What is a Genre? A Graph Unified Model of Categories, Texts, and Features

José Calvo Tello
Göttingen State and University Library

Introduction and Goal

Several theoretical models have been proposed to understand genres. The Aristotelian scholastic one imposes a taxonomic structure with texts belonging to one single genre, based on necessary and sufficient conditions. In the next level of this taxonomy, each category belongs to a more general one (particular text → war novel → historical novel → novel → fictional prose). This model has been criticized for failing when fitting real examples (Garrido Gallardo, 1988; Hempfer, 2014).

Two alternative models were proposed during the 20th century. First, the family resemblance (Wittgenstein and Schulte, 2013), which expects several shared traits among the members of a category. Second, the prototype theory, which highlights that some instances are better representatives of a category (Rosch, 1973; Rosch, 1975), like the works of Walter Scott for historical novels (Lukács, 1955; Henny-Krahmer et al., 2018).

Computational approaches have shown that genres can be classified to a certain degree using text-internal features (usually a multi-label task). Results vary: Some genres achieve better outcomes than others (Kessler et al., 1997; Stamatatos et al., 2001; Santini, 2011; Allison et al., 2011; Jockers, 2013; Underwood, 2014; Hettinger et al., 2016; Underwood, 2016). However, these theoretical and computational approaches have not been reconciled into a single model until now.

This proposal is the culmination of a series of analysis (Calvo Tello et al., 2015; Schöch et al., 2016a; Schöch et al., 2016b; Calvo Tello et al., 2017; Calvo Tello, 2018a) presenting a theoretical, computational, and visual graph-based model that fits several observations. It proposes a novel unified model for genre with further possibilities for description and interpretation of categories, regardless of periods or languages. Jupyter Notebooks, data and Python scripts are available online,1 for transparency about software, methodology, and parameters.

Data-Sets

Three corpora are used:

1. The Corpus of Novels of the Spanish Silver Age (related to the corpora presented in Schöch et al., 2019), with 358 works, manually enriched with literary phenomena (protagonist, plot, narrator, or ending) and linguistically annotated (grammar, semantics, textual types). Labels were extracted from several sources (manuals of literature, publishers, cover, or National Library, Calvo Tello, 2018b).

2. A collection of 848 classic French theater plays, with labels for subgenres (tragedy, comedy, etc., Fièvre, 2007; Schöch, 2017) and analyzed through topic modeling (see Notebook).

1 https://github.com/cligs/projects2020/tree/master/DH2020_genre_as_graph
3. The books of the Bible and its traditional genres (historical, law, prophetical, letters, etc.), with manual annotation about referred entities and communications (who communicates with whom how, Calvo Tello, 2019).

**Assembling the Graph**

The model is an abstract graph-based representation that can be analyzed and visualized as a network. It needs to fit six observations about genres stated in previous research. First: Texts can be associated with more than one genre (Santini, 2011; Calvo Tello, 2018b). Second: Some instances are better representatives of their category than others (Underwood, 2016; Henny-Krahmer et al., 2018). This can be formalized in a graph with two sets of nodes: Texts (green) which can be connected to any number of genres (purple). The edges are weighted with the proportion of sources relating both:

Third observation: Genres can be classified with internal distinctive features (Todorov, 1976; Garrido Gallardo, 1988; Underwood, 2014; Hettinger et al., 2016). This distinctiveness can be measured in classifiers’ weights, log-likelihood in linguistics, stylometric z-values, or statistical z-scores, here chosen. Next graph illustrates genres (purple) connected to their most distinctive semantic, textual, and literary features (blue):

![Graph of 3 Texts and their Subgenre Labels, Weighted with Proportions of Source](image)
The fourth observation asserts the difficulty of identifying necessary and sufficient conditions of literary genres (Croce, 1902; Chandler, 1997; Schröter, 2019). Genres require more flexible relations, next plotted as three sets of nodes: genres, texts, and features:
Some novels (like *Sonata de otoño*, green, bottom left) share linguistic features (vocabulary related to telephone) with subgenres (adventure) even when they are not labeled as such.

Fifth observation: some genres achieve higher classification results (Underwood, 2014; Hettinger et al., 2016; Calvo Tello, 2018b). This is seen in the graph: Genres with many distinctive features (are more accurately classified, like dialogue novels on the top) are pulled out to the periphery.

The sixth observation declares that some genres share greater similarity (Chandler, 1997; Underwood, 2014; Schöch, 2017). The model reflects this: The number of shared features and texts keep categories closer. War, historical, and adventure novels share more features and therefore, stay closer on the bottom.

Consequently, the model offers information about the extension (number of instances and best representatives), intension (distinctive features), classification results, and similarity of each genre.
Evaluation and Interpretation

The model can accommodate dozens of categories, features, and texts, with hundreds of nodes and edges:

However, digital models should not be interpreted before they are properly evaluated (Da, 2019; Jannidis, 2020). In this case, is the model accurately representing the results of the classification algorithms by their position in the network? As evaluation, the correlation between the centrality of nodes and classification’ scores is observed. The chosen measure is eigenvector centrality, which, in contrast to other centrality measures, maps correctly the perceptual peripheral position of genres like dialogue, episodio nacional, or erotic novel in the plot. The second variable is the F1-scores of the classification (cfr. Notebook and Calvo Tello, 2018b). These two variables show a strong negative correlation (r = -0.65***): The higher the results of the classification, the lower the centrality score in the graph.
Besides, categories treated normally together by scholars (naturalistic-realistic; historical-adventure; memoir-autobiographies) share a larger proportion of features and texts. Consequently, the model allows us to quantify the similarity of two genres and express it as a distance.

This model can be applied to any genre, regardless of period or language. The following plots show the analysis of the Bible and French plays, both with similar statistical correlations between centrality and classification accuracy (cfr. Notebook):

Figure 5: Network of Texts, Topics and Subgenres of French Plays
Conclusions

The proposed graph model fits several observations about genre. It unifies components of the prototype model and the family resemblance theory, offering visually the intension (internal features) and extension (instances and the best representatives) of each category. Besides, it allows two intuitive interpretations: The centrality as classification results, and the distance as similarity through shared features. It even opens further possibilities like community detection of genres or temporal graphs. It has been already applied and evaluated to several languages and periods, showing its potential for explaining the very complex phenomenon of genre.
References


