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 parameteres.

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 mozre 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:

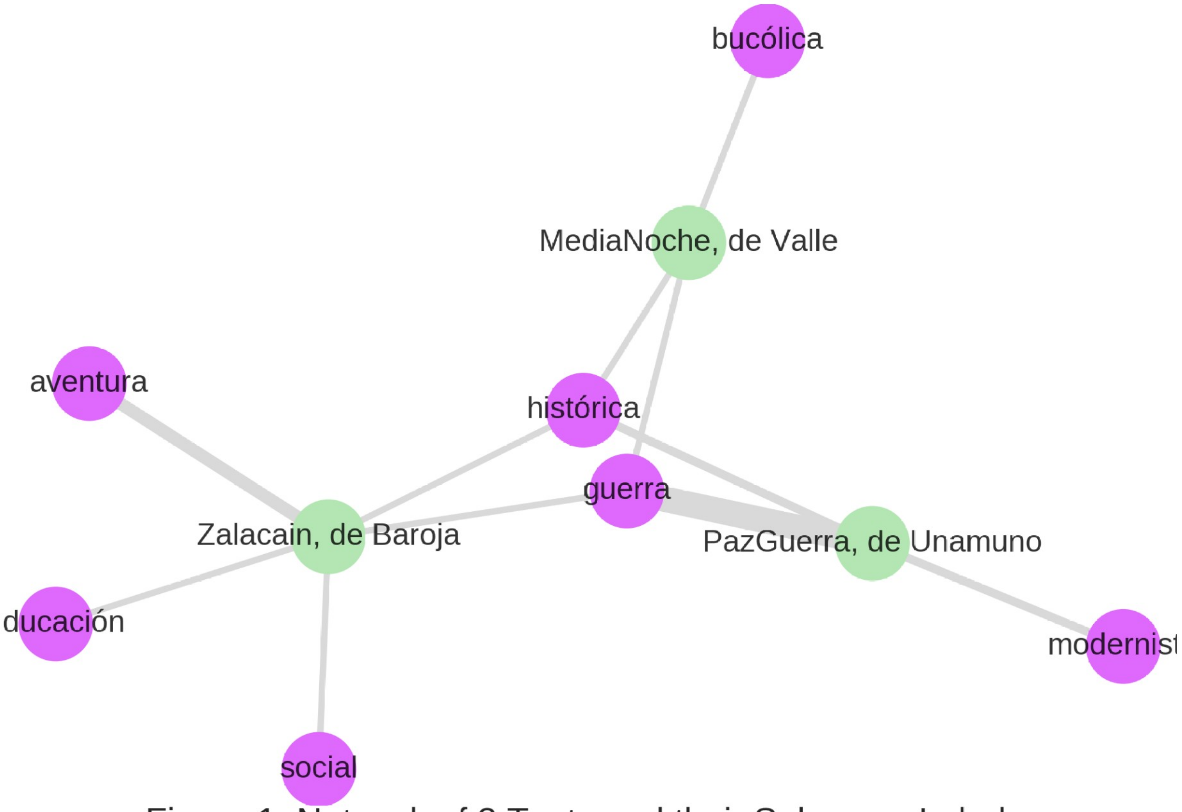


Figure 1: Network of 3 Texts and their Subgenre Labels, Weighted with Proportions of Source

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):

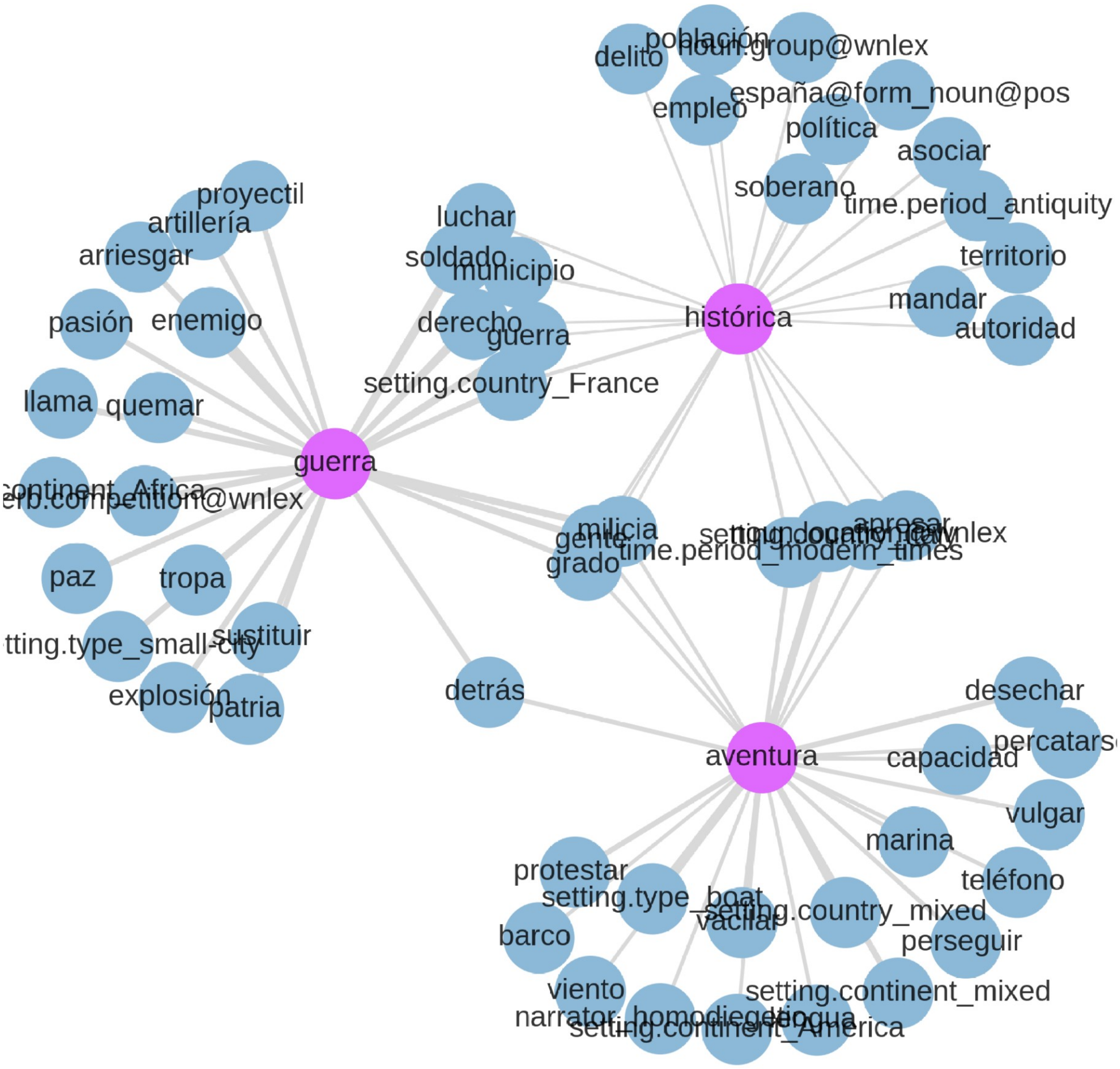


Figure 2: Network of 3 Subgenres and their Distinctive Features Weighted as Z-Scores

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:

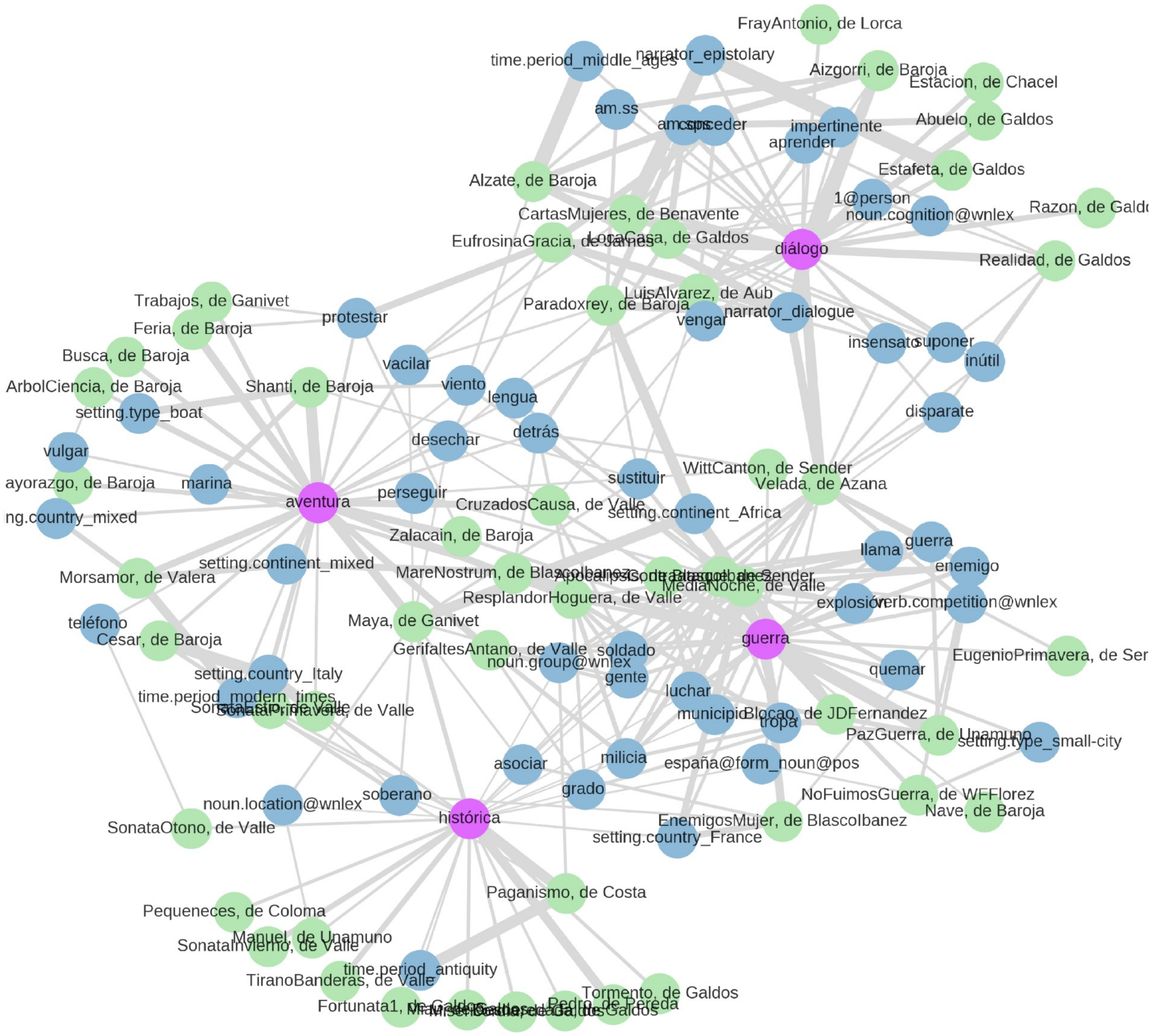


Figure 3: Network of Texts, Features and 4 Subgenres
(adventure, historical, war and greguerías)

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:

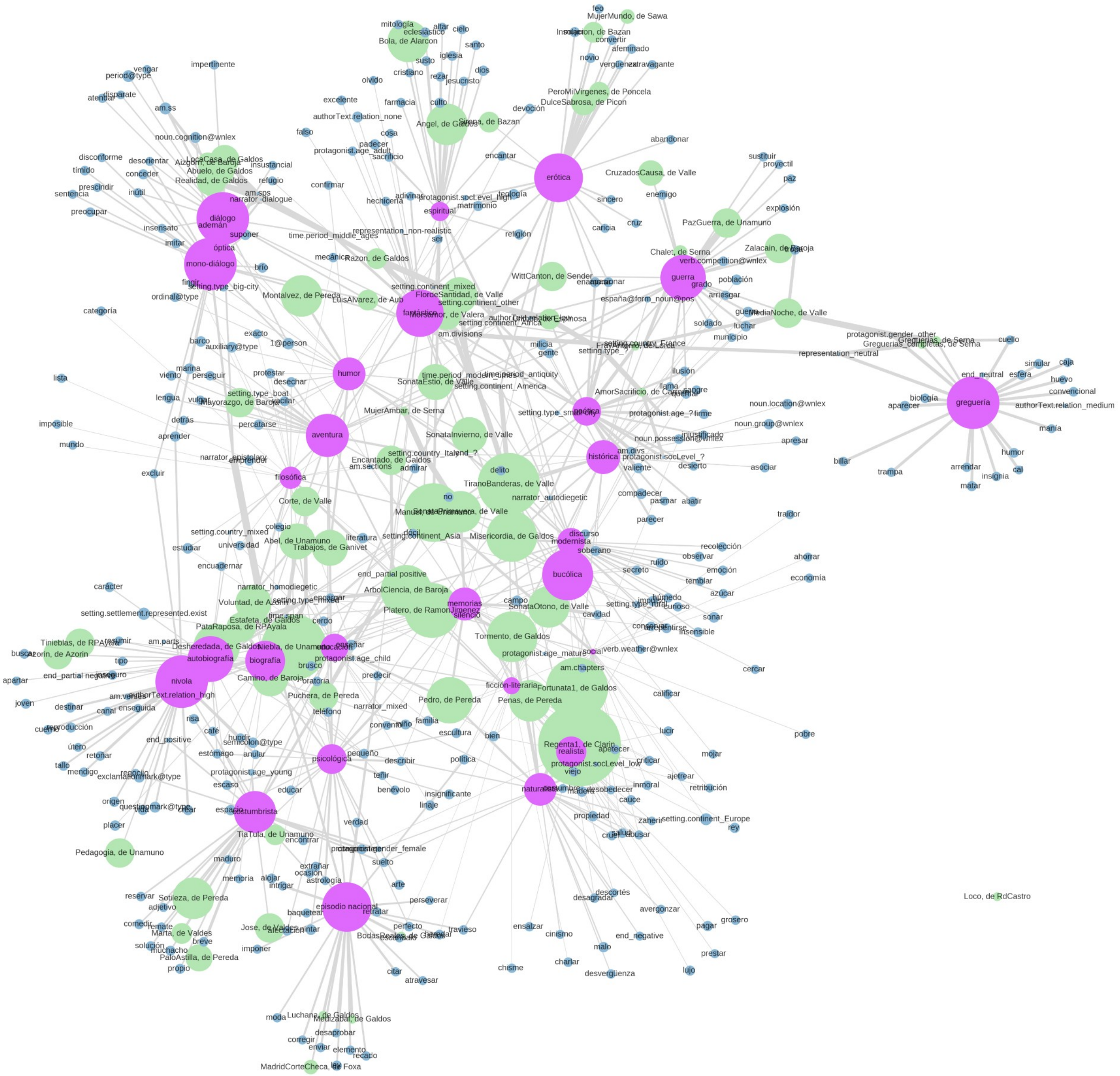


Figure 4: Network of all Subgenres, Texts, and Features of Spanish Novels

However, digital models should not be interpreted before they are properly evaluted (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):

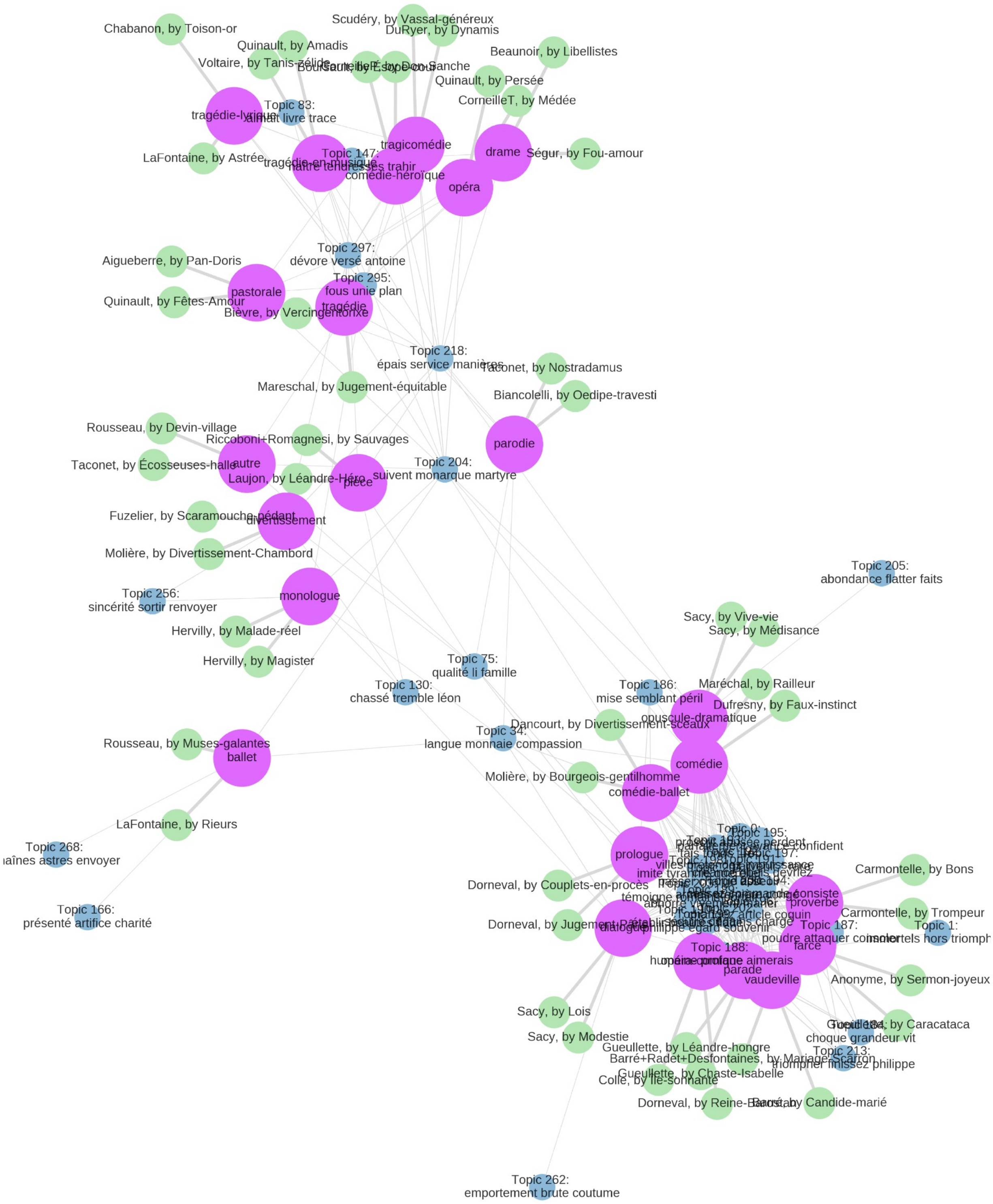
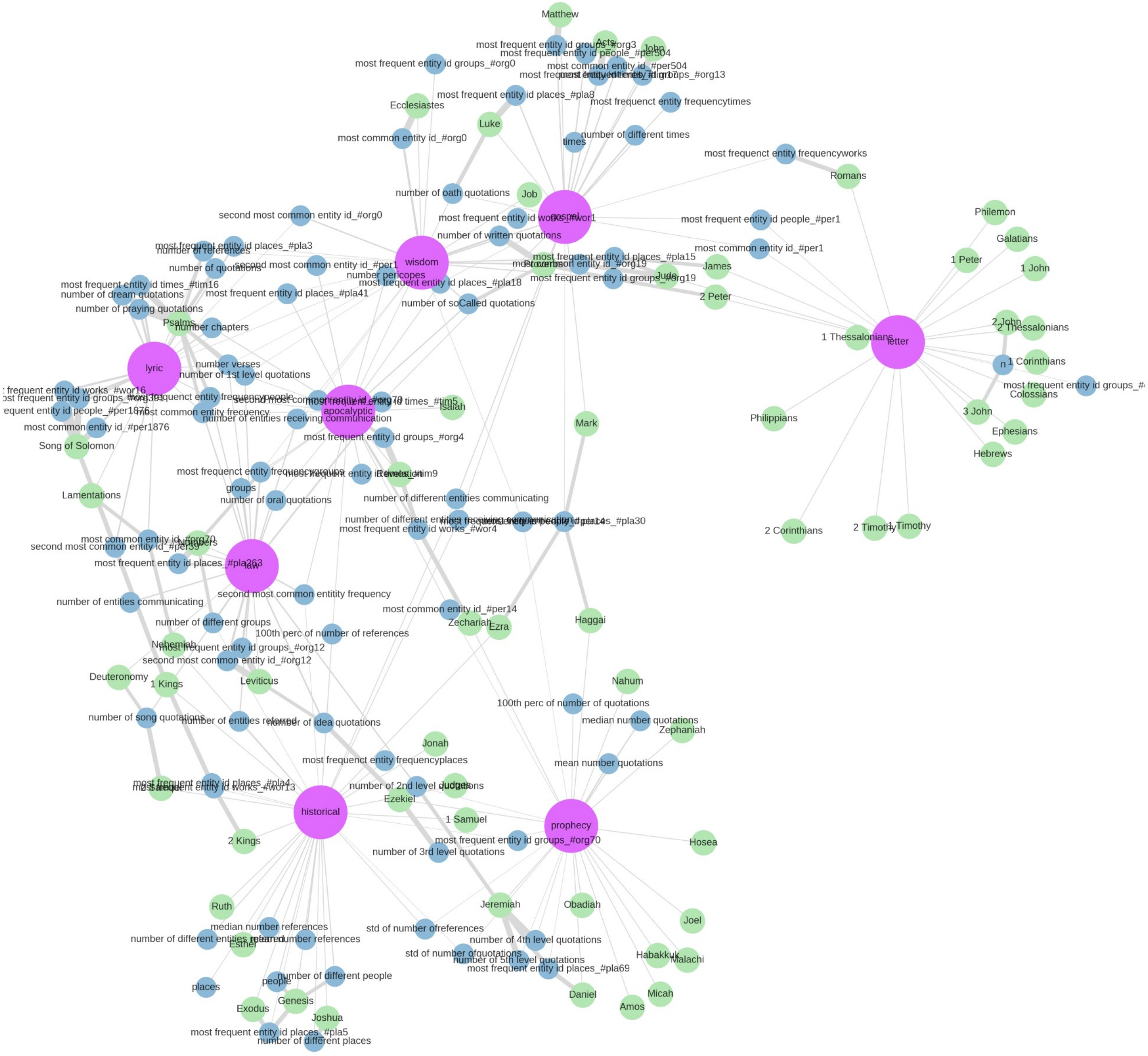

Scudéry, by Vassal-généreux
DuRyer, by Dynamis
Chabanon, by Toison-or
Quinault, by Amadis
Beaunoir, by Libellistes
Voltaire, by Tanis-zélide
Boursault, by Ésope-cour
Quinault, by Persée
CorneilleT, by Médée
Topic 83: aimant livre trace
tragédie-lyrique
tragicomédie
drame
Ségur, by Fou-amour
LaFontaine, by Astrée
Topic 147: naître tendresses trahir
tragédie-en-musique
comédie-heroïque
opéra
Topic 297: dévore versé antoine
Aigueberre, by Pan-Doris
Topic 295: fous unie plan
pastorale
Quinault, by Fêtes-Amour
tragédie
Bièvre, by Vercingentorixe
Topic 218: épais service manières
Taconet, by Nostradamus
Mareschal, by Jugement-équitable
Biancolelli, by Oedipe-travesti
Rousseau, by Devin-village
Riccoboni+Romagnesi, by Sauvages
parodie
Taconet, by Écosseuses-halle
autre
Laujon, by Léandre-Héro
pièce
Topic 204: suivent monarque martyre
Fuzelier, by Scaramouche-pédant
divertissement
Molière, by Divertissement-Chambord
Topic 205: abondance flatter faits
Topic 256: sincérité sortir renvoyer
monologue
Sacy, by Vive-vie
Sacy, by Médisance
Hervilly, by Malade-réel
Hervilly, by Magister
Topic 75: qualité li famille
Topic 130: chassé tremble léon
Topic 186: mise semblant péril
Maréchal, by Railleur
Dufresny, by Faux-instinct
opuscule-dramatique
Dancourt, by Divertissement-sceaux
Topic 34: langue monnaie compassion
Rousseau, by Muses-galantes
ballet
comédie
Molière, by Bourgeois-gentilhomme
comédie-ballet
LaFontaine, by Rieurs
Topic 268: aînes astres envoyer
Topic 0:
Topic 195:
Topic 197:
prologue
Carmontelle, by Bons
Dorneval, by Couplets-en-proces
proverbe
consiste
Carmontelle, by Trompeur
Topic 166: présenté artifice charité
dialogue
Topic 187: poudre attaquer consister
Topic 1: immortels hors triomph
farce
Topic 188: humaine profane aimerais
opéra-comique
parade
vaudeville
Anonyme, by Sermon-joyeux
Sacy, by Lois
Sacy, by Modestie
Gueullette, by Caracataca
Topic 184: choque grandeur vit
Gueullette, by Léandre-hongre
Barré+Radet+Desfontaines, by Mariage-Scarron
Topic 213: triompher finissez philippe
Gueullette, by Chaste-Isabelle
Colle, by Île-sonnante
Dorneval, by Reine-Barocco
Barré, by Candide-marié
Topic 262: emportement brute coutume

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



gospel
wisdom
lyric
apocalyptic
law
historical
prophecy
letter
Matthew
Acts
John
Luke
Ecclesiastes
Job
Romans
Philemon
Galatians
1 Peter
1 John
James
Jude
2 Peter
1 Thessalonians
2 John
2 Thessalonians
1 Corinthians
Colossians
3 John
Ephesians
Hebrews
Philippians
Mark
Psalms
Song of Solomon
Lamentations
Isaiah
Revelation
Numbers
Zechariah
Ezra
Haggai
Nehemiah
Deuteronomy
1 Kings
Leviticus
Nahum
Zephaniah
Jonah
Ezekiel
1 Samuel
2 Kings
Hosea
Ruth
Esther
Jeremiah
Obadiah
Joel
Habakkuk
Malachi
Genesis
Exodus
Joshua
Daniel
Micah
Amos
2 Corinthians
2 Timothy
1 Timothy
most frequent entity id groups_#org3
most frequent entity id people_#per504
most common entity id_#per504
most frequent entity id groups_#org13
most frequent entity id groups_#org0
most frequent entity id places_#pla8
most frequent entity frequencytimes
most common entity id_#org0
number of different times
times
most frequent entity frequencyworks
number of oath quotations
second most common entity id_#org0
most frequent entity id works_#wor1
number of written quotations
most frequent entity id people_#per1
most common entity id_#per1
most frequent entity id places_#pla3
number of references
number of quotations
second most common entity id_#per1
number pericopes
most frequent entity id places_#pla15
most common entity id_#org19
most frequent entity id groups_#org19
most frequent entity id places_#pla18
most frequent entity id times_#tim16
number of dream quotations
most frequent entity id places_#pla41
number of praying quotations
number of socalled quotations
number chapters
number verses
number of 1st level quotations
most frequent entity id groups_#
most frequent entity id works_#wor16
most frequent entity id groups_#org101
most frequent entity id people_#per1876
most common entity id_#per1876
most frequent entity frequencypeople
most common entity frequency
number of entities receiving communication
most frequent entity id times_#tim5
most frequent entity id groups_#org4
most frequent entity frequencygroups
most frequent entity id times_#tim9
groups
number of oral quotations
number of different entities communicating
number of different entities receiving communication
most frequent entity id works_#wor4
most frequent entity id places_#pla30
most common entity id_#org70
second most common entity id_#per39
most frequent entity id places_#pla263
second most common entitity frequency
number of entities communicating
most common entity id_#per14
number of different groups
100th perc of number of references
most frequent entity id groups_#org12
second most common entity id_#org12
100th perc of number of quotations
median number quotations
number of song quotations
number of entities referred
number of idea quotations
mean number quotations
most frequenct entity frequencyplaces
most frequent entity id places_#pla4
most frequent entity id works_#wor13
number of 2nd level quotations
most frequent entity id groups_#org70
number of 3rd level quotations
median number references
number of different entities referred
number references
std of number of references
number of 4th level quotations
std of number of quotations
number of 5th level quotations
most frequent entity id places_#pla69
number of different people
people
places
most frequent entity id places_#pla5
number of different places

Figure 6: Network of Texts, Features and Genres of the Bibel

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

Allison, S., Heuser, R., Jockers, M. L., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment (Stanford Literary Lab, Pamphlet 1)*. Stanford: Standford Literary Lab.

Calvo Tello, J. (2018a). Rules against the Machine: Building Bridges from Text to Metadata. *DH2018*. México DF: ADHO, pp. 550–52.

Calvo Tello, J. (2018b). Genre Classification in Spanish Novels: A Hard Task for Humans and Machines?. *Data in Digital Humanities*. Galway: EADH https://eadh2018.exordo.com/programme/presentation/82.

Calvo Tello, J. (2019). *XML-TEI Bible*. Würzburg: More Than Books https://github.com/morethanbooks/XML-TEI-Bible.

Calvo Tello, J., Schlör, D., Henny-Krahmer, U. and Schöch, C. (2017). Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels. Montréal: ADHO, pp. 181–83 https://dh2017.adho.org/abstracts/037/037.pdf.

Calvo Tello, J., Schöch, C., Rißler-Pipka, N. and Kraft, T. (2015). Humanidades Digitales y estudios hispánicos en Alemania. *Voy y Letra*, **26**(1): 45–61.

Chandler, D. (1997). An Introduction to Genre Theory. http://visual-memory.co.uk/daniel/Documents/intgenre/chandler_genre_theory.pdf.

Croce, B. (1902). *Estetica Come Scienza Dell'espressione e Linguistica Generale*. Milano, Italy.

Da, N. Z. (2019). The Computational Case against Computational Literary Studies. *Critical Inquiry*, **45**(3).

Fièvre, P. (ed). (2007). Théâtre classique. Université Paris-IV Sorbonne http://www.theatre-classique.fr.

Garrido Gallardo, M. A. (1988). Una vasta paráfrasis de Aristóteles. *Teoría de los géneros literarios*. Madrid: Arco/Libros.

Hempfer, K. W. (2014). Some Aspects of a Theory of Genre. In Fludernik, M. and Jacobs, D. (eds), *Linguistics and Literary Studies/Linguistik Und Literaturwissenschaft. Interfaces, Encounters, Transfers/Begegnungen, Interferenzen Und Kooperationen*. Berlin: De Gruyter, pp. 405–22.

Henny-Krahmer, U., Betz, K., Schlör, D. and Hotho, A. (2018). Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane. *Kritik Der Digitalen Vernunft*. Köln, pp. 105–12 http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

Hettinger, L., Reger, I., Jannidis, F. and Hotho, A. (2016). Classification of Literary Subgenres. *Digital Humanities Im Deutschsprachigen Raum Konferenz*. Leipzig: Universität Leipzig, pp. 154–58 http://dhd2016.de/boa.pdf.

Jannidis, F. (2020). On the perceived complexity of literature. A response to Nan Z. Da. *Journal of Cultural Analytics* doi:10.22148/001c.11829. https://culturalanalytics.org/article/11829-on-the-perceived-complexity-of-literature-a-response-to-nan-z-da.

Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Kessler, B., Numberg, G. and Schütze, H. (1997). Automatic detection of text genre. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. (ACL '98). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 32–38 doi:10.3115/976909.979622. http://dx.doi.org/10.3115/976909.979622 (accessed 8 May 2013).

Lukács, G. (1955). *Der Historische Roman*. Berlin: Aufbau-Verlag.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In Moore, T. E. (ed), *Cognitive Development and the Acquisition of Language*. Academic, pp. 111–44.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, **104**(3): 192–233 doi:10.1037/0096-3445.104.3.192.

Santini, M. (2011). *Automatic Identification of Genre in Web Pages: A New Perspective*. Saarbrücken: LAP Lambert Academic Publishing.

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, **11**(2) http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Schöch, C., Calvo Tello, J., Henny-Krahmer, U. and Popp, S. (2019). The CLiGS textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI. *Journal of the Text Encoding Initiative* https://journals.openedition.org/jtei/2085.

Schöch, C., Henny, U., Calvo, J., Schlör, D. and Popp, S. (2016a). Topic, Genre, Text. Topics im Textverlauf von Untergattungen des spanischen und hispanoamerikanischen Romans (1880-1930). Leipzig: nisaba verlag, pp. 235–38 http://dhd2016.de/boa.pdf.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U. and Calvo Tello, J. (2016b). Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University and Paedagocial University, pp. 346–53 http://dh2016.adho.org/abstracts/31.

Schröter, J. (2019). Gattungsgeschichte und ihr Gattungsbegriff am Beispiel der Novellen. *Journal of Literary Theory*, **13**(2).

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, **26**(4): 471–97.

Todorov, T. (1976). The Origin of Genres. *New Literary History*, **8**(1): 159–170.

Underwood, T. (2014). Understanding Genre in a Collection of a Million Volumes, Interim Report. doi:10.6084/m9.figshare.1281251.v1. https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251

Underwood, T. (2016). The Life-Cycle of Genres. *Journal of Cultural Analytics*(1) http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

Wittgenstein, L. and Schulte, J. (2013). *Philosophische Untersuchungen*. 4. [Aufl.]. (Bibliothek Suhrkamp 1372). Frankfurt am Main: Suhrkamp.