Digitizing Pite Saami
Making the most of limited resources

Joshua Wilbur • Lecturer in Digital Linguistics • Institute of Estonian and General Linguistics

Digitizing Pite Saami
Making the most of limited resources

moving into digital domains

- Pite Saami: background
- texts and other resources
- NLP for Pite Saami
  (how is this even possible?)
- challenges and prospects

Arjeplog / Árjepluovve
(entering town from the west)

```
Arjeplog
Arjeplog        Arjeplog        +?

Árjepluovve
Árjepluovve     árrje+N+Cmp/SgGen+Use/Circ+Cmp#pluovve+N+Sg+Nom
Árjepluovve     Árjepluovve+N+Sg+Nom
```

linguistic analyses (using FST)

Pite Saami language

ISO 639-3 code: sje
Glottocode: pite1240

- Uralic➔Finno-Ugric➔Saami…Pite Saami
- spoken by ~**30** individuals from Arjeplog/Árjepluovve in Swedish Lapland
- almost all speakers are at least 50 years old
- hardly taught to younger generations
- Swedish dominates in everyday life
- all speakers are bilingual (Pite Saami and Swedish/*arjeplogsmål*)
- *official* orthography since 2019; further standardization on-going
- practically no media; a few children's books

Pite Saami language

ISO 639-3 code: sje
Glottocode: pite1240

- Uralic➔Finno-Ugric➔Saami…Pite Saami
- spoken by ~**30** individuals from Arjeplog/Árjepluovve in Swedish Lapland
- **almost all speakers are at least 50 years old**
- **hardly taught to younger generations**
- **Swedish dominates in everyday life**
- all speakers are bilingual (Pite Saami and Swedish/*arjeplogsmål*)
- *official* orthography since 2019; further standardization on-going
- practically no media; a few children's books

'critically endangered'

Pite Saami language

morphological structure:

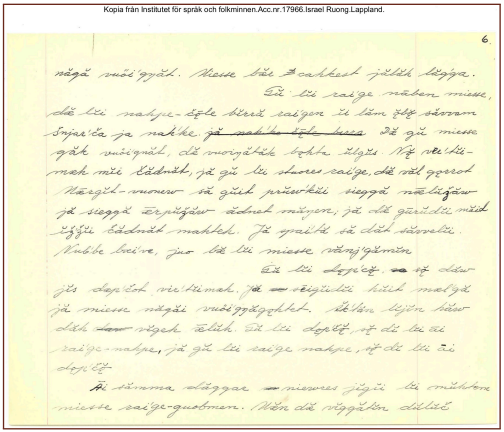
- mainly agglutinative
- complex but *systematic*
- extensive stem alternations due to consonant gradation, umlaut, allomorphy and metaphony

morphological structure of nouns:

stem + class-marker + case/number

juällge	juällg	e	-	‘leg/foot’
juolgev	juolg	e	v	‘leg/foot’ (ACC.SG)
juallgáj	juallg	á	j	‘leg/foot’ (ILL.SG)
julgij	julg	i	j	‘leg/foot’ (GEN.PL)
bägga	bägg	a	-	‘wind’
bieggav	biegg	a	v	‘wind’ (ACC.SG)
båtsoj	båts	o	j	‘reindeer’
buhtsuv	buhts	u	v	‘reindeer’ (ACC.SG)
vanas	vanas	-	-	‘boat’
vadnasijt	vadnas	i	jt	‘boat’ (ACC.PL)

Pite Saami heritage materials



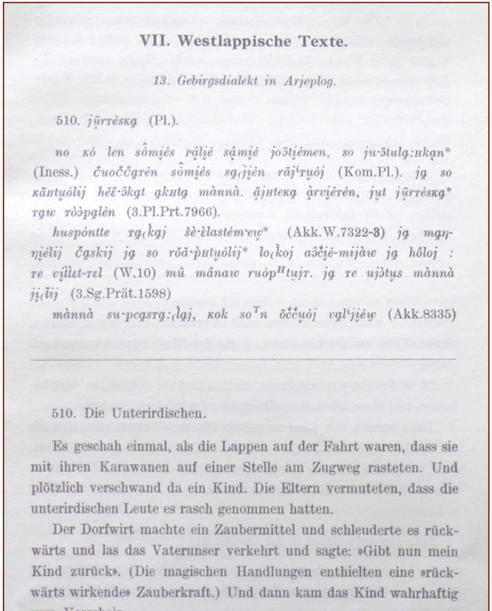
text by I. Ruong at ISOF

archived at the Swedish Institute for Language and Folklore (ISOF) in Uppsala:

- Israel Ruong left a large Pite Saami text collection, lexical items, paradigms, recordings
- smaller text collections and recordings by others are archived there, too

➜ *handwritten, mostly undigitized*

Pite Saami heritage materials



VII. Westlappische Texte.

13. Gebirgsdialekt in Arjeplog.

510. *jūrrėsкą* (Pl.).

*no кȯ len sômįės pą̊lįė så̊mįė joǯtįėmen, so ju·ǯtulą:ʜkąn** (Iness.) *čuoččąrėn sômįės są(jįėn rāj'rųȯj* (Kom.Pl.). *ją so кāʜtųȯlij hēž·ǯkąt ąкʜtą mànnà. ājʜteкą ąrvįėrėn, jųt jūrrėsкą* rąw ròǯpąlėn* (3.Pl.Prt.7966).

*huspȯntte rą(кąj šè·èlastėm·ew** (Akk.W.7322-3) *ją mąŋŋįėlij čąskij ją so rŏă·ᵽʜtųȯlij* lo(кoj aȏčįė-mijàw ją hȏloj : re vįllet-rel* (W.10) *mû mânaw ruȯpᴴtujr. ją re ujǯtųs mànnà ji(īij* (3.Sg.Prät.1598)

mànnà su·pcąsrą:(ląj, кok soᵀn ŏččųȯj vąl'jįėw (Akk.8335)

510. Die Unterirdischen.

Es geschah einmal, als die Lappen auf der Fahrt waren, dass sie mit ihren Karawanen auf einer Stelle am Zugweg rasteten. Und plötzlich verschwand da ein Kind. Die Eltern vermuteten, dass die unterirdischen Leute es rasch genommen hatten.

Der Dorfwirt machte ein Zaubermittel und schleuderte es rückwärts und las das Vaterunser verkehrt und sagte: »Gibt nun mein Kind zurück». (Die magischen Handlungen enthielten eine »rückwärts wirkende» Zauberkraft.) Und dann kam das Kind wahrhaftig

other texts:

- transcribed text collections by academics (I. Halász, E. Lagercrantz, J.-K. Qvigstad, etc.)
- several published texts in books and magazines (mainly by L. Rensund)

➜ *printed (often in FUT)*

text by M. Johansson
transcribed by E. Lagercrantz
in 1921

The Pite Saami Documentation Project

funded by *ELDP* (2008-2015)
digital documentation archived at ELAR (Berlin) and TLA (Nijmegen)

The Pite Saami Syntax Project

Syntactic Patterns in Pite Saami:
*A corpus-based exploration of 130 years of variation and change**

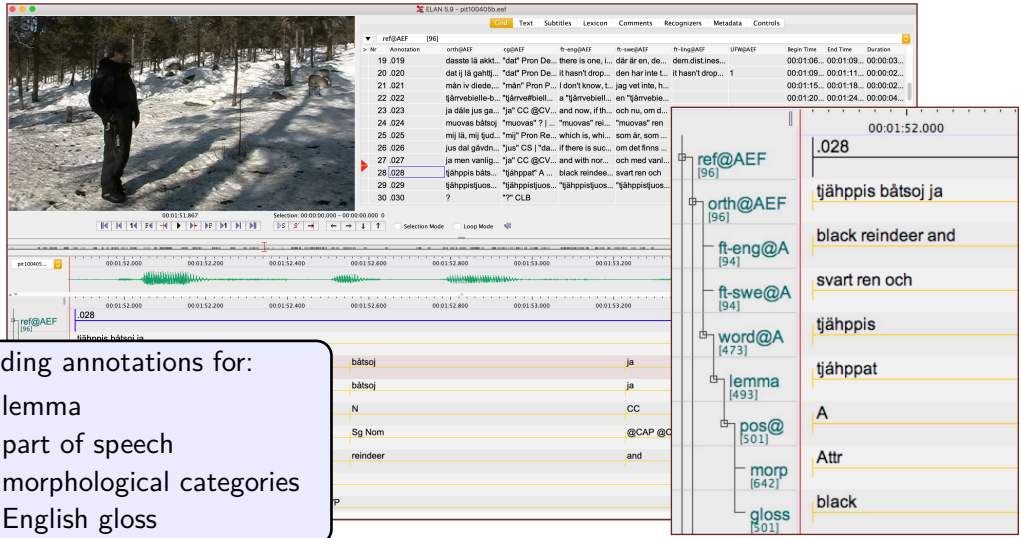
Goals

- Create a digital corpus with spoken-language texts spanning more than 100 years
  - ➜ about 60,000 tokens
  - ➜ automatic annotations for lemma, part of speech, morphology and English glosses (*in partial collaboration with Giellatekno*)
  - ➜ digital corpus available via ELAR and TLA
- corpus-based descriptions of syntactic structures

funded by *Deutsche Forschungsgemeinschaft*, 2016-2019

my Pite Saami corpus

in ***ELAN***; based on orthographic transcriptions; annotation files are XML



including annotations for:

- lemma
- part of speech
- morphological categories
- English gloss

summary of available texts

texts	*quantity*	*notes*
ISOF archive	thousands of pages, cards, etc.	mostly analogue and hand-written
other random texts	a few dozen	heritage texts in various orthographies; new texts in modern orthography
PSDP	~60 000 tokens	various degrees of annotation, orthography, genres

linguistics research about Pite Saami

Halász 1893 (Hungarian)

Lagercrantz 1926 (German)

Ruong 1943 (German)

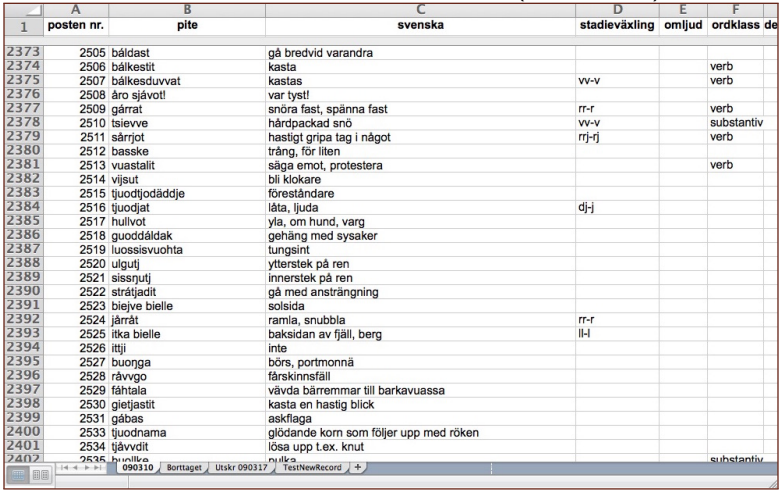
Lehtiranta 1992 (Finnish)

Wilbur 2014 (English)

Sjaggo 2015 (Swedish)

Pite Saami community

- language activists* ➜ wordlist (2008-2012)



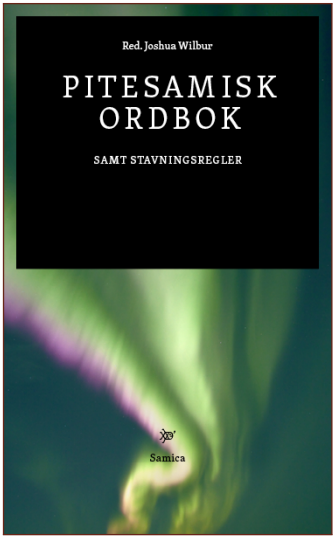
	A	B	C	D	E	F	
1	posten nr.	pite	svenska	stadieväxling	omljud	ordklass	de
2373	2505	båldast	gå bredvid varandra				
2374	2506	bålkestit	kasta			verb	
2375	2507	bålkesduvvat	kastas	vv-v		verb	
2376	2508	åro sjávot!	var tyst!				
2377	2509	gårrat	snöra fast, spänna fast	rr-r		verb	
2378	2510	tsievve	hårdpackad snö	vv-v		substantiv	
2379	2511	sårrjot	hastigt gripa tag i något	rrj-rj		verb	
2380	2512	basske	trång, för liten				
2381	2513	vuastalit	säga emot, protestera			verb	
2382	2514	vijsut	bli klokare				
2383	2515	tjuodtjodäddje	föreståndare				
2384	2516	tjuodjat	låta, ljuda	dj-j			
2385	2517	hullvot	yla, om hund, varg				
2386	2518	guoddåldak	gehäng med sysaker				
2387	2519	luossisvuohta	tungsint				
2388	2520	ulgutj	ytterstek på ren				
2389	2521	sissŋutj	innerstek på ren				
2390	2522	strátjadit	gå med ansträngning				
2391	2523	biejve bielle	solsida				
2392	2524	jårrát	ramla, snubbla	rr-r			
2393	2525	itka bielle	baksidan av fjäll, berg	ll-l			
2394	2526	ittji	inte				
2395	2527	buoŋga	börs, portmonnä				
2396	2528	råvvgo	fårskinnsfäll				
2397	2529	fåhtala	vävda bärremmar till barkavuassa				
2398	2530	gietjastit	kasta en hastig blick				
2399	2531	gåbas	askflaga				
2400	2533	tjuodnama	glödande korn som följer upp med röken				
2401	2534	tjåvvdit	lösa upp t.ex. knut				
2402	2535	hvalka	sulka			substantiv	

090310 | Borttaget | Utskr 090317 | TestNewRecord | +

‘Insamling av pitesamiska ord’

*N-H. Bengtsson, M. Eriksson, I. Fjällås, E-K. Rosenberg, G. Sivertsen, V. Sjaggo, P. Steggo & D. Skaile

The Pite Saami Lexicography Project



digital lexicographic database

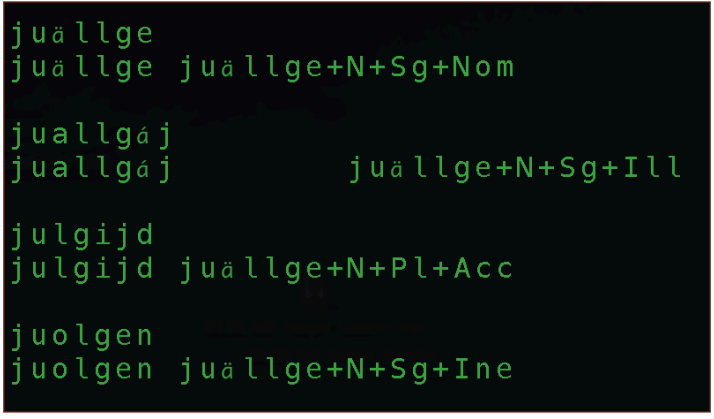
sjelex.keeleressursid.ee

originally funded by the Norwegian *Sametinget* and *Duaddara Ráffe* (2016)

NLP for Pite Saami

in collaboration with Giellatekno Center for Saami Language Technology

- Finite State Transducer (FST) for morphological parsing



github.com/giellalt/lang-sje/

NLP for Pite Saami

in collaboration with Giellatekno Center for Saami Language Technology

- Finite State Transducer (FST) for morphological parsing
- Constraint Grammar (CG) for syntactic disambiguation

```
"<men>"
  "men" CC @CVP MAP:580:CCasCNPCVPCAP
"<idtjin>"
  "ij" V Neg Prt Pl3
"<del>"
  "del" Adv
"<bårå>"
  "bårråt" V ConNeg SELECT:313:ConNeg3
;"bårråt" V Imprt Sg2 SELECT:313:ConNeg3
;"bårråt" V Ind Prs Sg2 SELECT:313:ConNeg3
"<dan>"
  "dat" Pron Dem Sg Gen SELECT:378:genB4Po
;"dat" Det Sg Gen SELECT:378:genB4Po REMOVE:414:NoDetWO-NPhead
;"dat" Det Sg Ill SELECT:378:genB4Po
;"dat" Det Sg Ine SELECT:378:genB4Po
;"dat" Pron Dem Sg Ine SELECT:378:genB4Po
"<sisste>"
  "sisste" Po
```

github.com/giellalt/lang-sje/

my Pite Saami corpus

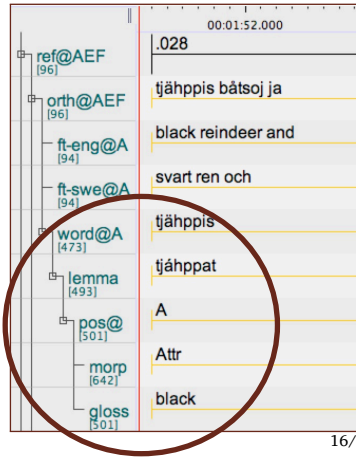
automatic corpus annotation*

using a script that:

1. tokenizes the orthographic representation
2. sends each token through FST
3. removes ambiguities using CG
4. adds an English gloss
5. inserts this output into ELAN

benefits:

- saves time
- avoids inconsistencies
- can be updated automatically



*see Blokland et al. (2015), Gerstenberger et al. (2017)

summary of available ~~texts~~ resources

resource	*quantity*	*notes*
ISOF archive	thousands of pages, cards, etc.	*mostly analogue and hand-written*
other random texts	a few dozen	*heritage texts in various orthographies; new texts in modern orthography*
PSDP	∼60 000 tokens	*various degrees of annotation, orthography, genres*
grammatical descriptions	6	*in Hungarian, German, Finnish, Swedish, English*
digital lexical database	∼7 700 entries (∼6 100 lemmas)	*regularly updated*
NLP (FST+CG)	–	*CG rather preliminary*

summary of available ~~texts~~ resources

resource	*quantity*	*notes*
ISOF archive	thousands of pages, cards, etc.	*mostly analogue and hand-written*
other random texts	a few dozen	*heritage texts in various orthographies; new texts in modern orthography*
PSDP		*n,*
grammatical descriptions		*in Hungarian, German, Finnish, Swedish, English*
digital lexical database	∼7 700 entries (∼6 100 lemmas)	*regularly updated*
NLP (FST+CG)	–	*CG rather preliminary*

an impressive amount of resources for such a small, critically endangered language

enabling NLP for Pite Saami

factors:

- all those resources (texts and others)
- relatively recent increases in:
  - state support for regional languages and dialects, especially in a European/Scandinavian context
  - private support for endangered languages
- NLP infrastructure already in development for closely related languages (i.e., Giellatekno)
- concurrent technical advances (NLP) and relevant research...

enabling NLP for Pite Saami

factors:

- all those resources (texts and others)
- relatively recent increases in:
  - state support for regional languages and dialects, especially in a European/Scandinavian context
  - private support for endangered languages
- NLP infrastructure already in development for closely related languages (i.e., Giellatekno)
- concurrent technical advances (NLP) and relevant research...
- more than a century of `engaged and motivated humans`:

in place	***needed***
native speakers • language learners • linguists	*language technologists*

enabling NLP for Pite Saami

factors:

- all those resources (texts and others)
- relatively recent increases in:
  - state support for regional languages and dialects, especially in a European/Scandinavian context
  - p
- NLP i
  (i.e., Giellatekno)
- concurrent technical advances (NLP) and relevant research...
- more than a century of `engaged and motivated humans`:

in place	***needed***
native speakers • language learners • linguists	*language technologists*

how much are *luck* and *coincidence* actual factors?

NLP and endangered languages

research on NLP methodologies to support documentary linguistics and under-resourced languages is not new, e.g.:

- Gerstenberger et al. (2017). "Instant annotations: Applying NLP methods to the annotation of spoken language documentation corpora"
- Gessler (2022). "Closing the NLP Gap: Documentary Linguistics and NLP Need a Shared Software Infrastructure"
- Ginn et al. (2024). "GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text"
- Moeller (2021). "Integrating machine learning into language documentation and description"
- Moeller & Hulden (2018). "Automatic Glossing in a Low-Resource Setting for Language Documentation"
- Moeller et al. (2018). "A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer"

but research on how LLMs can support this is only just beginning:

- Tanzer et al. (2024). "A Benchmark for Learning to Translate a New Language from One Grammar Book"

NLP and indigenous communities

➔ ***CARE data principles for working with indigenous data***

C collective benefit

- re/using data supports indigenous peoples, reflects community values

A authority to control

- indigenous nations should be *actively* involved in determining usage

R responsibility

- non-indigenous institutions must ensure the use of data supports the indigenous group(s)

E ethics

- indigenous ethics should inform the use of data across time

based on Carroll, S. R., et al. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous data futures"

outlook

challenges:

- making language technology **accessible** and **useful** for the community
- making language technology **valuable** (beyond being a nice symbolic gesture)
- accessing and incorporating **non-linguistic knowledge**
- implementing **C.A.R.E. principles**

outlook

challenges:

- making language technology **accessible** and **useful** for the community
- making language technology **valuable** (beyond being a nice symbolic gesture)
- accessing and incorporating **non-linguistic knowledge**
- implementing **C.A.R.E. principles**

prospects:

- Pite Saami presents a great opportunity for testing LLM development for under-resources languages: multiple modes of resources (texts, media, lexicons, linguistics research, extant NLP) for feeding into the LLM loop, aimed at supporting both research and the language community

an opportunity for other endangered, under-resourced languages, too?

Thanks!

joshua.wilbur@ut.ee