“Mansi corPŌS”: Archiving Mansi written materials published between 1937 and 2020

Csilla Horváth
Szeged, Hungary
naj.agi@gmail.com

Abstract

The paper presents the project “Mansi cosPŌS” aiming to archive the materials written in Mansi, published in the Soviet Union and the Russian Federation between 1937 and 2020. The paper briefly introduces the history of Mansi literacy, with special attention to the literary and ethnographic editions, transcribed in Cyrillic and published with the cooperation of Mansi speakers, as well as Mansi press. The presentation focuses on the previous attempts and projects aiming to archive the available written Mansi resources, the comparison of the results of the present and previous projects, the corpus created on the basis of materials archived in present project, as well as the future utilisation of the archived materials and the corpus, taking into consideration the viewpoint of both the scientific and the Mansi communities.

1 Introduction

The aim of the project “Mansi cosPŌS” aiming to archive the materials written in Mansi, published in the Soviet Union and the Russian Federation between 1937 and 2020. The word ‘pōs’ in the project’s name means ‘imprint, sign, number’ in Mansi, thus it has been closely connected to the Mansi concept of written communication since times prior to the European understanding of literacy in the region. Achieving this goal, both the largest archive of Mansi text published between 1937 and 2020 and the largest corpus of Mansi texts is going to be created. This Mansi corpus serves as a valuable resource of information for linguists working with the Mansi language. At the same time, the archived Mansi books give the Mansi speakers’ community a chance to become acquainted with its history of literacy. Since the books published during Soviet period were almost exclusively printed in Leningrad, they became accessible for Soviet and foreign researchers, but many of them never found their way to the Mansi speakers, thus without archiving and sharing it digitally they would be unknown even for the Mansi intelligentsia and language activists (Яныг утыт ос мāнь утыт. 2015.).

2 The status of the Mansi language

Mansi is an extremely endangered indigenous Uralic language, spoken in Western Siberia, especially on the territory of the Khanty-Mansi Autonomous Okrug. Among the approximately 13,000 people who declared to be ethnic Mansi according to the data of the latest Russian federal census in 2010, only 834 stated that they could speak the Mansi language, and the overall number of speakers of the language is as low as 938. Mansi is an indigenous minority language, without any kind of official status. The consequences of rapid urbanisation, namely, the altered lifestyle, the multiethnic environment, the Russian-dominated press and media, etc. intensify the influence of factors accelerating language shift and create new opportunities to support language revitalization attempts (c.f. Horváth, 2015). Although the domains of Mansi language use are becoming more numerous than before, they are still limited. Mansi is not an official language, either at the regional or the municipal level, and it is practically absent from official or semi-official domains such as legislation, public transport or street signage. Mansi has no economic significance either, thus it plays a marginal role in the business sphere or the labour market. Mansi has a small but growing importance in cultural and leisure activities, as well as the internet, and, compared to these domains, has a relatively strong position in education and family.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 16–19,
Online, March 2–3, 2021.

3 Written Mansi resources

3.1 The history of Mansi literacy

Although researchers had already been publishing Mansi texts prior to the date, the history of Mansi literacy is considered to have begun in the year 1931 (Чернецов, 1937: 168), when the Latin-based alphabet of the Mansi language was created in Leningrad, at the Institute of the Peoples of the North, together with the writing systems of other Siberian languages. The transition to Cyrillic writing system took place in 1937, and it has been in use ever since. Since 1937 the Mansi writing system has undergone only minor changes. Currently two slightly different variants are in use, one used in academic and pedagogical publications (dictionaries, traditional schoolbooks), the other in other media and in schoolbooks designed for heritage language learners.

3.2 Books in Mansi

Despite of the relatively long history of Mansi literacy, it is not farfetched and unrealisable to aim for the complete digitalisation of the Mansi publications According to a comprehensive catalogue of Mansi publications by 2007 (Волженина and Фетисова, 2007) and personal experiences since 2006, there only have been approximately 170-180 books published in the Mansi language since 1937. More than half of them are primers and other schoolbooks, while the other half consists of folklore collections (mainly tales), contemporary literature, Mansi translations of Soviet literature for children, Mansi translations of the Gospels and other religious texts.

3.3 Press in Mansi

The first newspaper completely written in Mansi has been published since February 1989, under the name "Luima Seripos". Now it appears every second week on 20 pages. The print version of the newspaper is in Mansi only, while on the official homepage articles can be found in Mansi and in Russian as well. The Mansi texts published cover various topics such as traditional lifestyle, folklore and short biographies, as well as different aspects of urban life. In addition to "Luima Seripos", the editorial board also published five issues of a journal for children, under the title "Khotalkve". Another monthly journal for children, "Vitsam" has been published since 2014 (c.f. Horváth, 2019).

4 Digitalised Mansi texts and Mansi corpora

A handful of projects has been reported to produce digitalised Mansi texts or Mansi corpora. The Mansi module of the project "Ob-Ugric morphological analyzers and corpora" was based on the digitalised Northern Mansi texts of three publications (cf. Fejes and Novák, 2010). Although these texts were proposed to be made available, they cannot be retrieved from the suggested location any more. The project used the Finno-Ugric transcription system. In the framework of the project "Eurobabel – Better Analyses on Endangered Languages: Ob-Ugric languages" 137 digitalised Northern Mansi texts appeared, most of them chosen from Artturi Kannisto's folklore collection, further texts were published of Bernát Munkácsi's works, other folklore publications, while some of them originally appeared in the Mansi newspaper Luima Seripos. A small proportion of the published texts are provided with English translation and glossing. Although reports about the project regularly mention the importance of accessibility for native speakers (e.g. Skribnik et al, 2011, Bakró-Nagy, 2014), the texts were published in IPA transcription, unintelligible for the majority of Mansi speakers.

In the project "FinUgReVita", a pilot corpus was created between 2013 and 2017. This corpus contains issues of the Mansi newspaper "Luima Seripos" published between 2013 and 2016. The corpus consists of 520,000 tokens, converted into XML format. (Horváth et al, 2017). The project "Mansi cosPŌS" aims to continue the work started by the researchers working with the Mansi language during the project "FinUgReVita".

5 Corpus compilation

5.1 Collecting the written sources

A smaller part of the Mansi books (especially the earliest Mansi publications) was accessed through the National Library of Finland. These publications were archived in the framework of the project "Fenno-Ugrica", although the previously freely available copies have been removed from the homepage. A larger part of the Mansi books (especially those published in the

1950s and after) are available for or in the possession of the author. The issues of the newspaper "Luima Seripos" published since 2013 are available on the webpage of the newspaper. The earlier issues of the "Luima Seripos" newspaper, all the five issues of "Khotalkve", and almost every issue of the children's newspaper "Vitsam" are in the possession of the author.

5.2 Processing the sources

The digitally available Mansi books and newspapers were downloaded manually, the non-digitally available resources are being scanned manually as well. Just as during the project "FinUgReVita", the documents are converted into text files with the help of Optical Character Recognition, segmenting long vowels into two characters (as described in Horváth et al, 2017: 63). For the OCR analysis the open source OCR engine "tesseract" (https://github.com/tesseract-ocr) is going to be used, as it is freely available, and has been used for processing resources of a similar period, cultural background and amount (Szabó et al, 2020, Kmetty et al, 2020). The text files are being proofread and normalised according to the transcription used in the Mansi press and contemporary educational publications. Initially the corpus is going to be transformed to XML format, extended with a unique identifier for every independent unit, indication of genre, author or Mansi translator of the unit. The XML file is going to apply special tags for named entities such as person names and locations.

5.3 Perspectives

Taking into consideration, that the corpus created in the project "FinUgReVita", based on the issues of "Luima Seripos" published during four years, consists of 520,000 tokens, we may conclude, that the corpus created in the project "Mansi corPŌS" will consist of at least 1.5 million tokens. Besides producing a comprehensive source of information, the corpus may also serve as the base of the final version of the Mansi morphological analyser described by Horváth et al (2017).

6 Conclusions

The project "Mansi corPŌS" aims to archive all the available Mansi materials published with Cyrillic transcription between 1937 and 2020. The corpus is going to contain the folkloric texts (volumes of recently collected legends, folktales, stories), the literary texts (children's literature, contemporary prose), the academic texts (papers, non-fiction publications for children), the Bible translations (Gospels, other biblical fragments). Taking into consideration the overwhelming significance of Mansi press, the project "Mansi corPŌS" pays special attention to the available issues of the only regularly appearing Mansi newspaper "Luima Seripos", published between 1989-2013 and since 2017, as well as the issues of sporadically appearing other newspapers (e.g. the children's newspaper "Khotalkve" and "Vitsam").

The project "Mansi corPŌS" is an individual project, at the moment it is not financed by or not allocated at any institution. The project undoubtedly does not offer any computational innovation, applying the methods of corpus building to an endangered language in order to create a comprehensive corpus nevertheless may be prospective. Although the final location of the corpus is not set yet, it is going to be freely accessible, thus may serve as a solid base for future language technology tools, as well as an easily searchable source of information for Mansi speakers.

References

Marianne Bakró-Nagy. 2014. Obi-ugor nyelvek digitális adatbázisa". PPT of the presentation held at the conference "A kulturális örökség és a humán tudományok innovációi a 21. században". http://www.babel.gwi.uni-muenchen.de/media/downloads/BNM_BABEL_obi-ugor.pdf

László Fejes, Attila Novák. 2010. Obi-ugor morfológiai elemzők és korpuszok. In: Tanács, Attila and Vincze, Veronika, eds. *VII. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem, pages 284-291.

Csilla Horváth. 2015. Beading and language class. Introducing the Lylyng Soyum Children Education Centre's attempt to revitalise Ob-Ugric languages and cultures. *Zeszyty Łużyckie* 48: 115-127.

Csilla Horváth. 2019. The "extraordinary thing": The only Mansi newspaper on online presence and social media practice. In: Pralica, Dejan and Šinković, Norbert, eds. *Digitalne medijske tehnologije i društveno-obrazovne promene 8*. Novi Sad: Univerzitet u Novom Sadu, 165-176.

Csilla Horváth, Norbert Szilágyi, Veronika Vincze and Ágoston Nagy. 2017. Language technology resources and tools for Mansi: An overview. In: *Proceedings of The 3rd International Workshop for Computational Linguistics of Uralic Languages*. Stroudsburg: The Association for Computational Linguistics, 56–65. http://dx.doi.org/10.18653/v1/W17-0606

Zoltán Kmetty, Veronika Vincze, Dorottya Demszky, Orsolya Ring, Balázs Nagy and Martina Katalin Szabó. 2020. Pártélet: A Hungarian Corpus of Propaganda Texts from the Hungarian Socialist Era In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2374-2381. https://www.aclweb.org/anthology/2020.lrec-1.290.pdf

Elena Skribnik, Veronika Bauer and Zsófia Kováts. 2011. The relevance of language documentation for language users. PPT of the presentation held at the *1st International Conference on Language Endangerment: Documentation, Pedagogy, and Revitalization"*. http://www.babel.gwi.uni-muenchen.de/media/downloads/relevance_of_language_documentation.pdf

Martina Katalin Szabó, Orsolya Ring, Balázs Nagy, László Kiss, Júlia Koltai, Gábor Berend, László Vidács, Attila Gulyás and Zoltán Kmetty. 2020. Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. *Historical Methods: A Journal of Quantitative and Interdisciplinary History.* https://doi.org/10.1080/01615440.2020.1823289

Волженина Светлана Юрьевна and Фетисова Галина Яковлевна. 2007. Издания на языках народов ханты и манси (1879-2006). ООО Баско, Екатеринбург.

Яныг утыт ос ма̄нь утыт. 2015. *Лӯима̄ сэ̄рипос* 2015/21, p. 16

Чернецов Валерий Николаевич. 1937. Мансийский (вогульский) язык. In: Прокофьев Георгий Николаевич, ed.: *Языки и письменность народов Севера I.* Государственное учебно-педагогическое издательство, Москва-Ленинград- 163-182.