Instant annotations in ELAN corpora of spoken and written Komi-Zyrian, an endangered language of the Barents Sea region (Russia)

^aGiellatekno – Saami Language Technology, UiT The Arctic University of Norway ^bDepartment of Uralic Studies, Uni Hamburg

Background

- ► ELAN¹ is a widespread GUI tool for ► transcribing and translating field recordings ▶ creating further annotations aligned to audio and video ► searching and analysing the resulting corpus data
- ► Morphosyntactic annotations are typically done ► manually in ELAN, or
- ▶ semi-manually in interaction with other tools
- ► NLP tools for low-resourced **written** languages exist, but they are rarely applied in **spoken** language documentation projects.

Promising alternative approach

- ► adapt NLP tools to small spoken languages
- ► avoid ineffective manual work
- ► create larger and deeper annotated corpora

ELAN-FST/CG Integration

Automated workflow for generating morphosyntactic analysis in ELAN

Using available Giellatekno² tools:

- ► Finite-State-Transducer (FST) for ► morphological analysis
- ► Constraint Grammar (CG) for
- ► dismbiguation and syntactic analysis

Test case with structurally uniform data from Komi-Zyrian (**160,000** speakers); full syntactic analysis and dependency tagging is in the works.³

Ongoing projects on smaller languages from the Barents region: ⊳Kildin Saami (500 speakers) ⊳Skolt Saami (500 speakers) \triangleright Pite Saami (30 speakers)⁴

Guiding principles

orthographic transcription system ► **computerized** annotation methods Ciprian Gerstenberger^a, Niko Partanen^b, Michael Rießler^c

Size	e of the dif	ferent spoke	en and writ	tten corpor	a for Kon	ni-Zyrian	
_anguage		Modality	Record	Recorded spea-		span	Tokens in
<komi<permic<finno-ugric<uralic)< td=""><td></td><td>kei</td><td>rs/writers</td><td>of</td><td>texts</td><td>corpus</td></komi<permic<finno-ugric<uralic)<>			kei	rs/writers	of	texts	corpus
Komi-Zyrian (Standard) 5		written		~2,500	1920-	2017	30,000,000
(omi-Zyrian (Izhva dialect)		t) spoken	~150		1844–2016		200,000
Comi-Zyrian (Udora dialect)		ct) spoken		~50		2013	40,000
Image: Note of the second s	0.050 Set Pr Pr Pr Pr 01:02:43.000 vr izva20150402-7-b.1326 vr izva20150402-7-b.1326 кан нёль баба чомъянум пыр опіса our chum four women lived permant нас в чуме постоянно четыре жени	Sele ▼ Proj ● O1:02:43.50 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000 • 01:02:43.000	о1:02:43.500 	I part- morphosy	OUTPUT: word tier emma tier of-speech t ntactic ana	ier lysis tier	01:02:45.000
utterance	 word@DAK-F lemma@DA pos@DAK morph@ ft-eng@DAK-F ft-rus@DAK-F 	1929 К-F-1929 -F-1929 DAK-F-1929 Pron Pers+PI1+GenStem In our chum four won 1929 У нас в чуме постоя	нёль Num Card+Abs Card+ nen lived permanently. нно четыре женщины жили.	баба V N N Der/MW Ind+ Ind+ Sg+Nom Sg+	чомъянум чом чомъя N A Sg+ Sg+In Sg+III+ Der/M	NUP NU Adv Po V WN+N+ Co	рны овны V nN Imprt+, Ind+Prt1+Sg3
		†	DISAMBI	GUATION	1	♥/	
ref@DAK-F-1929	01:02:43.000 kpv_izva20150402-7-b.1326	01:02:43.500	01:02:44.00	01:02:	44.500	01:02:45.000	01:02:4
orth@DAK-F-1929	Миян нёль баба чомъяным	пыр оліс.					-
word@DAK-F-1929	Миян	Нель	6868	чомъянум	пыр	оліс	
lemma@DAK-F-1929	ми	нель	бабны баба	YOM	пыр	ОВНЫ	
pos@DAK-F-1929	Pron Pore+Did+ConStem	Num :	V N IndaEu Inda Dr. Snahlom	N Saulaau DyDid Saulliu Dyf	Adv Po	V Ind + Dett + Se2	CLB
morph@DAK-F-1929	Pers+Pi1+GenStem	CarotAbs	Ind+Pu Ind+Pr Sg+Ivom	Sg+ine+PXPIT Sg+iii+PXP		Ind+Prt1+5g3	
ft-eng@DAK-F-1929	V use a use a portroquile use menunu use menu						
ft-rus@DAK-F-1929	у нас в чуме постоянно четыре женщины жили.						
IDEAL A	ми Pron Pers+PI1+GenStem In our chum four women live У нас в чуме постоянно чет NNOTATION	нёль Num Card+Abs d permanently. ыре женщины жили.	бабны баба V N Ind+Fu Ind+Pr Sg+Nom		Adv Po	овны V Ind+Prt1+Sg3	-

POSS-

-QUANT

Num

Pron

Миян нёль

Processing pipeline						
1. utterance extraction:	Python script					
2. tokenization:	Perl script					
3. morphosyntactic analysis:	FST					
4. disambiguation:	CG					
5. ELAN tier building:	Python script					





Advantages of **rule-based** morphosyntactic modelling for **endangered languages**: ► precise results of **automatic tagging** ▶ simultaneous creation of both a **tool** and a morphosyntactic description ► deployable for new (I)CALL technology for language revitalisation purposes

The Giellatekno **open-access infrastructure** includes dictionaries and rule-based grammars for several circumpolar (written) languages. It can be used for new (spoken/written) language projects easily.

Our approach challenges current manual practices in endangered language documentation projects.

^cFreiburg Institute for Advanced Studies, Uni Freiburg

Prospects

References

1http://www.mpi.nl/corpus/html/elan/ 2http://giellatekno.uit.no 3R. Blokland, N. Partanen, M. Rießler: "Language documentation meets language technology" Ongoing project funded by Kone Foundation (2017-2020)4 C. Gerstenberger, N. Partanen, M. Rießler, J. Wilbur (2017, in press): Utilizing language technology in the documentation of endangered Uralic languages Northern European Journal of Language Technology (NEJLT) 5http://komicorpora.ru