'Finally! My first shiny!': Vertical Text Mining Research Protocol and Multilingual Analysis of Player Community Discourse on *Pokémon Sword and Shield* on Twitter

Jérémie Pelletier-Gagnon
Université du Québec à Montréal, jrmie@ualberta.ca

Alexandra Dumont
Université du Québec à Montréal, alexandra.b.dumont@courrier.uqam.ca

Antoine Jobin
Université du Québec à Montréal, jobin.antoine@courrier.uqam.ca

Patrick Deslauriers
Université du Québec à Montréal, deslauriers.patrick@gmail.com

Maude Bonenfant
Université du Québec à Montréal, bonenfant.maude@uqam.ca

Abstract

In the context of the increasing use of participatory web platforms by both videogame player communities and game publishers, adapting research protocols in communication studies to integrate text mining tools and methodologies becomes key in understanding emergent discourses and practices on a large scale. Their usefulness has been proven by their integration in various domains in the humanities, but their use for the study of videogames remains scarce despite their nature as digital objects.

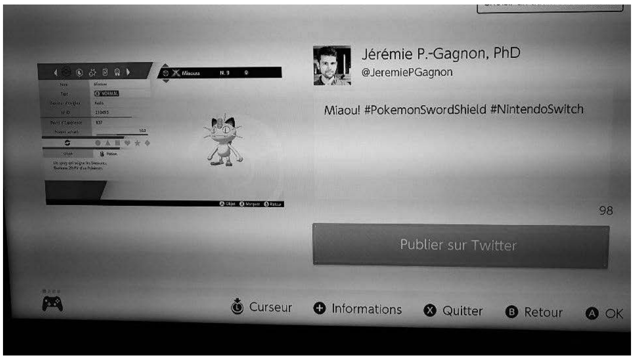
This paper provides an assessment of the effectiveness of several text mining techniques regarding their capacity to extract insights on standards and practices from a large corpus of textual data generated by fan communities online. A research protocol was formed and tested against a corpus comprised of the multilingual textual productions generated by fans during a period of five months using the hashtag #PokémonSwordandShield on Twitter. From the extraction of data to its statistical and textual analysis through frequency tables, word association network, and undirected topic modelling, the research team proposes an easily reproducible vertical research protocol developed for the purpose of research in the context of game studies and communication studies. It then applied it for the exploration of social media textual data through the analysis of fan discussion over the 'shiny' mechanic feature in *Pokémon Sword and Shield* (2019). Through comparative analysis, significant differences are observed between the Japanese, English, and French corpora in the distribution of three topics categories: 'affect', 'system', and 'promotion', but evidence of translinguistic textual input also suggests a certain degree of interconnection between linguistic regions. The paper concludes with an assessment and recommendations of tools and methods for further analysis, as well as a reflection on the influence of marketing campaigns on the production of new tweets upon the release of the game as community-led trends emerge and become dominant.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Context

The emergence of social media and participatory web structures has exerted a tremendous impact on communications between members of play communities, and the methods available to scholars to study these interpersonal interactions. Gamers have appropriated many of these platforms which are even integrated to the interface of home consoles, enabling users to produce and broadcast messages with the push of a button and comment on others' just as easily (Figure 1). Reddit, Facebook, Twitter, and other websites are now some of the major discursive spaces that both host these exchanges and, simultaneously, influence their form and substance. As they extend the social space (Nitsche 2008, 15) of the game, they also contribute to maintaining an active promotional voice online through the user-led indexation of content. It is in these spaces, prone to frequent and impulsive uses, that conventions and practices are being defined and negotiated. Accounting for this new paradigm poses challenges to traditional approaches to communication studies scholar and requires the adaptation of pre-existing methodologies and epistemological frameworks to the analysis of born-digital textual productions of players in online communication. With the objective of addressing this situation, this paper provides an assessment of the effectiveness of several text mining techniques used in a vertical research pipeline developed in the context of game studies and communication studies regarding their capacity to extract insights on standards and practices from a large corpus of textual data generated by fan communities online.

The use of text mining and natural language processing in literary studies (Moretti 2005; Moretti 2013; Jockers 2013; Jockers 2014) points at the benefits of the analysis of text and textual data on a massive scale, making possible the study of massive numbers of documents, the acquisition of new insights on well-study corpora, as well as the exploration of texts located on the margins of traditional cannons. Isolated work in game studies demonstrate the potential of these methodologies in providing new perspectives to core questions of the field (Ensslin 2012; Zagal et al. 2012; Kang, Yong and Hwang 2017; Faisal and Peltoniemi 2018; Pelletier-Gagnon 2018; Liang 2019; Mohri et al. 2019; Suomela 2019; Sapach 2020). However, the computer-assisted analysis of textual data has yet to be applied in order to study standards and practices born from player communications in online communities, in addition to doing so from a cross-linguistic comparative perspective, which would contribute to the development of a more global understanding of the reception of videogames by their publics.



Jérémie P.-Gagnon, PhD
@JeremiePGagnon

Miaou! #PokemonSwordShield #NintendoSwitch

Figure 1: Twitter integration on the Nintendo Switch

Objectives

With the objective of addressing this situation, this paper provides an assessment of the effectiveness of several text mining techniques regarding their capacity to extract insights on standards and practices from a large corpus of textual data generated by fan communities online. It proposes an easily reproducible vertical research pipeline developed for the purpose of research in context of game studies and communication studies tested through the examination of posts published on Twitter about the 2019 Nintendo Switch title *Pokémon Sword and Shield* (*Pokémon SaS*). This project is part of a broader effort on the assessment and categorization of natural language processing 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 a methodology to integrate communication-based game research with text mining-based analysis. The release of a game as anticipated as *Pokémon SaS* provided an ideal case study through which tools and methodologies could be defined and evaluated, 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 divergent discursive trends in the *Pokémon SaS* Twitter community representative of community standards and practices. 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 provide the best balance between 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 are created?
2) How can we mobilize text mining and data visualization techniques to monitor the evolution of 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?

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 around the game's release date of 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 consists of about 3.5 million tweets that include Twitter-specific data points and metadata. At the outset of the project in its current form, we built corpora structured on the axis of language to provide a comparative analysis of the communication between users. The team selected Twitter over other platforms such as Discord and Reddit for its global reach reflecting the game's international fanbase, as well as being one of Nintendo's foremost methods of addressing their customers directly. Our hypothesis is that the analysis of textual productions would reveal discourse discrepancies among different linguistic groups of Twitter users based on cultural and media environment factors. In addition, as evidenced in other studies of the use of social media by economic stakeholders in videogame promotion contexts (Kim and Chandler, 2018), we anticipated the discovery of quantitative evidence demonstrating the impact exerted by console manufacturers and software producers on the reception of the game based on social media campaigns. Keeping in mind the objective of providing an easily reproducible research protocol, detailed accounts of the steps taken in the process of data collection and analysis are also kept.

Twitter, Hashtags and Textual Data in Videogame Community Analysis

Since its launch in 2006, the microblogging platform Twitter has, over time, become one of the web's cornerstone for the sharing and accessing of information on all topics ranging from politics, science, and entertainment. It ranks among the Internet's top 15 most consulted websites (Kemp, 2019) with 500 million tweets generated every day by 134 million active users (Lin, 2019) in 2019. Gaming, however, is fast becoming one of the leading topics discussed on the platform. In 2018, the total number of tweets produced on videogame-related topics such as esports, industry showcases and new releases reached one billion for the first time (Chadha, 2020), by the end of the following year, this number would grow by an additional 20% fueled by a 15% increase in unique Twitter accounts writing about gaming (Chadha, 2019). Since the release of the first version of its API in 2013, which made possible the gathering and analysis of large-scale textual corpora, Twitter has become a valuable source of large-scale textual data for journalistic and scholarly enquiries. The increasing presence of gaming culture on the platform also hints at its potential as a research field in game studies.

However, Twitter, just like other online platforms, is not a neutral communication space. Previous work indicates that Twitter's interface and content-recommender algorithms encourage specific communicative behaviors. Twitter distinguishes itself from other participatory web platform by restricting the number of characters used in any single tweets to 280, and by the decentralized and public nature of its content; unlike other social networking platform, twitter is not necessarily bidirectional in its sharing of content between users as any user has access to anyone's tweets (Murthy 2012, 1061-1062). From a scholarly perspective, Twitter has been given an increasing level of attention as a new human communication platform, calling for the development of critical theory and mixed methods to overcome the challenges related to the study of its content and exchanges (Bruns and Burgess, 2012; Java et al., 2009; Kayser and Bierwisch 2016). Its technological affordances (character limit, likes, retweets, etc.) exert an important level of influence on textual productions which, based on popularity metrics, grants more visibility to certain types of publications and content. While it can be difficult to clearly frame Twitter as a platform from which community-like structures emerge due to its nature as a microblogging website rather than a social network favoring interpersonal interactions, Gruzd, Wellman and Takhteyev (2011) suggest that Twitter's asymmetrical design nevertheless leads to the creation of both 'real' communities, marked by strong interpersonal ties, and imagined ones in which individuals share culture, a vocabulary, a language, traditions, a feeling of belonging and commitment through mediated discourse (Calhoun, 2016).

Thus, in the context of communicative acts addressing mixes of known and unknown individuals (Marwick and Boyd, 2010 129) in a social context both concrete and imagined, tweeting can be interpreted as an act of performance (Murthy, 2012 1065) in which codifying the form and content of one's discourse according to community standards becomes key in its spreading to other users. Thus, while Twitter users are free to produce posts of any types, both its algorithms and user culture push forward certain types of forms and content.

These dynamics are further accentuated by 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, 2017) 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, 2019 4) that influence "use of language and or the creation of collectives," thus, sometimes, creating a sense of community of discourse (5). Andreas Bernard highlights marketing applications of the 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" (57), but also emphasizes its role as an identity forming device that both empowers and levels online discourse. In this context, content broadcast through the hashtag is of little importance; it is the hashtag phrase that contextualizes the utterance and gives it meaning. As an indexing tool, hashtags are also not uniform in their functions; if most of them could be seen as topics formed around specific events or objects, others are statements "created for expressive purposes" (Rathnayake and Suthers, 2018 3), adding additional semantic layers to a post. In sum, hashtags must be read in their context in order to properly grasp the type of content that it links together.

Twitter data, in conjunction with the hashtag system, has been used to study a variety of subjects from public health surveillance (Heaivilin et al., 2011) to the public opinion on the Syrian refugee crisis (Öztürk and Ayvaz, 2018). In the field of game studies, this type of material has been used to examine recent debates in gaming culture (Blodgett and Salter; Todd et al.) while industry researchers are exploring ways to integrate them "into the traditional game analytics pipeline" (Milambiling et al., 2019 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. In addition, in the context of an increasing production of tweets on videogames in recent years, there is little theoretical work investigating the relationship between digital play and microblogging platform practices, and their potential inter influence.

The *Pokémon* Series and Player Communication

The *Pokémon SaS* Twitter community provided an ideal case study to assess natural language processing techniques through the analysis of emergent fan online discourse on a large international scale. Since Game Freak's release of *Pokémon Red and Blue* in 1996, the series has been considered one of the most popular and influential global media franchises of all time (Bulbapedia, 2020). Every new iteration of the series is met with fond anticipation and excitement by its wide international fanbase as every aspect of the game becomes the object of intense discussion, rumors and speculation online. Released worldwide for the Nintendo Switch on November 15, 2019, *Pokémon Sword and Shield* further heightened expectations as it stands as the first mainline title to be released on a powerful home console rather than on an exclusively handheld device with Game Freak promising "higher-quality animation for the new game" (Cooper, 2019) and hinting at an unprecedented degree navigational freedom through a vast open world. Combined with the Switch's more robust Internet connectivity, native screen capture functionality, social media integration and newcomer-attractive hybrid handheld/living room play styles, the release represented an important technological and cultural paradigm shift for the series.

The cultural and media 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 to the older audience (Keogh, 2017; Zsila et al., 2018), in its relation to the environment (Bainbridge, 2014; Dorward et al., 2017), in its localization challenge (Iwabuchi, 2004; Katsuno and Maret, 2004) as well as the practice of ROM hacking (Barnabé, 2018). In Japan, anthropologist Nakazawa Shin'ichi discussed the videogame's unique regime of knowledge and communicative aspects as devices that empower children in creatively deciphering and interacting with their environment (1997).

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 developer Game Freak contributes to the creation of a sense of community among players (2017). By establishing special events where trainers can meet and interact, Game Freak takes part in the maintenance of their player base while at the same time creating marketing events. These various events, originating from the developers, are paired with several third-party platforms created by fans. These initiatives such as Smogon University allows players to create and share content with others. This study highlights the implication of Game Freak and *The Pokémon Company* in the success and retention of players to the *Pokémon* series community through the years. Kim, Merrill and Song relate how the feeling of presence fostered by *Pokémon Go* gameplay contributes to a sense of community among players (2020). Therefore, the act of exploring their city to catch *Pokémon* creates a sense of attachment from the player to their environment.

Thus, the *Pokémon* franchise, through its emphasis on interpersonal interactions built into its core gameplay mechanics and its active fanbase online, constitute an ideal case study to test the potential of text mining methodologies on a large and diversified corpus. Although the studies previously mentioned explored the notion of community, they are specifically centered on the game *Pokémon Go* as an augmented reality game. Work that closely examines the dynamics of player communities online around the release of a conventional *Pokémon* game on console has yet to be done. Considering the popularity and the importance of this franchise, the proposed methodology may provide complementary analytical tools to investigate how players engage each other through the framework of the game over time. Examining the specific case study of *Pokémon SaS* Twitter community textual production around the period of its release provides an opportunity to question and examine these issues from a multilingual comparative perspective beyond the sole emphasis on English material.

Study Methodology and Initial Observations

The Twitter data gathering process using the official Twitter API started in early September. With no access to the premium API, the team used the free client, thus enforcing a strict regimen of scraping tweets weekly. In the rare instances where data could not be acquired due to the time constraints, we employed alternative python packages to complete the corpus. Tweets were gathered based on their association to a list of hashtags that covers the English, French and Japanese version of the game's title, and thus does account for all tweets on the topic for the entire duration[1]. Tweets in other languages nevertheless entered the database through their multiple hashtag association, but they were omitted from the text analysis section. The final corpus contains about 3.5 million tweets with key metadata points such as retweets, favorite and follower counts, usernames, detected languages, mentions, and timestamp. Due to the nature of the API, user information and conversation ID could not be saved, and, as a result, limited the reach of the investigation.

Upon conclusion of the scraping process, we standardized the corpus using python scripts and Open Refine (2012) before separating it into three sub corpora to facilitate subsequent steps and independently study each linguistic region. The main corpus, which includes most metadata, was integrated into a Tableau file through which general data trends could be observed and to which relevant events surrounding the game's release were added to contextualize the following analysis.

Complementing the data collection, the team kept track of all announcements, leaks and rumors related to the game's release in a detailed calendar in order to properly contextualize the fluctuation of textual production in the

[1] Hashtags used in the query are the following: #PokemonSwordShield #PokemonSword #PokemonShield #ポケットモンスターソード #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.

later step of the project. Covering the period between July 2019 and January 31, 2020, two major types of events were logged: activities related to Nintendo's marketing campaign both for the game and its proprietary console platform, as well as influential topics and controversies emerging from the broader *Pokémon* series community during the period studied. Both categories included noteworthy events and debates from the Japanese, English, and French sides, of which the former demonstrated considerable differences from its 'global' counterpart. Using this document, created through an iterative process, we were able to distinguish trends in the corpus but also identify specific marketing strategies, and the general reaction to specific features such as the Curry Dex or the auto-saving function widely emphasized by Nintendo. This calendar allowed us to compare the corpus gathered through Twitter scraping to broader events from a qualitative perspective.

To generate initial insights, we used KH Coder (2015) to perform a range of different basic explorations, including word counts, crosstabs, and word association networks. This toolkit is one of the few text mining applications 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 concept of 'shiny *Pokémon*' as one of the most mentioned game mechanics of the corpus, comprising 7.86%, 7.64% and 6.63% percent of the words in the English, French, and Japanese corpora[1]. A word association network analysis also revealed major recurring semantic fields related to these terms, but which change in importance based on the corpora studied (Figure 2, 3 and 4). The most significant observation of this phenomenon is the presence and discrepancy of what can be best described as a 'live-streaming' semantic field composed of words like 'Twitch', 'Shiny Hunting', and others, which occupies a central place in the French and English corpora. Comparatively, the major semantic fields generated in the Japanese corpus are associated with language related to game functionalities such as 'raids', various attributes, and trades. The cross-tab analysis revealed that all corpora showed a similar pattern of increase in frequency distribution over time from month to month, suggesting an ever-growing interest in this feature after release (Figure 5, 6 and 7). These findings justified turning our focus for the rest of our investigation towards this concept to further detail discourse differences from a comparative perspective.

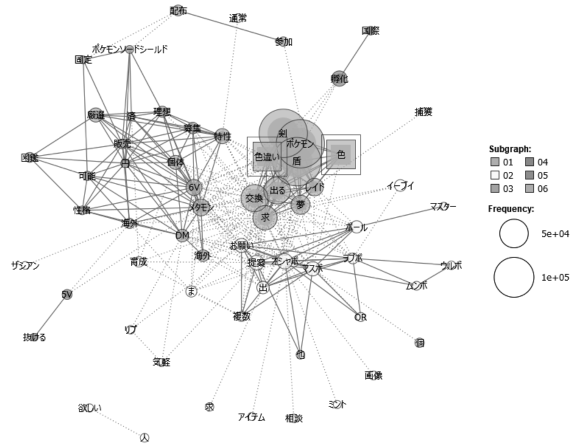


Figure 2: Shiny in the Japanese Corpus (jaccard 300)

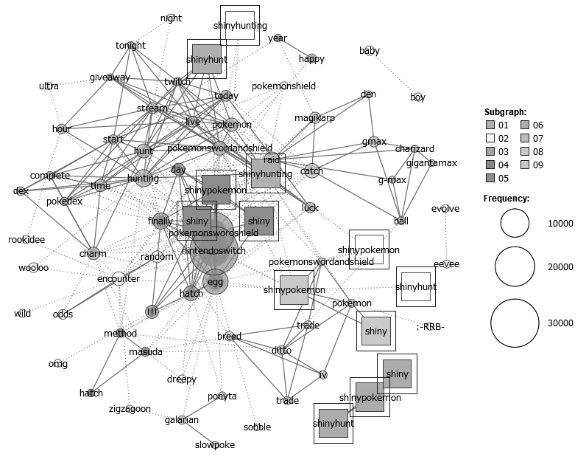


Figure 3: Shiny in the English Corpus (jaccard 300)

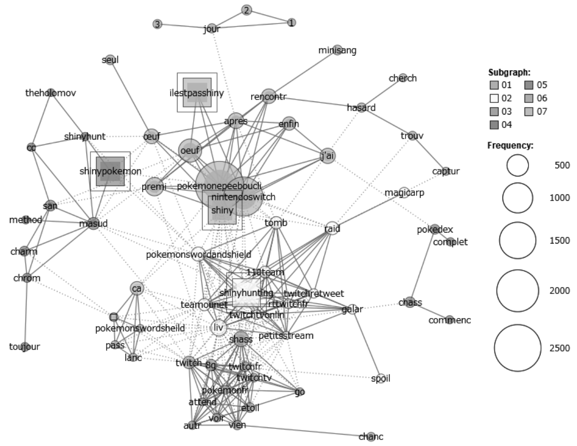


Figure 4: Shiny in the French Corpus (jaccard 300)

[1] 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.

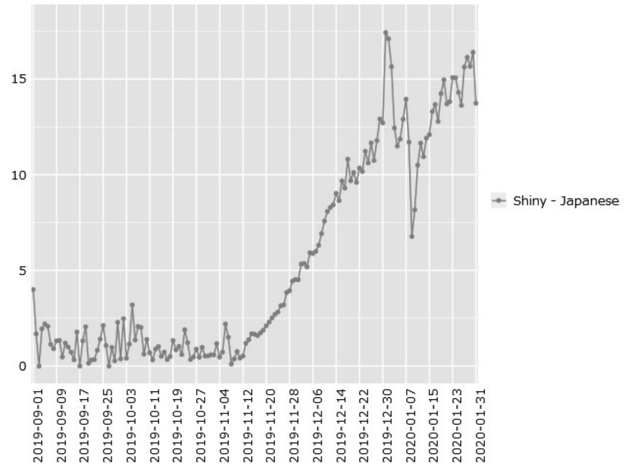


Figure 5: Crosstab table representing the frequency distribution percentage of the 'shiny' gameplay concept over time in the Japanese corpus

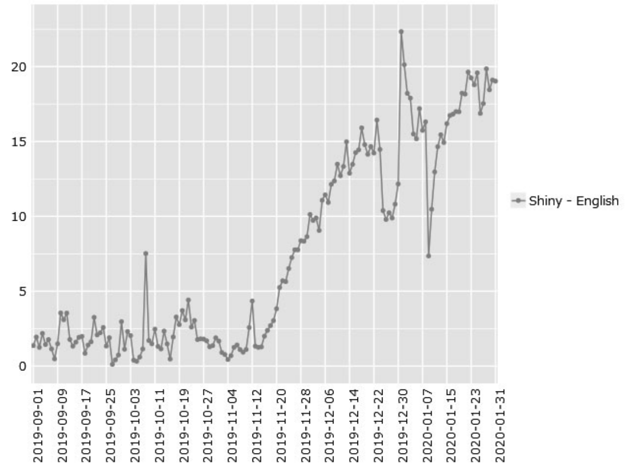


Figure 6: Crosstab table representing the frequency distribution percentage of the 'shiny' gameplay concept over time in the English corpus

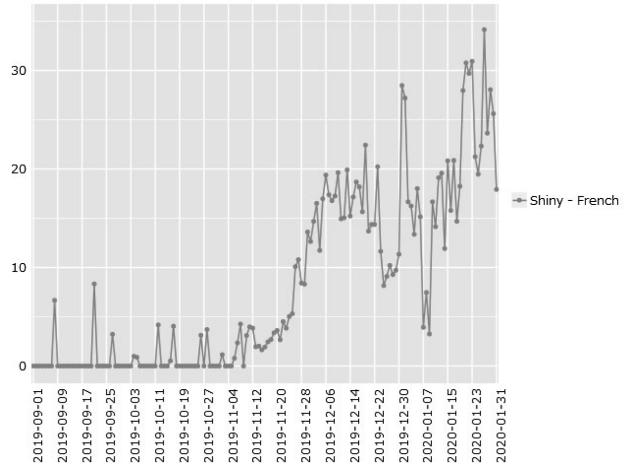


Figure 7: Crosstab table representing the frequency distribution percentage of the 'shiny' gameplay concept over time in the French corpus

The high frequency distribution of the shiny concept is meaningful when put in relation to the *Pokémon* series player community, its regime of values, and the affordance of Twitter as a communication platform. To briefly summarize the concept in terms of gameplay feature, shiny *Pokémon*, present in most games of the series, are similar in all aspects to regular *Pokémon* with the exception that they sport an alternative color pattern (Figure 8). Encountering a shiny *Pokémon* remains a rare occurrence with an initial appearance rate of 0,00024% (or 1/4096), odds that can, however, be improved by the acquisition of specific items, the use of specific breeding methods, or repetitively fighting specific *Pokémon*. 'Shinies', as they are also called, provide cosmetic late-game engagement elements that encourage player communication by sharing tips on how to maximize chances of encountering one. The acquisition of a shiny *Pokémon* generally involves many hours of repetitive play, while keeping in mind that there is no real guarantee of success. In the context of Twitter, these community dynamics are further accentuated as its short and impulsive affordance, as well as the hashtag system converge to valorize certain types of posts over others: the picture of a shiny *Pokémon* is likely to be rewarded with "likes." These tweets can be considered part of what Mia Consalvo (2007) calls videogame paratexts. Referring to the work of Pierre Bourdieu, Consalvo suggests that these paratexts allow players to grow their "game capital". According to her, the process of growing gaming capital is more than just playing videogames properly. It is also about developing a broad knowledge of the videogame industry, being aware of the latest gaming trends, knowing where to find reliable information about the tips and secrets of a particular game and being able to pass that knowledge on to others. Consalvo also supports the idea that paratexts have practically more influence on play than the game as such, in particular because they have a very specific educational function for players, that is to say that they teach them "how to play" and especially "how to be" as players that are part of a community.

In our case, the shiny concept, repurposed as a source of gaming capital or advertisement to bolster one's live-streaming viewership, 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 to produce and accumulate "public attention."

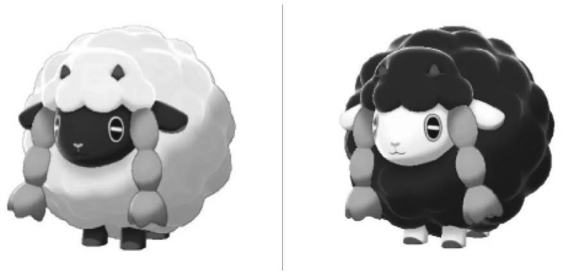


Figure 8:
Regular and shiny Wooloo
(nintendolife.comnews/2020/06/guide_pokemon_sword_and_shield_-_best_ways_to_catch_and_breed_shiny_pokemon)

We subsequently used the text mining package Mallet (McCallum, 2002) using lists of stop words (Le, 2016; Savand, 2020) adapted to a tweet-based corpus in order to generate 100 LDA topic models to group daily tweet production in dominant word groups in each linguistic corpus separately thus providing a more granular overview of the community's discursive trends. Many of these "bags of words" included some iteration of the shiny *Pokémon* concept along with other words. Since most of these topics were not directly comparable from a thematic perspective, we elected to classify them in one of three broad categories that emerged as a common feature in all corpora: (1), 'Affect', comprising of words indicative of expressive language indicating surprise, joy, and other emotions, as well as adjectives like 'cute' or 'adorable', (2), 'System', which include terms related to the playing of the game itself from the perspective of collection, maximization of assets, trade, battle and discussion of specific *Pokémon* attributes like statistics and nature types, and (3), 'Promotion', language related to marketing, influencers media, and words otherwise related to economic interest, which includes streaming, giveaway contests, and the resale of in-game assets on third-party websites. All topics were analyzed for classification in either three or none of these categories, and hybrid topics were classified in relation to the main trend shown therein. These concatenated entities provided a satisfyingly representative model from which a time-based multilingual comparison could be performed.

Analysis (1) - General Corpus: Language Distribution and Production Peaks

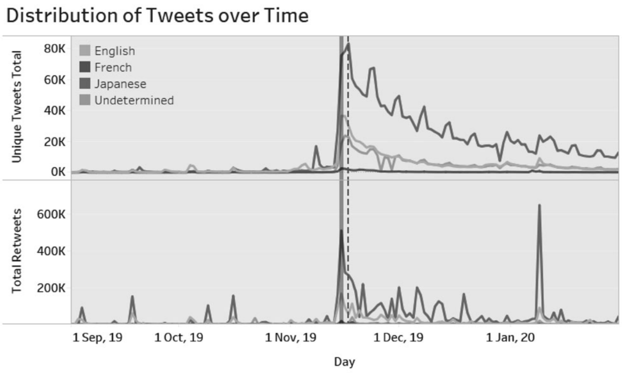


Figure 9: Distribution of Tweets over Time

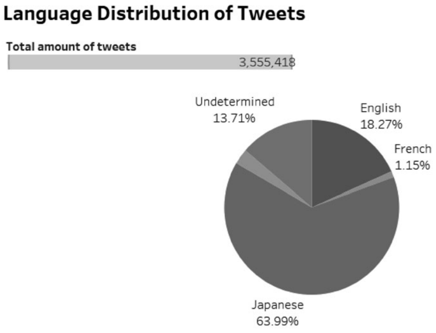


Figure 10: Corpus Language Distribution

Our first observation relates to the high volume of the Japanese corpus compared to the English and French ones (Figure 9 and 10). Of all tweets collected, 64% are tagged as written in Japanese, 18% in English, and about 1% in French, lower than the combined 'other' category set at under 3%. A fifth category, 'undetermined', representing 14% of collected tweets, identifies posts which do not include any text, and were thus ignored for the textual analysis process. Our initial postulate establishing the English corpus as the largest was eliminated early in the research process, which points to the need to develop cross-cultural perspective when studying online communities. Besides, it quickly became evident that linguistic regions did not evolve entirely independently from each other. The copresence of hashtags and text of multiple languages in many of the tweets gathered and topic models generated indicate freer translinguistic flow between different game communities than we had anticipated. Evidence of this phenomenon is most prominent in specific topic models such as the one assembling vocabulary related to fan art and illustrations. The broader *Pokémon Sas* Twitter community thus demonstrate evidence of a significant transcultural undercurrent centered on the Japanese language, which might consequently give rise to regimes of values defined by various influences and reinforces a broad transcultural outlook on such analysis.

Despite this situation, each linguistic corpus also demonstrates distinct discursive trends that speak of their community's inclinations in textual productions in relation to the degree of exposure to diverse economic stakeholders. For example, the Japanese corpus shows more influence by commercial interest in the sharing of content than in other languages. It also indicates regular textual production peaks every Sunday. In every case, however, game corporation-led events seem to drive retweet production but exert limited effect on the publication of new tweets by fans.

Several of the activity peaks in the *Pokémon SaS* hashtag are tied to direct marketing events, especially in the few weeks leading to the game's release. For example, the date of October 16, 2019, sees a significant increase in the amount of user-created content in the three languages observed. This date is associated with the release of a new trailer presenting the new 'Gigantamax' function. Reading the tweet production chart and Nintendo's marketing calendar, where significant actions were taken every five days in September and October, and every two days in November, it becomes evident that marketing discourse drives communication activity. However, the nature of its impact stays relative as data clearly shows that retweet activity shows extreme gains, but unique tweet production increases only slightly, suggesting a weaker degree of engagement. Nevertheless, this suggests that Nintendo and Game Freak are the principal creative force occupying the space of the hashtag, making it visible in preparation of its more active use after the game's release, where economic stakeholders' voices are diluted in the greater mass of publications.

The interrelation between marketing and content creation is even more significant while looking at the Japanese corpus because, as previously mentioned, the territory is the object of unique marketing campaign events (Figure 11). Some of them, including the publication of an AR camera filter reproducing *Pokémon* in the environment of its users as well as the cancelled holding of an event at the Pokémon Mega Center in Tokyo due to unexplained 'operational reasons' generated productivity peaks. The latter is significant because it generated a peak of unique tweet production discussing the event. The closest comparable community-led event in both the English (Figure 12) and French (Figure 13) languages is the denunciation campaign #Dexit where, in reaction to the lack of sufficient *Pokémon* assets in the game, fans called for the boycott of the game in protest. This event was immediately followed by a counter movement called #Thankyougamefreak where fans symbolically took the side of the game company amidst the barrage of criticism, invoking the positive impact that the franchise has had on their lives. Both events are marked by an increase of unique tweet production over retweet activity.

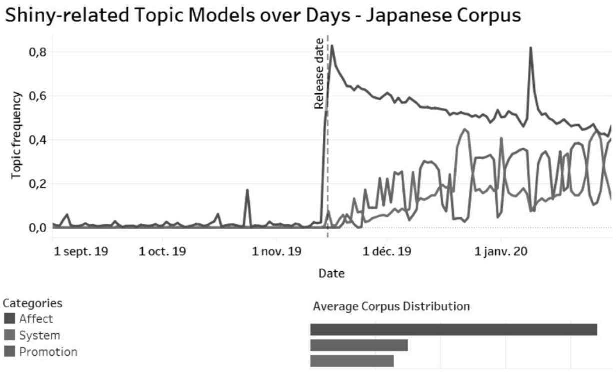


Figure 11: Shiny-related Topic Models over Days - Japanese Corpus

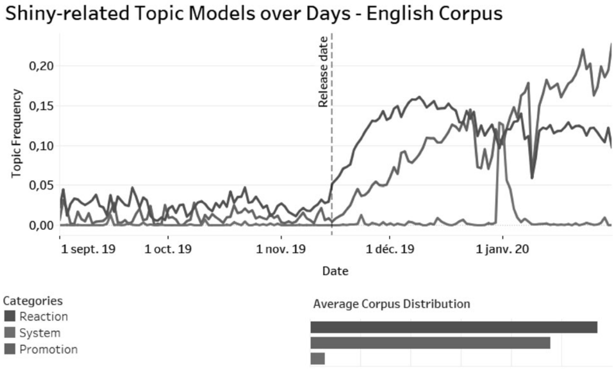


Figure 12: Shiny-related Topic Models over Days - English Corpus

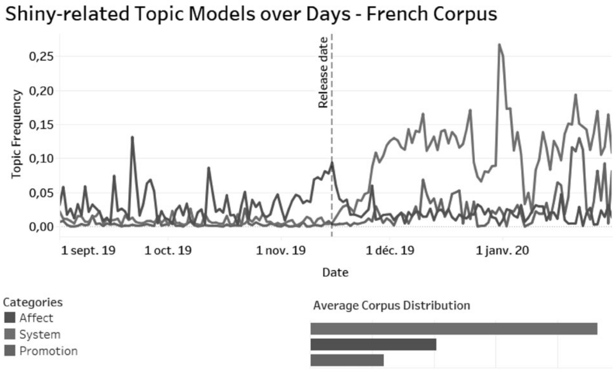


Figure 13: Shiny-related Topic Models over Days - French Corpus

Another point of comparison is the overall cross-linguistic text production over the five-month period. For all languages studied, the latter part of the corpus accounts for the majority of tweets produced, demonstrating a peak in production upon the game's release date followed by continuous decrease leading to the end of January 2020. The only notable event that slightly drives back tweet production is the January 9th announcement of the fall 2020 release of an expansion pass. However, this event does not influence the general downward tweet production tendency. Proportionally to the Japanese and English corpora, tweets produced in French would see a lower increase upon the game's release. The same corpus also shows a comparatively subdued downward production tendency, which is indicative of the more sustained engagement of a smaller pool of users.

Analysis (2) - "Shiny" Gameplay Mechanic and Topic Models

From the release of the game onward, mentions of the shiny *Pokémon* game mechanic increase in all corpora, following the general trend of textual production. However, as the latter quickly fades down, the former adopts the opposite direction, suggesting a higher degree of concentration of discourses around the concept over time probably tied to the late-game acquisition of shiny *Pokémon* and their desirability, but also by the anticipated positive reception from other users. 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. Many of the topics generated, including those that did not fit any of the categories, show unstable distribution, indicative of the highly disorganized and diverse ways in which the shiny gameplay concept was part of.

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 Japanese corpus, the most voluminous one, shows 'affect' as the initial dominating topic, although one that gradually diminishes over time after an initial 80% topic share characterizing tweets produced just after the game's release. Another tendency is the relative equivalence, and slow increase of 'system' and 'promotion' topics throughout the post release period up to January 31st where all three categories meet at comparable frequency levels. The surge of the 'promotion' category can be explained by the increasing number of tweets advertising the private sale of shiny *Pokémon* on online marketplaces such as RMT.club.

The English corpus indicates a similar point of departure with a more balanced pre-release tweet section regarding topic distribution, and a rise of the 'affect' category just after November 15th. However, the 'promotion' category would also quickly monopolize a higher share of the corpus later. Indeed, on January 1st a sudden paradigm shift occurs which sees all three topics at equal value followed by the clear rise of the 'promotion' category as the most discussed topic category. This event occurs concurrently with Nintendo's launch of a special in-game event where the Magikarp *Pokémon* type was given an increased spawn rate in the game's Wild Area, including shiny variants of the same *Pokémon*. Interestingly, this promotional event, the first of many of its kind that would be instilled in the game thanks to the console's constant Internet connectivity feature, tipped the balance towards advertisements and promotion as the most represented topic category.

While both Japanese and English corpora demonstrate very distinct shiny-related topic distribution between the pre-and post-release of the game, the French corpus represents an outlier case from this perspective. Indeed, while being equality diversified in topic frequency, the prerelease section indicates a high affinity for the 'affect category', which then drops significantly in the latter half of the examined timeline. The release of the game also does not seem to indicate a significant change in topic distribution, it is only two weeks later than we see the 'system' category take and keep the lion's share of the textual data, while maintaining at a relatively low 20% frequency. 'Promotion', however, mainly represented by the advertisement of live streams and influencer accounts, only becomes important in the last two weeks of the corpus. The importance of the 'system' topic category points at gameplay-related utilitarian discussions as being the norm in the French language corpus, upstaging affective language, mostly associated with the hype phenomenon leading to the game's release. In the context of a smaller textual pool, and a smaller sub community of interest, the association of the discussion on shiny *Pokémon* with more practical aspects of the game is indicative of a dedicated group mostly interested in securing rare game assets and achieving high levels of proficiency over the game's systems.

Conclusions and Further Developments

This project concerned the assessment of text mining tools and methodologies for the large-scale study of the reception of videogames using social media data. Doing so, we evaluated that content creation made by the *Pokémon Sas* community on Twitter was partly influenced by the promotional strategy of Nintendo and more specifically during the period preceding the release of the videogame, albeit mainly through retweets as opposed to the production of original publications. In addition, just before and following the release date, marketing-influenced discussions gradually gave way to emergent community conversations demonstrating an appropriation of the game through standards and practices. While Nintendo's impact is palpable in the database, it remains weak, short-timed, and does not alter the direction of long-term major discursive trends.

Combining different text analysis and data visualization techniques, we proposed a vertical research protocol for the analysis of social media textual data. Its application on the case study of the *Pokémon Sas* Twitter community allowed the identification of emergent fan discourse in regards their appropriation of the game's features. We focused on the analysis of the reception of the gameplay mechanics of shiny *Pokémon* and, using unsupervised topic modelling, we identified three main topic categories and their frequency rate over time. We posit that these topic categories represent broad tendencies in which the shiny *Pokémon* gameplay feature is discussed and appropriated. These findings are also framed in the context of the technological affordances and algorithms of Twitter as a platform that grand more visibility to popular posts, which further influence the formation of the *Pokémon* Twitter fan community as an imagined community. As such, we posit that for emergence of the shiny Pokémon game mechanic as a major discursive trend amongst Twitter user is the result of the complex interplay of the game's design, its actualization by players, and the affordances of the concerned communication platforms. These results are also specific to Twitter, and we anticipate that conducting similar research protocols on textual data gathered on other social media networks would yield different and complementary results.

We also examined differences in reception based on what we defined as linguistic regions on the platform. While results obtained indicate significant differences between them, we wish not to draw a conclusion that would reinforce essentialist discourses based on innate cultural characteristics and interests. Rather, they reflect the complex convergence of platform affordances and media environments that are specific to these regions. We also noted cross-linguistic participation of fans on social media textual production, which implications are difficult to quantify at this moment and calls for the development of more efficient language parsing methodology that better represents this phenomenon. While local media environments exert a significant influence in the direction of engagement, in digital contexts, discourse tend to blend and hybridize as linguistic communities clash and mix. Users engaged in more than one of such communities act as cultural brokers that enable the circulation of ideas from linguistic territory to another. With machine translation functionality integrated to most online platforms, translinguistic communication may become even more seamless between user groups, further linking discourses, communicative behavior, and play practices in the context of game culture. While this project aimed at identifying and comparing dominant tendencies among linguistic regions, future work should target the communicative behavior and the ramifications of the online engagement of these key users to explore this phenomenon further.

This investigation constitutes the project's first case study and should inform future iterations of the research protocol. The official Twitter API proved very constraining in terms of the rhythm of data acquisition, and alternatives will be explored to gain more scraping flexibility and data points. Secondly, subsequent work will focus on social metadata to identify community leaders and their impact more precisely. Finally, further research could focus on the analysis of vocabulary diversity and the adaptation of sentiment analysis tools to evaluate gaming-themed corpora to obtain complementary data that could provide different perspectives.

Acknowledgements

This project was funded by the Fonds de recherche du Québec and the Social Sciences and Humanities Research Council of Canada.

References

Assunção, Carina, Michelle Brown, and Ross Workman. 2017. "Pokémon Is Evolving! An Investigation into the Development of the Pokémon Community and Expectations for the Future of the Franchise." Press Start 4:1 17-35.

Bainbridge, Jason. 2014. "'It Is a Pokémon World': The Pokémon Franchise and the Environment." International Journal of Cultural Studies 17:4 399-414.

Barnabé, Fanny. 2018. "Between Freedom and Constraint: Rom Hacking of Pokémon Games." DiGRA JAPAN, Fukuoka.

Bernard, Andreas. 2019. Theory of the Hashtag. Cambridge: Polity Press.

Blodgett, Bridget Marie and Anastasia Salter. 2013. "Hearing 'Lady Game Creators' Tweet: #1reasonwhy, Women and Online Discourse in the Game Development Community." 14th Annual Conference for the Association of Internet Researchers.

Bruns, Axel, and Jean Burgess. 2012. "Notes Towards the Scientific Study of Public Communication on Twitter." In Science and the Internet, edited by Alexander Tokar, Michael Beurskens, Susanne Keuneke, Merja Mahrt, Isabella Peters, Cornelius Puschmann, Timo van Treeck and Katrin Weller. Düsseldorf: Düsseldorf University Press, 159-69.

Bulbapedia. 2020. "Pokémon Games." Bulbapedia, the Community-Driven Pokémon Encyclopedia. Accessed January 4. https://bulbapedia.bulbagarden.net/wiki/Pok%C3%A9mon_games.

Calhoun, Craig. 2016. "The Importance of Imagined Communities – and Benedict Anderson." Annual Review. Debats. Revista de Cultura, Poder i Societat 1 11–16.

Chadha, Rishi. 2020. "2019 Gaming on Twitter." Blog.Twitter.com. https://blog.twitter.com/en_us/topics/events/2019/2019-gaming-on-twitter.html

———. 2019. "Gaming Grabs the High Score on Twitter." Blog.Twitter.com. https://blog.twitter.com/en_us/topics/events/2019/Gaming-grabs-the-high-score-on-Twitter.html.

Consalvo, Mia. Cheating: Gaining Advantage in Videogames. Cambridge: MIT Press, 2007.

Cooper, Dalton. 2019. "Pokémon Sword and Shield Dev Faces Backlash Over Animations." Game Rant. November 13. https://gamerant.com/pokemon-sword-shield-animations-bad-game-freak-lied/.

Dorward, Leejiah J. , John C. Mittermeier, Chris Sandbrook, and Fiona Spooner. 2017. "Pokémon Go: Benefits, Costs, and Lessons for the Conservation Movement." Conservation Letters 10:1.

Ensslin, Astrid. 2012. The Language of Gaming. New York: Palgrave Macmillan.

Faisal, Ali, and Mirva Peltoniemi. 2018. "Establishing Video Game Genres Using Data-Driven Modeling and Product Databases." Games and Culture 13:1 20-43.

Gruzd, Anatoly, Barry Wellman, and Yuri Takhteyev. "Imagining Twitter as an imagined community". American Behavioral Scientist 55:10 1294-1318.

Heaivilin, N., B. Gerbert, J. E. Page, and J. L. 2011. Gibbs. "Public Health Surveillance of Dental Pain Via Twitter." Journal of Dental Research 90:9, 1047-51.

Higuchi, Koichi. 2015. "KH Coder: Quantitative Content Analysis or Text Mining." http://khc.sourceforge.net/.

Huynh, David. 2012. "Open Refine." https://github.com/OpenRefine/OpenRefine#contact-us

Iwabuchi, Koichi. 2004. "How 'Japanese' is Pokémon." Pikachu's Global Adventure: The Rise and Fall of Pokémon. Durham: Duke University Press, 53-79.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2009. "Why We Twitter: An Analysis of a Microblogging Community." Advances in Web Mining and Web Usage Analysis Conference, Berlin, Heidelberg.

Jockers, Matthew Lee. 2013. Macroanalysis: digital methods and literary history. Urbana, Chicago and Springfield: University of Illinois Press.

———. 2014. Text analysis with R for students of literature. New York: Springer.

Kayser, Victoria, and Antje Bierwisch. 2016. "Using Twitter for Foresight: An Opportunity?". Futures 84 50-63.

Kang, Ha-Na, Hye-Ryeon Yong, and Hyun-Seok Hwang. 2017. "A Study of Analyzing on Online Game Reviews using a Data Mining Approach: STEAM Community Data." International Journal of Innovation, Management and Technology 8:2 90-94.

Katsuno, Hirofumi and Jeffrey Maret. 2004. "Localizing the Pokémon Tv Series for the American Market." Pikachu's Global Adventure: The Rise and Fall of Pokémon. Durham: Duke University Press, 80-107.

Kemp, Simon. 2019. "Digital 2019: Global Internet Use Accelerates." We are social. https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates.

Keogh, Brendan. 2016. "Pokémon Go, the Novelty of Nostalgia, and the Ubiquity of the Smartphone." Mobile Media & Communication 5:1 38-41.

Kim, Jihyun, Kelly Merrill Jr, and Hayeon Song. 2020. "Probing with Pokémon: Feeling of Presence and Sense of Community Belonging." The Social Science Journal 57:1 72–84.

Kim, Yuna, and Jennifer D. Chandler. 2018. "How Social Community and Social Publishing Influence New Product Launch: The Case of Twitter During the Playstation 4 and Xbox One Launches." Journal of Marketing Theory and Practice 26:1-2 144-57.

Le, Van-Duyet. 2016. "Japanese-stopwords." Github repository. https://github.com/stopwords/japanese-stopwords.

Liang, Yuxi. 2019. "Fashion of Game Consoles Seen from Journalism Consideration on Game Reviews in 'Weekly Famitsu'." The Journal of Replaying Japan 1 79-92.

Lin, Ying. 2019. "10 Twitter Statistics Every Marketer Should Know in 2019 [Infographic]." Oberlo Blog. https://www.oberlo.com/blog/twitter-statistics.

Marwick, Alice E. and Danah Boyd. 2010. "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience." New Media & Society 13:1 114-133.

McCallum, Andrew Kachites. 2002. "Mallet: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edushttp://mallet.cs.umass.edus.

Milambiling, Lareina, Michael Katchabaw, and Damir Slogar. 2019. "Integrating Social and Textual Analytics into Game Analytics." in Data Analytics Applications in Gaming and Entertainment, edited by Günter Wallner, Boca Raton: CRC Press, 141-67.

Mohri, Hitomi, Fukuda Kazufumi, and Hosoi Koichi. 2019. "Research on the User Information Requirements of Video Games Resources for Subject Access - Quantitative Text Analysis of a Q&A Website -." The Journal of Replaying Japan 1 118-135.

Moretti, Franco. 2013. Distant Reading. New York: Verso.

———. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. New York: Verso.

Murthy, Dhiraj. 2012. "Towards a Sociological Understanding of Social Media: Theorizing Twitter." Sociology 46:6 1059-1073.

Nakazawa, Shin'ichi. 1997. Poketto no naka no yasei: Pokemon to kodomo (the Wilderness in the Pockets: Pokémon and Children). Tokyo: Iwanami Shōten.

Nitsche, Michael. 2008. Video Game Spaces: Image, Play, and Structure in 3d Game Worlds. Cambridge: MIT Press.

Öztürk, Nazan and Serkan Ayvaz. 2018. "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis." Telematics and Informatics 35:1 136-147

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

Pelletier-Gagnon, Jérémie. 2018. "'Very much like any other Japanese RPG you've ever played': Using undirected topic modelling to examine the evolution of JRPGs' presence in anglophone web publications." The Journal of Gaming and Virtual Worlds 10:2 135-148.

Rathnayake, Chamil, and Daniel D. Suthers. 2018. "Twitter Issue Response Hashtags as Affordances for Momentary Connectedness." Social Media + Society 4:1 1-10.

Sapach, Sonja. 2020. "Tagging My Tears and Fears: Text-Mining the Autoethnography." Digital Studies/le Champ Numérique, 10:1.

Savand, Alireza. 2020. "stop-words." Github repository. https://github.com/Alir3z4/stop-words

Serebii.net. Shiny Pokémon. https://serebii.net/swordshield/shinypokemon.shtmlhttps://serebii.net/swordshield/shinypokemon.shtml.

Suomela, T., Chee, F., Berendt, B., & Rockwell, G. 2019. "Applying an Ethics of Care to Internet Research: Gamergate and Digital Humanities." Digital Studies/le Champ Numérique 9:1.

Zagal, José P., Noriko Tomuro, and Andriy Shepitsen. 2012. "Natural Language Processing in Game Studies Research: An Overview." Simulation & Gaming 433 356-373.

Zsila, Ágnes, Gábor Orosz, Beáta Bőthe, István Tóth-Király, Orsolya Király, Mark Griffiths, and Zsolt Demetrovics. 2018. "An Empirical Study on the Motivations Underlying Augmented Reality Games: The Case of Pokémon Go During and after Pokémon Fever." Personality and Individual Differences 133 56-66.