Corpus of Literary Mari – Rudimentary Instructions

Jeremy Bradley (University of Vienna), jeremy.moss.bradley@univie.ac.at

Last updated 26 May 2022

---

1. Introduction

1.1. About this guide

This document is not a comprehensive manual to all functions of the corpus! Rather it is an introduction of its most basic functionalities so that you, as a complete beginner, can learn to use it in a rudimentary fashion, and then build upon this foundational expertise via trial and error. It was created in combination with a video tutorial found at corpus.mari-language.com. Everything laid out in this tutorial is illustrated with practical examples in the tutorial; Section 3 is an overview of the linguistic questions examined in the video.

1.2. About the corpus

This corpus was created by scholars from Ghent (Alexandra Simonenko), Helsinki (Jack Rueter, Niko Partanen), Moscow (Anna Volkova), Munich/Vienna (Jeremy Bradley), Tromsø (Trond Trosterud), Turku (Jorma Luutonen), and Yoshkar-Ola (Andrey Chemyshev, Gennadiy Sabantsev, Nadezhda Timofeeva). As of now it contains 57.38 million tokens of Meadow Mari texts from different genres (fiction, non-fiction, law, news, science) and covers over a century of Mari literacy.

This is an annotated corpus: one can search the texts in the corpus by grammatical information, etc. This annotation is at this point fairly rudimentary, and the disambiguation is currently not fully reliable. Caution and common sense is thus advised when using the corpus: please account for both false negative results (i.e., situations in which you do not find the pattern you are looking for due to incorrect interpretation in the corpus) and false positive results (i.e., the corpus returns erroneous results in addition to the desired ones) by making one's searches as unambiguous as possible, by checking on alternative interpretations of relevant forms, by manually processing the output of the search, etc. – see Section **Error! R eference source not found.** for some examples.

1.3. Useful links

GiellaLT documentation:	giellalt.github.io/lang/common/Korp_usage.html
Mari-English dictionary:	dict.mari-language.com
Mari-Finnish dictionary et al.:	muter.oahpa.no/
Mari-Russian dictionary et al.:	dict.fu-lab.ru/
Keyboard layout, orthography:	www.copius.eu/ortho, tech.mari-language.com
Regular expression sandbox:	regex101.com/

1.4. Searches and annotations

When you launch the corpus, you will first have to specify your type of search. If you use the **Simple** search, you can search for text fragments in the whole text fragments, e.g., Ик тымык ото шога мемнан элыште, and find all occurrences in the entire corpus. In the results returned by the search mask, you can already see what annotation and metadata the corpus has. For example, if you click on элыште in the first sentence, the output will look like this:

Corpus	
Fiction texts	
text attributes	
domain:	fiction
title:	Иксанова шыже блюз
date:	2010-01-01
word attributes	
part-of-speech:	noun
grammatical analysis:	N.Sg.Ine
dependency relation:	X
baseform:	эл

If you wish to search by these data, want to search for patters, or want to build flexibility into your search, you must use the **Extended** search. You can choose what fields you wish to search in the first dropdown menu, and what/how you want to search this field in the second one.

In the first drop-down menu, important points are:

- **word**: look for the word form as realized in the text. You can make your search case insensitive by clicking on **Aa** to the right of the text box. In a case insensitive search, the query элыште would not only find элыште, but also Элыште, ЭЛЫШТЕ, etc.
- **part-of-speech**: here you can specify that you wish to look for nouns, verbs, adverbs, etc. through a dropdown menu.
- **grammatical analysis**: here you can search for tags indicating the grammatical form of a word, e.g., in this example Sg for singular and Ine for inessive. An overview of suffixes used is given below in Section 2.
- **Dependency relation**: the syntactic role of the word in the clause (e.g., subject, object) – here, this parameter is not defined.
- **baseform**: the dictionary form; searching for эл 'country' in this field will find all inflected forms of this stem.
- **date:** the date when a resource was created.

In the second drop-down menu, you can choose if you want to match the whole word, just the beginning or end of the word, part of a word, or use regular expressions (**regexp**): a description of a search pattern that allows you to encode alternation, etc. Regular expressions can be incredibly powerful; you can find numerous tutorials online teaching you to use these. They can allow you to find all forms of a stem or suffix even when the annotation in the corpus is not fully reliable.

1.4.1. Multiple restrictions on search

By clicking on **+** in the bottom left corner of the search box, one can search for tokens satisfying multiple conditions (i.e., condition 1 AND condition 2). For example, if one searches for:

baseform	is	ава
and		
grammatical analysis	contains	Px

… will return all forms of ава ‘mother’ marked with a possessive suffix.

By clicking on **or**, one can search for tokens satisfying one of multiple conditions. One can combine AND conjunctions and OR conjunctions, as in this example:

baseform	is	айдеме
and		
grammatical analysis	contains	Ine
or		
grammatical analysis	contains	Ill
or		
grammatical analysis	contains	Lat

… will return all forms of айдеме ‘person’ marked with local case ending (inessive, illative, or lative). Note the scope: or-relations are always grouped at a lower level, i.e., in the example above, the four conditions are interpreted as follows: (CONDITION 1 **and** (CONDITION 2 **or** CONDITION 3 **or** CONDITION 4)).

1.4.2. Searching for patterns

By clicking on the **+** symbol to the right of the search box, one can add another search box. This allows users to search for patterns. Individual search boxes can be dragged and dropped to alter the ordering or deleted by clicking the **x** in the top right corner. If one leaves a search box empty in an added field, the search will accept any word in this position.

For example, the following search will return all occurrences of a genitive-marked word, followed by any word, followed by a possessive suffix (e.g., илышын тӱрлӧ этапыштыже life.GEN different stage.INE.PX3SG ‘in the different stages of life):

grammatical analysis contains Gen	word is <Any word>	grammatical analysis contains Px

By clicking on the gear icon (⚙) you can specify how many repetitions of a search query should be admissible in a spot. If you set 0 as the lower limit, this means that the search query is optional. For example:

grammatical analysis contains Gen	word is <Any word> Repeat 0 to 3 times	grammatical analysis contains Px

This query will find both поэмын геройжо poem.GEN hero.PX3SG ‘the hero of the poem’ (i.e., an example with nothing between the genitive form and the possessive-marked noun – 0 repetitions of <Any word>) as well as республикын образований да наука министерствыже

republic.GEN education and science ministry.PX3SG ‘the republic’s ministry of education and science’ (3 repetitions of <Any word>).

1.5. Statistics

For each set of search results, there is a **Statistics** tab showing you, in both relative frequencies and absolute frequencies, all patterns found in the corpus that match your search query, sorted by their freqency. Frequencies of individual forms in texts beloning to different genres are displayed here as well. For example, if you search for all words with the base form пӧрт ‘house’, (an excerpt of) the results will be:

total_rows 286

word	Total	*Non-fiction texts*	Fiction texts	Law texts	News texts	Science texts	Wikipedia texts
Σ	**786.6** (45,131)	**728.5** (145)	**951.3** (6,555)	**0** (0)	**769.8** (37,981)	**740.7** (214)	**359.4** (236)
пӧрт	**260.2** (14,927)	**226.1** (45)	**304.6** (2,099)	**0** (0)	**255.7** (12,615)	**280.4** (81)	**132.5** (87)
пӧртыш	**81** (4,650)	**115.6** (23)	**107** (737)	**0** (0)	**77.9** (3,845)	**100.4** (29)	**24.4** (16)
пӧртыштӧ	**71.1** (4,080)	**65.3** (13)	**73** (503)	**0** (0)	**71.5** (3,530)	**45** (13)	**32** (21)
пӧртшӧ	**24.7** (1,420)	**55.3** (11)	**35** (241)	**0** (0)	**23.3** (1,151)	**17.3** (5)	**18.3** (12)
Пӧрт	**48.5** (2,783)	**50.2** (10)	**75.6** (521)	**0** (0)	**45.3** (2,233)	**55.4** (16)	**4.6** (3)

The absolute frequencies, given in parentheses, show how often a particular form or string of words is find in the corpus, with Σ in the first line showing the total count of all forms. Note how lowercase пӧрт and uppercase Пӧрт are treated as separate forms in this output. These statistics can be exported (e.g., as a .csv file that can be edited in a spreadsheet program such as Microsoft Excel) for further processing.

2. Paradigms and tags

This section illustrates how morphology handled by the corpus is annotated by it – you can use this as a reference as to which tags to use in your search query. Alternatively, you can search for one occurrence of the form you are interested in, examine how it is annotated in the corpus, and build a more abstract search pattern on its basis.

2.1. Nouns

Illustrated using the nouns пӧрт ‘house’; where an animate noun is needed, ава ‘mother’ is used instead.

2.1.1. Case suffixes

Nominative	пӧрт	N.Sg.Nom
Genitive	пӧртын	N.Sg.Gen
Dative	пӧртлан	N.Sg.Dat
Accusative	пӧртым	N.Sg.Acc
Comparative	пӧртла	N.Sg.Cmpr
Comitative	пӧртге	N.Sg.Com
Inessive	пӧртыштӧ	N.Sg.Ine
Illative	пӧртыш, пӧртышкӧ	N.Sg.Ill
Lative	пӧртеш	N.Sg.Lat

2.1.2. Possessive suffixes

1Sg	пӧртем	N.Sg.Nom.PxSg1
2Sg	пӧртет	N.Sg.Nom.PxSg2
3Sg	пӧртшӧ	N.Sg.Nom.PxSg3
1Pl	пӧртна	N.Sg.Nom.PxPl1
2Pl	пӧртда	N.Sg.Nom.PxPl2
3Pl	пӧртышт	N.Sg.Nom.PxPl3

2.1.3. Possessive suffixes

Plural (1)	пӧрт-влак	N.Pl.Nom
Plural (2)	пӧрт-шамыч	N.Pl.Nom
Short (local) plural	пӧртла (гыч, …)	N.LocPl.Nom
Sociative plural	авамыт	N.AssocPl.Nom

2.1.4. Combination of suffixes

Case + Possessive	пӧрт-влакет	N.Sg.Ine.PxSg1.So_CP
Number + Case	пӧрт-влакым	N.Pl.Acc
Number + Poss.	пӧрт-влакет	N.Pl.Nom.PxSg2.So_NP
All three	пӧрт-влакшым	N.Pl.Acc.PxSg3.So_NPC

Note how the analysis here includes a tag showing the ordering of number suffixes (N), case suffixes (C), and possessive suffixes (P) – as there is a lot of alternation in this domain, these tags allow users to find only word forms using one arrangement:

Case + Possessive	пӧртланем	N.Sg.Dat.PxSg1.So_CP
Possessive + case	пӧртемлан	N.Sg.Dat.PxSg1.So_PC

2.2. Pronouns

- Personal, e.g., мый ‘I’ > Pron.Pers.Sg1.Nom
- Demonstrative, e.g., тиде ‘this’ > Pron.Dem.Sg.Nom
- Interrogative, e.g., кӧ ‘who’ > Pron.Interr.Sg.Nom
- Indefinite, e.g., иктаж-кӧ ‘somebody’ > Pron.Indef.Nom
- Reflexive, e.g., шкемым ‘myself’ > Pron.Refl.Acc.PxSg1
- Reciprocal, e.g., икте-весыштлан ‘(they) each other’ > Pron.Pron.Recipr.Dat.PxPl3

2.3. Adjectives

сай ‘good’ >

Positive	сай	A.Attr
Comparative	сайрак	A.Comp

2.4. Adverbs

ончыч ‘before’>

Positive	ончыч	Adv
Comparative	ончычрак	Adv.Comp

2.5. Postpositions

шенгелне ‘behind’>

Base form	шенгелне	Adp.Po
With possessive suffix 1Sg	шенгелнем	Adp.Po.PxSg1
…	…	…

2.6. Verbs

толаш (verbal stem тол-) ‘to come’ >

2.6.1. Indicative

		Positive		Negative	
Present	**1Sg**	толам	V.Ind.Prs.Sg1	ом тол	V.Neg.Ind.Prs.Sg1 V.ConNeg
	2Sg	толат	V.Ind.Prs.Sg2	от тол	V.Neg.Ind.Prs.Sg2 V.ConNeg
	3Sg	толеш	V.Ind.Prs.Sg3	огеш ~ ок тол	V.Neg.Ind.Prs.Sg3 V.ConNeg
	1Pl	толына	V.Ind.Prs.Pl1	огына ~ она тол	V.Neg.Ind.Prs.Pl1 V.ConNeg
	2Pl	толыда	V.Ind.Prs.Pl2	огыда ~ ода тол	V.Neg.Ind.Prs.Pl2 V.ConNeg
	3Pl	толыт	V.Ind.Prs.Pl3	огыт тол	V.Neg.Ind.Prs.Pl3 V.ConNeg
Past I	**1Sg**	тольым	V.Ind.Prt1.Sg1	шым тол	V.Neg.Ind.Prt1.Sg1 V.ConNeg
	2Sg	тольыч	V.Ind.Prt1.Sg2	шыч тол	V.Neg.Ind.Prt1.Sg2 V.ConNeg
	3Sg	тольо	V.Ind.Prt1.Sg3	ыш тол	V.Neg.Ind.Prt1.Sg3 V.ConNeg
	1Pl	толна	V.Ind.Prt1.Pl1	ышна тол	V.Neg.Ind.Prt1.Pl1 V.ConNeg
	2Pl	толда	V.Ind.Prt1.Pl2	ышда тол	V.Neg.Ind.Prt1.Pl2 V.ConNeg
	3Pl	тольыч	V.Ind.Prt1.Pl3	ышт тол	V.Neg.Ind.Prt1.Pl3 V.ConNeg
Past II	**1Sg**	толынам	V.Ind.Prt2.Sg1	толын омыл	V.Ger.Gen V.Neg.Ind.Prs.Sg1
	2Sg	толынат	V.Ind.Prt2.Sg2	толын отыл	V.Ger.Gen V.Neg.Ind.Prs.Sg2
	3Sg	толын	V.Ind.Prt2.Sg3	толын огыл	V.Ger.Gen V.Neg.Ind.Prs.Sg3
	1Pl	толынна	V.Ind.Prt2.Pl1	толын огынал ~ онал	V.Ger.Gen V.Neg.Ind.Prs.Pl1
	2Pl	толында	V.Ind.Prt2.Pl2	толын огыдал ~ одал	V.Ger.Gen V.Neg.Ind.Prs.Pl2
	3Pl	толыныт	V.Ind.Prt2.Pl3	толын огытыл	V.Ger.Gen V.Neg.Ind.Prs.Pl3

Negative forms of улаш ‘to be’ (present tense):

1Sg	омыл	V.Neg.Ind.Prs.Sg1
2Sg	отыл	V.Neg.Ind.Prs.Sg2
3Sg	огыл	V.Neg.Ind.Prs.Sg3
1Pl	огынал ~ онал	V.Neg.Ind.Prs.Pl1
2Pl	огыдал ~ одал	V.Neg.Ind.Prs.Pl2
3Pl	огытыл	V.Neg.Ind.Prs.Pl3

2.6.2. Imperative

	Positive		**Negative**	
1Sg	-	-	-	-
2Sg	тол	V.Imprt.Sg2	ит тол	V.Neg.Imprt.Sg2 V.ConNeg
3Sg	т**о**лжо	V.Imprt.Sg3	ынже тол	V.Neg.Imprt.Sg3 V.ConNeg
1Pl	-	-	-	-
2Pl	т**о**лза	V.Imprt.Pl2	ида тол	V.Neg.Imprt.Pl2 V.ConNeg
3Pl	т**о**лышт	V.Imprt.Pl3	ынышт тол	V.Neg.Imprt.Pl3 V.ConNeg

2.6.3. Desiderative

	Positive		**Negative**	
1Sg	толнем	V.Des.Prs.Sg1	ынем тол	V.Neg.Des.Sg1 V.ConNeg
2Sg	толнет	V.Des.Prs.Sg2	ынет тол	V.Neg.Des.Sg2 V.ConNeg
3Sg	толнеже	V.Des.Prs.Sg3	ынеж тол	V.Neg.Des.Sg3 V.ConNeg
1Pl	толнена	V.Des.Prs.Sg1	ынена тол	V.Neg.Des.Pl1 V.ConNeg
2Pl	толнеда	V.Des.Prs.Sg2	ынеда тол	V.Neg.Des.Pl2 V.ConNeg
3Pl	толнешт	V.Des.Prs.Sg3	ынешт тол	V.Neg.Des.Pl3 V.ConNeg

2.6.4. Non-finite forms

Infinitive	толаш	V.Inf
Necessitive infinitive	толман	V.Inf.Nec
Active participle	толшо	V.Act.Prc
Passive participle	толмо	V.Pass.Prc
Future-necessitive participle	толшаш	V.Fut.Prc
Negative participle	толдымо	V.Neg.Prc
Affirmative instructive gerund	толын	V.Ger.Gen
Negative gerund	толде	V.Ger.Abe
Gerund of prior action	толмек ~ толмеке	V.Ger.Prf
Gerund of future action	толмеш ~ толмешке	V.Ger.Imprf
Gerund of simultaneous action	толшыла	V.Ger

2.7. Clitics

нуно ‘they’ > нунат ‘they too’	Pron.Pers.Pl3.Nom > Pron.Pers.Pl3.Nom.Foc_at
мый ‘I’ > мыяк ‘(well) I’	Pron.Pers.Sg1.Nom > Pron.Pers.Sg1.Nom.Foc_ak
Палем. ‘I know.’ > Палемыс. ‘Oh, I know.’	V.Ind.Prs.Sg1 > V.Ind.Prs.Sg1.Foc_ys
ӱлыкӧ ‘down’ > ӱлыкыла ‘downwards’	Adv > Adv.Weak
Ончо! ‘look’ > Ончо-я! ‘Oh look!’	V.Imprt.Sg2 > V.Imprt.Sg2.Foc_ja
Ончо! ‘look’ > Ончо-ян! ‘Oh look!’	V.Imprt.Sg2 > V.Imprt.Sg2.Foc_jan

2.8. Derivational morphology

-ан	йӱр ‘rain’ > йӱран ‘rainy’	N.Sg.Nom > N.Der_Poss…
-(ы)сЕ	жап ‘inside’ > жапысе ‘of a time’	N.Sg.Nom > N.Der_Rel.Attr…
-дымЕ	шӱм ‘heart’ > шӱмдымӧ ‘heartless’	N.Sg.Nom > N.Der_Priv.Attr…
-лык	теле ‘winter’ > телылык ‘for winter’	N.Sg.Nom > N.Der_Pur.Attr…
-алт^{I}	ышташ (-ем) ‘to do’ > ышталташ (-ам) ‘to be done’	V.Inf > V.Der_Refl.Inf
-(ы)кт^{II}	шочаш (-ам) ‘to be born’ > шочыкташ (-ем) ‘to give birth’	V.Inf > V.Der_Caus.Inf
аш	йӧраташ (-ем) ‘to love’ > йӧратымаш ‘love’	V.Inf > V.Der.Der_Nom
-дымаш	палаш (-ем) ‘to know’ > палыдымаш ‘ignorance’	V.Inf > V.Der.Der_NomNeg.N.Attr

3. Example searches in video tutorial

3.1. Simple searches

How does one search the corpus for individual words or phrases, e.g., a name, or a line from a poem?

3.2. Extended searches

How does one search the corpus for all tokens starting with a certain string of characters, for all inflected forms of a stem, etc?

3.3. Animate nouns in local cases

Prescriptive Mari grammars claim that animate nouns such as айдеме ‘person’ do not occur in the local cases (inessive, illative, lative). Here the video tutorial shows how counterexamples to this rule can be found in the corpus – how one can search for this noun in one of these three cases.

3.4. Government of a verb

Students of Mari might not know if the verb йыҥгырташ ‘to call’ (but also, and orginally, ‘to ring’, e.g., a bell) co-occurs with the accusative case (as in German), or the dative case (as in Russian). Here we show how the corpus can be used to check if the person being called is marked with the dative or the accusative – but it is also shown how even in the case of such a straight-forward question, the output from the corpus must be processed by hand before a meaningful inferences can be made. The person being called is marked with the dative, as the

output shows, but accusative forms do co-occur with this verb: temporal adverbials (e.g., йӱдым ‘at night’ < йӱд ‘night’), objects when the verb is being used in the meaning ‘to ring (a bell)’, etc.

3.5. Adjective – noun collocations; statistics

The Mari adjectives сылне and мотор are both extensively used and generally translated as ‘beautiful’. Here we illustrate how the corpus produces statistics on collocations: which word forms these adjectives most commonly co-occur with, and how their functional ranges can thus be disambiguated.

3.6. Allowing for distance between searched words

Here we show how search queries can be set up to allow for “gaps” between two searched elements, on the basis of adnominal possessive constructions of the type Анушын изаже Anush.GEN elder_brother.PX3SG ‘Anush’s elder brother’. It is shown how to allow for words (exactly 1 word, 1–3 words, 0–3 words) between the possessor and the possessum in this search pattern.

3.7. Adjusting for ambiguity

The complementizer манын ‘that’ (< literally ‘saying’) is used both in combination with statements (which can be in the indicative, possibly past-tense forms) and in final clauses in combination with the imperative third person. In the third person plural, the syncretism of imperative and simple past tense I forms in the second conjugation (and only the second conjugation!) can make a contrastive study of these structures difficult:

	толаш ‘to come’ (first conjugation	**ышташ ‘to do’ (second conjugation)**
Simple Past I, 3PL	тольыч	ыштышт
Imperative, 3PL	толышт	ыштышт

In the video demonstration we show how (1) one can, if one does not trust the disambiguation in these cases, set up a search pattern that practically undoes the ambiguation, allowing users to separate cases by hand, (2) one can set up a search pattern that only looks at select first-conjugation verbs not afflicted by this ambiguity.

3.8. Diachronic studies

Here we show how the corpus can be used to study diachronic change, on the basis of the complex conjunction молан манаш гын ‘because’ (literally ‘if you say why’), a relative novelty in the Mari language. We show how one can look for usages of the phrase in different decades of the 20th and 21st century.

3.9. Regular expressions

Here we show how regular expressions – search patterns with inbuilt flexibility – can be used to find words with a certain structure when they cannot be found based on the tagging alone. This is illustrated on the basis of the suffix -дыме ~ -дымо~ -дымӧ, which with its three vowel harmonic variants is attached both to verbal stems to attach negative participles (e.g., палаш 'to know' > палыдыме 'unknown') or to nominal stems to create privative adjectives ('-less', e.g., вӱр 'blood' > вӱрдымӧ 'bloodless'). It is shown how all verbs with this ending, including the vowel harmonic alternation, can be found using regular expressions.