insights and its level of accessibility throughout all stages of a text mining project (data gathering, cleaning, normalization, analysis, and visualization)? What are the characteristics of the textual data regarding the affordances of the platform on which they were created?

2. How can we mobilize text mining and data visualization techniques to monitor the evolution player communication on a specific aspect of a given game? What are the limits of these conclusions, and how can we contextualize these results in conjunction with qualitative data?

2.3. With these questions in mind, we set out to test the potential of text mining methodologies and our research protocol through the analysis of the textual content produced by Twitter users about *Pokémon SaS* between September 2019 and February 2020, a five-month period equally divided between period before and after the game’s release on November 15, 2019. The corpus targeted tweets in English, French and Japanese associated with the hashtags #PokemonSwordShield, #PokémonÉpéeBouclier and #ポケモン剣盾(#Pokemonswordshield) and their derivatives. The resulting collection is comprised of about 3.5 million tweets that include Twitter-specific data point and metadata. At the outset of the project in its current form, we formed corpora structured around the axis of language to provide a comparative analysis of the communication between users. Twitter itself was selected over other platforms such as Discord and Reddit for its global reach which reflects the game’s international fanbase, as well as being one of Nintendo’s foremost online marketing channel.
2.4. Our starting hypothesis included the following statements:

1) The convergence of specific aspects of the game and the technical affordances of the communication platform encourage the production and dissemination of specific types of content.

2) Marketing strategies exert significant pressure on communities’ textual production while contributing in the flattening of topic diversity.

3) Finally, bridging the language divide would better account for the variety of ways in which an individual text is received on a global level.

3. Twitter, Hashtags and Textual Data in Videogame Community Analysis

3.1. Twitter ranks among the top 15 most consulted websites (Kemp) with 500 million tweets generated every daily by 134 million active users (Lin) in 2019, which speaks of its dominant position as a participatory web platform. While users share information and opinions on various topics ranging from politics to food, gaming is one of the leading topics on the platform with approximately one billion tweets generated in 2018 about a variety of topics ranging from esports, conventions, or specific titles (Chadha).

3.2. The affordances of its interface has been described as shaping discourse by “intensifying pre-existing characteristics of an erosion of the private in which more quotidian aspects of our lives are publicly shared” in the face of online communication as a process of the construction of the self (Murthy 1065) performed through the production of codified tweets addressed to an imagined audience constituted by a mix of expected known and unknown individuals (Marwick and boyd 129). Tweeting is thus a communicative act framed in a specific social context both concrete and imagined.

3.3. The hashtag indexing technology, which allows users to access content attached to a given topic in a single click, has become a “near-universal Internet symbol” (Panko) directly related to the emergence of influencer marketing and mobilization around sociopolitical issues. Hashtags allow content to be broadcast beyond the boundaries of one’s own relational network and to a platform’s entire audience by relating it to a broader subchannel. They expose one’s publication far more directly into the public eye standing “as the production and accumulation of public attention” (Bernard 4) that influence “use of language and or the creation of collectives,” thus, in some cases, creating a sense of community of discourse (Bernard 5). Andreas Bernard highlights the marketing use of hashtag as “it focuses attention of potential customers on particular brands, products, services and business ideas in a manner that casually involves the
community’s own participation” (Bernard 57), but also emphasizes its role as an identity forming device that both empowers and levels online discourse.

![Figure 1: Twitter integration on the Nintendo Switch](image)

3.4. In the field of game studies, textual data has been used to examine recent debates in gaming culture (Blodgett and Salter; Suomela; Todd et al.) while industry researchers are exploring ways to integrate them “into the traditional game analytics pipeline” (Milambiling et al. 142) to extend the effectiveness of metric analysis. However, this type of textual data has yet to be mobilized alongside text mining methods to study the reception of a specific videogame by fan communities from a humanities perspective. Tweets can be considered part of what Mia Consalvo calls videogame paratexts: texts produced in parallel to a game that influence the way in which the latter is played from which players shape a “gaming capital” of a given interpretive community (Fish). Within the context of this project, we consider textual data gathered from Twitter through the #Pokémonswordandshield hashtag as empowering, but also levelling discourse in relation to its user community’s standards and conventions. While others were created over the course of the study’s period, we choose to circumscribe the project to this official hashtag to better situate our object of study. In sum, analyzing the evolution of the use of said hashtag over time in would provide information on how users discuss certain topics in a given context, how dominant discourses emerge and shape the community’s identity, and how players (re)produce these dominant discourses to gain gaming capital within the community’s conventions.
4. Pokémon and Player Communication

4.1. First released in 1996 by the developer Game Freaks, Pokémon became one of the most popular and influential global media franchises that include over 120 videogames, 8 different anime series, 19 movies and a multitude of licensed products (Bulbapedia). Therefore, the cultural object represented by this franchise has been the subject of much scholarly attention throughout the years. A number of these works focused on its nostalgia effect for older audiences (Keogh; Zsila et al.), in its relation to the environment (Bainbridge; Dorward et al.), in its localization challenge (Iwabuchi; Katsuno and Maret) as well as the phenomenon of ROM hacking (Barnabé). Anthropologist Nakazawa Shin’ichi discussed the game as a device empowering child to develop and negotiate an alternative worldview through peer communication.

4.2. In recent years, most research focused on the augmented reality game Pokémon Go, and specifically around the study of the development of communities around the game. Assunção, Brown and Workman underline how the game company Game Freaks contributes to the creation of a sense of community among players. By establishing special events where trainers can meet and interact, Game Freaks takes part in the maintenance of their fan base. This level of involvement was also observed in the case of Pokémon SaS: in the months leading up to the game’s launch, Nintendo produced several trailers and Nintendo Directs, which are longer videos showcasing one or many games. These videos or trailers often presented new characters, features, areas and storylines in the game. In total, 22 videos were published by Nintendo between the 27th of February 2019 and the game’s launch. Ten of those videos are in Japanese and twelve are in English. Some videos were uniquely produced for the Japanese or English audience. Eleven of those 22 videos were published in the two months leading up to the game’s launch.

4.3. Through its emphasis on interpersonal interactions built into its core gameplay mechanics, its active fanbase online and the marketing presence of its development studio, the Pokémon franchise constitutes an ideal case study to test the potential of text mining mythologies on a large corpus. Although the studies previously mention explored the notion of community, they are specifically centred on the game Pokémon Go as an augmented reality game. Work that closely examine the dynamics of player communities around the release of a conventional Pokémon games on console has yet to be done.
5. Study Methodology and Observations

5.1. The decision was made to elect Pokémon SaS Twitter community as a case study in August of 2019 and data gathering process using the official Twitter API started in early September. With no access to the premium API access, the team used the free version, thus enforcing a strict regimen of scraping tweets every week. In the rare cases where data could not be scraped due to the time constraints, an unofficial scraping python package was used to complete the database. Tweets were scraped based on their association to a list of hashtags that covers the English, French and Japanese version of the game’s title, and thus cannot account for all tweets on the topic for the entire duration. Tweets in Spanish, for example, are not included. Retweets were filtered out as their inclusion would have far exceeded our available computing capability. We opted to keep count of tweets' retweets to replicate their popularity in later parts of the projects. The resulting corpus is comprised of about 3.5 million tweets with key metadata points such as retweet, favourites and followers count in addition to usernames, detected languages, mentions, and timestamp. Due to the nature of the API, user information and conversation ID were not scraped, thus limiting the reach of the corpus.

5.2. Upon conclusion of the scraping process, the corpus was cleaned and normalized using a combination of python scripts and Open Refine. The corpus was also separated into three different sub corpora to facilitate subsequent steps. The main corpus, which includes most metadata, was integrated into a Tableau file through which general data trends could be observed and relevant events surrounding the game’s release were added to contextualize the following analysis. In addition, language-specific corpora were created to independently study each linguistic region.

[The general explorer file is accessible here.]

5.3. To generate initial insights, we used KHCoder to perform a range of different basic explorations, including word counts, crosstabs, and co-correlation networks. KHCoder is one of the few text mining software of the ones tested that natively supports Japanese in addition to Latin script languages and was thus preferred over other alternatives. This process led the team to identify the “shiny” keyword as a significant concept the corpus, comprising 7.86%, 7.64% and 6.63% percent of the English, French and Japanese corpora. A co-occurrence network analysis also revealed significantly different semantic fields related to these terms in all languages. In light of these findings, we

1 Hashtags used in the query are the following: #PokemonSwordShield #PokemonSword #PokemonShield #ポケモンモンスターソード #ポケットモンスターシールド #ポケモン剣盾 #ソードシールド #PokemonBouclier #PokemonEpee #PokemonEpeeBouclier.

2 These figures include concatenation of different words expressing the same concept. For example, the francophone users refer to the “shiny pokémon” concept both as “shiny” and “chromatique.” Japanese has a higher number of variations.
turned our focus to this concept, seeking to further detail it was talked about across the three linguistic regions under study, as well as before and after the game’s release.

Figure 2: Shiny in the French Corpus (jaccard 300)

Figure 3: Shiny in the Japanese Corpus (jaccard 300)
5.4. The high frequency distribution of the shiny concept is not insignificant in relation to the Pokémon player community, its regime of values, as well as the affordance of Twitter as a communication platform. Shiny Pokémon, present in most games of the series, are similar in all aspect to regular Pokémon with the exception that they sport an alternative colour pattern. Highly sought after by players in the later stage of the game, shiny Pokémon are rare with an appearance rate of 0.024% in regular random encounters. Shiny Pokémon provide cosmetic late-game engagement game element for dedicated players that encourages player communication and the demonstration of one’s dedication to the game. The acquisition of a shiny Pokémon is usually celebrated as an accomplishment. In the context of Twitter, these community dynamics are found further accentuated as its short and impulsive affordance, "retweet" and "favoured" functionalities, as well as the hashtag system converge to valorize certain textual productions over others. The shiny concept stands as the game mechanic most affected by these affordances as the platform provides the means to share one’s achievement with a broader user base and contribute in constructing one’s online presence and identity in relation to Pokémon SaS and its community.
5.5. The team subsequently used Mallet to generate 100 LDA topic models to classify individual tweets of each linguistic corpus separately thus generating a more granular perspective on the community’s discursive trends. Many of these “bags of words” included some iteration of the shiny Pokémon concept along with other words. Since comparing topics between linguistic corpora is not possible due to vast differences in data granularity, we elected to classify each topic in one of three broad categories that were found across linguistic corpora: affective response, maximization and trade, as well as marketing, streaming and influencers. These concatenated entities provided a satisfying baseline from which to perform a multilingual comparison and the evolution of discourse trends over time.

[The English topic modelling visualization file is accessible here.]
[The French topic modelling visualization file is accessible here.]
[The Japanese topic modelling visualization file is accessible here.]
6. Observations

6.1. Our first observation relates to the importance of the Japanese corpus in relation to the English and French ones. Our initial postulate establishing the English corpora as the most voluminous was eliminated early in the research process, which points to the need to develop cross-cultural perspective when studying online communities. In addition, it is not possible to assume that each of these languages evolves in a vacuum, as the copresence of hashtags of multiple languages in many tweets (in the French corpus primarily) indicate freer transnational flow between different game communities and, consequently, regimes of values defined by various influences. Nevertheless, each linguistic corpus also demonstrated distinct discursive trends that speak of their community’s inclinations in textual productions and degree of exposure to Nintendo’s marketing strategies. For example, the Japanese corpus shows more influence being exerted by marketing campaigns in the sharing of content than in other languages, as well as seeing regular textual production peaks regularly on every Sunday. In every case, however, game publisher-led events seem to drive retweet production, but has only limited effect on the production of new tweets by fans.

6.2. From the release of the game onward, mentions of the shiny game mechanic increase in all corpora, following the general trend of textual production increase. However, as the latter quickly fades down, the former follows the opposite trend in proportional corpus distribution, suggesting a higher concentration of discourses around the concept over time probably led by the late-game acquisition of shiny Pokémon, but also the anticipated positive reception that such posts would gain. In addition, and perhaps due to the unbalanced tweet distribution across the corpus, posts published before the release of the game demonstrates a higher degree of variation which then stabilizes after November 15th.

6.3. The previously mentioned classification of topic models also demonstrates the discursive makeup of each linguistic corpus in relation to the shiny Pokémon game mechanic. The English corpus indicates a clear preference for emotional responses while the topic of maximization is the closest second. Both French and Japanese corpora indicate that marketing and influencers, as different types of community leaders, are driving communication in their own community. In addition, only in the Japanese corpus can we observe a clear and constant evolution in discourse where the topic of maximization and trade slowly overshadows marketing-influenced Tweets as the main speech trend. Finally, the French corpus’s distinctive trait is the high degree of category variation and the constant presence of influencers and streamers in discursive patterns, which is symptomatic of a smaller volume of textual production, and the representation of community leaders in smaller communities.
7. Conclusions and Further Developments

7.1. The research protocol helped the team identify an area of interest, as well as build a model suitable to the analysis of a gaming community’s emergent regimes of value in different language-based interpretive communities. Our initial hypothesis on the effect of marketing campaigns, however, was discredited as the post-release section of corpora indicated the standardization of the semantic field as Nintendo’s interventions faded down. This phenomenon singles out game mechanics and the technical affordances of Twitter in content production and broadcasting as more significant factor influencing the reproduction of discourse online.

7.2. This project constitutes the first case study examined by the team and should influence the protocol and future methods of analysis in subsequent endeavours. The official Twitter API proved very constraining, and alternatives will be explored to gain more scraping flexibility and data points. Secondly, further work will focus on social metadata analysis to identify community leaders and their textual productions more precisely. Finally, the team will turn its attention to adapt sentiment analysis tools to study social media-based and gaming-themed corpora. We would welcome any recommendations or comments in relation to any of these points.

References


Bulbapedia. Pokémon Games. https://bulbapedia.bulbagarden.net/wiki/Pok%C3%A9mon_games


Panko, Ben. "A Decade Ago, the Hashtag Reshaped the Internet." Smithsonian.com 2017.

Pelletier-Gagnon, Jérémie. "'Very Much Like Any Other Japanese Rpg You've Ever Played': Using Undirected Topic Modelling to Examine the Evolution of Jrpgs’ Presence in Anglophone

