A Machine Learning Approach To Cross-Cultural Narrative Patterns in Comics

Maarten Jansen

Tilburg University

Abstract

This work investigates culture and cultural influence of visual narrative structures in comic books, using a machine learning approach. It is motivated by proof that comics are susceptible to adjustments stimulated by culture, and the scarcity of a machine learning approach in studies on comics. Our main contribution is a better understanding concerning cross-cultural differences in visual language, and evolvement over time from a novel perspective. We find this contribution by asking the following research question: To what extent can we predict continent, country, and decade of publication based on narrative patterns in comic books published in the last eight decades? The answer to this question is obtained by training a decision tree and naive Bayes classifier, using narrative patterns extracted from the Visual Language Research Corpus. The results demonstrate that there are cultural distinctions in patterns if we examine continent or country of publication. However, if we examine decade of publication we witness fewer characteristics pointing towards changes over time, unless the time period investigated is of sufficient length, for example 50 years.

Keywords: visual language, pattern extraction, cross-cultural, machine learning, comic books

A Machine Learning Approach To Cross-Cultural Narrative Patterns in Comics

Introduction

Superheroes are more popular than ever since the creation of the Marvel Cinematic Universe(MCU). A universe based on the superhero stories told in comic books from the past. This rich source of storylines is loved by fans of different cultures over the world and continues to be loved in the future. As more of these stories are created, and more comic books from different cultures make their way across the world, the style of drawing and storytelling between cultures becomes intermingled. This occurred multiple times in the past eight decades (Wong, 2006).

Dispersing comic books across multiple countries involves translating the texts to the corresponding native language. However, drawings are a general concept and do not have to be translated for them to be understood. A drawing of a lion will always be recognized as a lion, no matter where in the world you are, as long as the viewer knows what a lion looks like (Iyyer et al., 2017). This last part is important as the viewer needs to make an inference based on the knowledge available to him or her. These can be inferences based on world knowledge as well as inferences based on drawings occurring in a sequence, like in a comic book. This cognitive understanding of comic books is a widely studied subject in cognitive linguistics, which attempts to comprehend meaning in linguistic structure with respect to general cognition. This is where the analysis of comics has been exceedingly ubiquitous since the publication of Scott Mccloud's (1993) Understanding Comics (Cohn, 2012). Despite this growth in research on comics, a minor part of these unrelated works are inspired by a central theory of language. This is where the theory of "Narrative grammar" proposed by Cohn (2013), based on linguistics and visual languages, becomes significant. A visual language is, according to Horn (1998), a "language composed of tightly integrated textual and visual elements". Cohn (2013) establishes a visual narrative structure to describe how humans create meaning out of sequential images and has been developing this theory by analysing comic books in subsequent papers.

This study uses the visual language presented by Cohn (2011, 2013) as described by Cohn, Taylor, and Pederson (2017) to find narrative patterns across culture. We will do so by ascribing patterns to three properties of comic books, the decade of publication, the country of publication, and the continent of publication. The question we ask ourselves is as follows: To what extent can we predict the decade, country, and continent of publication based on narrative patterns in comic books published in the last seven decades? We attempt to answer this by asking the following sub questions as well: Which features are paramount for correct predictions? Which distinctions do we see between comics from various cultures? What patterns can we attribute to culture and cultural influence? And, how does the result relate to the visual language theory developed by Cohn (2011)? The answers to these questions support other studies and instigate more research.

Related work

In recent years, a significant amount of studies focussed on the cognitive understanding of comic books. Scott McClouds publication of Understanding comics gave way for a wave of research on comics (Cohn, 2012). The development of a visual language has particularly been advancing continuously over the past years. A visual language according to Horn (1998), is a "language composed of tightly integrated textual and visual elements". In recent years, Neil Cohn worked on the visual language theory where he explored topics such as visual language, visual narrative grammar, and page layout (*Visual Language Lab*, 2018). He specifically investigated cross-cultural differences and demonstrated that visual narratives in Japanese and American comics are affected by cultural practice, resulting in patterns, and suggested that pages from different types of comics have different systematic characteristics (Cohn et al., 2017; Cohn, Taylor-Weiner, & Grossman, 2012). This points towards cultural heterogeneities in visual narratives across the world. More proof of this is found by looking at differences in natural languages, like English or Dutch, where it was discovered that distinctions between languages and culture are expressed and found in comics (Tversky & Chow, 2017). Thus, we expect to encounter diverging properties in cross-cultural comics and deliver proof from an alternative perspective, more on this later.

Technological advancements and changes in market strategies, moving from little visual adaptation to much visual adaptation, allow for the production of various comic editions customized to local markets. Furthermore, visual adaptations may be caused by publishers persuaded by cultural conventions or taboos, such as visual censorship by Disney comics published in the Arab peninsula (Zanettin, 2018). As commercial success is of great importance for publishers to continue to exist, it becomes a logical choice to adapt comics to abide by local customs. Here we see adaptation of comics to comply with local culture. However, looking back at past trends, we see cultural changes in comics caused by comic books of a separate culture. Notably is the introduction of American comics to Asia in the 1950s and 1960s, causing work of the great comics artist Tezuka Osamu to become influenced by Walt Disney and Max Fleisher animations. At this time, Japanese comics called manga start to take of in Japan and begin to influence its neighbours, as well as Europe and North America in the 1990s (Wong, 2006). It is clear that comics are susceptible to changes motivated by culture, since the 1950s. However, it is most likely a reoccurring phenomenon since the conception of comics, worth of investigating.

The inception of a visual narrative grammar, describing basic narrative categories and their relationship to an established narrative arc, makes it possible to analyse narrative structure within comic books (Cohn, 2013). Since comics change over time, it makes sense that the visual narrative structure changes as well. An example of this are changes over time in American comics, moving from an emphasis on text and shorter structures to visually emphasised longer structures (Cohn et al., 2017). These changes are often accredited to the introduction of Japanese manga on the American market in the 1990s (Wong, 2006). Cohn et al. (2017) express these changes in comics through a framework which categorizes panels and sequences of panels to apply a higher level of understanding to visual narratives, similar to grammar for natural languages. For example, when looking at English sentences we can explain the structure by looking at verbs, nouns, adjectives etc. A similar approach is taken by the visual narrative grammar; a sequence of panels in a comic can be explained by attentional framing categories and semantic relationships. Attentional framing categories designate in which way panels accentuate attention, whereas semantic relationships portray associations between panels (Cohn et al., 2017, 2012). They are successfully used in highlighting cross-cultural differences, can illustrate what shifts in narrative structure have occurred over time, and can be used to describe comics from a higher level of abstraction (Cohn, 2015; Cohn et al., 2017, 2012). However, they are mostly used in a singular sense or to describe patterns, not to form patterns and check their occurrence in comics or classify comics based on patterns, which is the goal of this study.

In addition, studies on cultural distinctions exploring narrative patterns in comic books often use traditional statistical approaches as opposed to a machine learning approach, despite a near simultaneous growth of research on comic books and advancements in machine learning (Langley, 2011; Wong, 2006). Some recent examples comprise the use of neural networks to draw inferences between panels in comic book narratives (Iyyer et al., 2017), the extraction of structural lines in manga using a neural network (Li, Liu, & Wong, 2017), and the proposal of Manga FaceNet, a neural network to detect manga faces of various appearances (Chu & Li, 2017). They utilize neural networks, however, this study is the first time a machine learning approach is considered in combination with the visual language theory developed by Neil Cohn, and aims to inspect cross-cultural changes in narrative structure from an unprecedented perspective.

It is common in machine learning to test multiple classifiers given the no free lunch theorem (Caruana & Niculescu-Mizil, 2006). However, the focus of this study is not to optimize performance but to discover how and why labels are classified. Given that the how and why are more important than optimizing performance, a decision tree is considered best option since it is interpretable and allows to review the decision making process (Pedregosa et al., 2011). A second classifier, naive Bayes, is considered as it operates under the assumption of conditional independent features, thus permitting comparison with a decision tree classifier which operates under the assumption of conditional dependent features. The features of comic books are narrative patterns and are usually extracted through sequential pattern mining, a process to ascertain interesting subsequences in a sequential dataset (Fournier-Viger, Lin, Kiran, Koh, & Thomas, 2017). Sequential pattern mining algorithms are considered with the help of the Sequential Pattern Mining Framework(SPMF) data mining library (Fournier-Viger et al., 2016). However, these algorithms extract the same patterns using slightly different methods, only affecting computational performance. Since the dataset is small with only 300 sequences, the general sequential pattern mining algorithm is considered the best choice (Srikant & Agrawal, 1996).

In summary, narrative patterns have been changing over time and there is proof of cross-cultural differences. So far, mostly traditional statistical approaches have been used to analyse comics, nonetheless, this paper aims to take a machine learning approach. Thus, a decision tree classifier and naive Bayes classifier are considered, to better investigate the decision making process. The general sequential pattern mining algorithm is considered to extract patterns, to be used as features, from the data. This study gives answers to questions such as: To what extent can we predict decade, country, and continent of publication based on narrative patterns in comic books? Which distinctions do we see between comics from various cultures? And, what patterns can we attribute to culture and cultural influence?

Method

A significant part of a successful classification based method lies with the training data. The dataset used for this study is accumulated from multiple studies. These studies annotated comic books based on the visual narrative grammar laid out by Cohn (2013). As mentioned before, the visual narrative grammar describes a framework to categorize panels and sequences of panels to apply a higher level of understanding to visual narratives. The manner used to annotate the dataset are attentional framing and semantic relations. Attentional framing comprises four variations to describe panel contents in comic books. The variations include macro panels, depicting multiple interacting entities (Figure 1a); mono panels, demonstrating only singular entities (Figure 1b); micro panels, portraying less than a singular entity (Figure 1c); and amorphic panels, where no active entities are illustrated(Figure 1d) (Cohn et al., 2017).

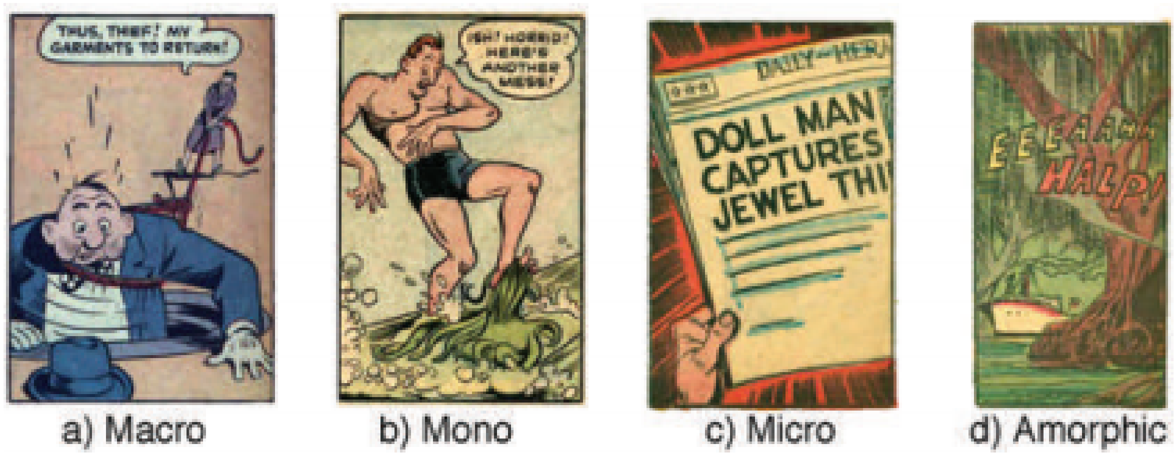

Figure 1. Examples of attentional framing categories. (a–b) come from Lady Luck by Klaus Nordling (1949), and (c–d) come from Doll Man by Bill Quackenbush (1950). (Image adopted from Cohn et al. (2017)).

Semantic relations explain the relationship between panels when presented in a narrative sequence. The relations between panels revolve around character changes, a change in characters from panel a to panel b; spatial changes, an alternation in locations between panels; time changes, permuting time from panel a to panel b; and causal changes, where panel b outlines the effect of a cause in panel a (Cohn et al., 2017).

Data structure and pre-processing

The dataset consists of 300 excel files containing sparse data representing the annotation of panels. Each file represented one comic book compromising around 300 comics in total. These files were merged and cleaned before the sparse annotations were transformed to narrative grammar sequences. An example can be seen in Table 1.

Table 1

An example of three panels and their annotation using attentional framing and semantic relationships.

Panel	Micro	Amorphic	Spatial	Causal
1	1		x	x
2	1			
3		1	1	1

Note. Additional coding is used in the complete dataset.

The semantic relation categories relate to inferences between panels. This property can be interpreted in Table 1 as following. When a relationship occurs this is denoted by a ‘1’ in the correct column, and refers to the relationship between the panel marked by a ‘1’ and the preceding panel. In Table 1 this is true for panel 2 and panel 3 where a spatial and causal relationship exists. Table 1 additionally exhibits the use of an ‘x’ at the first panel for the spatial and causal column. This is because semantic relations cannot exist prior to the first panel of a comic book. Figure 2 illustrates how data from Table 1 is represented by a sequence of attentional framing and semantic relationships. This process is applied to all comic books in the dataset involving more than 100,000 panels and their corresponding attentional framing and semantic relationships. The data distribution of panel coding can be found in Figure 3d.

Panel 1 Panel 2 Sem. rel. Panel 3

↓ ↓ ↙ ↘ ↓

['micro', 'micro', 'spatial', 'causal', 'amorphic']

Figure 2. The sequence of panels presented in Table 1 transformed to a sequence of attentional framing and semantic relationships. This transformation is applied to each comic book in the dataset and results in the list of narrative grammar sequences.

Feature extraction and selection

As mentioned previously, the combination of attentional framing and semantic relations transforms sequences of panels to sequences of attentional and semantic categories. These sequences are used to generate narrative patterns to train the predictive model. Generating narrative patterns is achieved by using the General Sequential Pattern mining algorithm(GSP).

GSP takes a sequence database and a minimum support threshold as input. A sequence database is, in this case, the list of narrative grammar sequences produced during pre-processing. A minimum support threshold is a value used to decide if a feature is important. The minimum support threshold is chosen by the user and is set at 30 percent for this study. It is set at 30 percent because we expect patterns belonging to a single culture to occur rarely when put in perspective to patterns occurring across the world. This means that a feature is deemed important when it occurs in at least 30 percent of the items in the dataset. GSP outputs a set of frequent sequential patterns and works as following according to Fournier-Viger et al. (2017).

Sequential pattern mining algorithms can be designated as being depth-first search or breadth-first search. The GSP algorithm is considered a breadth-first search algorithm, meaning it first scans the sequence database to find frequent 1-patterns, which are patterns containing a single item. The next step involves extending 1-patterns to generate 2-patterns, then to extend 2-patterns to create 3-patterns, continuing until no patterns can be produced. After each step of extending patterns, the algorithm evaluates which pattern is a sequential pattern by checking if the support is higher than the minimal support threshold. The support of patterns is calculated in two steps. First, for a pattern A, GSP checks if sub-sequences residing within pattern A are frequent as well. If pattern A contains a sub-sequence that is not considered frequent, pattern A cannot be frequent as well. This is called the downward-closure property which states that a sequence cannot be frequent if a subset of that sequence is infrequent (Tan, Steinbach, & Kumar, 2005). Second, if this is not the case, the GSP algorithm scans the sequence database to calculate the support of pattern A. If the support crosses the minimal support threshold, pattern A is frequent and considered output. This approach has successfully been applied to identify human behaviour patterns and continues to be used as base for other algorithms, such as an algorithm to monitor the pattern extraction process for meteorological data. (Bureva, Sotirova, & Chountas, 2015; He, Zhang, & Wu, 2018; Llerena et al., 2018).

In short, the GSP algorithm generates patterns from a sequence database and compiles all possible patterns while increasing pattern length at each pass. After each pass, the algorithm calculates the support of patterns across the entire dataset. If the support crosses a threshold of 30 percent, it is deemed important and elected as a feature. Past studies exemplify that the GSP algorithm is applicable in multiple fields of research, this is proof of the versatility and potential of the algorithm. To clarify, a pattern extracted by the GSP algorithm will now be referred to as a feature.

The features extracted by GSP are used to transform narrative grammar sequences to equal length vectors. For example, let us examine five features (F_1, F_2, F_3, F_4, F_5) where a sequence A contains F_1, F_3, and F_5, and a sequence B contains F_1 and F_2. We represent these features within narrative grammar sequences as (1, 0, 1, 0, 1) for sequence A and (1, 1, 0, 0 ,0) for sequence B. We do this to ensure equal length items when training and evaluating predictive models.

Additionally, we notice that the vectors do not acknowledge duplicate patterns and, therefore, do not include the frequency of a pattern within a comic book. A further remark worth mentioning is the restriction of the GSP algorithm. The algorithm compiles all possible patterns and is, therefore, computationally heavy.

The number of patterns(L) the algorithm considers can be calculated with the formula:

$$L = i^N$$

Where N is the maximum length of a pattern and i is the number of categories from which the algorithm can choose to fill the pattern. For this study the value for i is 8, and we choose a value of 6 for N causing the GSP algorithm to check the occurrence of 262144 patterns. Choosing a higher value for N exponentially increases the calculation time and is, therefore, not considered.

The GSP algorithm constitutes feature extraction and is the first part of feature engineering, feature selection is the second part. We expect some features to be more valuable than others, therefore, we apply recursive feature elimination combined with cross validation (RFECV) to produce the best features. Recursive feature elimination (RFE) determines features by recursively examining smaller and smaller sets of features. The algorithm does so by applying weights to features by training the classifier on the initial set of features, in this case the features extracted by the GSP algorithm. The least important features, the features with the smallest weight, are then dropped from the set of features. This procedure is recursively replicated until a desired number of features is reached.

Recursive feature elimination requires a predefined number of features to retain, however, it is unclear what the optimal combination or number of features are beforehand. These are found by applying cross-validation in combination with recursive feature elimination. For example, we run a 5-fold RFECV. For each split, the train set, produced by the cross validation, is transformed by RFE for N times, where N is the number of features. The classifier, a decision tree, is trained on the training set and evaluated on the test set. RFECV converges on the features producing the best evaluation score and returns features ranked by their importance.

Past studies demonstrate the versatile applicability of RFECV. For example, it has successfully been utilized in forecasting soccer injuries based on GPS data, the large-scale identification of birds by audio recordings, and in early detection of faults in heating, ventilation, and air conditioning systems (Chakraborty & Elzarka, 2019;

Lasseck, 2014; Rossi et al., 2018). In short, recursive feature elimination with cross validation is a pruning method recursively repeated in a cross-validation loop until it converges on a number of important features. We apply this method after fitting a model to filter features of high importance.

Data distribution and partitioning

As mentioned before, the data set consists of approximately 300 comic books published in various parts of the world. The data is divided by continent, country, and decade of publication. These, additionally, serve as target labels for the predictive model and can be found in Figure 3. Figure 3a displays the distribution for continent of publication where we see a slight bias towards Europe. Biased data means the model is easier to learn but less flexible, this causes it to have lower predictive performance on problems with high variance data. For continent of publication this should not cause issues, however, this is not the case with the distribution for decade of publication. Figure 3b indicates a strong bias towards comics published in the 2000s and 2010s. Here, we handle the risk of training a model inclined to predict the 2000s and 2010s more often than other decades. Furthermore, looking at Figure 3c, we notice a bias towards comic books published in the United States and some classes with low occurrence. The risk for decade of publication is valid for country of publication as well. Additionally, the target labels for country of publication consist of some classes with low occurrence, resulting in a risk of underrepresentation by the model. Classes with low occurrence in the dataset have a high probability of being ignored by the model in favor of other classes.

As we are dealing with an unequal distribution of classes, we apply 10-fold cross validation, which is discussed more under "Evaluation". The distribution for country of publication in Figure 3C additionally indicates a count considered too low for Swedish, German, and Spanish comic books. This causes them to be unsuitable to train the model and are, therefore, omitted.

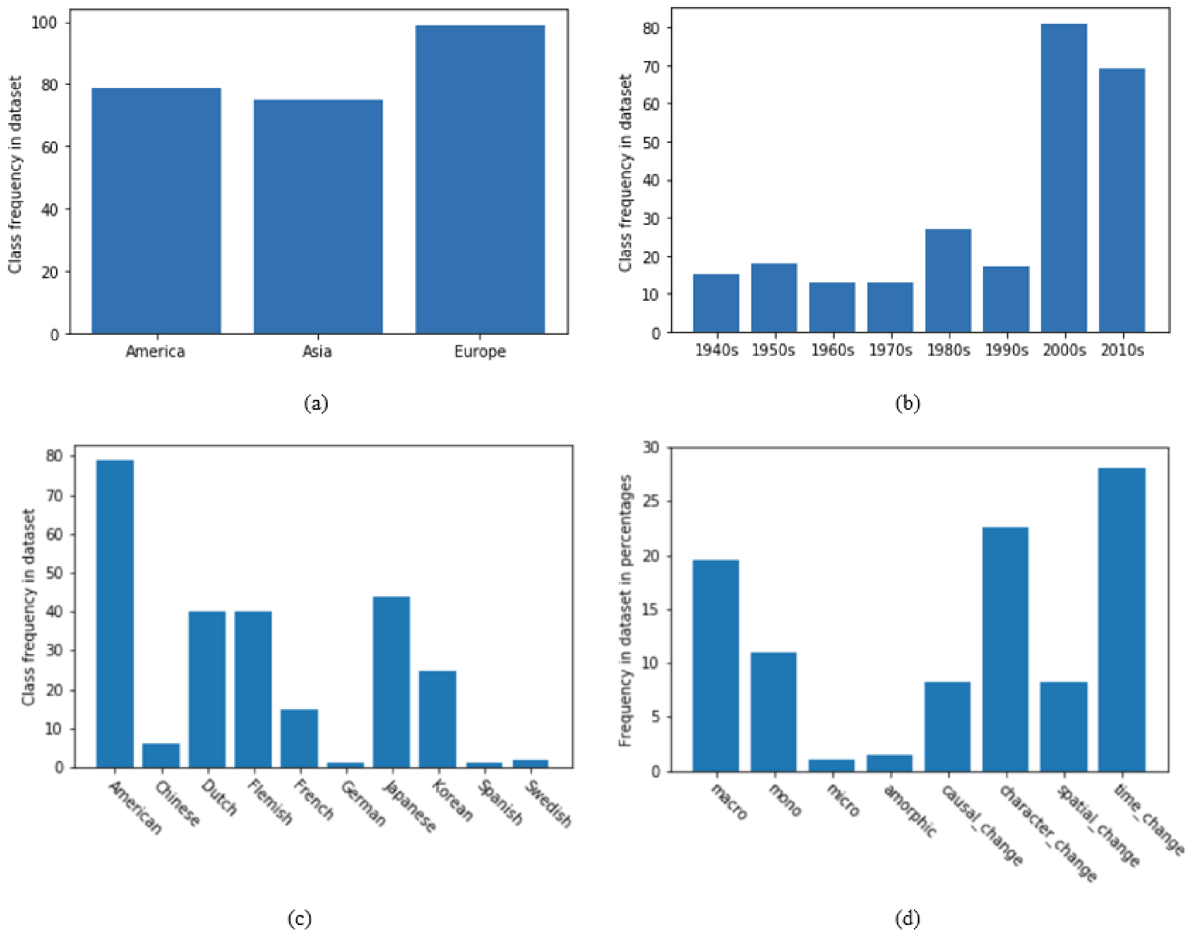

Figure 3. Data distribution. Figure 3a portrays the distribution for continent of publication, Figure 3b illustrates the distribution for decade of publication, Figure 3c demonstrates the distribution for country of publication, and Figure 3d displays the distribution for attentional framing and semantic categories in percentages.

Classifiers

To measure the extent to which we can predict continent, country, and decade of publication, we select a decision tree and naive Bayes classifier. A decision tree classifier is selected to allow the experimenter to interpret the decision making process. A decision tree classifier produces a tree-like graph of decisions and their outcomes. This is an important aspect concerning the importance of features for prediction. Another reason why a decision tree classifier is suitable is the structure of the data. Feature sequences, as explained previously, are ideal for a decision tree classifier given the decision making process. A naive Bayes classifier was selected in addition to a decision tree classifier. A naive Bayes handles features as if they are independent, we can use this to differentiate between decision tree predictions and naive Bayes prediction. The reason to do this, is to validate whether a decision tree classifier is making predictions we can expect from other classifiers, to make sure the combination of patterns does not lead to a bias.

Evaluation

The classifiers used in this study are evaluated by their classification accuracy. The accuracy is both the proportion of true positives and the proportion of false positives. These are examined by plotting true labels and predicted labels in a confusion matrix which permits the inspection of predictions for all labels. Unfortunately, there are no peer studies to compare model performance with. We, therefore, create a baseline by training a majority classifier, and train a decision tree and naive Bayes classifier on raw data without extracting features. The majority classifier always predicts the most frequent occurring class in the dataset. For example, when predicting continent of publication it always predict Europe. Training a decision tree and naive Bayes classifier on raw data supplies us with a baseline based on features and allows us to see if extracting features with the GSP algorithm increases performance.

We already stated that the data is unequally distributed among classes. This increases the risk of overfitting the model due to bias in the data. To avoid overfitting, 10-fold cross validation is implemented to ensure the models generalize well to unseen data. This means the data is partitioned in ten equal sized subsets. The models are trained against nine folds with one subset serving as a validation set. Model parameters are tuned simultaneously by utilizing a grid search. A grid search is provided with a grid of parameter candidates and exhaustively tests the classifier performance. Finally, the average fold accuracy of both classifiers is recorded.

Software

The experimental implementation was executed using python 3.6. We applied the following packages for data cleaning, pre-processing, and data analysis: To read files and handle data we used numpy, pandas, re, os, sys, and glob. To edit excel files and clean them for merging we used openpyxl, xlrd, and pyexcel. The GSP algoritm was adopted from Do Prado Lima (2017) and uses logging, multiprocessing, collections, and itertools. To pre-process data, save the data, and train predictive models we used sci-kit learn and pickle. To plot and visualize the results and decision trees we used Matplotlib and graphviz.

In addition to imported packages, we created support python files. These contained functions that we often used and can be found on GitHub. The programs excel, jupyter notebooks, and graphviz were used to program, visualize and view data. All python files can be found on GitHub (Jansen, 2018).

Experiments and results

To estimate to what extent we can predict continent, country, and decade of publication based on narrative patterns, we evaluate decision trees and naive Bayes classifiers. We additionally cover feature importance and feature distinctions between class labels to answer our question on features paramount for prediction. "Predicting continent of publication" covers the results of the decision tree classifying continent of publication, "Predicting country of publication" describes the outcomes of a decision tree classifying the country of publication, and "Predicting decade of publication" reports the product of the decision tree classifying the decade of publication. Finally, each section considers the performance of a naive Bayes classifier compared to the decision tree and will handle feature importance and distinction between class labels as well.

Predicting continent of publication

To measure the manner in which we can predict continent of publication, we train a decision tree and naive Bayes classifier. Feature vectors previously generated contain 42 features and are subjected to 10-fold cross validation. We then take the average accuracy across ten folds of training and validating to compare performance between the decision tree and naive Bayes classifier. In addition, we fashion a confusion matrix for both classifiers to analyse predictions in depth. The target label in this experiment is the continent of publication.

Table 2 portrays the majority baseline performance, the performance of models trained without feature extraction, and the performance of models with feature extraction and selection. The majority baseline is calculated with the utilization of GSP to extract features but without RFECV to select features. A checkmark for feature extraction indicates GSP is used to extract features. Whereas a checkmark for feature selection specifies RFECV is used to select features deemed most important. Both the decision tree and naive Bayes classifier are trained three times, with a different combination of feature extraction and selection each iteration. The results of all experiments are presented in the same format and can be interpreted in the same way.

Table 2

Average performance per classifier with and without feature selection for continent of publication.

Classifier	Feature extraction	Feature selection	Average accuracy
Majority baseline	✓	-	39.9
Decision tree	-	-	55.2
	✓	-	59.9
	✓	✓	63.1
naive Bayes	-	-	23.7
	✓	-	69.7
	✓	✓	75.6

Note. Here we see average accuracy across 10 folds. Note that feature selection causes an improvement for both classifiers despite the removal of features and that feature selection cannot be executed if feature extraction has not occurred.

Both classifiers outperform the majority baseline and the models trained without applying feature extraction, however, the naive Bayes classifier achieves the best results accuracy wise. This does not mean that the decision tree classifier is performing worse compared to the naive Bayes classifier. For example, when analysing important features, we see that the naive Bayes classifier makes prediction based on frequent occurring patterns. The more frequent a pattern occurs in a class, the more likely the naive Bayes classifier will predict said class. The decision tree, however, takes into account the combination of patterns. As a result, unlikely properties, like the absence of patterns, are utilized as well. The predictions of the decision tree and naive Bayes classifier translate to confusion matrices as seen in Table 3 and Table 4.

The confusion matrix in Table 3 portrays a comparison between decision tree predictions and actual labels. We see an overall better performance for predicting European and American publications and a slightly lower accuracy for Asian publications. However, compared to Asia and Europe, America has a larger amount of incorrect predictions. It appears the decision tree is having trouble making a distinction between Asian and American comic books. This is likely caused due to the influence of Japanese manga on American and western comics (Wong, 2006). When we examine the prediction of the naive Bayes classifier in Table 4 however, we see a shift compared to the decision tree predictions in Table 3.

Table 3

Confusion matrix of decision tree predicting continent of publication.

	Predicted labels		
Actual labels	America	Asia	Europe
America	**0.67**	0.17	0.17
Asia	0.45	**0.50**	0.05
Europe	0.23	0.07	**0.70**

Note. Here we see normalized classification accuracy on a scale of 0-1. Notable is that America and Asia are often missclassified as Europe.

Table 4

Confusion matrix of naive Bayes predicting continent of publication.

	Predicted labels		
Actual labels	America	Asia	Europe
America	**0.42**	0.08	0.50
Asia	0.18	**0.59**	0.23
Europe	0.03	0.03	**0.93**

Note. Here we see the same as Table 3 but for the naive Bayes classifier. Notice the drop in accuracy for Asia and Europe compared to Table 3.

From Table 4 we can conclude that the naive Bayes classifier has difficulties creating a distinction between European and American comic books. As mentioned, naive Bayes considers features to be independent which explains the shift in predictions. A decision tree makes a decision based on a sequence of patterns in this experiment and, thus, retains track of the chronological order of narrative patterns. This order is what gives comic books a unique footprint, making it more difficult to correctly predict the target label. As we will discuss in Feature importance and distinction for continent of publication, we know that the absence of a pattern can be as informative as the presence of a pattern.

In general, the naive Bayes classifier performs better for European and Asian publications but slightly worse for American publications. We additionally see a relative large number of incorrect predictions for Asian publications which again is explainable by the influence of manga on European and American comic books (Wong, 2006).

Feature importance and distinction for continent of publication

In this study features are the equivalence of narrative patterns in comic books. This is why we applied recursive feature extraction with cross validation(RFECV) to select the patterns paramount for prediction. RFECV yielded 25 patterns out of 42 with a reported accuracy of 67 percent to predict continent of publication. Table 5 illustrates a subsection of these patterns, a complete overview of important patterns can be found in Table A3 and Table A4 in the appendix.

Table 5

Section of important patterns and their occurrence for each continent

Narrative pattern	America	Asia	Europe
mono, character_change, mono	0.27	0.64	0.21
macro, character_change, spatial_change, amorphic	0.43	0.45	0.14
macro, causal_change, time_change, macro	0.76	0.45	0.92
macro, character_change, spatial_change, macro	0.58	0.29	0.53
macro, spatial_change, time_change, macro	0.20	0.08	0.84

Note. The data displays pattern occurrence for each class. For example, the first pattern occurs in 27 percent of comics published in America.

Table 5 portrays the normalized occurrence of a narrative pattern per target label. Here we can see that some patterns exist primarily in one class, whereas others are more equally distributed. For example, the last pattern in table 5 can be found in 84 percent of European comic books, and 20 and 8 percent of American and Asian comic books respectively. This indicates that it is primarily used in European countries and can be attributed to a European style of narration. Table 5 additionally exhibits patterns that illustrate near equal distribution across two classes such as the first, second, and fourth pattern. We conclude that these patterns display a “one vs. all” type behaviour where two classes contain near equal occurrence and the third class differs by either displaying a higher or lower appearance. For example, the second pattern of table 5 occurs in 43, 45, and 14 percent of comics published in America, Asia, and Europe respectively. The occurrence for America and Asia is virtually equal and only differs from the occurrence in European comics, hence we call this one occurrence versus all other occurrences or one vs. all. A one vs. all type pattern can only indicate two things. First, that the pattern often occurs within one class and is less frequent or does not occur at all in the leftover classes. Second, that it is less frequent or does not occur at all in one class and is, in addition, equally distributed across the leftover classes.

As mentioned, a pattern with "one vs. all" type behaviour occurs either frequent or infrequent for one of the classes. A higher occurrence indicates a higher importance, but a lower occurrence does not equal a lower importance, as an absence of a pattern excludes class labels as much as a presence includes them. We expect these so called “one vs. all” type patterns to be important in combination with other patterns. The key to prediction, in this case, is the combination of patterns occurring within a comic book to achieve a unique property. As naive Bayes handles features independently it explains the shift in predictions compared to the decision tree. Patterns occurring across all classes carry less meaning for naive Bayes, resulting in less discernible classes.

Country of publication

The goal of this experiment is to measure to what aspect we can predict country of publication and relates to our questions on culture, cultural distinctions, and cultural influence. This happens in the same procedure as the previous experiment. We train a decision tree and naive Bayes classifier, after which we compare performance and create a confusion matrix. The features are the same as used during the first experiment. One difference is that some target labels contained less than three entries in the dataset. To avoid a bias and interference with cross validation, we elected to omit these. The target label in this experiment is the country of publication. The average accuracy across all folds per classifier is summarised in table 6 and indicates that the models outperform the majority classifier and the models trained without extracting features.

As mentioned before, some country classes were omitted, this concerns comic books from Germany, Spain, and Sweden. Table 6 displays a near identical performance for the decision tree and the naive Bayes classifier. However, after applying RFECV we see a change in both models. The decision tree increases in accuracy as opposed to the naive Bayes classifier. This translates to the confusion matrix seen in Table 7.

Table 6

Average accuracy per classifier with and without feature selection for country of publication.

Classifier	Feature extraction	Feature selection	Average accuracy
Majority baseline	✓	-	32.0
Decision tree	-	-	31.5
	✓	-	36.9
	✓	✓	41.2
naive Bayes	-	-	11.8
	✓	-	37.5
	✓	✓	19.2

Note. The average accuracy across 10 folds. Feature selection causes an improvement for the decision tree and a deterioration for naive Bayes.

From an accuracy point of view we see, in Table 7, that the decision tree is struggling to classify between Dutch and Flemish, Flemish and French, Chinese and Japanese, and Japanese and Korean. Coincidentally, these countries are geographically close to each other, supporting a geography based style of narration. The model additionally has issues to classify American comic books, most notably with French, Japanese, and Korean comic books. On the other hand, naive Bayes illustrates the same difficulty classifying between Dutch and Flemish, and French and Japanese, see Table A1 in the appendix. Once again American comic books are classified across the spectrum signalling that American comic books probably contain a mix of patterns common found in non-American comic books.

Table 7

Confusion matrix of decision tree predictions for countries

	Predicted labels						
Actual labels	American	Chinese	Dutch	Flemish	French	Japanese	Korean
American	**0.58**	0.04	0.08	0.12		0.08	0.08
Chinese						1.00	
Dutch			**0.83**	0.17			
Flemish			0.42	**0.42**	0.17		
French	0.50			0.25	**0.25**		
Japanese	0.38		0.08	0.08	0.08	**0.38**	
Korean	0.12	0.12				0.12	**0.62**

Note. The data represents normalized classification accuracy on a scale of 0-1. Noteworthy, are the apparent difficulties in classifying Dutch and Flemish, among others.

Feature importance and distinction for country of publication

Once again we apply RFECV to select important features. This yields 5 patterns out of 42 with a reported accuracy of 49 percent to predict country of publication. Table 8 displays all patterns extracted and exhibits empty cells, indicating that the corresponding pattern did not occur for the designated class label. When we investigate Table 8, we notice that the frequency of all patterns is similar in Dutch and Flemish comic books. This could explain why the decision tree has issues creating a dissimilarity between these labels. Analysing further, we see that both American and French comic books do not contain outlining narrative patterns as often as pattern four, explaining why half of French comic books have been erroneously labelled as American.

The confusion matrix in Table 7 labels Chinese comic books as Japanese. Table 8 clarifies this inaccurate classification when we examine the first and third pattern. It appears that Chinese comic books only contain two out of five selected patterns, incidentally, these patterns occur most frequent in Japanese comic books as well. Additionally noticeable, are the patterns frequently seen in Japanese comic books in comparison to American, Dutch, Flemish, and French comic books. It appears that the patterns most common in Japanese comics are least common in American, Dutch, Flemish and French comic books.

Table 8

All important patterns and their occurrence for each country.

Narrative pattern	American	Chinese	Dutch	Flemish	French	Japanese	Korean
mono, character_change, mono	0.27	1.00	0.13	0.23	0.40	0.77	0.32
micro, character_change, time_change, macro	0.20		0.20	0.33	0.33	0.68	0.28
mono, character_change, spatial_change, amorphic	0.43	0.67				0.82	
macro, character_change, causal_change, time_change, macro	0.77		0.93	0.93	0.73	0.55	0.20
macro, time_change, macro, character_change, time_change, macro	0.18		0.95	0.93	0.60	0.23	

Note. The data displays normalized pattern occurrence for each country. Regard that smaller patterns occur less in Dutch and Flemish comics than larger patterns.

Examining geographical distance and the differences between pattern occurrence, it appears that the larger the distance between countries becomes, the larger the differences between pattern occurrence becomes. This, again, supports a geography based style of narration A final remark is the reason for Korean comic books to be relatively well predicted despite not having frequent patterns. This fact underlines the importance of absence of a pattern because the lack of the third and fifth pattern creates a large enough variance to generally predict the correct label. Based on Table 8 we expect the decision tree to predict Chinese or Japanese, and this is confirmed by the confusion matrix in Table 7.

Decade of publication

Previous experiments revolve around narrative patterns linked to continents or countries. The final experiment involves predicting the decade of publication and is connected to the frequency of patterns over time, which is connected to cultural influence. The goal of the experiment is to investigate the degree to which we can predict decade of publication using the same method as the previous experiments. Once more, we use the same features as the first and second experiment, and the target label in this experiment is the decade of publication. Table 9 presents the average accuracy across all folds per classifier, which we can compare to the models trained without extracting features and the majority baseline.

Table 9

Average accuracy per classifier with and without feature selection for decade of publication.

Classifier	Feature extraction	Feature selection	Average accuracy
Majority baseline	✓	-	31.5
Decision tree	-	-	25.0
	✓	-	28.2
	✓	✓	31.4
naive Bayes	-	-	9.2
	✓	-	22.4
	✓	✓	15.9

Note. Here we see average accuracy across 10 folds. Note that feature selection causes a different effect for both classifiers.

Compared to prior experiments we see a drop in accuracy for both classifiers. The majority baseline outperforms both classifiers, although the decision tree delivers equal performance when GSP(feature extraction) and RFECV(feature selection) are applied. A reason for this is that the cultural variance between countries is larger than the overall cultural variance between decades, making it more difficult to distinguish between decades. After applying RFECV we see the same change as the previous experiment in both models. The decision tree sees a small performance increase, whereas the naive Bayes classifier sees a sharp decrease in performance. This translates to the confusion matrix as seen in Table 10.

Table 10

Confusion matrix of decision tree predictions for decade of publication.

	Predicted labels							
Actual labels	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
1940s	**0.40**	0.20			0.40			
1950s			0.20		0.20	0.20	0.40	
1960s			**0.25**		0.25		0.25	0.25
1970s	0.50		0.25				0.25	
1980s	0.12						0.62	0.25
1990s	0.60			0.20		**0.20**		
2000s	0.08	0.04		0.04	0.04	0.08	**0.46**	0.25
2010s	0.05			0.10	0.05		0.38	**0.43**

Note. Data represents normalized classification accuracy on scale 0-1. The table illustrates more dispersed classifications across the true labels.

Table 10 displays the confusion matrix of the decision tree where we see that it often falsely predicts 2000s and 2010s. The confusion matrix confirms what we mentioned earlier, there is less variance between decades, making it problematic to create a clear distinction between classes. Another reason contributing to false positives is the variance from year to year. The narrative patterns in a comic book published in 1991 should be more similar to narrative patterns of 1992 than narrative patterns of 1999. Changes in narrative patterns over time happen gradually and are, therefore, troublesome to compare. The gradual change results in a larger difference the longer the period between decades where the opposite is true as well. The difference between now and the 1950s is larger than between now and ten years ago. This does not explain why the 2000s and 2010s are overly predicted, to do so we have to examine feature importance.

Feature importance and distinction for decade of publication

Selecting important features is done once more by applying RFECV. This yields 20 patterns out of 42 with a reported accuracy of 34 percent to predict decade of publication. Table 11 illustrates a section of patterns extracted and exhibits empty cells, indicating that the corresponding pattern did not occur for the designated class label. An overview of all patterns can be found in Table A6 and Table A7 in the appendix.

Table 11

A section of important patterns selected by RFECV. The pattern occurrence is portrayed for each decade of publication.

	Decade of publication							
Narrative pattern	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
macro, character_change, macro		0.22	0.31		0.26	0.37	0.51	
mono, character_change, macro					0.22		0.40	0.62
macro, character_change, causal_change, time_change, mono	0.87	0.72	0.77	0.69	0.67	0.76	0.58	0.84
macro, causal_change, time_change, macro, time_change, macro	0.33	0.44	0.77	0.69	0.44	0.53	0.25	0.17
macro, character_change, spatial_change, causal_change, time_change, macro	0.80	0.44	0.62	0.54	0.48	0.29	0.22	0.17

Note. Data is normalized and represents the proportion of comics of corresponding class containing the matching narrative pattern. Shorter patterns tend to occur more frequent later in time as opposed to longer patterns.

From Table 11 we can derive that shorter patterns are not as common in the past seven decades as longer patterns. However, they have been increasing in frequency since the 1980s, whereas longer patterns have been decreasing in frequency since the 1940s. This reinforces the notion that narrative structures have become decompressed over time. What is interesting is that across the majority of important patterns with a length greater than four, we see a decrease in occurrence originating in discrete decades yet always leading up to the 2010s. This signals that narrative structures are always changing over time perhaps following a trend in narrative constitution or due to cultural influence. An example of this is the fourth pattern in Table 11. This pattern recurs at an extended rate until it reaches its peak in the 1960s followed by a steady decline up until the 2010s. Finally, there are patterns maintaining a stable repetition in the past seven decades, such as the third pattern in Table 11. This perhaps means that some patterns are indispensable in narrative structure and cannot be decompressed by substituting with shorter patterns.

Discussion and Conclusion

This paper addresses the statement: To what extend can we predict continent, country, and decade of publication based on narrative patterns in comic books published in the last eight decades? To formulate an answer, we devised four sub questions: Which features are paramount for correct predictions? Which distinctions do we see between comics from various cultures? What patterns can we attribute to culture and cultural influence? And, how does the result tie into the visual language theory developed by Cohn (2011)? We used a dataset containing comic books of the past seven decades published across the world to answer these questions.

Which features are paramount for correct predictions?

This question is answered by looking at the results of applying recursive feature elimination for each experiment. The number of features was narrowed down from 42 to 25, 5, and 20 features for the first, second, and third experiment respectively. Omitting the leftover features resulted in an improved performance for the decision tree classifier in each experiment, confirming both the value of important features and the insignificance of the remainder. The features paramount for predicting labels of experiment one and three can be found in Table A3, A4, A6, and A7 in the appendix. Features of experiment two can be found in Table 8 under Feature importance and distinction for country of publication.

Which distinctions do we see between comics from various cultures?

Culture exists both country wise and continent wise (Featherstone, 1990). Let us, therefore, examine experiment one to discuss distinctions between comics from various continents. Commencing with experiment two, we notice that Flemish and Dutch comics contain the same, near equally frequent, narrative patterns. Comparing these to other countries, we see that shorter patterns are repeated more often in China and Japan, creating a distinction between Dutch and Flemish, and China and Japan. Continuing, we see a distinction between Korean comics and the remainder, due to a combination of absent patterns. If we then compare American and Japanese comics, we see that shorter patterns are more frequent in Japan and some longer patterns are more frequent in America. This creates a distinction between American and Japanese comics. Finally, we examine French comics and notice that these are strenuous to correctly classify as French comic books contain parallels with American, Dutch, and Flemish comics.

Now beginning with experiment one, we see that a majority of smaller patterns occur primarily in Asia; predominantly involving character changes as supported by Cohn et al. (2012), who observed that manga use more shifts between characters than American comics. This supports the inference we made at country level between Japan and America. Resuming our examination of patterns, we see that they are more likely to occur in Europe the longer they are. The opposite is true for America and Asia; here patterns are less likely to occur the longer they are. This contrast creates a distinction between Europe, and America and Asia, supporting our interpretations on country level distinctions. Finally, we have established distinctions between cultures on both country and continent level that support each other. We argue that distinctions on country level support distinctions on continent level and vice versa.

What patterns can we attribute to culture and cultural influence?

Let us compare narrative structure dissimilarity across decades with narrative structure dissimilarity between countries. We can make a clear distinction in narrative style between countries because geographical narrative styles are disjoint, and influence and change each other independently. It is different for narrative style dissimilarity across decades as this changes gradually over time and, therefore, exhibits less variance between classes. This is confirmed by the confusion matrices of the second and third experiment. The decision tree of the third experiment encounters more difficulties narrowing down classes than the decision tree of the second experiment.

As we already mentioned in feature importance and distinction for country of publication, the geographical distance between countries has a larger effect on differences in narrative structure than the temporal distance between decades. What does this mean for culture and cultural influence concerning narrative structure? It means that we see clear distinctions in narrative structure between countries. For example, feature importance and distinction for country of publication iterates the differences between Japanese and American, Dutch, Flemish, and French comic books, where patterns most common in Japanese comic books are least common in other comic books.

It is additionally mentioned that countries geographically close to each other have the tendency to utilize the same patterns but the larger the geographical distance the larger the dissimilarity in pattern occurrence. Does a large distance hinder cultural influence? No, according to (Wong, 2006) this is not the case. The results of the third experiment supports this as well. Experiment three accentuates cultural influence by looking at changes in pattern occurrence over time. The section on feature importance for decade of publication states that short patterns are not as common as longer patterns in the past seven decades, but have been growing in frequency since the 1980s. This reinforces the concept that storytelling has become decompressed over time since mass introduction of manga in America (Brienza, 2016; Goldberg, 2010; McCloud, 1996; Wong, 2006). In addition, the result of the third experiment supports the hypothesis of Cohn et al. (2017), who state that "changes over time might additionally occur to the structure of the narrative sequence, if meaning shifts more to the visuals".

Summarizing the answer to the question is that there is an important difference between patterns we can attribute to culture, and patterns we can attribute to cultural influence. The results of experiment one and two outline patterns we can attribute to culture by looking at their occurrence in continents and countries. Attributing patterns to cultural influence is difficult since this happens gradually over time. This is why we can only attribute patterns to cultural influence if we see a change caused by a trend over a substantial period of time; the results indicate a time period spanning multiple decades. In Table 11 we see patterns that increase or decrease in frequency over time but without a clear cause. We can attribute these to a cultural change but not necessarily to cultural influence without having additional world knowledge. There is one example where we can, the introduction of Japanese manga in America since the 1980s, which is followed by a change in narrative structure pointed out earlier.

How does the result relate to the visual language theory developed by Cohn (2011)?

So far, we have established features important for prediction, distinctions between cultures on country and continent level, and patterns we can attribute to culture and cultural influence. These relate to the visual language theory in the following aspects: first, features important for prediction consist of attentional framing categories and semantic relations and can be represented by a framework introduced by Cohn (2014) involving conjunctions. He expands this framework and posits that it could reveal dissimilarities in structure between cultures, which we demonstrate (Cohn, n.d.). Evidence for this are distinctions between cultures, and patterns we can attribute to culture and cultural influence. Second, patterns we can attribute to cultural influence change over time, confirming a shift towards decompressed, visual storytelling. We noticed that shorter patterns increased in frequency, whereas longer patterns decreased in frequency. These changes over time provide insight into how a visual language develops, like mentioned by Cohn et al. (2017).

Finally, Cohn (2018) presents similar results to this study by comparing the conjunctions mentioned earlier to illustrate cross-cultural variation. The analysis implies that separate narrative grammars are used by separate visual languages and point towards regularities in interactions. These separate narrative grammars and regularities provide the means to create distinctions between cultures as we have done with our experiments. On top of that, the results of Cohn (2018) support the results of our study and vice versa.

Recommendations for future research

The experiments contain some limitations that could be improved in future research. One such aspect is the GSP algorithm used to extract patterns from data. The GSP algorithm restricts the length of patterns because it is computationally heavy.

A consequence could be that some patterns, either paramount for prediction or connected to a culture, are left out. We advise that a different or perhaps adjusted GSP algorithm should be used to avoid this.

Another improvement is the use of a larger dataset containing more publications across the world. We encountered issues for certain classes when applying 10-fold cross validation. They did not contain enough items which caused them to be omitted. The same is true for trying to predict a combination of country and decade of publication. Adding publications from multiple countries in Europe, Asia, and America will aid in discovering variance between cultures, resulting in a larger contrast in cultural differences.

In addition to the limitations, the classifiers suggest significant improvement concerning accuracy. One possible solution could be to extract patterns for each continent with a high threshold for occurrence instead of extracting patterns from the complete dataset with a low threshold.

References

Brienza, C. (2016). *Manga in america: Transnational book publishing and the domestication of japanese comics.* Bloomsbury Publishing.

Bureva, V., Sotirova, E., & Chountas, P. (2015). Generalized net of the process of sequential pattern mining by generalized sequential pattern algorithm (gsp). In *Intelligent systems' 2014* (pp. 831–838). Springer.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Chakraborty, D., & Elzarka, H. (2019). Early detection of faults in hvac systems using an xgboost model with a dynamic threshold. *Energy and Buildings.*

Chu, W.-T., & Li, W.-W. (2017). Manga facenet: Face detection in manga based on deep neural network. In *Proceedings of the 2017 acm on international conference on multimedia retrieval* (pp. 412–415).

Cohn, N. (n.d.). Structural complexity in visual narratives: Theory, brains, and cross-cultural diversity. *Narrative Complexity and Media: Experiential and Cognitive Interfaces. Lincoln: University of Nebraska Press. Google Scholar*.

Cohn, N. (2011). A different kind of cultural frame: An analysis of panels in american comics and japanese manga. *Image & Narrative*, *12*(1), 120–134.

Cohn, N. (2012). Comics, linguistics, and visual language: The past and future of a field. In *Linguistics and the study of comics* (pp. 92–118). Springer.

Cohn, N. (2013). Visual narrative structure. *Cognitive science*, *37*(3), 413–452.

Cohn, N. (2014). The architecture of visual narrative comprehension: the interaction of narrative structure and page layout in understanding comics. *Frontiers in Psychology*, *5*, 680.

Cohn, N. (2015). Narrative conjunction's junction function: The interface of narrative grammar and semantics in sequential images. *Journal of Pragmatics*, *88*, 105-132.

Cohn, N. (2018). 15 visual language theory and the scientific study of comics. *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*, 305.

Cohn, N., Taylor, R., & Pederson, K. (2017). A picture is worth more words over time: Multimodality and narrative structure across eight decades of american superhero comics. *Multimodal Communication*, *6*(1), 19–37.

Cohn, N., Taylor-Weiner, A., & Grossman, S. (2012). Framing attention in japanese and american comics: cross-cultural differences in attentional structure. *Frontiers in psychology*, *3*, 349.

Do Prado Lima, J. (2017). *Gsp.* `https://github.com/jacksonpradolima/gsp`. GitHub.

Featherstone, M. (1990). Global culture: An introduction. *Theory, Culture & Society*, *7*(2-3), 1–14.

Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., & Lam, H. T. (2016). The spmf open-source data mining library version 2. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 36–40).

Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, *1*(1), 54–77.

Goldberg, W. (2010). The manga phenomenon in america. *Manga: An Anthology of Global and Cultural Perspectives*, 281–296.

He, Z., Zhang, S., & Wu, J. (2018). Significance-based discriminative sequential pattern mining. *Expert Systems with Applications*.

Horn, R. E. (1998). Visual language. *MacroVu Inc. Washington*.

Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J. L., Daumé III, H., & Davis, L. S. (2017). The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Cvpr* (pp. 6478–6487).

Jansen, M. (2018). *Vng.* `https://github.com/maartnjansn/VNG`. GitHub.

Langley, P. (2011). The changing science of machine learning. *Machine Learning*, *82*(3), 275–279.

Lasseck, M. (2014). Large-scale identification of birds in audio recordings. In *Clef* *(working notes)* (pp. 643–653).

Li, C., Liu, X., & Wong, T.-T. (2017). Deep extraction of manga structural lines. *ACM Transactions on Graphics (TOG)*, *36*(4), 117.

Llerena, L. A., Hurtado, J. A. B., Ordoñez, R. E. R., Carrillo, F. G. S., Naranjo-Santamaria, J., Barros, T. A., et al. (2018). Identification of human behavior patterns based on the gsp algorithm. In *International conference europe middle east & north africa information systems and technologies to support learning* (pp. 572–579).

McCloud, S. (1996). Understanding manga. *Wizard Magazine*, *56*, 44–48.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J., & Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, *13*(7), e0201264.

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *International conference on extending database technology* (pp. 1–17).

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Association analysis: basic concepts and algorithms. *Introduction to Data mining*, 327–414.

Tversky, B., & Chow, T. (2017, 11). Language and culture in visual narratives. *Cognitive Semiotics*, *10*. doi: 10.1515/cogsem-2017-0008

Visual Language Lab. (2018). `http://www.visuallanguagelab.com/papers.html`. (Accessed: 2018-12-18)

Wong, W. S. (2006). Globalizing manga: from japan to hong kong and beyond. *Mechademia*, *1*(1), 23–45.

Zanettin, F. (2018). Translating comics and graphic novels. *The Routledge handbook of translation and culture*, 445–460.

Appendix

Table A1

Confusion matrix of naive Bayes predictions for country of publication.

	Predicted labels						
Actual	American	Chinese	Dutch	Flemish	French	Japanese	Korean
American	**0.46**		0.08	0.08	0.21	0.04	0.12
Chinese		**0.50**				0.50	
Dutch			**1.0**				
Flemish			1.0				
French			0.25		**0.50**		0.25
Japanese	0.23		0.08		0.38	**0.31**	
Korean	0.12	0.12			0.12		**0.62**

Note. The data represents normalized accuracy of the naive Bayes classifier. Predictions are mostly Dutch and are less accurate than the decision tree classifier.

Table A2

Confusion matrix of naive Bayes predictions for decade of publication.

	Predicted labels							
Actual labels	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
1940s	0.40	0.20		0.20				0.20
1950s	0.40	0.40		0.20				
1960s			0.25	0.50		0.25		
1970s			0.50	0.50				
1980s	0.25	0.50		0.12			0.12	
1990s	0.20			0.80				
2000s	0.04	0.38		0.25		0.04	0.12	0.17
2010s		0.10		0.24			0.24	0.43

Note. The data represents normalized accuracy of the naive Bayes classifier. Predictions are biased towards pre-1980s comics and are less distinct as decision tree predictions.

Table A3

Part one overview of important features for continent of publication.

Narrative pattern	America	Asia	Europe
macro, character_change, mono	0.37	0.6	0.26
mono, character_change, macro	0.3	0.59	0.21
mono, character_change, mono	0.27	0.64	0.21
amorphic, character_change, spatial_change, macro	0.43	0.43	0.11
macro, causal_change, time_change, macro	0.76	0.45	0.92
macro, character_change, spatial_change, amorphic	0.43	0.45	0.14
macro, character_change, spatial_change, macro	0.58	0.29	0.53
macro, character_change, spatial_change, mono	0.59	0.4	0.4
macro, character_change, time_change, macro	0.68	0.61	0.98
macro, spatial_change, time_change, macro	0.2	0.08	0.84
micro, character_change, time_change, macro	0.2	0.52	0.27
mono, causal_change, time_change, mono	0.7	0.49	0.71

Note. Table A3, A4, A5 belong together. Here we see the first part of important features to predict continent of publication, extracted by applying RFECV. Shorter patterns occur frequent in Asia, longer patterns occur frequent in Europe, and America enjoys an equal distribution of patterns. This table contains short and medium length patterns.

Table A4

Part two overview of important features for continent of publication.

Narrative pattern	America	Asia	Europe
mono, character_change, spatial_change, macro	0.62	0.4	0.4
mono, character_change, time_change, mono	0.39	0.72	0.73
mono, spatial_change, time_change, mono	0.33	0.15	0.6
macro, character_change, causal_change, time_change, macro	0.77	0.43	0.89
macro, character_change, spatial_change, time_change, mono	0.47	0.24	0.9
macro, time_change, macro, time_change, macro	0.33	0.23	0.91
mono, character_change, causal_change, time_change, macro	0.78	0.56	0.75
mono, character_change, causal_change, time_change, mono	0.34	0.41	0.33
mono, character_change, spatial_change, time_change, macro	0.52	0.39	0.93

Note. Table A3, A4, A5 belong together. For a short description see Table A3. This table contains both medium length and long patterns.

Table A5

Part three overview of important features for continent of publication.

Narrative pattern	America	Asia	Europe
macro, character_change, spatial_change, causal_change, time_change, macro	0.46	0.07	0.42
macro, character_change, time_change, macro, time_change, macro	0.22	0.19	0.9
macro, time_change, macro, character_change, time_change, mono	0.33	0.31	0.77
macro, time_change, macro, character_change, time_change, macro	0.18	0.17	0.86

Note. Table A3, A4, A5 belong together. For a short description see Table A3. This table contains long patterns.

Table A6

First half of important patterns to predict decade of publication.

Narrative pattern	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
macro, character_change, macro		0.22		0.31	0.26	0.37	0.51	
macro, character_change, mono				0.46	0.22	0.24	0.43	0.61
mono, character_change, macro					0.22		0.4	0.62
mono, time_change, mono	0.53	0.44	0.92	0.77	0.7	0.88	0.69	0.74
amorphic, character_change, spatial_change, macro	0.27	0.22			0.37	0.35	0.36	
macro, character_change, spatial_change, macro	0.73	0.61	0.62	0.69	0.63	0.59	0.32	0.41
macro, character_change, spatial_change, mono	0.67	0.5	0.69	0.62	0.48	0.41	0.36	0.46
mono, character_change, spatial_change, mono	0.4	0.22	0.85	0.54	0.33	0.41	0.4	0.45
mono, character_change, time_change, mono	0.33	0.33	0.85	0.77	0.59	0.59	0.58	0.75
mono, spatial_change, time_change, mono	0.53	0.33	0.54	0.54	0.56	0.53	0.31	0.28
macro, character_change, causal_change, time_change, mono	0.87	0.72	0.77	0.69	0.67	0.76	0.58	0.84
macro, character_change, spatial_change, time_change, macro	0.93	0.67	1,0	0.77	0.81	0.71	0.48	0.58

Note. Table A6 and A7 belong together. Here we see the first half of important features to predict decade of publication, extracted by applying RFECV. We notice that shorter patterns commence to occur more frequent in comics since the 1940s. At the same time we see a decrease in frequency for longer patterns since the 1940s.

Table A7

Second half of important patterns to predict decade of publication.

Narrative pattern	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
macro, character_change, spatial_change, time_change, mono	0.67	0.5	0.77	0.69	0.74	0.82	0.53	0.42
mono, character_change, spatial_change, time_change, macro	0.8	0.61	0.92	0.85	0.67	0.82	0.6	0.51
mono, character_change, spatial_change, time_change, mono	0.6		0.54	0.38	0.48	0.59	0.33	0.36
macro, causal_change, time_change, macro, time_change, macro	0.33	0.44	0.77	0.69	0.44	0.53	0.25	0.17
macro, character_change, spatial_change, causal_change, time_change, macro	0.8	0.44	0.62	0.54	0.48	0.29	0.22	0.17
macro, time_change, macro, causal_change, time_change, macro	0.27	0.33	0.85	0.54	0.44	0.53	0.23	0.28
macro, time_change, macro, character_change, time_change, mono	0.4	0.39	0.85	0.62	0.56	0.71	0.43	0.45
mono, character_change, time_change, macro, time_change, macro	0.4	0.28	0.62	0.77	0.26	0.76	0.46	0.54

Note. Table A6 and A7 belong together. For a short description see Table 6. This table contains the majority of long patterns.