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

Virginia Partridge
University of Massachusetts Amherst
vcpartridge@umass.edu

Abstract

TODO

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 text from online platforms and social media 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 that 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 syntax.

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. By investigating topic models produced with no morphological preprocessing step, we demonstrate that morphological features can be evident in a topic and propose ways quantify this relationship. These observations cast doubt on whether preprocessing is necessary or if its use can actually obscure information. We also present post-processing of topics as an alternative to aid with interpretability by the end users. Finally, we compare various stemming and lemmatization treatments as preprocessing the corpus and contrast this with post-processing the keywords learned by models.

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 lemma-

tized 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 still adds 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 Methods

3.1 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 conflations.

In the word-based morphology framework, inflection is captured by triples consisting of the surface form (also called wordform) w, a lexeme signifying the pairing of the surface for with a 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 lexemes 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, as multiple slots may be realized with a single surface form. 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.

3.2 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

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

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

original rules for Russian[3] (Porter, 2001).

Mystem: This Yandex-owned tool is the most popular lemmatizer for Russian and can be used with or 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].

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).

3.3 Latent Dirichlet Analysis

Latent Dirichlet Analysis 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, taking a joint distribution of the observed and latent variables, you are finding 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.

3.4 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 obtain and 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

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

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

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

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 lemmas with the same grammatical features, for example different verbs conjugated in the first-person singular form or nominative case masculine nouns.

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

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.

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

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, in other words 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).

Coherence: We additionally report coherence, a measure of topic quality that relies on document co-occurence frequencies of word types, which 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.

4 Corpus

The selected RussianNovels[6] corpus is a collection of 101 Russian literary works from the 19th and

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

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.

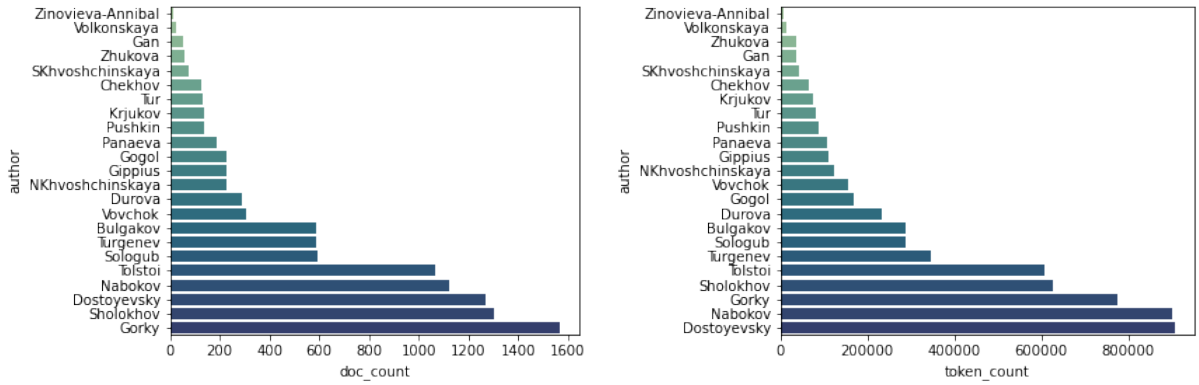

Figure 1: The number of documents per author used to train topic models (left) and the token counts by author before pruning (right).

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 3.2. 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).

5 Results

After pre-processing the corpus using each stemmer or lemmatizer, we trained 5 models for each number of topics $T \in \{50, 100, 250, 500\}$. We also train models on the corpus with no pre-processing treatment at all, other than pruning. Additionally, we did post-lemmatize the models without conflation treatment, using Mystem to produce topics with lemmas as the associated terms rather than surface-forms, but time did not allow for in depth investigation into these models.

5.1 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

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

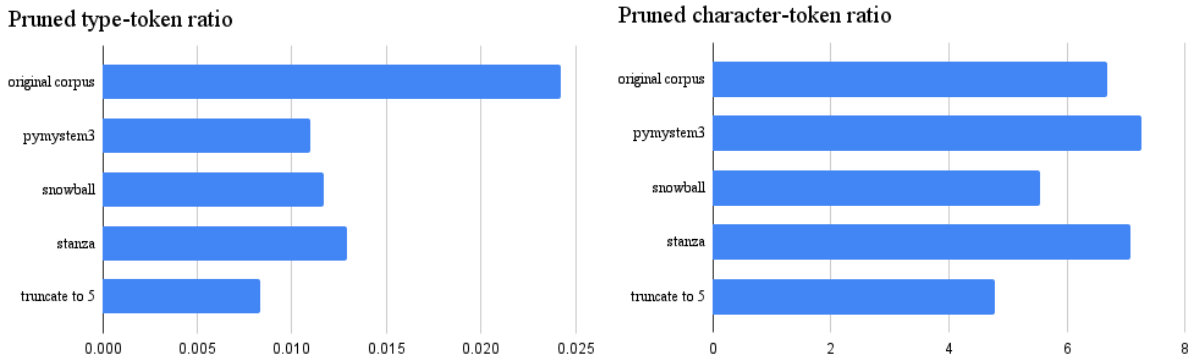

Figure 2: The ratio of word types to tokens for each conflation treatment after the vocabulary is pruned is shown left and demonstrates the strength of conflation method in reducing the vocabulary. Character to token ratio shows lemmatization returns longer normalized strings, while stemming shortens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

way that is expectedly similar to English. Both the author correlation measures plotted in Appendix A and anecdotal analysis of topics' key words find topics with both low and high author entropy.

Moreover, lemmatization and stemming with Snowball do not seem to affect whether author-specific topics or cross-cutting topics are learned, regardless of the number of topics. Compared to the other conflation treatments, truncation does reduce the number of topics with low author entropy, but only when the number of topics is low.

5.2 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 suprise is that Mystem is the next most agressive, followed by Snowball, then Stanza. Since both Mystem and Stanza map surface 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 feature matter in the corpus domain.

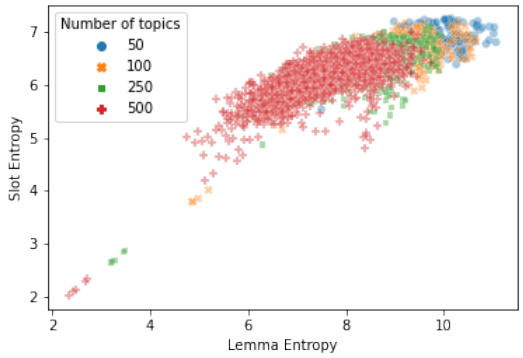

Figure 4: Lemma entropy vs slot entropy for models without lemmatization.

6 Future Work

Standardization of stop words for more consistent comparison between metrics Shannon-Jenson divergence Manage vocabularies between stemmed and unstemmed corpus, account for vocabulary size when producing metrics, especially coherence Stability

Corpora from other domains - Russian National Corpus and OpenCorpus

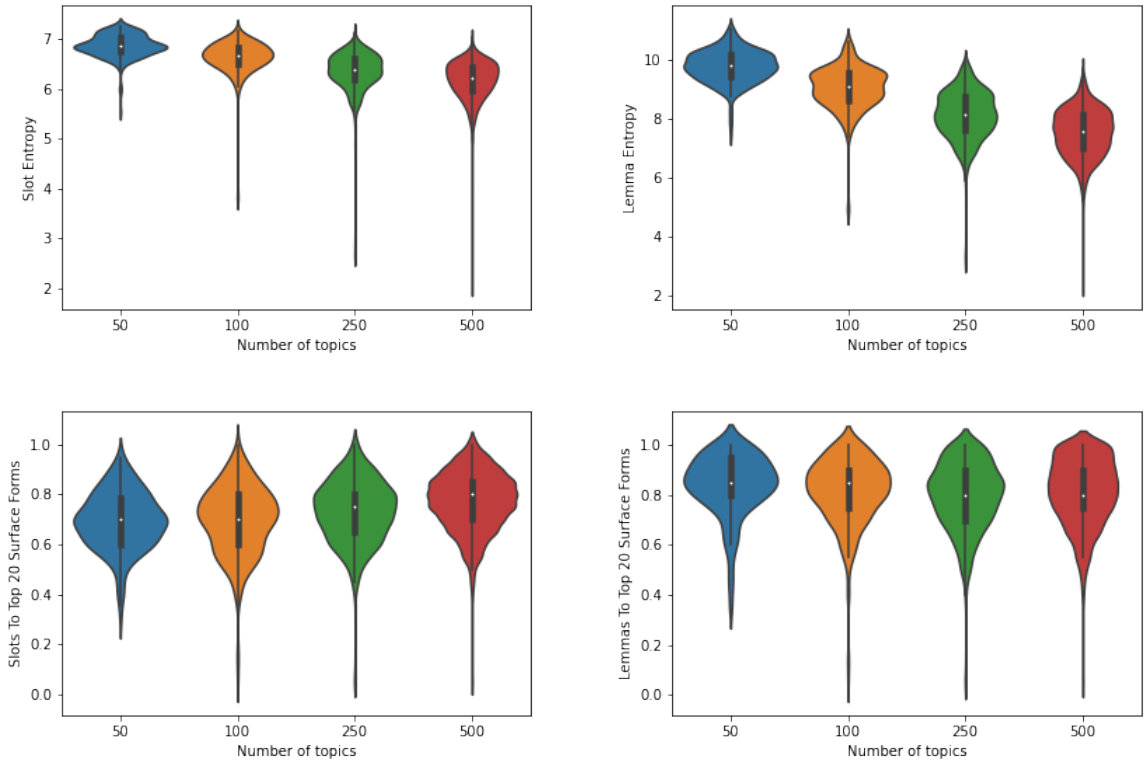

Figure 3: Values of metrics that capture morphological information in topics. These results are for topic models trained on the untreated corpus.

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.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):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. Http://www.cs.umass.edu/ mccallum/mallet.

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 D. Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *COLING*.

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

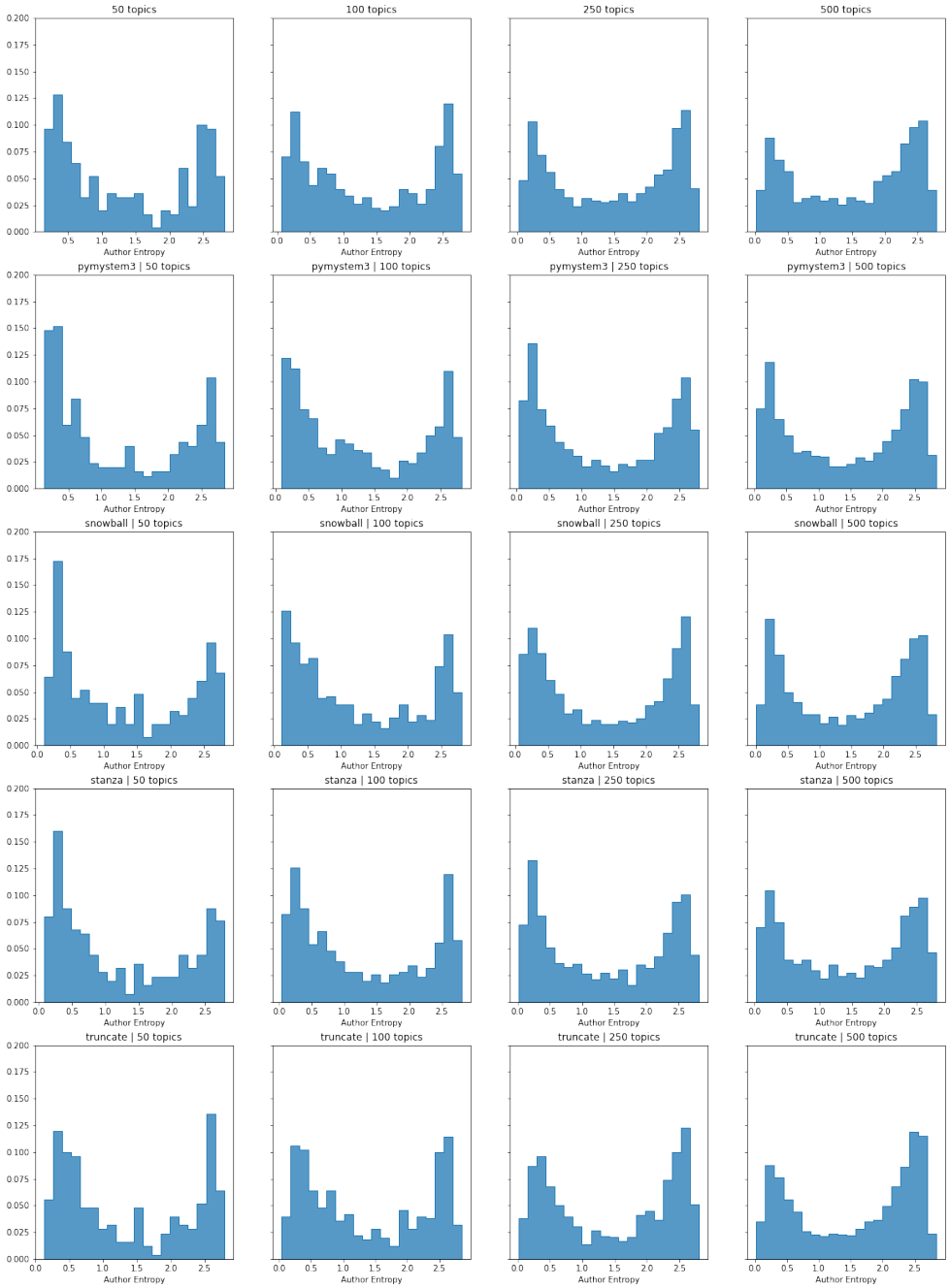

Figure 5: Author Entropy metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of conflation treament used.

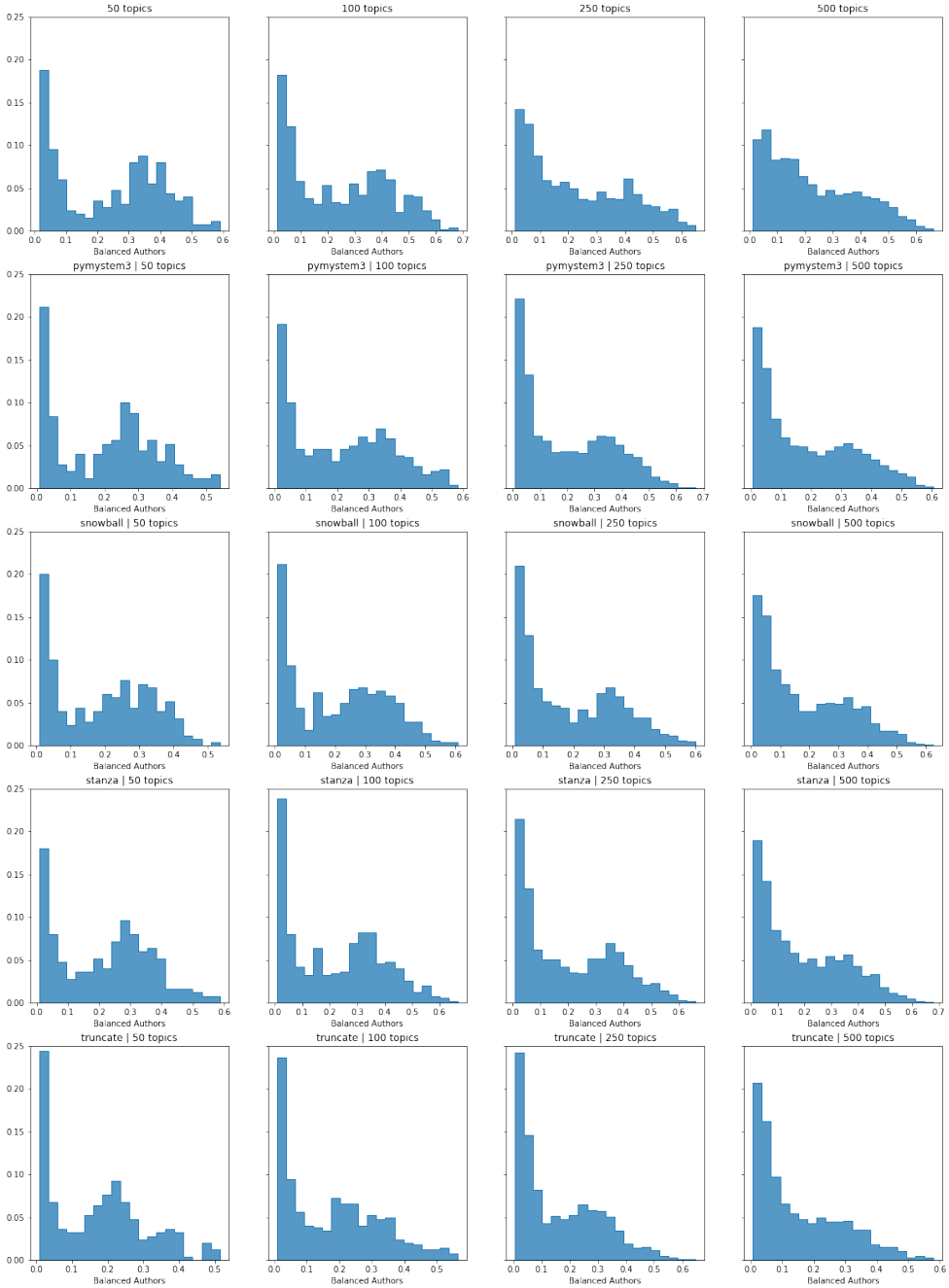

Figure 6: Balanced Author metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of conflation treament used.

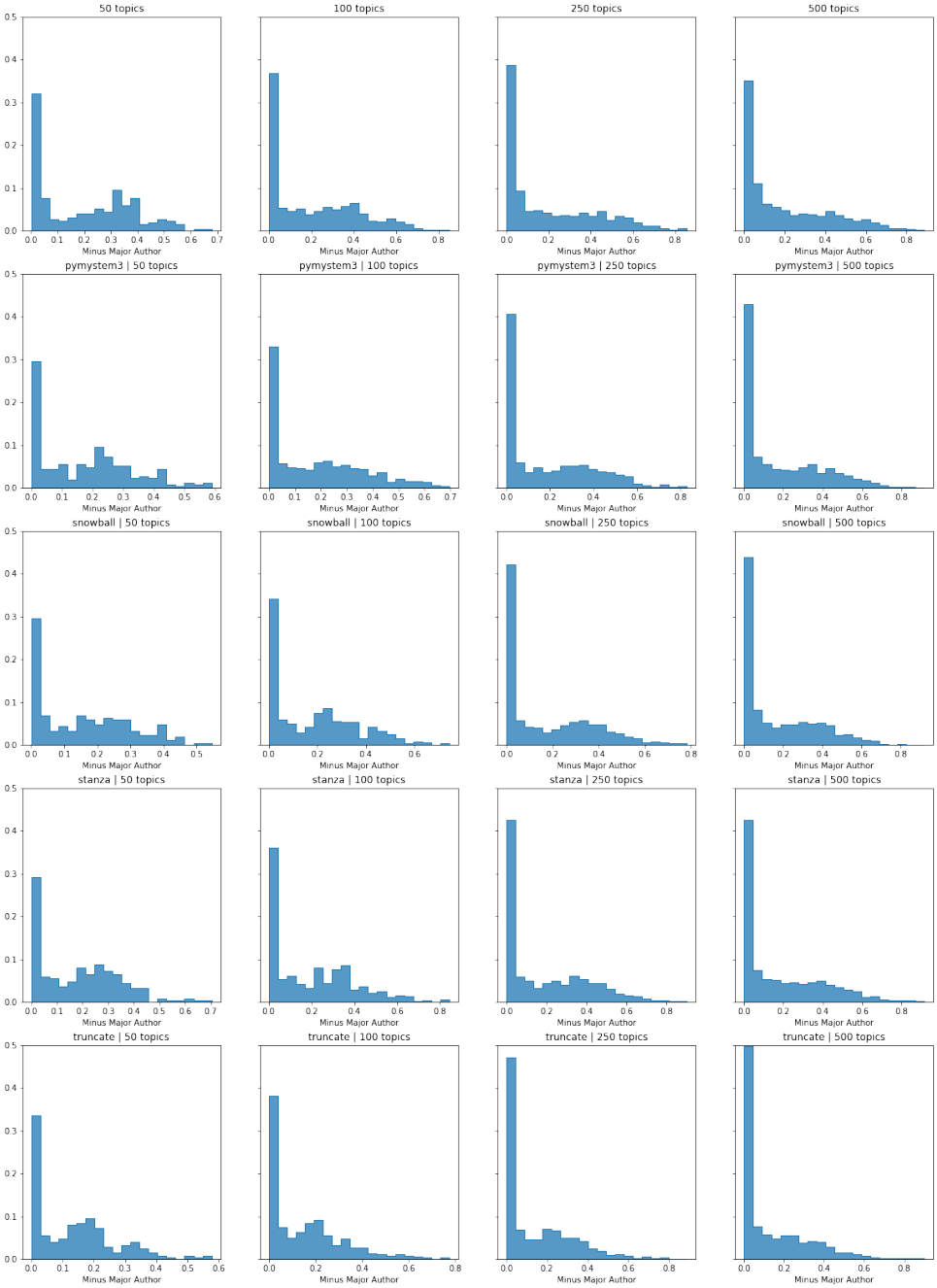

Figure 7: Balanced Author metric values for the 5 topic models trained for each number of topics $T \in \{50, 100, 250, 500\}$, broken down by the number of topics and the type of metrics conflation treament used.

B MALLET Diagnostics Metrics

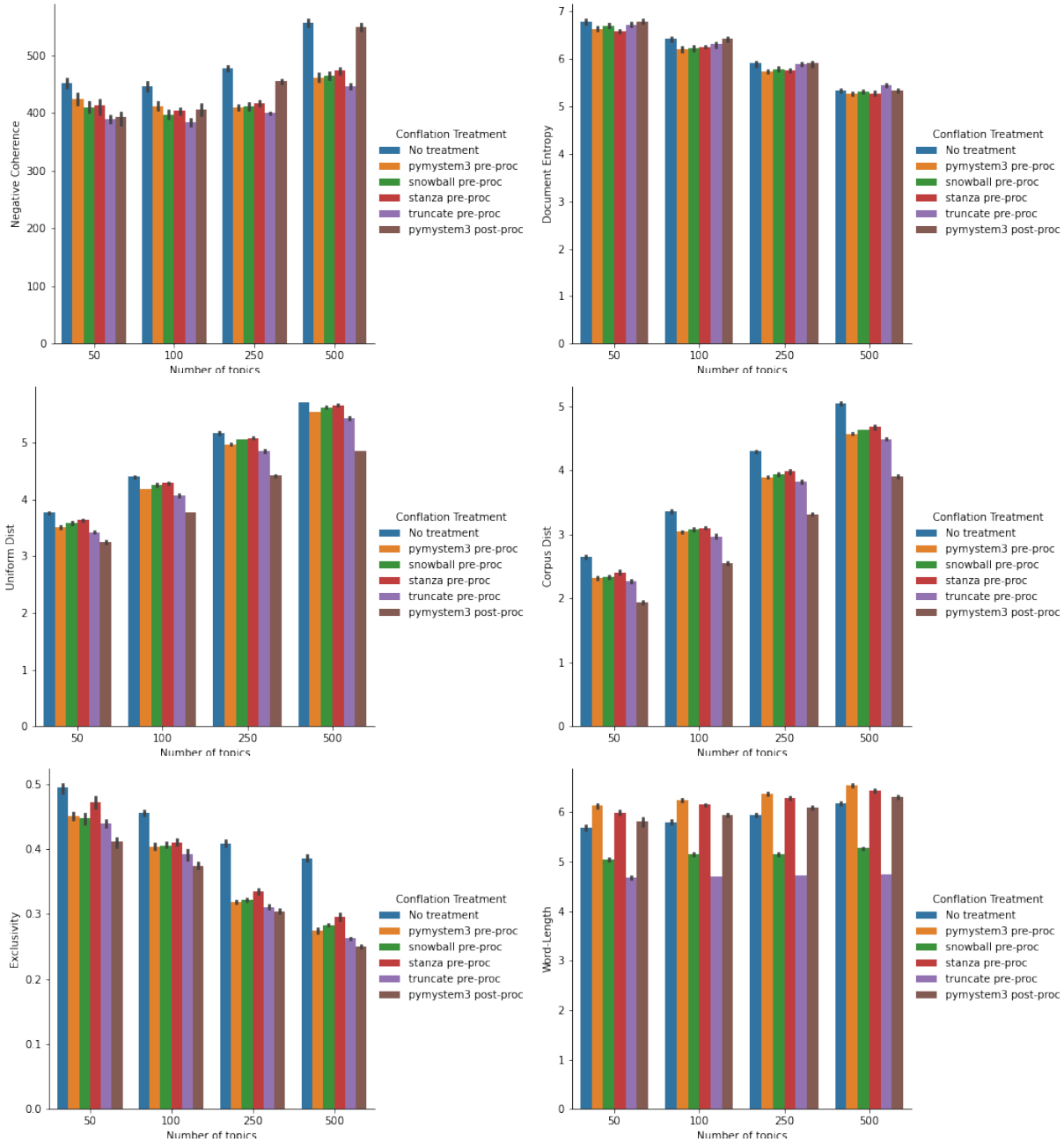

Figure 8: Comparing averages over of metrics produced by MALLET between conflation treatments. These are mainly provided as a sanity check. Note that word-length is much higher for lemmatizers than stemmers, as expected.

C Outputs of Lemmatizers and Stemmers

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

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.