Timur and the Mansi spellchecker

Csilla Horváth
University of Helsinki
csilla.horvath@helsinki.fi

Abstract

The article presents the results of an experiment involving the use of the Mansi FST and spellchecker created by the GiellaLT infrastructure. The Mansi are one of the indigenous peoples of the Russian Federation. The Mansi language is an endangered Uralic language primarily spoken in western Siberia, along the Ob River and its tributaries. The present article discusses the efficiency of the Mansi FST and spellchecker when used for translating Mansi literature from the 1950s.

1 Introduction

The article presents the results of an experiment of using the Mansi FST and spellchecker, created by the GiellaLT infrastructure on a text sample from a genre other than that for which the tool was designed.

Section 2 provides a brief overview of the Mansi language, its orthography and literature. Section 3 describes the experiment and its results. Section 4 presents the additional findings that were discovered during the experiment. Section 5 proposes conclusions.

2 Background

2.1 The Mansi language

The Mansi are one of the Arctic indigenous peoples of the Russian Federation. The Mansi are often still regarded as traditional communities living primarily on fishing, hunting, and gathering, and to some extent on reindeer breeding, however, as a result of industrialisation and urbanisation, taking place in their home region since the 1960s, the majority of the Mansi has been living in urban type of settlements, alienated from their once traditional lifestyle.

The Mansi language is a contested, endangered minority language. It is spoken mainly in Western-Siberia, along the Ob River and its tributaries. Four Mansi dialect groups were documented in the nineteenth century: Northern, Eastern, Southern, and Western Mansi, each of which had several (sub-)dialects. The Southern and Western dialects are already extinct, the Eastern dialect is either extinct or moribund. In this paper, Mansi language henceforth refers to the Northern Mansi variety. Mansi is used in both spoken and written form (cf. Virtanen and Horváth (2023)).

According to the results of the 2021 Federal Census, there were 12,308 Mansis living on the territory of the Russian Federation. 1,008 people claimed to use Mansi, altogether 951 of them were of Mansi ethnicity, while beside the Mansi, 8 Nenets, and 9 Khanties claimed to use the Mansi language. Nowadays, Mansi has its strongest position in the sphere of family language use, but since the turn of the century it has been introduced to new domains of language use as well, such as heritage language education, theatre and popular music, print, broadcast and social media (c.f. Horváth (2020, 2024, 2025)).

2.2 The written Mansi language

The literacy of the Mansi language, similarly to other indigenous languages of the Russian North, has been created in 1931. The Mansi standard orthography originally used Latin-based alphabet, then switched to a Cyrillic-based alphabet in 1937. Since 1937, the Mansi writing system has undergone several significant changes. In the earliest period, the Cyrillic transcription contained no special characters, and vowel length was not marked either. Later, perhaps at the beginning of the 1970s, a special character was introduced to denote the

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 68–71

velar nasal, while marking vowel length became widespread in the 1980s.

Currently, two slightly different variants of Mansi orthography are in use, one, more marginal, used in some of the academic and pedagogical publications (dictionaries, traditional schoolbooks), the other, more general, used in all other work, including print and broadcast media, social media, and schoolbooks designed for heritage language learners.

2.3 Mansi literature

According to a comprehensive catalogue of Mansi publications (Юрьевна and Яковлевна (2007)) and personal experiences, 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 Timur and his squad

3.1 A Mansi FST and spellchecker

In 2025, a Mansi spellchecker has been released, created with a small set of morphophonological rules (32 twolc rules) and a lexicon consisting of 12,000 Mansi entries, as well as a larger set of proper nouns. Tested on a newspaper corpus consisting of approximately 700k tokens, the transducer was able to cover 98.9 % The transducer was turned into a spellchecker.

The Mansi grammatical model was reported to contain 825 continuation lexica and 12,063 stems with an additional set of over 145,000 shared lexemes at GiellaLT for the annotation of 100% equivalents of Russian names and toponyms (see Rueter, 2024). For a presentation, see Rueter et al. (2025).

3.2 Timur

The short novel titled Timur and his squad was written by Arkadiy Gaydar was first published in Russian in 1940. The book tells the story of a gang of village kids who sneak around secretly doing good deeds, protecting families whose fathers and husbands are in the Red Army. The novel had a huge impact on young Soviet audiences.

The short novel was translated into several languages, it was published in Mansi in 1955. According to the bibliographical data, the translation was made by E. Rombandeeva and A. Zyrin. As Zyrin's name is unknown in the history of Ob-Ugric Studies, while a certain Aleksandr Alekseevich Zyrin (1923-2003) appears to be a Turkologist-Orientalist-translator, it is safe to conclude that the Mansi translation, similarly to several other translations of the era, was the work of Evdokiya Rombandeeva only.

As the Russian original dates back to 1940, while the Mansi translation comes from 1955, the text of the short novel may be freely used for testing the Mansi FST without violating copyright law.

3.3 The experiment

The Mansi translation of Gaydar's short novel was digitalised. First, a physical copy of the book was scanned, then the documents were converted into text files with the help of Optical Character Recognition. The text file was proofread, and a copy of it was also manually normalised according to the orthography used in the Mansi press and contemporary educational publications. Both the original and the normalised versions of the text file were analysed by the Mansi spellchecker.

3.4 Results

Rombandeeva's original translation was published in an era of Mansi literacy where neither language-specific characters nor diacritics for long vowels have been in use, and the marking of palatalisation differed from present-day orthography as well. As a result of this, the spellchecker could recognise only wordforms that would not contain long vowels or special characters.

The adjusted version of the Mansi text performed much better. Most of the missing forms were analysed incorrectly due to the stems missing from the lexicon. The missing noun stems generally belong to the semantic field of the Socialist era, consisting of Russian loanwords (e.g. military terms, plant names), while the missing verbs are Mansi verbs that have not been mentioned in the newspaper corpus.

4 Other findings

In Mansi the deminutive suffixes -кве and -рищ can be added not only to nouns, but also to any other part of speech except conjunctions (Ромбандеева, 2017: 209). A verbal stem augmented with these suffixes can take tense, mood, or voice markers. (Riese, 2001: 59). The phenomenon is classified by Kálmán as precative mood (Kálmán, 1989: 61), Rombandeeva, and Riese, following her, call it merely a form. The suffix -кве expresses the speaker's respect, tenderness, as well as joy, delight, pride, and pleasant emotional disposition toward the surrounding environment and objects (Ромбандеева, 2017: 268). According to Kálmán, the -ке- variant of the suffix appears in front of other suffixes (Kálmán, 1989: 61), while Rombandeeva regards the form a dialectal alternation (Ромбандеева, 2017: 268). Beside occasional mentioning of the phenomenon, no detailed paradigm of the possible forms is given, also other examples than nouns or verbs are missing. In the translation of Gaydar's novel, an example documents the suffix -кве on the negative particle.

Ӑтикве! — лӯньщалтахтынӗтэ ёл-пувим, ӗл-оим Женя лӑвыс.

ӑти-кве лӯньщалтахты-нӗ-тэ ёл-пув-им, ӗл-о-им Женя лӑв-ыс

no-Prec cry-PtcpIpfv-Px3Sg hold.back-PtcpPrf, run.away-PtcpPrf Zhenya say-Pst.3Sg

'No! – said Zhenya, holding back her sobs and running away.'

According to Rombandeeva, when the suffix is used in the imperative, it denotes a polite request, a desire rather than an imperative (Ромбандеева, 2017: 269). Despite the Rombandeeva's short description, precative suffixes rather appear to follow mood markers than preceeding them.

Китыт телеграмма ӧс воекелн, юрт ӧйкакве.

китыт телеграмма ӧс во-е-ке-лн, юрт ӧйка-кве.

second telegram too take-Imp-Prec-2Sg.Sg, friend old.man-Prec

'Take the second telegramm too, comrade.'

According to Rombandeeva, the Mansi existential negative particle ӑтим is the variant of the negative verb ӑти (Ромбандеева, 2017: 30), Murphy describes the situation the other way round (Murphy, 1968: 225), while Wagner-Nagy regards the two forms separate (Wagner-Nagy, 2011: 203). Neither of the authors mention that apparently ӑтим too can take the precative suffix -кве.

Турманыг та ӗмтыс, ӗтимас, а тав акваг та ӑтимакве.

турман-ыг та ӗмт-ыс, ӗтим-ас, а тав акваг та ӑтим-акве.

dark-trs thus become-pret-3Sg become.night-pret-3Sg but 3Sg still thus no-prec

'It's already became dark, night has fallen, still she is nowhere to be seen.'

Albeit minor details, these forms of negation and precative conjugation, were previously unattested in academic literature, also they appear more frequently than anticipated, as their use was considered to be marginal since the second half of the 20th century.

5 Conclusion

The Mansi FST and spellchecker, created by GiellaLT infrastructure, was aimed at language learners and language practitioners who would use the contemporary orthography of the Mansi language, and would create or process texts of the Mansi newspapers. The spellchecker provided excellent performance during the task. The experiment presented in this paper shows that the analyser and spellchecker achieve good results with texts from other genres too, although minor modifications to the model are still needed. While texts using older orthography can be adapted to the contemporary rules automatically rather than manually, the language model of the Mansi transducer requires extension with additional grammatical categories, and the lexicon, particularly the list of verbs, needs to be complemented.

Although the literary text corpus seems to be overshadowed by the newspaper corpus due to its size, the translation of youth literature was the second most common literary genre and the third most common genre of Mansi written texts. Moreover, as the experiment has proven, it supplements the language description both in terms of vocabulary and grammar. Processing literary texts with the Mansi FST proves to be beneficial, as such studies can lead to the discovery of previously undocumented grammatical phenomena.

References

Csilla Horváth. 2020. The vitality and revitalisation attempts of the Mansi language in Khanty-

Mansiysk. University of Szeged, Szeged.

Csilla Horváth. 2024. From the kitchen to pop culture: The role of Mansi heritage speakers in language shift and language revitalisation. Faits de langues, 54:197–212.

Csilla Horváth. 2025. The role of Ob-Ugric native speakers and heritage language speakers in creating Khanty and Mansi print, broadcast and social media. In Minority Language Media, pages 239–262. Palgrave Macmillan.

Bela Kálmán. 1989. Chrestomathia Vogulica. Budapest.

Lawrence W Murphy. 1968. Sosva Vogul grammar. Indiana University, Bllomington.

Timothy Riese. 2001. Vogul, volume 158 of Languages of the world Materials. Lincom Europa, München - Newcastle.

Jack Rueter. 2024. Testing and enhancement of language models (transducers) from GiellaLT (scientific blog). Scientific blog.

Jack Rueter, Csilla Horváth, and Trond Trosterud. 2025. A mansi fst and spellchecker. In Proceedings of the 9th Workshop on Constraint Grammar and Finite State NLP, pages 163–182, Tallinn. University of Tartu Library.

Susanna Virtanen and Csilla Horváth. 2023. Mansi. In The Uralic languages, 2nd edition, pages 665–702, London. Routledge.

Beáta Wagner-Nagy. 2011. On the Typology of Negation in Ob-Ugric and Samoyedic Languages, volume 262 of Suomalais-Ugrilaisen Seuran Toimituksia. Suomalais-Ugrilainen Seura, Helsinki.

Евдокия Ивановна Ромбандеева. 2017. Современный мансийский язык: Лексика, фонетика, графика, орфография, морфология, словообразование. Формат, Tymen.

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