BRADLEY, JEREMY

A Corpus-Based Analysis of Syntactic Structures: Postpositional Constructions in Mari

This paper aims to illustrate the potential of my work group's Meadow Mari morphological analyzer (BRADLEY 2011, 2015), when coupled with our recently-published 42,500+-headword-strong Mari–English dictionary (RIESE ET AL. 2014–), as a framework or model for corpus-based research on a 'small' Uralic language of Russia. For illustrative purposes, I have chosen the usage of verbal nouns traditionally called 'participles' in postpositional constructions, a typical feature of many languages of European Russia. While the existence of such constructions is plainly evident in existing grammars, the range of applications – which participles can be coupled with which postposition within which semantic constraints – is not detailed any further or delimited in any way. I will seek to show how the corpus infrastructure I have designed can be used to gain insights into this and similar matters; how such an infrastructure could be made more powerful in years to come; and that the study of other Uralic (and non-Uralic) languages would profit from the creation of infrastructures similar to those we have developed for Mari.

Keywords: Mari, postpositions, participles, corpus linguistics, morphological analysis.

1. Participle + Postposition

In the simplest terms, a *participle* can be defined as a „verbal adjective" (HASPELMATH 1995: 3), a „verb form… marking [a] relative clause…" (HASPELMATH – SIMS 2010: 86), or a "deverbal adjective that may retain some verbal properties" (id.: 337). While Uralic publications tend to use this term and its counterparts in other languages – German *Partizip*, Finnish *partisiippi*, Estonian *kesksõna* or *partitsiip*, Hungarian *melléknévi igenév*, and Russian *причастие* – in this manner, it is often better to regard it as a rather vague umbrella term used for polyfunctional forms. In numerous Uralic languages, forms labelled as participles in textbooks and grammars can be found used as adjectives, but are also used as verbal nouns with a considerably wider range of functions. Consider the following two usage examples from Mansi of the so-called perfect participle in *-m*:

Nyelvtudományi Közlemények 111: 255–275.
DOI: 10.15776/NyK.2015.111.13

Northern Mansi (BALANDIN 1960: 111)

(1)	*sun*	*wār-əm*	*χum*	*kol-n*	*śalt-əs*
	sled	make-PTCP.PERF	man	house-LAT	enter-PST.3SG

’The man that made a sled entered the house.’[1]

Northern Mansi (BALANDIN 1960: 115)

(2)	*taw*	*joχt-əm-e*	*urəl*	*potərt-as-uw*
	s/he	come-PTCP.PERF-PX3SG	about	speak-PST-1PL

’We talked about his/her arrival.’[2]

In the first example, the definitions given above hold: the participle is used in a relative clause; it serves as an adjectival attribute of the noun it follows, but preserves a verbal internal syntax by having an object. In the second example, the superordinate word is not a noun modified by the verbal noun, but rather an adposition combining with it. As constructions of this type are anything but rare in Mansi, it is not appropriate to label this sentence as a marginal non-prototypical usage of the participle: rather, the forms labelled as participles in Mansi are verbal nouns with a wide range of applications that includes prototypically participial functions, but is not restricted or defined by these. (For the sake of simplicity, I will continue referring to these forms as participles and will gloss them as participles – while acknowledging that the term is not entirely appropriate in many cases.)

The Mansi participle in *-m* is a reflex of Proto-Uralic **-m* (COLLINDER 1960: 266–269). Its cognates in numerous Uralic languages, while distinct regarding some of their semantic properties, exhibit similar behaviour with respect to their range of syntactic functions. The same duality seen in Mansi can also be found, for example, regarding its cognates in Khanty, Udmurt, Komi, and Mari. The following examples constitute pairs in which the first item shows the usage of the verbal noun in question as an adjectival participle and the second one, its usage in combination with a postposition.

Surgut Khanty (CSEPREGI 2011: 62)

(3)	*əj*	*Vatnə*	*må]i*	*åV*	*ăwt-əm*	*aŋkV-ət-a*	*jŏwət*
	one	during	past	year	cut-PTCP.PST	tree.stump-PL-LAT	come.PST.3SG

’Soon s/he got to some tree stumps cut in the previous year.’[3]

[1] ’Нарту делавший мужчина вошел в дом.’

[2] ’Мы говорили о его приезде.’

[3] ’Hamarosan az előző évben kivágott fatönkhöz érkezett.’

Surgut Khanty (from ZSÓFIA SCHÖN's unpublished corpus)

(4)	*wåńťənt-əm-am*	*ʌatnə*	*jom*
	pick.berries-PTCP.PST-1SG	during	rain.PST.3SG
	'It rained when I was picking berries.'		

Udmurt (CSÚCS 1998: 291)

(5)	*gi̮r-em*	*busi̮*
	plough-PTCP.PERF	field
	'ploughed field'	

Udmurt (Newspaper Idnakar, 27 November 2012, article *Makem duno aďamili̮ sakli̮k*)[4]

(6)	*tatč́i̮*	*li̮kt-em-ez*	*śari̮š*	*ogpol*	*no*	*e̮z*	*žaľa…*
	here	come-PTCP. PERF-PX3SG	about	one.time	also	NEG. PST.3SG	regret.CNG
	'S/he never regretted coming here once…'						

Komi-Zyrian (RÉDEI 1978: 113)

(7)	*pe̮žal-e̮m*	*ńań*
	bake-PTCP.PERF	bread
	'baked bread'[5]	

Komi-Zyrian (BEZNOSIKOVA ET AL. 2000: юӧртны)

(8)	*jue̮rt-ni̮*	*vo-e̮m*	*ji̮li̮ś*
	announce-INF	come-PTCP.PERF	about
	'to announce (one's) arrival'[6]		

Meadow Mari (BERECZKI 1990: 61)

(9)	*kural-me*	*mlande*
	plough-PTCP.PASS	land
	'ploughed land'[7]	

Meadow Mari (BERECZKI 1990: 62)

(10)	*βüt*	*kondə̑-m-em*	*godə̑m*	*už-ə̑m*
	water	bring-PTCP.PASS-PX1SG	when	see-PST1.1SG
	'I saw (it) when I was bringing water.'[8]			

[4] idnakar.info/udm/pumiskon-jos/880-makem-duno-adyamily-saklyk, accessed 2015-09-22.

[5] 'gebackenes Brot'

[6] 'известить о приезде'

[7] 'felszántott föld'

[8] 'mikor vizet hoztam, láttam'

This phenomenon is not restricted to the reflexes of Proto-Uralic *-*m*, nor is there anything exclusively Uralic about it. Verbal nouns/participles in Turkic and Mongolic languages, for example, exhibit a similar range of applications.

Tatar (LANDMANN 2014a: 75)

(11)	*bułəš-qan*	*dus-łar*
	help-PTCP.PERF	friend-PL

'friends that helped'[9]

Tatar (LANDMANN 2014a: 85)

(12)	*åβər-γan-əm*	*öṡ́ön*	*eš-kä*	*bår-a*	*åł-ma-də-m*
	fall.ill-PTCP.PERF-PX1SG	because	work-DAT	go-CVB	go-NEG-PST-1SG

'I could not go to work because I was ill.'[10]

Chuvash (LANDMANN 2014b: 72)

(13)	*pulăš-as*	*jultaš-sem*
	help-PTCP.FUT	friend-PL

'friends that will help'[11]

Chuvash (LANDMANN 2014b: 82)

(14)	*pajan*	*śumăr*	*śăv-as*	*pek*	*tuj-ăn-a-ť*
	today	rain	fall-PTCP.FUT	like	feel-PASS-PRS-3SG

'It feels as if it's going to rain today.'[12]

Buryat (SKRIBNIK 2003: 114)

(15)	*übhen-de*	*gara-χa*	*χün-üüd*
	hay-DAT	go-PTCP.FUT	person-PL

'people who [will] go to haymaking'

Buryat (SKRIBNIK 2003: 124)

(16)	*jaba-χa*	*deere-m*	*nüχer-ni*	*jer-ee*
	go-PTCP.FUT	above-PX1SG	friend-PX1SG	come-PST.3SG

'Just before I left, my friend came.'

This polyfunctionality can be considered to be a general property of „Eastern" verbal nouns/participles, and the usage of such forms in postpositional constructions can be seen as integral to „Eastern" syntax. Postpositions dominate over prepositions in all languages mentioned so far (cf. DRYER 2013b); with the

[9] 'Freunde, die geholfen haben'

[10] 'Ich konnte nicht zur Arbeit gehen, weil ich krank war'

[11] 'Freunde, die helfen werden'

[12] 'Es fühlt sich so an, als wollte es heute regnen'

exception of Komi (which is SVO), all of these languages have the basic word order SOV (cf. DRYER 2013a).

While the presence of this duality was easy to verify in the languages in question, detailed accounts of the mechanism, both cross-linguistic and with respect to individual languages, are hard to come by. Which participles are used in postpositional constructions in which languages, with what frequency, and within which semantic constraints? To my knowledge there has not been an areal typological study of these questions; old-fashioned descriptive grammars of Uralic and Turkic languages are not necessarily forthcoming regarding syntactic constructions. This paper will explore the prospect of delimiting the mechanism in one specific Uralic language – (Meadow) Mari – using a semi-automatic morphological analysis of a large body of texts, thus aiming to introduce corpus-linguistics methods into a field that has traditionally been based on text collections, but not the (semi-)automatic analysis of them using electronic tools. The focus here is firmly on introducing the methodology at hand, rather than on actually creating an extensive review of the feature itself – the data at my disposal at this point is not sufficient for this.

The following chapter will briefly introduce the participles treated in Mari grammars, will introduce some basic principles of their usage, and will mention what information can or cannot be found in the Mari grammars available to date. After that, I will briefly introduce the corpus infrastructure and show how it can be used to study this feature. Finally, I will touch upon prospects for improving this approach in the future.

2. Mari participles

Mari grammars generally list four different participles. All four of these can be used as adjectives and nouns alike, but two fundamental types of nominal usages should be distinguished:

– The usage of originally adjectival forms as nouns by means of conversion, e. g. the passive participle *paləme* 'acquaintance (< known)' < *pale-* 'to know' (RIESE ET AL. 2014–: палыме). As the dividing line between adjectives and nouns, in typical Uralic fashion, is comparatively vague in Mari (cf. ALHONIEMI 1985: 42), adjectives can in general be used as nouns. Hence, the occurrence of nominalized adjectival participles is comparatively trivial; nominal forms of participles of this type are noise that must be filtered out in the present study.
– The usage as fully fledged verbal nouns in manners that cannot be explained by conversion or a fuzzy dividing line between adjectives and nouns. For example, note the passive participle of an intransitive verb being used as a verbal noun that is the object of a transitive verb:

Meadow Mari (Raamatunkäännösinstituutti 2007: Luke 8:53)

(17)	*üdər*	*kolə-mə-m*	*pal-en-ət=at,*	…
	girl	die-PTCP.PASS-ACC	know-PST2-3PL=and	
	'… knowing that she [the girl] was dead.'			

The following subchapters will briefly introduce the individual participles found in Mari. An upper-case *E* denotes a vowel-harmonic vowel that can occur as either *e*, *o*, or *ö*.

2.1. The active participle in *-šE*

The active participle (cf. ALHONIEMI 1985: 136, BERECZKI 1990: 59–60) is neutral with regard to tense, i.e. its correct reading depends on the semantics (e. g. the telicity) of the verbal stem, the semantics of the modified noun, and the wider usage context.

Meadow Mari (RIESE ET AL. 2014–: илыше)

(18)	*ilə-še*	*kajək*
	live-PTCP.ACT	bird
	'live bird'	

Meadow Mari (RIESE ET al. 2014–: колышо)

(19)	*kolə-šo*	*karme*
	die-PTCP.ACT	fly
	'dead fly'	

The active participle is, by means of conversion, frequently used to form agent nouns, e. g. *tunəktəšo* 'teacher (< teaching)' < *tunəkto-* 'to teach', *tunemše* 'student (< studying)' < *tunem-* 'to study' (RIESE ET AL. 2014–: туныктышо, тунемше). Beyond this, it does not seem to be used nominally. I will revisit this statement in due course.

2.2. The passive participle in *-mE*

The passive participle (cf. ALHONIEMI 1985: 136–138, BERECZKI 1990: 60–61) is also neutral with regard to tense; the principles that apply to its interpretation are the same as those applicable to the active participle.

Meadow Mari (RIESE ET AL. 2014–: палыме)

(20)	*palə-me*	*ajdeme*
	know-PTCP.PASS	person
	'known person'	

Meadow Mari (RIESE ET AL. 2014–: колтымо)

(21)	*koltə-mo*	*serəš*
	send-PTCP.PASS	letter
	’sent letter’	

The passive participle can also be nominalized through conversion, e. g. *paləme* ’acquaintance (< known)’ < *pale-* ’to know’ (RIESE ET AL. 2014–: палыме).

In many situations the participle occurs as an attribute where a passive reading is not appropriate. Instead, it serves as a generic deverbal adjective form that is neutral with respect to voice. Participles are also formed from intransitive verbs in this function.

Meadow Mari (RIESE ET AL. 2014–: илыме)

(22)	*ilə-me*	*pört*
	live-PTCP.PASS	house
	’residential house’	

The participle is commonly used as a verbal noun, potentially marked for case and/or possessor, in a wide range of clauses. In this function as well, one should not read passiveness into the form; it is quite neutral with regard to voice. As the active participle is not used in this function, one could say that *-mE* as a verbal noun covers the range of both *-mE* and *-šE* as adjectival participles. The following examples show this form used in a subject and in an object clause:

Meadow Mari (RIESE ET AL. 2014–: кумдан)

(23)	*joča-n*	*kəčkər-al-mə-že*	*kumda-n*	*šergəlt-e*
	child-GEN	cry-MOM-PTCP.PASS-PX.3SG	wide-INS	be.heard-3SG
	’The child’s crying could be heard all around.’			

Meadow Mari (RIESE ET AL. 2014–: тунам)

(24)	*təj*	*məlanem*	*kert-m-et-əm*	*onč-əkto,*	…
	2SG	1PL.DAT	be.able-PTCP.PASS-PX.2SG-ACC	see-CAUS.IMP.2SG	
	’Show me what you can do…’				

As illustrated above, the participle can also occur in combination with postpositions. If the postposition governs a case other than the nominative, the participle takes the appropriate case suffix. For example, the postposition *köra* ’because of’ governs the dative case, and it is also used with the dative forms of the passive participle.

Meadow Mari (RIESE ET AL. 2014–: кӧра)

(25)	*tid-lan*	*köra*
	this-DAT	because.of
	’because of this’	

Meadow Mari (RIESE ET AL. 2014–: кӧра)

(26)	*čerlanə-mə-lan*	*köra*	*paša*	*deč*	*koraŋ-e*
	fall.ill-PTCP.PASS-DAT	because.of	work	from	leave-PST1.3SG
	’S/he left his/her job due to his/her illness.’				

The grammars cited above discuss and illustrate all of the functions shown here.

2.3. The future-necessitive participle in *-šaš*

The future-necessitive participle (cf. ALHONIEMI 1985: 138–140, BERECZKI 1990: 60) refers to things or persons that will, or will have to/be expected to, carry out an activity or that will be subject to an activity. This is to say, it is neutral with respect to voice, but has a temporal/modal value: it refers to actions that are subsequent to the point of reference, or that are expected at or after the point of reference.

Meadow Mari (ALHONIEMI 1985: 138)

(27)	*tol-šaš*	*una*
	come-PTCP.FUT	person
	’guest that will/should come’[13]	

Meadow Mari (ALHONIEMI 1985: 138)

(28)	*əštə-šaš*	*paša*
	do-PTCP.FUT	work
	’work to be done’[14]	

Forms nominalized through conversion can be found here too, but they are comparatively rare (as the participle itself is comparatively rare). The grammars cited here do not discuss these forms, but as adjectives can be nominalized by conversion in general, this is not strictly necessary.

Meadow Mari (Raamatunkäännösinstituutti 2007: Hebrews 1:14)

(29)	…	*utarə-maš-əm*	*nal-šaš-βlak-lan*	...
		rescue-NMLZ-ACC	take-PTCP.FUT-PL-DAT	
	’… for them who shall be heirs of salvation’			

[13] ’vieras, joka on tuleva, jonka on tultava’
[14] ’työ, joka on tehtävä’

Like the passive participle, this participle can also be used as a verbal noun and can also be coupled with postpositions in this function. The grammars do discuss this usage of the participle.

Meadow Mari (Raamatunkäännösinstituutti 2007: Matthew 24:43)

(30)	…	*Bor*	*tol-šaš-əm*	*pört*	*oza*	*pal-a*	*gən*	…
		thief	come-PTCP.FUT-ACC	house	master	know-3SG	if	

’… if the goodman of the house had known… the thief would come…’

Meadow Mari (Raamatunkäännösinstituutti 2007: Mark 13:11)

(31)	…	*mo-m*	*ojlə-šaš*	*nergen*	*ončə̆l goč*	*ida*	*azaplane.*
		what-ACC	say-PTCP.FUT	about	beforehand	NEG.IMP.2PL	be.worried.CNG

’… take no thought beforehand what ye shall speak…’

2.4. The negative participle *in -də̂mE*

The negative participle (cf. ALHONIEMI 1985: 140–141, BERECZKI 1990: 62) is the negated counterpart of all the affirmative participles given above. It is neutral with regard to voice and tense; its reading in these respects depends on the semantics of the lexemes and on context.

Meadow Mari (RIESE ET AL. 2014–: умылыдымо)

(32)	*umələ-dəmo*	*šomak*
	understand-PTCP.NEG	word

’incomprehensible word’

Meadow Mari (RIESE ET AL. 2014–: умылыдымо)

(33)	*umələ-dəmo*	*joča*
	understand-PTCP.NEG	child

’slow-witted child’

Nouns formed through conversion are possible here too, e. g. *palədəme* ’stranger (< unknown)’ < *pale-* ’to know’ (RIESE ET AL. 2014–: палыдыме).

This form can also be used as a more generic verbal noun and in different types of subordinate clauses, and can be connected with postpositions. The grammars cited here do not discuss these forms.

Meadow Mari (Raamatunkäännösinstituutti 2007: Luke 8:47)

(34)	*tunam*	*üdəramaš,*	*šoləp-eš*	*kod-ən*	*kert-dəmə-žə-m*	*už-ən…*
	then	woman	secret-LAT	stay-CVB	be.able-PTCP.NEG-PX3SG-ACC	see-CVB

’And when the woman saw that she was not hid…’

Meadow Mari (Raamatunkäännösinstituutti 2007: Matthew 17:20)

(35) *üšanə-dəmə-lan-da* *köra* ...
believe-PTCP.NEG-DAT-PX2PL because.of
'Because of your unbelief…'

3. A corpus-based analysis

The corpus infrastructure I will discuss in this chapter can be found at corpus.mari-language.com; a brief introduction to its capabilities, content, and user interface can be found in (BRADLEY 2015). When I was writing this paper, I had published an operational demo spanning 994,097 tokens under this address; I am planning to publish a more extensive resource at the same address in the future. While the body of texts available at that address might be greater by the time this article is read, and while the user interface might have been altered, I am confident that everything I describe in this paper will be reproducible in newer releases of the corpus infrastructure.

For the time being, the corpus infrastructure only covers the Meadow Mari literary norm. In due course, I intend to expand it to cover the second Mari literary norm, Hill Mari, as well. A prospective, but as yet unplanned inclusion of non-literary, dialectal texts will be discussed below.

The corpus infrastructure is in its essence a cross-integration of the following resources:

– A number of texts that are either in the public domain or for which I have been given permission to use in the project at hand: The texts from my work group's textbook „Oŋaj marij jzlme: A Comprehensive Introduction to the Mari language" (RIESE ET AL. 2012): 2,508 tokens; SERGEJ ČAVAJN's novel Elnet (ČAVAJN 1967): 63,918 tokens; a Mari translation of the New Testament (Raamatunkäännösinstituutti 2007): 127,717 tokens; all examples from the largest Mari–Russian dictionary to date (GALKIN ET AL. 1990–2005): 585,431 tokens; all example sentences from the Mari-English dictionary (RIESE ET AL. 2014–): 214,523 tokens. The total number of tokens is thus 994,097.
– A Mari morphological analyzer capable of handling productive Mari morphology, both inflectional and derivational, in its entirety. This tool can be found as a standalone application at morph.mari-language.com (> 'Analyzer'); its architecture and the analysis model at its core are discussed in detail in (BRADLEY 2011).
– A Mari–English dictionary covering 42,500+ headwords, found at dict.mari-language.com (RIESE ET AL. 2014–)
– A repository and user interface I designed for this application.

It should be noted that all of these resources use Cyrillic orthography. As I am otherwise using Finno-Ugric Transcription in this paper, the spelling of Mari in screen shots will not match the spelling used in example sentences elsewhere in the paper. I will give the individual examples taken from the corpus in Finno-Ugric Tanscription (the user interface does allow automatic transcription into Finno-Ugric Tanscription and IPA).

When texts are fed into the back end of the corpus infrastructure – which is not accessible to the general public – their individual sentences are run through the morphological analyzer. It, using the Mari–English dictionary as its lexical database, creates perfunctory interlinear glosses of individual sentences, one word at a time. The output created by the morphological analyzer for a single unambiguous word form looks like this (using the publicly accessible user interface at morph.mari-language.com > 'Analyzer'):

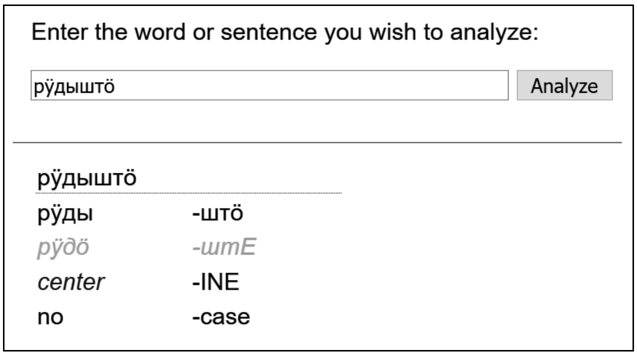

Figure 1
Morphological analysis of a single word
(screenshot from morph.mari-language.com)

The tiers of the interlinearization are as follows:

rüdə̑štö: the unglossed word, as it occurs in the input.

rüdə̑, *-štö*: the individual morphemes, as they are realized in the word in question (i.e. the morphs).

rüdö, *-štE*: the base form of the morphemes in question. For lexemes, this is the stem; the base forms of affixes are contained in the analysis model.

center, -INE: the gloss. For lexemes, these are taken from the lexical base. The first translation given in the lexicon is displayed; all translations are displayed as a tool tip if users hover the mouse cursor. The glossing abbreviations for suffixes are taken from the analysis model.

no, -case: the part of speech. Here again, information on lexemes is taken from the lexical base, whereas information on suffixes is taken from the analysis model.

No free translation – „in the center” – is given, as creating this automatically is not technically feasible at this point. For materials that are fed into the corpus infrastructure, English variants can however be included. This applied in the case of the example sentences from the Mari-English dictionary and the New Testament – where I inserted the appropriate verses of the King James Version as translations. Due to the existence of English „translations”, example sentences extracted from the corpus in this paper have been disproportionately taken from the New Testament.

When the analyzer processes entire sentences – as occurs when texts are inserted into the corpus infrastructure – it generally encounters forms that are ambiguous in one way or another. The following example shows the output of a random sentence, meaning „That grove stands on the shore of a large lake”, entered into the morphological analyzer. The circled numbers are not part of the output, but were added to allow easier reference to individual words of the glossing:

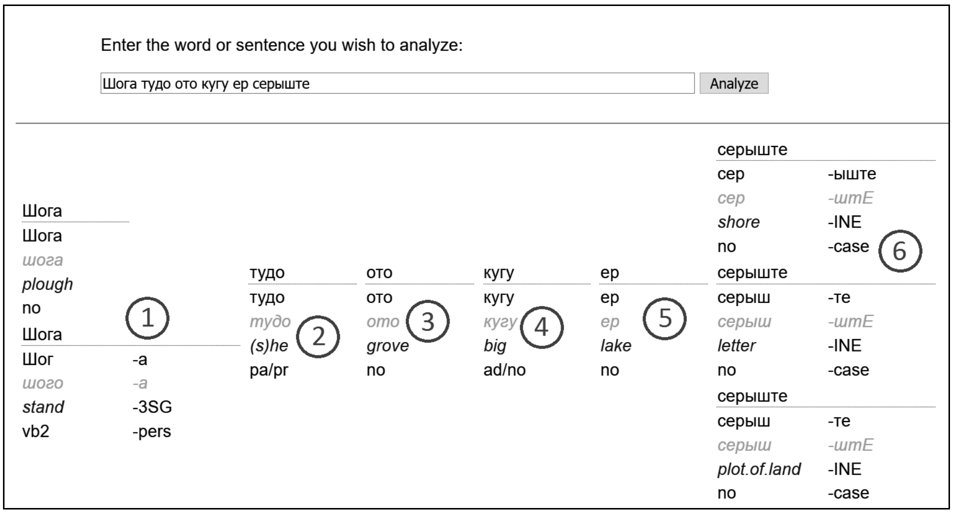

Figure 2
Morphological analysis of a sentence
(screenshot from morph.mari-language.com)

The morphological analyzer does not have any disambiguation mechanisms at this point. When it encounters a lexically or/and morphologically ambiguous form – such as ① *šoga* and ⑥ *serəšte* – all interpretations that would be valid given the morphological model and the lexical base are returned (in this case, the second interpretation – ’stand’ + -3SG – is correct in the case of ①, and the first interpretation – ’shore’ + -INE – in the case of ⑥. In addition, when words are ambiguous with respect to their part of speech – as ② *tudo* and ④ *kugu* are –

all possible classifications are given. In the case of polysemous lexemes such as ② *tudo*, the first translation given in the lexical base is given (which in this particular case is not appropriate – *tudo* can occur as a personal and a demonstrative pronoun; here it is a demonstrative pronoun ’that’).

When texts are fed into the corpus infrastructure, interlinearizations are saved in the repository in this format – including all interpretations that the morphological analyzer, blind to syntax and context, deems possible. Manual disambiguation is possible through the corpus infrastructure’s back end – authorized users can log in, select the correct form for morphologically ambiguous forms, and change the glosses if the wrong aspect of meaning was given for a lexeme (cf. BRADLEY 2015). I have done this for the 2,508 tokens taken from *Oŋaj marij jə̑lme* (RIESE ET AL. 2012), but not for the other materials found in the corpus infrastructure at present.

The corpus infrastructure enables users to search for grammatical patterns within all the resources fed into the infrastructure or within individual resources. This procedure is best illustrated by a practical example. Let us assume that we wish to search for the future-necessitive participle in *-šaš* (cf. ALHONIEMI 1985: 138–140, BERECZKI 1990: 60) in postpositional constructions. The simplest way to see how the structure in question is realized by the morphological analyzer – i.e. what exactly one should search for in the corpus – is to enter one example of the structure in question into the interface at morph.mari-language.com (> ’Analyzer’). For example, *paša ə̑štə̑šaš godə̑m* ’when (we are) supposed to be working’[15] (ALHONIEMI 1985: 140):

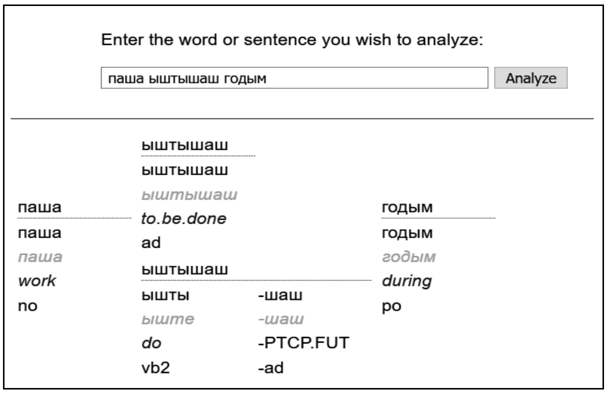

Figure 3
’when (we are) supposed to be working’
(screenshot from morph.mari-language.com)

15 ’… kun meidän on tehtävä työtä…’

The form *əštəšaš* is erroneously determined to be ambiguous by the analyzer, as the participle *əštəšaš* is known to the lexical base as an adjective. Hence, the analyzer recognizes the word both as an adjective without a defined internal structure and a participle derived from the verbal stem *əšte-* 'to do'. An analogous software tool for English would declare the word *tired* ambiguous for the same reason: it would recognize it as the adjective *tired*, and as the past participle of the verb *to tire*.

For my purposes, the complex reading is relevant. An abstraction of the pattern I am interested in would be: the gloss „-PTCP.FUT" in one word, followed by an item with the part-of-speech value „po" as the next word in the sentence. If one visits the page corpus.mari-language.com and clicks the button „[Search]" (or picks an individual resource first, to search only within it), a web mask appears that allows users to enter this data: „gloss", „equals", „-PTCP.FUT", „next word", „part of speech", „equals", „po". The remaining fields can be left empty and will be ignored by the program.

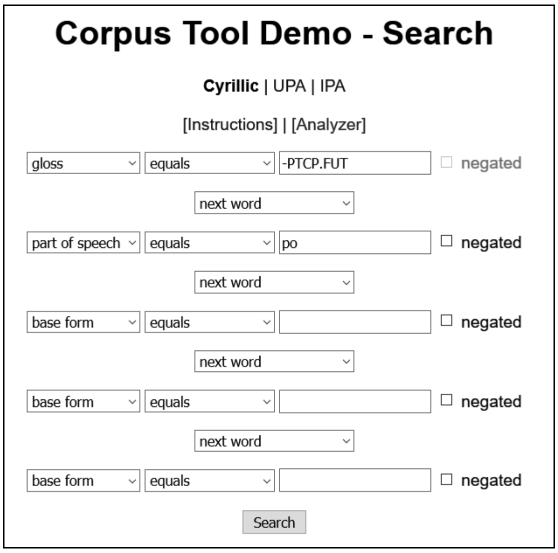

Figure 4
The pattern entered into the search mask
(screenshot from corpus.mari-language.com)

When searching the entire illustrative corpus, the program returns 256 hits for this particular structure. Depending on users' needs, they can now either copy individual usage examples of the structure at hand – as I did when writing the previous chapter – or make quantitative comparisons: how frequently is this participle used in participial constructions compared with other participles? Are there differences in the frequency of this construction between individual resources in the corpus?

If I search for all four participles detailed above in this construction, the preliminary number of hits is as follows:

-PTCP.ACT	969
-PTCP.PASS	5286
-PTCP.FUT	256
-PTCP.NEG	210

Table 1
Participles coupled with postpositions in corpus

These figures cannot be taken at face value. While the search for the future participle followed by a postposition seems to return mostly appropriate results (though these results would also have to be sighted in a true quantitative study), this is not always the case, and users must themselves distinguish between true positive results (i. e. sentences in which the feature at hand is correctly found) and false positive results (i. e. sentences in which the feature is erroneously reported by the software). This is plainly evident in relation to the active participle – „-PTCP.ACT" – which I previously claimed does not occur as a verbal noun, but which is nonetheless found coupled with a postposition 969 times. None of these hits are appropriate, as they all represent examples in which a nominalized adjectival participle is used as an agent noun or examples in which a passive participle was read into an ambiguous word form, but does not actually occur. Below is one example of each of these types of false positive results:

Meadow Mari (Raamatunkäännösinstituutti 2007: Mark 9:16)

(36) *iisus* *zakon* *tunəktə-šo-βlak* *deč̑* *jod-ən*
Jesus law teach-PTCP.ACT-PL from ask-PST2.3SG
'[Jesus] asked the scribes…'

Meadow Mari (RIESE ET AL. 2014–: мучко)

(37) *ilə̑š-em* *muč̑ko*
life-PX1SG through
'all my life'

In the first example, the participle *tunəktə-šo* 'teaching > teacher' is used as a noun, as it commonly is. This is due to conversion, and not to this participle being used as a verbal noun. In the second example, *ilə̑šem* is morphologically ambiguous: stripped from context, the form could also be read as *ilə-š-em* 'live-PTCP.ACT-PX1SG' (with the final vowel of the participle deleted by the possessive suffix). While this reading is not fitting in the given context, the morphological analyzer returns it and saves it into the repository of the corpus, and as a consequence the pattern under consideration is erroneously detected here.

For the negative participle, the results are fairly mixed: While some examples found are examples of true verbal nouns coupled with a postposition, the participle is clearly simply an adjectival form used nominally by means of conversion in other examples. Here is one true positive and one false positive example:

Meadow Mari (Raamatunkäännösinstituutti 2007: Hebrews 4:6)

(38) … *mut* *koləšt-dəmə-št-lan* *köra* …
word listen-PTCP.NEG-PX3PL-DAT because.of
'… because of unbelief:'

Meadow Mari (Raamatunkäännösinstituutti 2007: Mark 9:16)

(39) … *čən* *jumə-m* *palə-dəme-βlak* *deke* *ida* *kaj…*
true god-ACC know-PTCP.NEG-PL to NEG.IMP.2PL go.CNG
'… Go not into the way of the Gentiles…'

The 5286 constructions found with the passive participle, while mostly appropriate judging from a quick inspection, are so numerous that a visual examination of these by the end users would be difficult. It seems clear that the combination „passive participle + postposition" outnumbers all other combinations considered here by a factor of at least 20:1. As this particular construction is so exceedingly common, the corpus infrastructure is less necessary as a tool allowing users to find usage examples – these are easy enough to find simply by browsing through texts. However, the corpus infrastructure can be used to answer more detailed questions regarding the usage of the passive participle in combination with postpositions. For example:

– How commonly does the passive participle take a possessive suffix (marking the action's agent) in this construction? The appropriate search query for this question is:

„gloss"	„equals"	„-PTCP.PASS"
	„in same word"	
„part of speech"	„equals"	„-poss"
	„next word"	
„part of speech"	„equals"	„po"

This query returns 1657 hits – i. e. in roughly 30% of cases, the passive participle takes a possessive suffix in this construction. This number should be handled with care as I have not removed the false positive hits from either the 5286 total constructions or the 1657 construction with a possessive suffix.

– As mentioned above, the postposition *köra* governs the dative case and combines with the dative forms of the passive participle. Are there any other postpositions that combine with non-nominative forms of the passive participle? This can be determined with the following query:

„gloss”	„equals”	„-PTCP.PASS”
	„in same word”	
„part of speech”	„equals”	„-case”
	„next word”	
„part of speech”	„equals”	„po”

This search query returns 287 hits. While many of these are examples of the postposition *köra* coupled with the dative form of a passive participle, sentences in which the semantically similar postposition *βerč́* ’because of; for’ co-occurs with non-nominative forms can be quickly found as well. For example:

Meadow Mari (Raamatunkäännösinstituutti 2007: Mark 9:41)

(40)	…	*χristos-ən*	*ul-mə-lan-da*	*βerč́*	…
		Christ-GEN	be-PTCP.PASS-DAT-PX2PL	because.of	

’… because ye belong to Christ…’

The following search, motivated by this result, reveals a more complicated picture with respect to the postposition *βerč́*:

„gloss”	„equals”	„-PTCP.PASS”
	„next word”	
„base form”	„equals”	„верч”

This search yields 22 results. A quick visual inspection reveals that the passive participle is only in the dative in 4 of these cases and is in the nominative in all others. There seems to be some alternation regarding the government of *βerč́*, but the illustrative corpus lacks the depth and the metadata that would allow a competent analysis of such a fine aspect – the question of whether this alternation is determined by semantics, dialectal differences, language change, random chance, etc. The following query can, however, be used to determine that the postposition *köra* is not subject to the same alternation:

„base form”	„equals”	„кöра”	
	„previous word”		
„gloss”	„equals”	„-DAT”	„negated”

This search finds occurrences of *köra* where the previous word is not marked for the dative. No true positive results are returned by this query, i. e. *köra* only co-occurs with the dative in all materials in the corpus.

4. Prospects

This paper has illustrated how the corpus framework I have created has the potential of allowing the quantitative study of grammatical structures in the 'small' Uralic languages of Russia on a scale unprecedented to date. For comparatively rare constructions (e. g. negative participle + postposition), the infrastructure offers the possibility of quickly finding usage examples in a large body of texts. For comparatively common constructions (e. g. passive participle + postposition), it offers the possibility of studying the distribution of the feature contrastively: diachronically, dialectally, sociolinguistically, by genre, etc. The strength of this tool is determined by the range of texts available through it and the quality of the metadata attached to these texts. The illustrative corpus as it stands now is in no way representative and does not enable such quantitative studies for the time being – but the potential is there.

A number of possibilities suggest themselves as ways of making this application more powerful in the years to come. A matter of utmost priority would be to adapt the morphological analyzer to handle the second literary norm of Mari, Hill Mari. Numerous texts are available in Hill Mari – e. g. the recently published Hill Mari translation of the New Testament (Raamatunkäännösinstituutti 2014) – and these could potentially be fed into this corpus infrastructure. The inclusion of modern Eastern Mari newspaper texts from Bashkortostan – these exist in digital format, but copyright remains a problem – would ensure that three of the four dialect groups of Mari (cf. MOISIO – SAARINEN 2008: VIII) are covered in their contemporary forms to some extent.

Dialect text collections gathered by Finnish, Hungarian, and Mari scholars in the late 19th and early 20th centuries (e. g. GENETZ 1895, GENETZ 1889, PAASONEN – SIRO 1939, WICHMANN 1931, ALHONIEMI – SAARINEN 1983–1994, Beke 1957–1995) could in theory add a diachronic perspective to the corpus tool and could ensure that all four major dialect groups of Mari are covered. Should these texts ever be digitized, it would be conceivable that they, after manual interlinearization (as they do not follow literary standards, the morphological analyzer could not be easily applied to them), could be fed into the same corpus infrastructure, allowing users to compare the distribution of features in different dialects in the late 19th and early 20th centuries. For three of the four dialect groups (Meadow, Hill, Eastern), they would allow a comparison between a time point as early as 1885 (GENETZ 1895: VII) and today. This massive undertaking is not currently on my horizon.

Recent efforts by the National Library of Finland offer a more realistic perspective for the diachronic study of Mari with the infrastructure at hand in the near future. Thanks to the efforts of the National Library's employees, a wide range of materials found in various libraries of Russia have been scanned and made accessible at the address uralica.kansalliskirjasto.fi (National Library of Finland 2013–). These scanned materials include numerous issues of newspapers from the 1920s, 1930s and 1940s, including some from peripheral locations that in many cases are close to collection points of the text collections discussed above. The texts published to date represent three of the four dialect groups (Meadow, Hill, Eastern). It would be quite feasible to integrate these texts into the corpus infrastructure eventually, as they do adhere to literary norms. While these norms are not identical to the modern ones, they are compatible with these. Unfortunately, the digitized materials are not yet accessible in an adequate format to make this integration possible. Thus, it also remains a future prospect for the time being.

5. Acknowledgements

I am grateful to the Kone Foundation for funding the research project „Mari Web Project: Phase 2", in the course of which I undertook most of the research presented here.

I would like to thank ZSÓFIA SCHÖN for giving me access to her Khanty text collection and helping me gloss the Khanty examples taken from it. Likewise, CHRISTIAN PISCHLÖGER's help in finding Udmurt examples and glossing these was appreciated. I am also grateful to NIKO PARTANEN for sharing useful insights into Komi with me.

I would also like to thank TIMOTHY RIESE for proof-reading the paper and helping me transcribe the Mansi example sentences.

Glossing abbreviations

ACC	= accusative	MOM	= momentary
ACT	= active	NEG	= negative
CAUS	= causative	NMLZ	= nominalizer
CNG	= connegative	PASS	= passive
CVB	= converb	PERF	= perfect
DAT	= dative	PL	= plural
FUT	= future	PRS	= present
GEN	= genitive	PST	= past
IMP	= imperative	PST1	= past tense 1 (in Mari)
INE	= inessive	PST2	= past tense 2 (in Mari)
INF	= infinitive	PTCP	= participle
INS	= instructive	PX	= possessive suffix
LAT	= lative	SG	= singular

References

ALHONIEMI, ALHO (1985), Marin kielioppi. Apuneuvoja suomalais-ugrilaisten kielten opintoja varten X. Suomalais-ugrilainen Seura, Helsinki.

ALHONIEMI, ALHO – SAARINEN, SIRKKA (1983–1994), Timofej Jevsevjevs Folklore-Sammlungen aus dem Tscheremisschen (I–IV). Mémoires de la Société Finno-ougrienne 184, 199, 211, 219.

BALANDIN, A. N. [БАЛАНДИН, А. Н.] (1960), Самоучитель мансийского языка. Учпедгиз, Ленинград.

BEKE, ÖDÖN (1957), Mari szövegek. I. Akadémiai Kiadó, Budapest.

BEKE, ÖDÖN (1961a), Mari szövegek. III. kötet. Akadémiai Kiadó, Budapest.

BEKE, ÖDÖN (1961b), Mari szövegek. IV. kötet. Akadémiai Kiadó, Budapest.

BEKE, ÖDÖN (1995), Mari szövegek II. (Tscheremissische Texte II.) Berzsenyi Dániel Tanárképző Főiskola, Szombathely.

BERECZKI, GÁBOR (1990), Chrestomathia Ceremissica. Tankönyvkiadó, Budapest.

BEZNOSIKOVA, L. M. – AJBABINA, E. A. – KOSNYREVA, R. I. [БЕЗНОСИКОВА, Л. М. – АЙБАБИНА, Е. А. – КОСНЫРЕВА, Р. И.] (2000), Коми–роч кывчукӧр (Коми–русский словарь). Коми книжное издательство, Сыктывкар. [published online at dict.komikyv.ru/index.php/index/7.xhtml, accessed on 2015-09-23]

BRADLEY, JEREMY [БРЭДЛИ, ДЖЕРЕМИ] (2011), »Mari web project« и его марийский морфоанализатор. In: JUZYKAJN, T. V. ET AL. [ЮЗЫКАЙН, Т. В. И ДР. (ред.)] (eds), Языки меньшинств в компьютерных технологиях: опыт, задачи и перспективы. Министерство культуры, печати и по делам национальностей Республики Марий Эл, Йошкар-Ола. 82–89.

BRADLEY, JEREMY (2015), Corpus.mari-language.com: A Rudimentary Corpus Searchable by Syntactic and Morphological Patterns. Septentrio Conference Series 2015: 2, Septentrio Academic Publishing, University of Tromsø, Tromsø. 57–68 [published online at septentrio.uit.no/index.php/SCS/article/view/3468, accessed on 2015-09-23].

ČAVAJN, S. G. [ЧАВАЙН, С. Г.] (1967), Элнет. Марийское книжное издательство, Йошкар-Ола.

COLLINDER, BJÖRN (1960), Comparative Grammar of the Uralic Languages. Almqvist & Viksell, Stockholm.

CSEPREGI, MÁRTA (2011), Szurguti osztják chrestomathia (revised edition). SUA supplementum 6. [published online at www.babel.gwi.uni-muenchen.de/media/downloads/SzOCh_FUT_20110721.pdf, accessed on 2015-09-22].

CSÚCS, SÁNDOR (1998), Udmurt. In: ABONDOLO, DANIEL (ed.), The Uralic Languages. Routledge, New York. 276–304.

DRYER, MATTHEW S. (2013a), Order of Subject, Object and Verb. In: HASPELMATH – DRYER 2013.

DRYER, MATTHEW S. (2013b), Order of Adposition and Noun Phrase. In: HASPELMATH – DRYER 2013.

GALKIN, I. S. ET AL. [ГАЛКИН, И. С. И ДР. (ред.)] (eds) (1990–2005), Словарь марийского языка (I–X). Марийское книжное издательство/МарНИЙ, Йошкар-Ола [published online at dict.komikyv.ru/index.php/index/8.xhtml, accessed on 2015-09-22].

GENETZ, ARVID (1889), Ost-tscheremissische Sprachstudien – Sprachproben mit deutscher Übersetzung. Journal de la Société Finno-ougrienne 7.

GENETZ, ARVID (Hrsg.) (1895), Volmari Porkka's tscheremissische Texte mit Übersetzung. Journal de la Société Finno-ougrienne 13/1.

HASPELMATH, MARTIN (1995), The converb as a cross-linguistically valid category. In: HASPELMATH, MARTIN – KÖNIG, EKKEHARD (eds), Converbs in Cross-Linguistic Perspective – Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds. Mouton de Gruyter, Berlin. 1–55.

HASPELMATH, MARTIN – SIMS, ANDREA D. (2010), Understanding Morphology (Second edition). Hodder Education, London.

HASPELMATH, MARTIN – DRYER, MATTHEW S. (eds) (2013), The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig [published online at wals.info/chapter/85, accessed on 2015-03-02].

LANDMANN, ANGELIKA (2014a), Tatarisch – Kurzgrammatik. Harrassowitz, Wiesbaden.

LANDMANN, ANGELIKA (2014b), Tschuwaschisch – Kurzgrammatik. Harrassowitz, Wiesbaden.

MOISIO, ARTO – SAARINEN, SIRKKA (2008), Tscheremissisches Wörterbuch. Lexica Societatis Fenno-Ugricae 32, Helsinki.

National Library of Finland (2013–), Kansalliskirjasto Uralica. National Library of Finland, Helsinki [published online at uralica.kansalliskirjasto.fi, accessed on 2015-09-09].

PAASONEN, HEIKKI – SIRO, PAAVO (1939), Tscheremissische Texte. Mémoires de la Société Finno-ougrienne 78.

Raamatunkäännösinstituutti (2007), У Сугынь. Raamatunkäännösinstituutti, Helsinki.

Raamatunkäännösinstituutti (2014), У Согонь. Raamatunkäännösinstituutti, Helsinki.

RÉDEI, KÁROLY (1978), Syrjänische Chrestomathie – Mit Grammatik und Glossar. Studia Uralica 1, Verband der wissenschaftlichen Gesellschaften Österreichs, Wien.

RIESE, TIMOTHY – BRADLEY, JEREMY – GUSEVA, ELINA (2014–), Mari–English Dictionary. Department of Finno-Ugrian Studies, University of Vienna, Vienna [published online at dict.mari-language.com].

RIESE, TIMOTHY – BRADLEY, JEREMY – YAKIMOVA, EMMA – KRYLOVA, GALINA (2012), Оҥай марий йылме: A Comprehensive Introduction to the Mari Language (Version 2.1). Department of Finno-Ugrian Studies, University of Vienna, Vienna [published online at omj.mari-language.com, accessed on 2015-09-22].

SKRIBNIK, ELENA (2003), Buryat. In: JANHUNEN, JUHA (ed.), The Mongolic Languages. Routledge, New York. 102–128.

WICHMANN, YRJÖ (1931), Volksdichtung und Volksbräuche der Tscheremissen. Mémoires de la Société Finno-ougrienne 59.