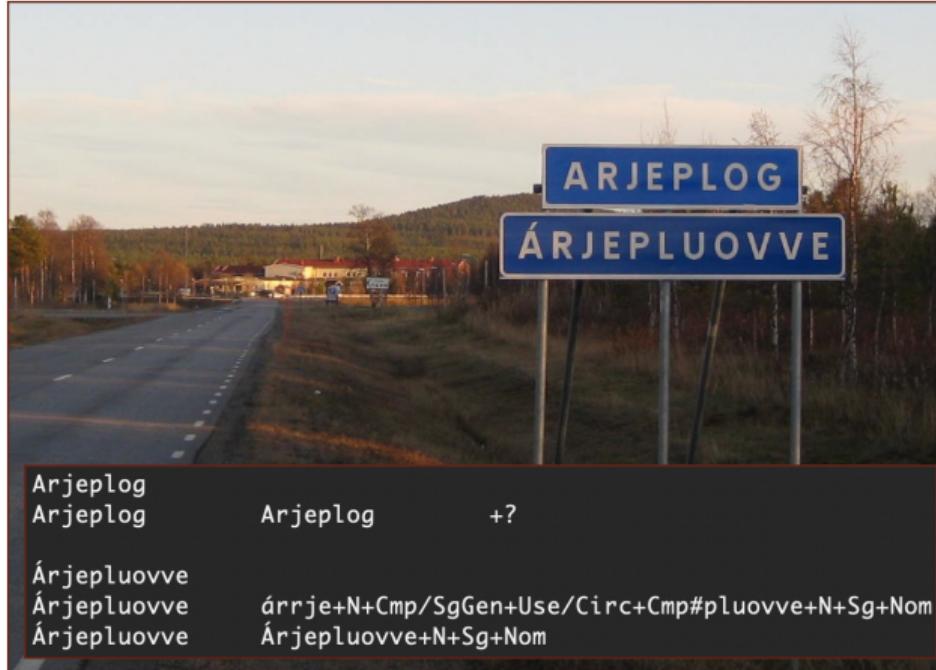


# Digitizing Pite Saami

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

linguistic analyses (using FST)

# Pite Saami language

- Uralic → Finno-Ugric → Saami... Pite Saami
- spoken by ~30 individuals from Arjeplog/Årjepluovve 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

ISO 639-3 code: sje  
Glottocode: pite1240



# Pite Saami language

- Uralic → Finno-Ugric → Saami... Pite Saami
- spoken by ~30 individuals from Arjeplog/Årjepluovve 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

ISO 639-3 code: sje  
Glottocode: pite1240

*'critically endangered'*



# Pite Saami language

morphological structure:

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

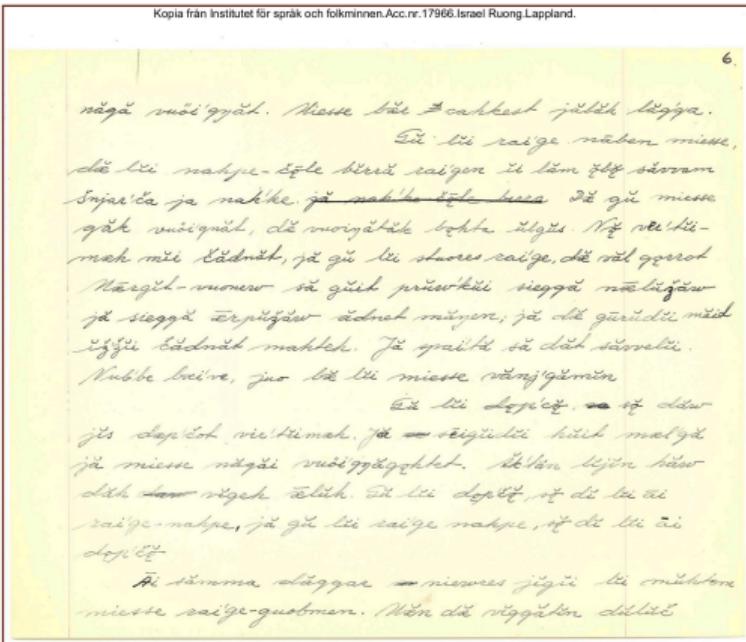
*morphological structure of nouns:*

**stem + class-marker + case/number**

morphological structure of nouns:			
<b>stem + class-marker + case/number</b>			
<i>juällge</i>	<i>juällg</i>	<i>e</i>	-
<i>juolgev</i>	<i>juolg</i>	<i>e</i>	<i>v</i>
<i>juallgáj</i>	<i>juallg</i>	<i>á</i>	<i>j</i>
<i>julgij</i>	<i>julg</i>	<i>i</i>	<i>j</i>
<i>bägga</i>	<i>bägg</i>	<i>a</i>	-
<i>bieggav</i>	<i>biegg</i>	<i>a</i>	<i>v</i>
<i>båtsoj</i>	<i>båts</i>	<i>o</i>	<i>j</i>
<i>buhtsuv</i>	<i>buhts</i>	<i>u</i>	<i>v</i>
<i>vanas</i>	<i>vanas</i>	-	-
<i>vadnasijt</i>	<i>vadnas</i>	<i>i</i>	<i>jt</i>
			'leg/foot'
			'leg/foot' (ACC.SG)
			'leg/foot' (ILL.SG)
			'leg/foot' (GEN.PL)
			'wind'
			'wind' (ACC.SG)
			'reindeer'
			'reindeer' (ACC.SG)
			'boat'
			'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. *jūrrēskä* (Pl.).

no kó len sômjës râljé sâmjé jođtjémen, so ju:tulg:ukqan\*  
(Iness.) čuoččqrén sômjës sq:jén râjtuoj (Kom.Pl.). *jg* so  
käntuolij hëč:kqt gkhtq männä. äjntekä ḡrjstérén, *jut jūrrēskä\**  
*tqwe* röðpglén (3.Pl.Prät.7966).

huspóntte tḡekqj šè·èlastém'euj\* (Akk.W.7322-8) *jg* mgŋ-  
njelij čgskij *jg* so röä·phuolij\* lo:kkoj aðtjé-mijåw *jg* hôloj :  
te vîltet-tel (W.10) mü mânaw ruop<sup>H</sup>tujt. *jg* te ujdtus männä  
jičij (3.Sg.Prät.1598)

männä su·pcqstrg:(lqj, kok so<sup>T</sup>n ðcčuoj vaglijiéw (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

Nachricht

*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 shows the ELAN 5.8 interface with the following components:

- Video Preview:** A video frame showing a person standing in a snowy, forested area.
- Transcript Table:** A table of annotations with columns for ref@AEF, Annotation, orth@AEF, cp@AEF, ft-eng@AEF, ft-swe@AEF, ft-ing@AEF, UFW@AEF, Begin Time, End Time, and Duration. The table lists 96 annotations from 00:01:09 to 00:01:24.
- Timeline:** A timeline at the bottom showing audio waveforms for different speakers and time markers from 00:01:52.000 to 00:01:53.200.
- Annotation Tree:** A hierarchical tree view on the right showing the structure of the annotations. The root node is .028, which branches into ref@AEF [96], orth@AEF [96], ft-eng@A [94], ft-swe@A [94], word@A [473], lemma [493], pos@ [501], morph [642], and gloss [501]. Each node contains its corresponding text transcription.

**Annotations for .028:**

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

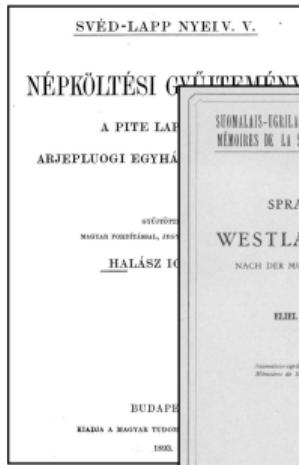
**including annotations for:**

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

## 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

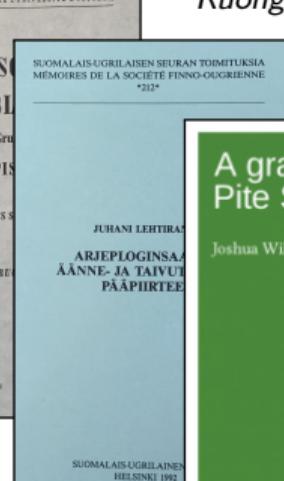
# linguistics research about Pite Saami



Halász 1893 (Hungarian)



Lagercrantz 1926 (German)



Ruong 1943 (German)

A grammar of  
Pite Saami

JUHANI LEHTIRANTA  
ARJEPLOGINSAÄÄNNE- JA TAIVUTTÄÄÄNNE-  
PÄÄPIIRTEE

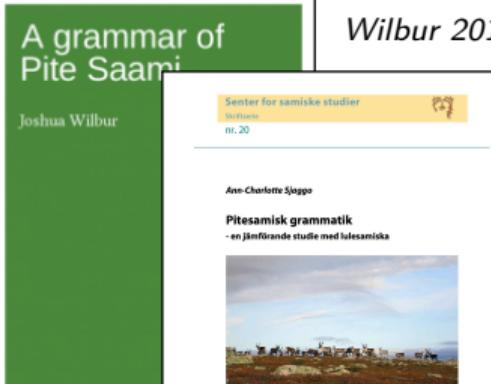
ISRAEL RUMKOFF

SUOMALAIS-UGRILAINEN  
HELSINKI 1992

Studies in Diversity Linguistics, No.

Lehtiranta 1992 (Finnish)

Wilbur 2014 (English)



Joshua Wilbur

Senter for samiske studier  
Innleiret nr. 20

Ann-Charlotte Sjaggo  
Pitesamisk grammatik  
- en jämförande studie med lulesamiska



Sjaggo 2015 (Swedish)

# Pite Saami community

- language activists\* → wordlist (2008-2012)

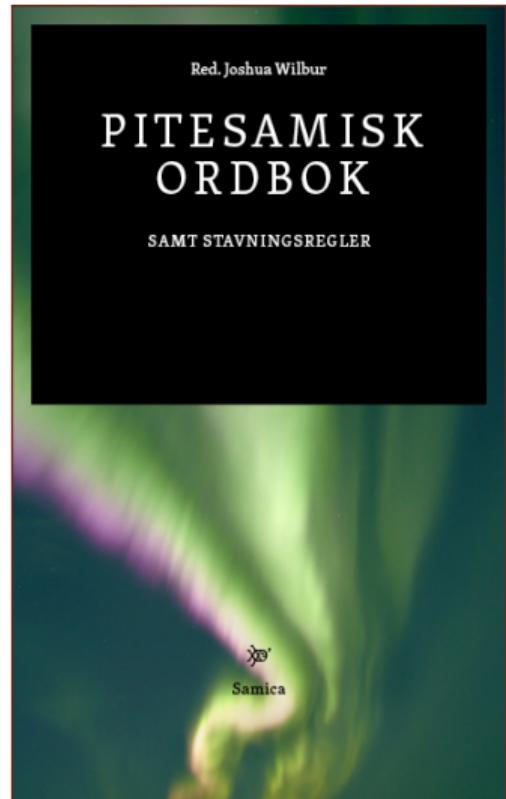
A 1	B posten nr.	C pite	D svenska	E stadieväxling	F omljud	ordklass	de
2373	2505	báldast	gå bredvid varandra				
2374	2506	bálkestít	kasta			verb	
2375	2507	bálkesduvvat	kastas		vv-v	verb	
2376	2508	åro sjávot!	var tyst!			verb	
2377	2509	gárrat	snöra fast, spänna fast	rr-r		verb	
2378	2510	tsievvé	hårdpackad snö	vv-v		substantiv	
2379	2511	sárrjot	hastigt gripa tag i något	rjj-rj		verb	
2380	2512	basske	trång, för liten				
2381	2513	vuastalit	säga emot, protestera			verb	
2382	2514	víjsut	bli klokare				
2383	2515	tjuodtjodäddje	föreständare				
2384	2516	tjuodjat	läta, ljuda	dj-j			
2385	2517	hullvot	yla, om hund, varg				
2386	2518	guoddáldak	gehäng med sysaker				
2387	2519	luossisvuohta	tungsint				
2388	2520	ulgutj	ytterstek på ren				
2389	2521	siissnjutj	innerstek på ren				
2390	2522	strátjadit	gå med ansträngning				
2391	2523	blejve bielle	solsida				
2392	2524	járrát	ramla, snubbla	rr-r			
2393	2525	itka bielle	baksidan av fjäll, berg	ll-l			
2394	2526	itjji	inte				
2395	2527	buonga	börs, portmonnä				
2396	2528	rávvgo	fårskinnsfall				
2397	2529	fáhtala	vävda bärremmar till barkavuassa				
2398	2530	giétjastit	kasta en hastig blick				
2399	2531	gábas	askflaga				
2400	2533	tjuodnama	glödande korn som följer upp med röken				
2401	2534	tjávdvit	lösa upp lex. knut				
2402	2535	búvillko	núlka			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

Bidumsáme Báhkogirrje  
Pitesamisk ordlista • Pite Saami lexical database

Sök:

pitesamiska:	ordklass:	svenska:	engelska:
%bå%	ordklass	svenska	English
stadieväxling:	omlijd:	stavelserantal:	stamförändring:
stadieväxling	omlijud	stavelserantal	stamförändring

sök | återställ | enkel sökning | regEx | info om jokertecken

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

- ahtselisbålás (II-I, -s-) SUAST ave. häufig regnskur eng. heavy rain shower
- arrambåtsoj (hts-ts, -buhtsu-) SUAST ave. fet ren vid gott hull eng. fat reindeer in good condition
- árbálomáj (arbálomá-) SUAST ave. änke man eng. widower
- átjábájjgå (jjg-jg) SUAST ave. rökskamp eng. puffball
- bebbmusbåtsoj (hts-ts, -buhtsu-) SUAST ave. matren eng. reindeer intended as food
- buoremus båddå FRAS ave. bästa stunden eng. best time
- buristemus båddå FRAS ave. väntetiden eng. welcome
- bådå VERB ave. du kommer eng. you come [båhet:2SG.PRS]
- bådådallat (II-I) VTRB ave. bli överraskad, bli ertappad eng. be surprised
- bådåv VERB ave. jag kommer eng. I come [båhetet:1SG.PRS]
- båddnåj SUAST ave. till bottens eng. towards the bottom [båddne:III.LS]
- båddne (ddn-dn) SUAST ave. bottens eng. bottom
- båddnje (ddnj-dnj) SUAST ave. make eng. husband
- båddå SUAST ave. tid, stund eng. time, while
- båddåtj (-tj-) SUAST ave. liten stund eng. a little while
- både lagabij FRAS ave. kom närmare eng. come closer
- både muv båkkto FRAS ave. kom om mig, kom förbi eng. stop by
- både siså FRAS ave. kom in eng. come in
- bådnå VERB ave. han, hon tvinner senträd eng. he/she twists tendon into string [bådnet:3SG.PRS]

bådå

ordklass	verb
svensk	du kommer
engelsk	you come
morfologi	2sg.prs -> båhet
stadieväxld.	omlijud
-	-
stavelserant.	stamförl.
2	-

0:00 / 0:01

version från: 2023-06-30 Multilang  
Om denna webbsida

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

```
juällge  
juällge juällge+N+Sg+Nom  
  
juallgáj  
juallgáj          juällge+N+Sg+Ill  
  
julgijd  
julgijd juällge+N+Pl+Acc  
  
juolgen  
juolgen juällge+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    "bårråt" V ConNeg SELECT:313:ConNeg3  
9 ;>"bårråt" V Imprt Sg2 SELECT:313:ConNeg3  
10 ;>"bårråt" 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
```

# 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

The screenshot shows a portion of the ELAN interface. On the left, a vertical timeline bar has a timestamp '00:01:52.000' at the top and '.028' below it. To the right of the timeline, several annotations are listed. A red circle highlights the 'word@A' annotation and its associated gloss 'black'. The annotations are as follows:

Annotation Type	Value	Gloss
ref@AEF	[96]	
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>
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		<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
  - p
- NLP i
  - (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

# 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](mailto:joshua.wilbur@ut.ee)

