# Quantifying Morphology in LDA Topic Models

**Virginia Partridge**
University of Massachusetts Amherst
vcpartridge@umass.edu

## 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, allowing analysis based on gold standard morphological features 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 occurs when documents share stylometric or genre-specific 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 precludes the formation of such a topic and applying it in post-processing would obscure 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 change 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 quality (Mimno et al., 2011) and Variation of information (VOI) (Meila, 2003) to show how stemming and lemmatization change the stability of topic assignments under over multiple experiments.

## 2 Related Work

The proposal for applying stemming in post- rather than pre-processing comes from work comparing the effects of various stemming approaches on En-

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

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

glish, evaluating 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 in pre-processing may be beneficial. However, they also suggest measuring the effects of lemmatization and do not rule out post-processing as an effective alternative.

Although we found little prior work on the effects of stemming in topic modeling for German, Rieger et al. present methods for improving the stability of LDA and detail 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.

Ample work in the field of linguistic typology has tried to quantify morphological complexity cross-linguistically. Baerman et al. recount various counting-based, entropy-based and description-based measures used to establish the generative power and degree of unpredictability in morphological systems (Baerman et al., 2015). Ackerman and Malouf present two such measures of complexity for inflectional morphology, *enumerative complexity*, capturing the size and nature of a language's inflectional paradigms, and *integrative complexity*, an entropy-based measure about predictability of surface forms (Ackerman and Malouf, 2013). Other researchers have used morphological analyzers and annotated data to empirically evaluate hypothesized trade-offs between enumerative complexity and integrative complexity (Cotterell et al., 2019), an approach we draw upon in this work.

## 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 $\phi_t$ is drawn from with a Dirichlet prior with concentration parameter $\beta$. Each document $d$ also has a multinomial distribution $\theta_d$ over the terms in the vocabulary, also drawn from a Dirichlet prior with concentration parameter $\alpha$. Viewing LDA as a generative process with a joint distribution of the observed and latent variables, find the $\phi_t$ and $\theta_d$ 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 $\beta$ and an asymmetric prior for $\alpha$ 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 first clarify terms for discussing morphological paradigms, following frameworks for quantifying morphological complexity used in linguistic typology and computational linguistics (Baerman et al., 2015; Ackerman and Malouf, 2013; Cotterell

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

| Token | Lemma | Translation | Slot 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. TIGER has more documents, but the documents in RNC are much longer.

| 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 |

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.

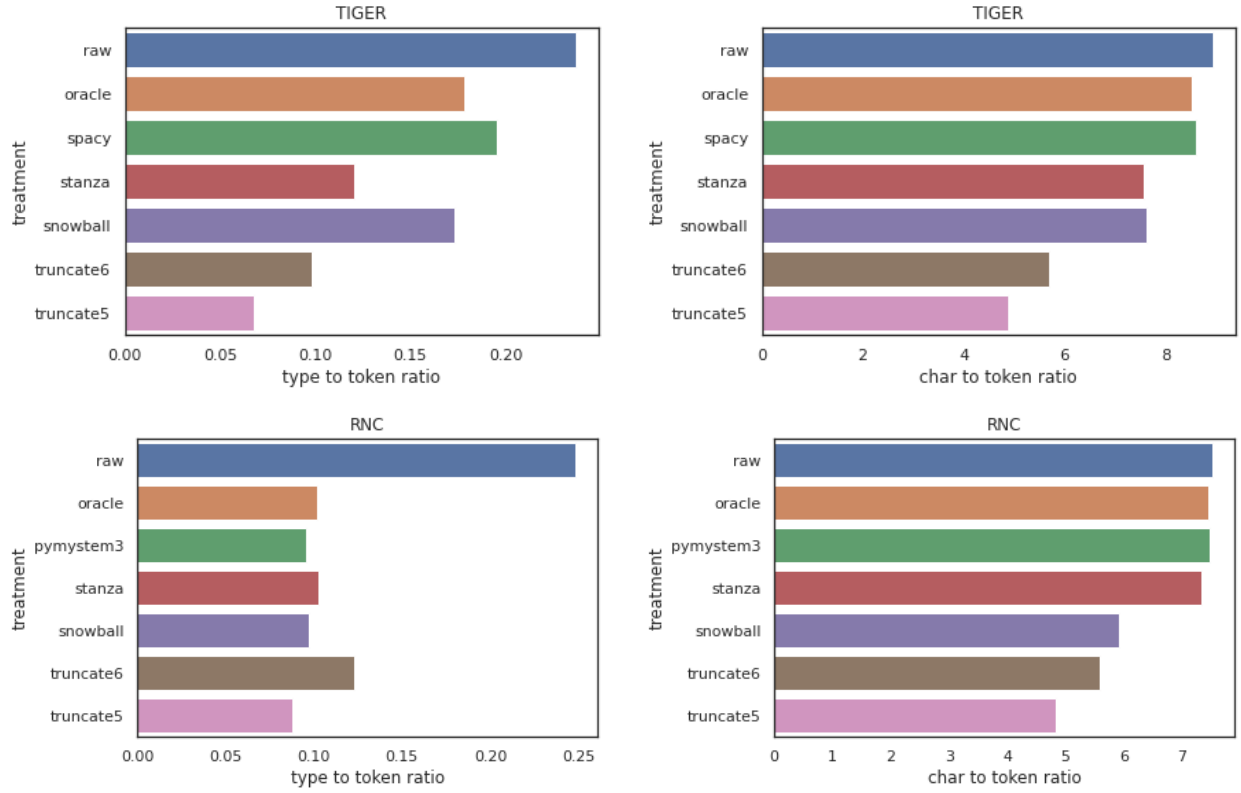
In the word-based morphology framework, inflection is captured by triples consisting of the surface form (also called wordform) $w$, a lexeme signifying the meaning and a slot $\sigma$, which can be understood as a set of "atomic" units of morphological meaning, also called inflectional features (Aronoff, 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., 2015; 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). *Lexical ambiguity* also comes in to play when a surface form is shared by two different lexemes. For example, стали 'stali' may be a form of the verb стать, *to become*, or the noun сталь, *steel*, (Sharoff and Nivre, 2011).

## 4 Corpora

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

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. Although the Snowball stemmer conflates both derivational and inflectional forms, the vocabulary size of its output is only slightly smaller than a lemmatizer's.




morphological 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. All LDA models were trained using downcased word types. We also trained experiments with the original surface forms, later referred to as 'raw' or 'untreated' experiments. This results in seven versions of each corpus, one for each stemming or lemmatization treatment.

[3] opencorpora.org

## 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 either 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) or be pre-trained machine learning models for part-of-speech and morphological feature tagging (Qi et al., 2020). Stemming methods make fewer 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]. Figure 1 shows that Stanza conflates more than the other German lemmatization treatments, oracle and spaCy, as its type-to-token ratio is much lower. This may be a consequence changing the tokenization step or Stanza producing 'unknown' as the lemma for terms missing from its model.

## 5.2 Evaluation metrics

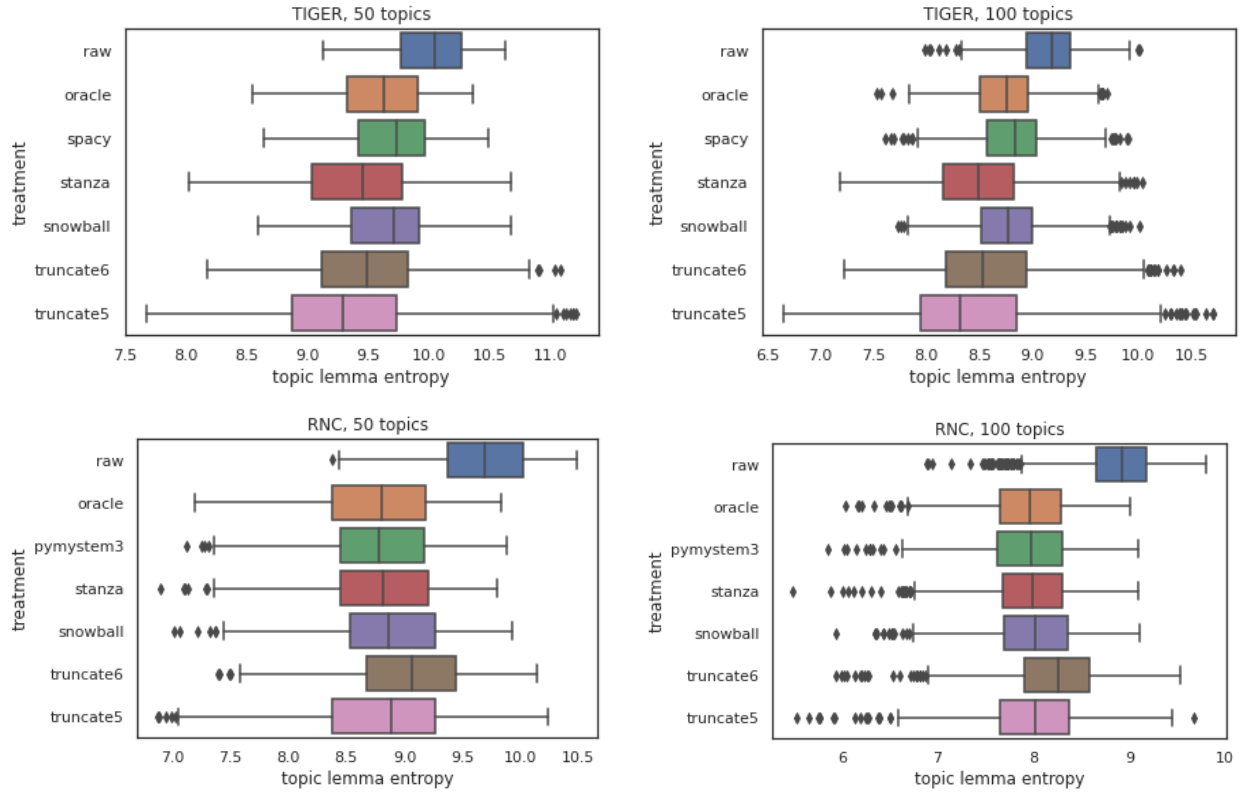
### 5.2.1 Entropy-based measurements

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 $\ell_w$ and slot $\sigma_w$. 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 $\sigma$, lemma $\ell$. 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 Author Entropy discussed in Thompson and Mimno (2018), where the morphology of the language is the metadata we

[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.




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. A slot could be any bundle of grammatical features marked by a language, from coarse part-of-speech to a token's full morphological analysis.

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

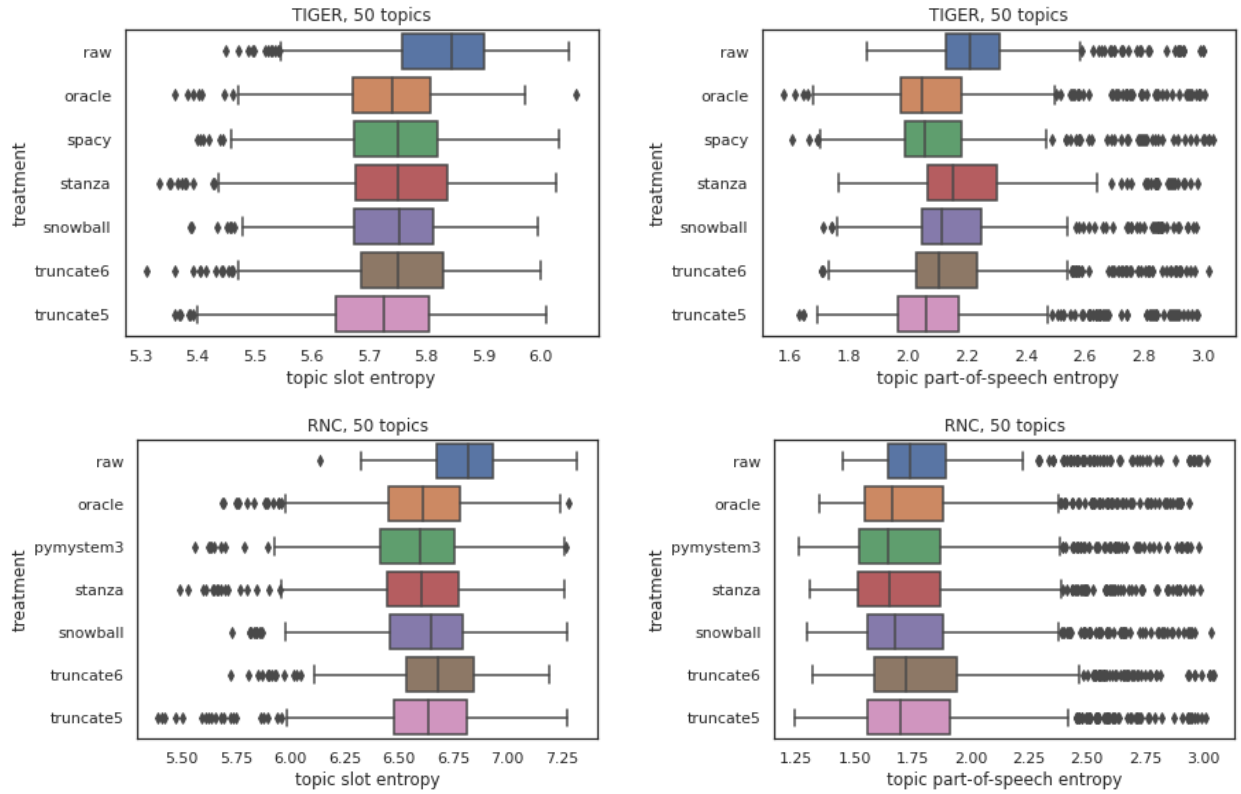
**Lemma entropy:** Similarly, we may want to know when a topic is dominated by a single lexeme, containing many grammatical forms of a single lexeme, but few other lexemes. For example, a topic may have many counts of different surface forms for each declension of a particular noun, its nominative, accustive, 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 if they reflect lexical and grammatical information known to every speaker of the language, but don't provide context 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.

Figure 3: Distribution of topic entropies computed using token allocations for full morphological slots (left) and coarse part-of-speech for 10 experiments of 50 topic models.




Similarly, a lexeme's allocations to a topic could be spread across many word forms such that no surface form of the lexeme appears in the keywords for the topic, despite the relatively high frequency of this lemma. Either of these problems occurring for many topics output by a model would suggest a need for lemmatization in post-processing.

**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. Each key term presented to the user represents a unique lexeme or multiple lexemes in cases of lexical ambiguity, determined by the annotated lemmas for each allocated instance of a term. If the set's size is closer to 1, different forms of the same lexeme are repeated in the keywords. There is no upper bound, since the level of lexical ambiguity will differ by language and corpus, but we posit this value to be around $n$ in ideal cases where post-processing is not necessary.

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

**Top $n$ lemmas:** A topic's most frequent lexemes may not always overlap with the lexemes of its most frequent surface forms. The set $L(k)$ is defined by taking the most frequent lemmas over all tokens allocated to a topic:

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

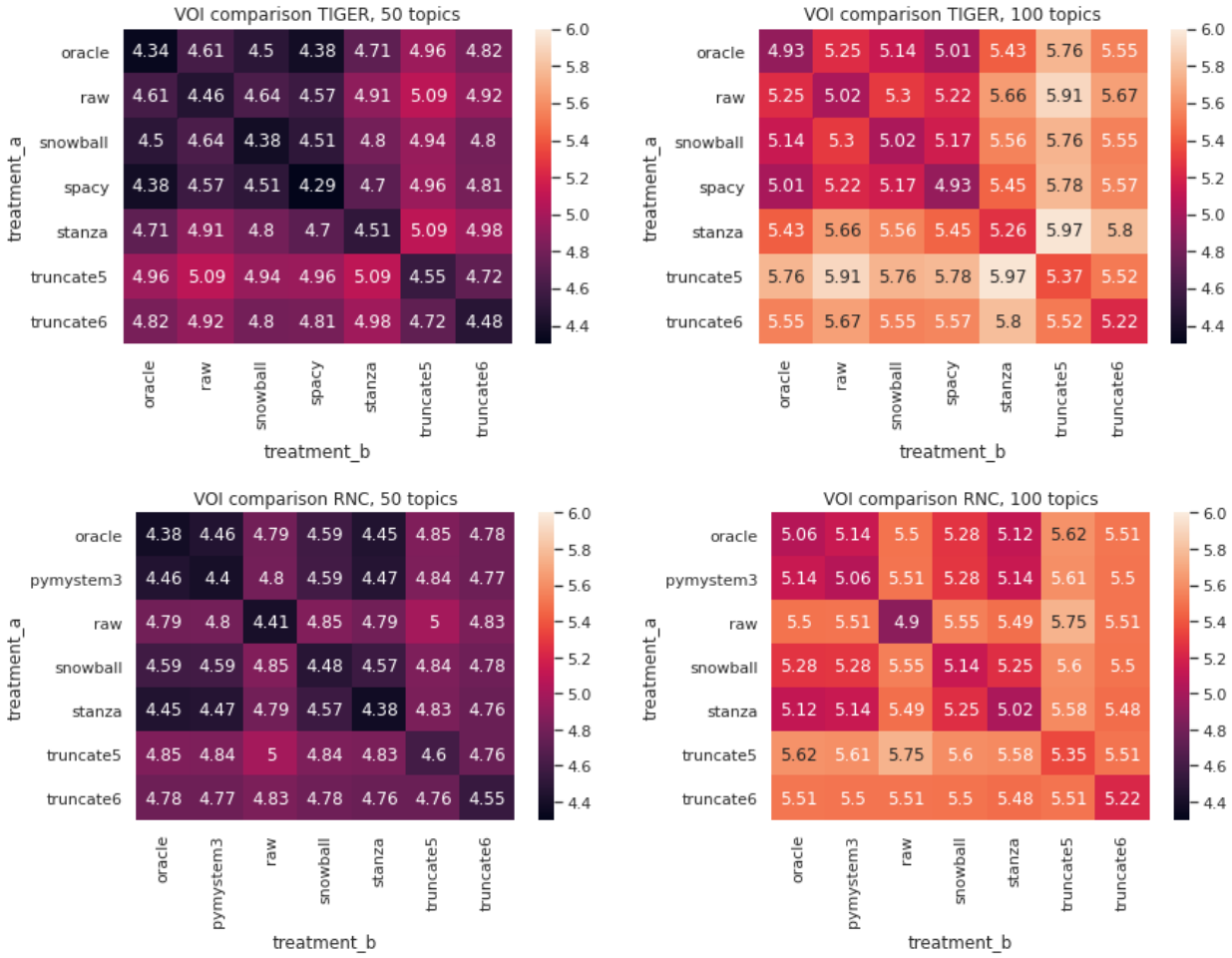
#### 5.2.3 Strength of treatment measurements

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), quantifies 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.

Figure 4: 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.




### 5.2.4 Topic Quality

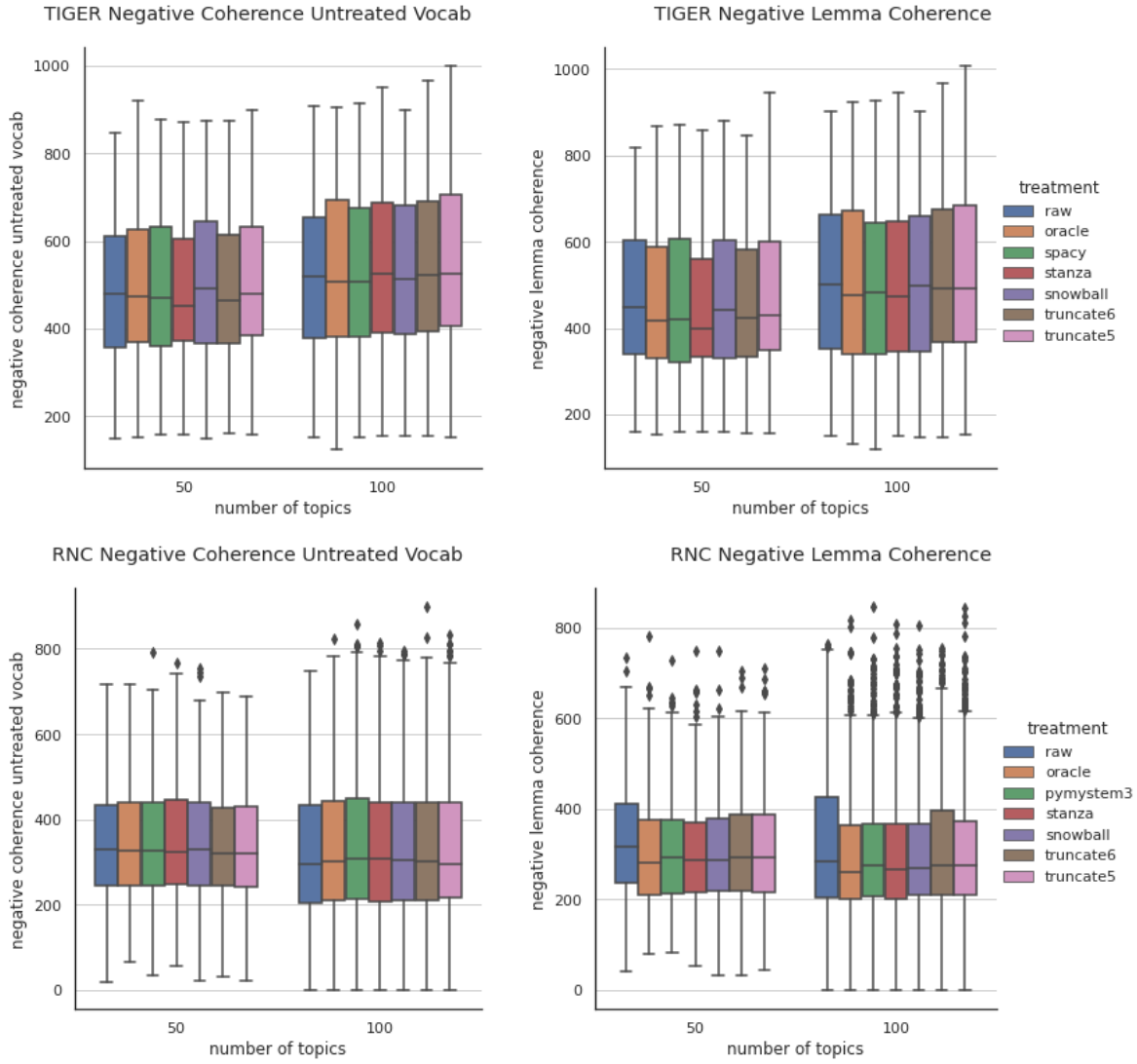
**Variation of information:** This symmetric metric allows for comparing different clusterings of the same dataset (Meila, 2003). Inherent randomness in the LDA algorithm will cause some variation across experiments, but VOI will be lower when clusterings are consistently similar over multiple runs. 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 to another treatment.

**Coherence:** An automatic metric computed from document co-occurrence of a topic's top terms, coherence 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 word 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 untreated word forms or lemmas as the vocabulary.

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

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.



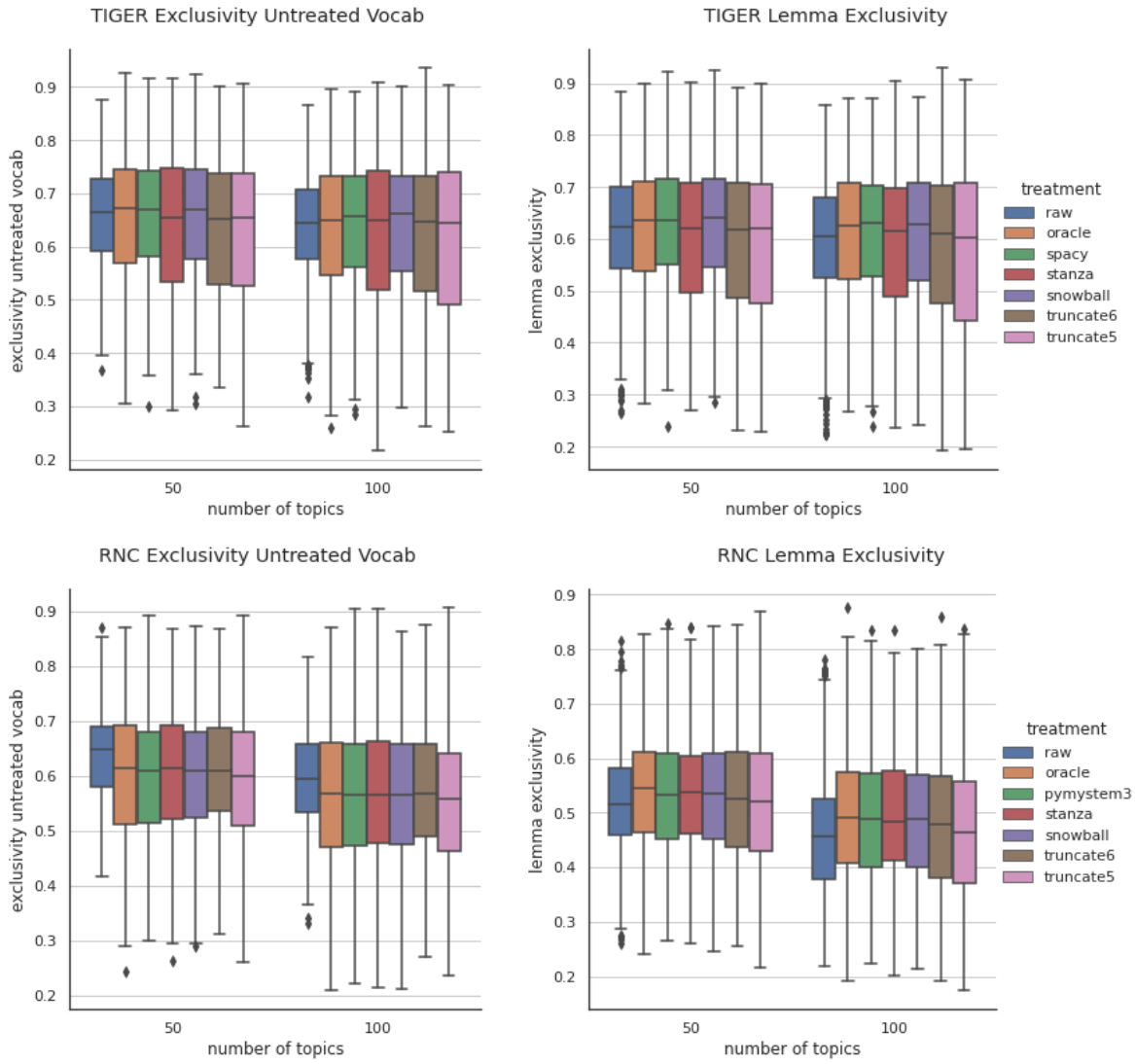

## 6 Results

### 6.1 Effects of conflation strength on morphological diversity of topics

Conventional wisdom says that lemmatization or stemming is performed in order to prevent topics from being dominated by multiple forms of a single lexeme. However, treatments with greater conflation strength actually seem to decrease the lemma diversity of topics. Ordering treatments by decreasing median lemma entropy (figure 2) exactly matches the ordering of treatments by increasing conflation strength using type-token ratio (figure 1). Pre-processing treatments decrease the diversity of topics' lemmas in similar proportion to the treatment's conflation strength.

Relatedly, the median morphological slot and part-of-speech entropies of topics also decrease under all pre-processing treatments compared to the untreated corpus, which can be seen in figure 3. Tokens that have the same lemma also belong to the same part-of-speech, so a reduction in lemma diversity corresponds to a topic having less diverse parts-of-speech. Although the topic may also be allocated more grammatical forms of a particular lexeme, the enumerative complexity of a single paradigm will always be smaller than the number of morphological slots in paradigms across all parts-of-speech in

Figure 6: 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.




the language. These metric values reflect that less diverse lemmas means fewer grammatical forms are available to a topic.

### 6.2 VOI and topic stability

Our findings regarding pre-processing treatments' impact on topic stability closely match Schofield and Mimno's (Schofield and Mimno, 2016), as shown in figure 4. A method's intra-treatment VOI is always lower compared to its inter-treament VOIs, and aggressive truncation methods have less stable cluterings than other treatments. We also see VOIs increase when training models with more topics.

Schofield and Mimno reported that light stemming (e.g. removing 's', Krovetz) improved stability of topic models on four English corpora. Those light stemmers and the WordNet lemmatizer frequently produced clusterings that were similar to the untreated corpus' allocations as well. Our results mirror these observations almost identically in German and to a lesser extent in Russian. On the German TIGER corpus, the oracle lemmatization, spaCy and the Snowball stemmer produce clusterings that are similar to each other, and they are also more similar to the untreated corpus results than Stanza or truncation treatments (as we noted earlier, Stanza's lemmatization may be problematic for the TIGER corpus). These treatments had the

lowest intra-treatment VOIs for German. Likewise, the RNC models have the most similar allocations between the lemmatizers and Snowball, but more significant differences between the untreated corpus and truncation methods. For 50 topics, the Russian lemmatization was as stable as no treatment, but this did not hold for 100 topics.

Lemmatization appears increase topic consistency compared to no treatment for German, but results for Russian are inconclusive. However, any improvements in intra-treatment VOI are slight and appear to be affected by the accuracy of the lemmatizer. Given these observations, pre-processing in order to gain slight improvments to reproducibility may not be justified due to the added computational and implementation costs of lemmatization.

### 6.3 Pre-processing and topic quality

Lemmatization also seems to improve the quality of topics for the TIGER corpus, based on coherence and the exclusivity of top terms, although there are no similar improvements seen with the Russian National Corpus. Negative coherence scores for each set of experiments are given in figure 5, where it can be seen that oracle lemmatization and spaCy slightly improve median coherence compared to the untreated TIGER corpus for both 50 and 100 topics. Truncation to 5 characters and Stanza lemmatization appear to improve coherence for TIGER with 50 topics, but this does not hold for 100 topics. Snowball and truncation to 6 characters have no effect on coherence.

Additionally, figure 6 shows that oracle lemmatization and spaCy also increase the median exclusivity of topics' top terms on TIGER, even on the untreated terms. This is evidence that lemmatization in pre-processing can help LDA know which topics to assign rare word forms to in a way that better respects document co-occurrence and increases the relative uniqueness of topics' top terms.

In contrast, the coherence scores on the RNC are nearly all the same, regardless of treatment and all treatments reduce median exclusivity on topics' terms compared to the untreated corpus. The small number of documents in the corpus may also be counfounding results.

Unsurprisingly, pre-processing treatments increase lemma coherence (figure 5) and lemma exclusivity (figure 6) on both TIGER and the RNC. However, increased lemma exclusivity may be an unavoidable affect of conflation, and it's not clear that lemma coherence would meaningfully correlate with human judgements on topic quality.

### 6.4 Utility of entropy metrics

The intention of lemma and slot entropy metrics was to help identify two types of potentially problematic topics in models trained using the untreated corpus. First, topics with few lemmas, but many repeated morphological forms of those lemmas, which would appear in the upper left corner of a graph plotting lemma entropy against slot entropy as in figure 2. Second, topics that encode grammatical information, where many lemmas are covered, but only appear with certain grammatical forms. These kinds of topics would appear in the lower right corner of the graph. Pre-processing would prevent their formation and any post-processing would obscure the grammatical information, hurting interpretability.

However, these hypotheses do not hold on closer examination of topics with extreme entropy values. Appendix A shows topics C and H have low slot entropy and high lemma entropy, but still have repeated grammatical forms of lemmas in their top terms. The interpretability of these topics may be improved by lemmatization in pre- or post-processing.

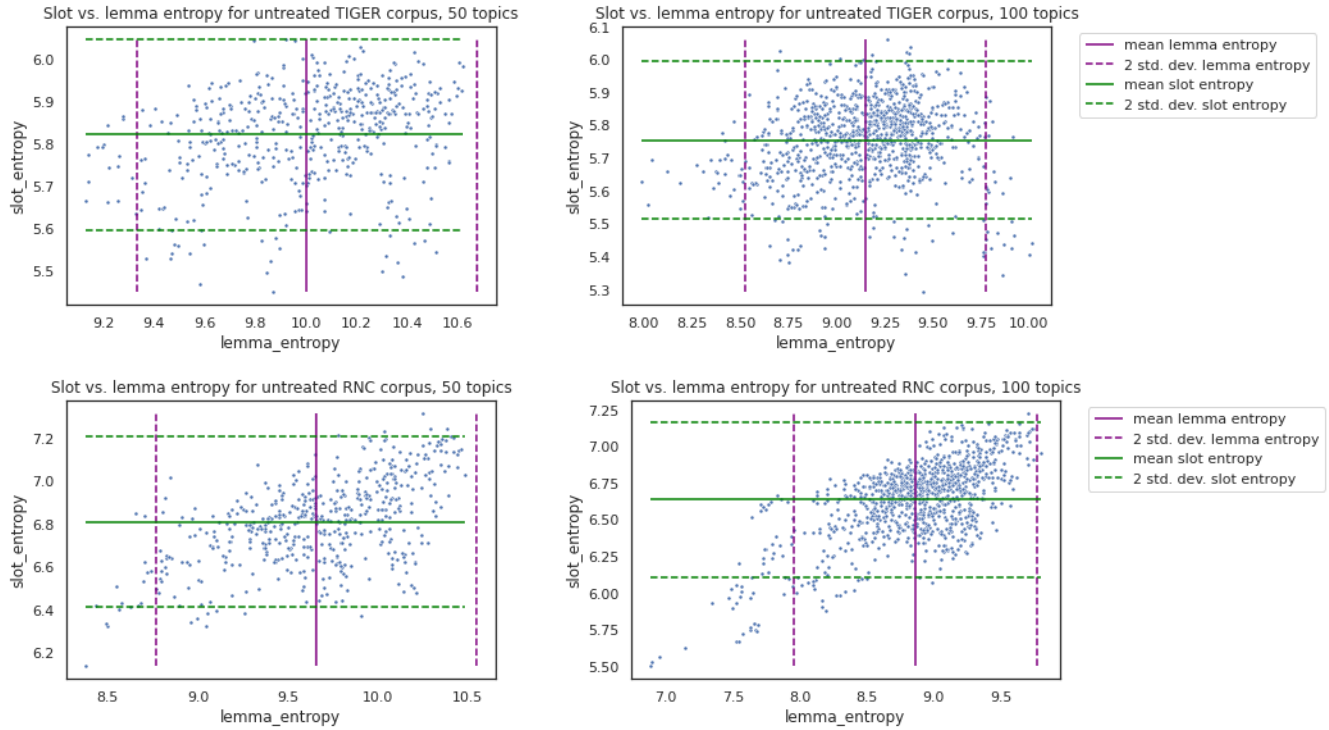
Additionally, topics with low slot entropies tend have proper names or acronyms as their top terms (A, B, F). This may be an effect of the way the annotation schema expresses morphological features for these parts-of-speech or even the result of other linguistic phenomena, such as stylistic conventions of the genre or salience determining which lemmas are more likely to appear as nominative subjects.

In any case, slot and lemma entropy do not appear to fulfill their hypothesized usefulness in quantifying topics' morphological diversity. Nonetheless, slot entropy may still have utility at finer granularity if used to target a specific morphological feature. For example, understanding the spread of 1st-person verbs may be a goal in stylometric analysis, but this is out-of-scope of these experiments.

### 6.5 Top lexemes and interpretability

Returning to the question of how morphology affects interpretability, we consider differences between the set of top lemmas of a topic, $L(k)$, and the set of lemmas expressed by the top terms, $K_\ell(k)$, since users typically work with lists of $n$ keywords identifying each topic. We lay out an approach for

Figure 7: 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 within experiments for the same corpus and number of topics.




visualizing these set differences for a group of topics in figure 8 and provide specific topic examples in appendix B, considering whether the presentation of these topics could be improved using post-processing. Recall that the $|L(k)| = n$ always, but $|K_\ell(k)|$ can theoretically be as small as 1 and its upper bound is determined by the language's lexical ambiguity across surface forms, but in practice it is frequently close to $n$. The distributions of $|K_\ell(k)|$ on topic models for the untreated corpora are shown in figure 9.

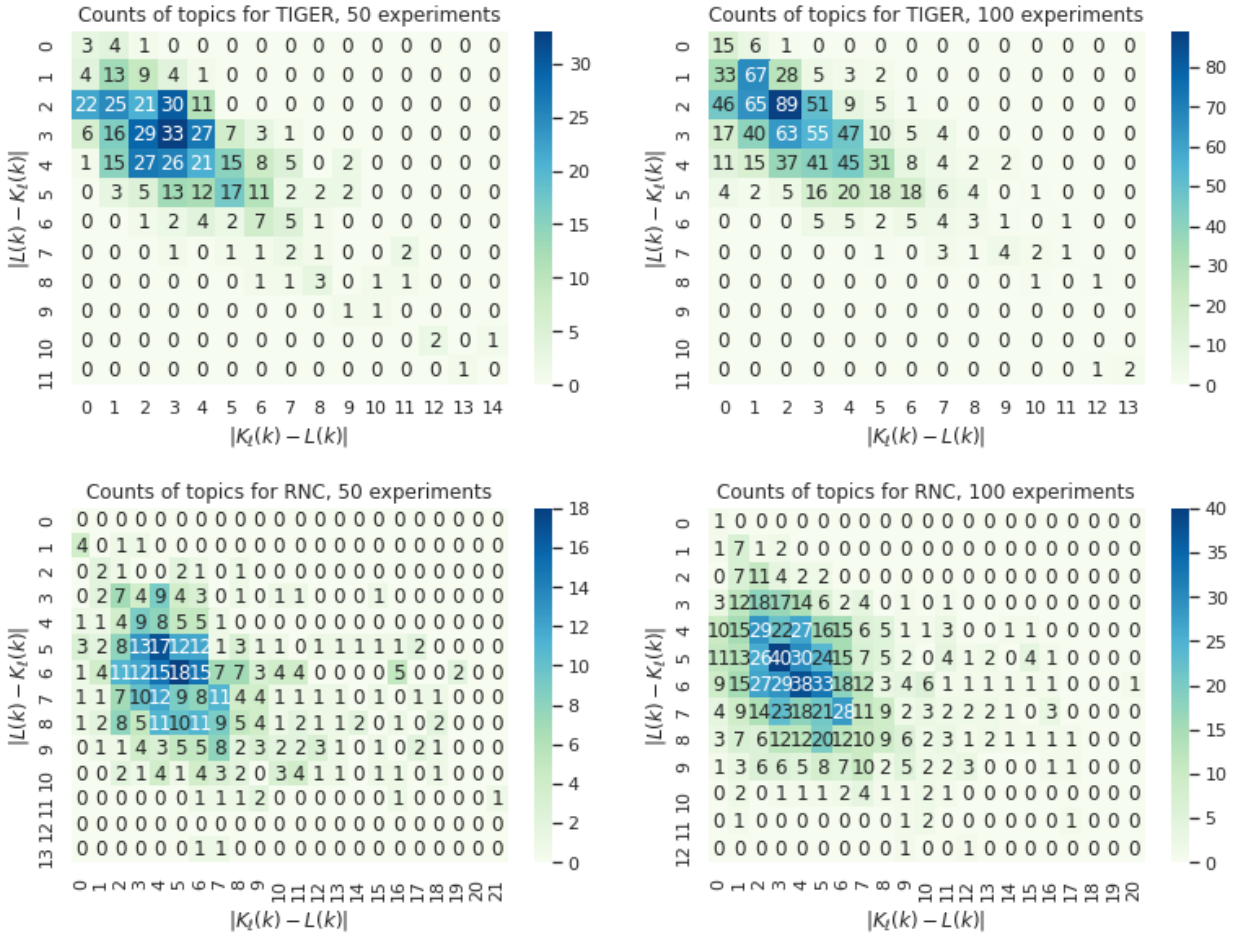
When $L(k) \cap K_\ell(k)$ is close to $n$, then there's a strong match between the top lemmas and the lemmas of the top terms. In other words, the set differences are small, so the top word types provide sufficient information for identifying the topic on their own. When many topics fall into this category, visualized in the top left of the heatmaps, there may be little motivation for lemmatization or stemming in pre-processing. The TIGER corpus experiments have more topics concentrated in this region than the RNC experiments.

If $|L(k) - K_\ell(k)|$ is small compared to $n$, while $|K_\ell(k) - L(k)|$ is large, this indicates that there aren't many repeated forms of lemmas in the top keywords, but that lexemes important to the topic may be obscured by having their instances spread across multiple rare forms. Counts of topics with this problem appear in the top right of the heatmap. Post-processing may be a solution to revealing these hidden lexemes to users.

Post-processing can be also applied to make topic presentations more concise in the opposite case, when $|L(k) - K_\ell(k)|$ is large and $|K_\ell(k) - L(k)|$ is small. This occurs when the top terms have repeated surface forms of the same lemmas, shown by topic counts in the lower left of the visualizations. Both the concentration of topics counts in the visualization and qualitative analysis of topics' key terms produced by the untreated corpus suggest this is a larger problem for RNC than for the TIGER corpus.

When both set differences are large, topic counts in the heatmaps' bottom right, both repeated grammatical forms and obscured lemmas could be occurring simultaneously. Additionally, top terms may be lexically ambiguous, meaning they belong to different lexemes depending on context. This also happens frequently with abbreviations. Consider that *mr* in most contexts would be the title *Mr.*, but *magnetic resonance* in medical settings. If the topic model doesn't learn to separate these contexts, then

Figure 8: 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 have repeated forms of the same lemmas, good candidates for post-stemming. Values in the lower right indicate large mismatches between those sets, a challenge for topic interpretability.



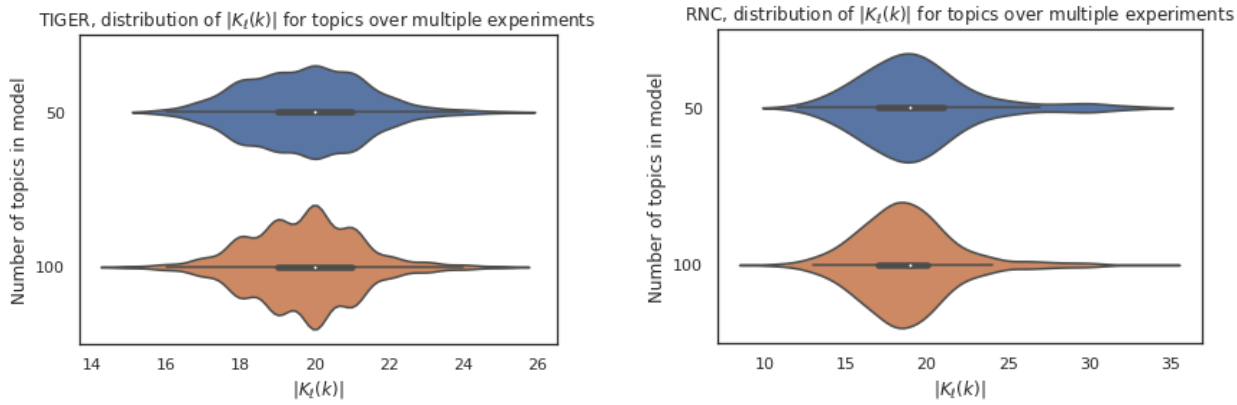

$|K_\ell(k)| > n$ and $|K_\ell(k) - L(k)|$ is affected as well. We see examples of this kind of lexical ambiguity in both TIGER and RNC topics. Post-lemmatizing the keywords in isolation would not resolve this type of lexical ambiguity, but lemmatizing the original documents to disambiguate, then computing top lemmas using the token allocations may provide a solution.

The size of $K_\ell(k)$ and set differences between $K_\ell(k)$ and $L(k)$ provide heuristics for determining which topics are candidates for post-processing. Although models on untreated RNC and TIGER corpora both exhibit topics with keyword lists complicated by morphology and lexical ambiguity, the problems are more pronounced for the Russian National Corpus. It's unclear from our limited experiments whether this pattern is intrinsic to the language or due to the particulars of these corpora.

## 7 Conclusions

In this work, we reproduced methodologies for measuring the effects of lemmatization and stemming in pre-processing on LDA topic models for a Russian and German corpus, comparing conflation strength and topic quality across 7 treatments. Additionally, we introduced several new metrics to quantify how much a topic is impacted by complex morphology. Our slot entropy metric seems too influenced by the size of languages' morphological paradigms for various parts-of-speech to be useful, but the counting-based metrics, top lemmas and lemmas in the top key terms, may be used to identify topics where post-lemmatization would improve the user-friendliness of the topic in some way.

Figure 9: The distribution of $|K_\ell(k)|$ when $n = 20$ for topics models trained on the untreated corpus over 10 experiments. Note that this value is has a larger range for the Russian corpus than the German one, but both have averages close to 20. The distribution is not significantly changed when the number of topics changes.




Our findings show that pre-processing treatments generally do not increase the diversity of lexemes expressed in topics in either language. In terms of model quality, more aggressive stemming preprocessing doesn't improve the quality of topic models for either corpus, but lemmatization may slightly improve topic consistency and coherence for German, while results for Russian are less conclusive. In contrast, topic models trained on the untreated RNC would benefit from post-lemmatization to improve conciseness and reveal more context in the key terms presented to users, but there's little motivation to doing so the German TIGER corpus.

Avenues for future work include expanding analysis to other languages, especially agglutinative ones and those outside the Indo-European family, or investigating if morphological slot entropy can applied to highlight stylometric or genre-specific features of a corpus.

## 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. 2015. *Understanding and Measuring Morphological Complexity*. 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.

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*.

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

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

## A Realization of Slot and Lemma Entropies in Sample Topics

Table 3: Example topics from 50 topic models on the untreated TIGER corpus with extreme values for lemma entropy and slot entropy.

| ID | Top 20 Key Terms | Top 20 Lemmas | Lemma Entropy | Slot Entropy | Comment |
|---|---|---|---|---|---|
| A | helmut mannheim schröder rudolf kohl lafontaine bundestag partei rau parteichef gerhard parteitag wahl sozialdemokraten spd scharping oskar cdu antrag delegierten | helmut mannheim schröder stellvertretend rudolf lafontaine sozialdemokrat bundestag partei ministerpräsident parteichef delegierter parteitag wahl spd scharping vorsitzend oskar cdu antrag | 9.58 | 5.47 | This topic has low lemma and slot entropies. It a clear topic about German politics in the mid-1990s, dominated by proper names of politicians and abbreviations for political parties. |
| B | regierung präsident demokraten kandidatur ellemann-jensen republikanern kongreß clinton washington bill gingrich republikaner kandidaten präsidenten powell us-präsident dollar lubbers partei usa | regierung kandidat präsident kandidatur ellemann-jensen demokrat kongreß clinton washington bill perot gingrich republikanisch republikaner powell us-präsident dollar lubbers partei usa | 9.26 | 5.7 | Very low lemma entropy, but moderate slot entropy. A topic about politics in the United States where proper names are quite obvious, but common nouns and adjectives appear more frequently than in A. |
| C | leute fast bleibt leben art gilt lassen meisten schließlich sehen eher stehen deutschen mal läßt steht alten menschen sogar land | bringen führen liegen leben halten alt lassen nehmen finden sehen deutsch stehen wissen mensch gelten zeigen grund müssen bleiben sogar | 10.36 | 5.5 | High lemma entropy, but low slot entropy. This appears to be a general topic with many terms that are likely common in German news articles, like *people, German, finally,* and many frequent verbs in 3rd person singular/plural or infinitive forms, *stand, see, live, leave*. Stemming verbs may make the topic's presentation more concise, but the top lemmas don't add new information about the topic. |
| D | regisseur stück bühne liebe uraufführung oper schauspieler theater konzert gruppe aufführung musik inszenierung ensemble publikum produktion karin szene komponisten frankfurt | regisseur abend stück bühne uraufführung gervaise komponist oper schauspieler theater aufführung bartók musik inszenierung kleist musikalisch ensemble publikum szene frankfurt | 10.49 | 5.82 | A topic with high lemma and slot entropy about opera and stage theater performances. The key terms are nearly all common nouns, with some proper nouns. |
| E | länder sowjetunion staaten milliarden rußland dollar japan gipfel ländern internationalen firmen usa ausländische mexiko iwf welt westlichen münchen liberalisierung land | international rußland westlich dollar staat japan milliarde gipfel firma usa mexiko organisation iwf russisch welt ausländisch kredit münchen liberalisierung land | 9.92 | 6.05 | A topic about international relations and finance with moderate lemma entropy and high slot entropy. Forms of *country* are repeated, but topic is otherwise identified by country names, common nouns and adjectives, such as *western, foreign, international*. Post-lemmatization may help here as it reveals *summit* and *credit*. |

Table 4: Example topics from 50 topic models on the untreated Russian National Corpus with extreme values for slot entropy and lemma entropy.

| ID | Top 20 Key Terms | Top 20 Lemmas | Lemma Entropy | Slot Entropy | Comment |
|---|---|---|---|---|---|
| F | петровна надежда мать прохожий алексей руки данильцев иннокентий наталья терехова анатолий adobe демидова солоницын алла звонит маргарита лиза смоктуновский игнат | открывать рука петровна надежда муж мать алексей данильцев иннокентий наталья терехова анатолий adobe демидова алла маргарита лиза смоктуновский звонить игнат | 8.38 | 6.14 | A topic with low lemma and slot entropy. This topic is dominated by proper names, mainly those of Soviet film actors, with some common nouns. |
| G | говорим шмаков сказала алеми первую женщин женщины екатерина мужчины жизель е ж сегодня данилова женщина лахова мужчин м ганапольский очередь | шмаков алеми журнал екатерина жизель метр говорить е сказать женский ж мужчина сегодня данилова женщина лахова м ганапольский очередь первый | 8.71 | 6.84 | This topic has low lemma entropy, but moderate slot entropy. Top terms include several grammatical forms of *woman* and *man*, proper names and some verbs. The topic is slightly incoherent, but may be related to women in Russian politics. |
| H | распространения ru клиентов компании компания проекта рынке сми рынка сети информации издания сбыта xgi adobe магазинов рекламы изданий интернет пользователей | ru издатель вирус компания пользователь издание магазин информация покупатель сайт клиент данные реклама распространение рекламный интернет продажа рынок товар сеть | 9.91 | 6.37 | A topic with low slot entropy and moderately high lemma entropy, which clearly about internet business and media. The terms are mainly common nouns, abbreviations and company names. Despite low slot entropy, forms of *company* and *market* are repeated in top terms. Post-lemmatization reveals useful lemmas, *buyer*, *website*, *selling*, *product*. |
| I | коротаев иль поэта народ начал читал стихотворение рубцов коля рубцова марья знал николая писателей николай витя стихи чтоб дело поэт | читать коротаев иль народ стихотворение река рубцов коля вологда стол писатель вологодский марья начать николай витя стихи чтоб бог поэт | 10.36 | 7.19 | This topic has high slot and lemma entropy. It is mostly nouns and verbs about poetry, but also has several proper names and common terms. There are no repeated lemmas in the top terms. |
| J | про недавно сколько добрый что-нибудь пожалуйста концерт покупатель покажите песни продавец поёт борнео песня девушка переписи михаил скажите радио стоит | про недавно сколько книга что-нибудь добрый пожалуйста концерт покупатель сказать продавец борнео песня петь показать девушка перепись михаил задорнов радио | 9.78 | 7.21 | A topic with high slot entropy and moderately high lemma entropy. It's fairly incoherent with common terms like *please* and *something*, but there are several terms related to music and sining: *concert, sing, song, radio*. There are no repeated grammatical forms of the top lemmas. |

## B Topics to Compare Top Lemmas with Lemmas in Top Terms

Table 5: These are sample topics from 50 topic models trained on the untreated TIGER corpus demonstrating how the difference between $K_\ell(k)$ and $L(k)$ can be used to identify topics as candidates for post-stemming. Bold marks lemmas unique to the respective set and italics mark lexically ambiguous terms. Matching color indicates surface forms may share lemmas.

| Top 20 key terms | $\lvert K_\ell(k)\rvert$ | Top 20 lemmas, $L(k)$ | $\lvert L(k) - K_\ell(k)\rvert$ | $\lvert K_\ell(k) - L(k)\rvert$ | Comment |
|---|---|---|---|---|---|
| **sogar** gilt **trotz fast deutschland eher weg** land menschen **gesellschaft** künnten *frage* steht deutschen ***meisten*** sieht **politik** lassen bleibt **arbeit** | 24 | **nehmen führen problem** können **politisch** sehen land halten gelten deutsch **stellen zeigen** bleiben **müssen neu geben** stehen frage lassen mensch | 14 | 10 | The words in this topic are mostly common, general terms with many verbs present in both top terms and top lemmas. This is likely an example of frequent verbs being spread across many grammatical forms where none of those individual forms are prominent enough to be a key term. |
| frage deutschland weg deutschen stellen **europa** politik politische gesellschaft sieht probleme menschen ***folgen*** könnten lassen **bevölkerung rolle** staat steht politischen | 24 | frage deutschland weg stellen mensch politik **land halten** gesellschaft **stark** problem **hoch** deutsch stehen **sozial** lassen können politisch staat sehen | 4 | 9 | This topic about general European politics seems slightly more interpretable under the top terms than the top lemmas. The set of lemmas adds the common verb *hold*, while excluding some terms like *Europe* and *population* that add context. The term *folgen* is lexically ambiguous between the verb *to follow* and the noun *consequence*. |
| frauen universität patienten krebs **alter körper** zellen studie bakterien arten krankheit apo menschen melatonin usa mensch medizin forscher wissenschaftler **licht** | 20 | frau universität patient krebs alter bakterie art studie **biologisch** krankheit zelle **risiko apo** melatonin usa mensch medizin forscher wissenschaftler arzt | 3 | 3 | This topic is appears to be about medical research, with terms like *university, cell, cancer, illness, researcher*. The top terms and lemmas present slightly different lexemes, but it's not clear if one group is more user-friendly. |
| rabin frieden ministerpräsidenten ministerpräsident israelischen tod mord rabins palästinenser israelis friedensprozeß jerusalem **ermordung** yitzhak peres israel amir arafat israels israelische | 16 | israelisch rabin frieden ministerpräsident tod **syrien** mord **attentäter** palästinenser **palästinensisch** friedensprozeß israeli jerusalem yitzhak peres israel **arabisch** amir arafat **jüdisch** | 5 | 1 | Both top terms and top lemmas clearly show this topic to be about the Israeli Palestinian conflict. There are some repeated grammatical forms in the key terms and the lemmas add additional context through the lexemes for *Syria, assassin, Palestinian, Arabic, Jewish*. Both sets clearly express the idea of the topic, but post-lemmatization could be used for conciseness. |

Table 6: These are sample topics from 50 topic models trained on the untreated Russian National Corpus demonstrating how the difference between $K_\ell(k)$ and $L(k)$ can be used to identify topics as candidates for post-stemming. Bold marks lemmas unique to the respective set and italics mark lexically ambiguous terms. Matching color indicates surface forms may share lemmas.

| Top 20 key terms | $\lvert K_\ell(k)\rvert$ | Top 20 lemmas, $L(k)$ | $\lvert L(k) - K_\ell(k)\rvert$ | $\lvert K_\ell(k) - L(k)\rvert$ | Comment |
|---|---|---|---|---|---|
| **стороны** ***д*** время **образом именно кроме** более **достаточно** является **наиболее** например работы часть ***т*** других **прежде** части можно решение между | 30 | время **любой результат** более **отношение процесс задача система иметь условие** другой работа **качество** часть **возможность** можно являться число решение между | 11 | 21 | The terms of this topic appear quite general, but the top lemmas reveal lexemes related to work, systems and processes, *relation, task, process, system, conditions, quality*, giving more context that this topic could be about work. There are ambiguous abbreviations, т and д, in the top terms, which lemmatization in post-processing could disambiguate. |
| обсе нато год россия войны лукин **территории** стран климов терроризмом **деньги** страны посадили америка **израиль** войска американцы **говорят** мнения война | 19 | **организация** обсе нато говорить американец год россия война лукин **тюрьма** мнение войско **ходорковский** климов **чечня военный** терроризм америка страна **конфликт** | 5 | 6 | This topic about international conflict is interpretable from the top terms, but an argument could be made for post-lemmatization. The key terms contain some common words, *money, say*, that are not present in lemmas. The top lemmas contain lexemes for *military, conflict, prison*. |
| **л** авторы **руси** литература стругацкие культуры русской **века** русские литературы **советской** интеллигенция фантастика литературе интеллигенции автор фантастику книг стругацких писателей | 14 | **журнал роман** книга **литературный** литература **запад** писать русский **восток проза** культура интеллигенция писатель **свобода** фантастика **интеллигент** автор **читатель** русь стругацкий | 10 | 4 | Interpretability for this topic about literature would likely be improved by post-lemmatization. The top terms contain repeated forms of *literature, Russian* and *intelligentsia*. The top lemmas have literary lexemes that aren't present in the top terms: *magazine/journal, novel, literary, prose, reader*. Both sets have some unique entries that are more general, *century, Soviet* in keyterms, and *east, west, freedom* in the top lemmas. |