

Please cite as:

Rogier Blokland, Marina Fedina, Niko Partanen, Michael Rießler. 2019. "Closing the cycle of language documentation, corpus building and corpus-based description of Zyrian Komi." Paper held at the conference *Descriptive grammars and typology. The challenges of writing grammars of underdescribed and endangered languages, University of Helsinki, 27–29 March 2019*. Date: 28.03.2019. (Version of 17.04.2019). Licence: CC-BY-ND



UPPSALA
UNIVERSITET



Kotimaisten kielten keskus

INSTITUTET FÖR DE INHEMSKA SPRÅKEN
INSTITUTE FOR THE LANGUAGES OF FINLAND

Universität Bielefeld

Closing the cycle of language documentation, corpus building and corpus-based description of Zyrian Komi



KONE FOUNDATION

Rogier Blokland (Uppsala University, Sweden)

Marina Fedina (The Finno-Ugric Laboratory for Support of the Electronic
Representation of Regional Languages, Syktyvkar, Komi Republic, Russia)

Niko Partanen (Institute for the Languages of Finland, Helsinki, Finland)

Michael Rießler (Bielefeld University, Germany)

Our Komi projects

1st project ('Izva Komi: Building an Annotated Digital Corpus for Future Research on Komi Speech Communities in Northernmost Russia'; 2014–2016)

- **Fieldwork** in several locations, inside and outside of the Komi Republic
- ~45 hours **transcribed**, 100 speakers (~80 hours recorded by us)
- ~400,000 tokens aligned at utterance level

2nd project ('Language Documentation meets Language Technology: The Next Step in the Description of Komi'; 2017–)

- Transcribed corpus > to Kielipankki
- Annotated portion in **Universal Dependencies treebank**
- Use of language technology **instead of manual annotation**

Our Komi grammar project

- Descriptive grammar with a **focus on syntax**
- Will be available **online**, written under **version control**
- All glossed examples in the grammar **connected to corpora**
- Allows examining **occurrences of a specific phenomenon** in corpora
- In addition to writing a ‘traditional’ grammar, another goal of ours is to **implement the grammar** with language technology



1967

Background

Zyrian Komi, our test language



Source: Komi mu newspaper, 6.12.2018



Source: OpenStreetMap.org

- **Uralic** language, spoken by **about 160 000 speakers** in northern Russia
- Relatively vital, but **definitely endangered** due to language shift to Russian
- Existing linguistic descriptions focus on **phonology and morphology**
- Orthographic standard, in which **much material** has been published
- Interest in preservation and further development, also in creating **language technology** for teaching and to **increase its functionality in the digital age**

Background NLP and Language Technology

Computational linguistics (NLP) = the analysis/generation of natural language

Language technology = the practical application of NLP

- spell checkers
- grammar taggers
- machine translation
- etc.



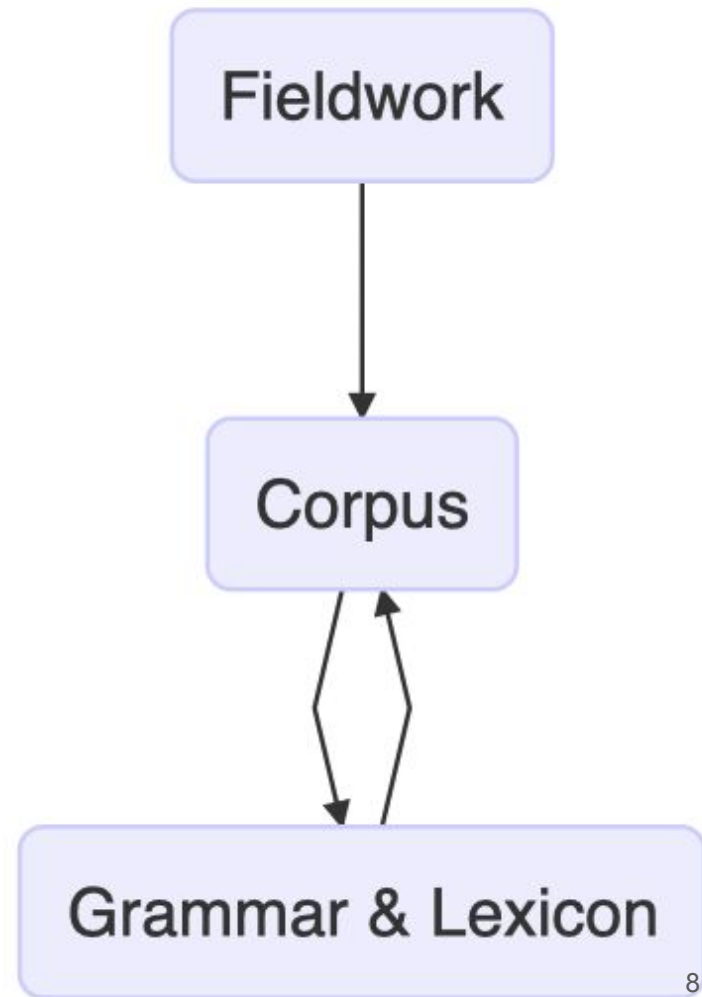
(North Saami; <http://divvun.no/korrektur/speller-demo.html>)

Workflow

- Our work relies on the NLP infrastructure available for Northern Eurasian endangered languages at **Giellatekno**. At the beginning of our workflow are **speech recordings with aligned transcriptions in ELAN**
- **Finite-State morphology (FST)** is used for rule-based modeling of stems and segmental affixes, as well as complex morphophonological rules
- Additional rules following **Constraint Grammar (CG)** are implemented for syntactic disambiguation and tagging dependency relations
- The **source code and documentation** is being developed using an SVN versioning system and is available under a GNU public license

Workflow: illustration

- **Fieldwork** (or “archive work”, digitization, etc.)
- ELAN, transcription, translation
- Writing **rules** FST, CG, maintaining lexicon
- Applying grammar into texts in ELAN
 - Manual corrections for selected texts
 - Improving the grammar



Workflow: example

ELAN 5.3 - kpv_izva20140404lgusevJA-fragment.eaf

File Edit Annotation Tier Type Search View Options Window Help

Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls

Volume: 100

kpv_izva20140404lgusevJA-fragment.wav
☐ Mute ☐ Solo

kpv_izva20140404lgusevJA.m4v
☐ Mute ☐ Solo

00:00:55.050 Selection: 00:00:55.050 - 00:00:58.020 2970

0:55.000 0:55.500 0:56.000 0:56.500 0:57.000 0:57.500

part [0]

ref@JAI-M-1939 [22]

note(ref@JAI-M-1939) [0]

orth@JAI-M-1 [21]

word@JAI-M [275]

lemma@J [321]

pos@JAI [383]

morph [597]

note(word) [0]

fl-eng@JAI-M [21]

fl-rus@JAI-M [21]

ref@NTP-M-198

kpv_izva20140404lgusevJA-b-381

Me nō cles cōsew zḡ lōbit, vīsytalō.

Me nō cles cōsew oz lōbitny, vīsytavny.

Pron CONJ Pcle ? Adv V V CLB V CLB

Pers+Sg1+N Pers+Sg1+C Neg+Ind+Prt Neg+Ind+Prt TV+ConNeg TV+Imprt+S TV+Ind+Prs+Sg3

I didn't love him at all, she says.

Говорит, я мол его совсем не любила.

FST / CG

We apply grammar-based (“symbolic”) NLP: the linguist writes a **formalized machine-readable** version of the grammar, and compiles it into a program capable of analyzing (and also generating) text input.



Finite-state transducer: modeling stems, affixes, linear morphology
(Shoebox/Toolbox/FLEEx/ELAN do this, too)

Two-Level-Morphology: modeling non-linear morphology
(Shoebox/Toolbox/FLEEx/ELAN **cannot** do this!)

Constraint Grammar: disambiguation and dependency tagging
(Shoebox/Toolbox/FLEEx/ELAN **cannot** do this!)

FST / CG: *Me чєру оѡ сѣѣ.* ‘I don't eat fish.’

Tagging (“glossing”)

сѣѣ сѣѣ+N+Sg+Nom

сѣѣ сѣѣны+V+ConNeg

сѣѣ сѣѣны+V+Imprt+Sg2

FST / CG: *Me черу оз сёй*. ‘I don't eat fish.’

Tagging (“glossing”)

сёй сёй+N+Sg+Nom

сёй сёйны+V+ConNeg

сёй сёйны+V+Imprt+Sg2

Disambiguation e.g. “IFF Rule”: ConNeg if Neg to the left

~~сёй сёй+N+Sg+Nom~~

сёй сёйны+V+ConNeg

~~сёй сёйны+V+Imprt+Sg2~~

FST / CG: *Me чери оз сёй.* 'I don't eat fish.'

Full analysis (incl. dependency structure)

ме	ме	+Pron+Pers+Sg1+Nom	@SUBJ>	#1->3
чери	чери	+N+Sg+Nom	@OBJ>	#2->4
оз	оз	+V+Neg+Ind+Prs+Sg1	@FAUX	#3->0
сёй	сёйны	+V+ConNeg	@IMV	#4->3
.	.	+CLB		#5->3



a-tree
zone=kpv

сетас
root
VERB

нигаыс
nsubj
NOUN

сӧмын
advmod
ADV

юкӧн
obj
NOUN

велӧдысьлы
obl
NOUN

.

пunct
PUNCT

Тайӧ
det
PRON

ичӧтик
amod
ADJ

ывлавыв
nmod
NOUN

"тайö" Pron Dem Sg Nom @X #1->2

"нига" N Sg Nom PxSg3 @SUBJ #2->3

"сетны" V TV Ind Fut Sg3 @X #3->0

"сөмьн" Adv @X #4->0

"ичӧтик" A Sg Nom @A< #5->6

"юкӧн" N Sg Nom @X #6->0

"ывлавыв" Der Der/ыв N Sg Nom @N< #7->8

"велöдысь" N Sg Dat @N< #8->8

"." CLB #9->9

Universal Dependencies

This page pertains to UD version 2.

ADV: adverb

Adverbs are words that typically modify [adjectives](#), [verbs](#) or other adverbs for such categories as time, place, direction or manner.

Examples

- [kpv] *бура* “well”
- [kpv] *дзык* “completely; really”
- [kpv] *уна* “much; many; a lot”

ADV in other languages: [\[bg\]](#) [\[bm\]](#) [\[ca\]](#) [\[cs\]](#) [\[da\]](#) [\[en\]](#) [\[es\]](#) [\[et\]](#) [\[eu\]](#) [\[fi\]](#) [\[fro\]](#) [\[fr\]](#) [\[ga\]](#) [\[grc\]](#) [\[hu\]](#) [\[hy\]](#) [\[it\]](#) [\[ja\]](#) [\[kk\]](#) [\[myv\]](#) [\[no-overview\]](#) [\[pcm\]](#) [\[pt\]](#) [\[ru\]](#) [\[sl\]](#) [\[sv\]](#) [\[tr\]](#) [\[uk\]](#) [\[u\]](#) [\[urj\]](#) [\[yue\]](#) [\[zh\]](#)

Treebank Statistics: UD_Komi_Zyrian-Lattice: POS Tags: ADV

There are 96 ADV lemmas (12%), 99 ADV types (9%) and 202 ADV tokens (10%). Out of 15 observed tags, the rank of ADV is: 4 in number of lemmas, 4 in number of types and 4 in number of tokens.

The 10 most frequent ADV lemmas: *нин, зэв, на, сӧмын, кыдзи, пыр, сідз, весиг, ӧдйӧ, ӧні*

The 10 most frequent ADV types: *нин, зэв, на, сӧмын, кыдзи, пыр, сідз, весиг, ӧні, кыдз*

The 10 most frequent ambiguous lemmas: *сӧмын* ([ADV](#) 7, [PART](#) 1), *пыр* ([ADV](#) 6, [ADP](#) 1), *водз* ([ADP](#) 3, [ADV](#) 3), *друг* ([ADV](#) 3, [NOUN](#) 1), *медся* ([ADV](#) 2, [PART](#) 2), *бӧрын* ([ADP](#) 1, [ADV](#) 1), *кор* ([SCONJ](#) 3, [ADV](#) 1, [NOUN](#) 1), *кутшӧм* ([ADV](#) 1, [PRON](#) 1), *кӧть* ([ADV](#) 1, [PART](#) 1, [SCONJ](#) 1), *мед* ([SCONJ](#) 2, [ADV](#) 1)

The 10 most frequent ambiguous types: *сӧмын* ([ADV](#) 7, [PART](#) 1), *друг* ([ADV](#) 2, [NOUN](#) 1), *медся* ([ADV](#) 2, [PART](#) 2), *Мыйла* ([ADV](#) 1, [SCONJ](#) 1), *бӧрын* ([ADP](#) 1, [ADV](#) 1), *кор* ([SCONJ](#) 3, [ADV](#) 1), *кӧть* ([ADV](#) 1, [SCONJ](#) 1), *мед* ([ADV](#) 1, [SCONJ](#) 1), *моз* ([ADP](#) 1, [ADV](#) 1), *став* ([DET](#) 4, [PRON](#) 3, [ADV](#) 1)

- *сӧмын*
 - [ADV](#) 7: Сэк жӧ удиті на воны *сӧмын* пипу рас весьтӧдз .
 - [PART](#) 1: Медым содтыны урожай идралӧмын ӧдъяс , колӧ не *сӧмын* сувтӧдны удж вылӧ став вундан машинаяс , но и приспособитны лобогрейкаӧ ытшкан машинаяс да используйтны найӧс тыр нагрукӧн урожай идралӧм вылӧ .
- *друг*
 - [ADV](#) 2: Мӧдлапӧлысь Присада ді вывса пожӧма яг весьтын *друг* тыдовтчис эшкын кодь кымӧр пласт , коді син водзын кутіс быдмыны .
 - [NOUN](#) 1: — Да мый тэ , *друг* , мый тэ !

Explicit references to sources, versions & licenses

The screenshot shows the ELAN 4.9.3-1 software interface. At the top, there is a menu bar with options: Grid, Text, Subtitles, Lexicon, Comments, Recognizers, **Metadata**, and Controls. Below the menu bar, there is a button labeled "Select Metadata Source..." and a text field indicating "No metadata source selected." To the right of this is a "Configure..." button.

The main window is divided into several sections. On the left, there is a video player showing a woman sitting and speaking. Below the video player, there is a timeline with a selection bar indicating the current time range: "Selection: 00:04:54.375 - 00:04:56.510 2135".

Below the video player, there is a list of transcription segments. The segments are organized into a table with columns for the segment name, the time range, and the transcription text. The segments are as follows:

Segment Name	Time Range	Transcription Text
krv_izva20150703-01-b.wav	00:04:55.000 - 00:04:56.000	Unsupported compression type for Wave file
f-ang<MSF-F-1968 [1]	00:04:55.000 - 00:04:56.000	
f-rus<MSF-F-1968 [1]	00:04:55.000 - 00:04:56.000	
ml<VPC-M-1993 [51]	00:04:57.000 - 00:04:58.000	krv_izva20150703-01-b-099
orth<VPC-M-1993 [51]	00:04:57.000 - 00:04:58.000	Кутешме сындатунгич танад айлыны из?
word<VPC-M-1993 [309]	00:04:57.000 - 00:04:58.000	Кутешме сындат танад айлыны из ?
f-ang<VPC-M-1993 [51]	00:04:57.000 - 00:04:58.000	So did you have some kind of difficulties?
f-rus<VPC-M-1993 [51]	00:04:57.000 - 00:04:58.000	Какие-то трудности у тебя были нет?
ml<IGT-F-1996 [93]	00:04:58.000 - 00:04:59.000	
orth<IGT-F-1996 [93]	00:04:58.000 - 00:04:59.000	
word<IGT-F-1996 [896]	00:04:58.000 - 00:04:59.000	
f-ang<IGT-F-1996 [93]	00:04:58.000 - 00:04:59.000	
f-rus<IGT-F-1996 [93]	00:04:58.000 - 00:04:59.000	

At the bottom of the interface, there is a large text area showing the transcription of the speech. The text is in Russian and includes the following sentences:

Кутешме сындатунгич танад айлыны из?
 Кутешме сындат танад айлыны из ?
 So did you have some kind of difficulties?
 Какие-то трудности у тебя были нет?

Below this, there is a table with two columns: the first column contains the transcription text in Russian, and the second column contains the transcription text in English. The text in the first column is:

krv_izva20150703-01-b-100
 Сындатунгич то, что нава прбста мена из пнматичны воддзык.
 Сындат то, что нава прбста мена из пнматичны воддзык.
 Difficulties with that they just didn't understand me earlier.
 Трудности в том, что они раньше просто не понимали меня.

The text in the second column is:

krv_izva20
 Ме ны пн
 Ме ны
 Я их пнчм

Source: IKDP corpus,
the Language Bank version 1.0 (coming soon)
kpv_izva20150703-01-b

Челәф школаҗасын сөмын вермаспы зев ічөтїка тōдмавын, сені оз уфїтны велавны велōдчыҗас—кыҗ еөкō колō тōдмавын воҗ. Сыхōд поҗō тōдманы сөмын ғырые школаҗасын—Унїверейтҗасын. Сїҗō школаҗасас ембе аслыс пōлөс җōртōдҗас-керкаҗас. Сїҗō җōртōдҗасас чукōртмабе быдеама пōлөс уҗалан кōлүҗҗасө. Сїҗō җōртōдҗасыс шуөҗны лабораторїаҗасө. Быдмōҗҗас, пемōҗҗас велōдны ембе бōрҗасын зорїҗҗас; җүҗас, тҗҗас-бердын аслыс пōлөс керкаҗас (лабораторїаҗас, стантсїаҗас).

Поҗō зев унатор ывлавылые тōдмавын лабораторїаҗасын, ещанїк кōлүҗōн—сы-понда ковмас тōдманы воҗдōр нїҗакōд. Нїҗа отсалас гōгōрвоны мед-ғырые ывлавыв петкōдчөмҗасө. Нїҗа индас кор, кытї ывлавылые поҗō тōдмавын. Сїҗкōн поҗō зев уна тōдмавын лабораторїаҗасын. Ставө җерт он понды тōдны. Таҗō нїҗаыс җїжōма сөщōм җōзлы. Унаөн оз вермыны всекавын некущōм лабора-торїаҗō—налы отсōҗ-пыҗҗы лōө таҗō нїҗа.

Танї абу став пїҗка, кїмїҗа, биолоҗїа. Танї ембе сөмын мед-колан торҗас. Таҗō нїҗаыс сетас сөмын ічөтїк җүкōн ывлавыв велōдылы.

Source: Fenno-Ugrica collection
<http://urn.fi/URN:NBN:fi-fe2014070132058>

Pros and cons

	Sketch grammar	Implemented grammar
Learning curve	✓ Common methodology	✗ Steep learning curve (command line tools)
Application	✗ Cannot be applied as a corpus annotation tool	✓ Can be directly applied as a corpus annotation tool
Time investment	! Describing rules and paradigms is time-consuming	! Programming grammar-based NLP is time-consuming
Audience of the grammar	! Comparative linguists, specialists in the respective language (perhaps also community members, teachers, etc.)	! Computational linguists (less likely to be community members, teachers, etc.)

Summary

- Our approach is fundamentally different **in practice** from common grammaticographical approaches in Documentary Linguistics
- It is not different **in its aims**:
 - ensuring that there are resources for this language (basic documentation)
 - providing sufficient analyses for linguistic research
 - creating something that is useful for the community (> creates preliminaries for digital infrastructure)
- In its **results**, it goes beyond common approaches and **links Documentary Linguistics closer** to established NLP methods in corpus linguistics

References

- Blokland, Rogier; Niko Partanen, Michael Rießler, & Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora. *Proceedings of ComputEL-3 2019*, 24–30.
<https://scholar.colorado.edu/scil-cmel/vol2/iss1/5>
- Gerstenberger, Ciprian; Niko Partanen, Michael Rießler, & Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 29–47.
<http://www.nejlt.ep.liu.se/2016/v4/a03>
- Beesley, Kenneth & Lauri Karttunen. 2003. *Finite State Morphology*. CSLI.
- Fedina, Marina & Olga Kuzivanova. 2013. Le statut officiel et social du komi sur le territoire de la République de Komi. Histoire et situation actuelle. *Études finno-ougriennes* 45, 1–13.
- Karlsson, Fred; Arto Voutilainen, Juha Heikkilä, & Arto Anttila (eds). 1995. *Constraint Grammar: A language-independent system for parsing unrestricted text*. Mouton de Gruyter.
- Partanen, Niko; Rogier Blokland, KyungTae Lim, Thierry Poibeau, & Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. *Proceedings of UDW 2018*, 126–132. <https://aclanthology.info/papers/W18-6015/w18-6015>
- Partanen, Niko; KyungTae Lim, Michael Rießler, & Thierry Poibeau. 2018. Dependency parsing of code-switching data with cross-lingual feature representations. *Proceedings of IWCLUL 2018*, 1–17.
<https://aclweb.org/anthology/papers/W/W18/W18-0201>
- Trosterud, Trond. 2006. Grammatically based language technology for minority languages. *Lesser-known languages of South Asia*, ed. by Anju Saxena et al. Mouton de Gruyter, 293–316.



Аттьö! Kiitos! Thank you!

