EuPaC Hackathon, Pilsen 2025: SCRIPTUM Dataset

TOPIC MODELLING: Samizdat & Exile Literature, 1968 - 1989

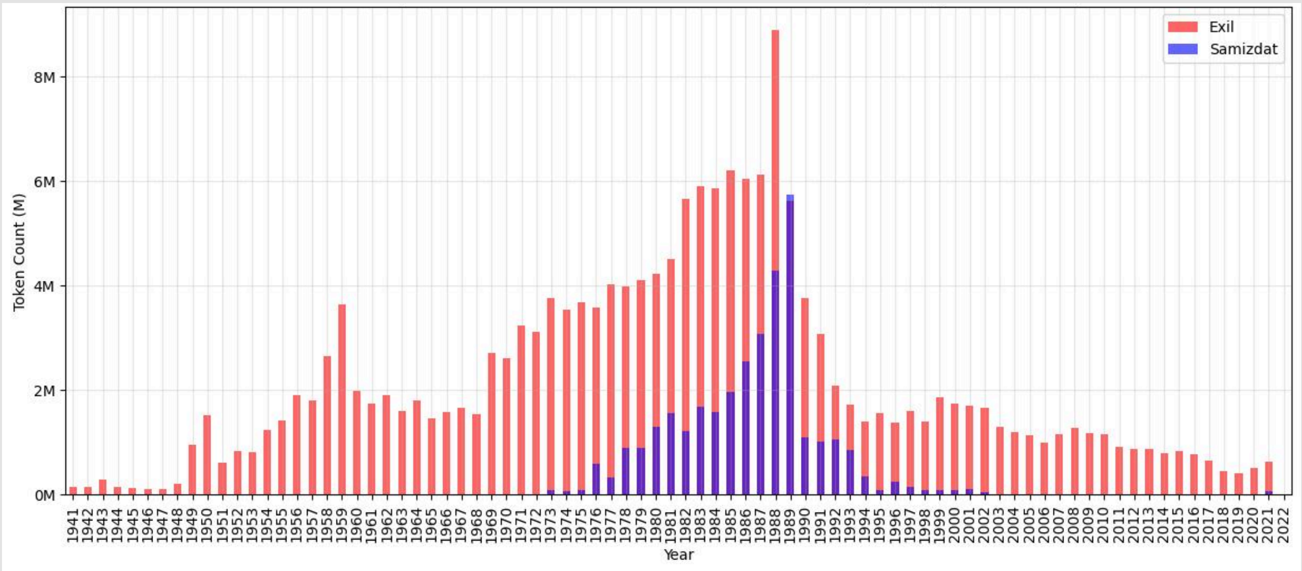
Historical Background

The Era of Normalisation:

Since 1948, the Communist party in Czechoslovakia (ČSSR) held all state power which meant nation-wide persecution, censorship, and police surveillance. In Spring 1968, ČSSR, amongst many other nations, made an effort to liberate their regimes from the strict rule of the Soviet Union. In fall 1968, the Soviet Union and other Warsaw Pact members invaded and crushed these efforts. The period that followed is known as the ***Normalisation***, where much of the ČSSR population became disillusioned about Communism and apathetic about the state of affairs. Certain individuals / groups throughout ČSSR, however, took it upon themselves to keep the Czechoslovak literature and spirit alive through **Samizdat** and **Exile** literature — texts that were illegal to produce and distribute, and thus, those involved risked persecution.

Dataset: SCRIPTUM

Fig. 1: Token Count Per Year of Samizdat and Exile Publications (1941-2021)



Note. this plot includes the texts whose publication dates were unknown but in the preprocessing of the data were recovered.

State of the Data:

- Established in 2007, and supported through the efforts of *Exodus*, a protected work group for people with disabilities.
- Consists of a large collection of Samizdat and exile journals (19th-21st century) stored in the library *Libri prohibiti*. Biggest collection of its kind.
- Single journal issues per file (pdf or djvu, 10.801 files, OCR'ed)

N.B.: the original texts were produced via type-writer and / or screen-printing, depending on the writer's / copier's access to technology and the associated risk of using it.

Samizdat = RUS: *самиздат* = "produce" + "yourself"; texts produced or distributed illegally within Czechoslovakia.

Exile / Tamizdat = LAT: *exsilium* = "to banish" / "leave (in)voluntarily"; mainly text production and distribution by Czechoslovakian exiles. Also includes text produced in Czechoslovakia and distributed abroad.

check out our topic model

Aims

1. We seek to identify and model the evolution of key topics in Samizdat and exile literature from **1968 (Prague Spring & Warsaw Pact Invasion) to 1989 (Velvet Revolution, resulting in the divorce of Czechoslovakia and signifying the end of the Communist regime**). We expect such an analysis of the corpus will elucidate what key concepts and ideas lay central to this underground movement of political dissent that has come to be so symbolic of this period in Czechoslovak history.
2. We aim to address the soundness of the common perception that **"female writers mostly didn't discuss [female topics]"** (Jiřina Šiklová) via exploratory work.

Methods of Analysis

***Data Limitations.* Addressed:** Missing publication years were filled; empty or nearly empty files were excluded; magazine titles excluded in textual data. **Not addressed:** Files are journal issues, thus a diverse set of topics are present; risk of incoherence in topic modeling. Files include bad OCR.

Overall 5 642 datapoints were used for analysis.

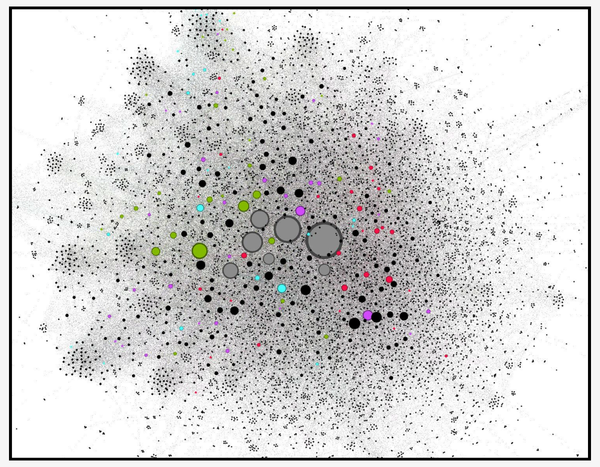
PRIMARY EXPLORATION — Topic Modelling:

- Cutting edge approach
  Semantic Signal Separation conceptualizes topics as axes of semantic space, allows for multiple topics per document and works in a multilingual setting (Kardos et al., 2025).
- (not only) Czech document embeddings
  BGE-M3, the top performing Sentence Transformer on Czech language tasks in MMTEB was used to encode the texts. Trained on multilingual data, can handle the other samizdat languages.

SECONDARY EXPLORATION — 'Female Topics' Mapping: Extraction & Lemmatisation (identifying "zen*" from the corpus), sorting collocations into categories of interest and Network Analysis of women related topics

Results

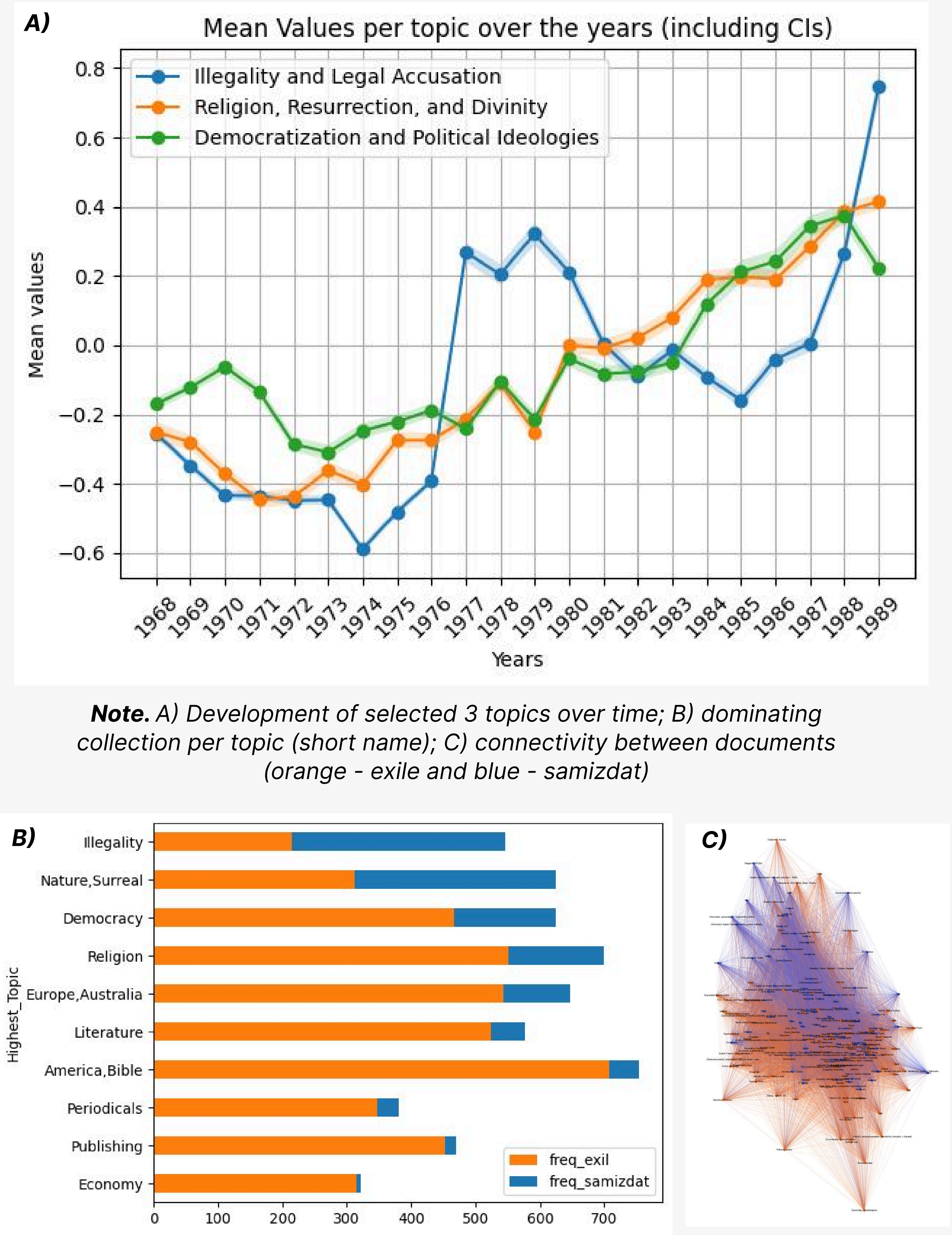
SECONDARY EXPLORATION — Representation of 'Female Topics'



- similarities in distribution of relevant collocation of, e.g., "žena", "ženský" from several domains of interest
- collocations related to political/ social issues (red) are clustered in specific journals

Fig. 2: One-mode projection of bipartite network between journals and terms surrounding root "žen" as related to women. Draft from NetworkX visualized in Gephi (ForceAtlas2)

PRIMARY EXPLORATION — Topic Modelling of Literature Types



Note. *A) Development of selected 3 topics over time; B) dominating collection per topic (short name); C) connectivity between documents (orange - exile and blue - samizdat)*

Discussion

Our primary exploration provides us with elementary insight into the key topics discussed in the texts. Our topic modelling yielded 10 topics which highlight discussions of: **democracy, justice, nature, literature, religion,** etc. Our secondary investigation of **women-related topics** found discussions of social and political questions specific to females to be clustered in specific journals.

Future endeavours working with SCRIPTUM may benefit from e.g. segmenting the journals into individual articles, using the table of contents as a guide. Additionally, it may be of use to consider what topics were discussed outside of this time period, i.e. how the socio-political situation may have shaped literary interests prior to the installation of the Communist regime as well as following its collapse. Alternative avenues for exploration include looking at the evolution of individual metaphor applications f.e. "greengrocer".

Special Thanks: Vojtěch Kaše, University of West Bohemia, Libri prohibiti, Exodus

Group Members: Eva Kvindt, Jan Kostkan, Stefanie Dach, Elena Hamidy, Aiswarya Roy, Bennet Hanke, Lui Walz, František Kalvas, Václav Steiner