Fieldwork and Grammaticography in a Digital World

```
! ===========
!! !!Vowel rules
! ===========

!! ! metaphony

"Default VH" !regressive vowel height assimilation in V1 with i/u in V2"
!! __@RULENAME@__
    Vx:Vy <=> [#|.#.] Cns:* _ Cns:+ VHtrig Cns:* Dummy:* Vow:* %>:0 ;
    where Vx in ( á a ä å )
          Vy in ( ä i e u )
          matched ;

"Default VH for 4syllables" !(ignores first foot, otherwise same as above
!! __@RULENAME@__
    Vx:Vy <=> [#|.#.] Cns:* Vow:+ Cns:+ Vow:+ Cns:+ _ Cns:+ VHtrig Cns:*
    where Vx in ( á a ä å )
          Vy in ( ä i e u )
          matched ;
```

Joshua Wilbur
Freiburg Research Group in Saami Studies • Universität Freiburg
Descriptive Grammars and Typology • University of Helsinki
28 March 2019

Fieldwork and Grammaticography in a Digital World

Overview

- background
- fieldwork
- grammaticography
- other advances
- outlook

BACKGROUND
(aka: contextualization)

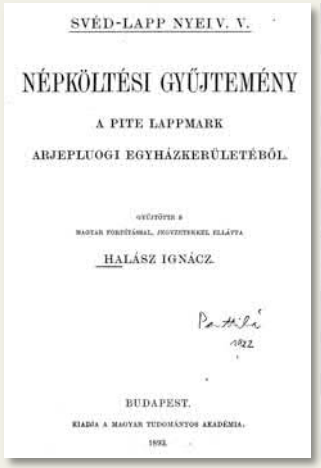
Pite Saami

- Uralic > Finno-Ugric > Saamic
- spoken by **~40** individuals from Arjeplog/*Árjepluovve* in Swedish Lapland
- aka: Arjeplog-Saami, *bidumsámegiella*
- nearly all speakers are at least 50
- all speakers are bilingual (Pite Saami and Swedish/*Arjeplogsmål*)
- no *official* orthography (yet...), but a working standard
- no media
- Swedish dominates everyday life
- hardly being passed on to younger generations

Pite Saami

larger linguistic studies:

- Halász 1893 (in Hungarian)
- Lagercrantz 1926 (in German)
- Ruong 1943 (in German)
- Lehtiranta 1992 (in Finnish)
- Wilbur 2014 (in English)
- Sjaggo 2015 (in Swedish)



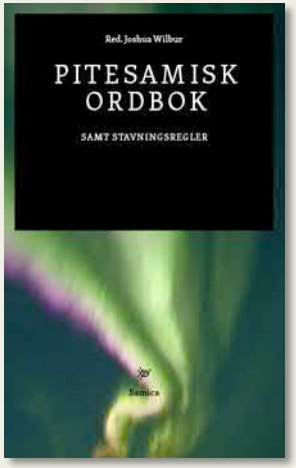
Pite Saami

larger linguistic studies:

- Halász 1893 (in Hungarian)
- Lagercrantz 1926 (in German)
- Ruong 1943 (in German)
- Lehtiranta 1992 (in Finnish)
- Wilbur 2014 (in English)
- Sjaggo 2015 (in Swedish)

other materials:

- extensive collection of heritage materials (ISOF, Uppsala)
- dictionary (Pite Saami -> Swedish/English)
  and proposed orthographic rules (2016)
- online lexical database
- online orthographic rules (including spellchecker (in beta))
- smart phone app in the works



Pite Saami

larger linguistic studies:

- Halász 1893 (in Hungarian)
- Lagercrantz 1926 (in German)
- Ruong 1943 (in German)
- Lehtiranta 1992 (in Finnish)
- Wilbur 2014 (in English)
- Sjaggo 2015 (in Swedish)

other materials:

- Extensive collection of heritage materials (ISOF, Uppsala)
- dictionary (Pite Saami -> Swedish/English)
  and proposed orthographic rules (2016)
- online lexical database
- online orthographic rules (including spellchecker (in beta))
- smart phone app in the works

recent linguistics projects:

- Documentation (2008-2015; materials archived at ELAR and TLA)
- Lexicography (2016)
- Syntactic structures (2016-present)

Pite Saami

larger linguistic studies:

- Halász 1893 (in Hungarian)
- Lagercrantz 1926 (in German)
- Ruong 1943 (in German)
- Lehtiranta 1992 (in Finnish)
- Wilbur 2014 (in English)
- Sjaggo 2015 (in Swedish)

other materials:

- Extensive collection of heritage materials (ISOF, Uppsala)
- dictionary (Pite Saami -> Swedish/English)
  and proposed orthographic rules (2016)
- online lexical database
- online orthographic rules (including spellchecker (in beta))
- smart phone app in the works

recent linguistics projects:

- Documentation (2008-2015; materials archived at ELAR and TLA)
- Lexicography (2016)
- Syntactic structures (2016-present)

Pite Saami

-> each fieldwork situation is unique!

- significant aspects of mine include:
  - an accessible modern technological infrastructure on-site
  - a previous history of linguistics work
  - extensive language technology tools for closely-related languages
  - messy but extant orthographic “tradition” when I started

FIELDWORK in a digital world

tools for fieldwork

- in the old days: notebook and pencil
- nowadays:
  - recording equipment
  - laptop
  - digital backup capacity (even in the cloud)
  - transcription software (ELAN)
  - mobile phones
  - social media
    (e.g.: for staying in contact, data source)
  - grammaticography software
    (e.g. FLEx for interlinearization)
  - language technology

  …

data collection and fieldwork

- modern, affordable digital recording technologies (especially video) allow fieldworkers to capture much more than just language, but the entire human event
  - more complete documentation, potentially useful beyond linguistics*

why not use:

- body cameras
- drones
- surround-sound microphones
- 360° cameras
- 3-D cameras

...

*cf. Rießler & Wilbur 2017

(re-)collecting old data
(heritage harvesting)

- OCR (optical character recognition)

34 SAMEFOLKETS EGEN TIDNING

Päŋka kädde

Män kalkav mujhtalit sämow tulutj kiektow, mav män mädde palen kulliv iehtjan ähtjest ja tjidtj-ajjasf, ku män lijiv mannam. Tat suptsas ja tajst Vuona-lantajst, kumse Tjidtjaka samoh ålen juhtin ja kiesijt urrun. Ja tan tihta mån tav muihtajv tjallet, ku tål kiesse ja ruvvaj påhtemin. Lij sämes sapme, mij pruvh kuj kiesen jåhtet Stuormon Vuona rijhkan. Tab lin pär kålbmå almatja, tat same-pundj, su ahka, ja säj inika aktav nejtav. Te lij tål kiesse, ja kåhte lij tål tan Stuormo vakken. Ja iello kuotoj taju krasse-pulkojn ja tan vakken jåhkå-katten. Te vuolkepa tah kuoktes vulus Vuonan vadtset, ja tat näjta pahtsa iktuk kåhtaj ielov vuoidnet. Tah kuoktes årojka vuonan mådde pejveh, ja ku te vist pajas vadtsaika, ja ku putijka kåte vujdnusij, te vuojdnepa såj, aht iello lij kujht kåte kuoran livan. Såj vättsijka. Ja ku te kåte lahka putijka, mav såj te äska åttjojka vuoidnet? Tat iello, mav säj jåhkijka lij livvatemen — tat lij kalle urrum livva iello — mån mij lij sjaddam tajna ielojn ja nejtajn? Såj vujnijka, aht näjta lij tjuhkim ielov ja mannam kåte kuorraj ja livutahtam ielov ja alkam altojt påhtjet. Ja ku lij påhtjemin, te lij atja-rajte jala ajja-rajte tsapmestam tav livva ielov ja påhtjej nejtav jamas. Åpå sijta iello ja almatj vil aj fieraj tasa tan kåddaj.

Ja iellam pahtsam ienap puhtsuh viessot, ku tah mah lin ulkolin kuohtomin jala »skånärtämen» — näv kukte mija varen pruvhkujin hållåt, ja tah iellam ienap ku nakan nälje-låhk hägga.

Tan kuoktasij sjattaj lussis vajve ja umårretis kiesse-pejveh. Näjta lij håhkånam ja kajhka iello latkanam, aht såj iepa mahtam jåhtet. Ja tat rajest kuhtjun tav kiettev Päŋka kädden, man nala lij åpå sijta iello ja näjta atja-raitest tsapmestuvvum jamas.

Mån aj mujhtav mu manna palest, aht vuoras almatjah kiesen, ku kalkajme ielov rejdnuhit, tjårkå vakotin, aht kåsse atja-rajte lä mannamin, de ij kalka padjat ielov tjåhkånit jala livutit ajnat tiptet suojbma ja luopoj kuohtot, tasak rajte vassa. Ja ij kalkam stivrit ielov nievres ietnamij, kånne läh pakteh ja juovah ajnat stivrit jalka ietnamij, kånne ij tarpahå iello tjåhkånit. Ja almatj ietj, vajku ajja ahtselis rassjo påhta, te idtjij åttjo rasjost pahtarit ja tjakŋat kärke-lapa vuollaj, tat lij iello-tjärråkav jåkå rasta. Ja te kalkaj sån ietj vist rasta mannat, men te vältij jåhkå su, ja säkkoj näjta tan jåhkåj. Skajte-kietjen lij jåhkå tjaskam rupmahav kaddaj. Ja pallin almatjah råkkev tan skajtaj ja javestin rupmahav tasa. Mån mädde jakeh tat rajest vässo almatjah kullin, ku tat näjta juojkaj tan skajte-kietjen ja hålåj: »Jåkå skåvvå kul mu pär pieljehtuhtij.» Ja lerrin almatjah tav vuolev ja lierahtin puolvast puolvaj.

Suptsas atja-rajte pirra

Tav lä mu tjidt-ajja mujhtalam. Lij sämes sapme mij iktuk urutij ja kulijt pivtij. Kåtatjav inij kånne viesoj. Sämes palen ku lij iktuk kåten, te påhta sisa kåhtaj amas stuorra ålmaj, mån nubbe juolke lä ånep. Nuppen kietan lä stuorra ruovte-karre ja tidno ja nuppen kietan stuorra vähtjer. Ja te hålla sapmaj: »Ihta tål tån mu tåptå?» Sapme pallaj ja vastetij aht idtjij sån täptå. Tä hålla tat stuorra ålmaj: »Mån läv tål tat atja-rajte. Ku kullapihtit mån läv mannamin, tä tsapmav tajna vähtjerijn tan ruovte-karraj, ja ku tidnuj tsapmav te altakastav, ja rassjo påhta taste, ku karest lejhkiv tjatsev.» Ja te vit hålla: »Ku kullapihtit mån läv påhtemin, te ehpit åttjo aktakav tahkat jala parkat, ajnat verhtipehtit vuortet tasak mån läv mäddel mannam. Mån iv vuokahä, aht tijah kalkapihtit nakanav tahkat, ku mån läv mannulakan. Tav kalka tån hållåt kajhka ietja almatjijta.

Ja te mån aj mujhtav, aht tulutj vuoras almatjah lin nävht mårahkah manajta, jus idtjin sjavot åro, ku atja-rajte jala »Balv-ajja» lij mannamin.

Lars Rensund.

Från Jokkmokk har insänts följande:

Muohta-kuolto.

Tiima kita lii aktii mårhmes mano manŋŋekietjen tjiervas kuolto. Vuostak muohtii åvvå ätnakav pieiven, ja te iehketis alhkii pieggat. Ja taina kå lii muohttam åtå muohttakav, te sjattai muohta-kuolto.

Mii liima vitas mah aniima ällokåtev Ultevistuoddarin. Kå piegga karrai iehketis, te mii vierhtiima tjadnat kåtev sämes kärkai ja tsakkatit ätnårisait ja täddalit låuta-maddariit. Idja-kuoutelin mii vierhtiima padjanit ja kamatit, taina ká piegga lii näv karram, atte

Renkalvarna fångades med metspön

Skogssamerna från Västra Kikkejaur samlas varje sommar i Tjatnis vid Baktälve för märkning av årets renkalvar. Därvid tillämpar man ett av Gällivaresamerna prövat system vid infångandet av kalvarna. I stället för lasso har man använt sig av långa spön, i vars ytterstа ända en repögla anbringats, och på detta sätt fångas renkalvarna i bakbenen.

Två skogssamer från Västra Kikkejaur voro för några år sedan på studieresa till Gällivare, där skogssamernas sommarrenskötsel studerades. Där ha samerna nämligen gått in för den intensiva renskötseln och bl. a. infört det nya ovannämnda systemet ifråga om kalvarnas infångande. Detta har bl. a. den förtjänsten, att kalvarna inte skiljas från vajorna, enär renarna inte skrämmas av vinande lasso. Renarna förhålla sig lugnare och när kalvarna följa vajorna helt stillsamt, är risken för felmärkning nästan helt utesluten. Framför allt har lappfogden varit intresserad av att det nya systemet prövas, enär det har sina givna förtjänster. Gällivaresamerna ha även gått in för att följa renarna natt och dag, vilket gör renarna lugnare och tamare.

Det var givetvis ingen lätt sak för samerna att på en gång övergå från lassokastningen till »metandet», och i slutet av märkningen i Tjatnis nyttjade man det gamla beprövade sättet. Rengärdet var nämligen i största laget, vilket gjorde, att »metandet» blev allt för tidsödande.

I Tjatnis håller det på att växa upp en lappstad i miniatyr. Där har nämligen byggts sex kåtor i stil med de, som finnas i lappstaden i Arvidsjaur, två härbren och tre rengärden.

Hur dränգen Lars blev en stor renpatron

Följande »historia» har mina förfäder berättat för mig, och de gjorde detta så, att berättelsen fick sannolikhetens prägel.

Händelsen förlägges till cirka tre kilometer norr om Arakanolo, där rengärdet Riebei-bäive-gädde ligger. Lars var dräng åt en rik skogssame vid namn Olof — i dagligt tal kallad Wuolla. En natt, då Lars var ute och vallade renhjorden, mötte han två näpna sameflickor — wieternäidah. De var båda klädda i »svaltjah», ytterplagg, sydda av beredda renskinn, som färgats röda med den naturfärg, som framställts av albark. Flickornas härflätor var långa och prydda med vackra pärlband av runda pärlor. Dessutom hade dessa töser i sina långa flätor färdigsnodda sensnören, som samer förr i tiden använde som sytråd vid olika sömnader. De hade också var sin hund med sig. Den ena av flickorna frågade Lars, vilkendera han helst önskade sig: en fågelhund eller en renhund. Lars ville naturligtvis ha renhunden, som han också genast fick som gåva.

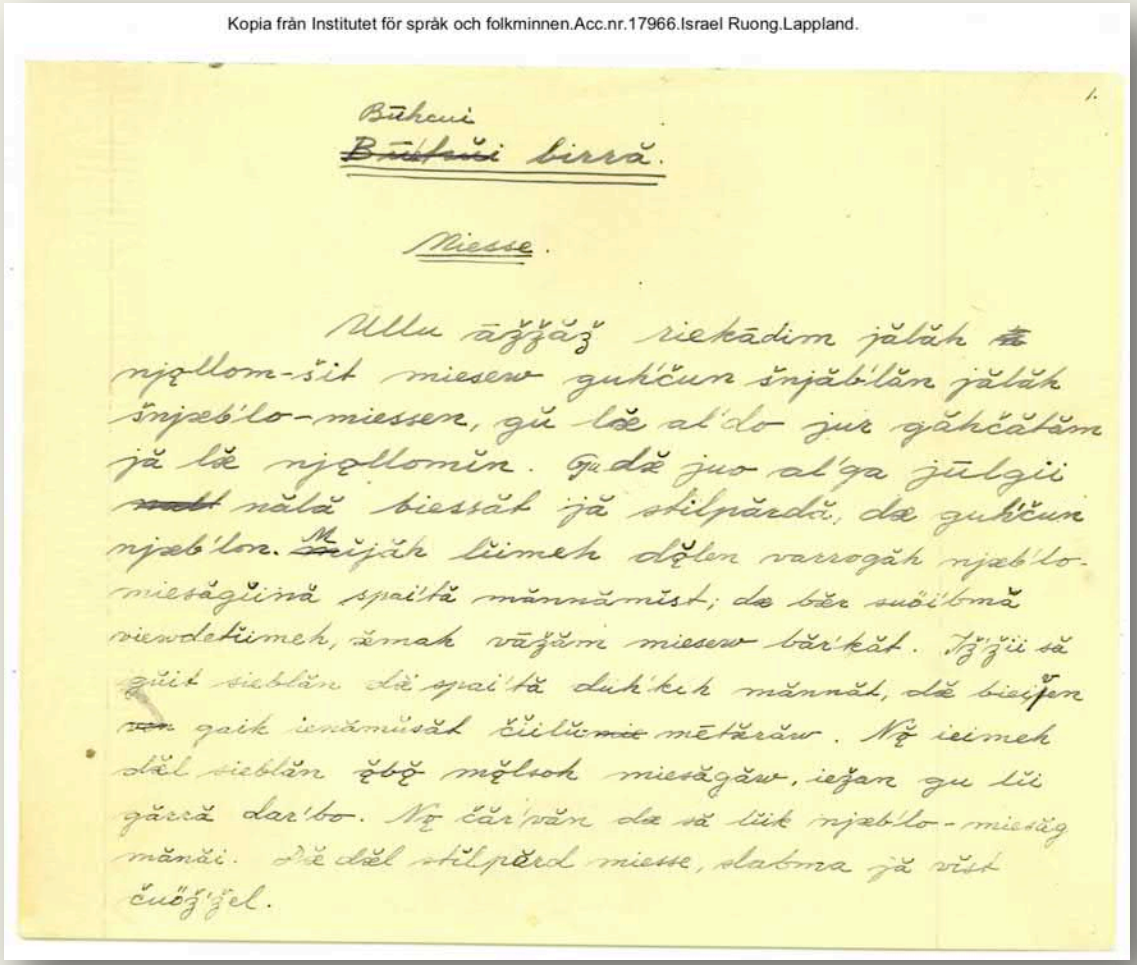
Då berättas det, att hans renar började föröka sig så snabbt, att han på några år blev en stor renhusbonde. Till att börja med hade han visserligen ett fåtal renar, som han fått som dränglön under årens lopp, men sedan han blev ägare till den underbara renhunden, visste han inte ordet av, innan han blivit

- embedded text
  (more than just scanning!)
- can be exported
  (e.g. to ELAN)
- can be part of a corpus*

*cf. Partanen & Rießler 2019

(re-)collecting old data
(heritage harvesting)

- HTR (handwritten text recognition)



- embedded text
  (more than just scanning!)
- can be exported
  (e.g. to ELAN)
- can be part of a corpus*
- *much more complex than OCR, thus it currently requires much more training data before it's useful*

*cf. Transkribus project (Kahle 2017); also Blokland et al 2019 for a brief discussion

```
352 ! ===========
353 !! !!Vowel rules
354 ! ===========
355
356 !! ! metaphony
357
358 "Default VH" !regressive vowel height assimilation in V1 with i/u in V2" ! á:ä a:i ä:e å:u ^O:0
359 !! __@RULENAME@__
360     Vx:Vy <=> [#|.#.] Cns:* _ Cns:+ VHtrig Cns:* Dummy:* Vow:* %>:0 ;
361     where Vx in ( á a ä å )
362           Vy in ( ä i e u )
363           matched ;
364
365 "Default VH for 4syllables" !(ignores first foot, otherwise same as above)
366 !! __@RULENAME@__
367     Vx:Vy <=> [#|.#.] Cns:* Vow:+ Cns:+ Vow:+ Cns:+ _ Cns:+ VHtrig Cns:* Dummy:* Vow:* %>:0 ;
368     where Vx in ( á a ä å )
369           Vy in ( ä i e u )
370           matched ;
```

GRAMMATICOGRAPHY
in a digital world

brief history of grammaticography

- 1/3 of the Boasian trilogy

…

- Payne 1997, Mosel 2006, Aikhenvald 2015, etc.
- Nordhoff 2008 *Electronic Reference Grammars for Typology: Challenges and Solutions*
- **Implemented grammars (incorporation in corpus and computational linguistics)**

digital tools for grammaticography

- Toolbox, FLEx
  - good for concatenative morphology
    - *play, play-s, play-ed, play-er, play-er-s*
  - not so good for non-linear morphology
    - *s**i**ng, s**i**ng-s, s**a**ng, s**u**ng*

*What do you do when non-linear morphology is the **default** in your language?*

digital tools for grammaticography

- Toolbox, FLEx

*What do you do when non-linear morphology is the **default** in your language?*

	SG	PL
NOM	*juällge*	*juolge*
GEN	*juolge*	*julgij*
ACC	*juolgev*	*julgijt*
ILL	*juallgáj*	*julgijda*
INESS	*juolgen*	*julgijn*
ELAT	*juolgest*	*julgijst*
COM	*julgijna*	*julgij*
ABESS	*juolgedak*	*juolgedaga*
ESS	*juallgen*	

juällge ‘foot/leg’

digital tools for grammaticography

- Toolbox, FLEx
- other, digital approaches...

	SG	PL
NOM	*juällge*	*juolge*
GEN	*juolge*	*julgij*
ACC	*juolgev*	*julgijt*
ILL	*juallgáj*	*julgijda*
INESS	*juolgen*	*julgijn*
ELAT	*juolgest*	*julgijst*
COM	*julgijna*	*julgij*
ABESS	*juolgedak*	*juolgedaga*
ESS	*juallgen*	

juällge 'foot/leg'

4 stem allomorphs:
juällg-
juolg-
juallg-
julg-

*What do you do when non-linear morphology is the **default** in your language?*

implemented grammars

- aka “precise” grammars
  - self-validating
- computer-processable
  - but only borderline human-readable (at least from a traditionalist perspective)
  - computational linguists, typically HPSG
- **analyze** linguistic structures
- implementation --> **parse and tag a corpus**

Siegel et al. 2016

cf. new *Language Science Press* series “Implemented Grammars”

implemented grammar (FST/CG) for Pite Saami

- Giellatekno infrastructure:
  the Research group for Saami language technology at University Tromsø
  - FST − Finite State Transducer[1]
  - CG − Constraint Grammar[2]
- **automatic** annotations in ELAN…

[1]Beesley & Karttunen 2003; [2]Didriksen 2007–2018, Karlsson 1990; Karlsson et al. 1995

implemented grammar (FST/CG) for Pite Saami

infrastructure:

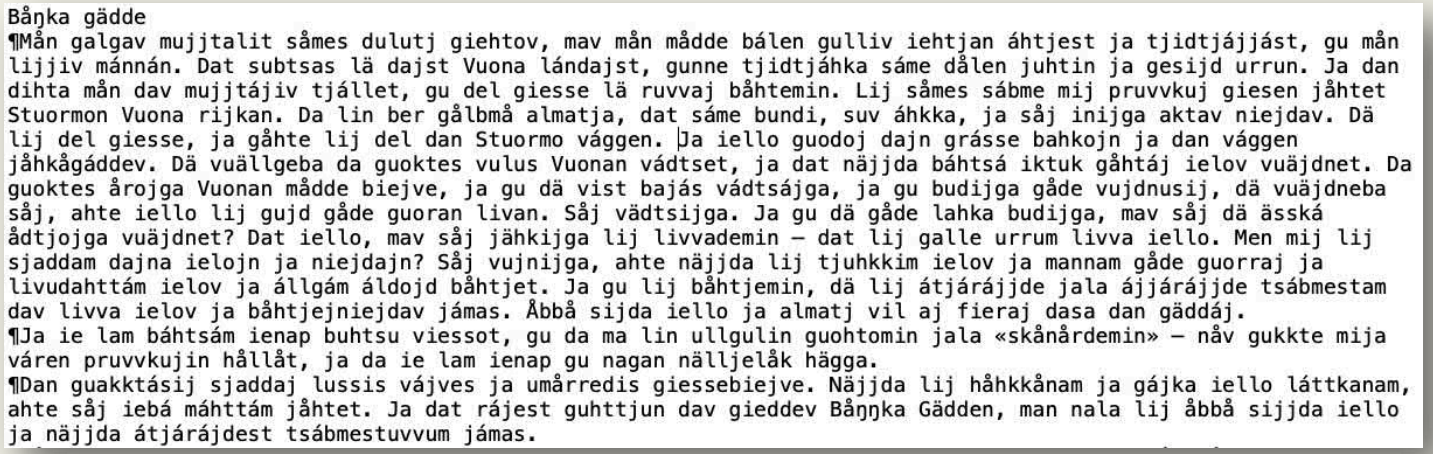
Finite State Transducer (FST) → for analyzing wordforms

Constraint Grammar (CG) → for removing ambiguities in FST output

formalism:

- `lexc` (lexicon, PoS, linear morphology)
- `twolc` (non-linear morphology)
- `cg3` (syntax)

Uses orthographic standard!



Båŋka gädde
¶Mån galgav mujjtalit såmes dulutj giehtov, mav mån mådde bálen gulliv iehtjan áhtjest ja tjidtjájjást, gu mån lijjiv mánnán. Dat subtsas lä dajst Vuona lándajst, gunne tjidtjáhka sáme dålen juhtin ja gesijd urrun. Ja dan dihta mån dav mujjtájiv tjállet, gu del giesse lä ruvvaj båhtemin. Lij såmes sábme mij pruvvkuj giesen jåhtet Stuormon Vuona rijkan. Da lin ber gålbmå almatja, dat sáme bundi, suv áhkka, ja såj inijga aktav niejdav. Dä lij del giesse, ja gåhte lij del dan Stuormo väggen. Ja iello guodoj dajn grásse bahkojn ja dan väggen jåhkågáddev. Dä vuällgeba da guoktes vulus Vuonan vádtset, ja dat näjjda báhtsá iktuk gåhtáj ielov vuäjdnet. Da guoktes árojga Vuonan mådde biejve, ja gu dä vist bajás vádtsájga, ja gu budijga gåde vujdnusij, dä vuäjdneba såj, ahte iello lij gujd gåde guoran livan. Såj vädtsijga. Ja gu dä gåde lahka budijga, mav såj dä ässká ådtjojga vuäjdnet? Dat iello, mav såj jåhkijga lij livvademin – dat lij galle urrum livva iello. Men mij lij sjaddam dajna ielojn ja niejdajn? Såj vujnijga, ahte näjjda lij tjuhkkim ielov ja mannam gåde guorraj ja livudahttám ielov ja állgám áldojd båhtjet. Ja gu lij båhtjemin, dä lij átjárájjde jala ájjárájjde tsábmestam dav livva ielov ja båhtjejniejdav jámas. Åbbå sijda iello ja almatj vil aj fieraj dasa dan gäddáj.
¶Ja ie lam báhtsám ienap buhtsu viessot, gu da ma lin ullgulin guohtomin jala «skånårdemin» – nåv gukkte mija váren pruvvkujin hållåt, ja da ie lam ienap gu nagan nälljelåk hägga.
¶Dan guakktásij sjaddaj lussis vájves ja umårredis giessebiejve. Näjjda lij håhkkånam ja gájka iello láttkanam, ahte såj iebá máhttám jåhtet. Ja dat rájest guhttjun dav gieddev Båŋŋka Gädden, man nala lij åbbå sijjda iello ja näjjda átjárájdest tsábmestuvvum jámas.

implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

formalism:

- `lexc` (lexicon, PoS, linear morphology)
- `twolc` (non-linear morphology)

Output analyses:

```
juällge
juällge juällge+N+Sg+Nom

juallgáj
juallgáj          juällge+N+Sg+Ill

julgijd
julgijd juällge+N+Pl+Acc

juolgen
juolgen juällge+N+Sg+Ine
```

implemented grammar (FST/CG) for Pite Saami

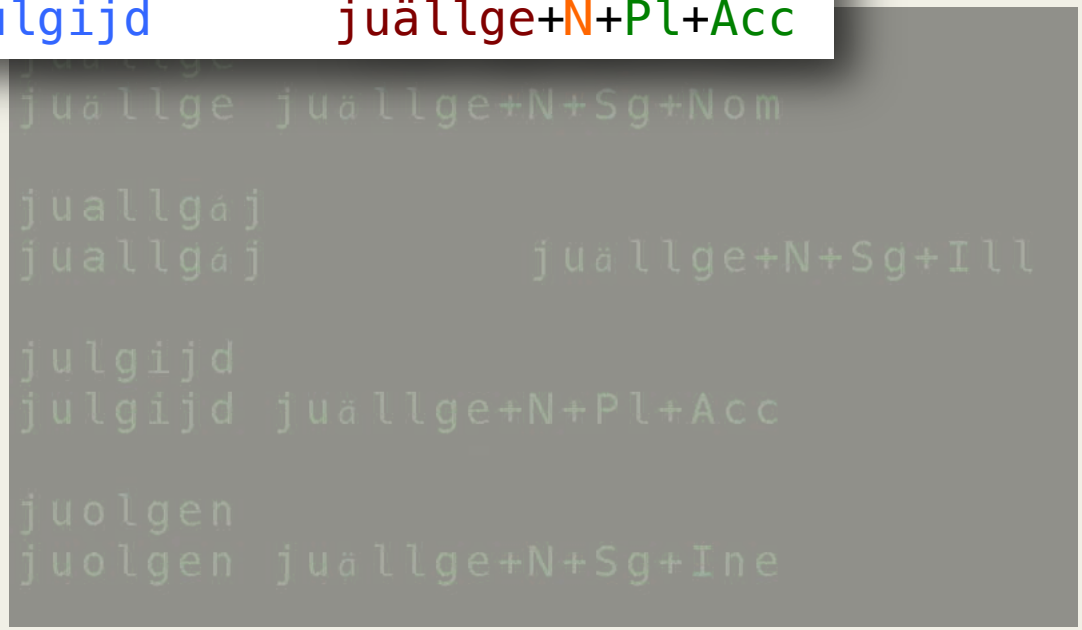
infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

input: wordform

output: wordform lemma+PoS+Morphology

```
juällge
juällge      juällge+N+Sg+Nom

julgijd
julgijd      juällge+N+Pl+Acc
```



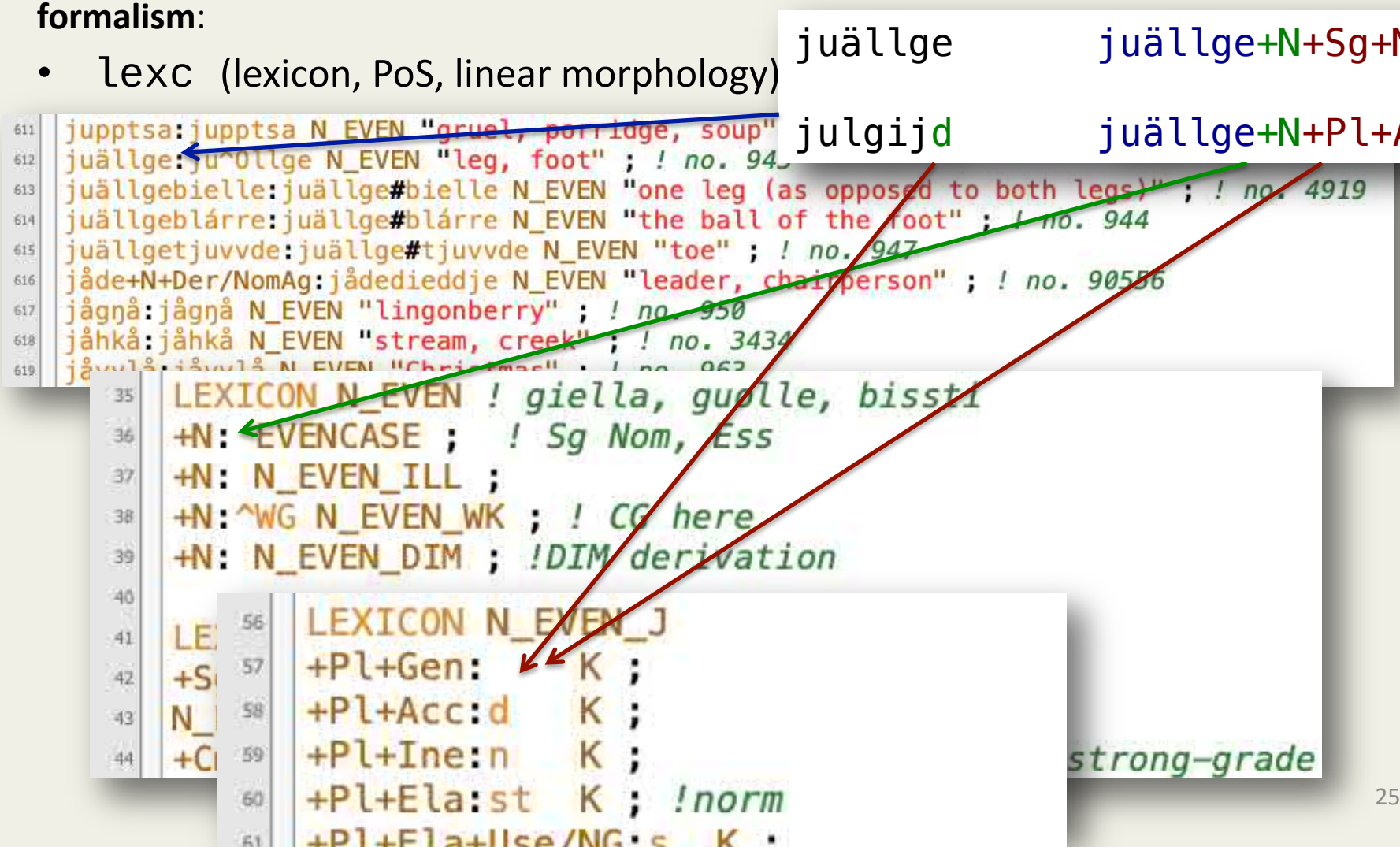
implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

formalism:

- `lexc` (lexicon, PoS, linear morphology)



implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

formalism:

- `twolc` (non-linear morphology)

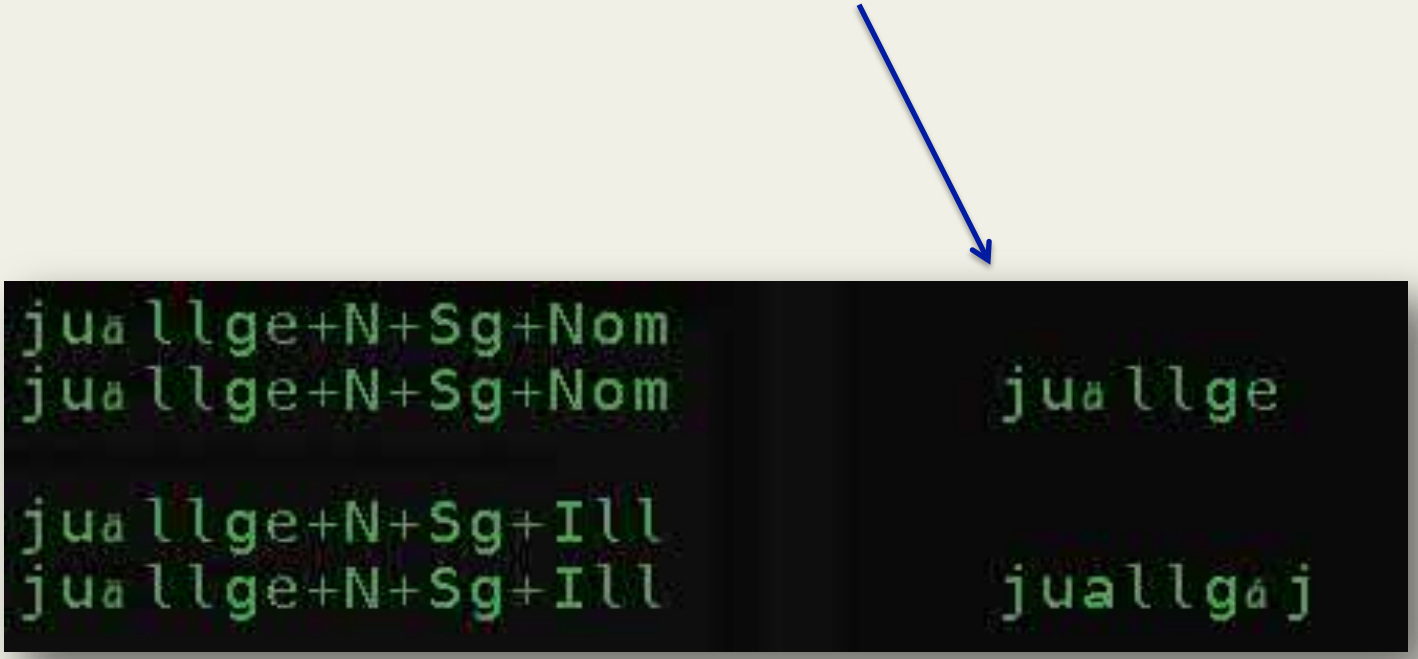
```
juällge        juällge+N+Sg+Nom

julgijd        juällge+N+Pl+Acc
```

```
"Consonant Gradation for xxy:xy"
!! __@RULENAME@__
Cx:0 <=> Vow:+ Cx _ Cy Vow:+ Cns:* ( Dummy: ) %^WG: ;
     where Cx in ( l l l l l l l l l l m m m m )
           Cy in ( j b d f g k p s v b g p s )
          matched ;

"Default VH" !regressive vowel height assimilation in V1 with i/u in V2" ! á:ä a:i ä:e å:u ^O:0
!! __@RULENAME@__
    Vx:Vy <=> [#|.#.] Cns:* _ Cns:+ VHtrig Cns:* Dummy:* Vow:* %>:0 ;
    where Vx in ( á a ä å )
          Vy in ( ä i e u )
          matched ;

"Default UA in G3" ! u^O:ua (juallgáj, luákkta); always with word-initial C
!! __@RULENAME@__
%^O:a <=> Cns:* u _ G3 [:a|:o|:y|:ä|:å|:á] \%^WG: ;
          Cns:* u _ noCG [:a|:o|:y|:ä|:å|:á] Cns:* %^G3:* %^UAUML: ;

"Special UÄ (VH) in G3" ! u^O:ua (juällge); always with word-initial C
!! __@RULENAME@__
%^O:ä <=> Cns:* u _ G3 :e \%^WG: ;
          Cns:* u _ Cns:+ :e (j:j) %^G3: ; ! duädde<->duodde, guäddej<->guäddeja
          Cns:* u _ noCG :e Cns: %^UAUML: ; ! tjuädtjelit (not tjuodtjelit)

"Special VH for u^O" ! u^O:u0
!! __@RULENAME@__
%^O:0 <=> Cns:* u _ Cns:+ VHtrig Cns:* Dummy:* %>:* ; !juällge<->julgij

"V2 E to I before j-suffixes" ! guolle -> gulij
```

implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for **generating** wordforms



(it works in both directions)

implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

BUT: how to deal with **morphologically ambiguous** wordforms? (disambiguation)

implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

Constraint Grammar (CG) → for removing ambiguities in FST output

formalism:

- `lexc` (lexicon, PoS, linear morphology)
- `twolc` (non-linear morphology)
- `cg3` (syntax)

```
328 #NP (incl. pronouns) preceding certain Po is Gen
329 SELECT:genB4Po Gen IF (*1C Po BARRIER NoNP);
330 REMOVE:NoPoWithoutGen Po IF (NOT *-1 Gen BARRIER NoNP);
331 #NP (incl. pronouns) following certain Pr is Gen
332 SELECT:genAfterPr Gen IF (*-1C Pr BARRIER NoNP);
333 REMOVE:NoPrWithoutGen Pr IF (NOT *1 Gen BARRIER NoNP);
```

example: rules describing dependency between adpositions and genitive case

implemented grammar (FST/CG) for Pite Saami

infrastructure:

Finite State Transducer (FST) → for analyzing wordforms

Constraint Grammar (CG) → for removing ambiguities in FST output

formalism:

- `lexc` (lexicon, PoS, linear morphology)
- `twolc` (non-linear morphology)
- `cg3` (syntax)

```
#NP (incl. pro
SELECT:genB4Po
REMOVE:NoPoWit
#NP (incl. pro
SELECT:genAfte
REMOVE:NoPrWit
```

output (analyses)

```
"<mån>"
    "mån" Pron Pers Sg1 Nom
"<lev>"
    "lä" V Ind Prs Sg1
"<julgij>"
    "juällge" N Pl Gen SELECT:329:genB4Po
;   "juällge" N Pl Com SELECT:329:genB4Po
"<nanne>"
    "nanne" Po
;   "nanne" Adv REMOVE:313:NoAdvIfPo
```

disambiguation example

nala *gähttjat* *tjurvij* *daj*

disambiguation example

nala	*gähttjat*	*tjurvij*	*daj*
onto	look+INF	antler+GEN+PL	DET+GEN+PL
		antler+COM+PL	DET+COM+PL
			PRON+GEN+PL
			PRON+COM+PL

FST output:

```
nala
nala     nala+Po

gähttjat
gähttjat         gähttjat+V+Inf

tjurvij
tjurvij  tjårrve+N+Pl+Gen
tjurvij  tjårrve+N+Pl+Com

daj
daj      dat+Det+Pl+Gen
daj      dat+Det+Pl+Com
daj      dat+Pron+Dem+Pl+Gen
daj      dat+Pron+Dem+Pl+Com
```

disambiguation example

daj *tjurvij* *nala* *gähttjat*

‘to look at those antlers’ [pit100405b.011]

FST output:

```
nala
nala    nala+Po

gähttjat
gähttjat        gähttjat+V+Inf

tjurvij
tjurvij tjårrve+N+Pl+Gen
tjurvij tjårrve+N+Pl+Com

daj
daj     dat+Det+Pl+Gen
daj     dat+Det+Pl+Com
daj     dat+Pron+Dem+Pl+Gen
daj     dat+Pron+Dem+Pl+Com
```

disambiguation example

daj	*tjurvij*	*nala*	*gähttjat*
da-j	tjurvi-j	nala	gähttja-t
DET-GEN.PL	antler-GEN.PL	onto	look-INF

‘to look at those antlers’ [pit100405b.011]

FST output:

```
nala
nala     nala+Po

gähttjat
gähttjat          gähttjat+V+Inf

tjurvij
tjurvij  tjårrve+N+Pl+Gen
tjurvij  tjårrve+N+Pl+Com

daj
daj      dat+Det+Pl+Gen
daj      dat+Det+Pl+Com
daj      dat+Pron+Dem+Pl+Gen
daj      dat+Pron+Dem+Pl+Com
```

CG syntactic disambiguation:

- postpositions govern genitive NPs

```
SELECT Gen IF (*1C Po BARRIER NoNP);
```

- pronouns are not embedded in an NP

```
REMOVE Pron IF (*1C N BARRIER NPNH);
```

implemented grammars

pros:

- entirely digital (easy copying, versioning, etc.)
- computer-processable
- can analyze AND generate (useful for practical tools, e.g. teaching apps)
- accuracy can be tested on *real empirical* data
- prose can be included (as `<!—comments-->`)
- **further use in other, digital applications...**

cons:

- requires significant technical knowhow to learn and to implement
- not very human-readable, especially for non-specialists
  - prose is only included as `<!--comments-->`
  - not ideal for standard average typologists
  - not even close to ideal for most non-linguists

further use in other, digital applications...

- spell-checkers
- grammar-checkers
- teaching materials (e.g. apps)

...

Johan Lasko ja suv áhka Hiŋga viessomájge birra

The life history of Johan Lasko and his wife Hinga

sje19210000a-lagercrantz1957a-426

by Maria Johansson, originally transcribed by Eliel Lagercrantz in 1921.

The Finno-Ugric transcriptions and German translations were originally published as text 426 in Eliel Lagercrantz, 1957. *West- und südlappische Texte*. Lappische Volksdichtung, volume 1. Suomalais-ugrilaisen Seuran Toimituksia 112. Helsinki: Suomalais-Ugrilainen Seura; these are published by permission of Suomalais-Ugrilainen Seura.

All other transcriptions and analyses copyright Joshua Wilbur and the Pite Saami Documentation Project, licensed by CC BY-NC-SA 4.0.

Lexical, wordclass and morphological analyses as well as English glosses were derived automatically, thanks in part to the Giellatekno infrastructure and to scripting assistance by Iris Perkmann and Ciprian Gerstenberger.

Johan Lasko ja suv áhka Hiŋga viessomájge birra.

FUT: Johan Lasko ja sù ähkà Hinka viessụöm-äjkjė pirrạ.

EN: The life history of Johan Lasko and his wife Hinga.

DE: Über die Lebensgeschichte des Johan Lasko und seiner Frau Hinga.

¶

Hiŋŋga lij riegádam Árjepluove nuortabiele suoknon, jábeп 1844, ja suv álmáj aj.

FUT: Hiŋka lij rì · èkàtạm ärjạ-pluvjė nǒǒ · rнtạ-pì:èljė su:oknuon, jâpjėn 1844, ja sù olmaj âj.

EN: Hinga was born on the north side of Arjeplog parish in 1844, and her husband as well.

DE: Hinga war im J. 1844 im Nordteil der Gemeinde Arjeplog geboren und so auch ihr Mann.

Sån lij nuorap ietjas áhkast

FUT: soᵀn lij nù · òrap ì · èčạs ähkàst.

EN: He was younger than his wife.

DE: Dieser war jünger als seine Frau.

áhkka
N Sg Ela
grandmother, wife

further use in other, digital applications...

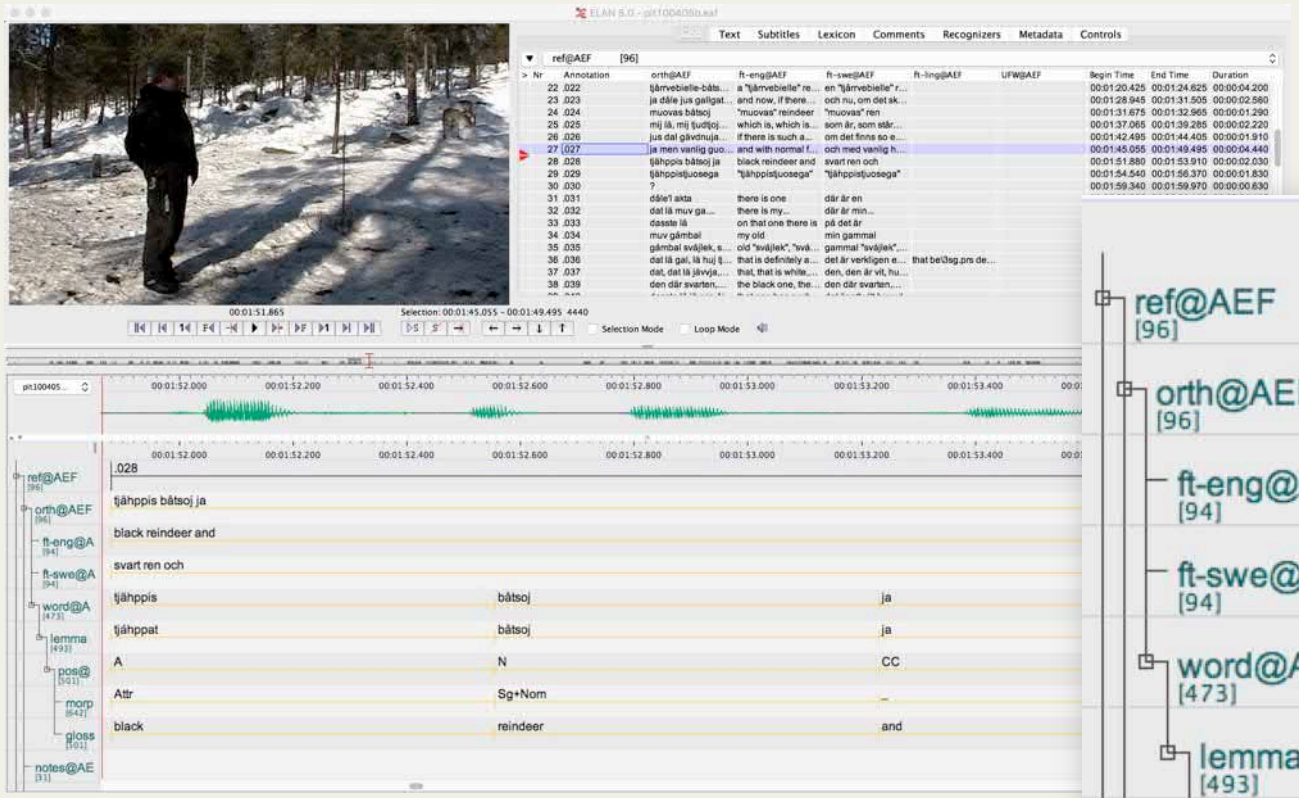
- spell-checkers
- grammar-checkers
- teaching materials (e.g. apps)

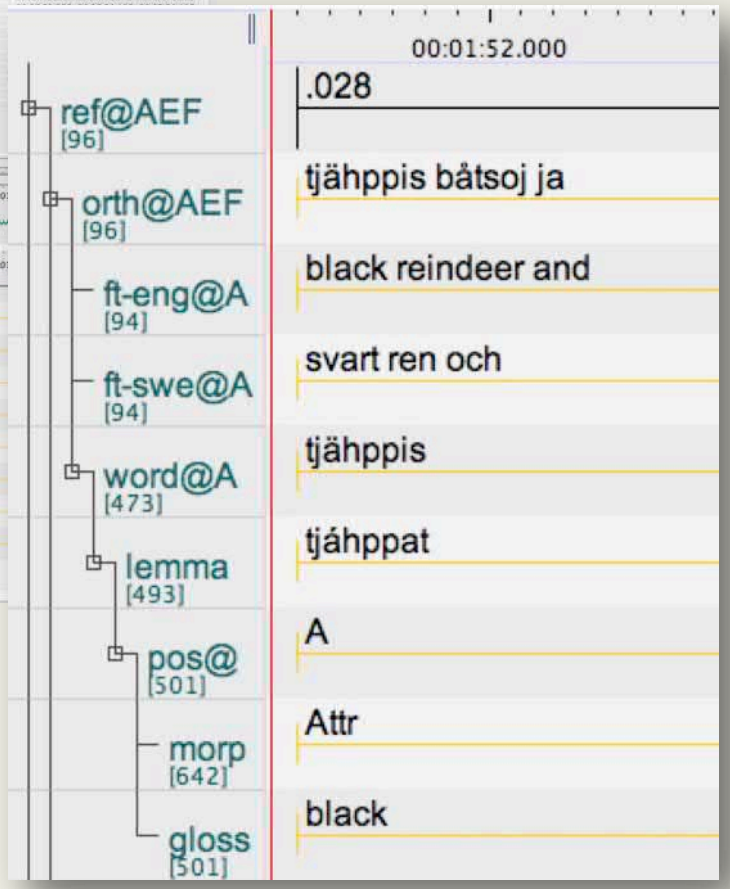
…

- in documentary linguistics / endangered language descriptions
  - automatic tokenization and annotation for corpora

further use in other, digital applications...

- tier structure in ELAN corpora (Freiburg-style)





including annotations for:

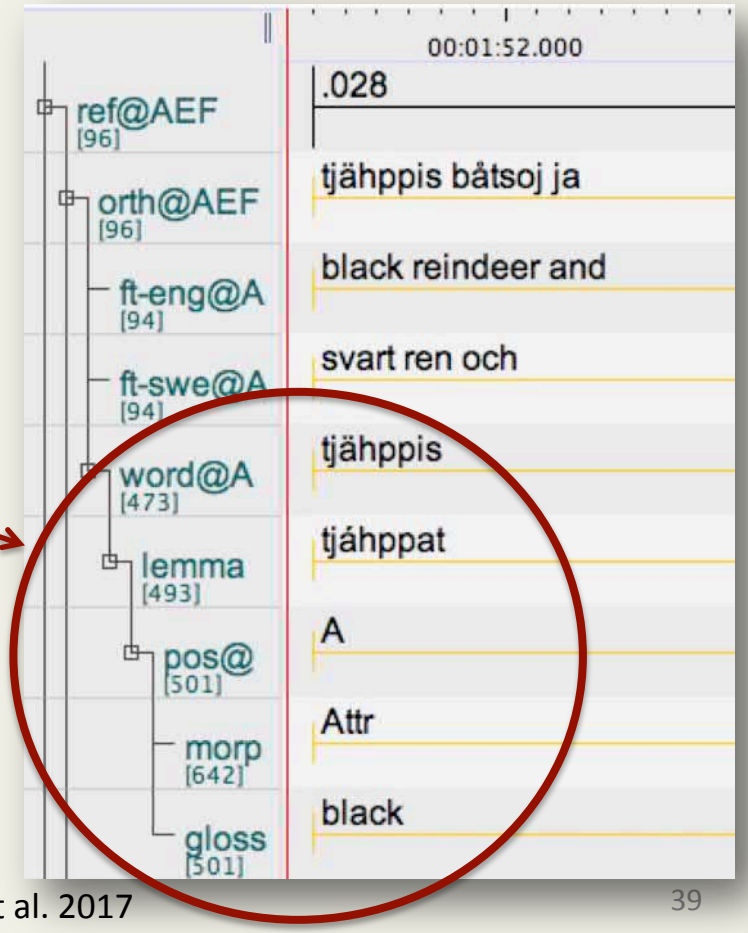
- Lemma
- Part of speech
- Morphological categories
- Gloss

further use in other, digital applications...

- tier structure in ELAN corpora (Freiburg-style)

corpus building/extension using a **script**[1] 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

More details in talk at 11:30 in room 13
by Blokland, Partanen and Rießler

[1]cf. Blokland et al. 2015; Gerstenberger et al. 2016; Gerstenberger et al. 2017

summary of digital grammaticography

requires:

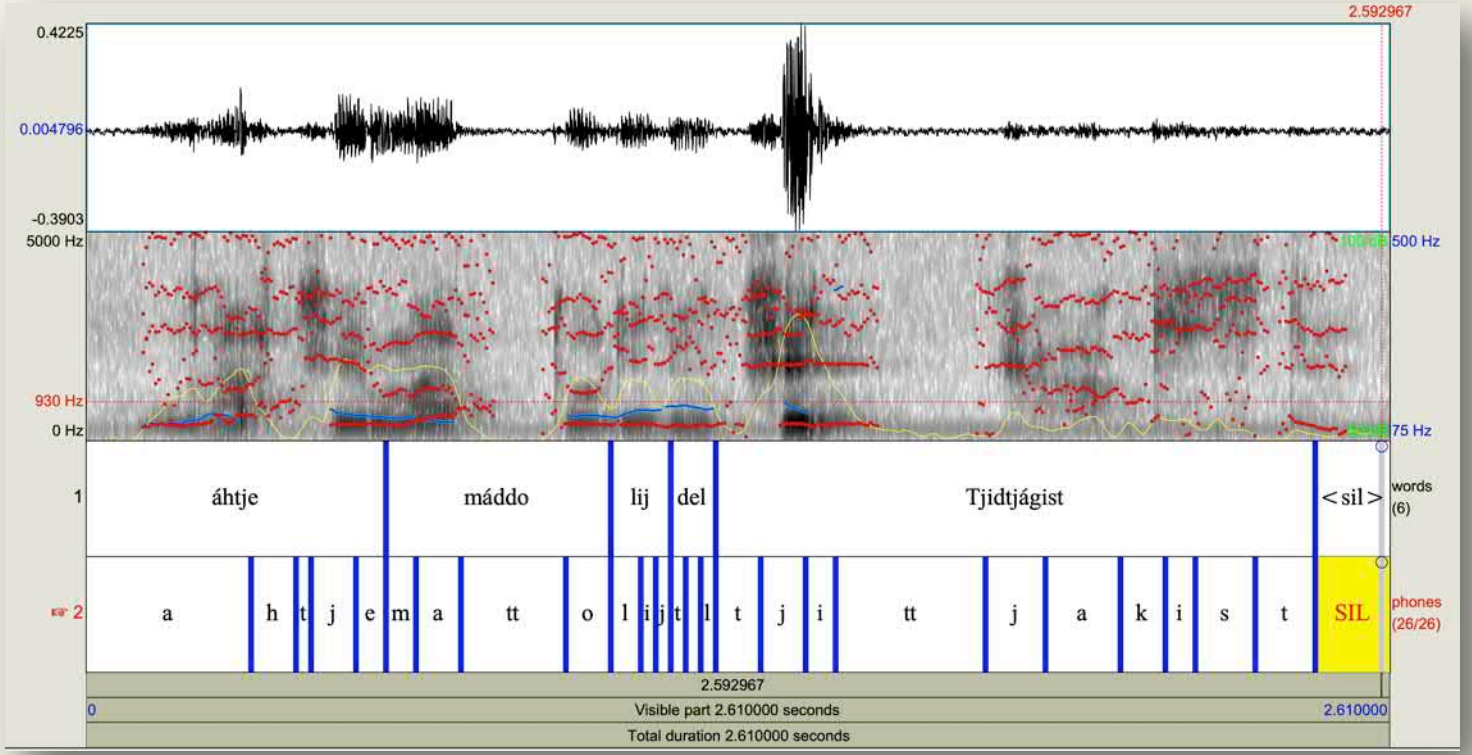
- time to learn the formalism and set up the infrastructure
- understanding of grammatical structures
- string-based representation of language

main benefits:

- can be freely accessible online
- possibility to publish (hopefully getting academic recognition, cf. LangSciPress series)
- export data for use in other tools and disciplines
    - spell-checker
    - lexicographic materials (including smart phone apps)
    - corpus building
    - teaching materials
    - increased status for the language
    - more accessible to other disciplines, e.g. via text search

main drawbacks:

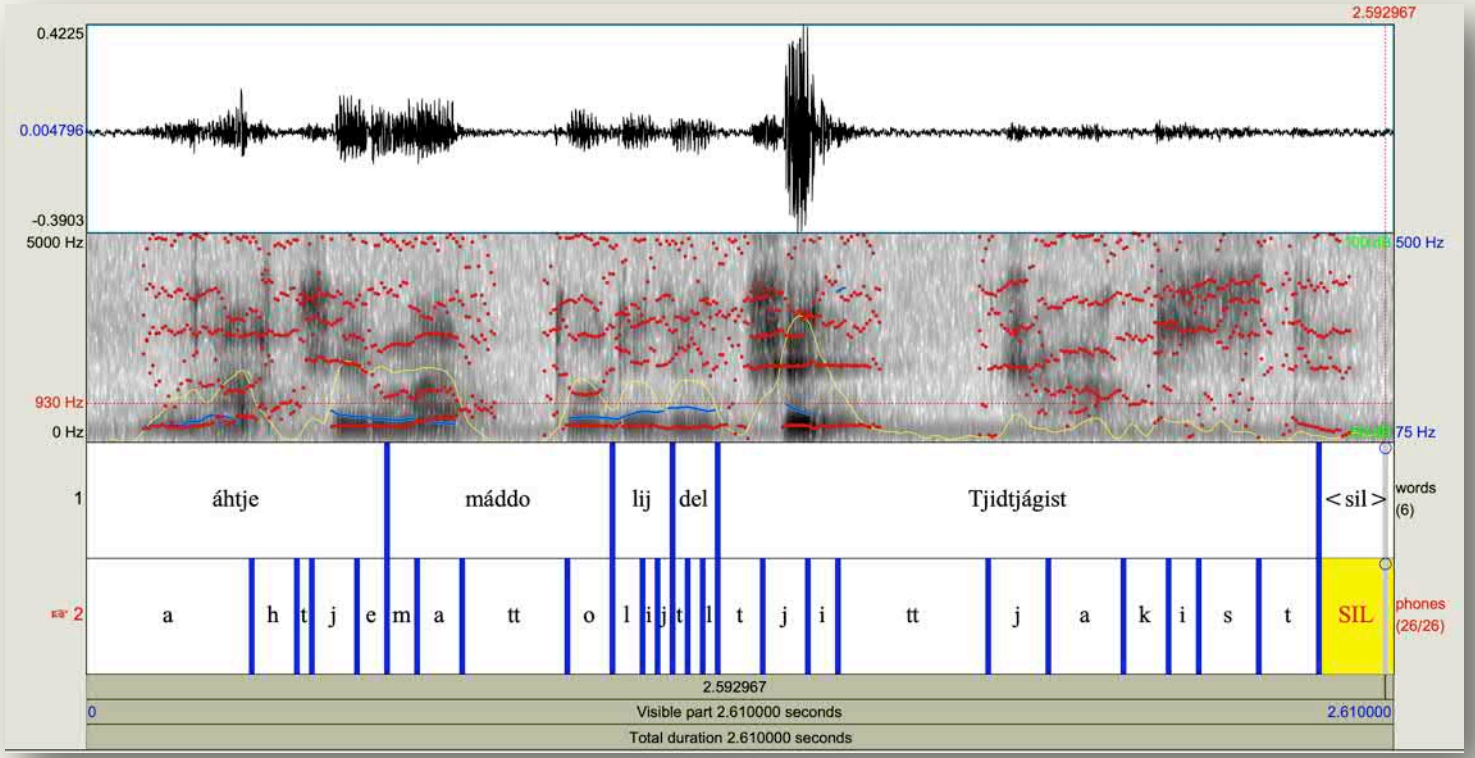
- not terribly human-accessible
- not taught traditionally in General Linguistics programs



OTHER ADVANCES
in digital technologies

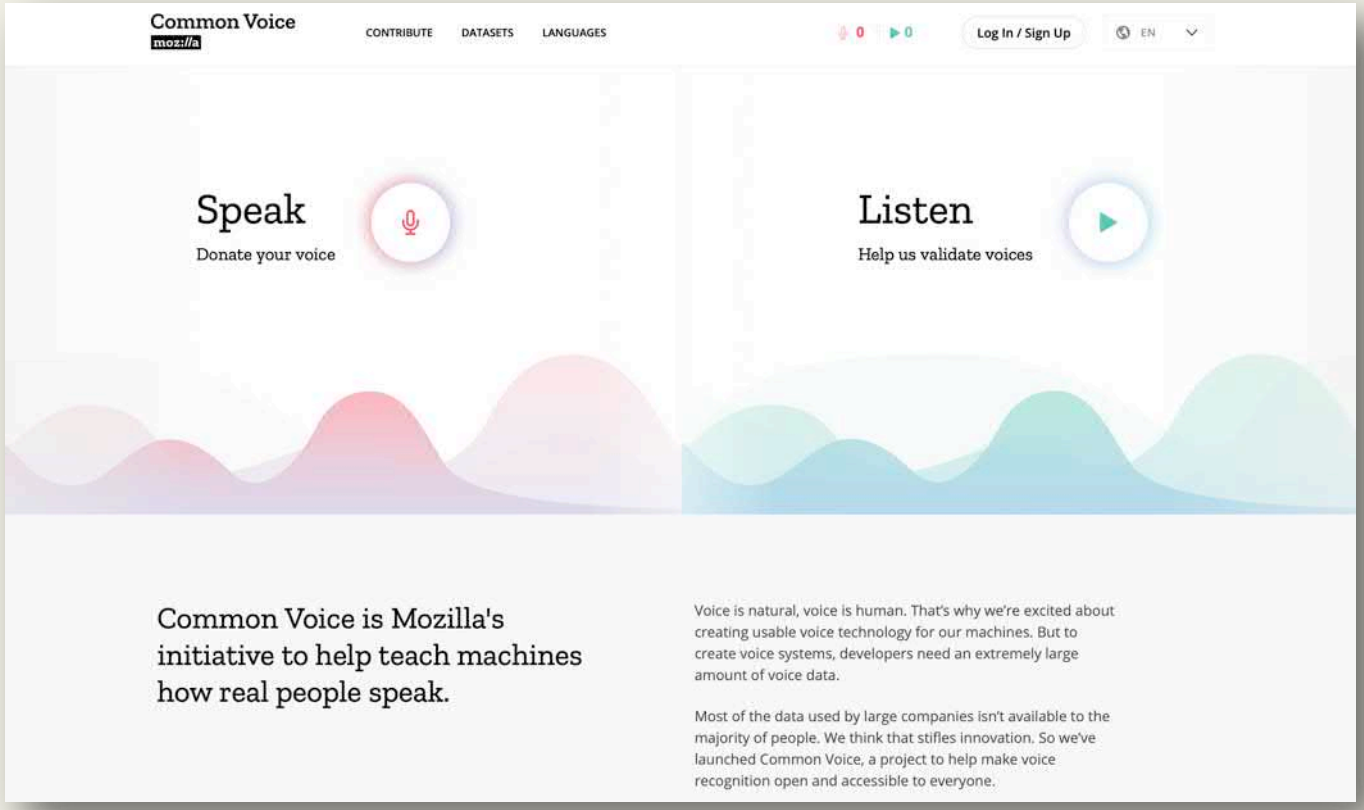
new language technologies

- automatic segmentation, e.g.:
  - Autosegmenteerija 2.0
    - Estonian autosegmentation forced-alignment tested on Pite Saami with surprisingly accurate results:



new language technologies

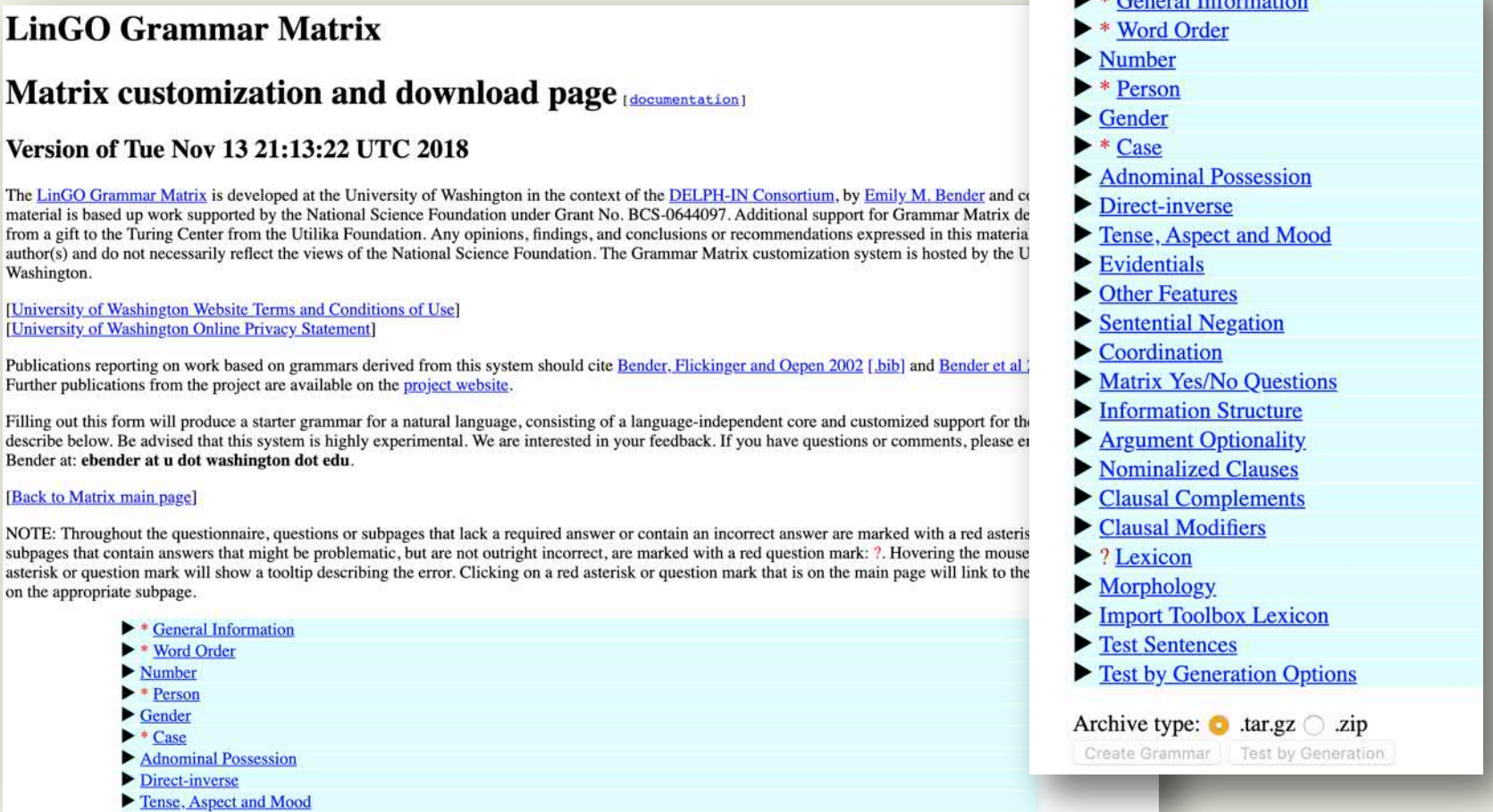
- speech recognition, e.g.:
  - Common Voice (moz://a) in community development for a number of smaller languages (e.g.: Erzya, Komi-Zyrian, ...)



new language technologies

- automatic implemented grammar production
  - LinGO Grammar Matrix

    http://matrix.ling.washington.edu/customize/matrix.cgi



LinGO Grammar Matrix

Matrix customization and download page [documentation]

Version of Tue Nov 13 21:13:22 UTC 2018

The LinGO Grammar Matrix is developed at the University of Washington in the context of the DELPH-IN Consortium, by Emily M. Bender and c
material is based up work supported by the National Science Foundation under Grant No. BCS-0644097. Additional support for Grammar Matrix de
from a gift to the Turing Center from the Utilika Foundation. Any opinions, findings, and conclusions or recommendations expressed in this materia
author(s) and do not necessarily reflect the views of the National Science Foundation. The Grammar Matrix customization system is hosted by the U
Washington.

[University of Washington Website Terms and Conditions of Use]
[University of Washington Online Privacy Statement]

Publications reporting on work based on grammars derived from this system should cite Bender, Flickinger and Oepen 2002 [.bib] and Bender et al
Further publications from the project are available on the project website.

Filling out this form will produce a starter grammar for a natural language, consisting of a language-independent core and customized support for th
describe below. Be advised that this system is highly experimental. We are interested in your feedback. If you have questions or comments, please e
Bender at: **ebender at u dot washington dot edu**.

[Back to Matrix main page]

NOTE: Throughout the questionnaire, questions or subpages that lack a required answer or contain an incorrect answer are marked with a red asteris
subpages that contain answers that might be problematic, but are not outright incorrect, are marked with a red question mark: ?. Hovering the mouse
asterisk or question mark will show a tooltip describing the error. Clicking on a red asterisk or question mark that is on the main page will link to the
on the appropriate subpage.

- * General Information
- * Word Order
- Number
- * Person
- Gender
- * Case
- Adnominal Possession
- Direct-inverse
- Tense, Aspect and Mood

- * General Information
- * Word Order
- Number
- * Person
- Gender
- * Case
- Adnominal Possession
- Direct-inverse
- Tense, Aspect and Mood
- Evidentials
- Other Features
- Sentential Negation
- Coordination
- Matrix Yes/No Questions
- Information Structure
- Argument Optionality
- Nominalized Clauses
- Clausal Complements
- Clausal Modifiers
- ? Lexicon
- Morphology
- Import Toolbox Lexicon
- Test Sentences
- Test by Generation Options

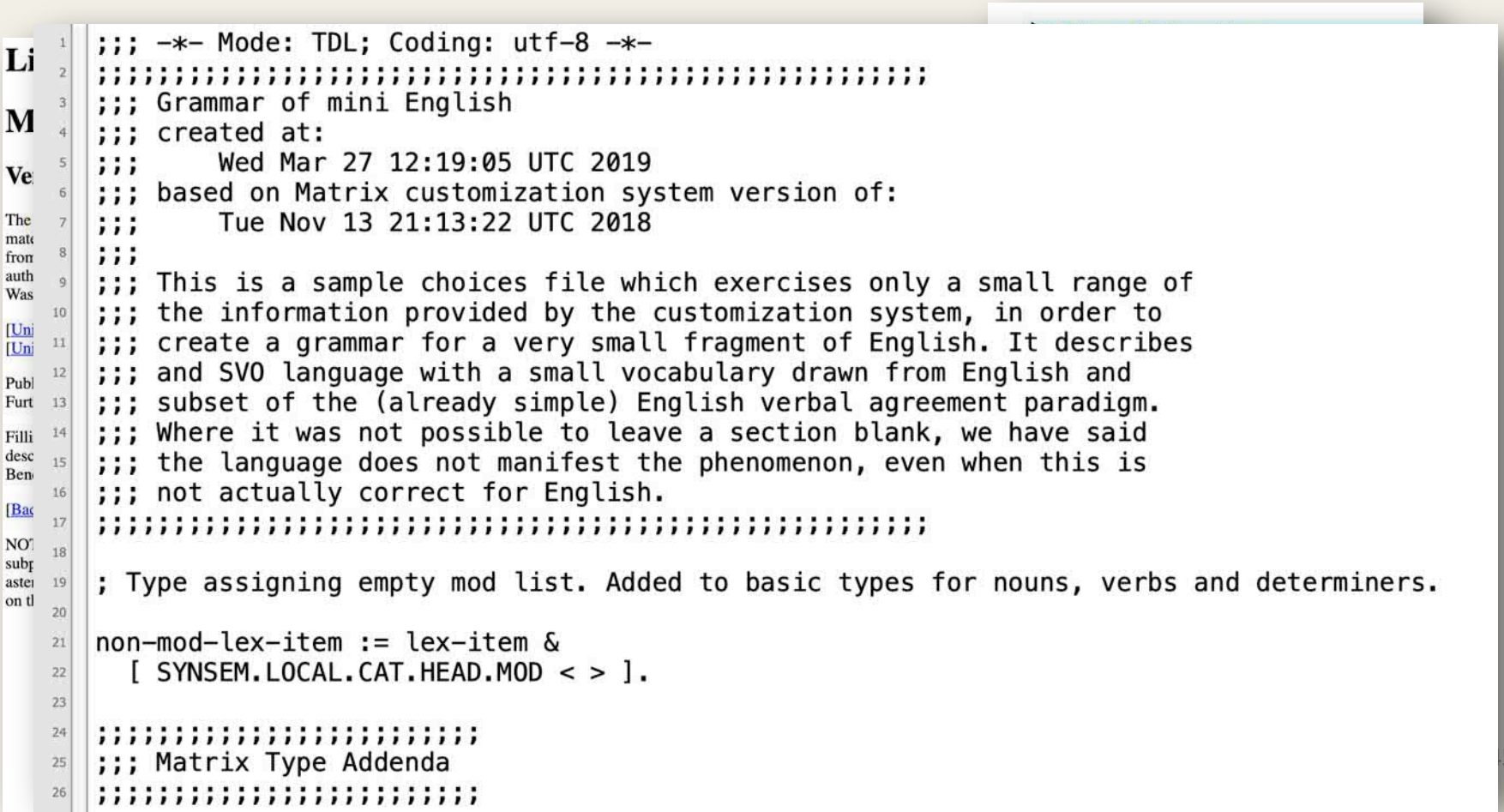
Archive type: .tar.gz .zip

Create Grammar Test by Generation

new language technologies

- automatic implemented grammar production
  - LinGO Grammar Matrix

    http://matrix.ling.washington.edu/customize/matrix.cgi

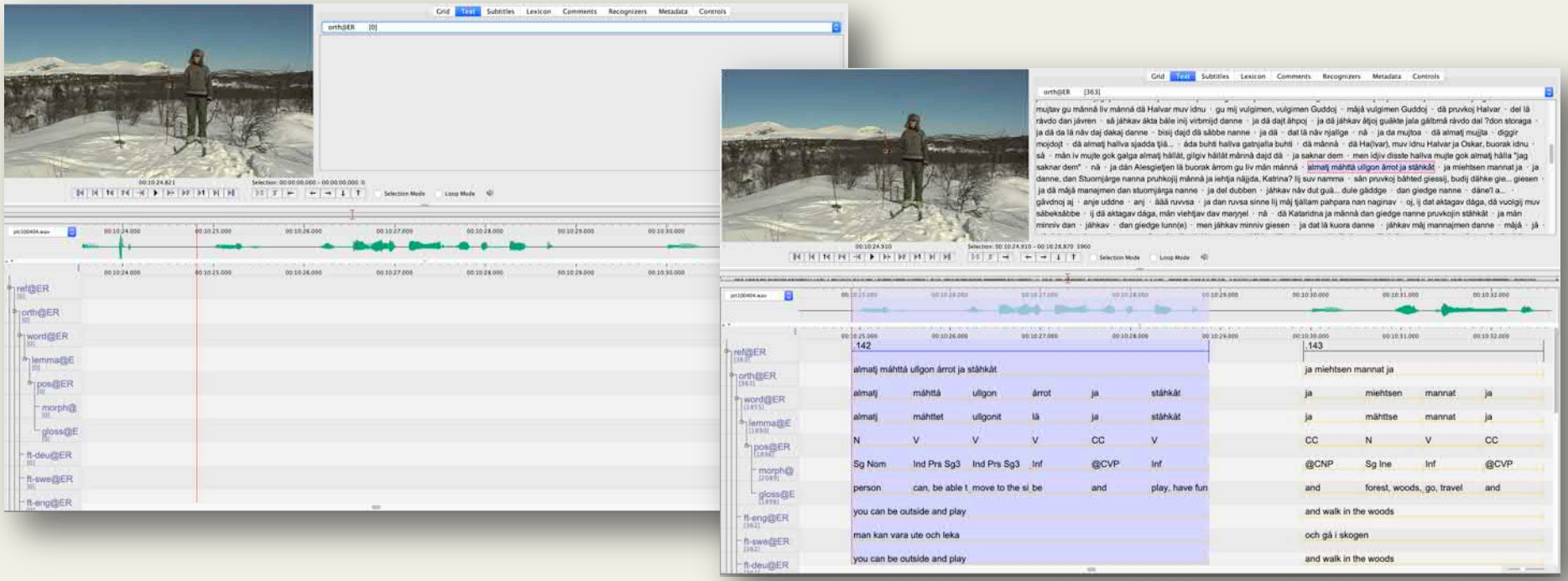


```
;;; -*- Mode: TDL; Coding: utf-8 -*-
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
;;; Grammar of mini English
;;; created at:
;;;     Wed Mar 27 12:19:05 UTC 2019
;;; based on Matrix customization system version of:
;;;     Tue Nov 13 21:13:22 UTC 2018
;;;
;;; This is a sample choices file which exercises only a small range of
;;; the information provided by the customization system, in order to
;;; create a grammar for a very small fragment of English. It describes
;;; and SVO language with a small vocabulary drawn from English and
;;; subset of the (already simple) English verbal agreement paradigm.
;;; Where it was not possible to leave a section blank, we have said
;;; the language does not manifest the phenomenon, even when this is
;;; not actually correct for English.
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

; Type assigning empty mod list. Added to basic types for nouns, verbs and determiners.

non-mod-lex-item := lex-item &
  [ SYNSEM.LOCAL.CAT.HEAD.MOD < > ].

;;;;;;;;;;;;;;;;;;;;;;;;
;;; Matrix Type Addenda
;;;;;;;;;;;;;;;;;;;;;;;;
```

new speech technologies

- relevant technologies being developed continuously
- leading to a significant increase in efficiency for corpus building

  -> better grammatical descriptions



OUTLOOK

outlook

- digital tools can provide powerful advantages for both fieldwork and (especially) grammaticography and documentation
- *but*: they require knowhow that goes beyond a typical linguist's training
- I'm not saying this is for everyone, and realistically only parts will be relevant for a few – the point is:
  Digital technologies should be considered, too!

References

Aikhenvald, Alexandra Y. (2015). *The art of grammar. A practical guide*. Oxford: Oxford University Press.

Beesley, Kenneth R. & Lauri Karttunen (2003). *Finite State Morphology*. Stanford: Center for the Study of Language and Information.

Blokland, Rogier, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, & Joshua Wilbur (2015). "Language documentation meets language technology". In: *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Proceedings of the workshop*. Ed. by Tommi A. Pirinen, Francis M. Tyers, & Trond Trosterud. Septentrio Conference Series 2015:2. Tromsø: The University Library of Tromsø, pp. 8–18.

Blokland, Rogier, Niko Partanen, Michael Rießler, & Joshua Wilbur (2019). "Using computational approaches to integrate endangered language legacy data into documentation corpora. Past experiences and challenges ahead". In: *Proceedings of the Workshop on Computational Methods for Endangered Languages*. Vol. 2. Honolulu: Association for Computational Linguistics, pp. 24–30.

Didriksen, Tino (2007–2018). *Constraint grammar manual. 3rd version of the CG formalism variant*. GrammarSoft ApS.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, & Joshua Wilbur (2016). "Utilizing language technology in the documentation of endangered Uralic languages". In: *Northern European Journal of Language Technology* 4, pp. 29–47.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, & Joshua Wilbur (2017). "Instant annotations. Applying NLP methods to the annotation of spoken language documentation corpora". In: *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2017)*. Ed. by Tommi A. Pirinen, Michael Rießler, Trond Trosterud, & Francis M. Tyers. St. Petersburg: Association for Computational Linguistics, pp. 25–36.

Halász, Ignácz (1893). *Népköltési gyűjtemény. A Pite Lappmark arjepluogi egyházkerületéből*. Vol. 5. Svéd-Lapp Nyelv. Budapest: Magyar tudományos akadémia.

Kahle, Philip, Sebastian Colutto, Günter Hackl, & Günger Mühlberger (2017). "Transkribus. A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 04, pp. 19–24.

Karlsson, Fred (1990). "Constraint Grammar as a framework for parsing unrestricted text". In: *Proceedings of the 13th International Conference of Computational Linguistics*. Ed. by Hans Karlgren. Vol. 3. Helsinki, pp. 168–173.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, & Arto Anttila, eds. (1995). *Constraint Grammar. A language-independent system for parsing unrestricted text*. Natural Language Processing 4. Berlin: Mouton de Gruyter.

Lagercrantz, Eliel (1926). *Sprachlehre des Westlappischen nach der Mundart von Arjeplog*. Suomalais-ugrilaisen Seuran Toimituksia 55. Helsinki: Suomalais-Ugrilainen Seura.

Lehtiranta, Juhani (1992). *Arjeploginsaamen äänne- ja taivutusopin pääpiirteet*. Suomalais-ugrilaisen Seuran toimituksia 212. Helsinki: Suomalais-Ugrilainen Seura.

Mosel, Ulrike (2006). "Grammaticography. The art and craft of writing grammars". In: *Catching language. The standing challenge of grammar writing*. Ed. by Felix Ameka, Alan Dench, & Nicholas Evans. Trends in linguistics: studies and monographs 167. Berlin: Mouton de Gruyter, pp. 41–68.

Nordhoff, Sebastian (2008). "Electronic Reference Grammars for Typology: Challenges and Solutions". In: *Language Documentation and Conservation* 2.2, pp. 296–324.

Partanen, Niko & Michael Rießler (2019). "An OCR system for the Unified Northern Alphabet". In: *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2019)*. Tartu: Association for Computational Linguistics, pp. 77–89.

Payne, Thomas E. (1997). *Describing morphosyntax. A guide for field linguists*. Cambridge: Cambridge University Press.

Rießler, Michael & Joshua Wilbur (2017). "Documenting endangered oral histories of the Arctic. A proposed symbiosis for language documentation and oral history research, illustrated by Saami and Komi examples". In: *Oral history meets linguistics*. Ed. by Erich Kasten, Katja Roller, & Joshua Wilbur. Exhibitions and Symposia. Fürstenberg: Kulturstiftung Sibirien, pp. 31–64.

Ruong, Israel (1943). *Lappische Verbalableitung dargestellt auf Grundlage des Pitelappischen*. Uppsala: Almqvist och Wiksell.

Siegel, Melanie, Emily M. Bender, & Francis Bond (2016). *Jacy. An Implemented Grammar of Japanese*. CSLI Studies in Computational Linguistics. Stanford: CSLI Publications.

Sjaggo, Ann-Charlotte (2015). *Pitesamisk grammatik. en jämförande studie med lulesamiska*. Senter for samiske studiers skriftserie 20. Tromsø: Septentrio Academic Publishing.

Wilbur, Joshua (2014). *A grammar of Pite Saami*. Studies in Diversity Linguistics 5. Berlin: Language Science Press.

Wilbur, Joshua, ed. (2016). *Pitesamisk ordbok samt stavningsregler*. Samica 2. Freiburg: Albert-Ludwigs-Universität Freiburg.

Gijtov adnet!

gijtov	*adnet*
gijto-v	adne-t
thank-ACC.SG	have-PL.IMP

with special thanks to
Michael Rießler, Niko Partanen, Rogier Blokland and Ciprian Gerstenberger
for ideas, collaboration and inspiration

Joshua Wilbur
Pite Saami Syntax Project
Freiburg Research Group in Saami Studies
joshua.wilbur@skandinavistik.uni-freiburg.de