Predicting Canonization: Comparing Canonization Scores Based on Text-Extrinsic and -Intrinsic Features

Judith Brottrager[1], Annina Stahl[2] and Arda Arslan[2]

[1]*Technical University of Darmstadt, Institute of Linguistics and Literary Studies, Dolivostraße 15, 64293 Darmstadt*

[2]*ETH Zurich, Social Networks Lab, Weinbergstraße 109, 8092 Zurich*

Abstract

The majority of literary texts ever written are hardly known, read, or studied today, and belong to the so-called "Great Unread". Theories of canonization predominantly focus on sociocultural processes of selection which culminate in the formation of a canon, but say little about how the texts themselves contribute to canonization. In this paper, we propose an operationalization for canonization, which is then used to build a classifier that predicts a canonization score for a text by considering text-intrinsic features only. Working on a historical corpus of English and German texts, which includes both canonical and "unread" works, the results show that a canonization score based on text-inherent features has weak correlations with a canonization score based on text-extrinsic features.

Keywords

literary texts, canonization, text classification

1. Introduction

A major promise of computational literary studies has been the inclusion of the so-called "Great Unread", i.e. those texts that have been previously underrepresented in literary history and which are hardly known, read, or studied today. Large-scale analyses have shown, however, that the argumentative strength of "distant reading" approaches [14] does not lie in including all of what Algee-Hewitt et al. [1] call the "archive", i.e. all published texts preserved in libraries and archives, but in contextualizing the available sample and the population in question [19]. One way of contextualizing the relationship between the "Great Unread"—in other words non-canonical texts—and highly canonical works is to explicitly address their degree of canonization.

The canon as such is not easily definable or even palpable: De- and recanonizations of texts prove that their canonical status is not fixed, but changes over time. Additionally, the criteria for evaluating literature vary between genres [16, 8] and over time [9] and can thus not be universally operationalized. Recent contributions in the field of canon studies stress the complex combination of selection and interpretation processes which are influenced by both literary and non-literary factors [16]. Canonization is seen as the result of an interplay between sociocultural, discursive, and institutional powers, which is only partially understood [22].

CHR 2021: Computational Humanities Research Conference, November 17–19, 2021, Amsterdam, The Netherlands

judith.brottrager@tu-darmstadt.de (J. Brottrager); annina.stahl@gess.ethz.ch (A. Stahl); arda.arslan@gess.ethz.ch (A. Arslan)

0000-0002-3108-8936 (J. Brottrager); 0000-0001-5456-9815 (A. Stahl)

These theories of canonization usually neglect the aspect of literary quality and there is almost no research on the textual aspects of literary judgment [8]. Winko [22] points to the fact that even though the question whether canonical texts share certain textual features that differentiate them from others is hardly addressed, literary scholars often implicitly name certain qualities of texts as the reason for their status in the canon. Similarly, computational literary approaches have often not discussed canonization in direct relationship with textual features, but have either been limited to metadata [1] or have addressed related concepts of distinction in the literary market, as, for example, popularity (see 2).

Our goal is to examine the relationship between a canonization score based on text-extrinsic features, as, for example, available editions and references in secondary literature, and a corresponding score based on text-intrinsic features only. For this, each text in our corpus comprising both canonized and non-canonized texts is assigned a score which reflects the likelihood of the text belonging to the canon, which is also used for the evaluation of our models. If the classification based on text-inherent features is successful, then we can assume that the canonization of texts is, as Winko [22] assumes, to a certain degree linked to textual qualities.

2. Previous Work

As mentioned above, existing quantitative studies either investigate canonization in metadata analyses by modeling indicators of prestige and popularity [20, 1] or by looking at related concepts of distinction. For example, Cranenburgh, Dalen-Oskam and Zundert [6] use textual features to predict the literariness of modern Dutch novels, which was previously determined in a comprehensive survey of readers. They find that a model that combines document embeddings and topic modeling best predicts literariness. Their results further indicate that novels that are perceived as especially literary tend to deviate from the norm by having higher semantic complexity, and that these literary novels use certain words and topics more frequently. Working with the same dataset, Jautze, Cranenburgh and Koolen [10] use topic modeling [5] and show that novels that are perceived as highly literary tend to use a more diverse range of topics. Ashok, Feng and Choi [3] predict literary success, measured by download counts from Project Gutenberg, discovering that some style characteristics only explain literary success for some genres, while others are universal indicators, and find some evidence that readability and literary success are negatively correlated.

3. Corpus

Our corpus comprises $1,153$ novels and narratives in English and German (606 and 547, respectively), covering the Long 18^{th} and the Long 19^{th} century of British and German language literary history. This time span from 1688–1914 avoids culture-specific temporal limits and encompasses great changes in literary production and consumption. We expect this comparative bilingual approach to be especially productive because canonization processes differ significantly between these two literary traditions [9]. During corpus preparation, we systematically adapted an approach proposed by Algee-Hewitt and McGurl [2], which aims at achieving representativeness for literary corpora by moving from a "found corpus" to a "made" list of text, working with best-of and bestseller lists and expert surveys. By doing so, Algee-Hewitt and McGurl [2] combine three different tiers of the literary production: a more exclusive canon, popular and financially successful texts, and a more diverse group of works added at the sug-

Table 1
Odds of a text being canonized per unit

Feature	Odds
Complete/Collected works	2.5795504
Student editions	2.0427817
Exclusive literary histories	1.9767068
Academic literature	1.6519360
Specialized secondary sources	0.5234495

gestion of experts in Postcolonial and Feminist Studies. Analogously, we identified secondary sources, narrative literary histories, anthologies, and more specialised academic monographs that represent these tiers of literary production and used them as bibliographies for our corpus.[1]

4. Canonization Score

In order to be able to take canonization into account, we had to find a reliable operationalization. As neither the canon nor canonization processes are fixed or agreed upon, we have decided to implement a canonization score which reflects the likelihood of a text belonging to the canon. By defining the canonization score as a likelihood, we account for the flexibility of canon formations.

Based on theoretical background provided by Heydebrand and Winko [9], we formalized the following characteristics of a canonized text: A text is more likely to be canonized if (1) there is an edition of the complete or collected works of the author, (2) student editions of the text are available, and if it is mentioned in (3) exclusive narrative literary histories and anthologies and (4) other academic literature.[2] For the computation itself, we made use of the bibliographical information we had gathered during the corpus compilation (for example, the number of times a specific work was mentioned in exclusive literary histories) and additional data taken from the respective national bibliographies and selected publishing houses.

Building on the conceptualization of the canonization score as a likelihood, we then identified minimum and maximum values, i.e. those texts that are extremely unlikely or extremely likely to be considered to be canonized. For the minima, we again used the bibliographic information collected during the corpus compilation to identify those texts that were mentioned in only one highly specialized secondary source. These specialized sources, which deal with literature by marginalised authors and genres, reference texts that are likely to be known and read only by an expert audience, which makes it extremely unlikely for them to be canonized. In contrast to the mass of completely forgotten texts, however, they are at least in some form remembered and available. To represent this difference in the score, we set the minimum score to 0.05. The maxima were defined with the help of university reading lists: the more often a text was mentioned on different reading lists, the higher its score. In our final model, works referenced on more than 60% of reading lists were attributed a score of 1.0, those mentioned on between 30-60% a score of 0.8, and all others with at least one mention a score of 0.6. Following this

[1] As expected, not all texts listed were already digitized and we retro-digitized some of the missing texts ourselves, focusing on adding more diversity to the corpus (by adding not yet represented authors, female authors, authors from geographical peripheries, and niche genres).

[2] Algee-Hewitt et al. [1] similarly formalize different canon notions by relying on entries in the *Dictionary of National Biography*, the MLA Bibliography, and Stanford PhD exam lists.

determination of our training data, we trained a logistic regression model (which again reflects the score's conceptualization as a likelihood) on these texts and their respective canonization features, i.e. the four characteristics mentioned above plus a count representing the number of references to the text in highly specialized secondary sources.

All features used have a significant impact on the discrimination between canonized and non-canonized texts. Table 1 summarizes how a text's odds of being canonized change per unit. For binary characteristics, as, for example, an existing student edition, this means that the odds of being canonized are 2.04 to 1 if a text is available in such an edition. For counts, as for example the mentions in exclusive literary histories, each reference raises the odds by 1.98.

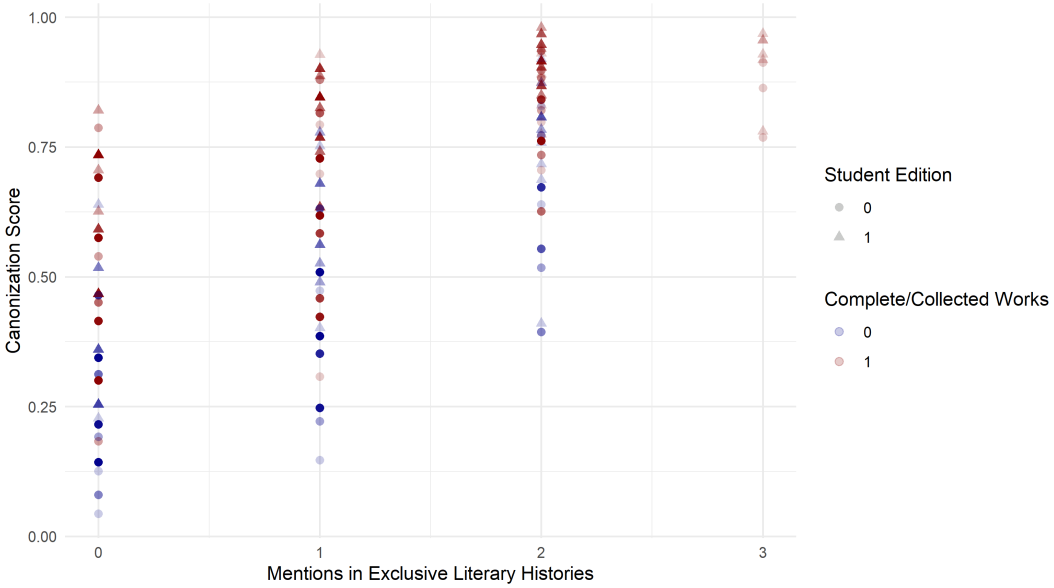


Figure 1: Distribution of canonization scores based on complete/collected works, student editions, and mentions in exclusive literary histories

The resulting model was then used for the prediction of the canonization scores for the entire corpus. Figure 1 shows the scores for all texts and depicts how they are determined by the existence of complete/collected works or student editions and the number of references in exclusive literary histories; the points are transparent so that clusterings and overlaps are identifiable. Overall, the upper end of the scale is dominated by texts by established and well-researched authors (as they are published in complete/collected works), that are also likely to be taught in schools and at universities (as they are published as student editions). The lower range of the scale is dominated by texts which are not part of a narrowly defined literary history (as they are not mentioned in exclusive literary histories) and whose authors are under-studied.

5. Methods

5.1. Approach

Having established a measure of the degree of canonization for each text in our corpus, we used these scores as the ground truth for a model that predicts canonization scores solely from

textual features.

We extracted a range of features from our texts that cover different aspects of style and content. In a preparatory step, we converted all texts to lowercase and replaced German-specific characters. In order to increase the number of data points on which we would train the model, we split the documents into chunks of 200 sentences using spaCy, a library for natural language processing, for sentence tokenization. Features were then calculated for either chunks or full documents, depending on the nature of the feature. In section 5.2, feature extraction is explained in more detail.

Using Support Vector Regession (SVR) as the regression model, we tested several combinations of features and dimensionality reduction techniques for each language separately with a 10-fold cross-validation. All works of an author were part of the same fold in order to avoid overfitting to an author's characteristics. We selected the model with the highest Pearson correlation coefficient (Pearson's r) between the canonization scores and the predicted scores. The p-value of the correlation coefficient was calculated by taking the harmonic mean of the p-values of the folds [21]. We included chunk- and document-level features both separately and in combination, either by adding the document-level features to each chunk, or by taking the average of all chunks per document. For dimensionality reduction, we tried PCA, including enough components so that 95% of the variance was explained, as well as SelectKBest from scikit-learn with either mutual information regression or F-regression as the scoring function, and retained 10% of features.

In addition to running the classification with all texts, we conducted two experiments with the texts that served as training data for the canonization scores. In our first approach, we trained the model on all texts and validated it only on these non-canonized and highly-canonized cases, and in the second approach, we both trained and validated the model on the extreme cases.

5.2. Features

We used a wide range of established features from micro- to macro-textual levels, covering character-based, lexical, and semantic characteristics.[3] Starting on the level of individual characters, the feature set comprises the ratio of various special characters, as, for example, punctuation marks. On the lexical level, we have included the tf-idf[4] of a word if it occurs in at least 10% of documents and is among the 30 words with the highest tf-idf for at least one document, as well as n-gram-based features, such as the 100 most frequent uni-, bi-, and trigrams, and the ratio of all unique uni-, bi-, and trigrams and their entropy. The type-token ratio and the ratio of stopwords are proxies for a chunk's lexical diversity, the Flesch reading ease score [7] for its readability. Additionally, the average word and paragraph length, the text length of a chunk of 200 sentences, and the average and maximum number of words per sentence are used as features. We also created a doc2vec embedding [12] for each chunk, treating chunks as separate documents,

For the modeling of semantic complexity, we implemented four variations of distances between chunks, which were introduced by Cranenburgh, Dalen-Oskam and Zundert [6]. The document vectors obtained by embedding techniques are interpreted as points in a vector space so that the Euclidean distance can be calculated between them. Representing in-text semantic

[3] An overview of all features used can be found in the appendix.

[4] Term frequency - inverse document frequency

similarity,[5] semantic coherence,[6] semantic similarity to other texts,[7] and semantic overlap with other texts,[8] these variations enable a diversified look at text similarities. We calculated all four semantic complexity measures with both doc2vec and Sentence-BERT (SBERT) embeddings, which is a modification of BERT that better captures semantic similarity between sentences [15].

6. Results and Discussion

For both languages, using PCA for dimensionality reduction yielded the highest correlation coefficients in the cross-validation.[9] For the English texts, the combination of text-level features and averages across all chunks performed best with a Pearson's r of 0.242, and chunk-level features delivered the best results on the German texts with a r of 0.285. Table 2 shows the results for using SVR, PCA, and different feature levels, and Figure 2 shows the canonizations scores versus the predicted scores. Limiting the training and/or test data to only non-canonized and highly canonized texts produced correlation coefficients that were similar to those from the full dataset. This can be seen as an indication that the canonization scores inferred for the texts between the extreme cases are reliable.

These weak correlations between the predicted and the actual canonization scores lead to the conclusion that a model of canonization based on text-extrinsic features and a model based on text-intrinsic features are only weakly interconnected. A closer look at some examples shows, however, that some interesting systematic shifts between the models can be observed: While for both the English and the German corpus, texts with the 10% highest canonization scores were on average published during the first half and middle of the 19$^{\text{th}}$ century (1853 and 1834, respectively), texts with the 10% highest predicted scores center around 1873 in the English and 1805 in the German corpus. This can be seen as an indication that what is actually captured is the closeness to central literary periods: Texts written in the Victorian Age dominate the highest predicted scores for the English corpus; texts from the Goethezeit (1770-1830) those for the German corpus.

7. Conclusion and Future Work

Building upon the theoretical framework of canonization theories and previous studies focusing on indicators of literary distinction, our approach offers a quantitative operationalization of

[5] *Intra-textual variance* is calculated by summing over the distances between the document chunks and the centroid, which is obtained by averaging over the chunk vectors. A high intra-textual variance means that the chunks are very semantically dissimilar to each other.

[6] *Stepwise distance* is similar to intra-textual variance, but instead of calculating the distance of chunks from the centroid, the distance between consecutive chunks is calculated, indicating how rapidly a text's semantic content changes.

[7] *Inter-textual variance* is measured with the outlier score, which is the distance of a text, represented as the centroid of its constituting chunks, to its nearest neighbour.

[8] The *overlap score* measures which fraction of the k nearest neighbours of a text's centroid are chunks that belong to that very text, with k being the number of chunks in the text.

[9] In order to allow for an evaluation of feature contribution, we also included SelectKBest from scikit-learn in the cross-validation, which assigns a score to each feature using a scoring function, and then only keeps the k features with the highest score. However, it performed worse than PCA in terms of correlation, so we chose a model with PCA instead.

Table 2
Results (SVR, PCA, various feature combinations)

	Correlation	
	English	German
All documents		
Document	0.184***	0.218***
Chunk	0.164***	0.285***
Document + average of chunks	0.242***	0.230***
Document + all chunks	0.239***	0.191***
Full training and and reduced test data		
Document	0.223**	0.258**
Chunk	0.142**	0.331***
Document + average of chunks	0.267**	0.233**
Document + all chunks	0.155**	0.300**
Reduced training and test data		
Document	0.289***	0.207**
Chunk	0.197***	0.430***
Document + average of chunks	0.347**	0.269***
Document + all chunks	0.182	0.289***

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

canonization and some initial analyses of the relationship between a metadata-based conceptualization of canonization and text-inherent features. Overall, our results indicate that this relationship is very limited. There are, however, some trends on a smaller scale that call for more detailed analyses.

In the next stage of our project, we will focus on those texts whose text-extrinsic and intrinsic canonization scores differ widely. By doing so, we will be able to further investigate the patterns of deviations. This step will also include an evaluation of the implemented features and an analysis of their individual impact on the predictions.

Moreover, dividing the texts into cohorts based on the publication date will help us explore the difference between similarity to other canonized texts and canonization itself, as this would level out the dominance of certain periods in the canon. These cohorts would also allow for a more theory-based description of canonization processes, because, as Heydebrand and Winko [9] have shown, value judgments and evaluative systems are highly adaptive and flexible.

On a methodological level, our approach could be improved by adding features that require more complex language processing, as, for example, Ashok, Feng and Choi [3] have done by including the distribution of part-of-speech tags, syntactic production rules, or sentiments. Finally, as we are working on historical texts from 1688-1914, the language models would have to be trained on or adapted for historical language.

Acknowledgments

This work is part of 'Relating the Unread. Network Models in Literary History', a project supported by the German Research Foundation (DFG) through the priority programme SPP 2207 Computational Literary Studies (CLS). Special thanks to Ulrik Brandes and Thomas Weitin for their feedback and support and to our anonymous reviewers for their invaluable

input and suggestions.

References

[1] M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti and H. Walser. "Canon/Archive. Large-scale Dynamics in the Literary Field". In: *Pamphlets of the Stanford Literary Lab* 11 (2016). URL: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

[2] M. Algee-Hewitt and M. McGurl. "Between Canon and Corpus: Six Perspectives on 20th-Century Novels". In: *Pamphlets of the Stanford Literary Lab*. Pamphlets of the Stanford Literary Lab 8 (2015). URL: http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

[3] V. Ashok, S. Feng and Y. Choi. "Success with style: Using writing style to predict the success of novels". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Ed. by A. for Computational Linguistics. Seattle, Washington, 2013, pp. 1753–1764. URL: https://aclanthology.org/D13-1181.pdf.

[4] C. Bentz, D. Alikaniotis, M. Cysouw and R. Ferrer-i-Cancho. "The Entropy of Words–Learnability and Expressivity across More than 1000 Languages". In: *Entropy* 19.6 (2017), p. 275. DOI: 10.3390/e19060275.

[5] D. M. Blei, A. Y. Ng and M. I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3 (2003), pp. 993–1022. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[6] A. van Cranenburgh, K. van Dalen-Oskam and J. van Zundert. "Vector space explorations of literary language". In: *Language Resources and Evaluation* 53.4 (2019), pp. 625–650. DOI: 10.1007/s10579-018-09442-4.

[7] R. Flesch. "A new readability yardstick". In: *The Journal of Applied Psychology* 32 (3 1948), pp. 221–33.

[8] M. Freise. "Textbezogene Modelle: Ästhetische Qualität als Maßstab der Kanonbildung". In: *Handbuch Kanon und Wertung: Theorien, Instanzen, Geschichte*. Ed. by Rippl, Gabriele and Winko, Simone. Stuttgart, Weimar: J.B.Metzler, 2013, pp. 50–58.

[9] R. von Heydebrand and S. Winko. *Einführung in die Wertung von Literatur*. Paderborn, München, Wien, Zürich: Schöningh, 1996.

[10] K. Jautze, A. van Cranenburgh and C. Koolen. "Topic Modeling Literary Quality". In: *Digital Humanities 2016: Conference abstracts*. Ed. by M. Eder and J. Rybicki. Jagiellonian University & Pedagogical University, Kraków, 2016. URL: https://dh2016.adho.org/abstracts/95.

[11] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov and P. G. Demidov. "A Survey on Stylometric Text Features". In: *25th Conference of Open Innovations Association (FRUCT)*. Helsinki, 2019, pp. 184–195. DOI: 10.23919/fruct48121.2019.8981504.

[12] Q. Le and T. Mikolov. "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31 st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Bejing, China, 2014, pp. 1188–1196. URL: proceedings.mlr.press/v32/le14.pdf.

[13] M. M. Mirończuk and J. Protasiewicz. “A recent overview of the state-of-the-art elements of text classification”. In: *Expert Systems with Applications* 106 (2018), pp. 36–54. DOI: 10.1016/j.eswa.2018.03.058.

[14] F. Moretti. *Distant Reading*. Verso, 2013.

[15] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Ed. by A. for Computational Linguistics. Hong Kong, 2019, pp. 982–3992. DOI: 10.18653/v1/D19-1410.

[16] G. Rippl and S. Winko. “Einleitung”. In: *Handbuch Kanon und Wertung: Theorien, Instanzen, Geschichte*. Ed. by Rippl, Gabriele and Winko, Simone. Stuttgart, Weimar: J.B.Metzler, 2013.

[17] Rippl, Gabriele and Winko, Simone, eds. *Handbuch Kanon und Wertung: Theorien, Instanzen, Geschichte*. Stuttgart, Weimar: J.B.Metzler, 2013.

[18] E. Stamatatos. “A Survey of Modern Authorship Attribution Methods”. In: *Journal of the American Society for Information Science and Technology* 60.3 (2009), pp. 538–556. DOI: 10.1002/asi.21001.

[19] T. Underwood. *Distant Horizons – Digital Evidence and Literary Change*. The University of Chicago Press, 2019.

[20] T. Underwood and J. Sellers. “The Longue Durée of Literary Prestige”. In: *Modern Language Quarterly* 77.3 (2016). DOI: 10.1215/00267929-3570634.

[21] D. J. Wilson. “The harmonic mean p-value for combining dependent tests”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.4 (2019), pp. 1195–1200. DOI: 10.1073/pnas.1814092116.

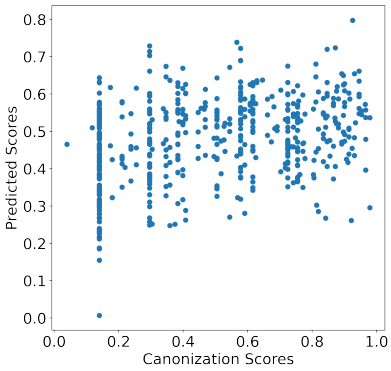
[22] S. Winko. “Literatur-Kanon als invisible hand-Phänomen”. In: *Literarische Kanonbildung*. Ed. by H. L. Arnold. München: edition text + kritik, 2002, pp. 9–24.

A. Overview of Features

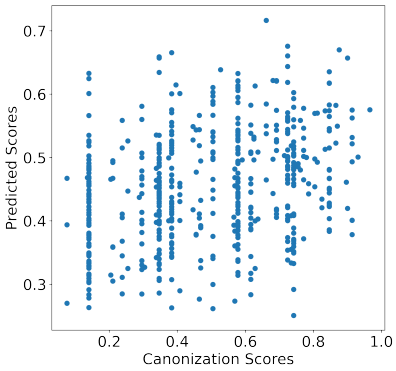
Table 3
Text Features for Prediction

Type	Feature	Source	Text Level
Character	Character frequency	[11]	Chunk
	Ratio of punctuation marks		
	Ratio of whitespace		
	Ratio of digits		
	Ratio of exclamation marks		
	Ratio of question marks		
	Ratio of commas		
	Ratio of uppercase letters		
Lexical	n-grams	[11]	
	Unigrams (100 most frequent)		Document
	Bigrams (100 most frequent)		Document
	Trigrams (100 most frequent)		Document
	Ratio of unique word unigrams		Chunk
	Ratio of unique word bigrams		Chunk
	Ratio of unique word trigrams		Chunk
	Unigram entropy	[4]	Chunk
	Bigram entropy	[4]	Chunk
	Trigram entropy	[4]	Chunk
	Ratio of stopwords		Chunk
	tf-idf	[13]	Document
	Type-token ratio	[1]	Chunk
	Average number of words per sentence	[11]	Chunk
	Max. number of words per sentence		Chunk
	Average word length	[18]	Chunk
	Average paragraph length	[18]	Chunk
	Text length per 200 sentences		Chunk
	Flesch reading ease score	[7]	Chunk
Semantic	Intra-textual variance	[6]	Document
	Stepwise distance	[6]	Document
	Outlier score	[6]	Document
	Overlap score	[6]	Document
Embedding	Doc2Vec	[12]	Chunk

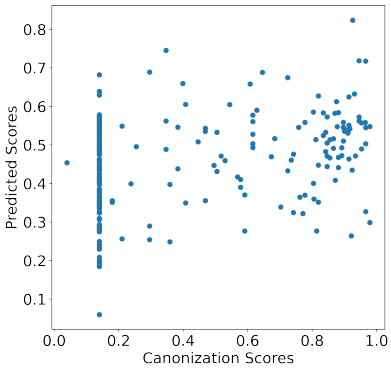
B. Predicted Scores



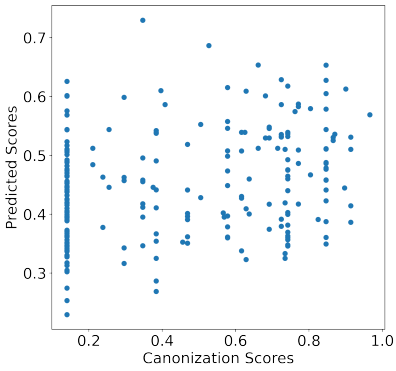
(a) English, all documents, book + average chunk features



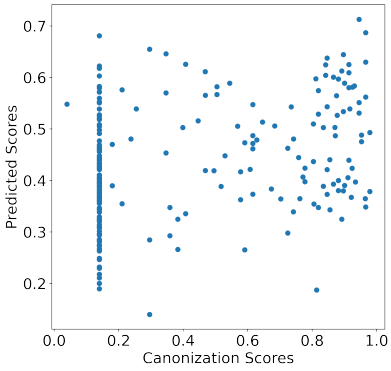
(b) German, all documents, chunk features



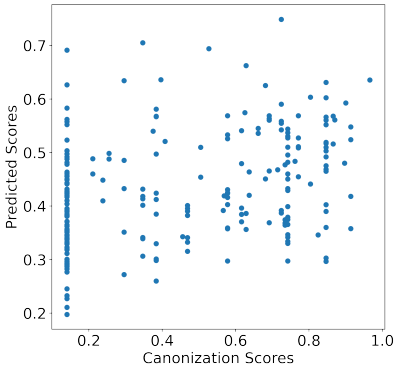
(c) English, full training and reduced test data, book + average chunk features



(d) German, full training and reduced test data, chunk features



(e) English, reduced training and test data, book + average chunk features



(f) German, reduced training and test data, chunk features

Figure 2: Canonization scores vs. predicted scores (SVR, PCA, best feature level for each language)