Mari morpheme order revisited: a corpus-based analysis

Luan Hammer and Jeremy Bradley
University of Vienna

Abstract

Morpheme order in Mari declension has been extensively studied in the past, but attempts to explain the large amounts of alternation found here have been constrained by the difficulty of accessing sufficient data to properly elucidate the complexities in this domain. The paper at hand examines the prospect of using the Corpus of Literary Mari, created by an international workgroup around Trond Trosterud and his colleagues and hosted by Giellatekno, and other recently published resources on Mari to efficiently access vast amounts of data to quantitatively study this subject in a manner that had not previously been possible.

Keywords: Corpus, Mari language, suffix order, possessive suffixes

1. Mission statement

The unusually large freedom the Mari language(s) of European Russia afford their speakers as regards the arrangement of case (CX), possessive (PX), and number suffixes (NX) in nominal morphology have long puzzled and perplexed scholars and language learners alike. While not all six possible arrangements of these are permissible, substantial variation can be encountered:

(1) Meadow Mari (Corpus of Literary Mari)

a. *joltaš-em-βlak-lan*	b. *pire-βlak-et-lan*	c. *joča-βlak-lan-že*
friend-1SG-PL-DAT	wolf-PL-2SG-DAT	child-PL-DAT-3SG
‘to my friends’	‘to your wolves’	‘to his/her/theirSG children’
(PX-NX-CX)	(NX-PX-CX)	(NX-CX-PX)

Reference materials (including those co-authored by authors of this paper) will often try to handwave this alternation away by labelling competing forms as equivalent, but linguists and language teachers describing multiple forms as equivalent are generally tacitly admitting (intentionally or not) that they simply do not understand the factors governing the alternation – be they on the level of morphosyntax, semantics, information structure, or dialectology.

Jorma Luutonen’s 1997 dissertation “The variation of morpheme order in Mari declension” is an impressive attempt to elucidate this question. It gives an exhaustive and satisfying overview on the constraints in this domain and gives a comprehensive quantitative overview of the respective frequencies of different arrangements, though the factors governing this variation remained elusive – as they do today.

More than two decades later, this question deserves re-examination due to revolutions in the quantitative study of morphologically rich minority languages, driven forward especially by the Tromsø-based Giellatekno work group under the leadership of Trond Trosterud. While Giellatekno’s foundational mission pertains to guaranteeing language technology for the Saami languages, its open infrastructure – and especially the broad horizon of its founder and leader – has allowed other language and scholarly communities to profit from its framework. For Mari, this has meant the creation of the Corpus of Literary Mari (henceforth CLM) by a work group including Trond Trosterud, Jorma Luutonen, and many others, a first release of which, covering 57.38 million tokens of Meadow Mari texts, was soft launched in December 2020 at gtweb.uit.no/u_korp/?mode=mhr, with meta information on the project and instructions published at corpus.mari-language.com. Texts are taken from non-fiction texts, fiction texts, legal texts, scientific texts, news texts and Wikipedia texts, and they span over a century of Mari literacy (1912–2018). In this paper, we will explore the prospect of using new corpus infrastructures as novel tools to revisit enigmas

 Nordlyd 46.1: 59–74, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6373

such as the alternation in suffix order in Mari declension. Given that this should only be understood as a pilot study, we are currently restricting ourselves to the dominant Meadow Mari literary norm, which is better covered by the corpus at this point than the secondary literary norm, Hill Mari.

2. Variation in suffix order in the Meadow Mari nominal paradigm

Mari is a Uralic language, or cluster of languages, spoken by several hundred thousand speakers mainly in the titular Republic of Mari El and by a sizable diaspora in the Ural Mountains. Most reference materials and dialectal surveys distinguish between four dialect groups (Meadow, Hill, Eastern, Northwestern); two literary norms (Meadow, Hill) are widely used today. There is an ongoing debate among native speakers and linguists on the status of Hill Mari, if it should be considered a language or a dialect – cf. Trosterud & Alikov (1994) for a discussion of this problem and a comparison with the discourse surrounding Bokmål and Nynorsk in Norway. This chapter will restrict itself to the Meadow Mari literary norm; all language data in this chapter is in Meadow Mari.

Mari is, together with neighbouring Uralic and Turkic languages, generally classified as a member of the *Volga-Kama Sprachbund* in which a large amount of linguistic convergence can be observed. Mari shows not just lexical but also significant structural influence from two Turkic languages, Chuvash and Tatar. The dominant contact language today is, unsurprisingly, Russian; Russian structural influence on Mari is salient as well.

As is typical of the languages of this region, Mari morphology is rich and highly concatenative, i.e., morpheme boundaries can almost always be easily and unambiguously identified; portmanteau morphemes are virtually non-existent. Mari morphosyntax does not have dual forms. Nominal stems can take number suffixes (e.g., *olma* ‘apple’ > *olma-βlak* apple-PL ‘apples’, see below about other plural markers), case suffixes (e.g., *olma-lan* apple-DAT ‘to the apple’), possessive suffixes (e.g., *olma-na* apple-1PL ‘our apple’), and word-final clitics (e.g., *olma=t* apple=ADD ‘also the apple’). It should be noted that the 3SG and 2SG possessive suffixes have secondary usages where they serve more as determining elements indicating a topic, a contrast, or that something is part of a salient whole (cf. Simonenko 2014, Riese et al. 2019: 60); in this function the possessive suffixes behave like clitics and generally appear in word-final position; they can also be used in combination with other possessive suffixes in possessive function (e.g. *ačа-t-še* father-2SG-3SG ‘as for your father’, CML).

The permissible combinations of suffixes belonging to these types (excluding clitics) and alternation encountered in the language are illustrated in Tables 1 and 2, with Table 1 showing the Mari noun *olma* ‘apple’ in all cases (singular and plural) and Table 2 adding the possessive suffix of the second person singular to it. This data was generated using paradigm.mari-language.com, with some forms added that, thanks to the Corpus of Literary Mari, we can assume to exist. It should be noted that the permissible combinations and their frequencies differ between different possessive suffixes, i.e., one should not extrapolate from the possessive suffix 2SG to other possessive suffixes.

Case	Singular	Plural
Nominative	olma	olma-βlak
Genitive	olma-n	olma-βlak-ə̂n
Dative	olma-lan	olma-βlak-lan
Accusative	olma-m	olma-βlak-ə̂m
Comparative[1]	olma-la	olma-βlak-la
Comitative	olma-ge	olma-βlak-ge
Inessive	olma-šte	olma-βlak-ə̂šte
Illative	olma-š(ke)	olma-βlak-ə̂š(ke)[2]
Lative	olma-š	olma-βlak-eš

Table 1: Paradigm of the Mari noun *olma* ‘apple’

Case	Singular	Plural
Nominative	olma-t	olma-t-βlak ~ olma-βlak-et
Genitive	olma-t-ə̂n	olma-t-βlak-ə̂n ~ olma-βlak-et-ə̂n
Dative	olma-t-lan ~ olma-lan-et	olma-t-βlak-lan ~ olma-βlak-et-lan ~ olma-βlak-lan-et
Accusative	olma-t-ə̂m	olma-t-βlak-ə̂m ~ olma-βlak-et-ə̂m
Comparative	olma-t-la ~ olma-la-t	olma-t-βlak-la ~ olma-βlak-et-la ~ olma-βlak-la-t
Comitative	olma-t-ge	olma-t-βlak-ge ~ olma-βlak-et-ge
Inessive	olma-št-et	olma-βlak-ə̂št-et ~ olma-t-βlak-ə̂šte
Illative	olma-šk-et	olma-βlak-ə̂šk-et ~ olma-t-βlak-ə̂š(ke)
Lative	olma-š-et	olma-βlak-eš-et ~ olma-t-βlak-eš

Table 2: Paradigm of the Mari noun *olma* ‘apple’ with possessive suffix 2Sg

Different suffix orders depending on the specific case, as encountered here, is not exclusive to Mari; it can also be encountered for example in Moksha Mordvin, Permic, and Southern Samoyedic (Honti 1995, Tauli

[1] The comparative case denotes a likeness or similarity (e.g., *olma-la* apple-CMPR ‘like an apple’) and is not to be confused with the comparative degree (e.g. *joškargə̂-rak* red-COMP degree ‘redder’). Presumably motivated by this terminological confusion, other sources refer to this case as ‘equitative’, ‘similative’, or ‘modal’.

[2] The illative suffix has a short form and a long form; the short form can only be used if the illative suffix is the final suffix attached to a stem.

1953). What is unusual, however, are the many degrees of freedom, esp. in the dative and comparative cases.

While Mari generally uses plural markers relatively sparingly, with formally singular forms being used when plural semantics can be assumed from context (cf. Riese et al. 2019: 55–57), four different plural markers can be used in literary Meadow Mari. In addition to the plural suffix *-βlak* that is shown in the tables, Mari has plural forms in *-šamə̑č*, *-mə̑t*, and *-la*. The distribution of *-βlak* and *-šamə̑č* seems to be governed by dialectal distribution, with both forms used in literary language today. The other plural markers have clearly defined distinct functions: *-mə̑t* is an associative/heterogenous plural marker 'x and those associated with x, x & co.', e.g., *ača-mə̑t* father-PL.ASS 'father and those with him'. This suffix can be preceded by possessive suffixes (e.g., *ača-t-mə̑t* father-2SG-PL.ASS 'your father and those with him') but only possessive suffixes used in non-possessive functions can follow it (*Irina-mə̑t-še* Irina-PL.ASS-3SG 'as for Irina and co.', *ruš spekul'ant-mə̑t-et* Russian profiteer-PL.ASS-2SG 'that Russian profiteer and his buddies', CLM). The plural in *-la* on the other hand is primarily usually used with inanimate nouns in a local meaning in combination with local case endings and spatial adverbs (e.g., *pört-la-šte* house-PL.LOC-INE 'in the houses'). While it is uncommon for this plural suffix to co-occur with possessive suffixes outside of their non-possessive functions (Riese et al. 2019: 59), examples can be found in which the possessive suffix follows the plural and case suffixes (*ojlə̑maš-la-št-em* story-PL.LOC-INE-1SG 'in my stories', CLM).

Some fundamental observations can be made regarding restrictions in the ordering of suffixes, irrespective the large variation that can still be observed, as illustrated in the tables:

1. Case suffixes always follow number suffixes.
2. Possessive suffixes can either precede or follow number suffixes.
3. The genitive, accusative, and comitative suffixes occur in the final position; only clitics and possessive suffixes used as clitics in a non-possessive function can follow them (e.g., *mə̑j-ən-že*[3] 1SG-GEN-3SG 'as for mine', *βolgə̑də̑-m-žo* light-ACC-3SG 'as for the light', CML)
4. The local case markers – inessive, illative, lative – precede possessive suffixes. A rare exception is when a possessive suffix precedes the number suffix; in this case, the local case suffix can be encountered in final position technically following the possessive suffix (e.g., *užaš-ə̑že-βlak-ə̑šte* part-3SG-PL-INE 'in its parts', CML).
5. In the dative and comparative case, possessive suffixes can either precede or follow case suffixes.

As regards the variation, Jorma Luutonen's 1997 dissertation provides statistical data on the frequency of different variants. Elina Guseva and Philipp Weisser touched on the topic in an article where they explain the different morpheme order of structural and local cases in Meadow Mari by postsyntactic operations (Guseva & Weisser 2018). In general, existing case studies on the order of morphemes for specific languages mainly concern verbs (e.g., Rice 2000 for Athapaskan languages, Caballero 2010 for Choguita Raràmuri) and apart from the aforementioned works by Luutonen and Guseva & Weisser we are not aware of any case studies on suffix order variation that are dedicated specifically to nouns.

In this paper, we will revisit the variation described in the literature using new corpora of Mari, on the one hand to re-examine Luutonen's findings, on the other hand to look for possible additional information and explanations. The main research questions are: Which of the variants are used the most frequently? Are there differences between the frequencies of the use of the variants between different possessive suffixes or cases? Did the distribution of the variants change over time? What motivates the usage of one or another suffix order variant?

[3] It should be noted that the placement of the possessive suffix third person singular, here realized as *-že*, mirrors the placement of the homophonous Russian clitic *že*, which yields the possibility of a structural influence from Russian. However, other realizations of the possessive suffix, such as *-žo* in the following example – it can thus be reasonably assumed that Mari speakers perceive these as possessive suffixes of the third person singular, even if the range of usages has been influenced by Russian.

3. A closer look at Mari corpora

The Corpus of Literary Mari (CLM) was already introduced in the introduction. The second resource we have utilized are the Meadow Mari corpora compiled by Timofey Arkhangelskiy and his colleagues as part of a larger endeavour to create corpora for languages of the Volga-Kama Region (henceforth VK Corpora; other corpora as of the moment of this writing published by this work group cover Erzya, Moksha, Udmurt, Komi-Zyrian; a general overview can be found at volgakama.web-corpora.net); this resource consists of two basic corpora: a general corpus of literary texts (henceforth VK-MAIN), and the Social Media Corpus (henceforth VK-SMC) consisting mainly of texts curated from the platform *vkontakte* and also including Russian texts and ample amounts of code mixing (for further information on the development of the Social Media Corpora see Arkhangelskiy 2019). Table 3 shows the token counts, as of the compilation of this paper, of the three main resources at our disposal.

CLM	gtweb.uit.no/u_korp/?mode=mhr	57.38 million
VK-Main	meadow-mari.web-corpora.net/meadow-mari_corpus	5.53 million
VK-SMC (Mari only)	meadow-mari.web-corpora.net/meadow-mari_social_media	3.59 million

Table 3: Tokens in the corpora used as of 20 October 2021

Table 4 summarizes some main structural differences between CLM on the one hand, and the VK Corpora on the other.

CLM	VK Corpora
Mari lemma forms are provided as part of the analysis, no translations of these are provided.	Russian translations of lemmas are provided as a part of the analysis.
Words are tagged in the analysis, but not segmented.	The analysis indicates morpheme boundaries.
For morphologically ambiguous word forms, only the form deemed the most likely by the software is provided and found by the search engine.	For morphologically ambiguous word forms, all interpretations yielded by the software are provided and found by the search engine.
Allows for diachronic analyses.	Do not allow for diachronic analyses.
Does not allow for sociolinguistic comparisons, but allows for genre comparisons (e.g., fiction vs. non-fiction).	Allow for sociolinguistic comparisons (literary language vs. social media).

Table 4: Some attributes of the infrastructures used

As an additional tool for analysing data is Vienna-based Mari Web Project's (MWP) morphological analyser found at morph.mari-language.com which yields fully-fledged interlinearisations (including all ambiguity) showing: morph realization, base morpheme (due to the concatenative nature of Mari morphology mostly, but not always, the same as the realization), gloss (including English translation), part of speech/type of suffix. Table 5 illustrates how the same forms are annotated in the tools under consideration. Highlighting indicates correct interpretations of the form at hand.

Glossed by hand	CLM	VK Corpora	MWP
(1) йолташем-влаклан *joltaš-em-βlak-lan* friend-1SG-PL-DAT 'to my friends'	part-of-speech: noun grammatical analysis: N.Pl.Dat.PxSg1.So_PNC dependency relation: HNOUN baseform: йолташ	**йолташем-влаклан** **йолташ** N anim, hum йолташ-ем-влак-лан STEM-1SG-PL-DAT gr: pl, dat, 1sg trans_ru: товарищ; друг	йолташем-влаклан йолташ -ем -влак-лан *йолташ -ем -влак-лан* *friend* -1SG -PL -DAT no -poss -num -case
(2) ачатше ača-t-še father-2sg-3sg 'as for your father'	part-of-speech: noun grammatical analysis: N.Sg.Nom.PxSg2.Foc_Poss dependency relation: HNOUN baseform: ача	**ачатше** **ача** N anim, hum ача-т-ше STEM-2SG-3SG gr: sg, nom, 2sg, 3sg trans_ru: отец; свёкор	ачатше ача -т -ше ача -ет -жЕ father -2SG -3SG no -poss -poss
(3) пӧртыштем *pört-ə̑št-em* house-INE-1SG 'in my house'	part-of-speech: noun grammatical analysis: N.Sg.Ine.PxSg1.So_CP dependency relation: ADVL→ baseform: пӧрт	**пӧртыштем** 1. **пӧрт** N пӧрт-ыште-м STEM-LOC-1SG gr: sg, loc, 1sg trans_ru: дом 2. **пӧрт** N пӧрт-ышт-ем STEM-3PL-1SG gr: sg, nom, 3pl, 1sg trans.ru: дом 3. **пӧртыш** N пӧртыш-те-м STEM-LOC-ACC gr: case.comp, sg, loc, acc trans.ru: вертячка 4. **пӧртыш** N пӧртыш-те-м STEM-LOC-1SG gr: sg, loc, 1sg trans.ru: вертячка 5. **пӧрт** N пӧрт-ыште-м STEM-LOC-ACC gr: case.comp, sg, loc, acc trans_ru: дом	пӧртыштем пӧрт -ышт -ем *пӧрт -штЕ -ем* house -INE -1SG no -case -poss пӧртыштем пӧртыш -т -ем *пӧртыш -штЕ -ем* coenurosis -INE -1SG no -case -poss

(4) шомакланда *šomak-lan-da* word-DAT-2PL 'to/for your word'	part-of-speech: noun grammatical analysis: N.Sg.Dat.PxPl2.So_CP dependency relation: X baseform: шомак	**шомакланда** 1. **шомакланаш** V шомаклан-д-а STEM-CAUS-NPST.3SG gr: npst, 3, sg, caus, caus_t trans_ru: разговаривать; ругаться 2. **шомак** N шомак-лан-да STEM-DAT-2PL gr: sg, dat, 2pl trans_ru: слово	шомакланда шомак -лан -да *шомак* *-лан* *-да* *word* -DAT -2PL no -case -poss
(5) енланат *eŋ-lan=at* person-DAT=ADD 'also to/for a person'	part-of-speech: noun grammatical analysis: N.Sg.Cmpr.PxPl1.So_CP.Foc_at dependency relation: HNOUN baseform: ең	**ең-ланат** 1. **ең** N anim, hum ең-ла-на-т STEM-SIM-1PL-ADD gr: add, sg, sim, 1pl trans_ru: человек 2. **ең** N anim, hum ең-лан-ат STEM-DAT-ADD gr: add, sg, dat trans_ru: человек	ең-ланат ең -лан -ат *ең* *-лан* *-ат* person -DAT -and ad/no -case -enc ең-ланат ең -ла -на -т *ең* *-ла* *-на* *-ат* person -COMP -1PL -and ad/no -case -poss -enc ең-ланат ең -ла -на -т *ең* *-ла* *-на* *-ат* person -PL -1PL -and ad/no -num -poss -enc ең-ланат ең -ла -н -ат *ең* *-ла* *-н* *-ат* person -PL -GEN -and

Table 5: Annotation in the tools under consideration

Example (1) represents an unambiguous case: here all three infrastructures only return one (correct) interpretation. In the case of example (2) as well, the three infrastructures agree on the (only) correct interpretation of a word – a stem marked with two possessive suffixes, one in a possessive function and one in a non-possessive function – but it should be noted that CLM glosses the non-possessive 3sg suffix distinctively, as **Foc_Poss**. In example (3), MWP returns an interpretation that is technically admissible, but unlikely to an extent that its exclusion is desirable ('in my house' vs. 'in my coenurosis' – a type of tapeworm infection), while VK Corpora return several interpretations that are not admissible (e.g., those in which the inessive suffix is purportedly followed by the accusative suffix, which is not permissible in Mari grammar), but among them the correct interpretation. In (4), CLM and MWP return only the correct interpretation, while VK Corpora returns the correct interpretation beside an inadmissible one. In example (5), however, CLM returns only an incorrect interpretation, while VK Corpora and MWP return several interpretations, including the correct one.

This illustrates how from a user[4] perspective, a cross-integration of these three resources in which the strengths of all three infrastructures are combined (CLM's sturdy morphological model, VK Corpora's handling of ambiguity, MWP's cross-integration with a Mari-English lexicon and interlinearisations) would be desirable. Optimally users could choose if they wish their searches to account for ambiguity (i.e., include all possible interpretations as in VK Corpora and MWP) or not (i.e., include only the interpretation deemed most likely by the software as in CLM). It is situationally dependent if it is desirable for the data to be

[4] Here "user" is defined as "author(s) of this paper".

disambiguated by good, but not 100% reliable, mechanisms or not: for naïve users looking at common structures, this disambiguation is desirable, but for linguists looking at rare structures which one cannot assume to be disambiguated appropriately, it is not. Such linguists might instead have to rely on regular expressions (regular expression searches are supported by both CLM and VK Corpora) to look for certain endings, but this requires greater technical competence and greater familiarity with the Mari language.

These potential problems notwithstanding, both CLM and VK Corpora afford users with the possibility to search the grammatical tags shown in Table 5, and even though CLM only provides tagging and no division into morphemes, these tags include information on the ordering of morphemes (e.g., **So_PNC** for *person suffix* > *number suffix* > *case suffix*) and can thus be utilized for the task at hand.

4. Determining the frequencies

Upon a perfunctory investigation of our results, we could determine that many incorrect analyses throughout our survey were a function of frequent ambiguities pertaining to a small number of oftentimes very widely used lexemes that could also be interpreted as a shorter lexeme with a possessive suffix: *una-βlak* guest-PL 'guests' was misinterpreted as **u-na-βlak* new-1PL-PL 'our new ones', *ušem-βlak* 'unions' was misinterpreted as **uš-em-βlak* mind-1SG-PL 'my minds', *urem-βlak* 'street' was (presumably) misinterpreted as **ur-em-βlak* squirrel-1SG-PL 'my squirrels', *keremet-βlak* evil_spirit-PL 'evil spirits' was misinterpreted as **kerem-et-βlak* rope-2SG-PL 'your ropes', etc. To avert problems caused by these lexemes, we excluded them and several other ones that created ambiguity (*koman* 'layered', *sös* (element of old name for October) from all searches (e.g., for *kerem* 'rope' by adding to the search query that the baseform is not **керем**). This list is surely not exhaustive but based on the output of the unrestricted searches we can assume that they are by far the most numerous culprits of mistakes of this type.

4.1 Overall Frequencies

In this section we aim to provide raw data on the basic (absolute) frequencies of different arrangements. For this task CLM was used due to the vastly larger amounts of data.

Case suffix and possessive suffix

There is only variation for two grammatical cases here: the dative in *-lan* and the comparative in *-la*. The survey is exacerbated by the homonymity afflicting the relevant suffixes, with the comparative suffix notably being homonymous with the plural of local meaning *-la* (see Section 1). Table 6 illustrates the arrangement of the dative suffix and comparative suffix with different possessive suffixes as returned by the software. To find, for example, all purported occurrences of the dative suffix followed by the possessive suffix first person singular (i.e., the ordering CXPX), one must search for tokens where the grammatical analysis contains **Dat.PxSg1.So_CP**. In the third person singular for the ordering CXPX, one must also add results returned by **Dat** and **Foc_Poss** as not to miss possessive suffixes used in a non-possessive function. In many cases, however, the false positive results – i.e., results returned that do not actually show the desired structure – greatly outnumber the legitimate results, as can be quickly determined by perusing the outputs and especially by viewing the most commonly found forms. These fields are highlighted in grey. It must also be noted that false negatives – i.e., tokens that should fall into these categories but are erroneously classified as something else – are by default excluded in this overview.

	Dative		Comparative	
	PXCX	**CXPX**	**PXCX**	**CXPX**
1SG	11,409	460	1,650	14,292
2SG	8,396	485	1,318	265
3SG	46,270	4,643	5,148	2,353
1PL	1,518	1,124	77	17,030
2PL	331	2,157	304	160
3PL	7,880	403	456	3,624

Table 6: Dative and Comparative (singular), PXCX vs. CXPX

Due to the high number of erroneous results, in spite of the exclusion of problematic stems, care must be applied before drawing conclusions. While the raw data shows PXCX as more common than CXPX in dative with 1PL, the critical mass of results are misinterpreted (e.g., the systematic erroneous analysis of inflected feminine patronyms such as *Petrovna-lan* Petrovna-DAT 'to [given name] Petrovna' as **Petrov-na-lan* Petrov-1PL-DAT 'to our Petrov'), with only comparatively few correctly analysed forms (e.g., *el-na-lan* country-1PL-DAT 'to/for our country') – one can thus safely assume that, in the first person plural, CXPX outnumbers PXCX here in practice. For the comparative, erroneous results greatly outnumber legitimate results regardless of the search pattern (e.g. *salam* 'greeting; hello' erroneously interpreted as **sa-la-m* scythe-CMPR-1SG 'like my scythe'), but legitimate occurrences of the arrangement PXCX can be encountered, especially if one restricts oneself to animate nouns and especially kinship terms, e.g., *aβa-m-la* mother-1SG-CMPR 'like my mother', *kokaj-na-la* aunt-1PL-CMPR 'like our aunt', and as expected CXPX can be encountered in the third person singular when the suffix is used in a non-possessive function, e.g. *ruš-la-že* Russian-CMPR-3SG 'in Russian, on the other hand'.

Despite the uncertainties encountered here, the following conclusions can be drawn:

- In the dative, the arrangement PXCX is dominant in all persons but 1PL and 2PL, where CXPX is dominant. The greatest variation can be found in 3SG, where PXCX dominates greatly but an ample body of CXPX examples can be found. Here the alternation can be assumed to be determined by the function of the possessive suffix: when used possessively, PXCX is dominant, while non-possessively used suffixes occur in the arrangement CXPX.
- An explanation is required for the unusual, but not rare, deviations from the dominant suffix orderings as represented by forms such as *el-na-lan* country-1PL-DAT 'to/for our country'.
- Legitimate usage examples of comparative suffix with possessive suffixes seems to be rare, but it is easier to find examples of the arrangement PXCX than CXPX.

Number suffix and possessive suffix

As the associative plural in *-mə̂t* and the plural of local meaning in *-la* are not subject to any notable variation (see Section 1), our investigation here is restricted to the plural suffixes *-βlak* and *-šamə̂ć*. To find all cases of the possessive suffix first person singular co-occurring with a plural suffix but no case suffix (i.e., nouns in the nominative) and ordering PXNX, one must search for all tokens with a grammatical analysis containing **Nom.PxSg1.So_NP**. If one wishes to restrict oneself to one suffix or the other, one must furthermore search for tokens containing their Cyrillic realizations, **влак** or **шамыч**, as the tagging does not distinguish between them. As these suffixes are not subject to any allomorphy, this is not problematic. In the third person singular for NXPX, one must again add results returned by **влак** / **шамыч** co-occurring with **Foc_Poss** in the grammatical analysis to include non-possessive usages of this possessive suffix.

In general, this point of investigation is afflicted by considerably less ambiguity, given that the plural suffixes *-βlak* and *-šamə̂ć* are not subject to homonymity and given that they are orthographically preceded by a hyphen. Table 7 shows the frequency of the different arrangements for all possessive suffixes and both plural suffixes, with problematic stems excluded.

	-βlak		*-šamə̑ć́*	
	PXNX	**NXPX**	**PXNX**	**NXPX**
1SG	2,924	473	611	158
2SG	1,452	558	188	279
3SG	14,835	1,496	1,087	465
1PL	2,979	282	215	86
2PL	376	69	39	21
3PL	444	164	27	50

Table 7: Number suffixes and possessive suffixes, PXNX vs. NXPX, with problematic stems excluded

For both *-βlak* and *-šamə̑ć́*, the ordering PXNX is considerably more common than NXPX – with the exception of 2SG and 3PL in the case of *-šamə̑ć́*, where the ordering NXPX is considerably more common. A perfunctory analysis of the results suggests a possible explanation for 2SG: numerous results show the possessive suffix used in a non-possessive function, and non-possessive usage of these suffixes is known to coincide with the final placement of the suffix. It would thus suggest itself that the non-possessive usage of 2SG is typical of the very same dialects for which the suffix *-šamə̑ć́* is typical. We have no good explanation for NXPX in 3PL at this point.

Number suffix, Possessive suffix, case suffix

Even though we, based on previous research, had made some a priori assumptions on the inadmissibility of certain suffix arrangements, for the sake of completeness, here we will investigate all theoretically possible permutations of possessive suffixes, number suffixes, and case suffixes. Given the sparsity of data even in an exceedingly large corpus when looking at unusual combinations, we have not distinguished between the plurals in *-βlak* and *-šamə̑ć́* here. To find all examples of the genitive suffix followed by the plural suffix followed by the possessive suffix 1SG (i.e., the ordering CXPXNX), one must search for grammatical analyses containing **Gen.PxSg1.So_CPN**. As above, problematic stems are systematically excluded from the search queries. Given the small size of the output, we could manually investigate the results and separate appropriate findings from erroneous ones; Table 8 only shows the findings we considered correctly identified.

Once again, the comparative with its ending *-la* was the most problematic, specifically when looking at 1PL with the arrangement NXCXPX: the overwhelming mass of the 1,010 results returned by this query were false analyses of dative forms with the additive clitic *=at*, e.g., *pašajeŋ-βlak-lan=at* worker-PL-DAT=ADD ‘also to/for the workers’ was erroneously analysed as *pašajeŋ-βlak-la-na=t* worker-PL-CMPR-2PL=ADD ‘also like our workers’. To exclude all false analyses of this type (i.e., all word forms ending in *-lanat*), we had to utilize regular expressions, as there is not currently a “does not include” functionality in the search mask. Such an addition would be desirable. After the exclusion of these forms, none of the remaining forms were plausible comparative forms.

Only for the genitive, accusative, and dative cases could we find enough data to allow for a meaningful analysis. In all cases PXNXCX and NXPXCX are admissible, with the first variant dominating over the second. There seem to be significant differences in the ratio between the two arrangements depending on the person, but we do not at this point dare to assume if this is noise in the data or if these differences are significant and due to a functional explanation currently eluding us (e.g., the comparatively high share of NXPXCX forms in 2SG: does this alternation relate to the non-possessive usage of this suffix even when the suffix is not used in the final position in either arrangement?)

In the dative, examples of the arrangement NXCXPX can be found as well, especially in 3SG. Here it seems likely that the usage of this arrangement relates to the non-possessive usage of this suffix.

		PXNXCX	PXCXNX	NXCXPX	NXPXCX	CXPXNX	CXNXPX
GENITIVE	1SG	196	0	0	21	0	0
	2SG	46	0	0	22	0	0
	3SG	1,421	0	0	38	0	0
	1PL	725	0	0	67	0	0
	2PL	39	0	0	18	0	0
	3PL	40	0	0	13	0	0
ACCUSATIVE	1SG	517	0	0	60	0	0
	2SG	328	0	0	224	0	0
	3SG	4,024	0	0	643	0	0
	1PL	485	0	0	96	0	0
	2PL	106	0	0	31	0	0
	3PL	125	0	0	55	0	0
COMITATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	3	0	0	2	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
DATIVE	1SG	334	0	1	34	0	0
	2SG	148	0	3	37	0	0
	3SG	1,614	0	32	28	0	0
	1PL	252	0	5	1	0	0
	2PL	41	0	8	0	0	0
	3PL	68	0	0	8	0	0
COMPAR.	1SG	4	0	0	1	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	1	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
INESSIVE	1SG	0	0	4	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	7	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
ILLATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	1	0	0	0	0	0
LATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0

Table 8: All arrangements of PX, NX, and CX for all non-nominative cases

4.2 Diachronic Comparison

Our next point of investigation is changes in the observed frequencies over the course of time. Only CLM with its large time depth affords the means to do this in a meaningful way. For the pilot study at hand, we restricted ourselves at comparing texts from before and after the year 2000. Given the sparsity of unambiguous forms with many arrangements, we are restricting ourselves to comparisons for which we can find sufficient data to render a comparison meaningful. We also did not differentiate between the plural suffixes *-βlak* and *-šamə̑č* here.

Case suffix and possessive suffix

	Before 2000		After 2000	
	PXCX	CXPX	PXCX	CXPX
1SG	4,158	125	4,925	178
2SG	3,015	170	3,328	190
3SG	13,368	429	18,875	664
1PL	101	259	97	392
2PL	8	394	62	256
3PL	479	9	2,770	11

Table 9: PXCX ~ CXPX in the dative, before and after 2000

Number suffix and possessive suffix

	Before 2000		After 2000	
	PXNX	NXPX	PXNX	NXPX
1SG	1,324	287	1,944	269
2SG	611	298	882	346
3SG	4,257	523	9,913	669
1PL	1,694	230	3,115	102
2PL	93	24	280	53
3PL	216	97	204	83

Table 10: PXNX ~ NXPX, before and after 2000

Number suffix, Possessive suffix, case suffix

		Before 2000		After 2000	
		PXNXCX	NXPXCX	PXNXCX	NXPXCX
GENITIVE	1SG	81	14	215	5
	2SG	163	8	202	9
	3SG	305	14	950	15
	1PL	362	39	470	20
	2PL	8	14	25	3
	3PL	11	7	25	5
ACCUSATIVE	1SG	70	8	408	36
	2SG	212	70	302	107
	3SG	1,194	226	2,351	302
	1PL	455	53	1,051	31
	2PL	32	17	69	10
	3PL	55	20	54	21
DATIVE	1SG	114	13	233	15
	2SG	72	9	253	20
	3SG	451	10	985	12
	1PL	175	1	424	0
	2PL	11	0	27	0
	3PL	19	5	39	2

Table 11: All arrangements of PX, NX, and CX (excerpt), before and after 2000

The data indicates a decrease in variation over the passage of time, with PXNX having an 85% share before 2000, but a 91% share after 2000. We are not confident that other trends that can be observed are particularly noteworthy, given the small amount of data.

4.3 Literary language vs. social media

In this section we will compare forms found in VK-MAIN and VK-SMC to compare variation as found in literary Meadow Mari with variation found in Mari as it is used on social media platforms. We restricted ourselves to analysing word forms not considered ambiguous by this infrastructure, which without a doubt excluded a lot of otherwise useful data, esp. considering the false analyses oftentimes included in these infrastructures as illustrated in Section 2. Here again we have not distinguished between the plural suffixes *-βlak* and *-šamâč* due to the relative paucity of data.

Case suffix and possessive suffix

	Literary		Social Media	
	PXCX	**CXPX**	**PXCX**	**CXPX**
1SG	843	0	1,014	3
2SG	142	2	874	6
3SG	4,942	0	1,970	0
1PL	42	159	532	182
2PL	8	82	90	426
3PL	1,045	2	496	1

Table 12: PXCX ~ CXPX in the dative, literary and social media texts

Number suffix and possessive suffix

	Literary		Social Media	
	PXNX	**NXPX**	**PXNX**	**NXPX**
1SG	508	39	269	40
2SG	59	25	37	49
3SG	6,328	144	1,066	129
1PL	996	12	335	87
2PL	86	1	53	30
3PL	147	19	58	24

Table 13: PXNX ~ NXPX, literary and social media texts

Number suffix, Possessive suffix, case suffix

		Literary		Social Media	
		PXNXCX	**NXPXCX**	**PXNXCX**	**NXPXCX**
GENITIVE	**1SG**	53	2	4	0
	2SG	3	3	1	0
	3SG	553	4	69	3
	1PL	240	4	55	4
	2PL	12	1	2	0
	3PL	7	9	3	0
ACCUSATIVE	**1SG**	149	4	16	3
	2SG	18	8	10	18
	3SG	1,213	71	167	45
	1PL	132	9	63	20
	2PL	38	0	42	20
	3PL	25	10	19	5
DATIVE	**1SG**	91	3	27	2
	2SG	12	1	7	4
	3SG	609	2	56	2
	1PL	83	0	40	0
	2PL	22	0	14	0
	3PL	20	1	4	3

Table 14: All arrangements of PX, NX, and CX (excerpt), literary and social media texts

For the most part, the same arrangements are predominant in both genres, but the dominance is weaker in social media texts – that is to say, there is more variation in these. In the case of the dative, PXCX has a 97% dominance in VK-MAIN but only 89% in VK-SMC. 2SG is especially notable here: here the arrangement CXPX becomes dominant. The non-possessive usage of this suffix might serve as an explanation here, considering that its usage seems to be considered dialectal and colloquial. In sharp contrast to this, in the case of 1PL, the arrangement PXCX is dominant in social media (75%), while in literary texts CXPX dominates (79%). This might be interpreted as a case of paradigmatic levelling (with the deviation of 1PL and 2PL from other persons previously noted seemingly disappearing in colloquial speech), but curiously the same phenomenon cannot be observed in 2PL. Further investigation is necessary here.

When looking at combinations of three suffixes, an increase in variation can be observed as well. For the accusative, PXNXCX dominates with 94% in VK-MAIN, but only 74% in VK-SMC.

5. Outlook and conclusions

The results yielded by our corpus-based survey line up with the statistics assembled by Jorma Luutonen over 20 years and published in his 1997 dissertation.

As regards our desire for functional explanations of this variation, the usage of possessive suffixes 2SG and 3SG not only as possessive markers but also as determining elements of sorts serves as a plausible explanation for some of the variation encountered. It does not, however, serve as an explanation for variation encountered in relation to other suffixes, nor can it explain the three-way distinction PXNXCX ~ NXCXPX ~ NXPXCX.

We cannot honestly claim to have made serious headway into explaining this variation, but the binary question raised at the outset of this study – if novel electronic tools such as the corpora under investigation here can serve as tools of analysis in this domain – can be answered with a resounding "yes". Along the way, however, we encountered shortcomings in the infrastructures we were using that are quite systematic,

and thus can be assumed to be quite straight-forward to fix: if forms are analysed incorrectly as a rule, one need only change that rule to improve the output.

It would be desirable to revisit the data contained in the corpora with some concrete hypotheses as regards the alternation encountered that can be verified or falsified based on the data. While we do not at this point have these, there are starting points for future investigations we are considering.

Does the sound structure of the base word and suffix matter?

Could phonological or prosodic elements be influencing the choice of one version or another? That the possessive suffixes 1PL (*-na*) and 2PL (*-da*) deviate from all other possessive suffixes seems notable in this respect as these two are phonologically very similar to one another, but different from others: they have an onset consonant followed by the vowel *a*, and always form a distinct generally stressed syllable of their own when attached to a stem. When investigating the variation that can be found, it might be worth investigating if systematic differences can be observed based on the sound structure of the stems taking suffixes.

Does the function of the dative matter?

The dative case has a wide range of meanings in Mari (cf. Alhoniemi 1985: 52–54): it can mark the indirect object of an action ('I gave a book to my friend'), a benefactor ('I baked a cake for my squirrel[5]'), or a purpose ('I went to the well for water'). It can be directly governed by the argument structure of a syntactically superordinate verb (e.g., 'to help', 'to call (by phone)'). Does the function of the dative have an influence on the choice of suffix arrangement? This seems especially worthy of investigation given the unusually large variation encountered in the dative case.

Does the sentence structure matter?

Syntactic rules in Mari can be observed to lose their rigidity over distance within an example – i.e., they are subject to saliency effects. The following example sentence, taken from a Mari textbook, seems to violate a Mari grammatical rule according to which quantifiers such as *βiċ́* 'five', *šuko* 'many', or *ikmə̑ńar* 'a few' co-occur with singular rather than plural forms (cf. Riese et al. 2019: 56).

(2) Meadow Mari (Riese et al. 2017: 168)

ikmə̑ńar	joltaš-em,	poškud-em	da	rodo-βlak-em
a_few	friend-PXSG1	neighbour-PXSG1	and	relative-PL-PXSG1

'a few of my friends, neighbours and relatives'

Consultations with native speakers confirmed that this is not a typo; the plural marking of *rodo* 'relative' is admissible here in spite of its co-occurrence with *ikmə̑ńar* 'a few'. However, our native informants rejected **ikmə̑ńar rodo-βlak* as completely ungrammatical: only the distance between the quantifier and the quantified, i.e., the quantifier no longer being salient when the quantified is verbalized, makes the usage of a plural suffix admissible. Likewise, the salience of the possessor might influence the manner in which possessive suffixes are verbalized. Future investigations could investigate suffix ordering in relation with the distance between possessor and possessum in a clause.

[5] One reviewer objected to this example sentence as a squirrel seemed like a semantically unlikely benefactor in this example. However, both authors of this paper have at some point shared their lives with a squirrel. As part of the revision process, one of the authors, to ensure the naturalness of this example sentence, baked a cake for their squirrel.

Glossing abbreviations

1	1st person	DAT	dative
2	2nd person	GEN	genitive
3	3rd person	INE	inessive
ACC	accusative	LOC	local
ADD	additive	NX	number suffix
ASS	associative	PL	plural
CMPR	comparative case	PX	possessive suffix
COMP	comparative degree	SG	singular
CX	case suffix		

References

Alhoniemi, Alho. 1985. *Marin kielioppi*. Apuneuvoja suomalais-ugrilaisten kielten opintoja varten 10. Helsinki: Suomalais-Ugrilainen Seura. ISBN: 951901988X

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. *Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*, 125–140. Tartu. https://doi.org/10.18653/v1/W19-0311

Caballero, Gabriela. 2010. Scope, Phonology and Morphology in an Agglutinating Language: Choguita Raràmuri (Tarahumara) Variable Suffix Ordering. *Morphology* 20(1): 165–204. https://doi.org/10.1007/s11525-010-9147-4

Guseva, Elina and Philipp Weisser. 2018. Postsyntactic reordering in the Mari nominal domain: Evidence from Suspended Affixation. *Natural Language & Linguistic Theory* 36(4): 1089–1127. https://doi.org/10.1007/s11049-018-9403-6

Honti, László. 1995. Zur Morphotaktik und Morphosyntax der uralischen/finnisch-ugrischen Grundsprache. In *Congressus Octavus lnternationalis Fenno-Ugristarum 10.-15.8.1995. Pars I. Orationes plenariae et conspectus quinquennales*, edited by Heikki Leskinen, pp. 53–82. Gummerus, Jyväskylä. ISBN: 9529066848

Luutonen, Jorma. 1997. *The variation of morpheme order in Mari declension.* Mémoires de La Société Finno-Ougrienne 226. Helsinki: Suomalais-ugrilainen seura. ISBN: 9525150011

Rice, Keren. 2000. *Morpheme order and semantic scope: word formation in the Athapaskan verb* (Cambridge studies in linguistics 90). Cambridge University Press, Cambridge/New York. ISBN: 9780521024501 https://doi.org/10.1017/CBO9780511663659

Riese, Timothy, Jeremy Bradley, Monika Schötschel, and Tatiana Yefremova. 2019. *Mari (марий йылме): An Essential Grammar for International Learners. [Draft version]*. University of Vienna. grammar.mari-language.com

Riese, Timothy, Jeremy Bradley, Emma Yakimova & Galina Krylova. 2017. *Оҥай марий йылме: A Comprehensive Introduction to the Mari Language.* 3.2. Vienna: Department of Finno-Ugrian Studies, University of Vienna. omj.mari-language.com

Simonenko, Alexandra. 2014. Microvariation in Finno-Ugric possessive markers. In *Proceedings of the 43rd annual meeting of the North East Linguistic Society*, edited by Hsin-Lun Huang, Ethan Poole, and Amanda Rysling, pp. 127–140. Graduate Linguistics Student Association, Amherst, MA. ISBN: 149751066X

Tauli, Valter. 1953. The sequence of the possessive suffix and the case suffix in the Uralian languages. *Orbis* II:2. 397–404.

Trosterud, Trond and Valerij Alikov. 1994. Марийское двуязычие у мари. In *Volgalaiskielet muutoksessa: volgalaiskielten symposiumi, Turussa 1.–2.9.1993*, edited by Arto Moisio and Jaana Magnusson, pp. 111–117. Turun Yliopiston Suomalaisen ja Yleisen Kielitieteen Laitoksen julkaisuja 45. University of Turku. ISBN: 9512901765