A Brother Karamazov: Quantifying Morphology in Topic Modeling of Literary Russian

Virginia Partridge
University of Massachusetts Amherst
vcpartridge@umass.edu

Abstract

For topic modeling on English, stemming has not been shown to improve model quality on quantitative measures, but topics with repeated forms of same keyword are undesirable to present to users. However, English has little inflectional morphology, so there may be more motivation to use stemming and lemmatization when training unsupervised topic models on a language with rich inflectional morphology. Using a collection of literary Russian texts, we present ways of measuring the morphological features of topics. We discuss the benefits of post-processing as a way of resolving redundant topic keywords, finding that post-processing can actually obscure grammatical information in topics and should be applied selectively.

1 Introduction

Latent Dirichlet Analysis (LDA) is a widely adopted approach for unsupervised topic modeling and has been used across disciplines for exploring themes and trends in large document collections. LDA has been applied to explore the ever-growing variety of English and Russian text from online platforms and to analyze language changes in academic fields over time (Koltsova and Koltsov, 2013; McFarland et al., 2013; Vogel and Jurafsky, 2012; Mitrofanova, 2015). Assuming a bag-of-words approach, LDA produces latent topics as multinomial distributions over words and each topic is viewed as being generated by a mixture of topics (Blei et al., 2003; Steyvers and Griffiths, 2007).

However, what happens when words in this bag-of-words approach are themselves are complex? We turn to topic modeling on Russian, a flective language with rich paradigms for nouns, adjectives and verbs (Wade et al., 2020). Russian's inflectional morphology increases the sparsity of words' surface forms in the collection, but it's unclear to what extent this sparsity impacts the interpretability and usefulness of topics. Stemming and lemmatization treatments are typical text preprocessing steps for topic modeling, even for English, which has relatively little inflectional morphology, but there is a lack of empirical evidence that these treatments improve the models from the perspective of human interpretability or quantitative measures of topic quality (Schofield and Mimno, 2016).

Furthermore, conflation of surface word forms may mask phenomena of interest to researchers. Topic modeling is a popular tool for exploring gender bias in corpora (Vogel and Jurafsky, 2012; Devinney et al., 2020), and many languages, Russian included, have inflectional morphology that marks gender. By normalizing tokens to a single form, topics learned in LDA won't distinguish between Russian's feminine, masculine and neuter word forms, which may or may not be desirable depending on the domain and researchers' goals. The situation in Russian is even more nuanced than the oft cited English example "apple", the company, as opposed to "apples", the fruit (Schofield and Mimno, 2016), as the different surface forms in Russian do share an underlying lexical word sense, but their variation results from requirements of the language's grammar.

In this work we explore baseline performance of LDA for topic modeling on a Russian literary corpus and report both quantitatively and qualitatively on the resulting topics. We first establish that topic modeling in Russian behaves similarly to English in terms of correlation with corpus metadata, regardless of the stemming or lemmatization approach. Without lexical conflation in Russian, we can imagine two possible extremes: either topic models learn lexeme-specific topics or they learn topics that demonstrate a particular grammatical feature in their morphology. By investigating topic models produced with no morphological preprocessing step, we see evidence for morphological features in topics and propose ways to quantify this relationship. These observations cast doubt on whether lemmatization is beneficial or if it obscures useful information. Finally, we address post-processing of topics' keywords as an alternative to aid with interpretability for end users.

2 Related Work

Probabilistic topic modeling has been applied on Russian text data from academic fields, social media, and Wikipedia articles (Mitrofanova, 2015; Koltsova and Koltsov, 2013; May et al., 2016). Prior to the work on Wikipedia, little attention was given to the role of lemmatization on topic modeling in Russian, and corpora were lemmatized by default. In studying Russian Wikipedia, May et al. (2016) address the impact of lemmatization on topic interpretability via a word intrusion evaluation task, finding that lemmatization may be beneficial. However, they also suggest measuring the effects of lemmatization and do not rule out that lemmatizing in post-processing would also be effective.

The proposal for applying stemming in post-processing comes from work comparing the effects of various stemming approaches on English (Schofield and Mimno, 2016). After comparing the relative strengths, qualitative and quantitative impacts of rule-based and context-based stemmers for English, it was concluded that stemmers do not emprically improve LDA topic models and may even hurt topic stability. Post-processing may still value from the perspective of topic interpretability, avoiding repeating surface forms of the same lexeme in topics' key word lists and presenting users with concise results.

3 Background

3.1 Latent Dirichlet Analysis

LDA uses the observed frequencies of vocabulary terms within documents to infer the *latent*, or hidden, distributions of topics over words and topic assignments for each document. Once a number of topics T is selected, the multinomial distributions $\phi_1, ...\phi_T$ define the distribution of each topic t over the vocabulary terms. Each ϕ_t is drawn from with a Dirichlet prior with concentration parameter β. Each document d also has a multinomial distribution θ_d over the terms in the vocabulary, also drawn from a Dirichlet prior with concentration parameter α. Viewing LDA as a generative process with a joint distribution of the observed and latent variables, find the ϕ_t and θ_d that maximize the likelihood of the corpus if you were to assign tokens to documents using the marginal distributions over topic assignments for the terms in each document. Gibbs Sampling allows estimation of the posterior for the joint topic distribution conditioned on the observed term frequencies by directly assigning topics to each token in the corpus, iteratively sampling topics and updating topic assignments (Steyvers and Griffiths, 2007; Blei et al., 2003; Schofield and Mimno, 2016).

Following Wallach et. al (2002), we will use a symmetric prior for β and an asymmetric prior for α with the MALLET's Gibbs Sampling implementation to train topic models (Wallach et al., 2009; McCallum, 2002). These parameters are optimized every 20 iterations after the first 50, the burn-in period. The Gibbs sampling implementation in MALLET allows us to directly inspect the topic assignments at the level of each token in a document.

3.2 Framework for Morphological Complexity

We will first clarify terms for discussing Russian's morphological paradigms, following frameworks for quantifying morphological complexity used in linguistics and computational linguistics (Baerman et al., 2015b; Cotterell et al., 2019). We draw a distinction between *derivational* morphology, the process by which new words are formed through changing meaning or part of speech, and *inflectional* morphology, which can be simplistically understood as verb paradigms to capture subject-verb agreement or noun declensions for case and grammatical gender. For our purposes here, we are primarily interested in the equivalence classes formed by normalizing inflectional morphology, to use an English example, conflating "respond" and "responds", rather than "respond" and "responsiveness", although aggressive stemming methods will do both types of conflation.

In the word-based morphology framework, inflection is captured by triples consisting of the surface form (also called wordform) w, a lexeme signifying the meaning and a slot σ, which can be understood as a set of "atomic" units of morphological meaning, also called inflectional features (Aronoff, 1976; Sylak-Glassman et al., 2015; Cotterell et al., 2019). A lemma is the surface form used to look up the lexeme in a dictionary, such as the infinitive verb form. Measurements of the size of a lexeme's morphological paradigm capture *enumerative complexity*, the number of distinct surface forms for a particular part-of-speech (Cotterell et al., 2019). A lexeme's mapping between slots and the surface forms is not always straightforward outside the context of a sentence, as multiple slots may be realized with a single surface form (see the example of большой 'big' in table 2). This type of morphological complexity is called *syncretism* and is common in Russian noun and adjective declensions (Baerman et al., 2015a; Milizia, 2015).

There are two morphological tagsets commonly used for natural language processing of Russian. The first originates with Zalizniak's grammatical dictionary and is used in the Russian National Corpus. A detailed explanation of these tags can be found on the website of the Russian National Corpus[1]. The second is the Universal Dependency tagset, which has been applied to the Russian dependency treebank SynTagRus (Sharoff and Nivre, 2011; Lipenkova and Souček, 2014; McDonald et al., 2013). Both of these tagsets express slots as a list of grammatical features and the two tagsets can be mapped to each other, although, as we shall see, their conflation classes (assigned lemmas) for Russian verbs differ significantly.

4 Methods

4.1 Stemmers and Lemmatization Treatments

Following Schofield and Mimno (2016), we distinguish between rule-based stemmers, which are deterministic, but only remove endings and do not map to lemmas, and context-based lemmatizers, which rely on a dictionary of word forms paired with outputs from a part-of-speech tagger to produce lemmas (Schofield and Mimno, 2016; Sharoff and Nivre, 2011). Rule-based methods make no distinction between inflectional and derivational morphological processes, leading to word types, conflation classes of terms, whose original surface forms may cover several lemmas.

Truncation: This simple baseline method trims surface forms to the first n characters (Schofield and Mimno, 2016). We truncate with $n = 5$.

Snowball Stemmer: This stemmer was introduced as a rigorous framework for implementing stemming algorithms for a variety of languages. We utilize the NLTK implementation[2] with the original rules for Russian[3] (Porter, 2001).

Mystem: This Yandex-owned tool is the most popular Russian lemmatizer and can be used without part-of-speech tags. Pairing a finite state machine algorithm for stemming with the Zalizniak grammatical dictionary for morphological tags, this system outputs a list of possible lemmas and slots for a given token input. The system also produces probabilities for each lemma and slot based on word frequency statistics, although the source corpus for these probabilities is not clear (Segalovich, 2003). This is not truly a context-based lemmatizer, as it does not use part-of-speech tags to disambiguate between lemmas or to assign a single slot to a syncretic surface form, but the word frequencies do represent some kind of contextual prior. We use the python wrapper for Mystem, pymystem3[4]. Notably, Mystem is as fast as the Snowball stemmer, while producing a normalized lemma form that is more interpretable for users. All slot and lemma entropy measurments produced in this paper are based on Mystem's morphological analysis for the most likely lexeme of the surface form.

Stanza: This toolkit implements full neural pipelines for processing raw text, including tagging morphological features using bidirectional long short-term memory networks and lemmatizing an ensemble of dictionary based and seq2seq methods (Qi et al., 2020). Because each step of the pipeline depends on the output of the previous step, in order to use Stanza for morphological tagging and lemmatization, we also use its tokenization and sentence-splitting. We use the Stanza model trained on the SynTagRus treebank[5]. Unlike Mystem, Stanza always produces a single lemma and morphological slot, the disambiguation step is included within the model.

Other well-known lemmatizers for Russian include TreeTagger, CSTLemmatiser, pymorphy2 (May et al., 2016; Sharoff and Nivre, 2011; Korobov, 2015). We opted not to use these here due to time constraints, as they are either difficult to install (TreeTagger, CSTLemmatiser) or very slow (pymorphy2).

[1] https://ruscorpora.ru/new/en/corpora-morph.html

[2] https://www.nltk.org/api/nltk.stem.html

[3] http://snowball.tartarus.org/algorithms/russian/stemmer.html

[4] pythonhosted.org/pymystem3/pymystem3.html

[5] https://universaldependencies.org/treebanks/ru_syntagrus/index.html

4.2 Evaluation metrics

When it comes to topic modeling on Russian, we would like to quantify the trade-offs between topic interpretability and loss of information that is linked to a surface form's morphology. We can apply Mystem or Stanza to retrieve the most likely morphological analysis for a surface form w in the vocabulary V to find a lemma ℓ_w and slot σ_w. Using the token-level topic assignments from Gibbs Sampling as our surface form w, we follow Thompson and Mimno (2018) in viewing single topic assignments for each surface form as a data table with columns: surface form w, topic assignment z, slot σ, lemma ℓ. For a given topic k, we obtain the joint count of the slots for the topic $N(\sigma, k)$, the counts of the lemmas for a topic $N(\ell, k)$ and the marginal count variable for a topic $N(k)$. Also note that $\text{argmax}_{w \in V} N(w, k)$ denotes the top key words or surface forms for the topic.

Morphological slot entropy: The goal of this metric is to measure the concentration of slots within a given topic, a proxy for the enumerative complexity of the topic. Does a topic have a concentration of only a few morphological features or does it have a wide spread of the language's inventory of features? This metric is similar to Author Entropy discussed in Thompson and Mimno (2018), where the morphology of the language is the metadata we are attempting to capture, rather than the author of a document (Thompson and Mimno, 2018). Topics that have low slot entropy would contain wordforms with the same grammatical features, for example different verbs conjugated in the first-person singular form or nominative case masculine nouns.

$$\begin{aligned} H(\sigma|k) &= \sum_{\sigma} P(\sigma|k) \log_2 P(\sigma|k) \\ &= \sum_{\sigma} \frac{(N(\sigma, k))}{N(k)} \log_2 \frac{(N(\sigma, k))}{N(k)} \end{aligned} \quad (1)$$

Lemma entropy: Similarly, we may want to know when a topic is dominated by a single lexeme, containing many grammatical forms of a single lexeme, but few other lexemes. For example, a topic may have many counts of different surface forms for each declension of a particular noun, its nominative, accustive, dative, etc... forms or even high counts for a single surface form, but relatively low counts of surface forms for any other lemma. Topics with very low lemma entropy may not be particularly useful to end users, as they reflect lexical and grammatical information known to every speaker of the language, but may not provide specific information about the corpus, other than the presence of a particular lexeme.

$$\begin{aligned} H(\ell|k) &= \sum_{\ell} P(\ell|k) \log_2 P(\ell|k) \\ &= \sum_{\ell} \frac{(N(\ell, k))}{N(k)} \log_2 \frac{(N(\ell, k))}{N(k)} \end{aligned} \quad (2)$$

Ratio of slots to top n key terms: Here we are capturing whether a topic makes a particular grammatical feature obvious when results are displayed to the user. Although this doesn't account for the case when paradigms are syncretic on different dimensions, for example comparing first declension to third declension nouns (Wade et al., 2020), when this value is low, it suggests that the user will notice some sort of grammatical pattern within the topics.

$$R_{\sigma}(k) = \frac{|\{\sigma_w | w \in \{n\,\text{largest} N(w, k)\}\}|}{n} \quad (3)$$

Ratio of lemmas to top n key terms: This metric is targeted at understanding how concise the presentation of a topic's key terms is to a user. When the value is close to 1, each surface form presented to the user represents a unique lexeme. When this value is low, different forms of the same lexeme are repeated, the situation that has motivated the use for lemmatization or stemming to begin with.

$$R_{\ell}(k) = \frac{|\{\ell_w | w \in \{n\,\text{largest} N(w, k)\}\}|}{n} \quad (4)$$

Type-token ratio: Following Schofield and Mimno (2016), this corpus-level metric measures a stemmer or lemmatizer's conflation strength. It is found by taking the ratio of the number of word-type equivalence classes produced by the treatment (the post-treatment vocabulary size $|V|$) to the token counts for the corpus (Schofield and Mimno, 2016).

Character-token ratio: This metric, also from Schofield and Mimno (2016), measures the aggressiveness of stemmers in trimming surface forms to a root form. It measures the average length of the tokens in the corpus after the stemming treatment. Because lemmatizers map surface forms to a normalized lemma instead, this metric isn't as meaningful for lemmatization.

Author entropy: Determining whether the topics correlate with known metadata provides a sanity check on the models. This metric, introduced by Thompson and Mimno (2018) measures how evenly a topic's tokens are spread across authors. We should expect to see both author-specific topic and general topics that are common to many authors (Thompson and Mimno, 2018).

Exclusivity: Exclusivity quantifies the relative uniqueness of the top keywords in a topic. It is high when the terms most frequently generated by a topic are rarely generated by other topics in the model (Bischof and Airoldi, 2012).

Coherence: We additionally report coherence in figure 9. A measure of topic quality that relies on document co-occurence frequencies of word types, coherence has been reported to agree with human evaluations of topic quality (Mimno et al., 2011). However, further work is required to adjust this measurement for the smaller vocabulary sizes resulting from conflation treatments (Schofield and Mimno, 2016). The coherence results reported here cannot be interpreted as evidence for lemmatization helping or hurting the models' quality.

5 Corpus

The selected RussianNovels[6] corpus is a collection of 101 Russian literary works from the 19th and 20th centuries by 23 authors. The collection primarily consists of novels and novellas, but there are some plays (Chekhov and Sologub) and short stories (Gogol) as well. Duplicated works and multiple versions of the same work by different translators were removed. The deduplicated version of the corpus is available on Github [7] with a change log.

Each work was subdivided into passages at least 500 tokens long, where tokens are determined by a simple non-whitespace pattern. This resulted in a corpus of 10,305 documents, broken down by author in figure 1. Next, the corpus was re-tokenized using a regular expression capturing strings of alphabet characters of any length, with punctuation allowed between alphabet characters. Token-level statistics are given in table 1. We produced separate versions of the corpus for each treatment described in section 4.1. After conflation treatments, the resulting terms were pruned to a maximum document frequency of 25% of the corpus and minimum term frequency of 5 occurences in the entire corpus. We chose these pruning settings as a reasonable alternative to manually determining a stopword list, as that process can be challenging and subjective (Schofield et al., 2017).

6 Results

After preprocessing the corpus using each stemmer or lemmatizer, we trained 5 models for each number of topics $T \in \{50, 100, 250, 500\}$. We also trained models on the corpus with no preprocessing treatment at all other than pruning and performed post-lemmatization on these models, using Mystem to produce topics with lemmas as the associated terms rather than surface-forms.

6.1 Qualitative Observations

At the most basic level, we want to check to what extent topic modeling on this Russian literary corpus produces sensible results, showing some topics that correlate with metadata and others that capture general themes not specific to a particular author or time period. Author entropy does not directly measure model quality, in fact users often want topic models not correlated with known metadata (Thompson and Mimno, 2018). However, it can at least reassure us that LDA performs on Russian in a way that is expectedly similar to English. Both the author correlation measures plotted in Appendix A show that stemming and lemmatization do not significantly affect whether author-specific topics or cross-cutting topics are learned, regardless of the number of topics.

Anecdotal analysis of topics' keywords also confirms topics obviously associated with specific authors and coherent general topics with a unified theme like 'school', 'travel', or 'nature'. Some examples are described in detail in appendices D and E. The models trained on untreated data also learn topics that are specific to particular morphological features, which we will revisit in detail in 6.3.

6.2 Lemmatizer Choice Matters

When examining the relative strengths of the stemmers and lemmatizers, laid out in figure 2, functional differences between the two lemmatizers are revealed. As expected, truncation, having the least type-token ratio, is the most aggressive treatment in terms of producing the largest word-type equivalence classes. The surprise is that Mystem is the next most aggressive, followed by Snowball, then Stanza. Since both Mystem and Stanza map sur-

[6] `https://github.com/JoannaBy/RussianNovels`

[7] `https://github.com/ginic/RussianNovels/tree/cleanups`

Stemming/Lemmatization Treatment	Unpruned number of tokens	Unpruned vocabulary size	Number of tokens after pruning	Vocabulry size after pruning	Processing time for treatment (minutes)
No treatment	6081210	319459	3321349	80540	-
Pymystem3	6084073	80163	2976804	32648	4.7
Snowball	6081203	108533	3070870	35938	5
Stanza	6097070	119602	2980649	38435	211.2
Truncate to 5	6081210	59469	3357584	27994	0.25

Table 1: Token level corpus statistics for each lemmatization and stemming treatment. Differences in unpruned token counts are due to Pymystem3 and Stanza using their own tokenization and Snowball normalizing some single characters to empty strings.

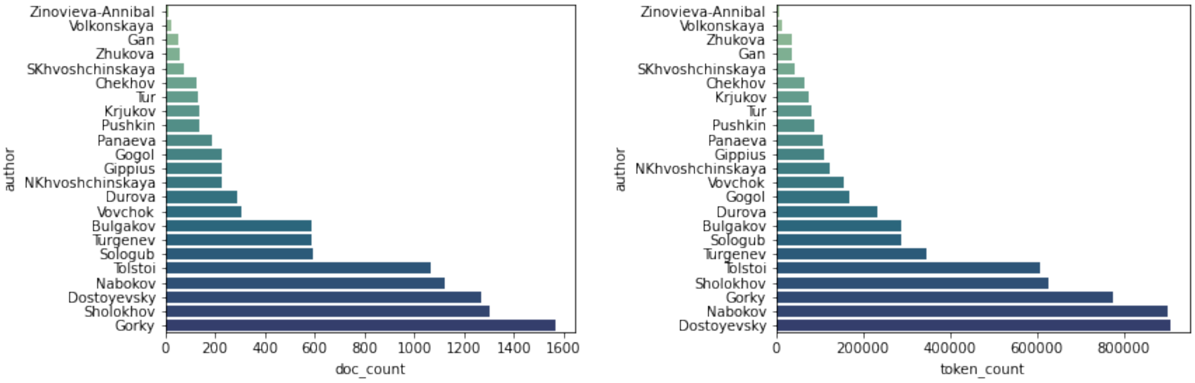

Figure 1: The number of documents per author used to train topic models (left) and the token counts by author before pruning (right).

face forms to a normalized dictionary lemma, we expect them to be roughly equivalent in conflation strength.

This seemingly counter-intuitive result exposes a difference in the way that Mystem and Stanza handle verbal aspect. Nearly all Russian verbs have two infinitive forms, one for the imperfective aspect and one for the perfective aspect (Wade et al., 2020). Mystem treats all conjugations of the both imperfective and perfective aspects as surface forms of the same lexeme, mapping to the imperfective inifitive as the lemma. In contrast, Stanza maps to separate lemmas, imperfective forms to the imperfective infintive and perfective forms to the perfective infinitive. A clarifying example is given in table 2. As a non-native speaker, it's difficult predict what impact this distinction would have on topic interpretability, but the takeaway is that not all lemmatizers are equal. The choice of lemmatizer is not obvious and requires consideration of how the implementation groups the language's grammatical features and which grammatical features matter in the corpus domain.

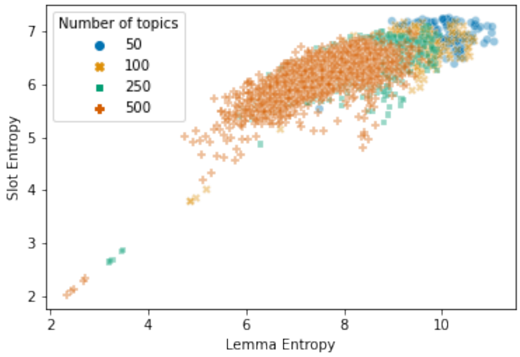

Figure 4: Lemma entropy vs slot entropy for models trained without lemmatization. The seemingly direct relationship between them is due to Mystem expressing syncretism.

6.3 Meaning and Utility of Slot Measures

Another result appearing contradictory at first glance is that slot entropy grows with lemma entropy for a topic. Given the initial assumption that lemmatization is done to remove redundant surface forms within a topic, we might expect that topics with low lemma entropy would have high slot entropy, a situation where topics are represented by many surface forms of a few lexemes.

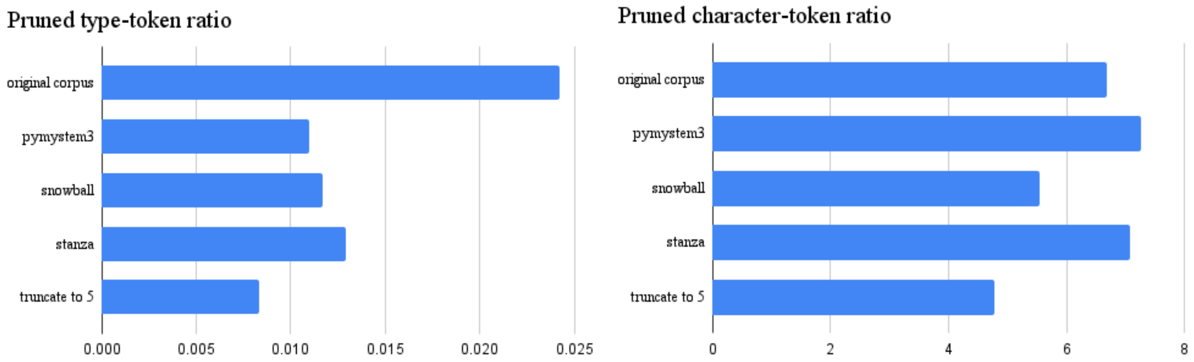

Figure 2: The ratio of word types to tokens for each conflation treatment after the vocabulary is pruned is shown left and demonstrates the strength of conflation method in reducing the vocabulary. Character to token ratio shows lemmatization returns longer normalized strings, while stemming shortens them.

Conversely, when there are many lexemes, high lemma entropy, we might expect some repeated morphological forms, low slot entropy. Contrary to our expectation of an inverse relationship between lemma entropy and slot entropy, figure 4 shows a more direct relationship. This may be partially explained by Mystem's behaviour with respect to syncretism. Recall that Mystem's morphological analysis does not disambiguate in cases where the surface form is syncretic, it simply provides all possible analyses. When a topic covers more lexemes, there are also more opportunities for those lexemes to follow different morphological paradigms and patterns of syncretism. When Mystem is used to analyze topics, the number of slots for a topic will increase as the variety of declension and conjugation types in the topic increases. Because we are not disambiguating the morphological tags based on context, the slot entropy measurement reflects Russian's syncretism in addition to the grammatical features captured for the topic.

Despite the ambiguity around the true morphological features of each tag, the slot entropy and ratio measures produced from the Mystem representation is useful when it is low. Take the examples from a 100 topic model in appendix D, table 3. Topics that consist mostly of non-Russian terms, such as the 'French' topic E, will have extremely low slot entropy and lemma entropy, since every term is assigned as the same *UNKNOWN* slot and *UNKNOWN* lemma, making it easy to identify these kinds of topics which aren't usually interesting to users.

The more interesting case is when slot entropy is low, but not a total outlier. For 100 topic models, the median value for slot entropy is about 6.5, as shown in figure 3. The example topics D and F have some of the lowest slot entropy measures for the model from which they were selected, but have high lemma to keyword ratios, so there are almost no repeated lemmas within each topic. Topic D consists almost entirely of verbs in the past tense feminine form and F of verbs in the second-person singular, including some overlapping lexemes, 'to talk','to ask', 'to sit', 'to look' and 'to begin'. Topic H also has some overlap with these lemmas, but in the past tense masculine form. If these topics were post-stemmed to a normalized form without considering the stem entropy, they would become nearly indistinguishable, a confusing situation to present to users. When stemming is done in preprocessing, it becomes impossible to learn these kinds of morphologically differentiated topics.

Whether users desire topics that reflect grammatical information, as opposed to topics split strictly on word sense relatedness, depends on the use case and domain, but slot entropy, even with ambiguous tags, can help to uncover such topics when there is no preprocessing. Since resources may not be available for a full morphological analysis of a large corpus, or you may only have topic keywords out of context, it's reassuring that slot entropy has meaning despite ambiguity.

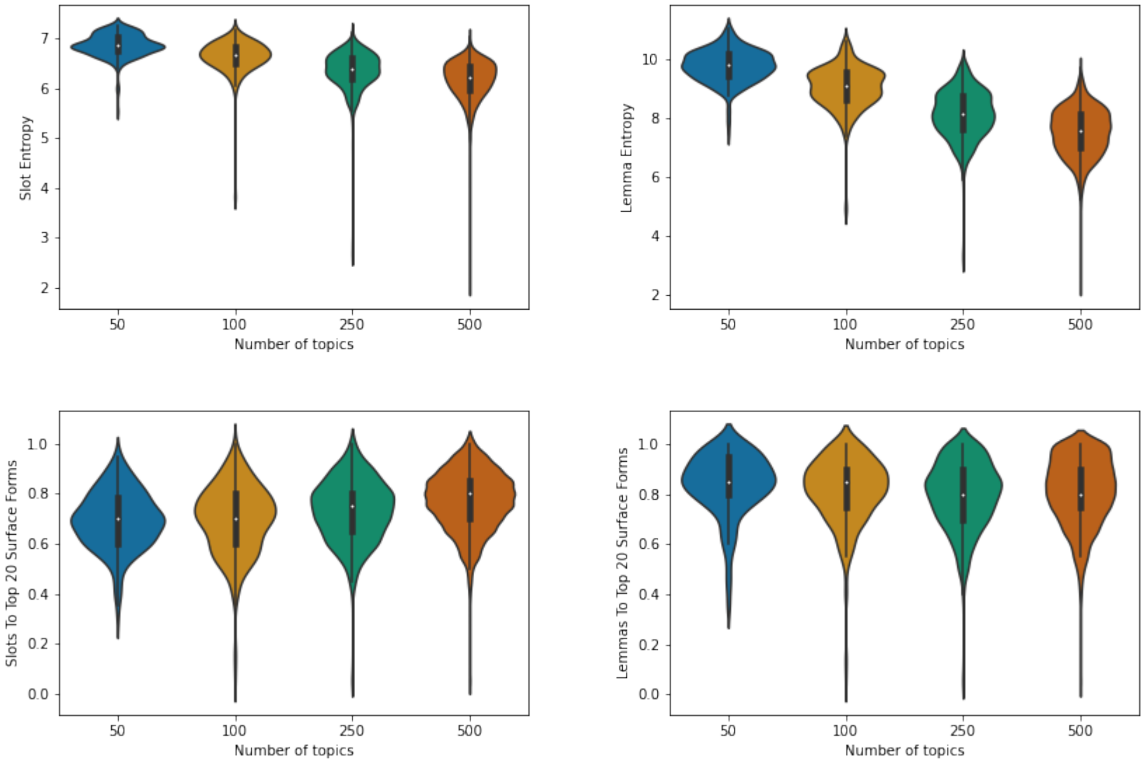

Figure 3: Values of metrics that capture morphological information in topics. These results are for topic models trained on the untreated corpus.

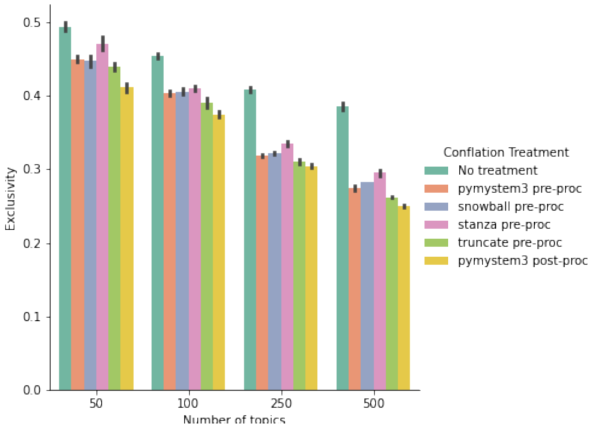

Figure 5: Average exclusivity of topics with over all experiments for each conflation treatment.

6.4 Models Don't Learn Lexeme-Specific Topics

The motivation behind post-stemming is to prevent showing users topics that consist of multiple forms of the same lemma. We are able to identify topics that would benefit from post-stemming by using the topics using the ratio of lemmas to keywords. Take topics I and J in appendix C figure 3, which have the second and third lowest lemma to keyword ratios of the 100 topic model investigated. Topic I has forms of the possessive pronoun мой repeated 12 times, making it difficult to interpret in any way. After post-stemming, it becomes obvious that this topic contains common words and isn't coherent. Topic J is clearly about school and school children, but there are redundant forms of 'boy' and 'school'. By post-lemmatizing, more related words, like 'principal' and 'student', are revealed.

These are clear cases where post-lemmatizing is useful, but the problem doesn't seem to be common. The average lemma to keyword ratio across all numbers of topics is around 0.8, meaning 16 of the 20 keywords are unique. Having such a low proportion of repeated keywords seems tolerable. Additionally, figure 5 demonstrates post-lemmatizing reduces exclusivity significantly. If lexemes were concentrated within the keywords of a single topic, then there shouldn't be such a large reduction in exclusivity. Coupled with the earlier observation that models do learn topics that express morphological features, a blanket application of post-lemmatization seems unwise, although applying it selectively based on some lemma to keyword threshold may be appropriate.

7 Conclusions

Anecdotally, LDA produces topic models that largely appear coherent and reasonable for this Russian literary corpus. Regardless of which conflation treatment is used, and even if none is used, both author-specific topics and thematic topics are found. Through morphological analysis of the terms for each topic and using a measure of morphological slot entropy, we established that models trained on corpus without preprocessing can also capture specific grammatical features. Topics with repeated surface forms of the same lemma are also present, but may not pose as large a challenge to usability as previously believed. Although post-lemmatizing provides clean, concise topic keywords, it obscures any grammatical information encoded in topics, which may lead to more confusion for users. We suggest that post-lemmatizing may be applied selectively to topics with a low ratio of lemmas to top keywords.

8 Future Work

More rigorous empirical comparisons are needed to discern the statistical differences between topic models under different stemming and lemmatization approaches. First, this requires revisiting the vocabulary pruning approach. Although using term and document frequency thresholds avoided the subjectivity of building a stopword list, not having a fixed vocabulary shared between the models over the various conflation methods made it difficult to compare models produced from post-processing, preprocessing, and the original untreated corpus. Fixing the vocabulary will allow us to compare topics empirically using measures of coherence, stability and variation of information.

Having uncovered Mystem's ambiguity with respect to syncretism, we would like to use a context-aware tagger, such as Stanza or TreeTagger, to gain more insight into what grammatical features topics may be capturing. This could also be explored using a corpus with manually annotated morphological tags. OpenCorpora[8] and the Russian National Corpus are good candidates for further exploration. These corpora also contain news and legal subsections, allowing us to ascertain whether our findings hold outside of the literary domain.

[8] opencorpora.org

References

M. Aronoff. 1976. *Word Formation in Generative Grammar*. Linguistic inquiry monographs. MIT Press.

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015a. *Understanding and Measuring Morphological Complexity*. Oxford University Press, USA.

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015b. Understanding and measuring morphological complexity: An introduction. In Matthew Baerman, Dunstan Brown, and Greville G Corbett, editors, *Understanding and Measuring Morphological Complexity*, chapter 1. Oxford University Press, USA.

Jonathan M. Bischof and Edoardo M. Airoldi. 2012. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 9–16, Madison, WI, USA. Omnipress.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.

Olessia Koltsova and Sergei Koltsov. 2013. Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet*, 5.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Janna Lipenkova and Milan Souček. 2014. Converting Russian dependency treebank to Stanford typed dependencies representation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 143–147, Gothenburg, Sweden. Association for Computational Linguistics.

Chandler May, Ryan Cotterell, and Benjamin Van Durme. 2016. An analysis of lemmatization on topic models of morphologically rich language.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel A McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D Manning, and Daniel Jurafsky. 2013. Differentiating language usage through topic models. *Poetics*, 41(6):607–625.

Paolo Milizia. 2015. Patterns of syncretism and paradigm complexity: The case of old and middle indic declension. In Matthew Baerman, Dunstan Brown, and Greville G Corbett, editors, *Understanding and Measuring Morphological Complexity*, chapter 8. Oxford University Press, USA.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Olga Mitrofanova. 2015. Probabilistic topic modeling of the russian text corpus on musicology. In *International Workshop on Language, Music, and Computing*, pages 69–76. Springer.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to Apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Serge Sharoff and Joakim Nivre. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. pages 657–670, Moscow, Russia. Dialogue: Computational Linguistics and Intellectual Technologies.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

T. Wade, D. Gillespie, S. Gural, and M. Korneeva. 2020. *A Comprehensive Russian Grammar*. Blackwell Reference Grammars. Wiley.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981.

A Author Correlation Metrics with Conflation Treatments

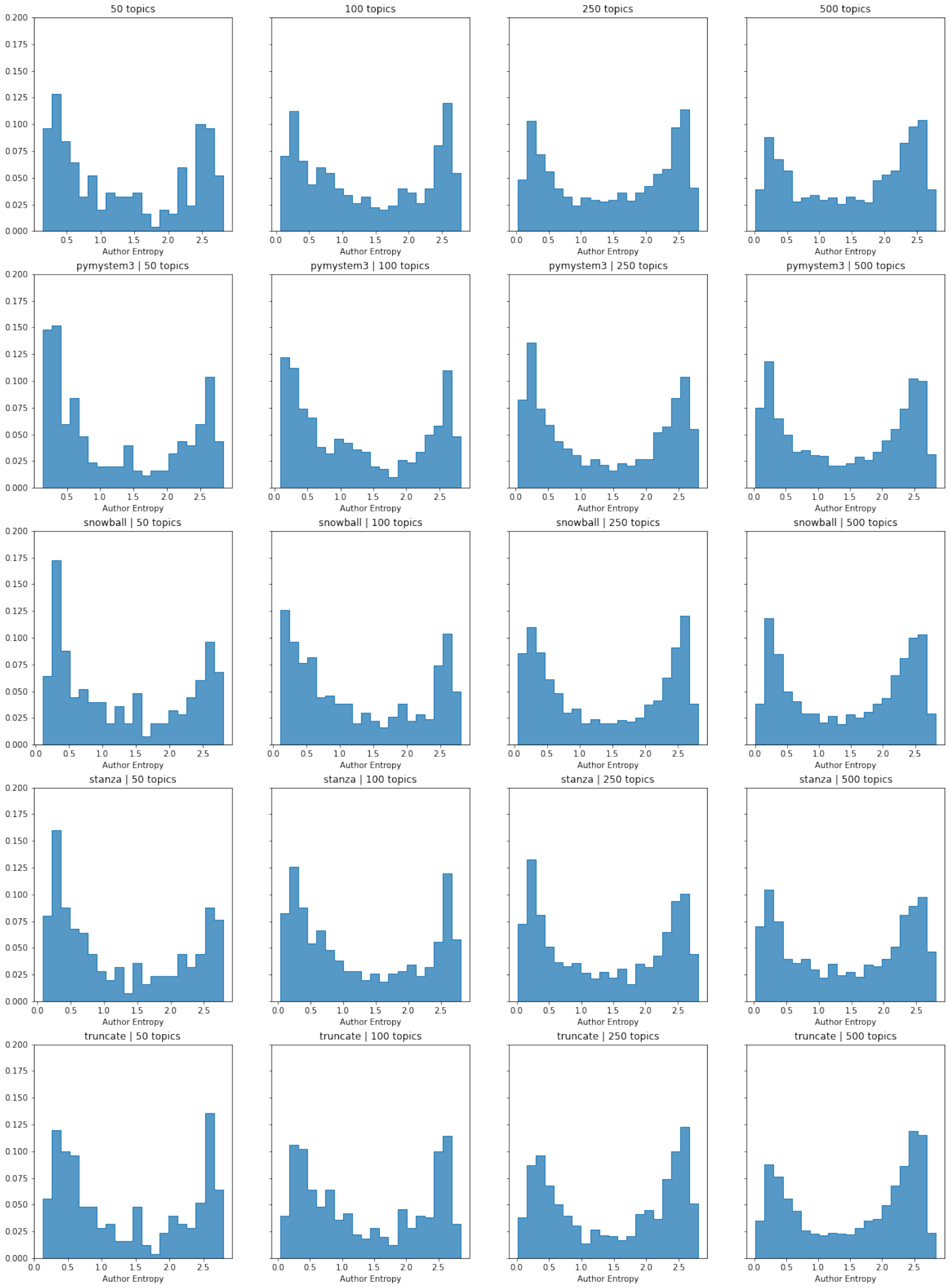

Figure 6: Author Entropy metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of conflation treatment used.

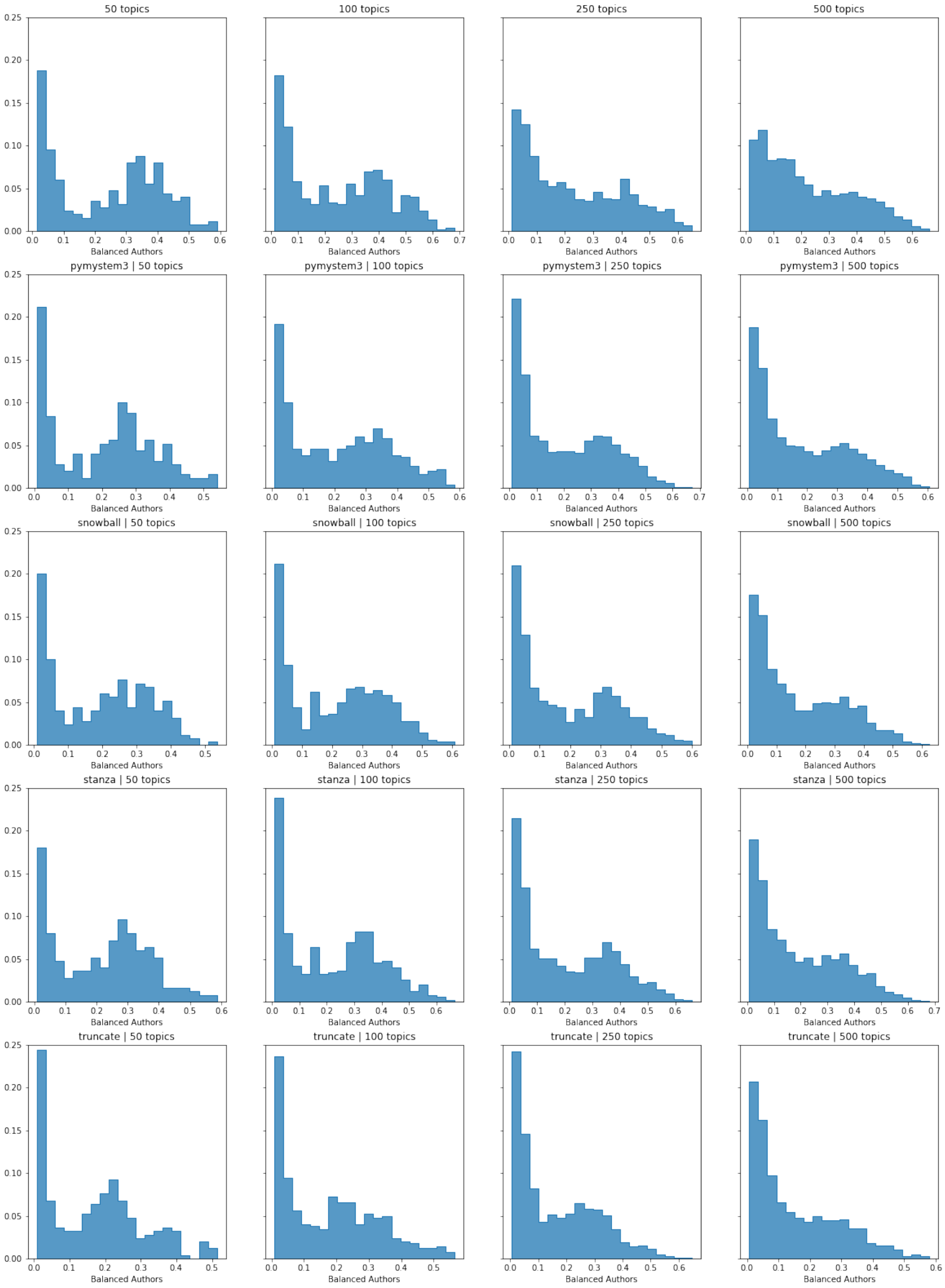

Figure 7: Balanced Author metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of conflation treatment used.

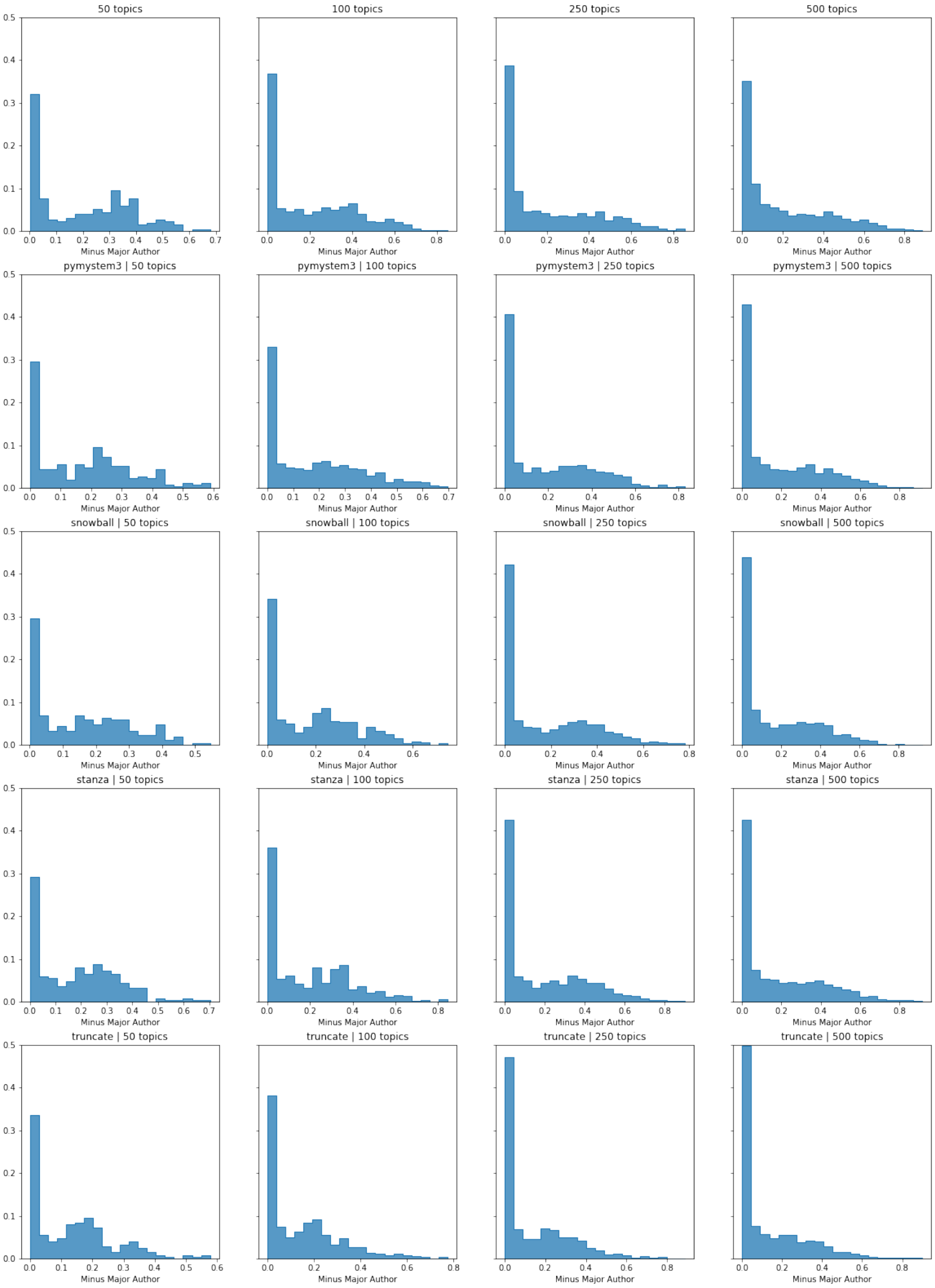

Figure 8: Minus Major Author metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of metrics conflation treatment used.

B MALLET Diagnostics Metrics

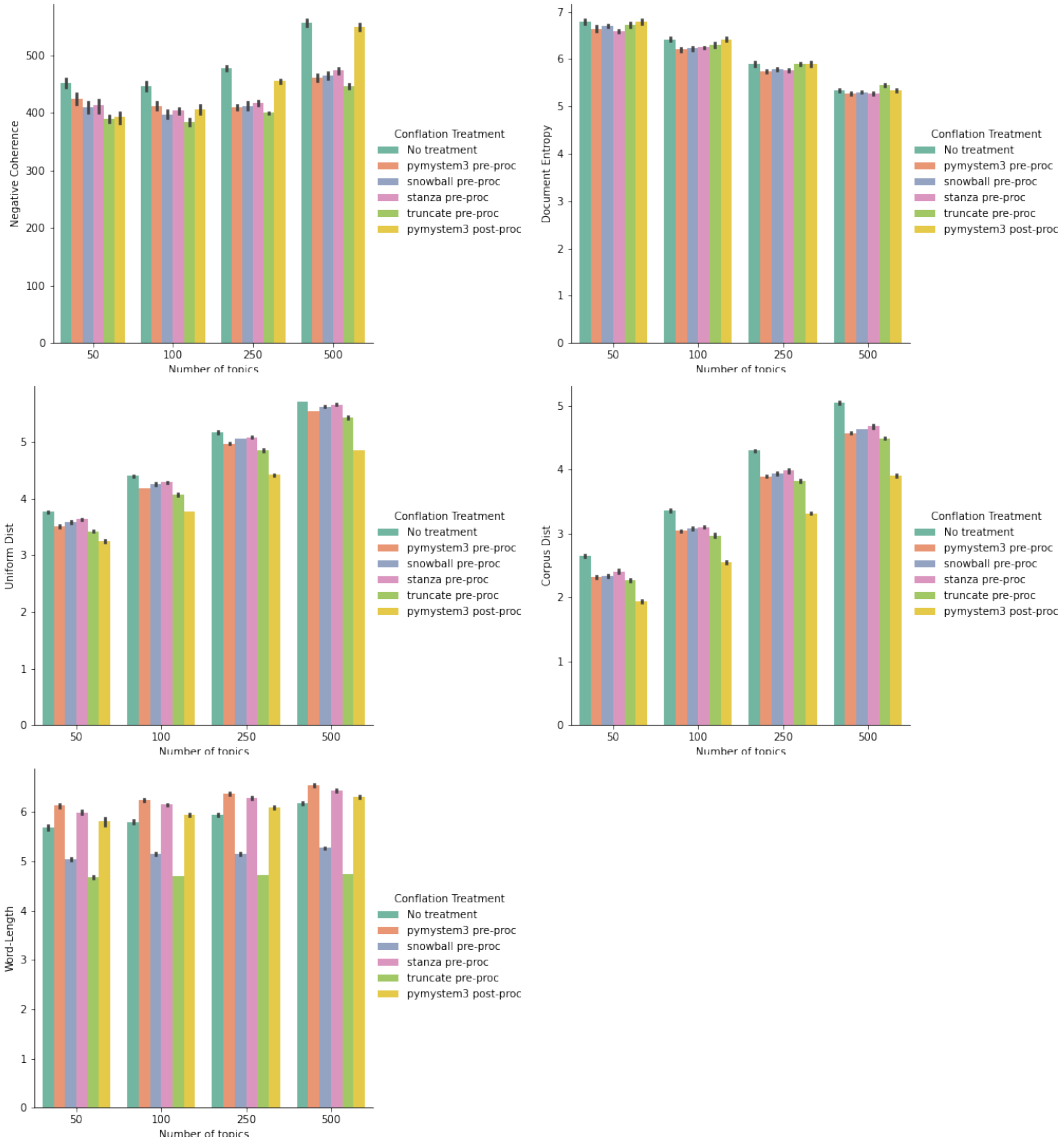

Figure 9: Comparing averages over of metrics produced by MALLET between conflation treatments. These are mainly provided as a sanity check. Note that word-length is much higher for lemmatizers than stemmers, as expected.

C Morphological Analysis Produced by Lemmatizers

Table 2: Contrasting lemmatization and stemming outputs for various surface forms. Observe that Mystem output captures syncretism and that Stanza and Mystem return different lemmas for perfective verbs. Note that Mystem outputs are translated from Russian abbreviations.

Surface form	**Translation and morphological features**	**Mystem**	**Stanza**	**Snowball**
ответ	‘answer’ Noun, Masculine, Singular, Inanimate Nominative or Accusative case	ответ Noun,Masc,Inan (Acc,Sing|Nom,Sing)	ответ Noun Animacy=Inan, Case=Nom, Gender=Masc, Number=Sing	ответ
ответим	‘we will answer/have answered’ Intransitive Verb, Perfective, 1st person, Plural, Future tense Indicative or Imperative mood	отвечать Verb,Intransive (Pl,Imperative,1st Pers,Perf|NonPast,Pl,Indicative,1st Pers,Perf)	ответить Verb Aspect=Perf, Mood=Ind, Number=Plur, Person=1, Tense=Fut, VerbForm=Fin, Voice=Act,	ответ
отвечать	‘to answer’ Intransitive Verb, Imperfective, Infinitive	отвечать Verb,Intransive Inf,Imp	отвечать Verb Aspect=Imp, VerbForm=Inf, Voice=Act,	отвеча
ответить	‘to answer’ Intransitive Verb, Perfective, Infinitive	отвечать Verb,Intransive Inf,Perf	ответить Verb Aspect=Perf, VerbForm=Inf, Voice=Act,	ответ
большой	‘big’ Adjective, Sing Masc, Animate: Nominative Masc, Inanimate: Nominative, Accuastive Fem: Genitive, Prepositional, Dative, Instrumental	большой Adjective (Acc,Sing,Full,Masc,Inan|Acc,Sing,Full,Masc,Inan|Nom,Sing,Full,Masc| Prep,Sing,Full,Fem|Dat,Sg,Full,Fem|Gen,Sing,Full,Fem|Instr,Sing,Full,Fem)’	большой Adjective Case=Nom,Degree=Pos,Gender=Masc,Number=Sing	больш

D Comparison between untreated and post-stemmed topics

Table 3: A selection of topics from a 100 topic model without any stemming treatment. Within a topic, surface forms with the same lemma are marked with the same color.

ID	Top 20 Keywords	Author Entropy	Slot Entropy	Lemma Entropy	Slots to Keywords	Lemmas to Keywords	Comment
A	клим лидия клима макаров климу варавка иноков спивак туробоев лидии дронов лютов мать чтоб пред макарова дмитрий алина нею диомидов	0.1486	6.4539	8.6838	0.65	0.8	Mostly names from *The Life of Klim Samgin* by Maxim Gorky
B	цинциннат м-сье пьер родион директор цинцинната родриг цинциннату иванович адвокат марфинька камере роман стол эммочка дверь марфиньки камеры библиотекарь камеру	0.1926	6.1703	8.2002	0.7	0.8	Names and themes from *Inviation to a Beheading* by Vladimir Nabokov
C	солнце небо сад ветер воды солнца берегу казалось вокруг здесь землю земли вода ними воздух далеко тени лес около саду	2.5730	6.3960	9.7248	0.7	0.8	Words about nature: ‘sun’, ‘river’, ‘garden’, ‘water’, ‘riverbank’, ‘sky’, ‘forest’,‘earth’
D	говорила сама спросила стала думала хотела могла нею знала начала мать продолжала одна пошла вышла видела сидела встала села смотрела	2.7266	6.0929	9.6380	0.5	1	The feminine reflexive pronoun ‘herself’, ‘mother’, and many past tense feminine verbs: ‘talked’, ‘asked’, ‘became’, ‘wanted’, ‘thought’...
E	фр de la vous le et un je a франц c’est mon pas que il les англ mais est ma	1.8941	3.8524	4.9799	0.15	0.15	French words and ‘France’.
F	говорит говорю смотрит стоит идет видит кричит спрашивает сидит отвечает глядит девица начинает лежит берет молчит хочет бежит плачет слышит	2.5775	5.8909	9.3936	0.4	0.95	Mainly verbs about talking and interacting, nearly all in the 3rd person.
G	петрович базаров аркадий николай павел василий иванович анна сергеевна петровича промолвил евгений одинцова катя базарова фенечка аркадия заметил аркадию франц	0.7604	6.2496	8.2076	0.55	0.85	Nearly all names and patronymics with two past tense masculine verbs, ‘uttered’ and ‘noticed’.
H	продолжал ответил начал проговорил посмотрел нему голосом голос встал подошел глазами заметил заговорил улыбнулся отвечал сел прибавил обратился тотчас сейчас	2.5037	6.2502	9.4994	0.5	0.9	‘Voice’ and past tense masculine verbs of talking and interacting: ‘continued’, ‘began’, ‘spoke’, ‘looked’...

I	моей моя мое моего мои мою ко мною моих сердце моим мной моем жизни нам моему сердца могу жизнь наши	2.5911	6.8899	8.6942	0.9	0.4	The first person possessive pronoun мой ‘my’ repeated many times. The words ‘life’ and ‘heart’ also appear.
J	мальчик учитель мальчика учителя детей школы дети лет мальчики мальчишка школе мальчиков классе гимназии школу уроки девочка урок ребенок товарищей	2.3820	6.7395	8.8690	0.7	0.55	A topic about school, but мальчик ‘boy’ is repeated. The other repeated words are ‘teacher’, ‘lesson, ‘child’.

Table 4: The same topics from table 3 but with post-stemming applied

ID	Top 20 Keywords	Author Entropy	Slot Entropy	Lemma Entropy	Slots to Keywords	Lemmas to Keywords	Comment
A-post	клим лидия макаров варавка инок спивак дронов мать туробой алина дмитрий диомидов томилин лют чтоб маракуев пред дядя нехаева девушка	0.1486	4.5456	8.6617	0.5	1	Mostly names from *The Life of Klim Samgin* by Maxim Gorky
B-post	цинциннат пьер м-сье родион директор родрига камера марфинька иванович адвокат стол койка эммочка роман стул крепость дверь библиотекарь давать казнь	0.1926	4.1709	8.1850	0.45	1	Names and themes from *Invitation to a Beheading* by Vladimir Nabokov
D-post	спрашивать говорить становиться сам мать один она думать отвечать муж выходить мочь хотеть пойти улыбаться начинать знать голос продолжать свой	2.7266	4.4762	9.6360	0.6	1	The stemmed version of D, but all verbs are now infinitive forms, so the feminine subject information is lost.
E-post	фр de la vous le et un je a франц c’est mon pas que il les англ mais est ma	1.8941	2.8536	4.9795	0.15	0.15	French words and ‘France’.
F-post	говорить стоять кричать идти смотреть сидеть спрашивать глядеть видеть отвечать слышать девица начинать лежать плакать бежать смеяться ходить приходить берет	2.5775	4.5224	8.5917	0.55	1	Mainly verbs about talking and interacting, nearly all in the 3rd person.
G-post	петрович базаров аркадий николай павел василий анна сергеевна иванович промолвить катя евгений одинцова фенечка петр замечать брат отец франц отвечать	0.7604	4.5113	8.1489	0.5	1	Nearly all names and patronymics, plus three verbs ‘to utter’, ‘to notice’, ‘to answer,’ and two common nouns, ‘father’ and ‘brother’.

H-post	свой отвечать голос лицо улыбаться продолжать глаз стол садиться начинать вставать обращаться говорить подходить посмотреть улыбка хотеть проговаривать замечать взгляд	2.5037	4.3617	9.5059	0.45	1	Possessive pronoun свой, some common nouns, ‘voice’, ‘face’, ‘eye’, ‘table’, ‘smile’, and many verbs about interacting, ‘to answer’, ‘to continue’, ‘to begin’, etc...
I-post	мой наш сердце я жизнь ко мы мочь любовь твой слеза душа один год имя свет быть видеть леон друг	2.5911	4.3650	8.6968	0.7	1	These are very common words, particularly pronouns, but also ‘love’, ‘tears’, ‘soul’, ‘name’ and the infinitive forms of ‘to be’ and ‘to see’.
J-post	мальчик учитель ребенок школа урок мальчишка класс девочка товарищ учиться год гимназия ученик маленький учить мальчиков директор училище взрослый отец	2.3820	3.9490	8.8672	0.45	0.95	More words related to school children are visible after post-stemming, ‘to study’, ‘year’, ‘student’, ‘little’, ‘principal’, ‘adult’, ‘father’.

E Topics from models trained on preprocessed data

Table 5: These are sample topics from models with 100 topics trained on various preprocessed versions of the corpus. We see interpretable topics that are both author specific and general. Words within the same topic that likely share a lemma are highlighted in the same color. The aggressive, simpler stemmers do not conflate these because they cannot capture consonant alternation patterns in Russian verb conjugations.

ID	Top 20 Keywords	Treatment	Author Entropy	Comment
K	логин анна клавдия мотовилов андозерский шестов валя молин мальчик город ж баглаев ермолин коноплев юшка шест дубицкий крикунов гомзин заговаривать	Mystem	0.149863	Mostly names from *Bad Dreams* by Fyodor Sologub, but also ’boy’, ’city’
L	цинциннат м пьер сье родион директор родрига марфинька камера иванович адвокат эммочка койка стена роман друг стул паук крепость проговаривать	Mystem	0.1779	Names and themes from *Invitation to a Beheading* by Vladimir Nabokov
M	слеза сердце плакать бросаться подходить тихо ужас сила умирать упасть грудь бледный вставать прощать останавливаться слышать быстро дрожать подымать бросать	Mystem	2.7315	Words about sadness: ‘tears’, ‘heart’, ‘to cry’, ‘to throw oneself’, ‘horror’, ‘strength’
N	сад дерево лес цветок солнце трава лист белый зеленый куст старый тень дорожка поле здесь земля аллея гора зелень скамейка	Mystem	2.6638	Words about nature: ‘garden’, ‘tree’, ‘flower’, ‘sun’, ‘grass’, ‘leaf,’ ‘white’, ‘green’
O	прокуратор пилат иуда иешуа город левий гость ершалаим сад арестант секретарь афраний игемон балкон столб дворец матвей луна дорога солнце	Stanza	0.34331	Words related to biblical themes: ‘Pontius Pilate’, ‘Judea‘, ‘Jerusalem’, ‘garden’. This is a topic highly correlated with Mikhail Bulgakov from imagery in *Master and Margarita*.
P	вода солнце лес берег дерево дорога сад белый небо река земля трава далеко куст поле гора зеленый лист ветер старый	Stanza	2.5371	Words about nature: ‘water’, ‘sun’, ‘riverbank’, ‘tree’
Q	генерал город чиновник имя петербург губерния губернатор губернский помещик служба служить сын дворянин весьма господин чин русский советник мундир предводитель	Stanza	2.5535	Words about the civil service and governance: ‘general’, ‘city’, ‘official’, ‘Petersburg’, ‘province’, ‘governor’, ‘service’
R	письм писа написа бумаг записк чита получ пишет конверт пер переда взял пиш лист бумажк строк буд соб прочел почерк	Snowball	2.5884	Stems about writing, sending and reading letters.

S	клим лид самгин варавк макар лют алин инок спивак туробо дрон варвар дмитр макаров томилин пред дьякон девушк брат снов	Snowball	0.1525	Names from *The Life of Klim Samgin* by Maxim Gorky
T	самги варва марин клим снова револ дроно тагил пред чтоб безбе толст кутуз ивано затем любаш дуняш чорт пробо вспом	Truncate	0.1981	Names from *The Life of Klim Samgin* by Maxim Gorky with some roots of common words mixed in.
U	лошад мужик кучер коляс крыль подъе телег прика стари экипа поеха колес барин ехать извоз запря приех станц бричк карет	Truncate	2.5469	Roots about traveling, ‘horse’, ‘coachman’, ‘wheel’.
V	письм бумаг запис напис писал получ писат бумаж кабин карма подпи проче доста проси конве посла пишет стол читат адрес	Truncate	2.6027	Roots about writing and letters.