Exploring interview collections with the help of named entity linking and topic classification

Júlia Egyed-Gergely, Judit Gárdos, Anna Horváth, Enikő Meiszterics, Róza Vajda
TK KDK - Centre for Social Sciences Research Documentation Centre

László Kovács, Balázs Pataki, András Micsik
SZTAKI DSD - Institute for Computer Science and Control Department of Distributed Systems

PROJECT GOAL

- Create an intuitive and explorative search interface for the social science interview archive by preprocessing it using AI/NLP tools

ASPECTS

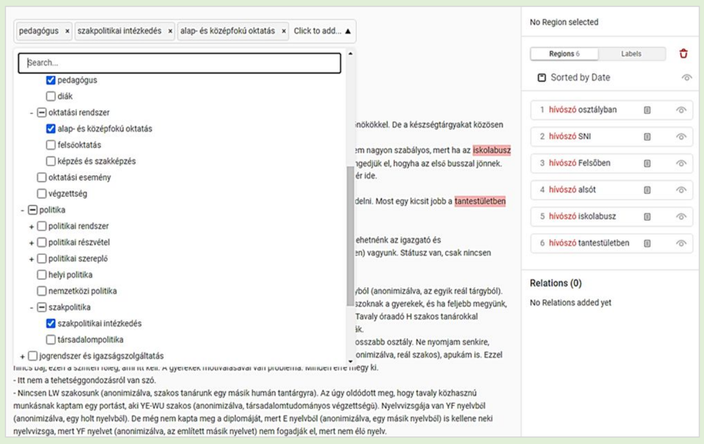
- integration of new principles (Open Science, FAIR)
- meeting current research requirements: machine-based qualitative studies
- testing Hungarian NLP tools / creating a suitable Hungarian NLP tool
- testing Named Entity Recognition tools

STARTING POINT

- 6000+ social science interviews
- Mixed formats: audio transcripts, typescripts, RTF/TXT files, etc.
- Semi-structured texts
- Very informal speech
- Wide range of themes
- Non-uniform speech transcription methods

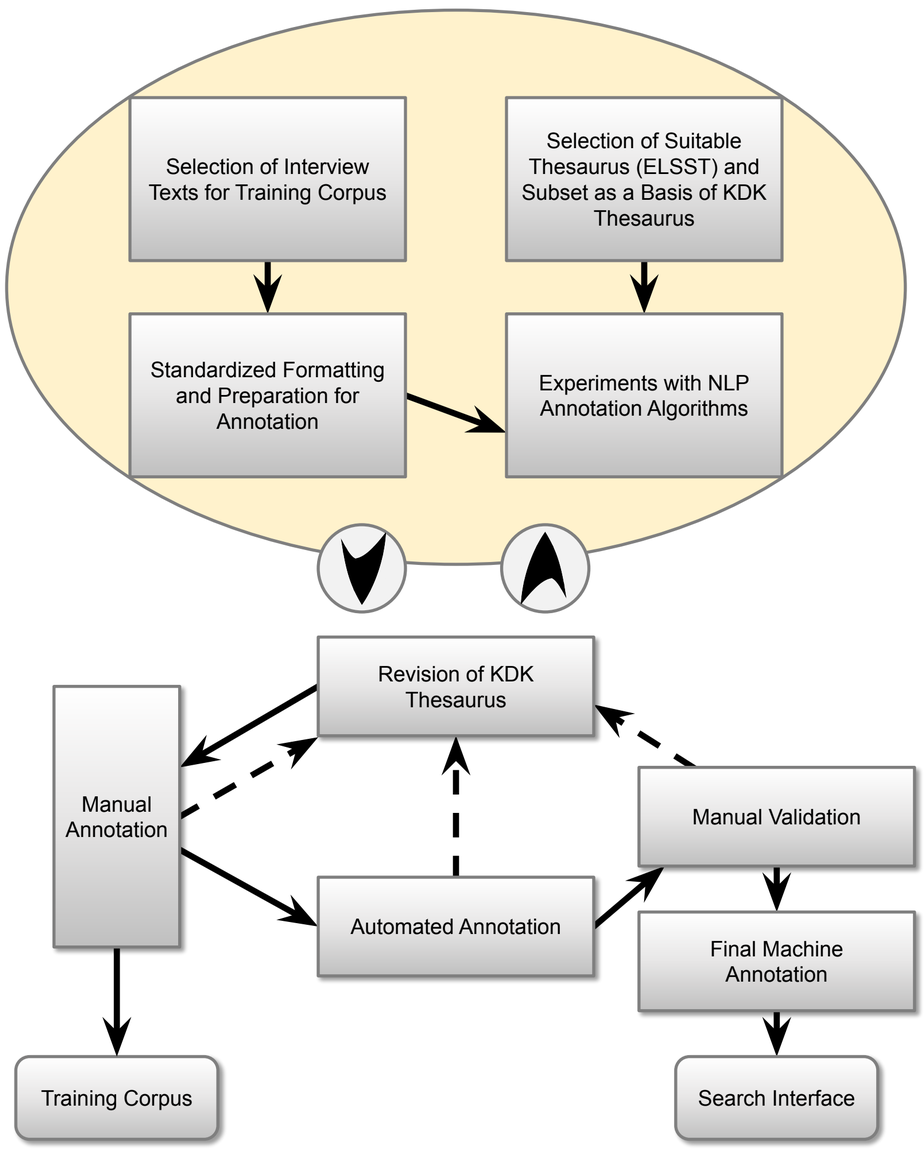
TRAINING CORPUS

- 39 interviews
- 1183 pages
- sections of 2,000-5,000 characters



Voices of the 20th Century
Archive and Research Group

Label Studio



KDK THESAURUS

Must represent:

- fields and major themes of sociological research
- contents of our archives
- issues in Hungarian history and society

Interventions (purposeful ethnocentrism):

→ downsizing thesaurus ELSST (~3400 to 220 terms)

→ introducing/prioritizing/specifying certain terms
e.g. 'language competence', 'privatization', 'reform', 'liberation', 'regime change'

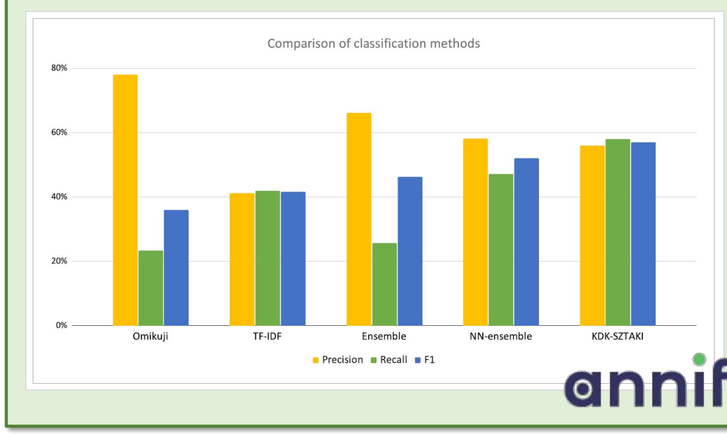
→ creating hierarchical structure (3 levels)

cessda ELSST Thesaurus

HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

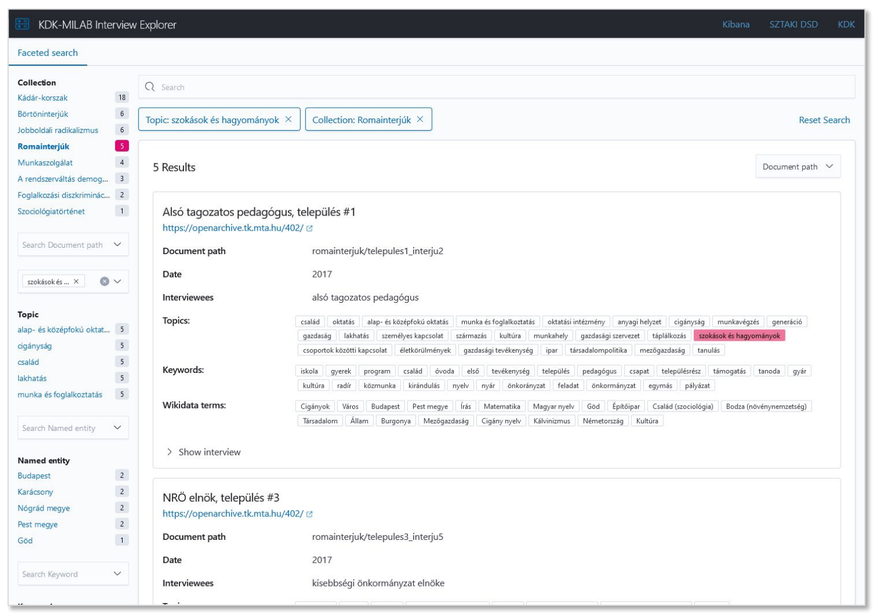
EXPERIMENTS

- 7 classification methods:
  - Fasttext, Omikuji, TF-IDF, MLLM, Stwsfa, Ensemble, KDK/SZTAKI
- 4 named entity recognition methods:
  - HuSpaCy, emtsv, HuBERT, KDK/SZTAKI
- 1 enhanced wikification method

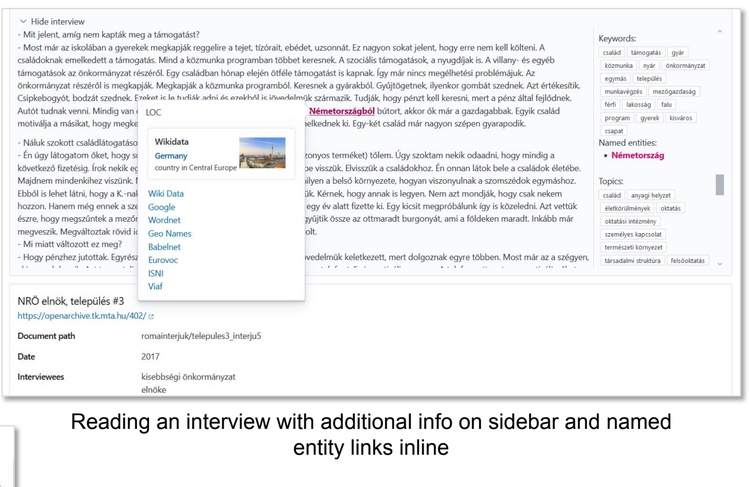


RESULTS

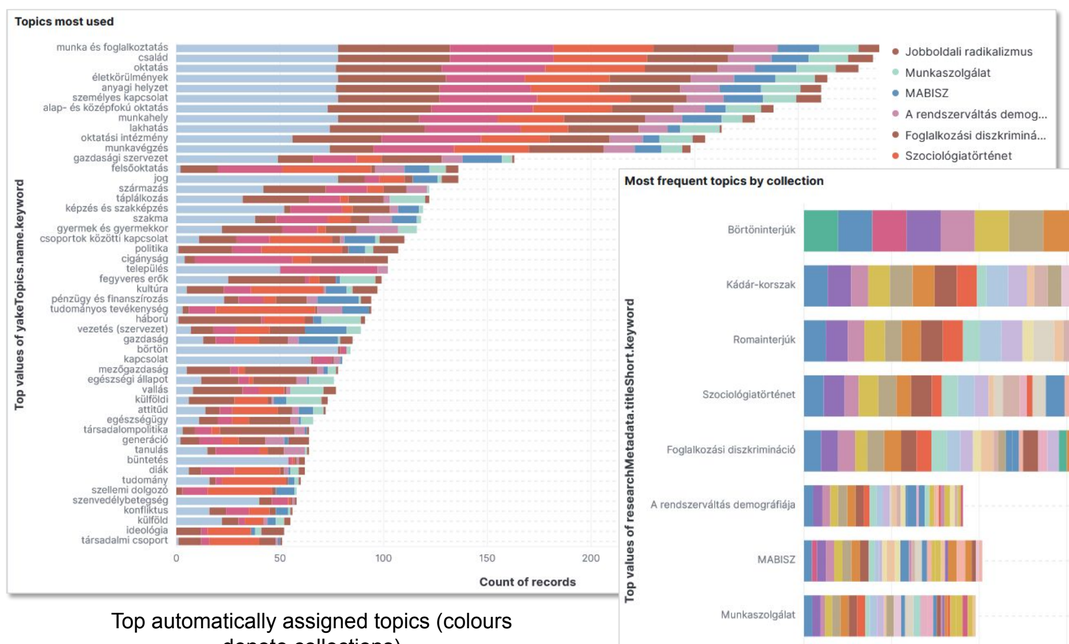
- KDK thesaurus
- Training corpus
- 8000+ sections of 360 interviews processed with our NLP methods
- Experimental user interface with faceted search, augmented reading and analytical diagrams



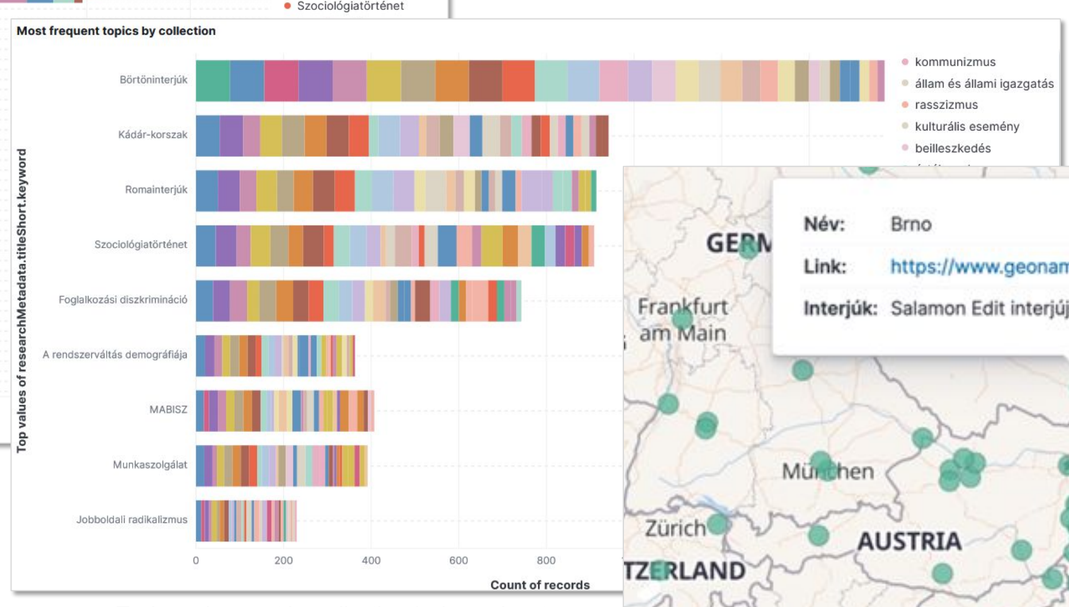
Faceted search in interviews based on NLP assigned keywords



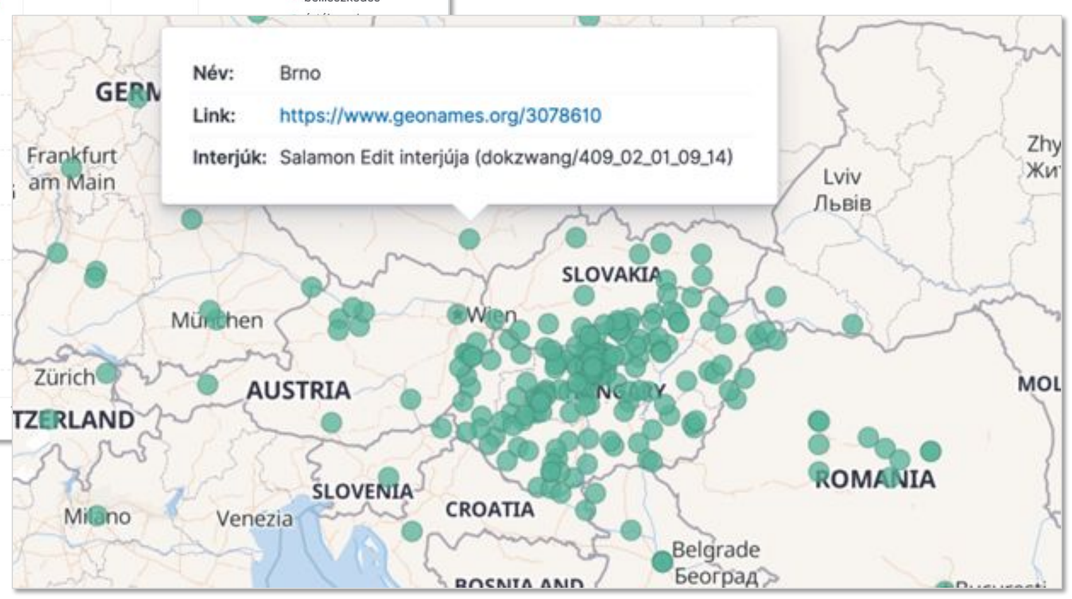
Reading an interview with additional info on sidebar and named entity links inline



Top automatically assigned topics (colours denote collections)



Topic assignment by collection (colours denote topics)



Interview mentions of location entities on a map (with interview title and identifier)

Sponsored by:

ARTIFICIAL INTELLIGENCE National Laboratory

Contacts:
Judit.Gardos@tk.hu
Andras.Micsik@sztaki.hu

tkkdk Centre for Social Sciences Research Documentation Centre

SZTAKI Institute for Computer Science and Control Department of Distributed Systems

https://kdk.tk.hu/en/natural-language-processing-project