HØGSKOLEN I OSLO
OG AKERSHUS

Named entities in indexing:

A case study of TV subtitles and metadata records

Anne-Stine Ruud Husevåg
Anne-Stine.Husevag@hioa.no

PhD Student
Oslo and Akershus University College of Applied Sciences
Department of Archivistics, Library and Information Science

HØGSKOLEN I OSLO
OG AKERSHUS

The Norwegian Broadcasting Corporation (NRK) archive

– a hidden treasure

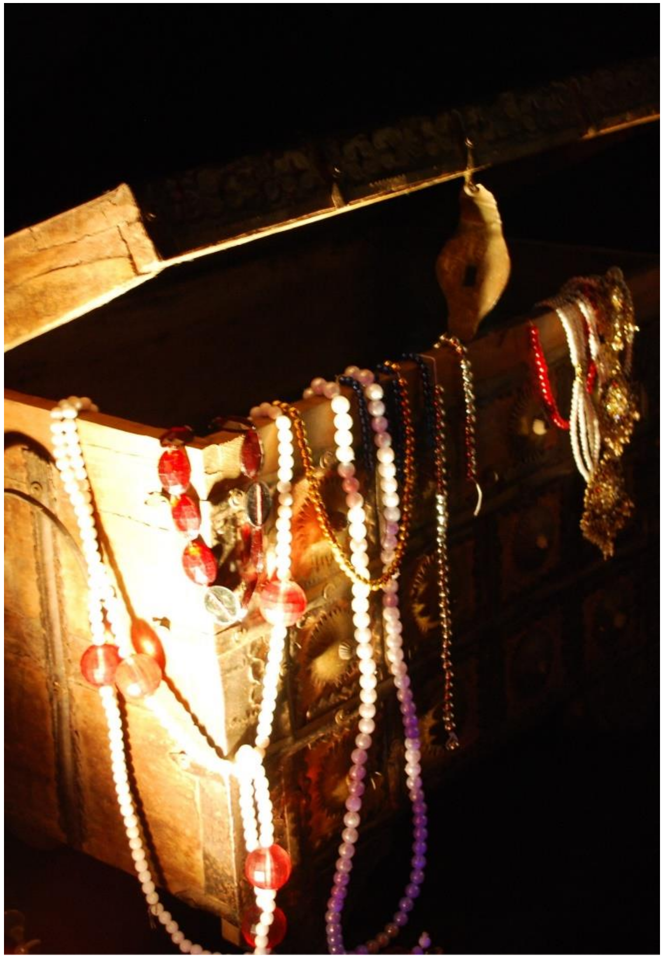
Photo by Nebojsa Mladjenovic

The Norwegian Broadcasting Corporation (NRK) archive

- NRK have multiple projects running that makes old material available to the public.
- Everything NRK has produced or financed since 1960 shall be published in NRK Web TV.

Photos: Screenshots from NRK, early testing of TV (1947)

The Norwegian Broadcasting Corporation (NRK) archive



Digitalization have opened the treasure chest.

We now have the opportunity to relive nostalgic moments and revisit historic experiences…

…if we can find them.

Photo: Tom Praison

Information retrieval in the NRK archive

- Searching today: mainly program titles

Textual sources that can be used for retrieval

- Written descriptions of old material
  - Found in a free text field in old metadata records
  - Considerable variation in length and quality
  - Written for internal use
  - Not always suitable for the public
  - No longer produced due to time costs

- Subtitles for the deaf and hard of hearing

Extraction of valuable entities to improve indexing

- Film archive users are often searching for named people, events, places and other named entities.
- Density of named entities in subtitles: 5 %
- Density of named entities in metadata records: 20 %

These findings are consistent with other studies in both Norwegian and English, showing that named entities are used more in summaries and book indexes than in texts similar to natural language.

Even when full text (here in the form of subtitles) are available, named entities should be an important focus of annotation. Recognizing named entities is a task that can be done automatically to support indexing and retrieval in a knowledge organisation system.

Salient entities

All that glitters is not gold—
Often have you heard that told.

Photo: leigh49137 @ Flickr, Quote by Shakespeare, The Merchant of Venice (Act II - Scene VII - Prince of Morocco)

Salient entities

Assuming that the entities in the metadata records are the best document descriptors, we must try to find those entities in the subtitles. Named entities mentioned only in passing or used as examples will not help information retrieval in a knowledge organisation system.

This is the reason why I have chosen to analyse the named entities found in both metadata and subtitles for the same programs in different genres.

Named entities from subtitles found in metadata

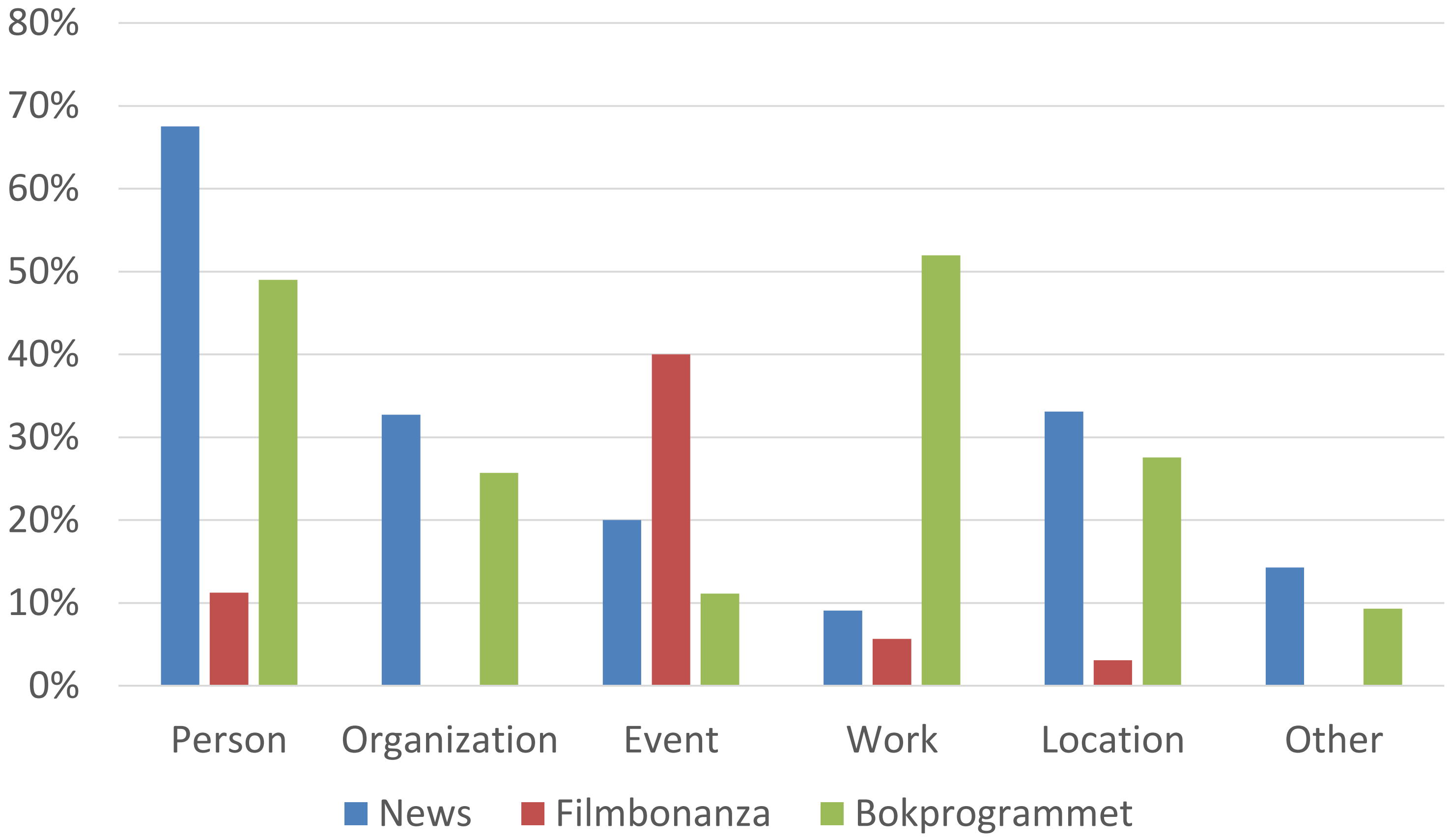
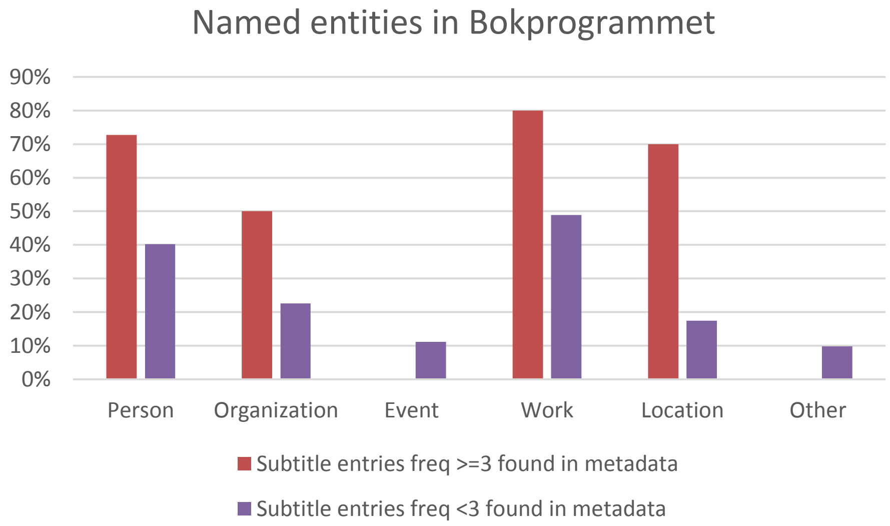
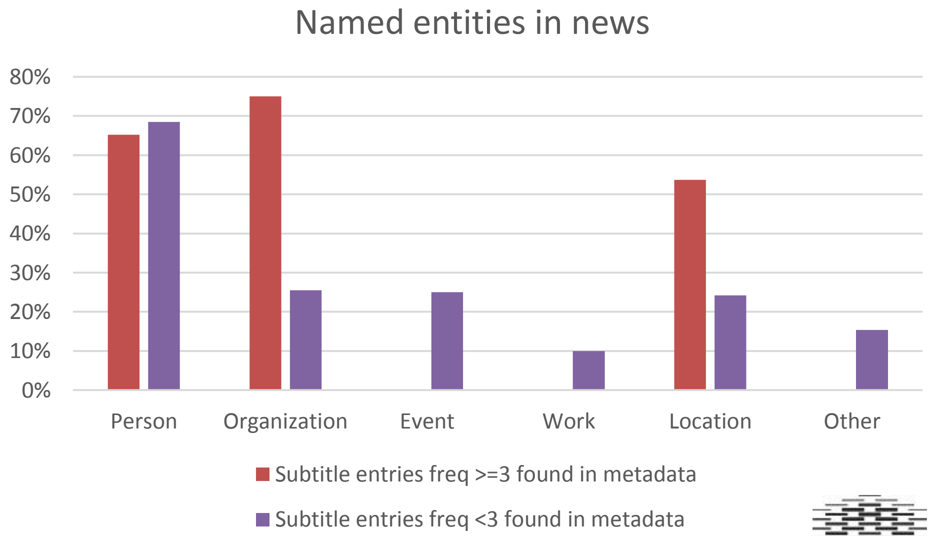


Fig. 1. Percentage of NEs from subtitles from different TV shows found in metadata, arranged by entity type.

The importance of frequency

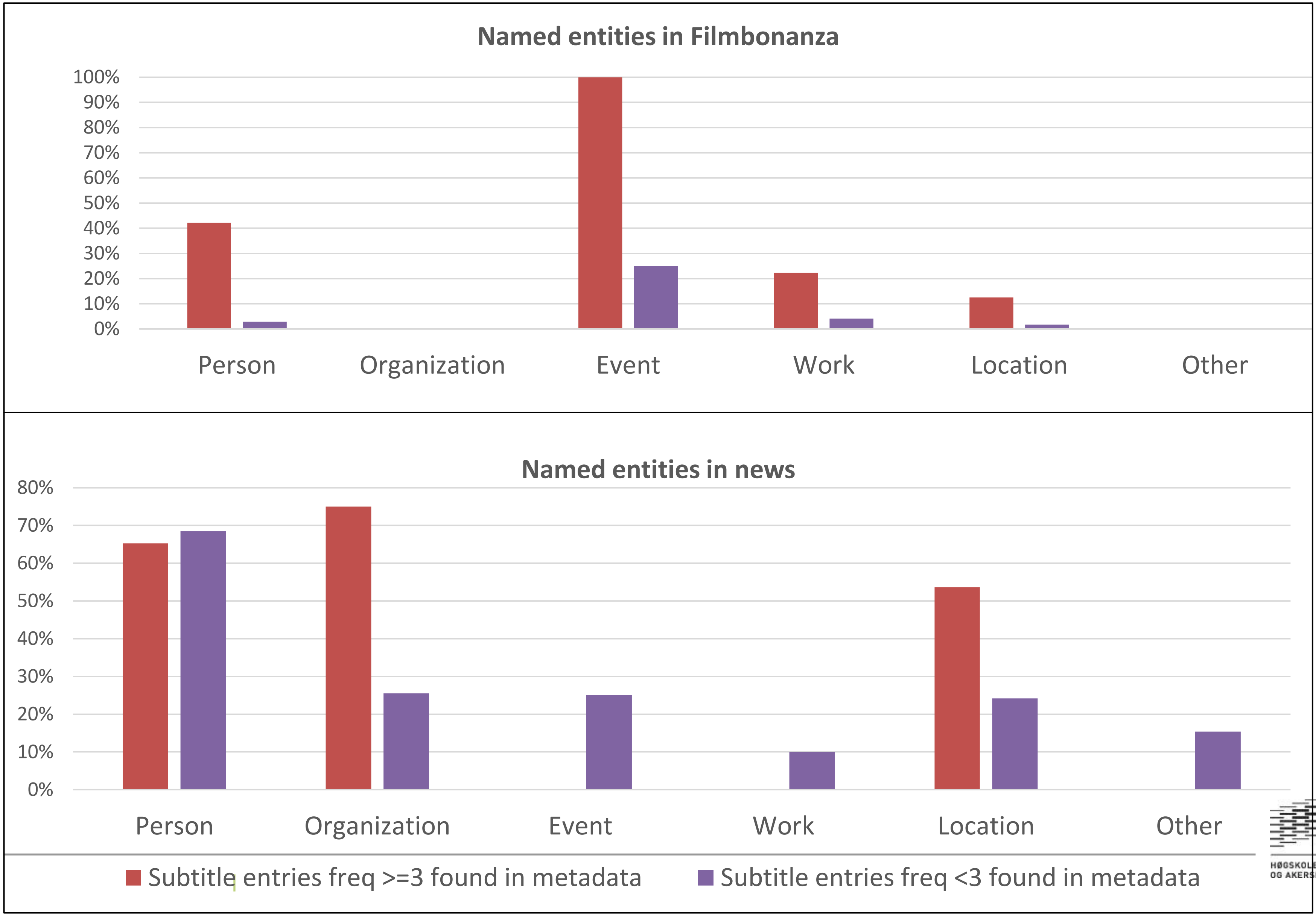
In order to find out if the frequency of the named entities in the subtitles was of importance for the likelihood of librarians choosing the named entities to represent the program in the metadata records, this figure separates the named entities occurring three times or more from the less frequently mentioned.





HØGSKOLEN I OSLO OG AKERSHUS

The importance of frequency



Conclusions

- Norwegian texts have a lower density of named entities than English texts
- The density of named entities in metadata (20 %) is much higher than in subtitles (5%), implying that named entities are better content descriptors than other parts of speech
- **Person, location** and **organization** are the most prominent entity types in news
- High frequent entities of the entity types **person**, **location** and **work** are important content descriptors of culture programs.
- Differences in frequencies have a higher discriminatory value for some entity types
- Personal names are often considered important regardless of frequency, especially in news material.

Future work

The overarching topic in this project is to explore the usefulness of Named Entity Recognition as a method for facilitating automatic indexing of broadcast material. My research questions are as follows:

- Are named entities more salient content descriptors than other words?
  - To what extent do users use named entities when searching for and describing content?
  - To what extent do trained librarians use named entities when describing content?
  - Are some entity types more salient than other types, and is this different in different genres and material?
  - What characterises salient entities?
- What methods are best suited to extract the most salient named entities from subtitles?
  - Will these approaches find the same entities that users use when searching?
- How can recognition of named entities improve linking between documents?
  - To what extent are users interested to learn more about named entities mentioned in a program?
  - Will similarity measures based on named entities alone give a different result than similarity measures based on descriptions in metadata records or the full text from subtitles?

Thank you
for your attention!