Quantifying Morphology in LDA Topic Models

Virginia Partridge
University of Massachusetts Amherst
vcpartridge@umass.edu

Abstract

Latent Dirichlet allocation (LDA) is a popular approach for probabilistic topic modeling, frequently applied in many disciplines for exploring themes and trends in large document collections. Because LDA assumes a bag-of-words approach, stemming or lemmatization are common pre-processing steps in preparing corpora for topic modeling, despite little evidence that these improve topic quality for English. Recent work has suggested post-processing topics so that they are more interpretable by end users may be a better approach.

There is more motivation to apply stemming or lemmatization for topic modeling on languages with complex inflectional morphology, due to concerns that rare word forms will either be randomly assigned to topics or cause lexemes that are important to a topic's interpretability to be obscured. We present several metrics designed to quantify the morphological and lexical complexity of topics learned by LDA, with a focus on identifying topics that may benefit from post-processing. We then use these metrics to analyze LDA topic models trained with Gibbs sampling on Russian and German corpora, comparing the effects of different stemmers and lemmatizers.

1 Introduction

Latent Dirichlet allocation (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 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? 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). To understand the effects of these treatments on languages with more inflectional morphology, we train topic models on the German TIGER corpus[1] (Brants et al., 2004) and the Russian National corpus[2] (RNC) (Apresjan et al., 2006). These corpora have high quality morphological and syntactic annotations, which allow for analysis based on gold standard morphological analyses and lemmatization.

There are many ways which choices about stemming or lemmatization could affect the quality and interpretability of topic models and these effects are not mutually exclusive. First, in the absence of any morphological treatment, a topic model may learn to concentrate all the surface forms of a particular lexeme in a single topic. A topic identified by repeated forms of the same lexeme is not likely to be useful to end users, making it a good candidate for post-stemming to reveal more lexemes and therefore more context for the topic. Alternatively, a topic could encompass many lexemes, but few grammatical forms. Our hypothesis is that this may occur when a documents share stylometric similarities, such as the top keywords for a topic with dialogue-heavy documents being first-person and second-person verb forms. Applying stemming or lemmatization in pre-processing would prevent the formation of such a topic and applying it in post-

[1] https://www.ims.uni-stuttgart.de/en/research/resources/corpora/tiger/

[2] https://ruscorpora.ru/old/en/corpora-morph.html

processing would obscure the stylistic information encoded by the grammatical form.

To examine the extent to which these issues arise, we adapt measures of morphological complexity to analyzing LDA topics produced by Gibbs sampling, quantifying the ways in which inflectional morphology influence topic models and identifying topics where morphology complicates interpretability. So long as reliable morphological analyses are available, these methods can be applied cross-linguistically, which we demonstrate using topic models for TIGER and RNC.

Pre-processing treatments could also hurt the quality of topic models in terms of reproducibility or topics' semantic coherence. Following Schofield and Mimno's work, we use topic coherence as a stand-in for human judgements of topic quality (Mimno et al., 2011) and Variation of information (VOI) (Meila, 2003) to show how stemming and lemmatization change tokens' topic assignments under various pre-processing treatments over multiple experiments.

2 Related Work

The proposal for applying stemming in post-processing comes from work comparing the effects of various stemming approaches on English evaluated on likelihood of a held-out test corpus, topic coherence and clustering consistency with VOI (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 stemming in pre-processing does not empirically improve LDA topic models and may hurt topic stability. Post-processing is still be valuable from the perspective of topic interpretability, avoiding repeating different surface forms of the same lexeme in topics' key word lists and presenting users with concise results.

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 post-processing as an effective solution.

Although we found little prior work on the effects of stemming on topic modeling for German, Rieger et al. present methods for improving the stability of LDA and detail results of experiments on a German newspaper corpus (Rieger et al., 2020). Their focus is on increasing the reproducibility of LDA topic models by choosing the initial token allocations for Gibbs sampling after comparing multiple LDA models and they limit pre-processing to removal of stopwords and punctuation. Schofield and Mimno also explicitly measured stability of token allocations by using VOI to compare the stemming treatments, finding that certain types of stemming could increase the impact of random initialization, hurting reproducibility.

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

Table 1: Examples from of morphological analyses annotated in RNC and TIGER. Morphological features are consistent in the annotation schema allowing slots to be compared across topics and word types.

Token	Lemma	Translation	Annotation	Explanation
нормально	нормальный	*normal*	A=n,sg,brev	Adjective, neuter, singular, short-form (Russian has long and short form adjectives)
слышала	слышать	*[she] heard*	V,ipf,tran=f,sg,act,praet,indic	Verb, imperfective aspect, transitive, feminine singular subject, past tense, indicative mood
texanishe	texanish	*Texan*	ADJA,Pos,Nom,Sg,Masc	Adjective,positive grade, nominative case, singular, masculine
kennt	kennen	*knows*	VVFIN,3,Sg,Pres,Ind	Verb, finite form, 3rd person, singular, present tense, indicative mood

Table 2: Corpus statistics after removing short documents, stopwords and punctuation.

Corpus name	# documents	# tokens	Average doc length (tokens)	Unique surface forms	Unique lemmas
TIGER	1260	310,925	247	73,633	55,498
RNC	394	319,991	812	79,413	32,487

MALLET allows us to directly inspect the topic assignments at the level of each token in a document.

3.2 Framework for Morphological Complexity

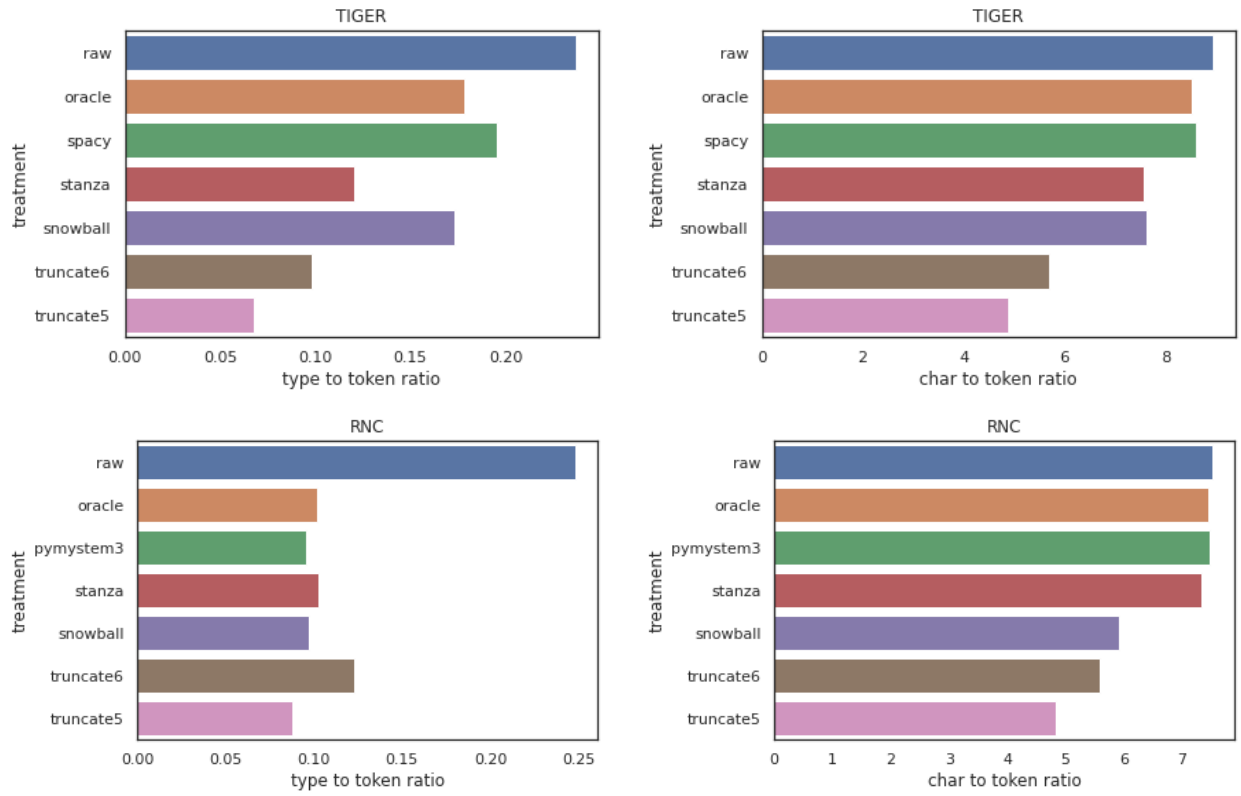
We first clarify terms for discussing morphological paradigms, following frameworks for quantifying morphological complexity used in linguistics and computational linguistics (Baerman et al., 2015b; Ackerman and Malouf, 2013; Cotterell et al., 2019). We draw a distinction between *derivational* morphology, 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 (Ackerman and Malouf, 2013). 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. This type of morphological complexity is called *syncretism* and is common in Russian noun and adjective declensions (Baerman et al., 2015a; Milizia, 2015). German also demonstrates syncretism in verbs between infinitive and present tense plural indicative forms and in adjective agreement for noun case and gender (Crysmann, 2005).

4 Corpora

In order to perform the desired analysis of the topic models, we needed corpora with high quality annotations of lemmas and morphological features and

Figure 1: Effects of treatment strength on TIGER and RNC, type-to-token ratio (left) and character-to-token ratio. Truncating to 5 characters is the most aggressive treatment. For German (top), Stanza appears to be over-lemmatizing significantly, conflating much more than necessary compared to the oracle from the corpus, while SpaCy is slightly under-lemmatizing. The Russian lemmatizers, Mystem and Stanza, have similar conflation strength to the oracle.



documents long enough for training topic models. For German, we selected the TIGER corpus version 2.2 (Brants et al., 2004), a set of newspaper articles from the Frankfuter Rundshau. The *public*, texts from newspapers and magazines, and *speech*, transcripts of radio and television interviews, portions of the Russian National corpus (RNC) were chosen for Russian. In both these corpora, morphological slots are annotated as consistently ordered lists of the features, as shown in table 1. We also considered using OpenCorpora[3], a crowd annotated Russian corpus, but this proved difficult to use for our purposes as syncretic word forms are not fully disambiguated in the annotations.

Particular care was taken in pre-processing, as evaluation metrics are sensitive to token counts and vocabulary sizes. Due to the nature of the annotations, both corpora are pre-tokenized. Documents with less than 100 tokens were excluded, then punctuation and stopwords from a fixed list were removed. Corpora statistics after these pre-processing steps are given in 2. Finally, each token was stemmed or lemmatized according to a selected method described in section 5.1. We also trained experiments with the original surface forms, later referred to as 'raw' or 'untreated'. This results in seven versions of each corpus, one for each stemming or lemmatization treatment.

5 Methods

For each pre-processing treatment described below, ten LDA topic models are trained for both 50 and 100 topic models in MALLET. This allows us to compare evaluation metrics across multiple experimental runs, reducing the chance that any observed effects result fron randomness in training.

5.1 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 can rely on a dictionary of word forms paired with outputs from a part-of-speech tagger to produce lemmas (Schofield and Mimno, 2016; Sharoff

[3] opencorpora.org

and Nivre, 2011) or may be pre-trained machine learning models for part-of-speech and morphological feature tagging (Qi et al., 2020). 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.

Oracle: This treatment consists of taking the lemma as annotated from the corpus, standing in as a highly accurate lemmatizer.

Truncation: This simple baseline method trims surface forms to the first n characters (Schofield and Mimno, 2016). We truncate with $n = 5$ and $n = 6$.

Snowball Stemmer: This stemmer was introduced as a rigorous framework for implementing stemming algorithms for a variety of languages. We utilize the NLTK implementation[4] with the original rules for Russian[5] and German[6] (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 influential Zalizniak grammatical dictionary for morphological tags (Zaliznyak, 1977), 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[7]. Notably, Mystem is as fast as the Snowball stemmer, while producing a normalized lemma form that is more interpretable for users.

spaCy: SpaCy v.3[8] supports different kinds of rule-based or dictionary lookup lemmatizers, depending the language[9]. Their German pipeline[10] uses a dictionary lookup, reporting lemmatization accuracy of 73% on data which include TIGER. We do not use spaCy for Russian.

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). Typically, Stanza models operate as a full NLP pipeline from tokenization to tagging output, however because we need to compare the output of each treatment, we used Stanza to lemmatize a single token at a time, which may hurt the accuracy of lemmatization and morphological tagging (see 1). For Russian, we use the Stanza model trained on the SynTagRus treebank[11], which has the RNC as a subset, and for German, we use Hamburg Dependency treebank model[12].

5.2 Evaluation metrics

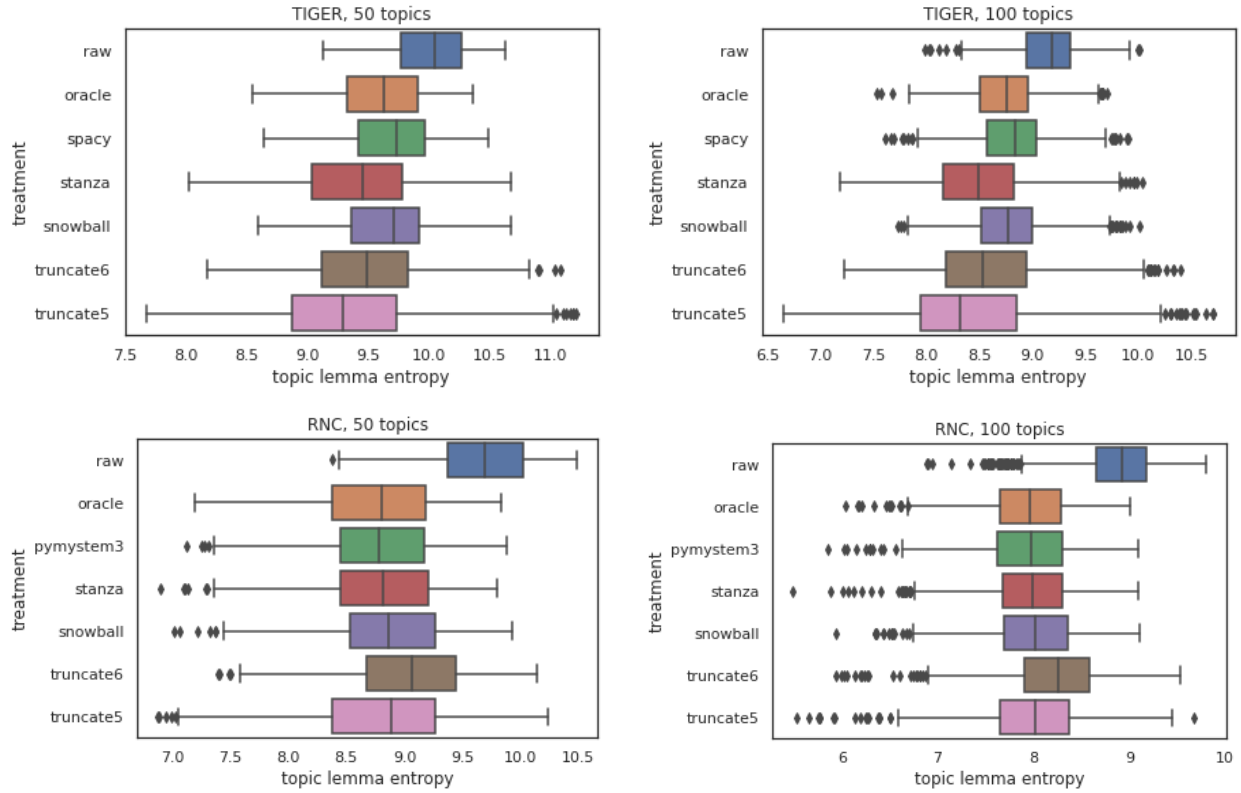
5.2.1 Entropy-based measurements

We would like to quantify the trade-offs between topic interpretability and loss of information that is linked to a surface form's morphology. The annotated corpora give the morphological analysis for a surface form w as 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 k, 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 keywords 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 Au-

[4] https://www.nltk.org/api/nltk.stem.html
[5] http://snowball.tartarus.org/algorithms/russian/stemmer.html
[6] http://snowball.tartarus.org/algorithms/german/stemmer.html
[7] pythonhosted.org/pymystem3/pymystem3.html
[8] spacy.io
[9] https://spacy.io/api/lemmatizer
[10] https://spacy.io/models/de#de_core_news_lg
[11] https://universaldependencies.org/treebanks/ru_syntagrus/index.html
[12] https://universaldependencies.org/treebanks/de_hdt/index.html

Figure 2: The distribution of lemma entropies for topics computed using the allocations of tokens over experiments different numbers of topics and pre-processing treatments, with median and interquartile range marked. The treatments that conflate more terms together result in topics with lower lemma entropy. The large ranges of lemma entropy values for truncation stemmers are a likely a reflection of how much these treatments overstem and understem by their nature, since they are not tied to the language's lexicon or morphological paradigms.



thor 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. The range for this metric is affected by the size of morphological paradigms for various parts-of-speech in a language.

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

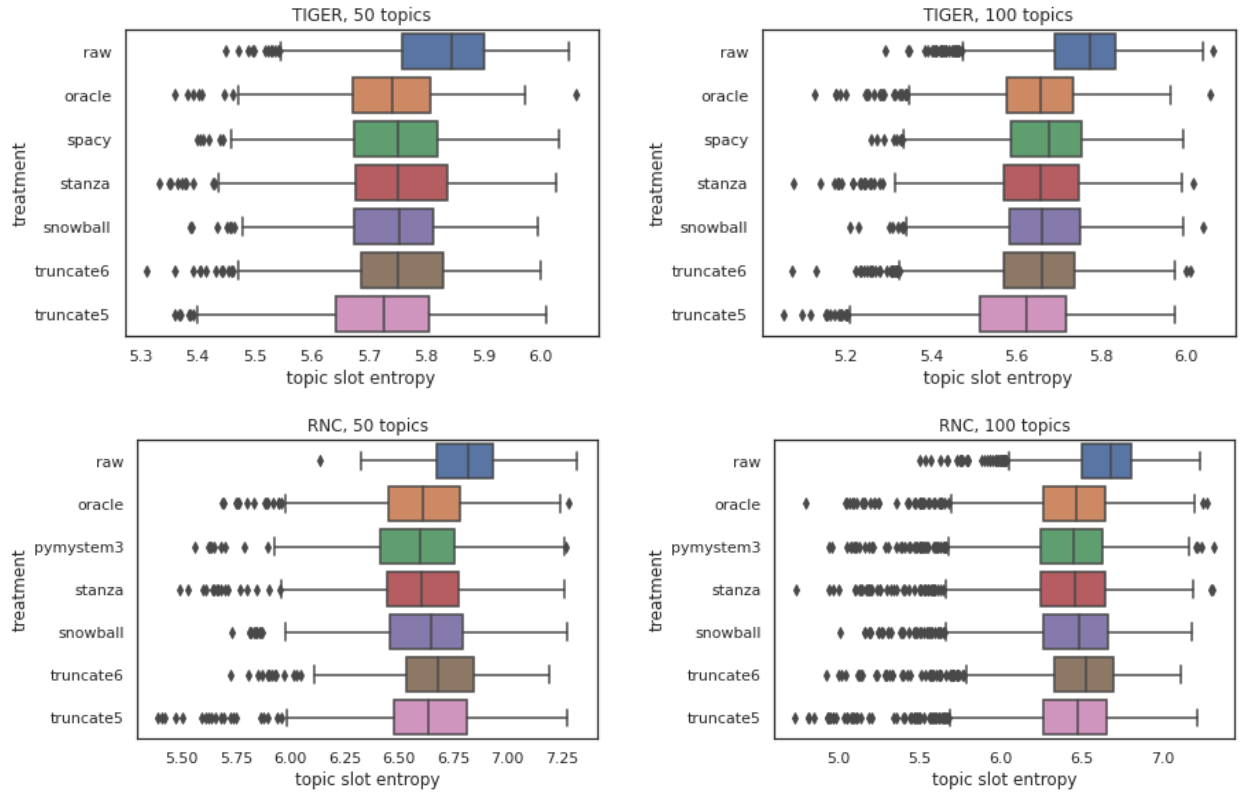
Lemma entropy: Similarly, we may want to know when a topic is dominated by a single lexeme, containing many grammatical forms of a single lexeme, but few other lexemes. For example, a topic may have many counts of different surface forms for each declension of a particular noun, its nominative, accustive, 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.

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

5.2.2 Counting-based measurements

In practice, topics are often identified by keywords, the most frequently allocated terms to a topic. However, without pre-processing it's possible that the set of top n keywords consists of many surface forms of the same lexeme, obscuring forms of other lexemes that could be useful to identifying the topic. Similarly, it's possible that a lexeme's allocations

Figure 3: The distribution of slot entropies for topics computed using the allocations of tokens over experiments with the various pre-processing treatments, with median and interquartile range marked.



to a topic are spread across many word forms, such that no forms of the lexeme appear in the keywords for the topic, even though this may be the most frequent lemma allocated for the topic. Both of these problems occurring simultaneously for many topics would suggest a need for post-processing treatment.

Lemmas expressed by top n key terms: This set is targeted at understanding how concise the presentation of a topic's key terms is to a user. When its size is close to n, each key term presented to the user represents a unique lexeme or multiple lexemes in cases of lexical ambiguity. If the set's size is closer to 1, different forms of the same lexeme are repeated in the keywords.

$$K_\ell(k) = \{\ell_w | w \in \{n \,\text{largest} N(w,k)\}\} \quad (3)$$

Top n lemmas: A topic's most frequent lexemes may not always overlap with the lexemes of its most frequent surface forms. Comparing the differences between the most a topic's most frequent lexemes, $L(k)$, defined below, and $K_\ell(k)$, will reveal topics where morphology impacts interpretability.

$$L(k) = \{\ell | \ell \in \{n \,\text{largest} N(\ell,k)\}\} \quad (4)$$

5.2.3 Strength of treatment measurements

These measurements quantify the aggressiveness of stemming or lemmatization.

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). Smaller values indicate more tokens are conflated to the same word type by the treatment.

Character-token ratio: This corpus-level 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.

5.2.4 Topic Quality

Variation of information: This symmetric metric allows for comparing different clusterings of the same dataset (Meila, 2003). Inherent randomness

in the LDA algorithm will cause some variation across experiments, but VOI will be lower when clusterings are consistently similar over multiple runs. Viewing topics as clusterings of tokens, we follow Schofield and Mimno (2016) in distinguishing *intra-treatment* VOI to quantify topic stability over experiments using the same pre-processing treatment, from *inter-treatment VOI*, used to compare the effects of one treatment on clustering to another treatment.

Coherence: An automatic metric computed from document co-occurence of a topic's top terms, coherence that has been show to correspond with human judgements on topic quality (Mimno et al., 2011). Because coherence is sensitive to vocabulary size, we calculate coherence based on the surface forms of the untreated tokens (Schofield and Mimno, 2016) and based on the lemmas of tokens, *lemma coherence*, in order to have a fair comparison between treatments.

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). This metric can also be modified to quantify the relative uniqueness of lemmas to a topic. We rely on topic exclusivity computed by MALLET [13] using either the original tokens or lemmas as the vocabulary.

6 Results

6.1 Conflation strength of treatments

6.2 Utility of entropy metrics

6.3 Counting top lexemes

6.4 VOI and topic stability

7 Future Work

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429 – 464.

Jurij D. Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid L. Iomdin, Andrei Sannikov, and Victor G. Sizov. 2006. A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *LREC*.

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.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2:597–620.

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.

Berthold Crysmann. 2005. Syncretism in German: A unified approach to underspecification, indeterminacy, and likeness of case. *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*.

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

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.

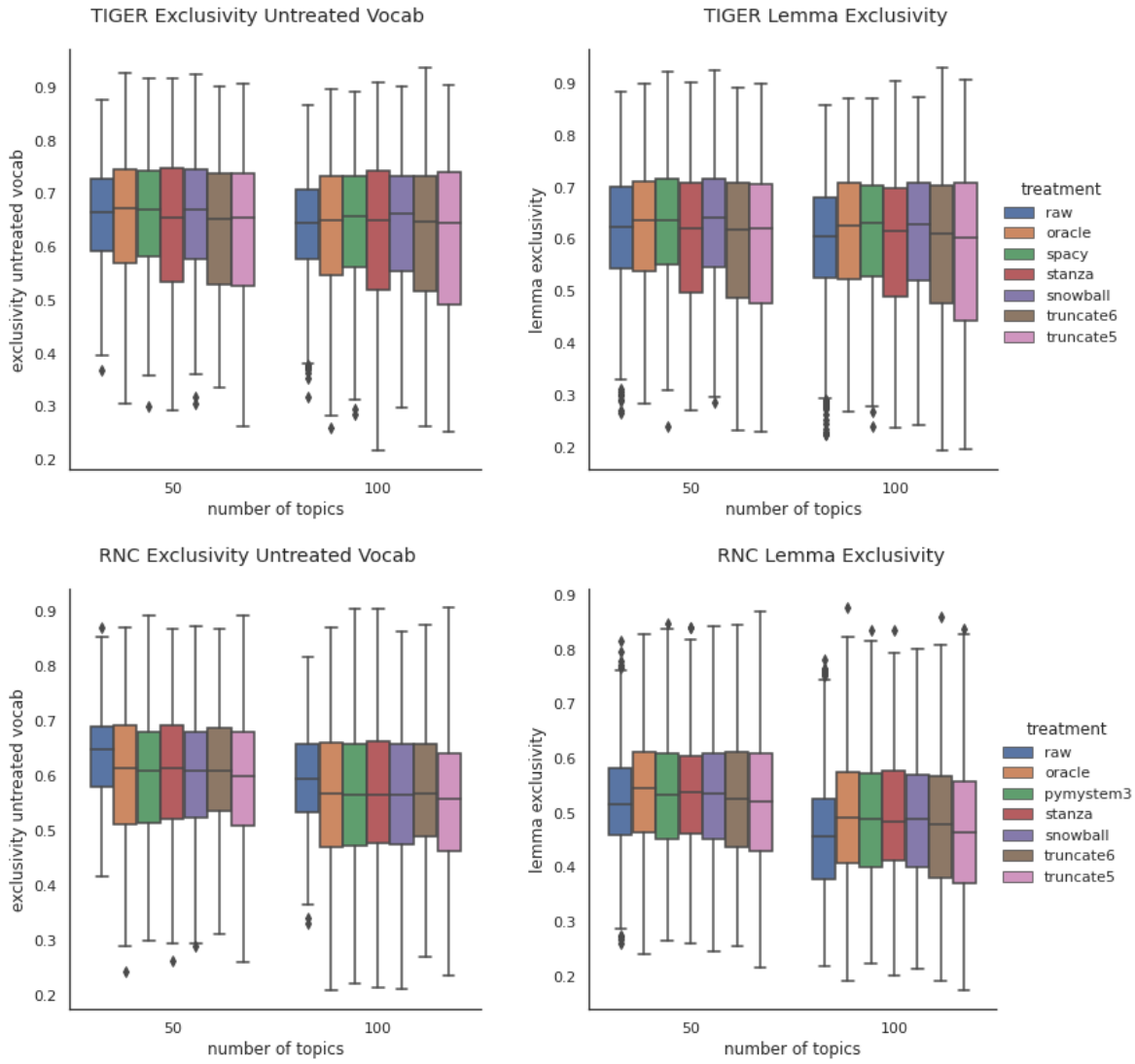
Daniel A McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D Manning, and Daniel Jurafsky. 2013. Differentiating language usage through topic models. *Poetics*, 41(6):607–625.

Marina Meila. 2003. Comparing clusterings by the variation of information. In *COLT*.

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.

[13] https://mallet.cs.umass.edu/diagnostics.php

Figure 4: Exclusivity computed with word types from the untreated vocabulary and lemma exclusivity over 10 experiments for each treatment. Higher values mean that topics' top terms do not overlap with other topics' top terms. Plots show median and interquartial range. Lemmatization and stemming increase exclusivity of word types for the German TIGER corpus, but do not have a consistent effect on the Russian National Corpus. Word types are more exclusive than lemmas in the untreated corpus, a difference which is more pronounced in the Russian corpus.



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.

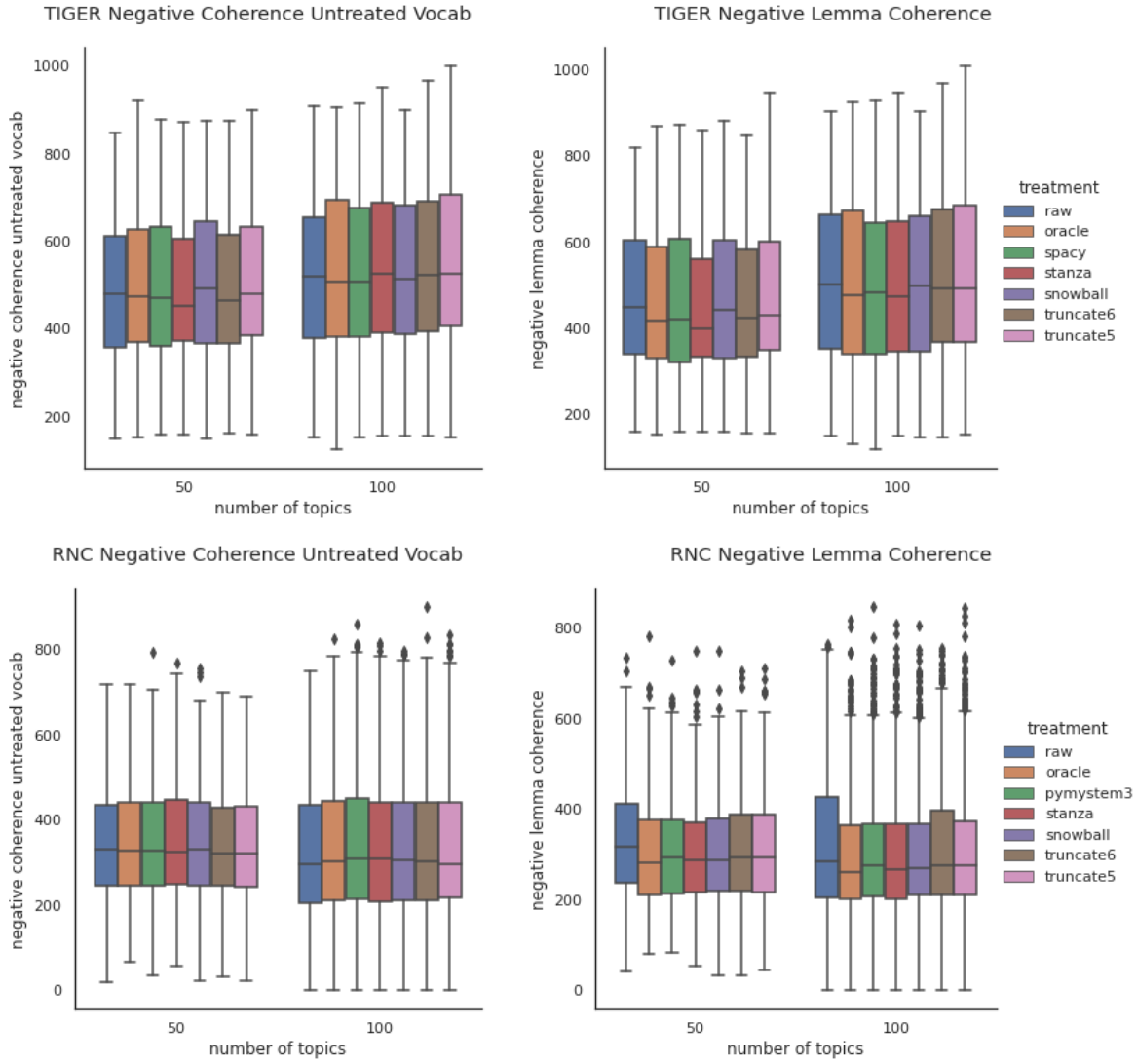
Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

J. Rieger, Jörg Rahnenführer, and Carsten Jentsch. 2020. Improving latent dirichlet allocation: On reliability of the novel method LDAprototype. *Natural Language Processing and Information Systems*, 12089:118 – 125.

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*.

Figure 5: Negative coherence computed using topic assignments for tokens using the word types in the original vocabulary (left) and lemmas over 10 experiments for each treatment. Plots show median and interquartile range. Lower values may correspond to more coherent topics according to human judgements. Lemmatization seems to increase slightly coherence for the German TIGER corpus (top), but results for the Russian National Corpus are inconclusive.



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.

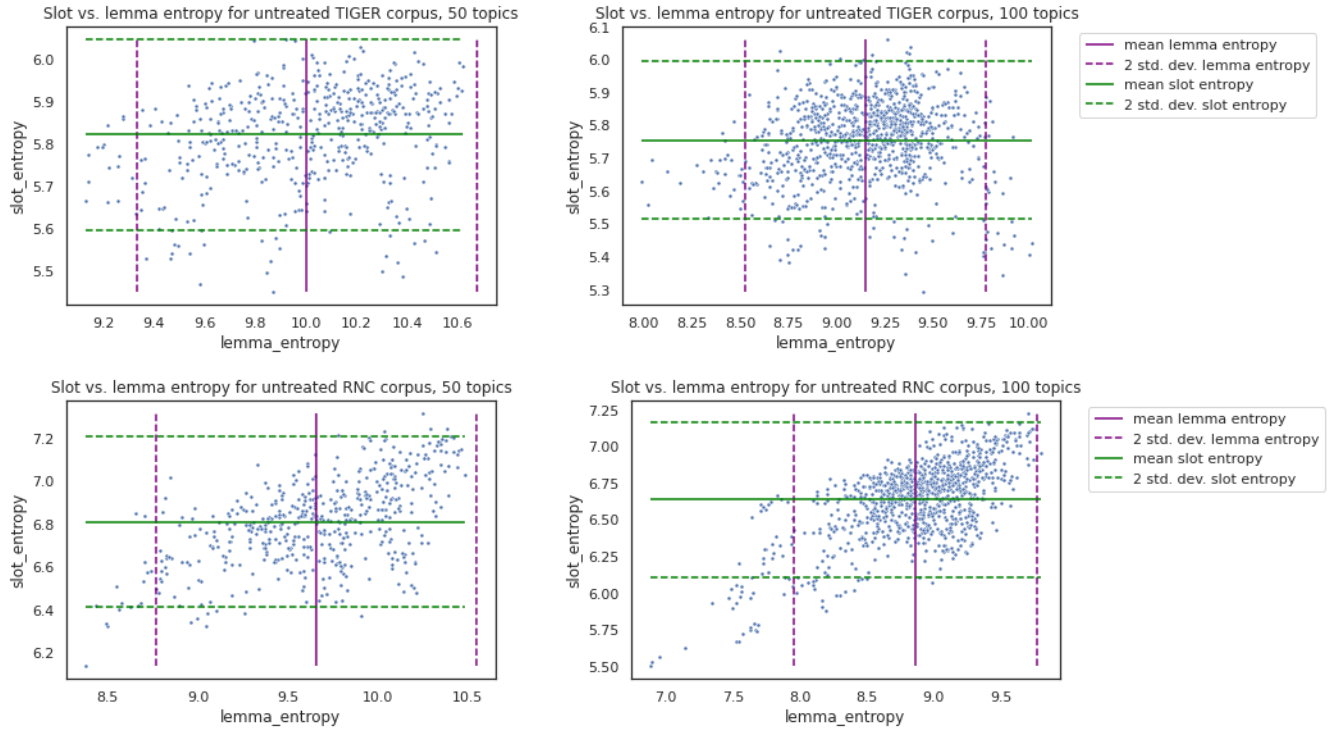
John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (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.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter.

Figure 6: Using the measurements from 10 50-topic models (left) and 10 100-topic models trained on the untreated word types, we plot a topic's lemma entropy (x-axis) vs its slot entropy. Solid lines show the mean and dotted lines indicate two standard deviations of the mean.



In *Advances in Neural Information Processing Systems*, pages 1973–1981.

Ayndrey Zaliznyak. 1977. *Grammaticheskij Slovar' Russkogo Jazyka [Grammatical Dictionary of the Russian Language]*.

Figure 7: Comparison of the top 20 lemmas for each topic and the lemmas covered by the topic's top 20 key terms, over 10 experiments on the untreated corpus. The cells show the number of topics with the corresponding set differences between $L(k)$ and $K_\ell(k)$. Values in the upper left indicate topics with high overlap in lemmas of the top terms and the topic's most frequent lemmas. Values in the lower left are topics where the top terms are obscured by the top terms, good candidates for post-stemming. Values in the lower right indicate large mismatches between those sets, a challenge for topic interpretability.

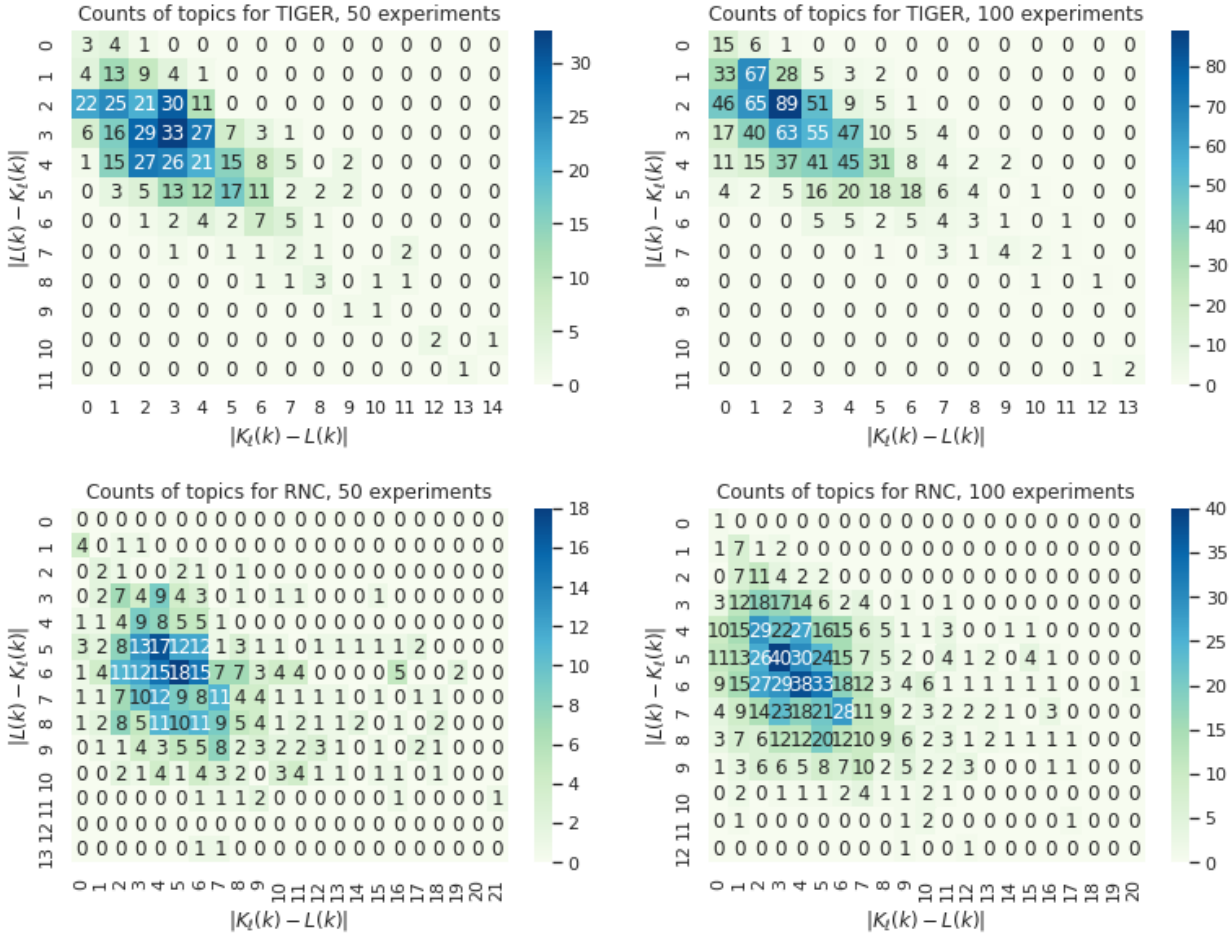


Table 3: These are sample topics from 50 topic models trained on the untreated corpora demonstrating how the difference between $K_\ell(k)$ and $L(k)$ can be used to identify topics as candidates for post-stemming.

$K_\ell(k)$	$L(k)$	$\lvert L(k) - K_\ell(k)\rvert$	$\lvert K_\ell(k) - L(k)\rvert$	**Comment**

Figure 8: Variation of information between pre-processing treatments averaged over pairwise comparison of 10 experiments for each treatment. Lemmatization methods have the lowest intra-treatment VOI, except for 100 topics in the RNC, where the no treatment gives the lowest value. Inter-treatment VOIs between truncation and lemmatization treatments are the high. Snowball has more overlap with lemmatization methods than simple truncation or no treatment.