Discovering a Shared Past: Topic Modelling of Austrian Historical Newspapers

Lucija KRUŠIĆ BROZIĆ
Institute Centre for Information Modelling and Digital Humanities
University of Graz
lucija.krusic@uni-graz at

Abstract

This study examines topics in Austrian historical newspapers published in the second half of the 19th century, specifically "Das Vaterland" and "Neue Freie Presse", which are part of the ANNO collection (Austrian National Library, 2021). The primary aim is to identify and analyse topics related to national minorities, labour, and migration in the multi-ethnic Habsburg Empire. The study outlines the process of data collection, pre-processing, and Dynamic Topic Modelling (DTM) using BERTopic. The results demonstrate that BERTopic, enriched with seed words, successfully identifies topics related to migration, national minorities, labour, and nationalistic movements. DTM revealed that historical events could be identified through the frequency of these topics at specific points in time, highlighting its effectiveness in the topic-specific corpus-building process. These results will serve as a foundation for constructing a migration corpus of Austrian newspapers, enabling further in-depth analyses in the future.

Keywords: historical newspapers, Topic Modelling, Dynamic Topic Modelling, BERTopic, Natural Language Processing

1. Introduction

After the so-called "refugee crisis" of 2015, migration became a topic of heated debate in the media (Bischof & Rupnow, 2017) and was widely researched in the fields of Digital Humanities and Natural Language Processing (Heidenreich

et al., 2019). At the time, Austria was one of the largest recipients of migrants in the EU, but this is by no means an isolated event in Austrian history. The modern and pluralistic society that Austria is today has been built over the past centuries through repeated waves of migration (Bischof & Rupnow, 2017). The refugees from Croatia and Bosnia in the 1990s, the "guest workers" from Yugoslavia in the 1960s, the Hungarians, Czechs, and Slovaks in the 1950s only tell the story of migration after the Second World War. However, migration was also an important force in the pre-World War I period, when the Habsburg Empire was multi-ethnic and characterized by internal migration.

As Bischof and Rupnow (2017) argue, making migrants and minorities visible in historical narratives is a prerequisite for the acceptance and recognition of social diversity. This study aims to contribute by shedding light on minority narratives in Austrian media during the second half of the 19th century through a quantitative analysis of periodicals popular at the time, the liberal "Neue Freie Presse" and the Catholic-conservative "Das Vaterland". This is achieved through the process of Topic Modelling, the automatic detection of hidden thematic structures in texts (Blei, 2012).

So far, Topic Modelling of migration discourse in German has been carried out on contemporary (Heidenreich et al., 2019; Czymara & Klingeren, 2021; Erhard et al., 2022) and historical newspapers (Obichler & Pfanzelter, 2021) using traditional topic models, predominantly LDA. This study contributes to the field of Digital Humanities by applying Dynamic Topic Modelling (DTM) to a historical corpus, using a BERT-based topic model.

Unlike traditional probabilistic topic models, this method considers the temporal dynamics of the data, allowing the analysis of the evolution of topics over time. The contribution leverages BERTopic (Grootendorst, 2022), a transformer-based Topic Modelling technique for DTM that outperforms traditional probabilistic methods by taking into account the context of words in a sentence.

The discovery of topics in historical newspapers can be considered the first step in the process of topic-specific corpus building (Wilson Black, 2023; Oberbichler & Pfanzelter, 2021). Such a corpus can later be used for a more detailed analysis using other techniques from the NLP family, e.g., Sentiment Analysis. The insights gained from this analysis aim to contribute to the understanding of Austria's minority and migration history and also to the field of Digital Humanities in relation to historical texts in German.

2. Related Work

Migration, the spatial movement from one place of residence to another, is a central societal characteristic that fluctuates with changing political and economic circumstances (Steidl, 2021). Austria has a rich history and tradition of migration due to its political and economic circumstances, resulting in a diverse modern society (Bischof & Rupnow, 2017). However, the concept of Austria as a country shaped by migration is often not acknowledged in contemporary media and political discourse (Bischof & Rupnow, 2017). This discourse can be openly anti-immigrant at times (Rheindorf & Wodak, 2018). Bischof and Rupnow (2017) argue that to contextualize modern attitudes towards migration, the complex history of migration must be considered, dating back to the internal migration of the multi-ethnic Habsburg Empire.

According to Steidl, Stockhammer, and Zeitlhofer (2007), the term migration was first defined in the 19th century as the crossing of administrative borders. Internal migration, referring to movements within the Austrian part of the Habsburg Monarchy, was the most common form of migration in the Habsburg Empire in the second half of the 19th century (Steidl & Stockhammer, 2007). According to Komlosy (2004), internal migration accounted for 99% of all mobility in the Habsburg Monarchy, with the majority occurring within the Austrian provinces. It was primarily short-distance, with people moving from rural to urban areas such as Vienna, Prague, Celje, and Klagenfurt. Notably, Vienna experienced a population growth of 100% between 1840 and 1880 (Hahn, 2000). During this period, inhabitants from Hungary, Bohemia, Moravia, and Southeast Europe followed specific migration patterns based on their professions and the labour market (Steidl, 2017). For instance, Steidl provides an example of young Slovenian women who moved to Gorica and Trieste in search of employment as domestic workers or nurses. These processes were encouraged by the new Austrian constitution of 1967, which guaranteed freedom of movement for people and property.

Most historical research on migration during this period relies on census records for internal migration and ship passenger records for overseas migration (Steidl, Stockhammer & Zeitlhofer, 2007) as well as migrant quarantine records (Pesalj, 2019).

However, newspapers published in this period can offer a new perspec-

tive on the experiences and attitudes towards migration and migrants in the Habsburg Empire. King and Wood (2001) note that migration in the media is often subtly referenced through economic, employment, class, and ethnic community issues rather than explicitly named.

Hahn (2000) examines the concept of foreignness and the perception of labour migrants as foreigners in the Habsburg Empire. According to Hahn, the newspapers of the 19th century, whether party or church-oriented, often portrayed immigrants stereotypically, using terms such as "coarse", "wanton", "rogue", or the "plague" (Hahn, 2000).

Okey (2007) explores the representation of South Slavs in the newspaper "Neue Freie Presse". The paper closely followed the nationalism movement and often stereotyped South Slavs as "passive" and "excitable dreamers, ready to quarrel about trivialities" and "easily roused". The study also discusses the implications of these dynamics for the broader Austro-Hungarian relationship, highlighting the newspaper's fluctuating stance on the topic of Slavic nationalism.

In more recent times, the trend of digitization of historical newspapers has enabled the expansion of existing research and the development and application of new research methods (Wilson Black, 2023). The digitized editions serve as source material for historical research and as data for quantitative research in the field of Digital Humanities. Topic Modelling is one of the methods used to study narratives and discourses in journalistic texts. It has been a popular method in the Digital Humanities for some time, allowing humanities scholars to answer specific qualitative research questions through quantitative topic models (Nikolenko et al., 2017). Stemming from the field of Natural Language Processing, it is an unsupervised method which aims to uncover hidden thematic structures in unstructured text collections through probability distribution (Blei, 2012).

Within the family of Topic Models, Latent Dirichlet Allocation (LDA) is the most frequently used in the Digital Humanities, for example to analyse classical and Enlightenment dramas (Schöch, 2017) and Enlightenment periodicals (Völkl et al., 2022). As such, it is often considered the standard Topic Modelling approach (Zhu, 2022). It is an unsupervised probabilistic model based on Bayes' theorem, where a document is considered to be a random mixture of "topics", probability-weighted lists of words that together can be interpreted as a topic or idea (Blei, 2012).

LDA has also been used to analyse migration discourse in contemporary and historical newspapers in German. Czymara and Klingeren (2021) investigate the "migration crisis" in Europe between 2015 and 2017 by employing Structural Topic modelling, a method that combines LDA with metadata to improve word assignment to topics within a corpus. They find topics such as "economy", "crime", and "terrorism" appearing in relation to migration. Erhard et al. (2021) make use of the same method to investigate immigration in German newspapers over a timeframe between 2001 and 2016, finding topics such as "arrival", "violence", and "family" related to migration. Hartnett (2019) explores discourses on *Willkommenskultur* (welcome culture) in German newspapers in the context of terms "Flüchtling" (refugee), "Ausländer" (foreigner), and "Einwanderer" (migrant) using a combination of keywords and LDA.

When it comes to historical newspapers, Oberbichler and Pfanzelter (2021) use a combination of keywords and LDA and annotation to create a training corpus of return migration from the US to Austria. They base their work on four Austrian newspapers (Neue Freie Presse, Arbeiter Zeitung, Kronen Zeitung, and Innsbrucker Nachrichten), published between 1850 and 1950. The newspapers were provided by the Austrian National Library (ONB) in the context of the NewsEye project. They identify discourses such as "enhancement of return migration", "benefits and dangers of return migration", "uselessness", and "delusion and disappointment". The discourse of "uselessness" refers to the tendency of the newspapers to highlight the notion that repatriates do not contribute to the labour market and economy of Austria-Hungary, as they believed the migrants would finally return to the US. The discourse of "delusion and disappointment" refers to the negative experiences of returnees, used in the media to avoid further emigration.

Wilson Black (2022) incorporates LDA in a similar manner into a bootstrapping approach (that also includes keyword search and annotation) to create a specialized corpus of philosophical writing in early colonial New Zealand newspapers.

These studies both demonstrate that Topic Modelling is an effective method not only for the explorative analysis of a corpus, but also for the creation of topic-specific, specialized corpora that can be used for further analysis. However, a limitation of LDA is that it does not account for the temporal aspects of journalistic texts (Marjanen, 2020).

On the other hand, Dynamic Topic Models constitute another family of topic models that enable the tracking of discourse development over a designated time period. Unlike traditional probabilistic topic models, this method takes into account the temporal dynamics of data, thus enabling the analysis of the evolution of topics over time. Marjanen et al. (2020) previously employed Dynamic Topic Modelling (using the ldaseq model, from the gensim library) to examine the emergence and development of topics in Finnish historical newspapers.

Currently, the most advanced method for DTM in BERTopic (Groostendorst, 2022) is based on BERT (Devlin, 2019), the most popular transformer model in recent years. BERTopic has previously been tested and compared with LDA and other topic models, outperforming them by 34.2% (Gan et al., 2024).

It uses contextual embeddings to capture the context of the input text, resulting in coherent topics. Although BERTopic has been widely used to analyse modern and historical texts, such as medical journals (Karabacak & Margetis, 2024), interviews with refugees (Sprenkamp et al., 2023), and forum posts (Omizo, 2024), its implementation of DTM has not yet been applied to historical German newspapers. This study uses BERTopic to apply DTM to Austrian historical newspapers, with the aim of identifying topics related to minorities, migration, and workers at different points in time in the second half of the 19th century.

3. Research Goals

So far, the migration in Austrian newspapers has not been explored using computational methods. Further, BERTopic has not yet been used to analyse historical newspapers. Therefore, this study aims to:

a) Identify topics and terms related to national minorities, labour, and migration in Austrian historical newspapers,
b) Identify the change and development of topics related to national minorities, labour, and migration over time,
c) Compare the periodicals “Neue Freie Presse” and “Das Vaterland” based on identified topics,

In order to answer the following research questions:

1) At which points in time were the topics related to national minorities, labour, and migration more frequent?
2) What are the differences in frequency of topics related to national minorities groups, labour, and migration between the two periodicals?
3) Is BERTopic an effective method to support building a topic-specific corpus?

4. Corpus

The corpus comprises issues from two periodicals, "Neue Freie Presse" (Österreichische Nationalbibliothek, 2023b) and "Das Vaterland" (Österreichische Nationalbibliothek, 2023a). These newspapers are part of the ANNO digital collection of newspapers, published by the Austrian National Library. The ANNO repository offers both scanned images of newspapers and the XMLs produced through the process of Optical Character Recognition (OCR).

These specific newspapers were chosen to constitute the corpus due to their timeliness and pivotal function in influencing and reflecting public discourse on migrants and national minorities.

"Das Vaterland" was a Catholic-conservative daily newspaper published in Vienna between 1860 and 1911. It was largely funded by the Bohemian aristocracy with the aim of propagating a monarchist and "specifically German, not Slavic identity" (Schachenmayr, 2018). The paper had a small readership of 5000 subscribers in 1898 and comprised only three to five pages.

The "Neue Freie Presse" was one of the most important newspapers published between 1867 and 1914. It was published twice daily and reflected the political opinions of the liberal bourgeoisie. It was largely centrist, anti-clerical, and supportive of the working class during the time of the Monarchy. According to Okey (2007), the paper was the primary source of information for German-speaking Austrians to learn about the Serbian, Croatian, and Bosnian inhabitants of the Monarchy. As a result, it played a significant role in shaping the public opinion on minorities. In 1867, it had 35,000 subscrib-

ers, which increased to 55,000 in 1901, giving it a much wider following and reach than "Das Vaterland".

This study analysed a selection of 876 issues, 440 from "Das Vaterland" and 436 from "Neue Freie Presse". However, "Das Vaterland" had an average of 7505 tokens, while "Neue Freie Presse" had an average of 17028 tokens. Only the first three pages of "Das Vaterland" and the first six pages of "Neue Freie Presse" dealing with foreign and national affairs and news were selected. The chosen editions of "Neue Freie Presse" and "Das Vaterland" were published on the first day of each month. The range of "Neue Freie Presse" editions included in this study spans from 1 September 1864 to 1 December 1900, while the range for "Das Vaterland" spans from 1 September 1860 to 1 July 1897.

Table 1. *Corpus information*

Newspaper	Number of issues	Average number of tokens per issue	Start date	End date
Das Vaterland	440	7505	1.9.1860.	1.7.1897
Neue Freie Presse	430	17028	1.9.1864	1.12.1900.

5. Methodology

5.1. Corpus Preparation and Pre-processing

The first step of the methodology involved the selection of newspapers. These publications were chosen for their opposing political leanings and availability through the ANNO repository. To track topics over time, one issue per month was selected from each newspaper, covering the period from 1860 to 1900. As this study focuses solely on reporting politics and news, only relevant pages were selected, excluding other newspaper content such as advertisements, weather reports, and death notices.

The newspapers were downloaded from the ANNO online repository using their API in XML format and transformed into plain text, which is a prerequisite for training the BERTopic model. The ElementTree XML Python library was used to perform the transformation, taking into account different XML schemas (PAGE and ALTO, in this case). According to rec-

ommendations for input document size for the sentence transformer all-MiniLM-L6-v2 (Reimers et al., 2019), which is the default in BERTopic, converting documents into sentences and paragraphs is preferred in order to avoid information loss.

Therefore, the issues were tokenized, which involved splitting them into sentences with an average length of 23 tokens. Sentences containing fewer than three tokens were excluded as irrelevant for further analysis.

Additional pre-processing steps included removing German stop words (such as 'er', 'des', 'und', 'der', etc.), numbers, and artefacts.

Although OCR noise is a well-known issue that can cause irregularities in results, post-processing of noisy OCR is currently still underway. At the time of writing, only 20% of the corpus has undergone experimental post-processing. The remaining 80% will be processed in the future. The post-processing was carried out using the multilingual hmByT5 model (Schweter, 2023), which is trained on historical European newspapers for German. The model was fine-tuned for OCR post-processing using the NewsEye READ-OCR corpus (Muehlberger & Guenter Hackl, 2019), which includes pages from "Neue Freie Presse" and "Das Vaterland". In this process, manually corrected data from the corpus was paired with the noisy OCR present in the ANNO repository. The fine-tuned model, available on the HuggingFace platform (Krušić et al., 2024), was evaluated using the SacreBLEU metric and achieved a score of 86.0. This outperformed the OCR in the ANNO repository, which received a significantly lower score of 64.4. Although the model is computationally expensive and requires long processing times, it represents a significant improvement over the original OCR. In the future, the model can be further improved through hyperparameter tuning to enable post-processing of the entire corpus.

5.2. Dynamic Topic Modelling

Dynamic Topic Modelling was performed using BERTopic (Grootendorst, 2022).

The first step involved creating global topic representations by fitting BERTopic on the entire corpus, without considering the temporal aspect.

This was achieved by:

1. converting each document into an embedding representation using a pre-trained language model

For this step, the default embedding model is used, all-MiniLM-L6-v2 (Reimers et al., 2019) and the parameter "language" is set to "multilingual".

2. clustering of embeddings and dimensionality reduction

This was done using default models, UMAP for clustering and HDBSCAN for dimensionality reduction. If desired, these models can be further fine-tuned.

3. creating topics from the clusters using c-TF-IDF (Grootendorst, 2022)

c-TF-IDF is the BERTopic adaptation of the TF-IDF algorithm to work at the cluster level rather than the document level, where each cluster is converted into a single document. This step can be further refined by adding seed words (Grootendorst, 2023). This allows domain-specific words to be weighted higher and used more frequently in topic representations.

In this process, two distinct BERTopic topic models were created and trained on texts from two newspaper titles in the corpus: one model for issues of "Neue Freie Presse" and another model for issues of "Das Vaterland". Each model was developed using seed words to increase the likelihood of discovering relevant topics. The seed words, inspired by Hartnett (2019) and Oberbichler and Pfanzelter (2021), included: 'Fremd' (foreign), 'Flüchtling' (refugee), 'Arbeit' (work), 'Einwanderer' (immigrant), 'Zuwanderer' (immigrant), 'Tschechen' (Czechs), 'Polen' (Poles), 'Kroaten' (Croats), 'Serben' (Serbians), 'Bosnien' (Bosnia), 'Slowaken' (Slovaks), 'Slowenen' (Slovenians), 'Türken' (Turks), 'Italiener' (Italians), and 'Deutsche' (Germans). Furthermore, the parameter seed_multiplier was set to 10 to amplify the significance of the seed words.

The second step involves Dynamic Topic Modelling, which incorporates the temporal aspect. This is achieved by multiplying the term frequency of documents at each time step with the previously calculated global IDF values, resulting in a local representation of each topic (Grootendorst, 2022).

To ensure that the topic representations are not negatively impacted, Grootendorst (2003) suggests using no more than 100 unique timestamps. The corpus consisted of 491 unique timestamps, which were divided into 60 equal-sized bins using the 'nr_bins' parameter in the 'topics_over_time' function.

This two-step approach allows for a comparison between newspapers at specific time points and ensures the identification of relevant topics.

5. Results

The Dynamic Topic Modelling (DTM) using BERTopic identified various topics across two Austrian historical newspapers, "Das Vaterland" and "Neue Freie Presse" covering the period between 1860 and 1900. The focus was on detecting topics related to minorities, labour, and migration.

5.1. Das Vaterland

Table 2 shows the ten most frequent topics detected in "Das Vaterland". The analysis revealed a range of subjects, with notable emphasis on Germans, the French, religion, and the army. The newspaper's Catholic-conservative political leaning is reflected in the detected topic of religion. The participation of Austria in significant military conflicts is reflected in the frequent appearance of the army-related topic. It is unsurprising that 1475 documents represent the topic of the army, given Austria's participation in the Austro-Prussian-Italian War (1866) and subsequent military defeat (Wawro, 1992). Dynamic Topic Modelling enables successful identification of this event through the increased frequency of this topic in 1866 (see Image 1). This war is also reflected in topics "Germans" (see Table 2) and "Italy" (refer to Table 3). The data includes representation of the Franco-German war (1870-71) through the frequency of topics "French" and "Germans" (refer to Table 2) and spikes in their frequency around 1866 and 1870 (refer to Image 1). The topic of "Elections" is a frequent topic, appearing in 914 documents. One frequently used term in this context is 'suffrage', which emerged in the late 19th century (Österreichische Nationalbibliothek, 2023).

Table 2. *Das Vaterland—most frequent general topics*

Topic Label	Most frequent terms in the topic (English translation)	No. of texts
Germans	'german', 'prussia', 'german', 'germany', 'prussian', 'prussian', 'berlin', 'german', 'austria', 'prussia'	2164
French	'france', 'paris', 'french', 'frances', 'french', 'frenchs', 'france', 'italy', 'german', 'parisian'	1994
Religion	'catholics', 'church', 'catholic', 'catholic', 'catholic', 'ecclesiastical', 'clergy', 'priest', 'state', 'work'	1739
Army	'army', 'soldiers', 'infantry', 'troops', 'military', 'artillery', 'regiments', 'regiment', 'battalions', 'battalion'	1475
Elections	'elections', 'voters', 'election', 'votes', 'candidates', 'electoral body', 'vote', 'voters', 'suffrage', 'electoral reform'	914
Russians	'russia', 'poland', 'russian', 'russian', 'russia', 'italy', 'russians', 'pol', 'petersburg', 'russian'	739

Consistently, there was a high frequency of topics dealing with minorities, labour, and lack of work found in the corpus (see Table 3). The topic of education is considered a migration-related topic as it also includes the terms 'German' and 'language of instruction'. During the latter half of the 19th century, nationalism gained momentum in the Habsburg Monarchy, and one of the primary objectives of national minorities was the introduction of minority languages in schools. In 1876, the administrative court of Austria permitted the establishment of 'nationality schools', where minorities were taught in their native languages (Prokopovych et al., 2019). The prominence of the topics "Czechs" and "Poles" (1220) and spikes in their frequency around 1865 and at the end of the century are consistent with their demands for constitutional rights. Hungary is a prominent topic throughout the period, with 916 documents dedicated to it. This period saw the Austro-Hungarian compromise of 1867, which facilitated the resolution of the Hungarian crisis and the development of dual Austria-Hungary (Heka, 2017). According to Komlosy (2004), this period saw the expansion of migration networks and short-term employment in construction, agriculture, and forestry. The topic "Jews", containing the term 'anti-Semitism' is prominent starting from 1880 (refer to Image 2), which could be connected with the permeation of anti-Semitism to Austrian politics after the stock market crash of 1873 (Britannica, 2024). The full table with a list of all the topics in "Das Vaterland" is

openly available[1].

Table 3. *"Das Vaterland"—Frequent topics related to migration*

Topic Label	Most frequent terms in the topic (English translation)	No. of texts
Education	'german', 'school', 'teacher', 'schools', 'elementary school', 'lessons', 'elementary schools', 'children', 'language of instruction', 'middle schools'	1955
Poles, Czechs, and Croatians	'polish', 'czechs', 'germans', 'prague', 'czech', 'prague', 'polish', 'young czechs', 'young czechs', 'croats'	1220
Hungary	'hungary', 'hungarian', 'hungarian', 'hungary', 'budapest', 'nation', 'magyar', 'budapester', 'crown', 'pol'	916
Jews	'Jews', 'German', 'Jewish', 'Jewish', 'anti-Semitism', 'Jew', 'anti-Semites', 'Israelite', 'Christian', 'Christians'	695
Labour	'work', 'worker', 'working time', 'workers', 'hours', 'employer', 'work', 'manufacturer', 'children', 'wage'	489
Minorities (especially Turkish)	'Turks', 'Serbs', 'Turkish', 'Turkey', 'Bosnia', 'Pol', 'Turkish', 'Pascha', 'Armenians', 'Jews'	471
Italy and Croatia	'italian', 'italian', 'italiens', 'italian', 'italians', 'italian', 'rome', 'veneto', 'croats', 'italians'	405
Lack of Work	'work', 'none', 'impossible', 'rarely', 'mistake', 'little', 'unknown', 'unfortunately', 'games', 'hardly'	394
Croatia	'croats', 'serbs', 'croatia', 'croatian', 'croatian', 'croatia', 'fiume', 'slavonia', 'croatia', 'slavonian'	163
Bulgaria	'bulgaria', 'bulgarians', 'bulgars', 'bulgarians', 'pol', 'bulgarians', 'turks', 'bulgarias', 'kaulbars', 'exarchs'	160

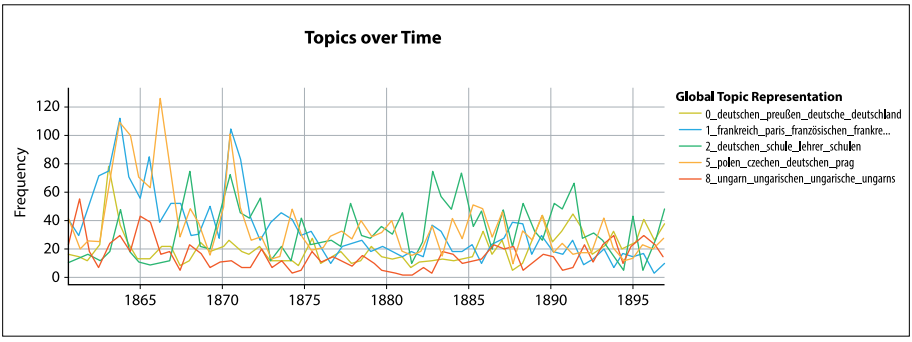

Figure 1. *"Das Vaterland"—Topics over time* (0_Germans, 1_French, 2_Education, 5_Poles, Czechs and Croats, 8_Hungary)

[1] Full table containing topics detected in "Das Vaterland": https://docs.google.com/spreadsheets/d/1ypehEMVLk8eug6GG0XMXbjr1a63pv-3lqxpD6nDqmS4/edit?usp=sharing

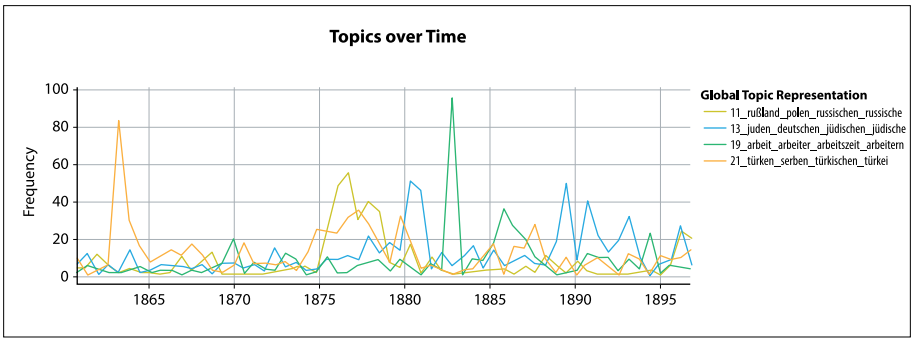

Figure 2. *"Das Vaterland"—Topics over time* (11_Russians, 12_Jews, 19_Labour, 21_Minorities)

5.2. Neue Freie Presse

The detected topics are largely similar to those found in "Das Vaterland", e.g., those dealing with education (although the term 'language of instruction' is missing), the army, Czechs and Poles, Hungary, Russians, and elections. Unlike in "Das Vaterland", religion is not a prominent topic. Instead, "Neue Freie Presse" contains a large number of topics dealing with crime, royals, the railway, and Russians (see Table 4). As this was a period marked by the expansion of railway, industrialization, and internal migration (Komlosy, 2004), the number of documents dedicated to building the railway are to be expected. The frequency of topics surrounding Russia and Russians is consistent with the Three Emperors' League event in 1873 (see Image 3), which constituted an agreement between Austria-Hungary, Russia, and Germany regarding the Balkan development and Russia's further involvement (Britannica, 2024).

Table 4. *"Neue Freie Presse"—most frequent general topics*

Topic label	Most frequent terms in the topic (English translation)	No. of texts
Crime	'accused', 'judge', 'crime', 'arrested', 'indictment', 'court', 'convicted', 'punishment', 'witness', 'law'	4064
Army	'army', 'troops', 'military', 'soldiers', 'revolver', 'infantry', 'artillery', 'battalions', 'corps', 'military'	3306
Royals	'king', 'majesty', 'prince', 'emperor', 'queen', 'prince', 'empress', 'royal', 'king', 'emperor'	2389

Topic label	Most frequent terms in the topic (English translation)	No. of texts
Railway	'railway', 'train', 'train', 'railroads', 'trains', 'locomotive', 'trains', 'wagons', 'work', 'station'	1567
Elections	'elections', 'voters', 'vote', 'choice', 'suffrage', 'vote', 'vote', 'electoral reform', 'electoral body', 'candidates'	1531
Russians	'Russia', 'Russian', 'Russian', 'Russians', 'Russen', 'Moscow', 'Poland', 'Petersburg', 'Russian', 'German'	1441

Regarding the development of migration topics over time, the topic of "Hungary" is fairly consistent over the time period, with one spike between 1865 and 1870 (see Image 3). This is consistent with the wave of internal migration (Komlosy, 2004; Hahn, 2000), and the identified spike is consistent with the passing of the Austro-Hungarian compromise of 1867 (Heka, 2017). Furthermore, the topic of "Czech and Poles", which spikes before and around 1870, is consistent with the results from "Das Vaterland" and follows their fight for constitutional rights. Other prominent topics are of Jews and anti-Semitism, minorities (including Turkish and Southeast European minorities), Italy, and workers' rights (see Table 5 and Image 4). The full table with a list of all the topics in "Neue Freie Presse" is openly available[2].

Table 5. *"Neue Freie Presse"—Frequent topics related to migration*

Topic Label	Most frequent terms in the topic (English translation)	No. of texts
Education	'school', 'teacher', 'schools', 'elementary school', 'pupils', 'lessons', 'elementary schools', 'middle schools', 'students', 'children'	3428
Czechs and Poles	'Czech Republic', 'Poland', 'German', 'Czech', 'Prague', 'Czech', 'Prague', 'Polish', 'Club', 'Czech'	2739
Hungary	'hungary', 'hungarian', 'hungarian', 'hungary', 'budapest', 'work', 'compensation', 'magyar', 'delegation', 'poland'	2633
Jews	'jews', 'jewish', 'anti-semites', 'israelite', 'anti-semitism', 'anti-semitic', 'jew', 'jewish', 'rabbi', 'israelite'	975
Italy	'italian', 'italian', 'italiens', 'italian', 'italians', 'italians', 'italian', 'sardinia', 'crispi', 'austria'	933
Minorities (especially Turkish)	'turks', 'turkey', 'turkish', 'turkish', 'stambulov', 'pasha', 'turk', 'serbs', 'turkish', 'turkish'	732

[2] Full table containing topics in "Neue Freie Presse": https://docs.google.com/spreadsheets/d/12z9WQBK6TspXgBZrVPhMGTMNuYdGXCjE3xb_yeJMbCo/edit?usp=sharing

Topic Label	Most frequent terms in the topic (English translation)	No. of texts
Southeast European Minorities	'bosnia', 'serbs', 'serbia', 'montenegro', 'serbian', 'herzegovina', 'serbian', 'serbian', 'turks', 'belgrade'	710
Workers' Rights	'work', 'worker', 'workers', 'strike', 'workforce', 'working time', 'worker', 'employer', 'wage', 'work'	672
Romanians	'romania', 'romanian', 'bucharest', 'romanian', 'romanias', 'romanians', 'pol', 'bucharest', 'serbs', 'italy'	455
Croatia, Serbia, Slovenia	'croatia', 'serbs', 'croatians', 'croats', 'slovenia', 'croatiens', 'croatians', 'croats', 'slavonians', 'croatians'	127

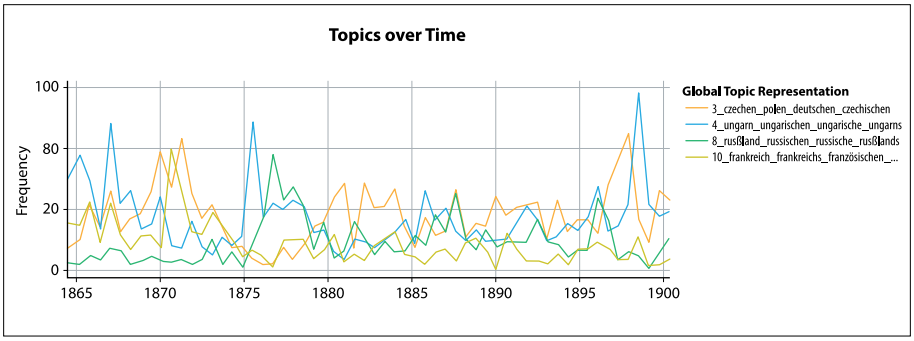

Figure 3. *"Neue Freie Presse"—Topics over time* (3_Czechs and Poles, 4_Hungarians, 8_Russians, 10_French)

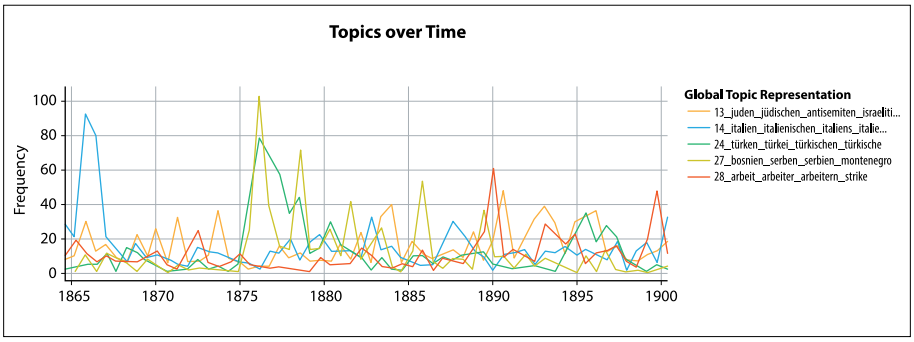

Figure 4. *"Neue Freie Presse"—Topics over time* (13_Jews, 14_Italy, 24_Minorities (Turkish), 27_Southeast European Minorities, 28_Workers' Rights)

Table 6. *Identified topics and associated texts*

Periodical	Topic Label	Text example (translated)
Das Vaterland	Poles, Czechs, and Croatians	The predominantly German regions, on the other hand, have recently shown a striking increase in the Czech working population. The reasons for this phenomenon were aptly explained by Plener himself in this year's Diet in his major speech on the occasion of Herbst's motion concerning the division of districts.
Das Vaterland	Jews	Jews living in dispersion in the Orient Occident, as far as the reports of loud experience within our circle of observation reach, have a decided aversion to the very work on which society is preferably based, namely agriculture and any craft requiring a strong labour force.
Neue Freie Presse	Southeast European Minorities	Even if necessity drives a Serb or Wallachian to work, he does not exert himself at all.
Neue Freie Presse	Hungary	Hungary, for example, where the arable farming population has been declining by millions for years. This also explains the emigration from eastern districts.

Comparing "Das Vaterland" and "Neue Freie Presse", the identified topics are similar (with the exception of religion and crime), appearing at similar time frames. A significant number of documents (labelled as -1 in BERTopic results) were identified as outliers, not fitting well into the main topics (see full tables). These documents are unlikely to relate to migration, minorities, or the primary areas of interest. Future research will involve a detailed examination of these outlier documents to understand their content and significance. However, the analysis successfully detected a large number of topics connected with minorities and work. The texts with relevant topics will serve as a basis for the topic-specific corpus building (see Table 6 for text examples). As the "Neue Freie Presse" was more represented in the corpus than "Das Vaterland", so is the number of texts containing the investigated topics higher.

6. Discussion

In this study, Dynamic Topic Modelling (DTM) using BERTopic, a transformer-based architecture, was employed to identify the frequency and development of topics in two Austrian historical newspapers. The focus was on

topics related to minorities, labour, and migration in the Habsburg Empire between 1860 and 1900. The methodology, which included downloading, transforming, pre-processing, and OCR post-processing the newspapers, as well as performing Dynamic Topic Modelling, resulted in two topic models enriched with the selected seed words. This approach allowed the topics to be tracked over time, ensuring a more precise and detailed study of the topics.

The seed words, focusing on terms related to migration and specific minority groups, enabled the successful identification of desired topics. As expected, "Das Vaterland" had a higher frequency of topics related to religion, while "Neue Freie Presse" had more topics related to royalty and crime. In both cases, the seed word approach yielded many desired topics related to national minorities, labour, and workers' rights. Compared to the methodologies used by Black (2023) and Oberbichler and Pfanzelter (2021), which involve keyword search and LDA in multi-stage iterative approaches, this methodology allows for the immediate integration of seed words into the transformer model. The BERT-based transformer model's capability to understand context resulted in a higher number of relevant, understandable topics. Additionally, the BERTopic's ability to output the assigned topic for each document in the corpus is advantageous and will be used in future analyses.

However, the use of seed words carries the potential risk of bias. Even with expert selection, the choice of seed words reflects subjective judgments, which can skew the identified topics. To address this, future research will include experiments comparing results with no seed words and various sets of seed words. Preliminary experiments indicate that models without seed words detect few to no migration-related topics, underscoring the necessity of carefully chosen seed words. These comparative studies will help assess the extent of bias introduced by seed words and inform strategies to minimize it.

The weights assigned to seed words, currently set to default values, also play a crucial role in the Topic Modelling process. Future research will involve systematic parameter tuning to optimize seed word weights. By experimenting with different configurations, we aim to reduce bias while maintaining topic relevance and accuracy.

Another area for future work is refining the nr_bins parameter, which divides the corpus into time slices for Dynamic Topic Modelling. It is currently set to bins of 5 to 10 years based on domain knowledge and practical considerations. While practical, this approach may not be appropriate for all corpora. Future efforts will include empirical testing and the use of specific domain knowledge to explore different bin sizes, establishing the optimal one for this corpus. Further, the identified outliers may be investigated in the future in order to understand their significance.

The infrequent appearance of the term 'migration' can be attributed to its formal definition emerging only in the 19th century (Steidl, Stockhammer & Zeitlhofer, 2007), making its usage in historical newspapers unlikely. However, as King and Wood (2001) note, migration is often subtly referenced through themes such as economic conditions, employment, social class, and community formation. Building on this, the scope of this research was expanded to include various facets of migration, including labour, education, nationalism, and minority rights. By employing Dynamic Topic Modelling, both explicit and implicit references to migration were identified. Even in the absence of specific terms like 'migration' or 'refugee', the topics related to minorities and labour identified in this study provide a foundation for analysing migration aspects and will serve as a basis for creating a specialized corpus focused on migration.

The findings of this study confirm that DTM is an effective method for analysing large historical corpora with a temporal aspect, such as newspapers. This confirms the results of previous studies in the field of Digital Humanities (Marjanen, 2020), as well as in other domains (Omizo, 2024; Sprenkamp et al., 2023). Furthermore, the results provide clear insights into internal migration and life in the Habsburg Monarchy between 1860 and 1900, illuminating minority perspectives of the time. The study successfully addressed the research questions and aims, particularly in identifying the frequency and evolution of topics related to national minorities, labour, and migration over time, and comparing the periodicals "Neue Freie Presse" and "Das Vaterland" based on the identified topics.

7. Conclusion

This study aimed to identify and analyse topics related to national minorities, labour, and migration in Austrian historical newspapers using Dynamic Topic Modelling (DTM) with BERTopic. By focusing on the period between 1860 and 1900, the results capture the temporal dynamics of the identified topics and compare their frequency in the "Neue Freie Presse" and "Das Vaterland." The methodology involved pre-processing and OCR post-processing to ensure the accuracy of the textual data. BERTopic successfully identified numerous topics related to migration, minorities, and labour, and the scope was expanded to include various facets of migration, such as education, nationalism, and minority rights.

DTM revealed that historical events could be identified through the frequency of these topics at specific points in time, with significant spikes between 1865 and 1870 corresponding to the Austro-Hungarian Compromise of 1867 and patterns of internal migration. The analysis highlighted the importance of education, particularly in relation to national minorities and their efforts to preserve their languages and cultures. Comparing the two newspapers, "Neue Freie Presse" exhibited a higher frequency of documents on these topics, covering a broader range of issues including crime and workers' rights, while "Das Vaterland" had a more focused coverage on religion and conservative viewpoints.

Overall, BERTopic proved to be an effective method for building a topic-specific corpus, providing insights into the socio-political landscape of the Habsburg Monarchy. This research contributes to the field of Digital Humanities by presenting the use of BERTopic on historical texts in Austrian German and demonstrating its effectiveness in this new context. The results highlighted differences in the way these newspapers covered issues related to minorities, labour, and migration, providing valuable insights into the historical media landscape. The texts with these topics will form a topic-specific corpus, facilitating further research and analysis. This study provides a solid foundation for future Digital Humanities research[3], contributing to a deeper understanding of Austria's socio-political history.

[3] All code and identified texts will be available in the following repository: https://github.com/lucijakrusic/TopiAnno

Acknowledgements

I would like to express sincere gratitude to Klara Venglarova and Raven Adam for their critical contributions in creating the post-processing model. My thanks also extend to Klaus-Jürgen Hermanik for his valuable assistance with the selection of seed words and to Georg Vogeler for methodological support and supervision.

References

BISCHOF, G., & RUPNOW, D. (Eds.). (2017). *Migration in Austria*. University of New Orleans Press. https://doi.org/10.2307/j.ctt1t89kvv

BLEI, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

CALLAWAY, E., TURNER, J., STONE, H., & HALSTROM, A. (2020). The Push and Pull of Digital Humanities: Topic Modeling the "What is digital humanities?" Genre. *Digit. Humanit. Q.* https://www.semanticscholar.org/paper/The-Push-and-Pull-of-Digital-Humanities%3A-Topic-the-Callaway-Turner/ac724521626695524805bfcb763caa89a04e77ed

CZYMARA, C. S., & KLINGEREN, M. V. (2022). New perspective? Comparing frame occurrence in online and traditional news media reporting on Europe's "Migration Crisis." *Communications*, *47*(1), 136–162. https://doi.org/10.1515/commun-2019-0188

DeMEESTER, T. R., & JOHNSON, L. F. (1975). Evaluation of the Nissen antireflux procedure by esophageal manometry and twenty-four hour pH monitoring. *American Journal of Surgery*, 129(1), 94–100. https://doi.org/10.1016/0002-9610(75)90174-9

DEVLIN, J., CHANG, M.-W., LEE, K., & TOUTANOVA, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

ERHARD, L., HEIBERGER, R. H., & WINDZIO, M. (2022). Diverse Effects of Mass Media on Concerns about Immigration: New Evidence from Germany, 2001–2016. *European Sociological Review*, *38*(4), 629–647. https://doi.org/10.1093/esr/jcab063

GAN, L., YANG, T., HUANG, Y., YANG, B., LUO, Y. Y., RICHARD, L. W. C., & GUO, D. (2024). Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic. In H. Lu & J. Cai (Eds.), *Artificial Intelligence and Robotics* (Vol. 1998, pp. 376–391). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9109-9_37

GLENN, J. K., & GOLDMAN, J. (1976a). Task delegation to physician extenders--some comparisons. *American Journal of Public Health*, *66*(1), 64–66. https://doi.org/10.2105/ajph.66.1.64

GLENN, J. K., & GOLDMAN, J. (1976b). Task delegation to physician

extenders--some comparisons. *American Journal of Public Health*, *66*(1), 64–66. https://doi.org/10.2105/ajph.66.1.64

GROOTENDORST, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. http://arxiv.org/abs/2203.05794

GROOTENDORST, M. P. (2003). *Dynamic Topic Modeling - BERTopic*. https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html

GROOTENDORST, M. P. (2023). *Seed Words - BERTopic*. https://maartengr.github.io/BERTopic/getting_started/seed_words/seed_words.html

HAHN, S. (2000). Inclusion and Exclusion of Migrants in the Multicultural Realm of the Habsburg "State of Many Peoples." *Histoire Sociale / Social History*. https://hssh.journals.yorku.ca/index.php/hssh/article/view/4568

HARTNETT, S. (2019). Willkommenskultur: A Computational and Socio-linguistic Study of Modern German Discourse on Migrant Populations. *Transit*, *12*(1). https://doi.org/10.5070/T7121043491

HEIDENREICH, T., EBERL, J.-M., LIND, F., & BOOMGAARDEN, H. (2020). Political migration discourses on social media: a comparative perspective on visibility and sentiment across political Facebook accounts in Europe. *Journal of Ethnic and Migration Studies*, *46*(7), 1261–1280. https://doi.org/10.1080/1369183X.2019.1665990

HEKA, L. (2017). Analiza Austro-ugarske i Hrvatsko-ugarske nagodbe: u povodu 150. obljetnice Austro-ugarske nagodbe. *Zbornik Pravnog Fakulteta Sveučilišta u Rijeci*, *38*(2), 855–880. https://doi.org/10.30925/zpfsr.38.2.7

KARABACAK, M., & MARGETIS, K. (2024). Natural language processing reveals research trends and topics in The Spine Journal over two decades: a topic modeling study. *The Spine Journal*, *24*(3), 397–405. https://doi.org/10.1016/j.spinee.2023.09.024

KING, R., & WOOD, N. (Eds.). (2001). *Media and migration: constructions of mobility and difference*. Routledge.

KOMLOSY, A. (2004). State, Regions, and Borders: Single Market Formation and Labor Migration in the Habsburg Monarchy, 1750-1918. *Review (Fernand Braudel Center)*, 27(2), 135–177. https://www.jstor.org/stable/40241597

KRUŠIĆ, L., RAVEN A. & VENGLAŘOVÁ, K. (2024). *hmbyt-NewsE-*

yeAnno (1.0). lukru/hmbyt-NewsEyeAnno

LI, C., CHEN, S., XING, J., SUN, A., & MA, Z. (2019). Seed-Guided Topic Model for Document Filtering and Classification. *ACM Transactions on Information Systems*, *37*(1), 1–37. https://doi.org/10.1145/3238250

MARJANEN, J., ZOSA, E., HENGCHEN, S., PIVOVAROVA, L., & TOLONEN, M. (2020). *Topic modelling discourse dynamics in historical newspapers* (arXiv:2011.10428). arXiv. http://arxiv.org/abs/2011.10428

MUEHLBERGER, G., & Guenter HACKL. (2019). *NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.)*. [object Object]. https://doi.org/10.5281/ZENODO.3387369

NIKOLENKO, S. I., KOLTCOV, S., & KOLTSOVA, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393

OBERBICHLER, S., & PFANZELTER, E. (2021). Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods. *Journal of Digital History*, *jdh001*. https://journalofdigitalhistory.org/en/article/4yxHGiqXYRbX

OKEY, R. (2007). The Neue Freie Presse and the South Slavs of the Habsburg Monarchy, 1867-1914. *The Slavonic and East European Review*, *85*(1), 79–104. https://www.jstor.org/stable/4214395

OMIZO, R. M. (2024). A Comparison of Topic Modeling Approaches Using Networked DiscussionForum Posts From the City-data.com Corpus. Journal of Open Humanities Data, 10, 16. https://doi.org/10.5334/johd.182

Österreichische Nationalbibliothek. (2021). ANNO *Historische Zeitungen und Zeitschriften*. ANNO Historische Zeitungen Und Zeitschriften. https://anno.onb.ac.at/

Österreichische Nationalbibliothek. (2023). *Women demand the right to vote – 1848 to 1918*. |

Österreichische Nationalbibliothek. https://www.onb.ac.at/en/more/ariadne-the-women-and-gender-specific-knowledge-portal/women-use-your-vote/women-demand-the-right-to-vote-1848-to-1918

Österreichische Nationalbibliothek. (2023a, February 24). *Das Vaterland*. ANNO Historische Zeitungen Und Zeitschriften. https://anno.onb.ac.at/info/vtl_info.html

Österreichische Nationalbibliothek. (2023b, February 24). *Neue Freie Presse*.

ANNO Historische Zeitungen Und Zeitschriften. https://anno.onb.ac.at/info/nfp_info.html

PESALJ, J. (2019). *Monitoring migrations: the Habsburg-Ottoman border in the eighteenth century* [Doctoral, Leiden University]. https://scholarlypublications.universiteitleiden.nl/handle/1887/70437

PROKOPOVYCH, M., BETHKE, C., & SCHEER, T. (Eds.). (2019). Language diversity in the late Habsburg empire. Brill.

REIMERS, N., ESPEJEL, O., CUENCA, P., & AARSEN, T. (2019). *all-MiniLM-L6-v2*. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

RHEINDORF, M., & WODAK, R. (2018). Borders, Fences, and Limits—Protecting Austria From Refugees: Metadiscursive Negotiation of Meaning in the Current Refugee Crisis. *Journal of Immigrant & Refugee Studies*, *16*(1–2), 15–38. https://doi.org/10.1080/15562948.2017.1302032

SCHACHENMAYR, A. (2018). Austrian Newspaper Coverage of the Cistercian Jubilee in 1898. *Cistercian Studies Quarterly*, *53*, 141–157.

SCHWETER, S. (2023). *hmbyt5-preliminary/byt5-small-historic-multilingual-span20-flax*. https://huggingface.co/hmbyt5-preliminary/byt5-small-historic-multilingual-span20-flax

SPRENKAMP, K., ZAVOLOKINA, L., ANGST, M., & DOLATA, M. (2023). Data-Driven Governance in Crises: Topic Modelling for the Identification of Refugee Needs. *Proceedings of the 24th Annual International Conference on Digital Government Research*, 1–11. https://doi.org/10.1145/3598469.3598470

Stadt Wien. (2023, April 4). *Neue Freie Presse*. Wien Geschichte Wiki. https://www.geschichtewiki.wien.gv.at/index.php?title=Neue_Freie_Presse&oldid=879140.

STEIDL, A. (2020). On Many Routes: Internal, European, and Transatlantic Migration in the Late Habsburg Empire. *Central European Studies*. https://docs.lib.purdue.edu/ces/2

STEIDL, A., & STOCKHAMMER, E. (2007). *Coming and leaving. Internal mobility in late Imperial Austria.* https://doi.org/10.57938/A59C74C6-D2D7-4163-ACAC-CED0071825AB

STEIDL, A., STOCKHAMMER, E., & ZEITLHOFER, H. (2007). Relations among Internal, Continental, and Transatlantic Migration in Late

Imperial Austria. *Social Science History*, *31*(1), 61–92. https://www.jstor.org/stable/40267929

VAN DAM-MIERAS, M. C., SLOTBOOM, A. J., PIETERSON, W. A., & DE HAAS, G. H. (1975). The interaction of phospholipase A2 with micellar interfaces. The role of the N-terminal region. *Biochemistry*, *14*(25), 5387–5394. https://doi.org/10.1021/bi00696a001

VÖLKL, Y., SARIĆ, S., & SCHOLGER, M. (2022). Topic Modeling for the Identification of Gender-Specific Knowledge. Virtues and Vices in French and Spanish 18th Century Periodicals. *JCLS (Journal of Computational Literary Studies)*, *1*(1). https://doi.org/10.48694/jcls.108

WAWRO, G. D. W. (1992). *The Austro-Prussian War: Politics, Strategy and War in the Habsburg Monarchy, 1859-1866.* (9315269th ed., Vols. 1 & 2). Yale University ProQuest Dissertations & Theses.

WILSON BLACK, J. (2023). Creating specialized corpora from digitized historical newspaper archives. *Digital Scholarship in the Humanities*, 38(2), 779–797. https://doi.org/10.1093/llc/fqac079

ZHAO, H., PHUNG, D., HUYNH, V., JIN, Y., DU, L., & BUNTINE, W. (2021). *Topic Modelling Meets Deep Neural Networks: A Survey* (arXiv:2103.00498). arXiv. http://arxiv.org/abs/2103.00498

ZHU, L., & CUNNINGHAM, S. W. (2022). Unveiling the knowledge structure of technological forecasting and social change (1969–2020) through an NMF-based hierarchical topic model. *Technological Forecasting and Social Change*, *174*, 121277. https://doi.org/10.1016/j.techfore.2021.121277