On the questions in developing computational infrastructure for Komi-Permyak

Jack Rueter
University of Helsinki
jack.rueter@helsinki.fi

Niko Partanen
University of Helsinki
niko.partanen@helsinki.fi

Larisa Ponomareva
University of Helsinki
dojegpl@gmail.com

Abstract

There are two main written Komi varieties, Permyak and Zyrian. These are mutually intelligible but derive from different parts of the same Komi dialect continuum, representing the varieties prominent in the vicinity and in the cities of Syktyvkar and Kudymkar, respectively. Hence, they share a vast number of features, as well as the majority of their lexicon, yet the overlap in their dialects is very complex. This paper evaluates the degree of difference in these written varieties based on changes required for computational resources in the description of these languages when adapted from the Komi-Zyrian original. Primarily these changes include the FST architecture, but we are also looking at its application to the Universal Dependencies annotation scheme in the morphologies of the two languages.

Дженыта висьталöм

Коми кылын кык гижан кыв: пермяцкöй да зырянскöй. Öтамöд коласын нія вежöртанаöсь, но аркмисö нія разнöй коми диалекттэзісь. Пермяцкöй кыв олö Кудымкар лапöлын, а зырянскöй – Сыктывкар ладорын. Пермяцкöй да зырянскöй литературнöй кыввезын эм уна öткодьыс, öткодьöн лоö и ыджыт тор лексикаын, но ны диалектнöй чертаэзлöн пантасьöмыс öддьöн гардчöм.

Эта статьяын мийö видзöтам эна кык кывліс ассямасö сы ладорсянь, мый ковсяс вежны лöсьöтöм зырянскöй вычислительнöй ресурсіс, медбы керны сыись пермяцкöйö. Медодз энö вежсьöммесö колö керны FST-ын, но мийö сідзжö видзöтам, кыдз FST лöсялö Быдкодь Йитсьöммезлöн схемаö морфология ладорсянь.

1 Introduction

The Komi language is a member of the Permic branch of the Uralic language family. By nature, it is a pluricentric language, which, in addition to having two strong written traditions (Komi-Zyrian and Komi-Permyak), can be divided into several varieties. Although some of the other varieties do exhibit written use, no actual new written standards seem to be emerging (Цыпанов, 2009). Both Zyrian and Permyak have long and established written traditions, with continuous contemporary use. There are also numerous dialect resources currently available, i.e. Пономарева (2016) for Northern Permyak dialects.

Komi-Permyak and Komi-Zyrian are extremely agglutinative, but the two standards have different tendencies in their criteria for the definition of an orthographic word. Inflection mainly involves the use of suffixes, which, in the case of nominals, are NP final. Hence, contextual dropping of the head noun means that formatives shift to the next rightmost constituent of the NP.

While Komi-Zyrian has a long tradition of computerized morphological analysis, as finite-state transducers have been developed for it since the mid 1990s (Rueter, 2000), the computational resources for Komi-Permyak have been the focus of less intensive work. This article is intended as a roadmap for further development of Komi-Permyak computational resources. The morpho-

Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages, pages 15–25
Wien, Austria, January 10 — 11, 2020.
https://doi.org/10.18653/v1/FIXME

logical features discussed in this paper have largely been already implemented in the FST in Giellatekno infrastructure, and the work has been primarily carried out by Jack Rueter. Ongoing work includes intense work on paradigmatic description by Larisa Ponomareva (Rueter et al., 2019a), which has been published within AKU-infrastructure. AKU is an abbreviation for *Avointa Kieliteknologiaa Uralilaisille/Uhanalaisille kielille* (Open language technology for Uralic/Endangered languages). Other projects that are directly associated with this are uralicNLP (Hämäläinen, 2019), Akusanat (Hämäläinen and Rueter, 2019b) and Ver'dd (Alnajjar et al., 2019) (see also *On Editing Dictionaries for Uralic Languages in an Online Environment*, in this publication). Forthcoming work includes the expansion of the initial Permyak treebank found in Universal Dependencies version 2.5 (Zeman et al., 2019), i.e. further work on what is scheduled for the next UD release, hence the underlying acuteness of further work with this often understudied, but central variety of Komi.

As the Komi-Zyrian finite state transducer has already reached a very advanced state, and the languages are so similar to one another, it is necessary to ask how far we can reuse the components of a Zyrian analyzer when working with Permyak. Although it has been suggested this kind of resource-sharing becomes most useful at higher levels of grammar, especially syntax Antonsen et al. (2010), in the case of very closely related languages the number of shared elements is considerable at all levels of the language. We understand this is a slippery road, and uttermost attention has to be paid to full respect of Permyak features and particularities, so that we do not simply force the Zyrian conventions upon Permyak. At the same time starting to develop a Permyak infrastructure from scratch feels like a missed opportunity to find some synchrony. In this article we attempt to describe all those particularities and pitfalls that have to be considered when one endeavors further the analysis of Permyak. Our approach is also in some sense comparable to that of (Pirinen, 2019).

As two Komi-Zyrian treebanks are also under continuous development (Partanen et al., 2018) it is particularly important to pay attention to Komi infrastructure at large. A recent survey of Uralic Universal Dependencies treebanks (Rueter and Partanen, 2019) showed that more work is needed to harmonize the annotation between languages, and working on closely related languages is certainly where similarity is most easily enforced but also most logically expected. In this context, it also has to be taken into account that several other smaller Uralic languages have had their own treebanks introduced in the past couple of years, e.g. Erzya (Rueter and Tyers, 2018), Karelian (Pirinen, 2019) and North Saami (Tyers and Sheyanova, 2017). This kind of work that concentrates more on manually annotated corpora complements the descriptive work on morphological analyzers extremely well. A well functioning morphological analyzer, however, seems to be one of the best starting points for further language technology, which provides a motivation for the work grounded in this paper.

Since this paper describes only the current, rather preliminary state of investigation, we have published the list of discussed features as an accompanying database (Rueter et al., 2020). This database is available online,[1] and can be extended as needed when a larger inventory of differing lexical items and syntactic constructions becomes available. Since we hope the comparative investigations reach other dialects and variants of the Permic languages, the database has been named accordingly with an optimistic mindset.

In this paper we have chosen to distinguish morphological suffixes from clitics with preceding hyphens. This will be achieved through the use of hyphens to set of morphological suffixes and equal signs to indicate clitics and other elements separated by a hyphen in the written norms. All examples in the paper, unless the source is given, have been created by Larisa Ponomareva.

2 Orthographic distinctions

During the development of the two Komi norms, a few orthographic distinctions have been made. These distinctions can be attributed to sub-dialect variation, on the one hand, and arbitrary spelling principles, on the other. The arbitrary spelling choices are simply orthographic, and do not necessarily relate to actual phonological differences in the languages.

Arbitrary choice involves the definition of word (i.e. written unit without white space) and the selection of background language form and letter combinations. It will be observed below that the Komi-Permyak converb paradigms are minimalis-

[1] https://langdoc.github.io/comparative-permic-database

tic in comparison to those of Komi-Zyrian. Komi-Permyak, on the one hand, tends to write separate words, and Komi-Zyrian tends to write single concatenated words, on the other. This can be exemplified in the converb paradigms for Komi-Permyak *-ик* /-ik/, which can only appear alone, in the singular illative or with possessive suffixes, whereas the Komi-Zyrian *-иг* /-ig/ converb also takes plural marking, cases, as well as numerous other elements *-ӧн* /-ən/, *-моз* /-moz/, *-тыр* /-tɨr/, *-тырйи* /-tɨrji/, *-тыръя* /-tɨrja/, *-кости* /-kosti/, *-коста* /-kosta/,... (Некрасова, 2000, p. 344–353)

Arbitrary character combinations can be illustrated best with two prominent paradigms: the geminate voiced palatal affricate is represented in Zyrian with *ддз ddz* but in Permyak this same affricate is rendered with *дзз dzz*. Consonants followed by a palatal glide and subsequent vowel are written using hard and soft sign combinations. In Zyrian, the norm is to use a soft sign following inherently soft consonants, whereas hard signs are used in other instances. In Permyak, on the contrary, the hard sign is used with specifically hard consonants, while the soft sign is used as default for other combinations.

One orthographic convention that works similarly in Permyak and Zyrian alike is *l : v* variation in stem-final position. This variation is not present in this form in any of the Komi-Permyak dialects, but as a literary convention it is shared with Zyrian standard. Permyak dialects, it will be noted, generally display a multitude of l-related subsystems (Баталова, 1982, p. 58).

Orthographic distinctions between the two Komi norms present few problems. On the one hand, computational distinctions are only attested in the use of the few paragogic consonants in alternate Permyak morphological forms. On the other, use of the NP plural in both variants appear to follow the same distribution, so any computational issue might only be found at the morpho-syntactic level.

3 Phonetical differences

The morpheme-final t / d correlation between Permyak and Zyrian is a prominent source of predictable morphological and lexical differences. This morpho-phonological difference is found word finally in the Permyak adjective *сьӧкыт* /ɕəkɨt/ and Zyrian *сьӧкыд* /ɕəkɨd/ 'heavy; difficult' as well as other corresponding pairs *-ыт* /-ɨt/ vs. *-ыд* /-ɨd/, Permyak and Zyrian respectively. It can also be observed in the causative derivation marker *-ӧт* /-ət/ vs. *-ӧд* /-əd/ in verb stems, such as *велӧтны* /velətnɨ/ and *велӧдны* /veləd nɨ/ 'to teach', Permyak and Zyrian respectively. The same correlation is also found in the comitative case ending *-кӧт* /-kət/ vs *-кӧд* /kəd/ and the possessive suffixes for the second persons singular and plural: *-ыт* /-ɨt/ vs. *-ыд* /-ɨd/, and *-ныт* /-nɨt/ vs. *-ныд* /-nɨd/. The same voiceless vs voiced correlation might also be detected in the converbs *-икӧ* /-ikə/ vs *-игӧ* /-igə/, Permyak and Zyrian respectively.

On a similar note, there is a correlation between Permyak *ӧ* /-ə/ and Zyrian *-ӧй* /-əj/ in first person singular possessive marking. In verbal morphology, the Permyak morpheme-final *ӧ* /ə/ of the first person plural marker *-мӧ* /-mə/ corresponds to Zyrian endings *-м* /-m/, whereas the Zyrian first person plural imperative usage might include both *-мӧ* /-mə/ and *-мӧй* /-məj/, as in *мунамӧй* /munaməj/ 'let's go'.

4 Morphological differences

Many of the morphological forms provide an illustration of where the human learner may have problems in comprehension while the computer has no problems in computation. There are, however, numerous ways of how a minor difference in one morphological form has a potential impact on ambiguities in other parts of the system.

In this section we go through most essential differences in Komi-Zyrian and Komi-Permyak morphology.

4.1 Paragogical consonants

Both Komi language forms have the same paragogical consonants, but their distribution is varied. In practice, the so-called paragogic consonants are present when the stem is followed by a suffix with an onset vowel, and it is absent when word-final or followed by a consonant (cf. Безносикова et al., p. 16).

Paragogic consonants may be present in Permyak but to a lesser extent than they are in Zyrian due to sub-dialect representation, i.e. many of the sub-dialects do not have them. Komi-Zyrian includes paragogic consonants in its nominal declension and derivation – approximately 0.07 percent of the 12,046 noun stems in the Zyrian transducer have paragogic consonants, but this is reduced to 0.024 once the diminutive/material formative *мор* /mor/ is removed. Komi-Permyak, in contrast, limits its use of paragogic consonants in declensions, and the number of Komi-Permyak stems with paragogic

Permyak	*кыв*	+	*вез* (← -йэз)	*кыввез*
	/kɨv/	+	/vez (← -jez)/	/kɨvvez/
Zyrian	*кыв*	+	*яс*	*кывъяс*
	/kɨv/	+	/jas/	/kɨvjas/
	‘word; language’	+	PL	‘words; languages’

Figure 1: Example plural of /kɨv/ ‘word; language’

Permyak	*му*	+	*эз*	*муэз*
	/mu/	+	/ez/	/muez/
Zyrian	*му*	+	*яс*	*муяс*
	/mu/	+	/jas/	/mujas/
	‘land; country’	+	PL	‘lands; countries’

Figure 2: Example plural of /mu/ ‘land; country’

consonants is smaller. The Komi-Permyak standard language recognizes the paragogic consonants *ӥ* /j/, *к* /k/ and *м* /m/ as alternative variants. The paragogic consonant *ӥ* /j/ is more common than *к* /k/ and *м* /m/, the latter two are found only in a limited set of stems, such as *син* /ɕin/ : *синм-* /ɕinm-/ ‘eye’, *кос* /kos/ : *коск-* /kosk-/ ‘lower back’, *мыш* /mɨʃ/ : *мышк-* /mɨʃk/ ‘back’. Thus the Komi-Permyak literary language supports the use of both *синмӧ пырӧ* /ɕinmə pyrə/ and *синӧ пырӧ* /ɕinə pyrə/ ‘gets in the eye’, where the analysis of *синмӧ* /ɕinmə/ and *синӧ* /ɕinə/ is eye.SG.ILL. (The paragogic *т* /t/ in the verb *локны* /lokɨ/ and *локт-* /lokt-/ ‘to arrive’ is the standard and cannot be left out of the paradigm in either of the literary languages.)

4.2 Plural formation

Phonological variation can be detected in the plural marking of NP heads, where the Zyrian normal plural marker involves the realization of *-яс* /-jas/, on the one hand, and the Permyak normal plural marker calls for either word-final consonant doubling (see fig 1) or, following a vowel, a simple *-эз* /-ez/ (see fig 2), on the other. Orthographically, the word-final consonant *ӥ* /j/ forms an exception to this, here the Cyrillic *е* /je/ without orthographic duplication of *ӥ* /j/ (see fig. 3).

Plural character duplication, which is the primary method of plural formation in Permyak, is also partially present in Zyrian dialects. This, however, is not accepted in the Zyrian written standard. Whereas Zyrian plural is formed with distinct suffix *-яс* /-jas/ (as illustrated in figures 1, 2, 3, above).

4.3 Possessive marking

Although singular possessive marking differs from Zyrian only through expected phonetic correspondence *t / d*, the plural forms display more complex assimilation. While the plural possessive forms in Zyrian are clearly segmentable, i.e. *поӈъяс* : *поӈъясыд* /ponjas/ : /pon-jas-ɨd/ dog-PL : dog-PL-2PSX, the corresponding forms for the second and third person in Permyak are often fused, i.e. *поннэз : поннэт* /pon-nez : pon-net/ dog-PL : dog-PL.2PSX (Лыткин, 1962). Forms with separate elements are, however, also possible. Both form types have already been implemented in the Permyak analyzer.

Permyak	*кай*	+	*ез*	*кайез*
	/kaj/	+	/jez/	/kajjez/
Zyrian	*кай*	+	*яс*	*кайяс*
	/kaj/	+	/jas/	/kajjas/
	‘bird’	+	PL	‘birds’

Figure 3: Example plural of /kaj/ ‘bird’

4.4 Cases

While both literary norms generally describe the number of cases as being sixteen or seventeen, a reality check might be required. The most recent and extensive presentation of Komi-Zyrian, it should be noted, indicates at least 23 cases with new ones appearing all the time (Некрасова 2000:59–62). One reason for this inconsistency is the definition of case: What is a case, and what kinds of combinations they can be used in when speaking of a single referent and a double referent (inclusive elliptic referent). Thus we can observe organic expansion of the local cases and diversion in case enumerations.

Both language norms have regular extensions of the approximative case *-лань* /-laɲ/ ‘towards X’. The case marker may take additional local case combinations, e.g. approximative + elative, in Permyak *-ланись* /-laɲ+ɨç/ and in Zyrian *-ланьысь* /-laɲ+ɨç/ ‘from on towards X’, which is actually just a more specific combination of semantics. Additional extensions mutual to both literary norms include the inessive, illative, prolative, terminative and egressive.

Diversity between Komi-Zyrian and Komi-Permyak is apparent in both phonetic variation and complementary distribution of morphology. This can be seen in regular nominal declension with regard to the prolative and terminative. The prolative *-ӧд* /-əd/ and translative *-ми* /-ti/, which are both regular declension in Komi-Zyrian, are only represented by a regular prolative *-ӧм* /-ət/ in Komi-Permyak. Albeit, an analogous transitive *-ми* /-ti/ is present present in Komi-Permyak in a few adpositions and adverbs, but it is not considered to be an independent case of its own.

Similarly, the two Komi-Permyak terminative cases in *-ӧдз* /-ədz/ and *-ви* /-vi/ are only represented by one terminative *-ӧдз* /-ədz/ in Komi-Zyrian. As a rule of thumb, we can say that the deviant Komi-Permyak *-ви* /-vi/ might be replaced in most places by *-ӧдз* /-ədz/, but research is still required to establish where the semantics of these two forms are distinct. Initially, it may be said that *-ви* /-vi/ can be used when indicating motion up to a boundary, whereas *-ӧдз* /-ədz/ implies both up to and passing that boundary.

Phonetic diversity is observed in the dative and elative cases. While the Zyrian dative is marked with *-лы* /-lɨ/, Permyak uses *-лӧ* /-lə/. Similarly, elative and ablative differ in their vowels. In Zyrian, the elative is marked with *-ысь* /-ɨç/ and the ablative with *-лысь* /-lɨç/, whereas in Permyak the corresponding forms are elative *-ись* /-iç/ and ablative *-лісь* /-liç/

When inspecting NPs where the head has been deleted because it can be derived contextually, as discussed in the WALS chapter on adjectives without nouns (Gil, 2013), it will be noticed that Komi-Permyak uses a special accusative form for the accusative adjective without a head noun in *-ӧ* /-ə/, while the Komi-Zyrian solution in the same context is *-ӧс* /-əs/, see in Examples 1 and 2 below.

(1) тэныт гӧрдӧ али вежӧ сетны?

tenɨt gərd-ə ali veʒ-ə ɕet-nɨ?
2SG.DAT red-ACC or yellow-ACC give-INF

'shall [I] give you the red one or the yellow one?' (Permyak)

(2) Тэныд гӧрдӧс либӧ вежӧс сетны?

tenɨd gərd-əs ʎibə veʒ-əs ɕet-nɨ?
2SG.DAT red-ACC or green-ACC give-INF

'shall [I] give you the red one or the green one?' (Zyrian)

This difference, although seemingly small, has many implications for possible morphological analysis of such adjective forms. It creates ambiguity between adjective accusative, illative and possessive forms in a way that is not at all present in Zyrian. In addition, the resulting syntactic structure will need very distinct Constraint Grammar rules (Karlsson, 1990).

4.5 Case and possessive marker ordering

Possessive suffixes and case endings in the Komi-Permyak standard may appear in varied order, as illustrated in Example 3.

(3) каньыстӧг : каньтӧгьяс

kaɲɨstəg kaɲtəgjas
cat-PXSG3-CAR cat-CAR-PXSG3

'Without his / her cat' (Permyak)

Similar phenomena are also attested in Komi-Zyrian but not to the same extent (cf. Некрасова 2000, pp.54–95). Instead of changing the order of tags in the transducer according to morpheme order, an additional tag set for suffix ordering +So/CP *case*, *possession* and +So/PC *possession*, *case* has been adapted, as in the description of the two Mari standards (mhr) and (mrj) by Jeremy Bradley, Jack Rueter and Trond Trosterud at Giellatekno. The idea of the extra tag is to retain tag ordering used in testing and constraint grammar construction. In the meantime, an extra tag is made available for possible grammar research.

4.6 Verbal morphology

Both Permyak and Zyrian have dialect variation in verbal morphology, but in Permyak orthography more variation is accepted. For example, first and second person finite verb forms have a possibility to omit the final *-ӧ* /-ə/ in all tenses, both *мунам* /mun-am/ and *мунамӧ* /mun-amə/, for example, have identical meaning 'to_go-1PL.PRES'. Similar variation is also present in Zyrian dialects, but in the literary language it is not accepted, and the Zyrian FST returns an additional error tag.

In the second past tense third person singular, a different kind of variation is present in which *мунӧма* /munəma/ and and*мунӧм* /munəm/ with both being accepted. In Zyrian, only the first variant is in the literary standard. This has some impact to the possible tags of corresponding participles. In the second past tense second person singular, however, variation is present in the two possible forms such as *мунӧмат* /munəmat/ and *мунӧмыт* /munəmɨt/ 2SG.PST2. Again, there is no difference in meaning. The latter form is directly comparable to the Zyrian form *мунӧмыд* /munəmɨd/ through a phonological difference already described above, see Section 3.

In the third person plural present the variation is similar, but with different elements: *мунӧны* /munəni/ and *мунӧн* /munən/ 'to_go-3PL.PRES'.

Again, there is no conceivable difference in meaning. The shorter form seems to be used more in the spoken language. This variation is not present in any form in Zyrian.

There are parts of Permyak verbal morphology that have no counterparts in the Zyrian standard language. One of the most frequent differing forms are the third person plural past and future indicative verb forms. In Permyak, the paradigm in past, present and future can be illustrated with the verb *мунны* /munnɨ/ 'to go', *мунісӧ* : *мунӧны* (or *мунӧн*) : *мунасӧ* /munisə/ : /munənɨ/ (or /munən/) : /munasə/. In Zyrian the corresponding paradigm would be *мунісны* : *мунӧны мунасны* /munisnɨ/ : /munənɨ/ : /munasnɨ/, which illustrates how forms with *-sə* are lacking.

Permyak past tense formation is more regular than Zyrian, which displays complex variation in possible homonymy for first and third person past tense forms of some intransitive verbs, such that *муні* /muni/ could be both a first or third person singular form. In Permyak, the only verb that displays this variation is *вӧвны* /vəvnɨ/ 'to be', whereas other verbs are regularly marked: *муні* /mun-i/ to_go-1SG.PST *муніс* /mun-is/ to_go-3SG.PST.

In the Permyak second past tense the form *мунӧмасьӧ* /munəmaɕ/ corresponds to Zyrian *мунӧмаӧсьӧ* /munəmaəɕ/. Here the morpheme suffixation in Zyrian is more transparent. Similar forms are also possible in Zyrian dialects, but they do not occur in the written standard. From the perspective of morphological analyzer construction, these forms pose no challenge.

Permyak connegatives are formed differently from their Zyrian counterparts, so that Permyak plural connegative is always marked with *-ӧ* /-ə/, e.g. *оз мунӧ* 'he/she does not go' : *озӧ мунӧ* 'they do not go' /oz munə/ : /ozə munə/. In Zyrian, the plural connegative would be formed as *оз мунны* /oz munnɨ/ 'they do not go', with the connegative form identical to the infinitive of the verb. In this detail, the Permyak connegative is less ambiguous than Zyrian, and i.e. some of the Constraint Grammar rules that disambiguate this currently in Zyrian would not be needed.

Another difference associated with connegatives is the second person plural negation verb forms *одӧ* /od/ and *одӧ* /odə/ in Permyak, which are distinct from their Zyrian counterparts *он* /on/ and *онӧ* /onə/. The same stem is also present in past tense forms, and regularly matches the past tense paradigm with stem initial э- /e-/. The variation in vowel in the end behaves as already described above.

Permyak imperatives have multiple forms not found in Zyrian. Forms created with *-me* /-ce/, e.g. *мунӧте* /munəce/ 'go-IMP.2PL' and *босьтӧте* /boɕtəce/ 'take-IMP.2PL', do not differ in their meaning from more common imperative forms, such as *мунӧ* /munə/ 'go-IMP.2PL' and *босьтӧ* /boɕtə/ 'take-IMP.2PL'. The former forms, however, may be more colloquial (Лыткин, 1962, 249). Forms marked with *-me -ce* are present in plural first and second persons.

Another type of imperative is formed with *-ко* /-ko/. In the orthography it is written with a hyphen. It is used in second person singular, and in the first and second person plural. This imperative has a softer meaning, more of a request than a command. We use the tag +Prec, as in precative[2]. This form is a direct parallel to the Russian *-ка* /-ka/, which also indicates a request, e.g. *возьмите-ка* /voʑmice-ka/ 'do take [it]'.

Related to imperatives, the optative is formed in Permyak written language with two particles *асьӧ* /aɕ/ and *мед* /med/. The former particle does not exist in Komi-Zyrian.

The converb system in Permyak displays some characteristics not found in Zyrian. One difference is uniquely the Permyak converb *-тӧн* /-tən/. It expresses simultaneous action of two verbs.

(4) муні сьывтӧн

mun-i *ɕɨv-tən*
go-1SG.PST sing-CNV

'I went singing' (Permyak)

Besides converb forms that are not marked for person, there are also forms with possessive suffixes. These, unexpectedly, occur with palatalization and gemination of the stem-final consonant, as in:

(5) муні сьывтӧнням

mun-i *ɕɨv-təɲːam*
go-1SG.PST sing-CNV.1SG

'I went singing' (Permyak)

In fact, this palatalization and concurrent gemination occurs in other possessive forms, too:

[2]https://glossary.sil.org/term/precative-mood

(6) увтöттяс

uvt-əcːas
under-PRL.PxSg3

‘(to go) under (something)’ (Permyak)

In this latter form the prolative and possessive suffix are not clearly separable, which again illustrates the more fusional morphology of Permyak when compared to Zyrian. (Looking back at the plural morpheme, it will be noted that palatalization is a distinguishing factor in the possessive forms)

Another converb that lacks a complete correspondence in Komi is the Permyak *-ик* /-ik/. In Zyrian there is a cognate converb *-иг* /-ig/, and this form also expresses simultaneous action as the Permyak *-төн* /-tən/ converb discussed above. There are, however, small differences between the languages. In Permyak the converb when not used as an unmarked complement is always used with the unambiguous illative case or the ambiguous illative case with possessive suffixing, whereas in Zyrian the instrumental is used in the forms that are not marked for possessor. In both languages, however, the possessive forms are deductively in the illative (as determined by the semantic use of the illative), and they are structurally formed in identical way, i.e. *мун-ікас* /mun-ikas/ go-CNV.ILL.3SG, Zyrian *мун-ігас* /mun-igas/ go-CNV.ILL.3SG ‘while going’

4.7 Derivational morphology

There are individual derivational morphemes that are present in Permyak but not in Zyrian. There is *-жуг* /-ʒug/ that forms pejoratives, and multiple diminutives such as *-ок* /-ok/, *-очка* /-oʧka/ and *-иньöй* /-iɲəj/.

In adjective formation, Permyak has several particular features. It is possible to form new adjectives from nouns with suffix *-овöй* /-ovəj/ (Лыткин, 1962, p. 14) Additionally, *-öв* /-əv/ forms excessive adjectives and adverbs, i.e. *ыджыт* : *ыджытöв* /ɨdʒːɨt/ : /ɨdʒːɨtəv/ ‘large : too large’.

There are also numerous derivation types that are found in Zyrian, but are not present in Permyak (Лыткин, 1962, p. 14) *-лун* /-lun/, *-шой* /ʃoj/ and *-ук* /-uk/. As corresponding forms do not exist, the analyzer should either provide no analysis for them, or possibly mark them with a tag indicating they are non-standard.

5 Clitics

Discourse clitic marking in Komi-Zyrian is a salient source of morphological ambiguity. While both *=сö* /=sə/ and *=мö* /=tə/ can be interpreted as clitics, they also represent the accusative case with third person singular and second person singular possessive marking, respectively. As these clitics do not occur in Permyak, such a homonymy is not present in the paradigm, making disambiguation of Permyak less problematic.

There are two discourse clitics commonly used in Permyak, *=му* /=tu/ and *=мо* /=to/. Both occur in the written standard, with their origin possibly in varied dialect distributions. In Zyrian dialects, a corresponding clitic in *=мо* /=to/ is also present, but the most important factor here is that, as explained above, while these clitics take the role of Zyrian *=сö* /=sə/ and *=мö*, they also make Permyak accusatives much less ambigious than those in Zyrian.

With the infinitive forms of Permyak verbs, a form identical to Zyrian *=мö* /=tə/ does occur (Баталова, 1975, p. 188), but the amount of ambiguity this introduces is not as problematic as what is seen in Zyrian.

Question marking in the two Komis presents a dichotomy of *=ö* /=ə/ in Komi-Zyrian and an independent particle *я* /ja/ in Komi-Permyak. Anticipation of a shallow-transfer translation system, raises the question of how these equivalent items might be designated for both languages regardless of orthographic conventions. (In Western tradition, the question is one of the four traditional sentence types, so there should be a way to address it in the code.)

6 Universal Dependencies

Work with the 2.5 release of the Komi-Permyak Universal Dependencies treebank (UD_Komi_Permyak-UH) has emphasized the need for consistency with the existing Zyrian treebanks. Since the Zyrian treebanks are relatively small, it is still easy to propose changes for both treebank sets, and future work with Zyrian also needs to be considered in the Permyak treebank.

As Permyak and Zyrian are very closely related languages, the development of different treebanks will certainly be mutually beneficial. There has been recent interest to use resources from related or contact languages in order to train tools such as dependency parsers (i.e. Lim et al., 2018), but, in the case of Komi-Zyrian, none of the languages

in the Universal Dependencies project have been particularly close to Komi, and the results have not been at so high a level that such models could have been applied in language documentation work. With Komi-Permyak and Komi-Zyrian, multilingual model training of this type may very well be worth the effort, as the grammatical structures and lexicon are largely shared. The benefits become particularly clear when attempts are made to process dialect materials in either language, as the distribution of features is in many ways different from those of the written standards (further discussion of which, unfortunately, is outside the scope of this paper).

The Komi-Permyak treebank, once again, underscores the need for a different approach to representative sentence selection. While large treebank projects are able to utilize large amounts of data with inherited but transferable annotation from other projects, small languages, such as Komi-Permyak and Komi-Zyrian, cannot really opt for statistical representation. Instead, it is proposed that features specific to the language be selected. Hence, part of the strategy for the initial release of the Komi-Permyak treebank was to feature numerals and their regular morpho-semantic use, e.g. both Komi standards have multiplicative-distributional numerals as well as ordinal-multiplicative numerals. Komi-Permyak, however, has an additional *a*-final numeral used in copula complement position to indicate the notion of a tallied sum, e.g.

(7) Деревняын олиссес нёля.

ɟerevɲa-in oliⷪːes ɲoʗ-a
village-INE dweller.PL four-A

'In the village, there are four people all together' (Permyak)

One approach could be to select example sentences from available Komi grammars, as this way it would be possible to make different grammatical phenomena fully represented. There are many features of Komi that are typologically relevant, but relatively rare, as already discussed in Partanen et al. (2018). These include, among other features, various stacked cases that occur only sporadically in all their realizations even in a very large written corpora.

7 Possibilities for resource reuse

While the morphological analyzer is still being developed for Permyak, with the groundwork for it largely copied from the existing Zyrian analyzer, special attention must be paid to the particularities of Permyak and the reduction of interference from the original Zyrian. One approach that needs further work is to ensure that both Permyak and Zyrian YAML tests are comparable in their coverage, which would also allow further automatic testing of how large the number of shared forms is. This, for example, would require the writing of YAML tests for Komi-Zyrian, which has few tests on the whole.

Permyak and Zyrian also share a extensive majority of their lexicon. This leaves the question open as to how exactly we should proceed with the management of the lexicographic data for these languages, i.e. while using tools such as Akusanat and Verdd (see i.e. Rueter and Hämäläinen, 2017; Hämäläinen and Rueter, 2019b). One also has to ask whether there are specific ways on how Permyak and Zyrian lexical resources should be connected to each other. This might be solved with cognate searching analogical to what has been used for Northern Sami and Skolt Sami cognates for establishing initial etymological associations (Hämäläinen and Rueter, 2019a). Russian loanwords, although differently adapted are largely shared. At present, this issue has been partially solved through the sharing of proper nouns mutual to nearly all languages written in Russian Cyrillics[3] (49,156 words) and additional adjectives shared by both Komi transducers [4] (~6000 words), whose content was initially introduced in FU-Lab for adjectives ending in *-ӧй* /-əj/. The shared kom-adjectives-russian-like.lexc file has preliminarily been selected on the pretext that the Komi letter *ӧ* cannot occur twice in a given Komi-Permyak stem. Further editing of this file will be required to remove Komi-Zyrian instances of *-ӧй* /-əj/ where the Russian equivalent would indicate a stressed vowel. When the Russian equivalent has a stressed *-о*, the Permyak variant is also *-о*.

It must be mentioned that through our meticulous work on Komi-Permyak analyzer, we have arrived at a situation where there are more YAML tests for Permyak than for Zyrian. It could be an interesting idea to make sure that Permyak and Zyr-

[3] gtsvn/giella-shared/urj-Cyrl/src/morphology/stems/urj-Cyrl-propernouns.lexc

[4] gtsvn/langs/kpv/src/morphology/stems/kom-adjectives-russian-like.lexc

ian tests contain the same lexemes with their matching analyses. The forms would be different, but this would allow comparing the paradigms from one more perspective. (In fact, this can be rendered rather easily by generating a separate full Zyrian YAML test for every lexeme addressed in the Permyak YAML tests, but it will also require native-like language knowledge for proof-reading. (Rueter et al., 2019b)) In addition, at least the forms that categorically do not exist in the Permyak should not be getting a reading, but the situation becomes more complicated with the forms shared by various Zyrian and Permyak dialects. (Here, we will need to use the descriptive YAML tests. As there are already three categories of YAML tests in the Giella infrastructure: dict[ionary], norm[ative] and desc[riptive]) Probably, some additional distinctions will be made between the descriptive and normative analyzers, with the first being less restricted, as has been done with Zyrian earlier. (Analogical work has been done in this vein with development of the Võro language YAML tests due to the extensively descriptive nature originally depicted in the transducer to cover various dialects (Iva and Rueter, 2020))

8 Conclusion

Based to our analysis, developing a Komi-Permyak FST based on the Komi-Zyrian FST is a worthwhile and relatively straightforward process. We also believe that there are ways to use such analyzers for better identification and quantification of the differences between these pluricentric varieties.

The approach taken in this paper, with a detailed description of the morphological differences encountered between the two norms, is believed to render a more legible work flow. Such a plan helps to formulate strategies for development and further work on the Komi-Permyak analyzer and treebanks.

One of the upcoming tasks is to extend this work from the literary languages into various dialects, as has already been done with the Zyrian analyzer. This will further complicate the relationship between work done on Permyak and Zyrian, as the feature isoglosses usually have distributions that do not follow the official language boundaries. Although smaller Komi varieties such as Zyuzdin and Yazva have some resources and recent publishing activities (for Yazva i.e. Паршакова, 2003), it is currently unclear in which forms the existing resources on these languages should be integrated into the infrastructure described here.

Our analysis is based on standard grammatical references to Komi-Permyak, so if there are features that need to be addressed further, they might be something that earlier literature has either neglected or failed to notice. Thus the development of a computational infrastructure becomes better anchored in the grammatical description of Komi-Permyak, and the relationship of these often remote, although closely connected activities, becomes more firmly established.

Acknowledgments

Jack Rueter has been able to participate in these developments while performing expertise work on Uralic languages for a FINCLARIN project at the University of Helsinki, Digital Humanities Department. Special thanks to the University of Helsinki for funding Rueter's travel.

Niko Partanen works within the project Language Documentation meets Language Technology: The Next Step in the Description of Komi, funded by the Kone Foundation, Finland. Special thanks to the University of Helsinki for funding Partanen's travel.

Larisa Ponomareva is presently working as a research assistant at the University of Helsinki, Digital Humanities with funding from the Finnish Social Insurance Institution (KELA).

References

K. Alnajjar, M. Hämäläinen, N. Partanen, and Jack Rueter. 2019. The open dictionary infrastructure for uralic languages. In *Электронная письменность народов Российской Федерации: Опыт, проблемы и перспективы. Материалы II Международной научной конференции (Уфа, 11–12 декабря 2019 г.)*, pages 49–51.

Lene Antonsen, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

David Gil. 2013. Adjectives without nouns. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen and Jack Rueter. 2019a. Finding Sami cognates with a Character-Based NMT Approach. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Mika Hämäläinen and Jack Rueter. 2019b. An open online dictionary for endangered uralic languages. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*.

Sulev Iva and Jack Rueter. 2020. rueter/aku-morph-voro: Basic adjectives, nouns and verbs.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLNG 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018. Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Jack Rueter and Mika Hämäläinen. 2017. Synchronized Mediawiki based analyzer dictionary development. In *International Workshop for Computational Linguistics of Uralic Languages*, pages 1–7. Association for Computational Linguistics.

Jack Rueter and Niko Partanen. 2019. Survey of Uralic Universal Dependencies development. In *Workshop on Universal Dependencies*, page 78. Association for Computational Linguistics.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2019a. rueter/aku-morph-komi-permyak: Basic nouns, verbs and pronouns.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2019b. rueter/aku-morph-komi-zyrian: Basic nouns, verbs and pronouns.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020. langdoc/comparative-permic-database: Comparative Permic Database.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. In *Proceedings of the 4th International Workshop on Computational Linguistics for Uralic languages*, pages 106–118. ACL.

Jack M. Rueter. 2000. Хельсинкиса университетын кыв туялысь Ижкарын перымса симпозиум вылын лыддьӧмтор. In *Пермистика 6 (Proceedings of Permistika 6 conference)*, pages 154–158.

Francis M Tyers and Mariya Sheyanova. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 66–75.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng,

Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Р. М. Баталова. 1975. *Коми-пермяцкая диалектология*. М.: Наука.

Р. М. Баталова. 1982. *Ареальные исследования по восточным финно-угорским языкам: Коми языки*. М.: Наука.

Л. М. Безносикова, Е. А. Айбабина, and Р. И. Коснырева. *Коми-роч кывчукӧр, publisher = Сыктывкар: Коми небӧг лэдзанін, year=2000*.

В. И. Лыткин. 1962. *Коми-пермяцкий язык: учебник для высших учебных заведений*. Кудымкар: Коми пермяцкое книжное издательство.

Г. Некрасова. 2000. Эмакыв. In Г. В. Федюнёва, editor, *Ӧнія коми кыв, морфология*. Россияса наукаяс академия, Коми наука шӧрин, Сыктывкар.

А. Л. Паршакова. 2003. *Коми-язьвинский букварь. Учебное издание*. Пермь: Пермское книжное издательство.

Л. Г. Пономарева. 2016. *Ойвывся коми-пермяккезлӧн сёрни*. М.: Быдкодь Отирлӧн кыввез.

Йӧлгинь Цыпанов. 2009. Перым кывъяслӧн талунъя серпас. *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne*, 258:191–206.