A Mansi FST and spellchecker

Csilla Horváth
Helsinki University
csilla.horvath@helsinki.fi

Jack Rueter
Helsinki University
jack.rueter@helsinki.fi

Trond Trosterud
UiT The Arctic University of Norway
trond.trosterud@uit.no

Abstract

The article presents a finite state transducer and spellchecker for Mansi, an Ob-Ugric Uralic language spoken in north-western Siberia. Mansi has a rich but mostly agglutinative morphology, with a morphophonology dominated by sandhi phenomena. With a small set of morphophonological rules (32 twolc rules) and a lexicon consisting of 12,000 Mansi entries and a larger set of propernouns we were able to build a transducer covering 98.9 % of a large (700k) newspaper corpus. Being a part of the GiellaLT infrastructure, the transducer was turned into a spellchecker. The most common spelling error in Mansi is the omission of length marks on vowels, and for the 1000 most common words containing long vowels, the spellchecker was able to give a correct suggestion as top-five in 98.3 % of the cases, and as first suggestion in 91.3 % of the cases.

1 Introduction

The article presents a finite state transducer and spellchecker for Mansi. Section 2 presents the Mansi language, its orthography and its grammar. Section 3 is the main part of the article, it presents the Mansi grammatacial model, discusses how it was made, and gives an evaluation of its performance. Section 4 presents and evaluates the Mansi spellchecker. Finally comes a conclusion.

2 Background

2.1 The Mansi language in society

Mansi is a severely endangered minority language, spoken mainly by the Mansi, an indigenous people of the Russian North. Genetically, it belongs to the Ob-Ugric branch of the Uralic language family. According to the latest Russian census data, approximately 1,000 people claimed to use the Mansi language (Перепись, 2020a), the vast majority of speakers reside on the territory of the Khanty-Mansi Automous Okrug - Yugra and the Sverdlovsk Oblast (cf. Перепись (2020b)). Four Mansi dialect groups were documented in the nineteenth century, each of which had several subdialects. While other dialect groups have become moribund, then extinct during the 20th century, varieties of the Northern Mansi dialect group are still in use, both in spoken and written form (cf. Virtanen and Horváth (2023)). The Mansi literary standard is based on the Sosva variety of Northern Mansi.

Regarding the status of the language, Mansi is an indigenous minoritised language spoken in Western Siberia, it has no official status, not at the regional nor at the municipal level. Mansi plays a minor role in its Russian-dominated, multi-ethnic and multilingual environment. Its situation is heavily affected by the loss of the traditional way of life and by rapid urbanisation. Mansi is barely present in official or semi-official domains, such as legislation, public transport or street signage, and due to its low economic significance, Mansi is also absent from the business sphere and only plays a marginal role in the labour market. 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, forthcoming)).

2.2 The development of the Mansi orthography

The Mansi language, in a way similar to that of other indigenous languages of the Russian North, has a history of literacy spanning almost a century. The first publications appeared in 1931, and

Proceedings of the 9th Workshop on Constraint Grammar and Finite State NLP, pages 32–37

originally Mansi was written in a Latin-based alphabet. This changed, however, with the transition to a Cyrillic-based alphabet in 1937 (Chernetsov, 1937, 168). Since 1937, the Mansi writing system has undergone minor changes. In the earliest period, the Cyrillic transcription contained no special characters, and vowel length was not marked either (as e.g. in Chernetsova (1938)). Later, a special character was introduced to denote the velar nasal (as e.g. in Balandin and Vakhrusheva (1972)), and vowel length has been marked with diacritics since the 1980s (as e.g. in Rombandeyeva et al. (1985). Currently, two slightly different variants of Mansi orthography are in use, one used in some of the academic and pedagogical publications (dictionaries, traditional schoolbooks), the other used in all other work, including print and broadcast media, social media, even schoolbooks designed for heritage language learners (Virtanen and Horváth, 2023, 667). For a more in-depth discussion of Mansi orthography, see Bradley and Skribnik (2021).

2.3 Mansi grammar

Mansi is a Uralic language, spoken mostly in Western Siberia. Typologically it forms a Sprachbund together with the neighbouring Khanty and Northern Samoyed languages, showing similar traits especially within the morphology, but also within syntax and morphophonology. It has a vowel system consisting of six vowels, each with a phonologically distinct short and long vowel, and an (only short) schwa. Mansi shows no vowel harmony and almost no vowel or consonant stem alternations. The morphophonological processes involved in Mansi inflection are mainly stem adjusting processes resulting from suffixes being attached to vowel or consonant stems.

Pronouns are inflected for 5 grammatical cases. The nouns have a six morphological (mainly adverbial) cases and a possessive declension. Verbs are inflected both for subject number and person and for object number, as well as for tense, mood and diathesis. Both nouns and verbs inflect for singular, dual and plural. There are also infinite forms, infinitive, gerund and participles.

Mansi is predominantly SOV, with adverbials allowed preverbally. The word order is not rigid, the verb may also be found sentence-initially in order to give it focus.

3 The Mansi Finite State Transducer

The Mansi grammatical analyser is made as two finite-state transducers, where the lower side of the *lexical* one corresponds to the upper side of the 74.40

The grammatical analyser is modeled in the GiellaLT infrastructure. For a presentation, see Moshagen et al. (2023).

3.1 Lexicon

At present, the Mansi grammatical model contains 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).

The attestation of actual Mansi words required a consensus. On the basis of the word forms found in our newspaper corpus, we concluded that at least all words with Mansi morphology would be treated as Mansi words. One of the original issues had been that there was a large portion of the text in quotes, so it was difficult to establish which word forms were being used in context.

Despite previous work with the vocabulary, it soon became apparent that verbal conjugation and noun declension paradigms often had more than one variant per cell of morphological analysis. The Mansi word for 'house' *kol* has two forms to represent the singular nominative form with first person dual possessive marking, e.g., *колмēн*, *колмен*.

Even though many of the variations became apparent in short versus long vowels, there were also instances where verbs with <y> stems in the infinitive took <a> stem variants. We do not want overlapping paradigms with multiple identical forms, which would be the result of simply joining multiple paradigms for a single verb type. Since duplicate identical interpretation defeats the advantage of fostering rule-based concise morphology, we limit the description of additional paradigmatic cells to precise annotation where a descriptive analyser will identify any extra forms. Thus, our work continues here with designing optimal representations of verb and noun inflection types that avoid duplication of individual forms. This work is carried out with full-scale test paradigms for each individual inflection type.

3.2 Morphology

Mansi verbal morphology includes the detachment of prefixes from their verbal stems when negative

particles are introduced with other possible particles. This entails the use of so-called flag diacritics, which are used in the description of languages with non-adjacent collocated morphology[1]

We use flag diacritics in the description of collective paired nouns in Mansi. Collective paired nouns tend to appear in combinations of kin terms. Such words are the equivalents of 'children' *āгитпыгит* (lit. *girls-boys*) and 'my parents' *омагумātягум* (lit. my.mothers-my.fathers). In both instances, the first and second components take identical morphology, i.e., in the word for 'children' the word for 'girl' *āги* takes the nominative plural marker *m*, which is repeated on the word for 'boy' *пыг*. The flag diacritics disallow any analyses other than tandem, identical readings. The word for 'parents' is rendered according to the same requirements, but here 'mother' *ома* takes morphological marking for nominative dual with a first person singular possessor.

```
омагум-ātягум
ома+N+Du+Nom+PxSg1+Cmp/Coll+Err/Orth-no-hyphen
+Cmp#ātя+N+Du+Nom+PxSg1
```

The flag diacritics used in the description of verbs with detachable prefixes are used at two points in the continuation lexica. First, they are given with the entire lemma with correlation to the verb prefix in the stem, and they have a continuation lexicon to allow for work with orthography, i.e., hyphenation or not, and the possibility for negation. The next continuation lexicon then provides for joining the prefixes to individual stems. This is accomplished with a strategy involving diacritic flags based on the value of the individual prefixes, as there are fewer prefixes than main verbs.

3.3 Morphophonology

Morphophonological processes were treated in twolc, where initially stem-alternating processes are described according to the shape of the stem and the affixes. In order to take control of this variation, meta-symbols were added at the stem and affix boundary, partly also to the suffixes. The following rule deletes the initial suffix vowel *а/я* whenever the stem is marked with the %VO%: trigger.

```
"%{аяØ%}:0"
%{аяØ%}:0 <=> %{VO%}:  %> _  ;
```

[1] see, e.g., pair verbs in Komi-Zyrian Rueter et al., 2021, reduplication in Lushootseed Rueter et al., 2023.

Triggers are used to describe the stem-final phonology of a word and help in the realisation of desired changes in the stem and suffixes. In this manner, the designer can choose which spellings or misspellings are derived by using continuation lexicon strategies.

The traditional description of Mansi nouns divides them into five distinct groups (c.f. Riese (2001) and Rombandeeva (2017)). At a first glance, Mansi nouns look like they might be described as a single set of words. As newcomers to Mansi language description, however, we aligned our approach to what tradition dictated. The five stem types are divided according to the way they end: (1) stems ending in the vowel *i*; (2) stems ending in other vowels; (3) stems ending in one consonant, (4) stems ending in consonant clusters and (5) stems with syncope. Also, types 3, 4 and 5 have variation according to the palatalisation of the final consonant. This simple breakdown, in fact, is not the entire picture: Syncope only applies to the high non-labial vowel with a single following consonant. Phonologically, this vowel is a schwa, but in Mansi orthography it is written with either <ы> or <y>. The number of stem-final consonant may be directly related to the presence of a vowel at the onset of some suffixes, and palatal versus non-palatal is an important factor when considering the suffix onset. To this end, we placed morpheme boundary triggers describing the word-stem phonology. Not all patterns are consistent with usage, for example, there are two types of syncopating stems, one is soft *SYNCS* and the other is hard *SYNCH*, but there are also words that might fit the syncope patterns that do not syncopate. Here we use *NOSYNCS* and *NOSYNCH* triggers as distinct from *VCS* and *VCH*, so we will be able to develop the modelling needed in the generation and analysis of misspellings in the use of syncope.

At Giellalt, it is encouraged that code and strategies be reused where possible. In practice, this means that time can be saved by applying solutions already found and applied in other language projects. Thus, a meander occurred in the development of the verbal paradigm with regard to the tagging of third person object marking on verbs.

3.4 Building the transducer

Our team consisted of people with professional knowledge of the target language, vast experience in the implementation of finite-state description,

testing, and spell checker development. This has meant that our contributions to the development of the analyser stem from complementary collaboration and the establishment of a mutual work flow. The language professional provides extensive paradigms for words in the language. The finite-state description is written by one researcher in constant consultation with the language professional and tester. The tester and spellchecking specialist leads the group, produces lists of lexemes not recognised by the analyser, and in collaboration with the other two workers helps to establish enhanced workflow strategies, such as having the language specialist make analysis notes for the lexemes misspelled most frequently. This is accomplished by remote meetings every other week, but it does not prevent the workers from contacting each other more often. Mansi native speakers assisted the team's work only occasionally, when explaining the unknown word forms, missing from the existing dictionaries.

Test paradigms are written for words representative of specific word classes. Nouns, for example, are divided according to traditional morphological descriptions, so that the resulting analysers can best fit the established norm. Since no one writes texts perfectly every time, and we hope the coding efforts will lead to invigoration in the language community with better perspectives in the future, we design the analyser so it will also recognise words regardless of their inconsistent spellings. In Northern Mansi, there are several factors contributing to misspelling. They stem from changes in the orthography involving the palatal *s*, an underdocumented use of long and short vowels, and multiple values for some cells in the paradigms. This is positive and means the orthography is still developing and will be for a number of decades to come. Our job is to make a description that allows writers and other language leeway.

Working as a group helps us to grow, especially, when we are trying to teach a new developer to become more self-sufficient, and when everyone is trying to keep language-independence at an optimal level. In Mansi, our solution for the orthographic representation is to use precomposed letters where-ever possible. This is due to the low development of UNICODE in the Cyrillic range, i.e., there are only two Cyrillic vowels precomposed with macrons. As such, we can only use the precomposed *ӯ* and *ӯ*, whereas the other long vowels are simply combinations with *U+0304*.

The absence of precomposed long vowels has meant that some corpus work and earlier code has been done using characters from the UNICODE Latin range (usually researchers), or non-UNICODE characters (media facilitators). To solve this problem, a keyboard specifically for Mansi that used precomposed Cyrillic-range characters where possible was built by Trond Trosterud. Our finite-state description of Mansi, as is the case with other languages, utilizes spell-relax strategies whereby Latin-range characters or non-standard character combinations can be recognized as their look-alike standard forms.

3.5 Evaluation

The development of the Mansi FST was done by testing against a newspaper corpus of 700000 words (Horváth et al., 2017). At present, the analyser recognises **98.86 %** of the words in the newspaper corpus. This impressive result is somewhat weakened by the fact that the analyser was developed on the same corpus. In contrast, however, words were added to the corpus based on their grammatical properties. Newspaper corpora, by their very nature, contain a vocabulary spanning many genres.

Proper nouns are a challenge for any language model. This grammatical model contains a language-independent set of 140000 names[2]. Restricting the test to words with initial capital letter (sentence-intitial words and names) weakens the coverage result from 98.86 % to 94.17 %. A weaker result is as expected, since names belong to an open category. Most of the missing words were either Russian words (Территории, 'territories', Утро, 'morning') or local names (Кантык-Ях).

Although our test corpus is both large (for an indigenous language) and representative for literary language use, it would no doubt have been relevant to test the speller against an unseen corpus. Unfortunately all available unseen text contained so much OCR errors that they made meaningful testing impossible.

4 Practical tools

The Mansi language model has been implemented as a spell checker, both online[3] and in Microsoft Word, 1.

[2] `https://github.com/giellalt/shared-urj-Cyrl/`
[3] https://divvun.org/proofing/online-speller.html

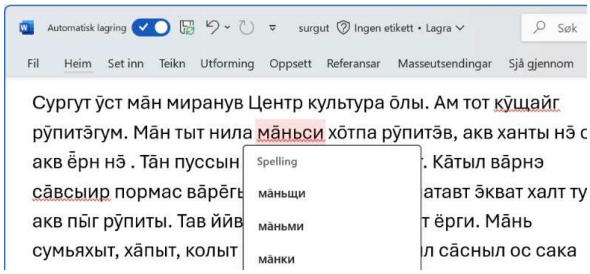

Figure 1: Mansi MS Word speller.

The overwhelmingly most common error in Mansi is use of the length mark for long vowels; the macron is often omitted where it should be or added where it does not belong. In order to test the spellchecker for its ability to correct length errors, we collected the words in the newspaper corpus containing long vowels, all in all 338937 words. In order to test the suggestion mechanism, we counted only the word forms that had an analysis, 12684 unique word forms. The long vowels were shortened, and pairs where shortening the vowel resulted in an existing word were removed. The resulting test suit contained 11064 word pairs. We tested both the full list and the list containing the 1000 most common long vowel words. The result is shown in table 1.

The result shows that the spellchecker is indeed capable of correcting this error type. Interestingly, the results from the most common words are better than for the whole material, this is probably because the rarer words were longer and therefore offered more possibilities for corrections.

Test	Words	1st pos	Top-5
Short-long	1000	91.30	98.30
Short-long	11064	86.77	96.72

Table 1: Testing error correction

5 Conclusion

Mansi is a language with a rich morphology but with relatively simple morphophonological processes. The main problem when modeling Mansi was to handle long-distance dependencies linked to prefixing, this was done with flag diacritics. As a result of this, a transducer with a relatively simple morphophonological component (32 rules) and a small lexicon (13.000 entries) and a large set of Russian and international names (145.000 entries) was able to give a text coverage above 98 %.

The transducer was turned into a spellchecker, and for the most common error type (omission of vowel length), it gave very good results, for the most common words 98.3 % of the suggestions were among the top-5 suggestions and 91.3 % were first suggestions. For other error types the results were not that good, here more work is needed.

Acknowledgments

Thanks to Sjur N. Moshagen for helping out with the implementation of the Mansi keyboard and proofing tools. Thanks to Mansi consultants for helping out with explaining unclear wordforms.

References

Aleksey Balandin and Matra Vakhrusheva. 1972. *Маньси букварь*. Просвещение, Ленингад.

Jeremy Bradley and Elena Skribnik. 2021. The many writing systems of Mansi: challenges in transcription and transliteration. In *Multilingual Facilitation*, pages 12–24.

Valeriy Chernetsov. 1937. Мансийский (вогульский) язык. In *Языки и письменность народов Севера I.*, pages 163–182, Москва, Ленинград. Государственное учебно-педагогическое издательство.

Irina Chernetsova. 1938. *Ловинтан магыс книга*. Государственное учебно-педагогическое издательство Наркромпроса, Ленингад.

Csilla Horváth. 2020. *The vitality and revitalisation attempts of the Mansi language in Khanty-Mansiysk*. Phd thesis, 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, forthcoming. 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*. Palgrave Macmillan.

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*, pages 56–65, St. Petersburg, Russia. Association for Computational Linguistics.

Sjur Nørstebø Moshagen, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. *The GiellaLT infrastructure: A*

multilingual infrastructure for rule-based NLP, volume 2 of *NEALT Monograph Series*, pages 70–94. NEALT.

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

Evdokija Ivanovna Rombandeeva. 2017. *Современный мансийский язык: Лексика, фонетика, графика, орфография, морфология, словообразование*. Формат, Tymen.

Evkodiya Rombandeyeva, Matra Vakhrusheva, and Klavdiya Saynakhova. 1985. *Māньси лāтынг*. Просвещение, Ленингад.

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

Jack Rueter, Mika Hämäläinen, and Khalid Alnajjar. 2023. Modelling the reduplicating Lushootseed morphology with an FST and LSTM. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (Americas-NLP)*, pages 40–46, Toronto, Canada. Association for Computational Linguistics.

Jack Rueter, Niko Partanen, Mika Hämäläinen, and Trond Trosterud. 2021. Overview of open-source morphology development for the Komi-Zyrian language: Past and future. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, page 62–72, United States. The Association for Computational Linguistics. International Workshop on Computational Linguistics for Uralic Languages 2021, EWPRF 2021 / IWCLUL 2021 ; Conference date: 23-09-2021 Through 25-09-2021.

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

Всероссийская Перепись. 2020a. Владение языками и использование языков населением. Part 5, Table 4. Accessed: 2023-04-01.

Всероссийская Перепись. 2020b. Владение языками и использование языков населением. Part 5, Table 17. Accessed: 2023-04-01.