



# Challenges in OCR today

Report on experiences from INEL

Niko Partanen

17/3/2017



|       |       |           |
|-------|-------|-----------|
| 1710» | 1710» | 1.8~      |
| 1711» | 1711» | 0.245455~ |
| 1712» | 1712» | 1.8~      |
| 1713» | 1713» | 1.05~     |
| 1714» | 1714» | 0.283333~ |
| 1715» | 1715» | 1.8~      |
| 1716» | 1716» | 3.92754~  |
| 1717» | 1717» | 0.20625~  |
| 1718» | 1718» | 0.211364~ |
| 1719» | 1719» | 0.267391~ |
| 1720» | 1720» | 0.585606~ |
| 1721» | 1721» | -0.3~     |
| 1722» | 1721» | -0.3~     |
| 1723» | 1721» | 1.8~      |
| 1724» | 1722» | 0.562558~ |
| 1725» | 1723» | 0.221739~ |
| 1726» | 1724» | 1.8~      |
| 1727» | 1725» | 3.81136~  |
| 1728» | 1726» | 1.8~      |
| 1729» | 1727» | 0.259459~ |
| 1730» | 1728» | 0.294545~ |

[hunalign](#)



```

\ref title:folklordolgan2000-30 page:328 section:1 par:1 sentence:1~
\tx Bîlîr ogonn'ordo:k eme:ksin hogotokko:n olorbutter, biek ogoto huōktar.~
\litv Былыр огонньордоок эмээксин hogotokkoон олорбуттар, биэк огото huоктар.~
\litr В старину жили одинокие старик со старухой, век-то не имели ребенка.~
\ref title:folklordolgan2000-30 page:328 section:1 par:1 sentence:2~
\tx Bi:rde turan ogonn'or eppit eme:ksiniger:~
\litv Биирдэ туран огонньор эппит эмээксинигэр:~
\litr Однажды старик говорит своей старухе:~
\ref title:folklordolgan2000-30 page:328 section:1 par:1 sentence:3~
\tx – Eme:ksin, bihigi ogolonuōkputugar bert.~
\litv – Эмээксин, биһиги оголонуокпутугар бэрт.~
\litr – Старуха, нам бы хорошо завести своего ребенка.~
\par~
\ref title:folklordolgan2000-30 page:328 section:1 par:2 sentence:4~
\tx Onuga eme:ksine eppit:~
\litv Онуга эмээксинэ эппит:~
\litr На это старуха сказала;~
\par~

```



EXMARaDA Panther Editor 3.5.2 [C:\Users\fon960\Documents\Doğan\Corpus\NLP\PoMA\_1964\FoxDeceiver\_PoMA\_1964\FoxDeceiver\_Rus.bib]

File Edit View Transcription Tier Event Timeline Format Help

12 [00:1] 13 [00:10.5] 14 [00:11.0] 15 [00:11.5] 16 [00:12.0] 17 [00:12.5] 18 [00:13.0] 19 [00:13.5] 20 [00:14.0] 21 [00:14.5] 22 [00:15.0] 23 [00:15.5] 24 [00:16.0] 25 [00:16.5] 26 [00:17.0] 27 [00:17.5]

|     |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
|-----|----------------------------------|---------------|-------------------|---------------|----------------------------------|----------------|-------------|---------------|---|--------------|-------------------|-----------------|--|---------|-------------------|---------------|
| ref | PoMA_1964_FoxDeceiver_Rk.001.003 |               |                   |               | PoMA_1964_FoxDeceiver_Rk.001.004 |                |             |               | PoMA_1964_FoxDeceiver_Rk.001.005            |              |                   |                 | PoMA_1964_FoxDeceiver_Rk.001.006               |         |                   |               |
| st  | Бу ого ачыктаан өлөрө булар.     |               |                   |               | Буруу хаһыл булар.               |                |             |               | — Мин эвентен элтин ооруюм, — деп.          |              |                   |                 | — Баай элтин булугор, баай драктары да         |         |                   |               |
| ts  | Bu ogo açtıktaan ölörö buolar.   |               |                   |               | Bunu haşıl bular.                |                |             |               | —Min en'gin kihi oğurūm,— dır,ş             |              |                   |                 | —Baş kihi bulūog, baş d'aktarı d'aktarı ilan   |         |                   |               |
| tx  | Bu                               | ogo           | açtıktaan         | ölörö         | buolar.                          | Bunu           | haşıl       | bular.        | —Min  | en'gin       | kihi              | oğurūm,—        | dır,ş  | —Baş    | kihi              | bulūog,       |
| nb  | ba                               | ogo           | açtıkta: a        | ölör-ö        | buol- a- r                       | ba- m          | haşıl       | bul- a- r     | min   | en'gi- n     | kihi              | oğur- uo- m     | dır- r   | baş     | kihi              | bul- uo- g    |
| wp  | ba                               | ogo           | açtıkta: a        | ölör- a       | buol- a- r                       | ba- nı         | haşıl       | bul- a- r     | min   | en- n        | kihi              | oğur- lık- m    | dır- r   | baş     | kihi              | bul- lık- g   |
| eo  | this                             | child.[NOM]   | hunger-CVB        | kill-CVB.SIM  | be-EP-3SG                        | this-ACC       | fox.[NOM]   | find-EP-3SG   | 1SG.[NOM]                                   | 2SG-ACC      | human.being.[NOM] | make-FUT-1SG    | say.PRS-3SG                                    | rich    | human.being.[NOM] | become-FUT-2  |
| et  | этот                             | ребёнок.[NOM] | проголодаться-CVB | убить-CVB.SIM | быть-EP-3SG                      | этот-ACC       | лиса.[NOM]  | найти-EP-3SG  | 1SG.[NOM]                                   | 2SG-ACC      | человек.[NOM]     | сделать-FUT-1SG | говорить.PRS-3SG                               | богатый | человек.[NOM]     | стать-FUT-2   |
| id  | dieres                           | kind-[NOM]    | Hunger.haben-CVB  | öten-CVB.SIM  | sein-EP-3SG                      | dieres-ACC     | Fuchs.[NOM] | finden-EP-3SG | 1SG.[NOM]                                   | 2SG-ACC      | Mensch.[NOM]      | machen-FUT-1SG  | sagen.PRS-3SG                                  | reich   | Mensch.[NOM]      | werden-FUT-2  |
| ln  |                                  | 1             | 2                 | 3             | 2-2                              |                | 1           | 2-2           | 1-1   | 1            | 2-1               | 3-2             | 1-2  | 2       | 2-1               | 3-2           |
| nc  | dempro                           | n-accuse      | v-verb            | v-verb        | v-v(in)-v-pr                     | dempro-proxase | n-accuse    | v-v(in)-v-pr  | petr-proxase                                | petr-proxase | n-accuse          | v-vtense-v-pr   | v-v-pr   | adj     | n-accuse          | v-vtense-v-pr |
| ps  | dempro                           | n             | v                 | v             | v                                | dempro         | n           | v             | petr  | petr         | n                 | v               | v  | adj     | n                 | esp           |
| SeR |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| SyF |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| ISI |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| Tsp |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| Foc |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| fe  |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |
| fg  | Dieser Junge starb vor Hunger.   |               |                   |               | Ein Fuchs fand ihn.              |                |             |               | "Ich mache dich wieder lebendig", sagte er. |              |                   |                 | "Du wirst ein reicher Mensch, ich verheirate d |         |                   |               |
| fr  | Этот мальчик от голода пропал.   |               |                   |               | Нашла его лиса.                  |                |             |               | — Я тебя человеком сделаю, — сказала.       |              |                   |                 | — Станешь богатым, женю тебя на богато         |         |                   |               |
| lv  | Этот мальчик от голода пропал.   |               |                   |               | Нашла его лиса.                  |                |             |               | — Я тебя человеком сделаю, — говорит.       |              |                   |                 | — Станешь богатым, женю тебя на богато         |         |                   |               |
| st  |                                  |               |                   |               |                                  |                |             |               |   |              |                   |                 |  |         |                   |               |

Done

[10:22:47] Panther Editor started

Segmentation: GEMRIC Player: DirectShow Player

10:22 14.03.2017



| Kamas (33 results) |                                     |                       |         |  |       |   |       |
|--------------------|-------------------------------------|-----------------------|---------|--|-------|---|-------|
| RegEx (T)          |                                     | Search:               |         | \b(M m)ān\b                                      |       |   |       |
| #                  | S                                   | Communication         | Speaker | Left Context                                     | Match | Right Context                                   | id[S] |
| 1                  | <input checked="" type="checkbox"/> | AA_1914_Mouse_flk     | AA      | a pada?, ija!" "Kalla? kăštəga?, kola pūla:...   | Mān   | kola ej amənzəliem. Let i?gō, kia?dolam."...    | AA    |
| 2                  | <input checked="" type="checkbox"/> | AA_1914_Mouse_flk     | AA      | "Šuktəlbiba?, šo?!" Kalla? kăštəgut, šuktə...    | Mān   | dīgən igem." Ne kubi, pa kabarbi, di t'uga...   | AA    |
| 3                  | <input checked="" type="checkbox"/> | AA_1914_Khan_flk      | AA      | aba?bi: "Šində dīgən amna?" Inebə ot't'erl...    | mān   | kūk no?bə tonəlla?bə ej surarga. To?gən...      | AA    |
| 4                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | Nūke amnobi, o?b n'it ibi. N'it ijaandə mā...    | Mān   | nejlēm." Ijat mālia: "Naga, man'əkkən ne ...    | AA    |
| 5                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | nejlēm." Ijat mālia: "Naga, man'əkkən ne ...     | Mān   | kallam ne pile?." Kalla? t'ūrbi, ijabə ba?la... | AA    |
| 6                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | də. "Ad'am! Aspa? edə?, uja pada?!" "Uj...       | Mān   | tetliem." N'i?də u?bdəbi, t'alaš šūškū ibi. ... | AA    |
| 7                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | N'e?le?bliet "I? t'ora?!" "Tān aspa?lə ko:l...   | mān   | ujam amnu?bi." Ne: "I? t'ora?! Miliem. O...     | AA    |
| 8                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | ularzəngən'ibe?! Ō?le it ularlə!" "Ši? ularza... | mān   | ularbə amnu?lədən." "İmbijle? amnədən? ...      | AA    |
| 9                  | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | ot's'ün Gūd'ər nūnne? ba?bi, u?la sa?məb...      | Mān   | nōrbəbiöm ši?n'ida, ularbə amnədən. Šān...      | AA    |
| 10                 | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | ābəj?le?bddə. Di bazo? šünə tirlölē. "I? t'...   | Mān   | ularəm milem." "Māna ej kere? tān ularla?...    | AA    |
| 11                 | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | ?! Mān ularəm milem." "Māna ej kere? tān...      | Mān   | imbin'im māna mibi." "No, i? t'ora?! Bar ...    | AA    |
| 12                 | <input checked="" type="checkbox"/> | AA_1914_Corpse_flk    | AA      | a?də Büz'e amna, šide ko?bdot amna. "A...        | Mān   | nükem ige. Pan tabəndə ularzəndə t'abola...     | AA    |
| 13                 | <input checked="" type="checkbox"/> | IA_1912_Trickster_flk | IA      | P'el simabə kajlaba?bi. "Tān kădə? mola?...      | Mān   | simam t'ezerdələm. Ō?le? māna, tān simal...     | IA    |
| 14                 | <input checked="" type="checkbox"/> | IA_1912_Trickster_flk | IA      | əbi. Dizeŋ šobii?. "Kădə? mola? š'a:mna? ...     | Mān   | ej š'a:mniām. Mān sāgər kuza igem." Ibii?...    | IA    |
| 15                 | <input checked="" type="checkbox"/> | IA_1912_Trickster_flk | IA      | "Kădə? mola? š'a:mna? i?bəl?" "Mān ej š...       | Mān   | sāgər kuza igem." Ibii? idəm sogənə sarbi...    | IA    |
| 16                 | <input checked="" type="checkbox"/> | NN_1914_Birds_flk     | NN      | əššət igel. Nen pi?meendə tülində siktölē? ...   | Mān   | toru inen toru kunda tora to?la? kunnim."       | NN    |

uke amnobi, o?b n'it ibi. N'it ijaandə mālia: "Mān nejlēm." Ijat mālia: "Naga, man'əkkən ne naga." **Mān** kallam ne pile?." Kalla? t'ūrbi, ijabə ba?la:mibi. Kandəga, kandəga. T'āgan to?gəndə ma? nuga. Šūbi

id[S] AA

## Lament [song]

Speaker: Avdakeya Andzhigatova

Language: Kamass [xas]

Recorded in 1914 by Kai Donner

Published in: A.J. Joki (ed.). 1944. *Kai Donners Kamassisches Wörterbuch*. Helsinki: Suomalais-Ugrilainen Seura. P. 87.

This publication is part of the [INEL project](#), 2016-

Glossed by Gerson Klumpp and Tiina Klooster

English translation by Gerson Klumpp

Kijen mille?biem sagər mājazaŋbə iʔbəle? kojobi.

gʃen mʲn-laʔbə-bi-m sagər mʲʃa-zAŋ-m iʔbə-lA? koʃo-bi  
where go-DUR-PST-1SG black.[NOM.SG] mountain-PL-NOM/GEN/ACC.1SG lie-CVB stay-PST.[3SG]

Where I was going, my black mountains stayed behind

Mille?bəne t'ūm kūk noʔts'əʔ özerle? baʔbi.

mʲn-laʔbə-NTA t'ū-m kūk noʔ-t-xiʔ özer-lA? baʔbə-bi  
go-DUR-PTCP land NOM/GEN/ACC.1SG green.[NOM.SG] grass NOM/GEN.3SG-INS grow-CVB throw-PST.[3SG]

The land which I walked was grown over with green grass

Sagər mājazaŋbə iʔbəle? kojobi, siri t'ālamzaŋbə maʔla? kojobiiʔ.

sagər mʲʃa-zAŋ-m iʔbə-lA? koʃo-bi siri d'elam-zAŋ-m  
black.[NOM.SG] mountain-PL-NOM/GEN/ACC.1SG lie-CVB stay-PST.[3SG] white.[NOM.SG] Sayan.mountains-PL-NOM/GEN/ACC.1SG

ma-lA? koʃo-bi-ʃəʔ

stay-CVB stay-PST-3PL

My black mountains stayed behind, my white Sayan mountains were left behind.



# Paradox with ABBYY

## Desktop version

- Good user interface
- Training new models easy, although not transparent
- Practical to do fast post-correction after OCR
- **No XML export**

## Engine version

- Used from command line
- Only pre-defined models
- Cannot be post-corrected in Abbyy Desktop
- Good ALTO XML export





## How is this possible?

- In defence of ABBYY, other OCR tools suffer with same
- Maybe it simply is difficult to manage user edits and word/letter position coordinates?

## Currently lost information

- Coordinate information on page
- Some formatting
- Footnotes are later added manually as notes
- Time is wasted!



# INEL

- Grammars, corpora and language technologies for Indigenous Northern Eurasian Languages ([website](#))
- A long-term project (18 years) funded by Academy of Sciences and Humanities in Hamburg
- The project was applied for by Prof. Dr. Beáta Wagner-Nagy, Dr. Michael Riessler and the management of HZSK
- Research:
  - Institut für Finnougristik / Uralistik at Universität Hamburg (IFUU)
- Infrastructure:
  - Hamburger Zentrum für Sprachkorpora ([HZSK](#))



# Why is this a problem? (1/2)

- Reconstructing paragraphs from plain text very unreliable
  - Word coordinate on page gives more cues
- Distinguishing different numberings from one another hard
  - Page numbers, chapter numbers etc.
- Accurate page information nice for citations



## Why is this a problem? (2/2)

- No chances to deal with more complex layout in the document
- Hard to make nice ebooks automatically!
  - Ebooks with broken paragraph structure annoying
- Hard to make nice digital facsimiles!
  - Relevant for older and rarer items
- Unnecessary information loss is never desirable!

# Future tasks

## Combining text and coordinates

- Text files in some collections are nicely corrected
- XML files contain coordinate info
  - It must be possible to do matching between these two files?

## Combining corpora for research purposes

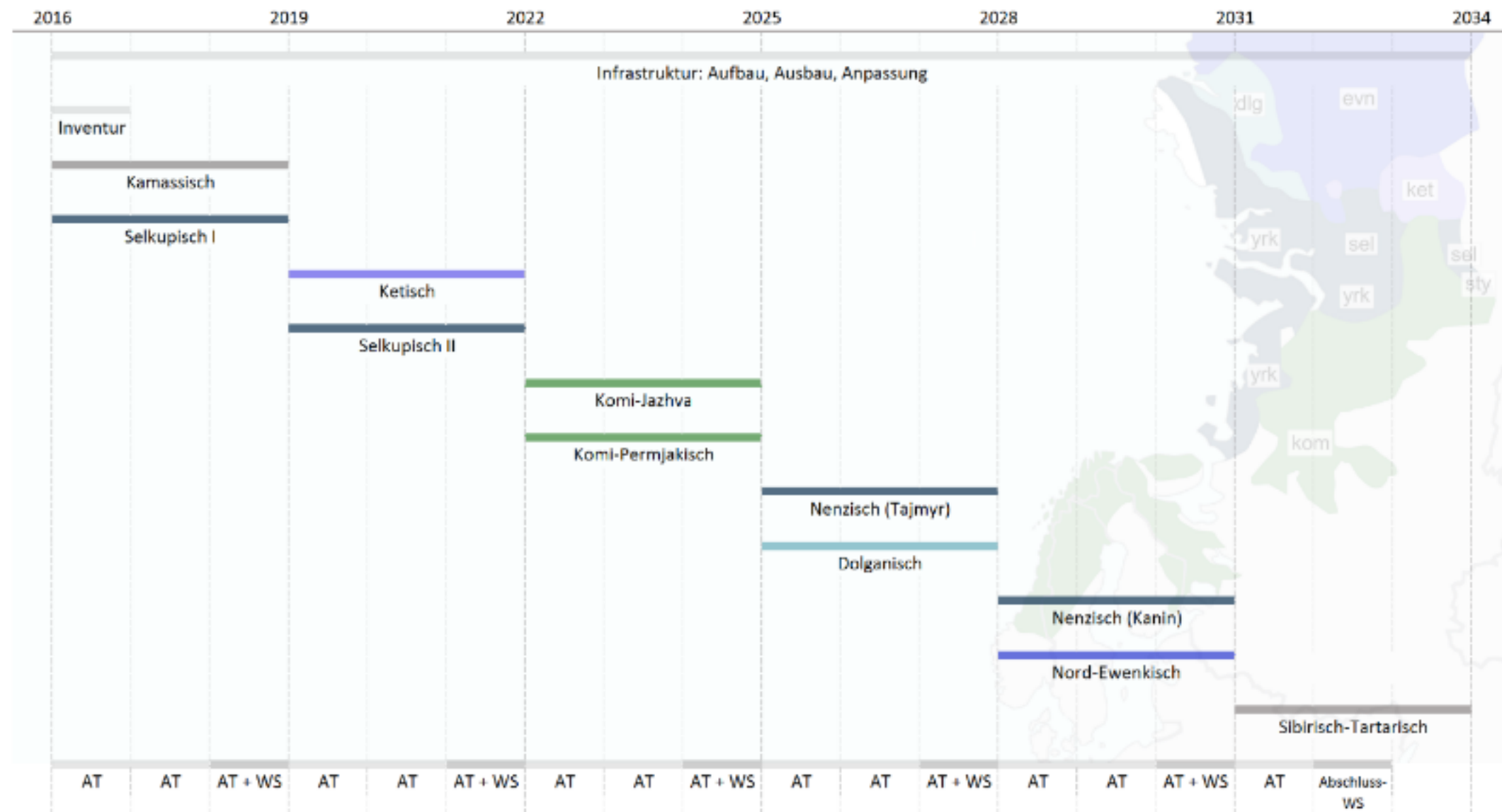
- Same speakers and writers (Albert Vanejev, Jevgeni Igušev)

Аттьö! Thank you! Спасибо!

CC-BY – Niko Partanen / INEL – 2017

More information and publication [here](#).





# Data first phase

- Kamas
  - Kamassisches Wörterbuch (Donner, 1944), glossing complete
  - Audio from last speaker, currently being transcribed
- Selkup
  - A.I. Kuzmina's archive at IFUU: field notes, recordings...
- Dolgan
  - Texts from existing publications (folklore)
  - Audio recordings from different sources





# Peculiarities of Dolgan case

- Variety of transcription systems
  - Transliterated to INEL conventions
- Some printed publications rather new
  - High print quality, easy OCR
- Sentence-level translation to Russian
  - Need for aligning
  - Russian translations often in internet already

# Basics of INEL OCR workflow

## Tools used

- We have been using ABBYY Finereader
- First goal is to bring texts into FLEx with metadata
  - Toolbox file as interchange format
- Audio is aligned later in EXMARaLDA
- Git is used as version control across work phases
  - Most relevant in EXMARaLDA stage
- Hunalign has been practical in alignment checking



## Overview to the current workflow



