Spoken Corpus of Karel Makoň

Oldřich Krůza
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
kruza@ufal.mff.cuni.cz

Abstract

We present the spoken corpus of Karel Makoň, a corpus of spontaneous speech by a single speaker in Czech language within the domain of Christian mystic, with partly automatic, partly manual transcription obtained through a web application made for correcting the output of ASR. This paper briefly outlines the system for processing the corpus and elaborates on the corpus itself. A baseline for topical analysis is presented.

1 Introduction

The spoken corpus of Karel Makoň is a collection of talks given in a circle of friends in the course of late 60's or early 70's till 1991. The recordings have been kept on magnetophone tapes until their digitization that took place between 2010 and 2012. A complete transcription was obtained using a dedicated ASR system and the work of the community around K.M.'s legacy. The corpus is about 1000 hours in total length, of which about 66 have been transcribed manually.

Most of the previous work was focused on acquiring the automatic transcription and development of a web interface (Krůza and Peterek, 2012) to allow users to access the talks and provide corrections to the existing transcription, in order to have high-quality transcription as well as to gather more training data.

The actual content of the corpus with respect to topics, references and statements, is on one hand quite well known because the the domain stays more or less consistent across the whole set. On the other hand, it is only known vaguely and a systematic effort to analyze it is yet to be carried out.

2 Author

The author of the talks, Mr. Karel Makoň (Hájek, 2007) *1912 †1993, was giving talks in a strife to share his awareness of the proverbial meaning of life and to give a manual to eternal life after his release from a concentration camp during the WWII.

His steep spiritual path began in early childhood when he was subject to repeated arm surgery without anesthesy. The pain the child suffered made him escape mentally from inside his body, which led him to what he calls "the yellow light".

His path culminated in 1939, after nine years of constant prayer for a stronger love to God, when he was deported to the nazi concentration camp in Sachsenhausen, as a university student. Here, he was promised certain death by a Gestapo warden. As the nazi was approaching K.M. to kill him, he gave up his life completely to God and the attacker unexplainably turned around in awe and fled.

It was in that moment that Makoň obtained enlightenment, accompanied with profound understanding of Christian symbolism. He was experiencing absolute freedom and abundance ever since, even in the concentration camp. He also understood the general spiritual mechanisms that conditioned his deliverance and devoted much of his subsequent life to enabling others to learn from his experience.

3 Giving and Recording the Talks

The mystical, religious and spiritual nature of his talks put K.M. in direct antagonism with the communist regime. The opposition was strictly unidirectional though, as Makoň was far from wasting his energy on a fight against it. His way of dealing with the conflict was by staying out of wide popularity. Hence, most of the talks have been given during regular meetings in various places of Czechoslovakia and later the Czech Republic, in a narrow circle of friends.

One of the fortunate consequences of the seclusion of the talks is that there is very little background noise on most of the recordings. Also, most of them have been taken by a single person who used the available state-of-the-art amateur recording technology and methods. This person was also very systematic in labeling and archiving the recordings. As a result, the majority of the base for digitization was a set of reel-to-reel tapes labeled with ordinal indexes and a set of cassettes labeled by year and ordinal index. The reel-to-reel tapes have been recorded in a generous speed of 9cm/second, which granted them a very good shield against the inevitable aging.

4 Processing

We don't have the resources needed to manually annotate the corpus like e.g. (Maekawa et al., 2000), but there are lay volunteers willing to work on the project for the cause of tending Makoň's legacy.

A complete transcription is essential or at least helpful for any kind of further processing. Therefore, I have focused on two major aspects of processing the material: 1. Making it accessible to the public and 2. acquiring a complete transcription.

4.1 Access to the Corpus

The speech recordings are available through two sources. Firstly, it is in the Clarin repository, which guarantees it will be easily available no matter what fate this project will take.

Secondly, I have created a web application for listening to the recordings directly in the browser. The web application is also capable of displaying the transcription synchronously. When there is an error in the transcription (and the automatic transcription has plenty of errors), the user can submit a correction.

4.2 Complete Transcription

Transcription of the whole corpus has been the main focus for the majority of the time spent on the project. The concept was to transcribe a few minutes from scratch, train a HMM-based ASR model on that, evaluate the bulk of the corpus and feed that to the web application, as a base for the users to correct.

The community managed to provide over 66 hours of high-quality manual transcriptions through the web application and which have been used gradually as training data. Now, the acoustically intact recordings are automatically transcribed well enough to allow reading along and to do full-text search.

The word error rate on our test set is currently 42% but it is probably lower on clean recordings and also most mistakes are in word endings, so the meaning is comprehensible. I am currently working on a DNN acoustic model (Hannun et al., 2014), so chances are the ASR score will soon rise.

4.2.1 Acoustic Quality

Despite of the relatively good recording technique and mostly quiet environment, the resulting acoustic quality varies wildly. Many recordings are clearly intelligible, while some suffer from various defects like 1. background noise in varying intensity and bandwidth, 2. altered speed, 3. clipping or faitness, 4. echo, 5. overlap of two recordings.

The varying acoustic quality is the most serious obstacle to the quality of the automatic transcription. But it is also detrimental to the manual transcription. To mitigate this, the web interface features since lately a user-controllable graphical equalizer, which can help especially in the case of narrow-band noise.

4.3 Secondary Processing Attempts

Some limited effort was invested into further processing beyond plain transcription. I have created a simple search engine over the transcription that yields deep links into relevant passages in the audio. Elasticsearch was used for this purpose, with Czech algorithmic stemmer.

Another effort was processing manually written indexes that are available for a fraction of the recordings. These have been partly hand-written, partly typed and are aligned using the positions of the device counter. Most of these are scanned and some limited experiments were carried out to apply OCR. The explicit mapping of the indexes with corresponding recordings and time-positions is yet to be done.

5 Topics

In general, the whole corpus deals with a single topic that can be summarized as a howto for entering the eternal life before the physical death. On a finer-grained look, we can identify recurring sub-topics, like

- interpretation of some passages in the New Testament, notably the parable of the prodigal son, the parable of the talents, and the Our Father prayer,
- milestone personal experiences, like the one in the concentration camp,
- explaining symbolism, like the apostles as symbols for human abilities,
- references to specific people, like St. Teresa of Ávila or Padre Pio.

A list of recurring topics and their identification in the sound files is a point of future work with similarities to (Skorkovská et al., 2011). We can take the current transcription as a source for a baseline. Topic is inherently a vague concept but for the baseline evaluation, we can take a look at some easily defined cases like named entities. The ease about them is that we can look for their specific word form and if it is not found, then it means the transcription is not covering it.

Of course, there are exceptions, like saying "the capital of Japan" instead of "Tokyo" but firstly, I chose terms where this is not likely to happen and secondly, this can only lower the recall of our baseline score, not the precision, and we have no way of telling the recall given no annotated data for this task anyway.[1]

I have chosen the following example topics: 1. **Lazarus**, 2. **Mithra, Mithraism**, 3. **Satan**, 4. **St. Teresa**, 5. **fairy-tale**. I chose fairy-tale (Czech: *pohádka*) as a recurring topic with a unifying word, for an example of a non-named-entity.

The search for stems of these words in the transcription yielded results summed up in Table 1. For terms where there were more than 20 hits, I have checked the first 20 but each from a different file.

Term	Query	hits in automatic transcription	expected hits in automatic transcription	hits in manual transcription	Precision
Lazarus	`lazar.*`	16	171	14	8/10
Mithra	`mithra.*|mitra.*`	0	0	0	n/a
Satan	`satan.*`	308	1575	129	16/20
St. Teresa	`terez.*`	882	1111	91	15/20
fairy-tale	`pohádk.*`	216	671	55	18/20
average precision					81.25%

Table 1: Text search hits in the transcription; precision only calculated from automatic part

The precision of over 80% seems quite promising and hints that the current quality of transcription allows for reasonable searching through the corpus. However, the hits were all clearly audible. I met none in an acoustically defect file, which leads to the hypothesis that those are not yet searchable to a distantly satisfying level.

7.6% of the total words in the corpus are manually transcribed. The column "expected hits" interpolates how many occurrences should appear given equal frequency. Oddly enough, the ratio of the expected and actual hits varies a lot among the search terms. It is almost certain that at least for some of them, the recall is very low.

[1] The 50-minute test set is way too small for this kind of task. Another source could be manual notes available to some of the recordings but preparing them for this purpose requires some more effort.

5.0.1 Correspondence between Talks and Books

K.M. was writing, translating and commenting books most of his life. We can assume that the book he was just working on influenced the topic of his talks. For one, he translated the book *El Castillo Interior*, English *The Interior Castle* by St. Teresa of Ávila in 1988. For most recordings, the year of their acquisition is known. Figure 1 shows how many hits for the name of the Saint are present in the transcription by recording year.

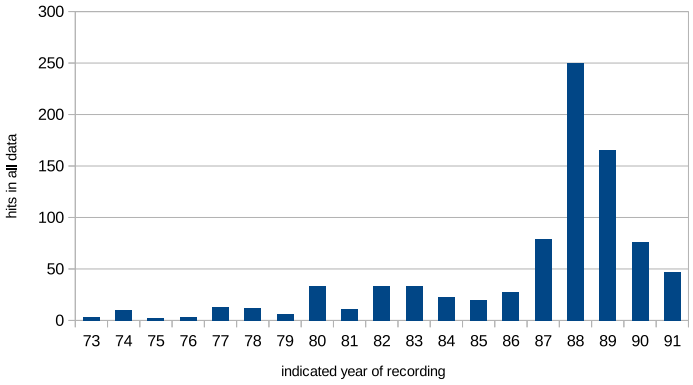

Figure 1: Number of hits for the query `terez.*` in the transcription by alleged year of the recording

The peak around the year 1988 supports the speculation that topics in the talks correlate with those of books written at the same time.

6 Conclusion

With its length of over 1000 hours spoken by a single speaker over three decades, the corpus of Karel Makoň is a unique piece in the Lindat/Clarin repository. The consistent domain of Christian mystic is another of its distinguishing features. A basis for detailed topical analysis is presented exploiting the existing transcription and the author's books.

Acknowledgments

The research was supported by SVV project number 260 453.

This work has been using language resources stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

[Hájek2007] Jurik Hájek. 2007. Český mystik karel makoň. *Dingir*, 2007/4:142–143.

[Hannun et al.2014] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[Krůza and Peterek2012] Oldřich Krůza and Nino Peterek. 2012. Making community and asr join forces in web environment. In *International Conference on Text, Speech and Dialogue*, pages 415–421. Springer.

[Maekawa et al.2000] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of japanese. In *LREC*. Citeseer.

[Skorkovská et al.2011] Lucie Skorkovská, Pavel Ircing, Aleš Pražák, and Jan Lehečka. 2011. Automatic topic identification for large scale language modeling data filtering. In *International Conference on Text, Speech and Dialogue*, pages 64–71. Springer.