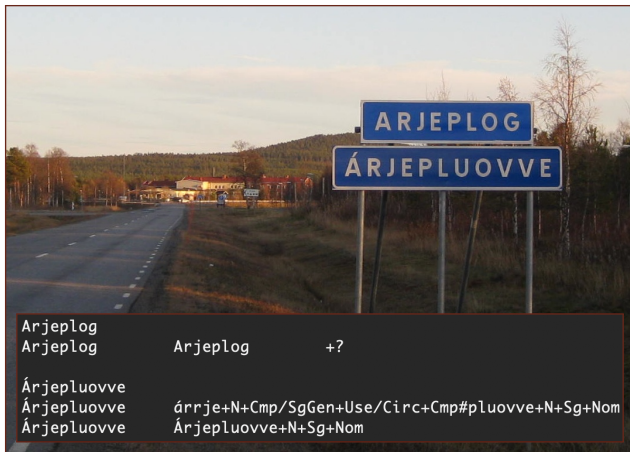


Digitizing Pite Saami

Making the most of limited resources



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	árnje+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:

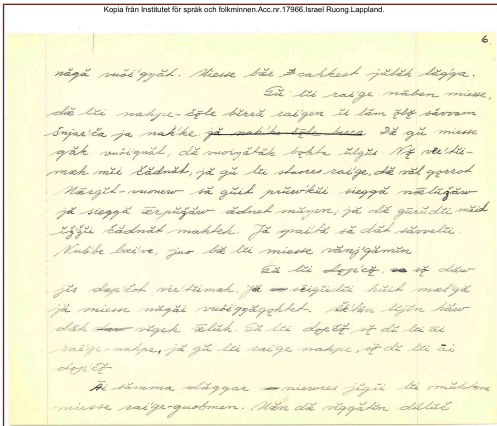
- mainly agglutinative
- complex but *systematic*
- extensive stem alternations due to consonant gradation, umlaut, allomorphy and metaphony

morphological structure of nouns:
stem + class-marker + case/number

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

Pite Saami heritage materials

Kopia från Institutet för språk och folkminnen. Acc.nr.17966.Israel Ruong Lappland.



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. *jūrreškq* (Pl.).

*no kó len sômjés vâljê sâmjê jo5tjêmen, so ju5tulq:nkqn**
(Iness.) *ĉuočĉgrên sômjés sq.jjên râj'ruój* (Kom.Pl.). *jq so*
*kântuólij hêĉ-škqt gkvtq mánna. âjntekq ârvjêrên, jyt jūrreškq**
rgw rôðrglên (3.Pl.Prt.7966).

*huspónte rg.kgj šê-èlastém'vew** (Akk.W.7322-3) *jq mqr-*
ñjêlij ĉqskij jq so rōā-pñtuólij lo.koj adĉjê-mijàw jq hōloj :*
te vjllst-rel (W.10) *mú mánav ruóp^Htujr. jq te ujðtus mánna*
jj.ii (3.Sg.Prät.1598)

mánna su-pcgrsq:lj, kok so^Tn ôĉ'ujó vgl'jjeŵ (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

my Pite Saami corpus

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

The screenshot displays the ELAN 5.9 interface. On the left, a video window shows a person in a snowy forest. The main window contains a list of annotations with columns for time, orthographic transcription, and English gloss. A detailed view on the right shows the hierarchical structure of an annotation, including layers for reference, orthography, morphology, and gloss.

Nr	Annotation	orth@AEF	cp@AEF	ft-eng@AEF	ft-swe@AEF	ft-Ing@AEF	UFW@AEF	Begin Time	End Time	Duration
19	.019	dasse lä akkt...	"dat" Pron De...	there is one, l...	där är en, de...	dem.dist.ines...		00:01:06...	00:01:09...	00:00:03...
20	.020	dat ij lä gahtj...	"dat" Pron De...	it hasn't drop...	den har inte l...	it hasn't drop...	1	00:01:09...	00:01:11...	00:00:02...
21	.021	mán iv diede...	"mán" Pron P...	I don't know, t...	jag vet inte, h...			00:01:15...	00:01:18...	00:00:02...
22	.022	tjårvebielle-b...	"járvebiell...	a "járvebiell...	en "járvebie...			00:01:20...	00:01:24...	00:00:04...
23	.023	ja dále jus ga...	"ja" CC @CV...	and now, if th...	och nu, om d...					
24	.024	muovas bátsoj	"muovas" ?	"muovas" rei...	"muovas" ren					
25	.025	mij lá, mij tju...	"mij" Pron Re...	which is, whi...	som är, som					
26	.026	jus dal gávdn...	"jus" CS	"da... if there is suc...	om det finns					
27	.027	ja men vanlig...	"ja" CC @CV...	and with nor...	och med vanl...					
28	.028	tjáhppis báts...	"tjáhppat" A ...	black reindee...	svart ren och					
29	.029	tjáhppisjuos...	"tjáhppisjuos...	"tjáhppisjuos...	"tjáhppisjuos...					
30	.030	?	*?" CLB							

including annotations for:

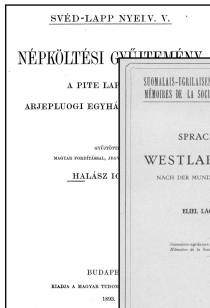
- lemma
- part of speech
- morphological categories
- English gloss

ref@AEF [96]	.028
orth@AEF [96]	tjáhppis bátsoj ja
ft-eng@A [94]	black reindeer and
ft-swe@A [94]	svart ren och
word@A [473]	tjáhppis
lemma [493]	tjáhppat
pos@ [501]	A
morp [642]	Attr
gloss [501]	black

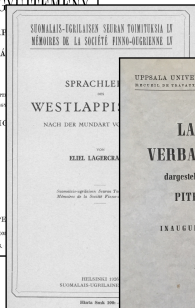
summary of available texts

<i>texts</i>	<i>quantity</i>	<i>notes</i>
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

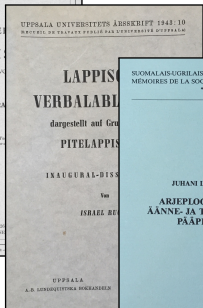
linguistics research about Pite Saami



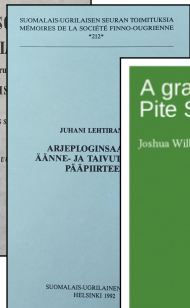
Halász 1893 (Hungarian)



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	báldast	gå bredvid varandra			
2374	2506	bálkestít	kasta			verb
2375	2507	bálkesduvvat	kastas	vv-v		verb
2376	2508	áro sjávot!	var tyst!			
2377	2509	gárrat	snöra fast, spänna fast	rr-r		verb
2378	2510	tsievve	hårdpackad snö	vv-v		substantiv
2379	2511	sárrjot	hastigt gripa tag i något	rr-j		verb
2380	2512	basske	trång, för liten			
2381	2513	vuastalit	säga emot, protestera			verb
2382	2514	vijsut	bli klokare			
2383	2515	tjuodtjodáddje	föreståndare			
2384	2516	tjuodjat	låta, ljuda	dj-j		
2385	2517	hullvot	yla, om hund, varg			
2386	2518	guoddáldak	gehäng med sysaker			
2387	2519	luossisvuohtha	tungsint			
2388	2520	ulgutj	ytterstek på ren			
2389	2521	sisnjutj	innerstek på ren			
2390	2522	strátjadit	gå med ansträngning			
2391	2523	biejve bielle	solsida			
2392	2524	járrát	ramla, snubbla	rr-r		
2393	2525	itka bielle	baksidan av fjäll, berg	ll-l		
2394	2526	ittji	inte			
2395	2527	buonga	börs, portmonnä			
2396	2528	rávvgó	fårskinnsfäll			
2397	2529	fáhtala	vävda bärremmar till barkavuassa			
2398	2530	gietjastit	kasta en hastig blick			
2399	2531	gábás	askflaga			
2400	2533	tjuodnama	glödande korn som följer upp med röken			
2401	2534	tjávvdit	lösa upp t.ex. knut			
2402	2535	hunnika	nulka			substantiv

'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

Bidumsáme Báhkogirrije
Pitesamisk ordlista • Pite Saami lexical database

Sök:

pitesamiska:	ordklass:	svenska:	engelska:
%bâ%	ordklass	svenska	English
stadväxling:	omjjud:	stavelserantal:	stamförlängning:
stadväxling	omjjud	stavelserantal	stamförlängning

sök återstäl enkel sökning regex info om sökertecken

Resultaten: (171 träffar; klicka på ord för detaljer)

- **ahtselisbäläs** (ll-l, -s-) *SUBST* *sv.* häftig regnskur *eng.* heavy rain shower
- **arrambåtsoj** (hts-ts, -buhtsu-) *SUBST* *sv.* fet ren vid gott hull *eng.* fat reindeer in good condition
- **árbálmáj** (arbálmá-) *SUBST* *sv.* änkeman *eng.* widower
- **áttjábájjgá** (jjg-jg) *SUBST* *sv.* röksvamp *eng.* puffball
- **bebbmusbåtsoj** (hts-ts, -buhtsu-) *SUBST* *sv.* matren *eng.* reindeer intended as food
- **buoremus báddá** *FRAS* *sv.* bästa stunden *eng.* best time
- **burist báhtem** *FRAS* *sv.* välkommen *eng.* welcome
- **bádá** *VERB* *sv.* du kommer *eng.* you come [báhtet:2SG.PRS]
- **bádádallat** (ll-l) *VERB* *sv.* bli överraskad, bli ertappad *eng.* be surprised
- **bádáv** *VERB* *sv.* jag kommer *eng.* I come [báhtet:1SG.PRS]
- **báddnáj** *SUBST* *sv.* till botten *eng.* towards the bottom [báddne:ILL.SG]
- **báddne** (ddn-dn) *SUBST* *sv.* botten *eng.* bottom
- **báddnje** (ddnj-dnj) *SUBST* *sv.* make *eng.* husband
- **báddá** *SUBST* *sv.* tid, stund *eng.* time, while
- **báddátj** (-tj-) *SUBST* *sv.* liten stund *eng.* a little while
- **báde lagabij** *FRAS* *sv.* kom närmare *eng.* come closer
- **báde muv bákkto** *FRAS* *sv.* kom om mig, kom förbi *eng.* stop by
- **báde sisa** *FRAS* *sv.* kom in *eng.* come in
- **bádná** *VERB* *sv.* han, hon tvinnar sentråd *eng.* he/she twists tendon into string [bádnat:3SG.PRS]

version från: 2023-08-30 kl. 23:00
Om denna websida

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

```
juállge
juállge juállge+N+Sg+Nom

juallgáj
juallgáj          juállge+N+Sg+Ill

julgijd
julgijd juállge+N+Pl+Acc

juolgen
juolgen juállge+N+Sg+Ine
```

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

```
1 "<men>"
2   "men" CC @CVP MAP:580:CCasCNPCVPCAP
3 "<idtjin>"
4   "ij" V Neg Prt Pl3
5 "<del>"
6   "del" Adv
7 "<bårå>"
8   "bårå" V ConNeg SELECT:313:ConNeg3
9 ;>"bårå" V Imprt Sg2 SELECT:313:ConNeg3
10 ;>"bårå" V Ind Prs Sg2 SELECT:313:ConNeg3
11 "<dan>"
12   "dat" Pron Dem Sg Gen SELECT:378:genB4Po
13 ;>"dat" Det Sg Gen SELECT:378:genB4Po REMOVE:414:NoDetW0-NPhead
14 ;>"dat" Det Sg Ill SELECT:378:genB4Po
15 ;>"dat" Det Sg Ine SELECT:378:genB4Po
16 ;>"dat" Pron Dem Sg Ine SELECT:378:genB4Po
17 "<sisste>"
18   "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

	00:01:52.000
ref@AEF [96]	.028
orth@AEF [96]	tjähppis båtsoj ja
ft-eng@A [94]	black reindeer and
ft-swe@A [94]	svart ren och
word@A [473]	tjähppis
lemma [493]	tjähppat
pos@ [501]	A
morp [642]	Attr
gloss [501]	black

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

summary of available ~~texts~~ resources

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

summary of available ~~texts~~ resources

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

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

native speakers • language learners • linguists

needed

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
 - Pite Saami
- NLP i (i.e., Giellatekno)
- concurrent technical advances (NLP) and relevant research...

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

- more than a century of engaged and motivated humans:

in place

native speakers • language learners • linguists

needed

language technologists

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

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`

