Tracing the Proliferation of Socialist Realism Doctrine in Latvian Periodicals: Case Study of "Literature and Art" and "The Flag"

Anda Baklāne¹, Valdis Saulespūrens²

¹National Library of Latvia, 3 Mukusalas St, Riga, LV-1423, Latvia
²National Library of Latvia, 3 Mukusalas St, Riga, LV-1423, Latvia

Abstract
The paper presents the results of a study on the dissemination of socio-political and aesthetic ideas of Socialist Realism doctrine in the Latvian periodicals Literature and Art (LitArt) and The Flag. Several programmatic, ideologically saturated articles that were published in The Flag in the 1940s were compared to the rest of the corpus to explore the proliferation and persistence of similar ideas in the course of following decades. Authors have employed methodologies commonly used for plagiarism detection: fingerprinting and the comparison of document similarity based on word embeddings and document similarity measures. In particular, three perspectives were used to examine the similarity and reuse of texts: comparison of matching 5-grams processed by the winnowing algorithm and comparison of documents based on the TF-IDF and Doc2Vec embeddings and cosine similarity metrics. To facilitate the analysis, the results were loaded in the open-source version of Neo4j graph database. The findings were further explored and evaluated qualitatively to identify the distribution of direct citations, frequently reused phrases and most similar documents.

Keywords
string similarity, document similarity, Latvian historical newspapers, Socialist Realism, discourse analysis

1. Introduction
This paper is part of a series of case studies aimed at researching the possibilities for implementing text similarity detection methodologies for the analysis of the collection of digitized historical newspapers of the National Library of Latvia.¹ Methods for identifying similarity in text documents have a broad range of applications, such as the detection of plagiarism [1] and studying the reuse of texts in historical newspapers [2] [3]. In this article, the authors investigate the usability of computational string and document similarity detection [4] [5] [6] in

¹In the earlier case study, the methodology of LDA topic modelling was explored; see: A. Baklāne, V. Saulespūrens, The application of latent Dirichlet allocation for the analysis of Latvian historical newspapers: Oskars Kalpaks’ case study. Nauka, tehnologii, innovacii, 2022, No.21, pp. 29-37; A. Baklāne, V. Saulespūrens, Latento Dirihle’ sadalijumu modela izmantojums laikraksta Latvijas Kareivis tematu analīze: Oskara Kalpaka gadijuma izpette, Letonica, 2022, No.47, pp. 150-166.
recognizing the dispersion of similar discourse in the corpus of historical newspapers. The case study is constructed as an exploration of the proliferation of the lexicon and utterances related to the discourse of aesthetical tenets of Social Realism in two Latvian cultural periodicals – The Flag and Literature and Art (hereafter LitArt).

In the field of discourse analysis, the examination of discourse is understood as the study of texts and other semiotic phenomena taking into account a wide range of semiotic features and especially paying attention to the construction of meaning above the level of a sentence. In addition to that, discourse analysis is not a formal and purely linguistic analysis of a text and other semiotic entities: it always involves studying language in the context of society, culture, history, institutions, identity formation, politics, and power [7]. Depending on the context, discourse can be understood as (a) meaning-making as an element of the social process; (b) the language associated with a particular social field or practice (e.g. ‘political discourse’); (c) a way of construing aspects of the world associated with a particular social perspective (e.g. a ‘neo-liberal discourse of globalization’) [7] [8]. When encoded in language and text, a particular discourse can manifest itself as a characteristic vocabulary and a set of stylistic and rhetorical devices that can be identified with different positions or perspectives of societal groups or actors. Discourse can be expressed in all dimensions of a text: semantic, syntactic, and pragmatic. In this case study, authors are focusing on the semantic aspects of the discourse by studying lexical features of texts on a phrase level, as well as looking at the overall similarity of documents based on word embeddings.

Socialist Realism is an artistic style and an aesthetic doctrine that was officially sanctioned and prevalent in the Soviet Union from 1932 to the mid-1980 [9]. According to the tenets of Socialist Realism, artworks should represent the objective reality and be realistic in style (all non-realist modernist styles are condemned as formalism), typical (present condensed ideal types of characters, classes, and circumstances), optimistic and progress-oriented, pedagogical, and profess loyalty to the Communist Party and its goals [10]. The introduction and dismissal of socialist realism did not occur simultaneously in all Socialist countries (e.g., in Latvia, it was fully implemented after the Second World War when Latvia was occupied by the USSR for the second time, not in the 1930s). Although in the USSR Socialist Realism officially remained a leading aesthetic standard and ideal until the 1980s, it was contested at different times in different countries throughout the 1950s and 1960s [11]. Hence, it is relevant to explore the temporal dynamics of the discourse.

The possibility of tracing the proliferation of Socialist Realism doctrine is a theoretical problem in itself that cannot be fully tackled in this article. On the one hand, the media discourse in the USSR is believed to be characterized by parrot-like repetitions of ideas and phrases pertaining to the doctrine. On the other hand, it has been often emphasized that the development of national literary traditions is a complex process and it is not always easy to discern the direct influence of Socialist doctrine from other influences and personal beliefs of writers and literary critics [12] [13]. What is the vocabulary of Socialist Realism? Which phrases are signaling the presence of the discourse? Complete formalization of the Socialist Realism discourse for to purposes of fully automated analysis, if possible, is beyond the scope of this case study. Instead, the authors of the paper have assumed a less ambitious approach: to look at the reuse of vocabularies and phrases through the lens of individual notable articles; quantitative analysis is supported by the qualitative evaluation of the results.
Table 1
Corpus statistics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>The Flag</th>
<th>LitArt</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issues</td>
<td>607</td>
<td>2573</td>
<td>3180</td>
</tr>
<tr>
<td>Articles</td>
<td>39 406</td>
<td>79 047</td>
<td>118 453</td>
</tr>
<tr>
<td>Raw token count</td>
<td>62 835 625</td>
<td>67 098 255</td>
<td>129 933 464</td>
</tr>
<tr>
<td>Clean lemma count</td>
<td>48 324 759</td>
<td>52 051 659</td>
<td>100 376 418</td>
</tr>
</tbody>
</table>

Various approaches of computational text analysis can be employed to pursue the task of studying lexical features of a discourse, for instance: searching for specific pre-selected words and phrases in the corpus, identifying topics by means of topic modelling, measuring the similarity of documents (i.e., TF-IDF, Jaccard distance, Cosine distance, Hamming distance, Levenstein distance, or other), or searching for direct re-publications of texts, citations, and uncited reoccurrences of passages.

In the examination of *The Flag* and *LitArt*, authors have explored two approaches: (1) the analysis of matching 5-grams, based on the application of document fingerprinting performed by the winnowing algorithm [5] [14]; (2) the analysis of document similarity, based on TF-IDF [15] and Doc2Vec embeddings [16] that both were further compared by Cosine similarity [17].

Each of the measurements explored highlights a different aspect of the similarity of texts: the comparison of 5-grams is geared towards identifying the reuse of phrases, not taking into account the overall similarity of documents; measuring the cosine similarity between documents, on the other hand, allows to trace the overall similarity of documents that may stem from the similarity of topics or similarity of the vocabulary preferred in a particular discourse, or both.

2. Corpus and seed articles

Two titles of periodicals were selected for the case study – monthly literary magazine *The Flag* (*Karogs, 1940-1995*) and the weekly newspaper *Literature and Art* (*Literatūra un Māksla*, hereafter: *LitArt, 1945-1994*) that covered a broad range of topics related to literature, visual arts, and architecture. *LitArt* was a literary, artistic, and political weekly newspaper of the creative unions of the Latvian Socialist Republic (LSSR); *The Flag* was a monthly literary magazine published by the Writers’ Union of the LSSR. Both periodicals were initiated in the 1940s and retained their status as the most important sources of information on current events in literature, art, and architecture until the 1990s (for corpus statistics see Table 1).

The corpus was derived from the digital collection of periodicals of the National Library of Latvia². During the digitization, the content of periodicals has been segmented on the level of individual articles; in most cases, the titles and authors of the articles were also detected and marked providing additional features for further in-depth analysis. The corpus was lemmatized

---
²Access to the digitized periodicals is provided through the sites http://periodika.lv/ and https://lndb.lv/; corpora used in this study are protected by copyright and can be accessed for research purposes on demand: https://dom.lndb.lv/data/obj/1282468.html; https://dom.lndb.lv/data/obj/1282469.html
with the Latvian natural language processing tool pipeline NLP-PIPE\textsuperscript{3} [18] and stopwords were removed. To mitigate the effects of optical recognition errors, all one-symbol words were removed, as well as words not recognized by the lemmatizer.

To construct the case study, five articles (hereafter: 'Seeds') were selected to represent the discourse of Socialist Realism:

- Seed 2: "Padomju Latvijas rakstnieku deklarācija (Declaration of the writers of Soviet Latvia)", Karogs, No.3 (1940), pp. 323–324.
- Seed 3: Vipers, B., "Socialistiskais reālisms mākslā (Socialist realism in art) ", Karogs, No.3 (1940), pp. 437–442.
- Seed 5: Rokpelnis, F., “Cīruli sauc cīnā (Larks are calling to battle)”, Karogs, No.1 (1943), pp. 84-86.

The seed articles were selected manually aiming to include articles that, based on a subjective judgement of the authors of this paper, contain ideologically charged Socialist vocabulary. In addition to that, the selection was designed to include articles according to the following criteria: (1) articles related to both literature and visual art; (2) theoretical writings as well as practical criticism; (3) works of renowned authors who were more likely to be cited in the following years. Since the scope of the analysis did not entail systematical research of the republishing patterns between \textit{The Flag} and \textit{LitArt} and the sample was very small, all seed articles were selected only from the early editions of \textit{The Flag}.

Seeds 1, 2, and 5 are pertaining to the field of literary studies, 3–4 are related to visual art. Seed 2 – "Declaration of the writers of Soviet Latvia" – is a short manifest that itemizes the aesthetic principles of Soviet writers and pledges loyalty to V.I. Lenin, J.V. Stalin, the Communist Party, and the Soviet People. Seeds 1 and 3 are programmatic theoretical accounts of the state and principles of Socialist art written by prominent scholars in their respective fields – literary theory and criticism (Andrejs Upīts) and art history and theory (Boriss Vipers). Seeds 4 and 5 are examples of art criticism: Seed 4 provides an account of the first art exhibition organized in the Latvian USSR during the first Soviet occupation (1940-1941), while Seed 5 is a literary review that discusses a poetry book of a notorious Soviet poet Jānis Sudrabkalns titled "Larks Are Calling to Battle".

3. Methodology

Matching of identical and near-identical text strings and document similarity measures were used as a proxy to recognize the distribution of discourse similar to the discourse represented in the seed articles. After the cleaning and lemmatization, the corpus was processed to acquire

\textsuperscript{3}NLP-PIPE: Latvian NLP Pipeline as a Service. Accessible: http://nlp.ailab.lv/
### Table 2

**Top 10 of sorted n-grams in the Corpus**

<table>
<thead>
<tr>
<th>3-grams</th>
<th>4-grams</th>
<th>5-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>to be, which, that</td>
<td>highly, worker, art, accomplishment</td>
<td>supreme, latvia, council, presidium, ssr</td>
</tr>
<tr>
<td>which, about, that</td>
<td>to be, which, about, that</td>
<td>highly, worker, art, accomplishment, ssr</td>
</tr>
<tr>
<td>time, same, that</td>
<td>Latvia, soviet, writer, union</td>
<td>highly, worker, lsr, art, accomplishment</td>
</tr>
<tr>
<td>also, to be, that</td>
<td>supreme, latvia, council, ssr</td>
<td>supreme, honororary, council, presidium, diploma</td>
</tr>
<tr>
<td>to be, about, that</td>
<td>great, october, revolution, socialist</td>
<td>anniversary, great, october, revolution, socialist</td>
</tr>
<tr>
<td>to be, that, all</td>
<td>highly, latvia, accomplishment, ssr</td>
<td>supreme, honororary, council, presidium, ssr</td>
</tr>
<tr>
<td>to be, something</td>
<td>supreme, council, presidium, ssr</td>
<td>highly, latvia, art, accomplishment, ssr</td>
</tr>
<tr>
<td>to be, that, or</td>
<td>but, time, same, that</td>
<td>literature, museum, art, rainis, history</td>
</tr>
<tr>
<td>to be, that, he</td>
<td>about, same, understand, self</td>
<td>academic, ballet, opera, theater, state</td>
</tr>
<tr>
<td>but, to be, that</td>
<td>highly, artist, accomplishment, stage</td>
<td>award, supreme, latvia, council, ssr</td>
</tr>
</tbody>
</table>

Three types of embeddings (vector representations of text strings) that would enable the computational comparison of documents: (1) 5-grams processed by the winnowing algorithm, (2) TF-IDF embedding, and (3) Doc2Vec embedding. To further map the similarity of documents two distance measures were used: (1) Jaccard similarity for 5-grams and (2) cosine similarity for tf-idf and Doc2Vec embeddings. To make the similarity more accessible to the qualitative analysis, matrices of the embeddings were loaded in the open-source version of Neo4j graph database. Finally, the results of the quantitative analysis were explored, evaluated, and interpreted qualitatively, zooming in to individual cases of matching 5-grams and clusters of most similar documents. In this process, the interactive Neo4j and Plotly visualizations as well as various lists of most common n-grams were immensely helpful to perform a closer inspection: a small selection of these figures and tables is included in this article.

Following Python libraries and other tools were used to perform the computations:

- **Pandas** – data wrangling;
- **NumPy** – numerical calculations;
- **Gensim** – Doc2Vec;
- **Scikit-learn** – tf-idf similarities;
- **Plotly** – visualizations;
- **Neo4J drivers** for Python.

The **winnowing algorithm** is a variation of a fingerprinting algorithm that converts text documents into a set of hash values called “fingerprints”; the winnowing algorithm is proven to be efficient for finding matches of a set length of text strings (n-grams) [14].

---

2. Pandas - open source data analysis and manipulation tool: [https://pandas.pydata.org/](https://pandas.pydata.org/)
7. Neo4J - NoSQL graph database: [https://neo4j.com/docs/api/python-driver/current/](https://neo4j.com/docs/api/python-driver/current/)
For the 5-gram embeddings, a dictionary of n-grams of size 5 was calculated for each document of the corpus before applying the winnowing algorithm; the words in the 5-grams were alphabetically sorted, hence allowing for more flexibility in finding matches. After the pre-processing of the corpus, all text strings consisting of 5 consecutive words matching the strings found in seed documents were identified in the corpus; the instances of the matching strings were analyzed qualitatively and visualized by using the Python Plotly library.

The length of n-grams was selected based on the explorative analysis of the document frequency of n-grams (where \( n = 3-8 \)) in all documents of the corpus (Figure 1). Subjective inspection of 1000 most common reoccurring 3-grams revealed that this index does not include phrases immediately useful for qualitative analysis of the discourse of socialist realism - this data entails mainly common combinations of words (e.g., "which also is", "all this is", "he is the one", "it is about" etc.) and some mentions of organisations and events, such as "Great Patriotic War" (13th most common 3-gram), "Opera [and] Ballet Theatre" (27th most common 3-gram), "First World War" (67th most common 3-gram) etc. The indices of 1000 most common n-grams with the count of \( n>3 \) contain a much larger proportion of the mentions of named entities - organisations, events, and honorary titles of awarded writers and artists (e.g., "People’s Artist [of the] Latvian SSR", "Highly Accomplished Artist [of the] Latvian SSR"). The analysis suggested that n-grams with \( n>3 \) contain information potentially relevant to the studying of the discourse, with 4-grams most prominently featuring complete or partial mentions of named entities and 5/8-grams increasingly revealing additional elements of the discourse. For \( \text{‘n’} \) larger than 4, the reoccurrence of phrases in the Corpus drops significantly (Figure 1), however, the amount of reoccurring 4-grams is large also due to the overlap of two or three 4-grams that are layered upon the named entities that are more than 4 words long (see table 2 for top 10 most common 3/5-grams). Thus, 5-grams were selected for the case study as a middle ground;
after initial inspection, 5-grams were assumed to represent a unit of language that is suitable not only for quantitative analysis but for subjective qualitative evaluation as well which was a relevant aspect within the framework of this study.

TF-IDF and Doc2Vec embeddings were used to prepare documents for further similarity measurements. **TF-IDF or “term frequency - inverse document frequency”** is a measure of word frequency in the documents where the importance of a term is inversely related to its frequency across documents in the whole corpus. Namely, after counting the frequencies of all words, the overall importance of words that appear in the majority of documents in the corpus is reduced - based on the premise that the most frequently used words are functional words that do not carry essential information about the subject discussed in a given document [15].

\[
\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D)
\]

- **TF(t, d):** Term Frequency, the number of occurrences of term \( t \) in document \( d \)
- **IDF(t, D):** Inverse Document Frequency, calculated as

\[
\text{IDF}(t, D) = \log\frac{N}{\text{DF}(t, D)}
\]

- \( t \): the term or word of interest
- \( d \): the document in which the term \( t \) appears
- \( D \): the collection of documents
- \( N \): the total number of documents in the collection \( D \)
- \( \text{DF}(t, D) \): Document Frequency, the number of documents in the collection \( D \) containing the term \( t \)

For the TF-IDF embedding, a word vector consisting of the top 2000 word lemmas was generated for each document. The documents were further compared by using the metrics of cosine similarity. **Cosine similarity** is a type of hierarchical agglomerative clustering, the normalized dot product between two vectors representing the compared document [15, 17] that is proven to be efficient for the comparison of text documents [6]. The results were analyzed qualitatively and visualized by using Neo4j network dependency graphs.

**Doc2Vec** is an unsupervised learning algorithm that represents each document by a dense vector which is trained to predict words in the document [16]. In contrast to the TF-IDF-based embedding, Doc2Vec creates vector representations not for de-contextualised individual words but for sentences. For the Doc2Vec embedding, a size 50 vector was generated for each document after training the corpus for 20 epochs (i.e., repetitions of the training process; each epoch represents one full iteration over all of the training data). The hyperparameters used were the default parameters suggested by the gensim library and based on the experiences of the original implementers of Doc2Vec. The results were further analyzed qualitatively and visualized by using Neo4j network dependency graphs.

Each of the approaches - the analysis of 5-grams and the comparison of the similarity of documents - provide a different perspective on the similarity of the articles of the corpus, allowing one to reflect on different aspects of the proliferation of discourses.
### Table 3
Number of documents per number of matching 5-grams in each seed document

<table>
<thead>
<tr>
<th>Seed 1</th>
<th>Seed 2</th>
<th>Seed 3</th>
<th>Seed 4</th>
<th>Seed 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>172: 1 doc.</td>
<td>73: 1 doc.</td>
<td>40: 1 doc.</td>
<td>67: 1 doc.</td>
<td>43: 1 doc.</td>
</tr>
<tr>
<td>30: 1 doc.</td>
<td>68: 1 doc.</td>
<td>8: 1 doc.</td>
<td>28: 1 doc.</td>
<td>17: 1 doc.</td>
</tr>
<tr>
<td>26: 1 doc.</td>
<td>46: 1 doc.</td>
<td>2: 23 docs.</td>
<td>3: 2 docs.</td>
<td>13: 1 doc.</td>
</tr>
<tr>
<td>5: 1 doc.</td>
<td>31: 1 doc.</td>
<td>1: 544 docs.</td>
<td>2: 21 docs.</td>
<td>11: 1 doc.</td>
</tr>
<tr>
<td>2: 2 docs.</td>
<td>30: 1 doc.</td>
<td>1: 212 docs.</td>
<td>8: 1 doc.</td>
<td></td>
</tr>
<tr>
<td>1: 69 docs.</td>
<td>29: 1 doc.</td>
<td>14: 1 doc.</td>
<td>5: 5 docs.</td>
<td></td>
</tr>
<tr>
<td>13: 1 doc.</td>
<td>1: 23 docs.</td>
<td>2: 5 docs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: 15 docs.</td>
<td>1: 88 docs.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: 23 docs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4. Results

#### 4.1. Matching 5-grams

Comparison of 5-grams is focused on identifying identical or near-identical text strings. In this case study, the method proved to be especially useful for detecting direct citations and commonly used phrases (at the same time, it does not account for the broader lexicon and topics of the articles).

The analysis of the 5-grams contained in the five seed articles demonstrated that all instances of 5-gram matches consisting of five or more than five 5-grams were direct quotes of the seed articles. Matches that entailed 1 to 4 5-grams in the majority of cases contained phrases that were present in the seed articles, however, were not used as direct citations from those articles. These phrases broadly fell into four categories:

- ideological aphorisms directly related to SR doctrine - expressions attributed to V.I. Lenin and J.V. Stalin ("nationalist in form, socialist in content");
- other idioms, stock phrases with no authorship ("nation’s best sons and daughters");
- named entities ("communist party", "soviet union");
- other common multi-word expressions ("it is self-evident that").

According to the analysis of the matching 5-grams (Table 3), longer direct citations consisting of 5 or more consecutive 5-grams were relatively rare: 2 to 8 cases of text re-use in later publications. For instance, the "Declaration of the writers of Soviet Latvia" (Seed 2) was cited in eight other articles in *The Flag* and *LitArt* where the length of the cited passage varied from 13 to 73 consecutive 5-grams (see Table 3). All citations occurred in the event of the anniversary of the founding of the Writers’ Union of the Latvian USSR. Similarly, the review of the first Soviet art exhibition in the Latvian USSR (Seed 4) was directly cited two times - on the dates when the exhibition was commemorated in later years. In the case of the literary review of the poetry collection "Larks are calling to battle" (Seed 5), the cases of the text re-use were not...
Figure 2: Most popular matching n-grams: "nationalist in form socialist in content"

direct citations of the words of the critic but the citations and re-publications of the lines of Jānis Sudrabkalns’s poems that were also cited in the seed article.

In the analysis of the reuse of individual phrases, two doctrine-related 5-grams stood out among other idioms.

Firstly, it was the phrase containing the words "nationalist in form, socialist in content", found in Seed 2 - "Declaration of the Latvian Soviet Artists". In the Declaration, we find the following lines: "The work of the Soviet man, his high socialist moral, unshakeable Bolshevik stance, and submission to the communist idea – i.e. Soviet life in its entirety – needs to be vividly represented by writers in their works. These works need to be Socialist in content and nationalist in form."

In other texts of the Corpus, the phrase "nationalist in form, socialist in content" was sometimes attributed to J.V. Stalin but sometimes identified as a commonly known and accepted truth. Outside the Flag and LitArt Corpus, in the body of Stalin’s work, the phrase appears in Stalin’s speech "On the Draft of Constitution of the U.S.S.R." where we find it in the following context: "The absence of exploiting classes..., the fact that power is in the hands of the working class..., and, finally, the flourishing national culture of the peoples of the U.S.S.R., culture which is national in form and Socialist in content - all these and similar factors have brought about a radical change in the aspect of the peoples of the U.S.S.R.11. Figure 2 demonstrates the distribution of the reoccurrences of the phrase in the course of time (each bubble represents one occurrence in an article; the size of a bubble represents the length of the article).

Another phrase that stood out as a notable signal of the ideological discourse was the adage "feelings, thoughts and will of the masses" that was used in the art review "Visual arts exhibition of Latvian USSR" (Seed 4). The author of the review proclaims: "Art belongs to the people. It

must with its widest stretching roots go out into the very thick of the broadest masses. It must combine the feelings, thoughts and will of the masses and uplift them. In this article and all later cases of the text re-use in the Corpus, the phrase is attributed to V.I. Lenin; although it is one of the most popular Lenin’s aesthetic ideas, it is not sourced from the writings of Lenin himself but is known from the recollections of Clara Zetkin\(^{12}\) (see Figure 3 for the distribution of occurrences over time).

In addition to phrases that could be directly traced back to the discourse promoted by Soviet ideologues, there were idioms that signaled the presence of an ideologically inclined discourse more subtly. For instance, the phrase “nation’s best sons and daughters” could not be attributed to a particular author or source, however, it was recognizable as a rhetorical device, used to persuade and influence readers’ opinions on the matter at hand. In the examples recovered in this case study, “nation’s best sons and daughters” were usually people who have suffered in the name of the Socialist revolution and Communist future. Initially used in the “Declaration of Writers of Soviet Latvia”, the phrase mostly appears in the 1950s (see Figure 4). It is interesting to note that in the use cases in the 1990s (after the regaining of Latvia’s independence), in one instance, the phrase is used ironically, ridiculing the Soviet discourse, while in the second example, we find the same rhetorical device that is now employed to support another (Latvian nationalist) discourse - to reference people who have suffered from the Soviet deportations.

The largest number of re-used 5-grams, nevertheless, fall into the category of common multi-word expressions, such as “it is self-evident that”, phrases that contain words “it can be said that...” and “consequently...”. In contrast to the ideologically significant phrases that are mostly used in the 1950s, common phrases are evenly distributed throughout all years up until the

---

1990s (see Figure 5).

Hence, although the analysis of 5 seed articles is not nearly sufficient to generalize about overall trends, so far, looking at the matches of 5-grams of a sample, we can support the hypothesis that frequent reiteration of the doctrine-related rhetoric was typical for the writings of the 1950s and by the 1960s was already declining. The analysis of a larger number of seeds, including texts outside the current corpus, could bring more evidence to test the claim.

It is relevant to notice that rhetorically significant phrases could be shorter than those consisting of five consecutive words. Already among the titles of the seed articles, we find a high-flown three-word phrase "towards new horizons" that is frequently found in the corpus and could be studied as a signifying element of a discourse. However, as shown in the preliminary examination of most common n-grams, the index of 3-grams contains a very large proportion of commonly used word combinations, hence, it is not particularly suitable for approaches where subjective qualitative inspection is involved. However, in other types of approaches, 3-grams should be also considered.

4.2. Document similarity

In contrast to the 5-grams, the TF-IDF and Doc2Vec embeddings are more focused on the overall similarity of the lexical content of the articles and are not as useful for identifying the reuse of specific text strings (Table 4).

The TF-IDF and Doc2Vec embeddings each provide a slightly different perspective on lexicons of articles: TF-IDF is based on the frequency counts of de-contextualized words while Doc2Vec embeddings are learned by the algorithm taking into account the context of sentences. The results of all three similarity metrics (5-grams, TF-IDF, Doc2Vec) are partially overlapping.
Figure 5: Most popular matching n-grams: "it is self-evident that"

Table 4
Number of similar documents per quantiles

<table>
<thead>
<tr>
<th>TF-IDF score</th>
<th>0.44 - 0.49</th>
<th>0.49 - 0.50</th>
<th>0.50 - 0.52</th>
<th>0.52 - 0.53</th>
<th>0.53 - 0.68</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed 1</td>
<td>37</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Seed 2</td>
<td>8</td>
<td>16</td>
<td>11</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Seed 3</td>
<td>12</td>
<td>20</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Seed 4</td>
<td>12</td>
<td>17</td>
<td>11</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Seed 5</td>
<td>8</td>
<td>18</td>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Doc2Vec score</th>
<th>0.64 - 0.66</th>
<th>0.66 - 0.67</th>
<th>0.67 - 0.69</th>
<th>0.69 - 0.71</th>
<th>0.71 - 0.78</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed 1</td>
<td>15</td>
<td>11</td>
<td>7</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Seed 2</td>
<td>14</td>
<td>11</td>
<td>13</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Seed 3</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Seed 4</td>
<td>22</td>
<td>12</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Seed 5</td>
<td>13</td>
<td>15</td>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4 shows the number of similar documents, divided into quintiles (i.e., groupings of documents in five sets from lower to higher similarity levels). Although the scores of TF-IDF and Doc2Vec embeddings are not entirely congruent, there is a tendency for results to overlap.

For faster visual exploration all three similarity metric scores were uploaded into Neo4J graph database as edges connecting vertices - documents. Figure 6 shows most similar documents to all 5 seeds, including all three types of embeddings (5-gram, TF-IDF, Doc2Vec) in the network visualization. The results are filtered with the cutoff being the top quintile for the corresponding metric. The Neo4J visualization tool allows the researcher to browse the relationships and
inspect similar articles interactively - Figure 6 is a snapshot from this environment.

The qualitative analysis of the similar articles reveals that in the majority of cases, the most similar documents are discussing the same topics as the seed articles. Regarding the distribution of similar articles across time, there were differences among the seeds. For instance, the majority of similar documents related to the Seed 5 ("Larks call to battle") were published in 1940s, i.e., close to the time of the publication of the seed article, followed by several publications in 1950s and very few in the later years. It can be hypothesized that this correlates with the dynamics of the popularity and relevance of poet Jānis Sudrabkalns and this poetry collection in particular.

There are fewer similar documents in the highest quintile for Seeds 1, 2, 3, and 4, compared to Seed 5, and the articles are published at different times from the 1940s to 1990s. Seeds 3 and 4, which are both discussing visual arts, share the largest number of similar documents. Documents similar to 3 and 4 are historical and theoretical accounts of the situation of visual art in Latvia, written in the style resembling Seed 3, which is not as much propagandist but rather a scholarly article devoted to the theory of Socialist Realism in art. Seed 2, the "Declaration of the Writers of Soviet Latvia" is surrounded, first, by documents published close in time to the original document and, second, by documents issued around the time of celebrating anniversaries of the founding of the Union of Latvian Soviet Writers: to a large extent, this last set of documents overlaps with the articles that were identified as containing direct citations in the analysis of matching 5-grams.

Overall, a close inspection of the relations of each individual seed article reveals that the document similarity networks, first of all, tell stories about the situational and historical relevance of particular subjects discussed in these documents - publication of a fashionable poetry book, the founding of the Latvian Soviet Writers’ Union, the opening of the first Latvian Soviet art exhibition. It is especially pronounced amidst the most similar articles that belong to quantile with the highest similarity scores.

When casting the net more widely and looking at documents with slightly lower similarity scores, there is more variation, however, it is difficult to judge to what extent the ideologically-driven stylistic and rhetorical devices are contributing to the similarity scores - perhaps, the subject matter of the document always plays the leading role. One can argue that the distinction between the subject matter and rhetorical devices is not that important here since both aspects are inseparably intertwined in the vocabularies of discourses and a particular Zeitgeist dictates simultaneously the topics and the rhetoric. A further meta-analysis of the similarity networks of all articles would be needed to find out whether there are global patterns of document similarity that could be at least in part attributed to the discourse change.

5. Conclusions

It was found that identifying matching 5-grams reliably works for detecting direct citations. Five and more matching 5-grams indicated direct citation (2 to 8 citations for each seed article).

5-gram matches with the length of 1 to 4 consecutive matching 5-grams often entail frequently used phrases: (1) ideological aphorisms directly related to SR doctrine ("nationalist in form, socialist in content"); (2) other idioms ("nation’s best sons and daughters"); (3) named entities ("communist party", "sovi union"); (4) multi-word expressions ("it is self-evident that").
Figure 6: Documents similar to Seed documents by top quintile scores TF-IDF (>0.53) or Doc2Vec (>0.71) (Neo4j visualization)

The TF-IDF and Doc2Vec results for documents most similar to the seed documents showed almost 90 percent overlap despite different approaches: tf-idf calculation performed on the word level, Doc2Vec on the phrase level. Articles selected as most similar to the seed documents: (1) are on a similar topic; (2) contain similar ideological vocabulary; (3) were published closer to the date of the publication of the seed article in several cases.

In the current form, TF-IDF and Doc2Vec-based document similarity results are suitable for exploratory analysis, further research would be required to obtain hard evidence on the proliferation of the discourse in time and across several newspapers. Compared to the methodology aimed at finding similar documents, analysis of the reuse of individual phrases and passages across time (i.e., analysis of matching n-grams) can provide faster results that are immediately usable for qualitative research.

The authors hypothesize that both approaches selected for the case study are very promising for further usage in discourse studies. The proliferation of the discourse can be traced by identifying direct citations of notable works and dissemination of individual eminent phrases, as well as by analyzing the similarity of documents. More quantitative and qualitative analysis is required to provide evidence-based generalizations on the overall proliferation of the Socialist Realism discourse in the given Corpus. Further studies could be based either on the analysis of a considerably larger number of seed articles carefully selected from the same corpus or other sources or performed as a meta-analysis of the similarity networks of all articles in the Corpus.

Acknowledgments

The study was conducted with the support of the project "Text Analysis Methods and Tools

References


