From Fieldwork Collections to Universal Dependencies

Course: NLP for Endangered Languages of the Amazon.
From a Uralic perspective.

Jack Rueter, Niko Partanen,
Mika Hämäläinen & Khalid Alnajjar

Role of fieldwork data in language technology

- Often represents endangered languages with limited resources
- Collected for various purposes:
  - Language documentation
  - Teaching
  - Grammar writing
  - Language revitalization
- Materials should be accessible at least for the language community members and researchers
- Some materials are sensitive and the access needs to be restricted, but usually not everything

Typical formats

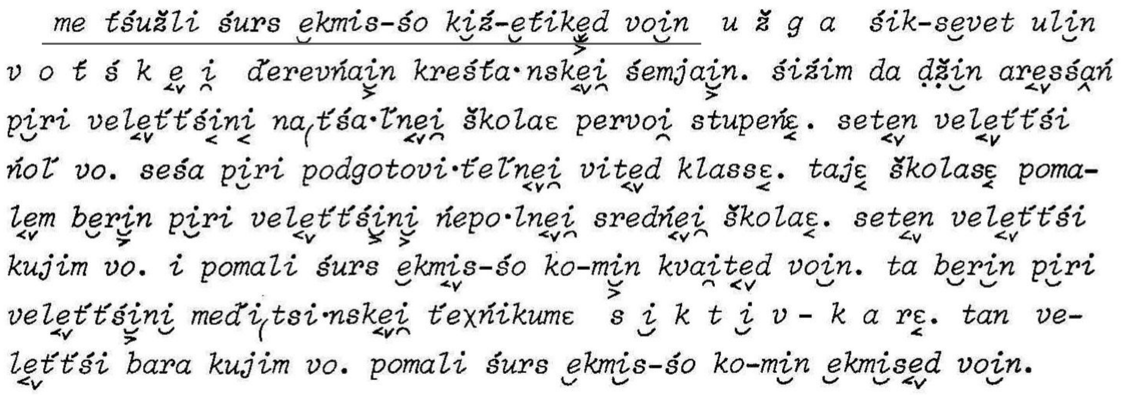
Transcriptions and translations in ELAN
Annotated texts & lexicon in FLEX
Various unpublished and published text collections, documents, notes etc.
Universal Dependencies treebanks increasingly common

The materials are usually in an onion-like structure:

- Lots of recordings that are not transcribed
- Some of the transcribed texts are translated
- Some of the texts are glossed
- A portion of those texts may be in UD

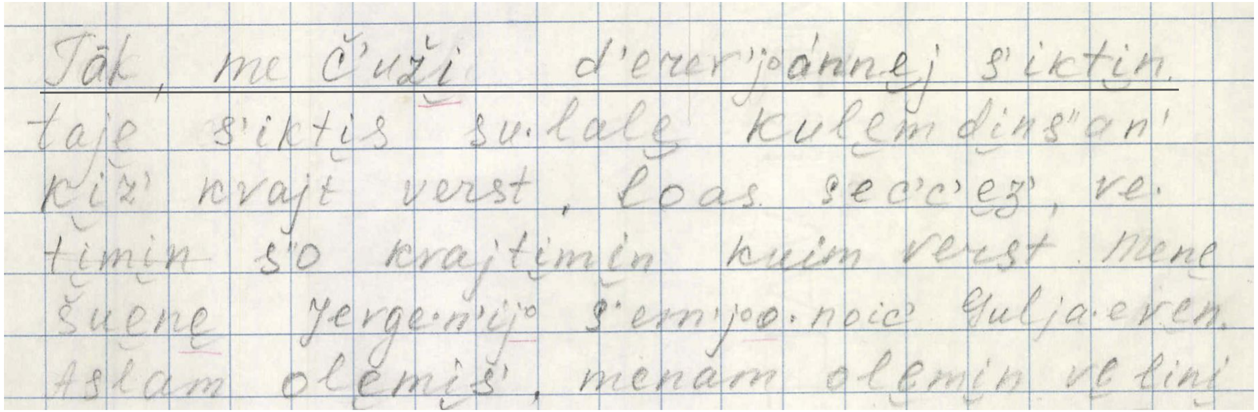
Our goal is to bring materials more effectively from the outer circle to the center

Possible approaches



me t́śužli śurs e̯kmis-śo ki̯ź-e̯tike̯d voi̯n u ž g a śik-se̯vet uli̯n v o t́ ś k e̯ i̯ đerevńai̯n kreśt́a·nske̯i̯ śemjai̯n. śižim da ḍži̯n are̯śśań pi̯ri vele̯ťťśi̯ni̯ na͡t́śa·ľne̯i̯ školaɛ pervoi̯ stupeńɛ̯. sete̯n vele̯ťťśi ńoľ vo. seśa pi̯ri podgotovi·ťeľne̯i̯ vite̯d klassɛ̯. tajɛ̯ školasɛ̯ pomale̯m beri̯n pi̯ri vele̯ťťśi̯ni̯ ńepo·lne̯i̯ sredńe̯i̯ školaɛ̯. sete̯n vele̯ťťśi kujim vo. i pomali śurs e̯kmi̯s-śo ko-mi̯n kvai̯te̯d voi̯n. ta beri̯n pi̯ri vele̯ťťśi̯ni̯ međi͡tsi·nske̯i̯ ťeχńikumɛ s i̯ k t i̯ v - k a rɛ̯. tan vele̯ťťśi bara kujim vo. pomali śurs e̯kmi̯s-śo ko-mi̯n e̯kmi̯se̯d voi̯n.

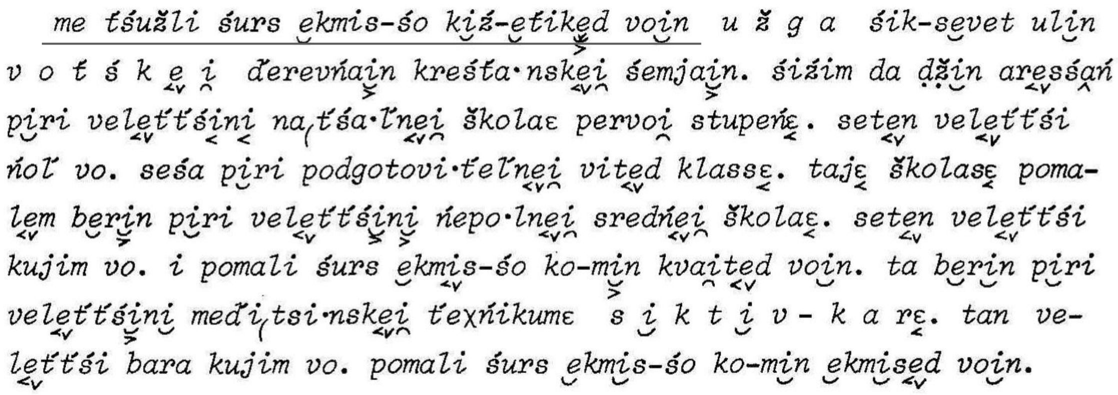
me ťśužli śurs e̯kmis-śo ki̯ź-e̯ťik̯ voi̯n…



Tāk, me č'uži d'erev'joannej s'iktin.
taje s'iktis su·lale kulemdins"an'
kiź krajt verst, loas seč'č'ez, ve·
timin s'o krajtimin kuim verst. Mene
šuene jerge·n'ijo s'em'joo·noič Guljaeven.
Aslam olemiš', menam olemin velini

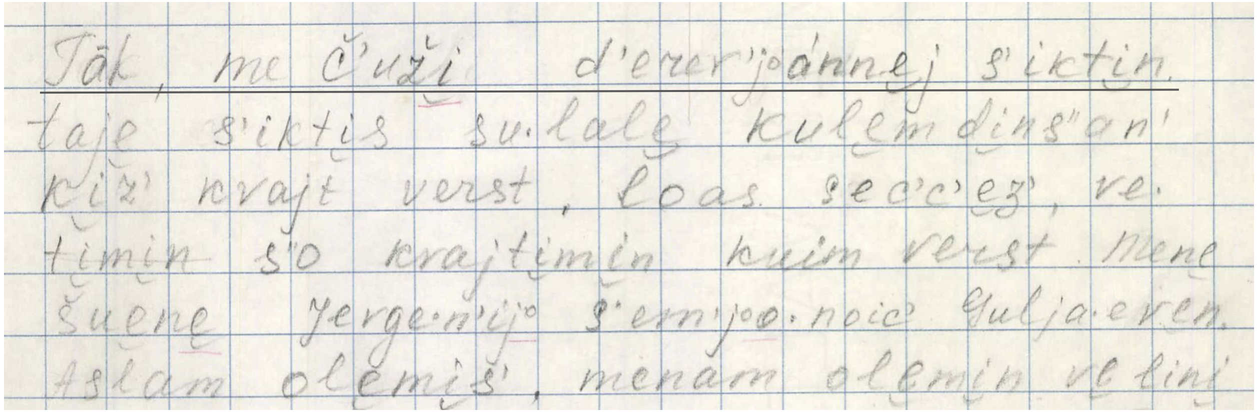
Tāk, me č'uži d'erev'jᵒannej s'ikti̯n.

Possible approaches



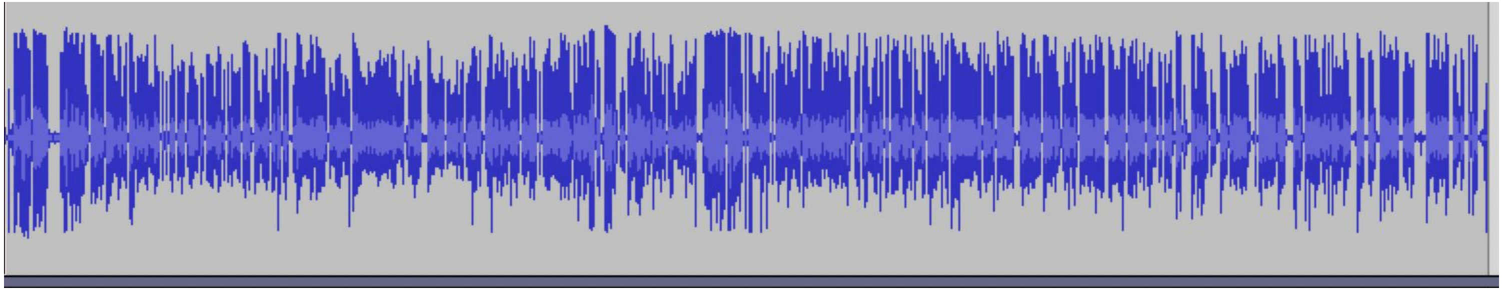
me tśužli śurs e̯kmis-śo ki̯ź-e̯tike̯d voi̯n užga śik-se̯vet uli̯n votśkei ďerevńai̯n kreśťa·nskei śemjai̯n. śiźim da dži̯n aressań pi̯ri veleťťśini na‿ťśa·ľnei školaɛ pervoi stupeńɛ. seten veleťťśi ńoľ vo. seśa pi̯ri podgotovi·ťeľnei vited klassɛ. tajɛ školasɛ pomalem be̯ri̯n pi̯ri veleťťśini ńepo·lnei sredńei školaɛ. seten veleťťśi kujim vo. i pomali śurs e̯kmis-śo ko-mi̯n kvai̯ted voi̯n. ta be̯ri̯n pi̯ri veleťťśini meďitsi·nskei ťeχńikumɛ siktiv-karɛ. tan veleťťśi bara kujim vo. pomali śurs e̯kmis-śo ko-mi̯n e̯kmise̯d voi̯n.

me t́śužli śurs e̮kmis-śo ki̮ź-e̮tik̮ voi̮n… → Ме чужлі сюрс ӧкмыс сё кызь ӧтик воын



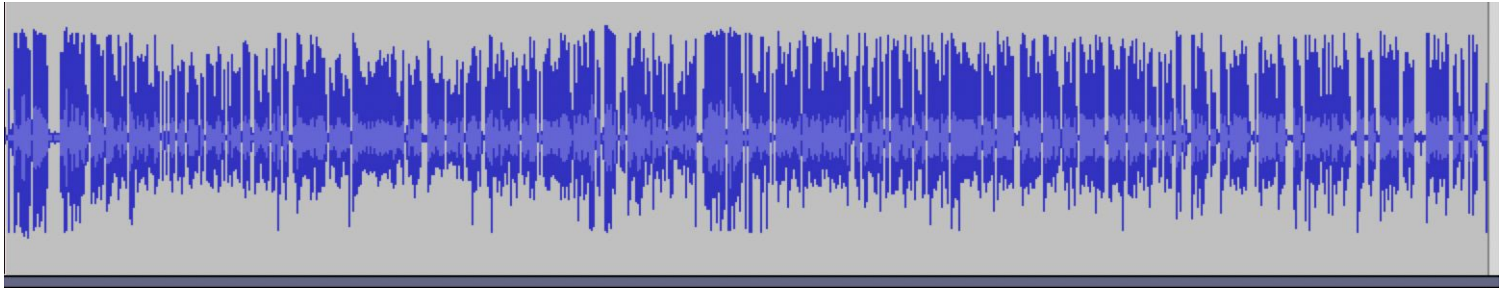
Tāk, me č'uži d'erev'joannej s'iktin.
taje s'iktis su·lale kulemdins"an'
kiz' krajt verst, loas seč'č'ez, ve-
timin s'o krajtimin kuim verst. Mene
šuene jergen'jo s'em'joo·noič Guljaeven.
Aslam olemis', menam olemin ve lini

Tāk, me č'uži d'erev'j^{o}annẹj s'iktin. → Так, ме чужи Деревяннӧй сиктын



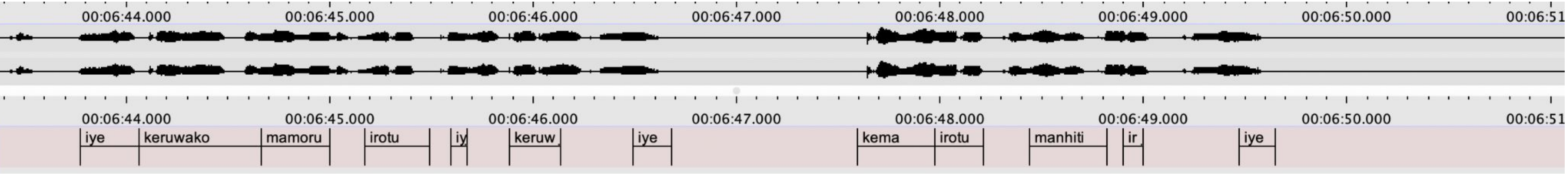
```
\ref katarok1 001
\tx Uwã  ata,      ata       keruwako iye  itokoru    ata.
\mb uwã  ata       ata       keruwako iye  itokoru    ata
\gl then we/us/our we/us/our then     then farm.field we/us/our
\ps PRT  1PL.PRON  1PL.PRON  PRT      PRT  N          1PL.PRON

\ft So, we, we... the farm field... we...
```



```
\ref katarok1 001
\tx Uwã  ata,       ata        keruwako iye  itokoru    ata.
\mb uwã  ata        ata        keruwako iye  itokoru    ata
\gl then we/us/our we/us/our then     then farm.field we/us/our
\ps PRT  1PL.PRON   1PL.PRON   PRT      PRT  N          1PL.PRON

\ft So, we, we... the farm field... we...
```



[Partanen et. al. 2020: Speech Recognition for Endangered and Extinct Samoyedic languages](#)
[Jonathan D. Amith et. al. 2021: End-to-End Automatic Speech Recognition: Its Impact on the Workflow in Documenting Yoloxóchitl Mixtec](#)
[Juho Leinonen et. al. 2021: Grapheme-Based Cross-Language Forced Alignment: Results with Uralic Languages](#)

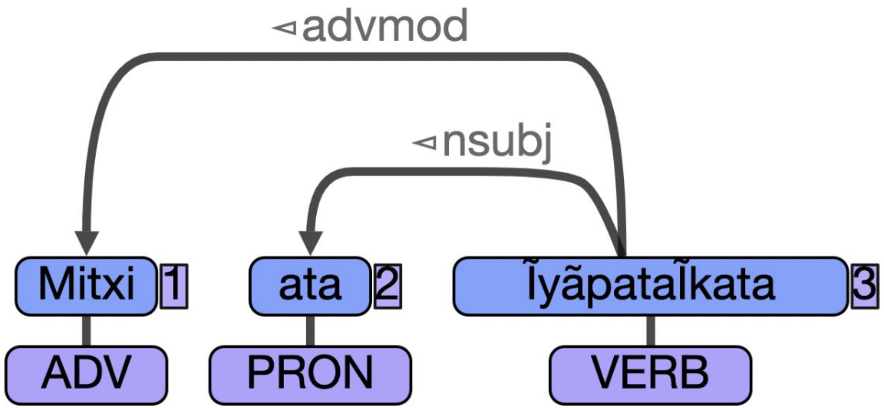
Further tasks closer to Natural Language Processing

Mitxi atha ĩyãpataĩkata

> first   we/us/our cut.down-INTENS   -VBLZ
> ADV.PRT 1PL.PRON Rt   -INTENS.SUF-VBLZ.SUF
> First we do the cutting down (of the trees and other plants)

Or predicting:

> ADV PRON VERB



Dependencies

<https://universaldependencies.org/u/dep/all.html>

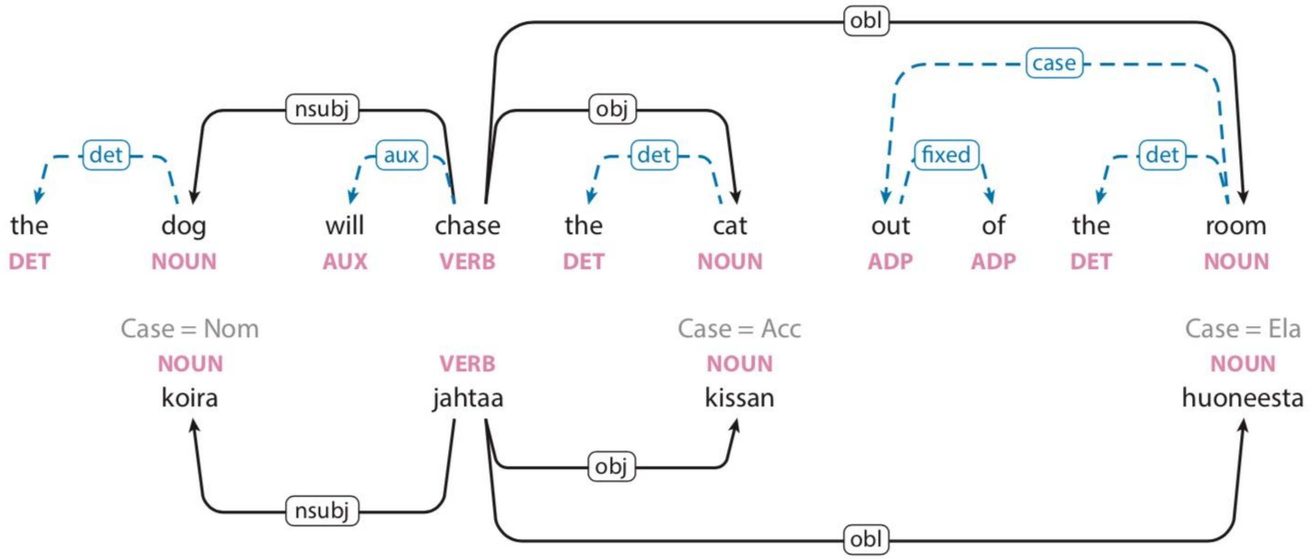


Figure from de Marneffe et. al. 2019

Interlinear glossing

tõt sluužbäi'ǧǧ leäi kuâhttlovitt ee'jj.

tõt	sluužb-äi'ǧǧ	leäi	kuâhtt-lo-vitt	ee'jj
PRON	N	V	NUM	N
DIST.SG.NOM	service-time.SG.NOM	be.PST.3SG	two-ten-five.SG.NOM	year.SG.GEN

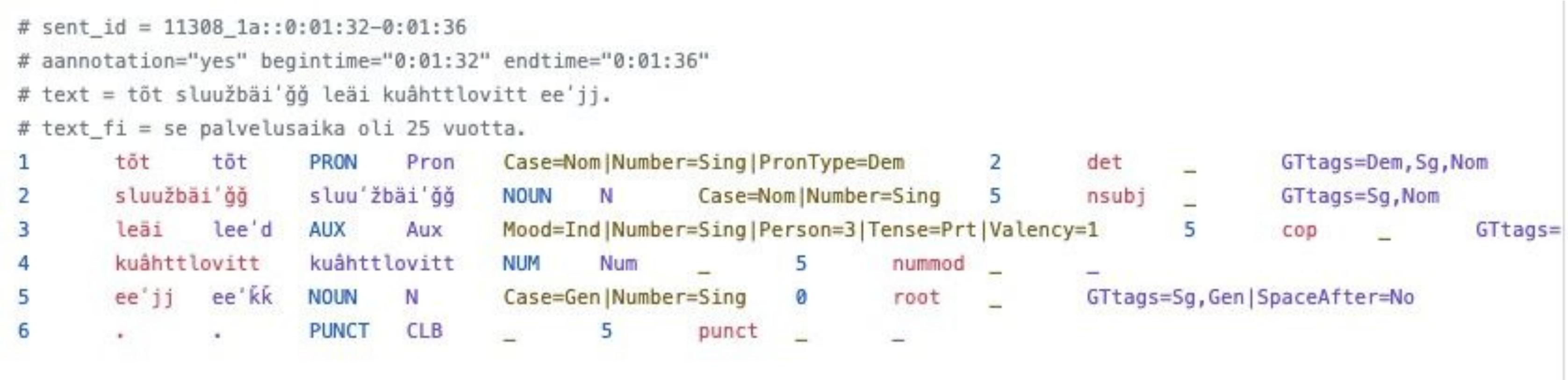
se palvelusaika oli 25 vuotta.
that service time was twenty-five years.

Interlinear glossing

tõt sluužbäi'ǧǧ leäi kuâhttlovitt ee'jj.

tõt	sluužb-äi'ǧǧ	leäi	kuâhtt-lo-vitt	ee'jj
PRON	N	V	NUM	N
DIST.SG.NOM	service-time.SG.NOM	be.PST.3SG	two-ten-five.SG.NOM	year.SG.GEN

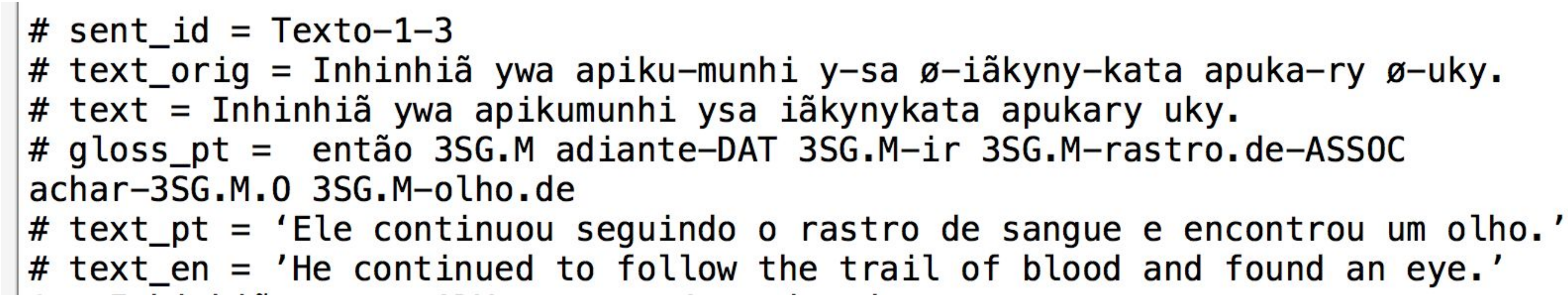
se palvelusaika oli 25 vuotta.
that service time was twenty-five years.



```
# sent_id = 11308_1a::0:01:32-0:01:36
# aannotation="yes" begintime="0:01:32" endtime="0:01:36"
# text = tõt sluužbäi'ǧǧ leäi kuâhttlovitt ee'jj.
# text_fi = se palvelusaika oli 25 vuotta.
1       tõt     tõt     PRON    Pron    Case=Nom|Number=Sing|PronType=Dem       2       det     _       GTtags=Dem,Sg,Nom
2       sluužbäi'ǧǧ     sluu'žbäi'ǧǧ    NOUN    N       Case=Nom|Number=Sing    5       nsubj   _       GTtags=Sg,Nom
3       leäi    lee'd   AUX     Aux     Mood=Ind|Number=Sing|Person=3|Tense=Prt|Valency=1       5       cop     _       GTtags=
4       kuâhttlovitt    kuâhttlovitt    NUM     Num     _       5       nummod  _       _
5       ee'jj   ee'ǩǩ   NOUN    N       Case=Gen|Number=Sing    0       root    _       GTtags=Sg,Gen|SpaceAfter=No
6       .       .       PUNCT   CLB     _       5       punct   _       _
```

Apurinã progress (new languages)

Research work going into UD



```
# sent_id = Texto-1-3
# text_orig = Inhinhiã ywa apiku-munhi y-sa ø-iãkyny-kata apuka-ry ø-uky.
# text = Inhinhiã ywa apikumunhi ysa iãkynykata apukary uky.
# gloss_pt =  então 3SG.M adiante-DAT 3SG.M-ir 3SG.M-rastro.de-ASSOC
achar-3SG.M.O 3SG.M-olho.de
# text_pt = ‘Ele continuou seguindo o rastro de sangue e encontrou um olho.’
# text_en = ’He continued to follow the trail of blood and found an eye.’
```

Since December 2019

Link for homework and exploring

https://github.com/rueter/nlp-for-endangered-languages

New paper about predicting interlinear glosses:

Diego Barriga Martínez et. al 2021: Automatic Interlinear Glossing for Otomi language

AmericasNLP workshop is generally worth checking out:

http://turing.iimas.unam.mx/americasnlp/program.html

Obrigado