Challenges in OCR today

Report on experiences from INEL

Niko Partanen

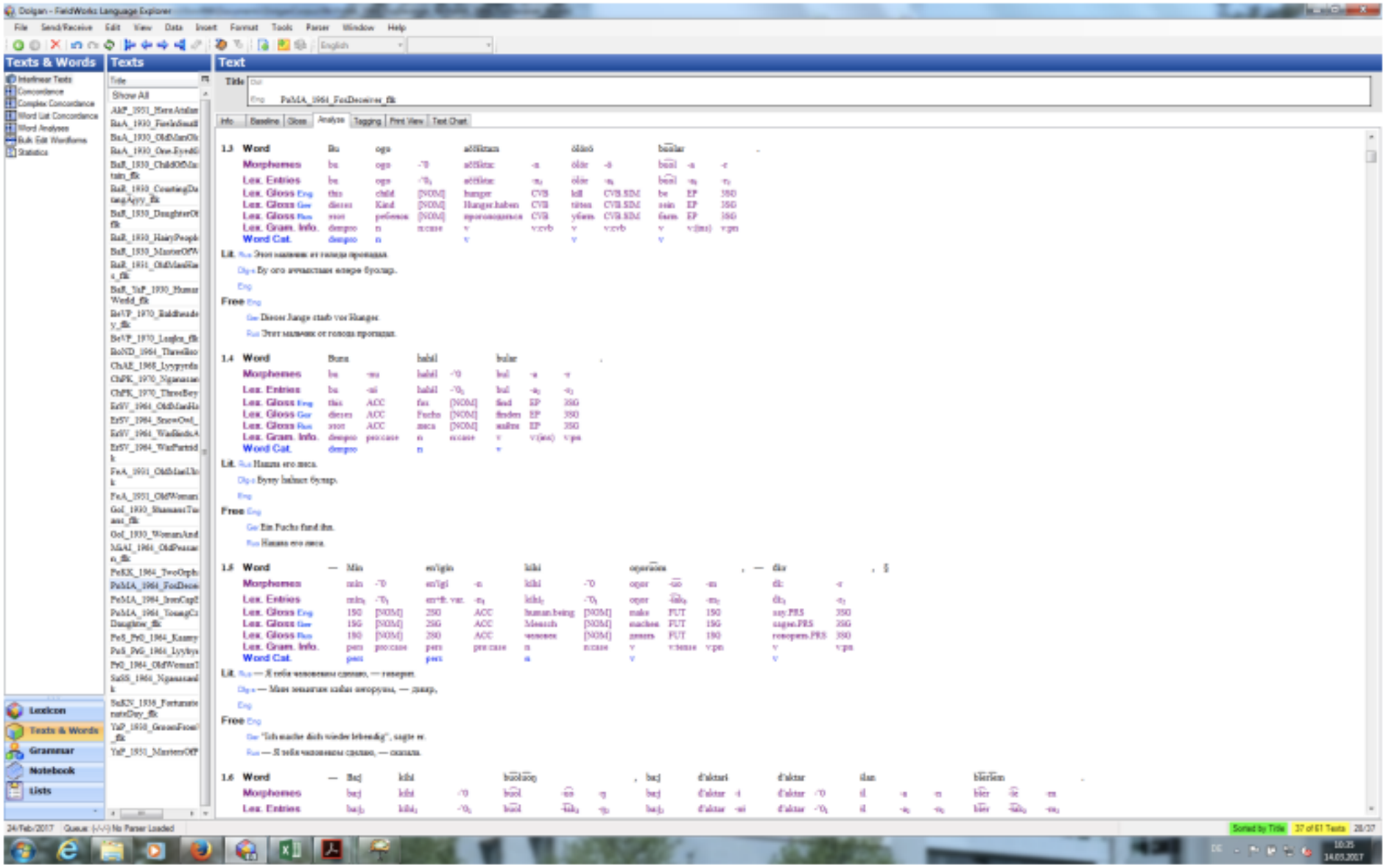
17/3/2017

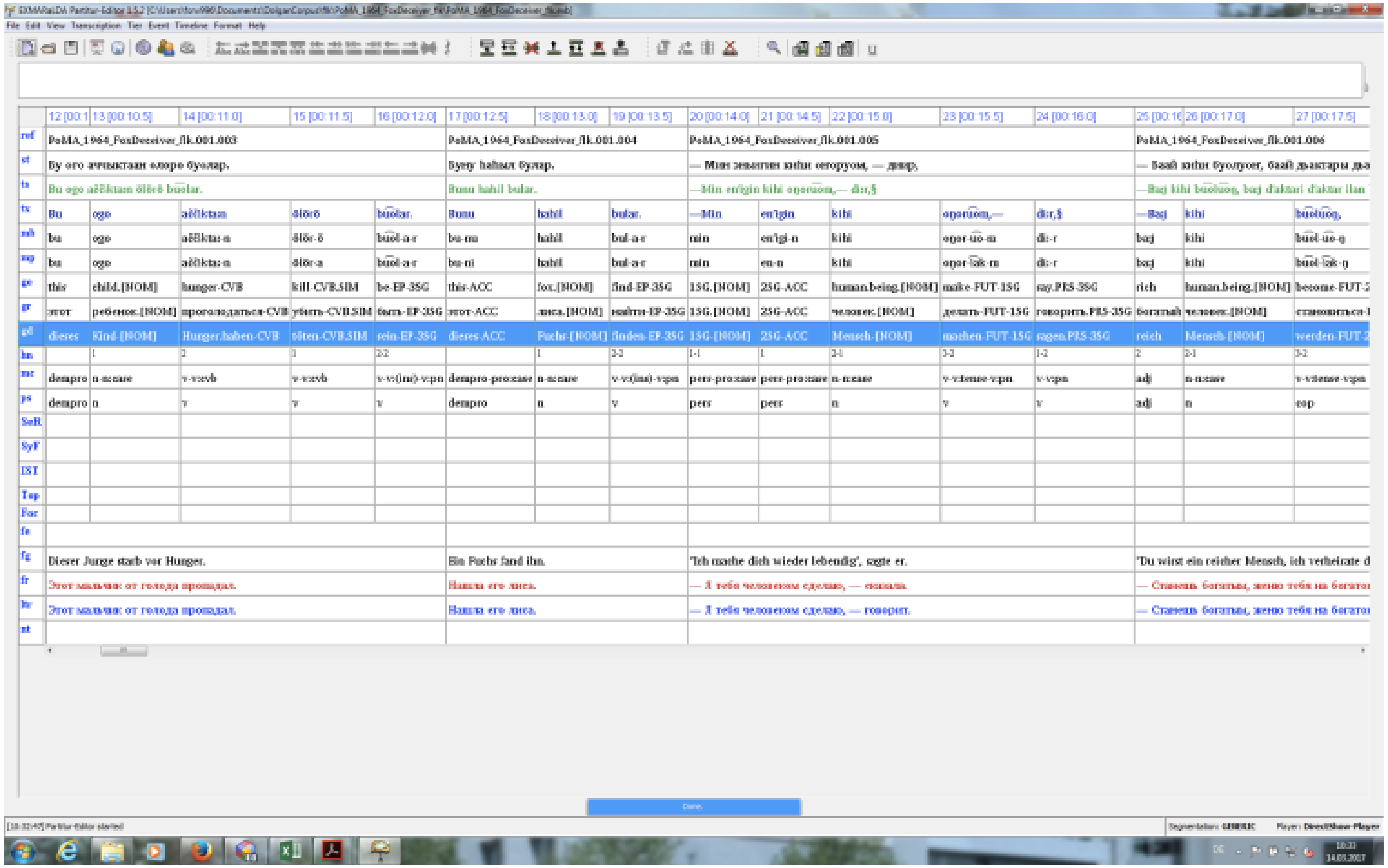
Былыр огонньордоок эмээксин һоготоккоон олорбуттар, биэк оҕото һуоктар. 1	В старину жили одинокие старик со старухой, век-то не имели ребенка. 1
Биирдэ туран огонньор эппит эмээксинигэр:	Однажды старик говорит своей старухе:
paragraph_break	paragraph_break
— Эмээксин, биһиги оголонуокпутугар бэрт.	— Старуха, нам бы хорошо завести своего ребенка.
paragraph_break	paragraph_break
Онуга эмээксинэ эппит:	На это старуха сказала;
paragraph_break	paragraph_break
— Һиккибитигэр оголоммотокпут, кайдак билигин оголонуок-путуй?	— Когда мы были молодыми, не появился у нас ребенок, как сейчас родим его?
Ол һатаныа дуо? — диэбит.	Сможем ли мы?
paragraph_break	paragraph_break
Ону огонньоро эппит:	Тут старик сказал:
paragraph_break	paragraph_break
— Мин киһи буларыттан атынык оголоноору гынабын, — диэбит.	— Я хочу завести ребенка не так, как у людей бывает.
paragraph_break	paragraph_break
378	x
Эмээксинэ диэбит:	Старуха сказала:
paragraph_break	paragraph_break
— Оголонорго туок куһагана буолуой?!	— Разве худо обзавестись ребенком?!
Оннук буоллагына тэриннэгин дии.	Раз так, то ты постарайся.
paragraph_break	paragraph_break
Маныага огонньоро илими ылбыт, кабыйакаанка туруорарларын курдук кытылга $	Тогда старик взял сеть, поставил ее как на куропаток, навесив на колья от $
Бу гынан баран: илиминиһигэр киирэн эккирии-эккирии: "Ohoo-эһээ!" — диэн $	Сделав так, вошел внутрь сети и, подпрыгивая на месте, заголосил: "Осоо-эс$
Огурдук үс күн быһыланар.	Вот так три дня делал.
Үһүс күнүгэр таксыбыта, илимигэр уол оҕо түбэһэн ытыы һытар.	На третий день вышел — в сеть мальчик маленький попался, лежит и плачет.
Маны огонньор дьиэтигэр эгэлэр.	Старик его домой принес.
paragraph_break	paragraph_break
Дьиэгэ киирбиттэрэ бу оголоро чороокута* да эмэһэтэ дээ бүтэй. 3	Внес в дом, [посмотрел], а у этого ребенка ни спереди, ни сзади отверстия $
Үөрбүт дьон маны дьэ матана курдук каалаан аһаталлар.	Обрадованные люди стали кормить его, словно суму [кожаную] набили.
Бүтэй киһи кайдак буолуой — һонно өлөн каалбыт.	Как же будет жить цельный человек, не имеющий отверстия? Тут же и умер.
paragraph_break	paragraph_break
Оннук кайа иччititтэн ого кэлэн барбыт диэн былыргылар кэпсииллэр.	Вот что рассказывали предки о том, как от духа горы был получен ребенок, н$
paragraph_break	paragraph_break
330	330

```
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 Bɨlɨr ogonn’ordo:k eme:ksin hogotokko:n olorbuttar, biek ogoto huo͡ktar.
\litv Былыр огонньордоок эмээксин һоготоккоон олорбуттар, биэк огото һуоктар.
\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 ogolonu͡okputugar 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
```


Dolgan – FieldWorks Language Explorer
File Send/Receive Edit View Data Insert Format Tools Parser Window Help
Texts & Words
Interlinear Texts
Concordance
Complex Concordance
Word List Concordance
Word Analyses
Bulk Edit Wordforms
Statistics
Lexicon
Texts & Words
Grammar
Notebook
Lists
Texts
Title
Show All
Text
Title
Info Baseline Gloss Analysis Tagging Print View Text Chart
Word
Morphemes
Lex. Entries
Lex. Gloss Eng
Lex. Gloss Ger
Lex. Gloss Rus
Lex. Gram. Info.
Word Cat.
Lit.
Free
Eng
Ger
Rus
24/Feb/2017 Queue: (-/-/-) No Parser Loaded
Sorted by Title 37 of 61 Texts 28/37


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 [00:1 26 [00:17.0] 27 [00:17.5]
ref PoMA_1964_FoxDeceiver_flk.001.003 PoMA_1964_FoxDeceiver_flk.001.004 PoMA_1964_FoxDeceiver_flk.001.005 PoMA_1964_FoxDeceiver_flk.001.006
st Бу ого ачыктаан өлөрө буолар. Буну һаһыл булар.
ts Bu ogo aččiktan ölörö buolar. Bunu hahil bular. —Min en'igin kihi oŋorüom,— diːr,§
tx Bu ogo aččiktan ölörö buolar. Bunu hahil bular. —Min en'igin kihi oŋorüom,— diːr,§ —Baj kihi buoluoŋ,
mb bu ogo aččiktaː-n ölör-ö buol-a-r bu-nu hahil bul-a-r min en'igi-n kihi oŋor-üo-m diː-r baj kihi buol-üo-ŋ
mp bu ogo aččiktaː-n ölör-a buol-a-r bu-nI hahil bul-a-r min en-n kihi oŋor-IAK-m diː-r baj kihi buol-IAK-ŋ
ge 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
gr этот ребенок.[NOM] проголодаться-CVB убить-CVB.SIM быть-EP-3SG этот-ACC лиса.[NOM] найти-EP-3SG 1SG.[NOM] 2SG-ACC человек.[NOM] делать-FUT-1SG говорить.PRS-3SG богатый человек.[NOM] становиться-
gd dieses Kind-[NOM] Hunger.haben-CVB töten-CVB.SIM sein-EP-3SG dieses-ACC Fuchs-[NOM] finden-EP-3SG 1SG-[NOM] 2SG-ACC Mensch-[NOM] machen-FUT-1SG sagen.PRS-3SG reich Mensch-[NOM] werden-FUT-2
mc dempro n-n:case v-v:cvb v-v:cvb v-v:(ins)-v:pn dempro-pro:case n-n:case v-v:(ins)-v:pn pers-pro:case pers-pro:case n-n:case v-v:tense-v:pn v-v:pn adj n-n:case v-v:tense-v:pn
ps dempro n v v v dempro n v pers pers n v v adj n cop
SeR
SyF
IST
Top
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 пропадал. Нашла его лиса. — Я тебя человеком сделаю, — сказала. — Станешь богатым, женю тебя на богато
ltr Этот мальчик от голода пропадал. Нашла его лиса. — Я тебя человеком сделаю, — говорит. — Станешь богатым, женю тебя на богато
nt
Done.
Segmentation: GENERIC Player: DirectShow-Player

Kamas (33 results)

RegEx (T) | Search: | \b(M|m)ăn\b

#	S	Communication	Speaker	Left Context	Match	Right Context	id[S]
1	☑	AA_1914_Mouse_flk	AA	a pada?, ija!" "Kalla? kăštəga?, kola pŭla:...	Măn	kola ej amənzəliem. Let iʔgŏ, kiaʔdolam."...	AA
2	☑	AA_1914_Mouse_flk	AA	"Šuktəlbiba?, šo?!" Kalla? kăštəgut, šuktə...	Măn	dĭgən igem." Ne kubi, pa kabarbi, dĭ t'uga...	AA
3	☑	AA_1914_Khan_flk	AA	aba?bi: "Šində dĭgən amna?" Inebə öt't'erl...	măn	kŭk noʔbə tonəlaʔbə ej surarga. Toʔgən...	AA
4	☑	AA_1914_Corpse_flk	AA	Nŭke amnobi, oʔb n'it ibi. N'it ijaandə mă...	Măn	nejləm." Ijat măliа: "Naga, man'əkkən ne ...	AA
5	☑	AA_1914_Corpse_flk	AA	nejləm." Ijat măliа: "Naga, man'əkkən ne ...	Măn	kallam ne pileʔ." Kallaʔ t'ŭrbi, ijabə baʔla...	AA
6	☑	AA_1914_Corpse_flk	AA	də. "Ad'am! Aspaʔ edəʔ, uja padaʔ!" "Uj...	Măn	tetliem." N'iʔdə uʔbdəbi, t'alaš šŭškŭ ibi. ...	AA
7	☑	AA_1914_Corpse_flk	AA	N'eʔleʔbliet "Iʔ t'oraʔ!" "Tăn aspaʔlə ko:l...	măn	ujam amnuʔbi." Ne: "Iʔ t'oraʔ! Mĭliem. O...	AA
8	☑	AA_1914_Corpse_flk	AA	ularzaŋgən'ibeʔ! ÖʔIe it ularlə!" "Šiʔ ularza...	măn	ularbə amnuʔlədən." "Ĭmbijleʔ amnədənʔ. ...	AA
9	☑	AA_1914_Corpse_flk	AA	öt's'ün Güd'ər nünneʔ baʔbi, uʔla saʔməb...	Măn	nörbəbiŏm šiʔn'idə, ularbə amnədən. Šăn...	AA
10	☑	AA_1914_Corpse_flk	AA	ăbəjʔleʔbddə. Dĭ bazoʔ šünə tĭrlölie. "Iʔ t'...	Măn	ularəm milem." "Măna ej kereʔ tăn ularlaʔ...	AA
11	☑	AA_1914_Corpse_flk	AA	ʔ! Măn ularəm milem." "Măna ej kereʔ tăn...	Măn	ĭmbin'im măna mibi." "No, iʔ t'oraʔ! Bar ...	AA
12	☑	AA_1914_Corpse_flk	AA	aʔdə Büz'e amna, šide koʔbdot amna. "A...	Măn	nükem ige. Pan tabəndə ularzaŋdə t'abola...	AA
13	☑	IA_1912_Trickster_flk	IA	P'el simabə kajlabaʔbi. "Tăn kădəʔ molaʔ...	Măn	simam t'ezerdələm. Öʔleʔ măna, tăn simal...	IA
14	☑	IA_1912_Trickster_flk	IA	əbi. Dĭzeŋ šobiiʔ. "Kădəʔ molaʔ š'a:mnaʔ ...	Măn	ej š'a:mniam. Măn săgər kuza igem." Ibiiʔ...	IA
15	☑	IA_1912_Trickster_flk	IA	"Kădəʔ molaʔ š'a:mnaʔ iʔbəlʔ" "Măn ej š...	Măn	săgər kuza igem." Ibiiʔ idəm sogənə sarbii...	IA
16	☑	NN_1914_Birds_flk	NN	əššət igel. Nen piʔmeendə tülində siktöleʔ ...	Măn	toru inen toru kunda tora toʔlaʔ kunnim."	NN

üke amnobi, oʔb n'it ibi. N'it ijaandə măliа: "Măn nejləm." Ijat măliа: "Naga, man'əkkən ne naga." "**Măn** kallam ne pileʔ." Kallaʔ t'ŭrbi, ijabə baʔla:mbi. Kandəga, kandəga. T'ăgan toʔgəndə maʔ nuga. Šŭbi

id[S] | AA

Lament [song]

Speaker: Avdakeya Andzhigatova
Language: Kamas [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 mīlle?biem sagər mājazaŋbə i?bəle? kojobi.

gijen	*mīn-la?bə-bi-m*	*sagər*	*mǟja-zAŋ-m*	*t'ībə-lA?*	*kojo-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

Mīlle?bəne t'ūm kūk no?ts'ə? özerle? ba?bi.

mīn-la?bə-NTA	*t'ü-m*	*kük*	*no?-t-zi?*	*özer-lA?*	*ba?bdə-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 i?bəle? kojobi, sīri t'ālamzaŋbə ma?la? kojobii?.

sagər	*mǟja-zAŋ-m*	*t'ībə-lA?*	*kojo-bi*	*sīri*	*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?	*kojo-bi-jə?*
stay-CVB	stay-PST-3PL

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

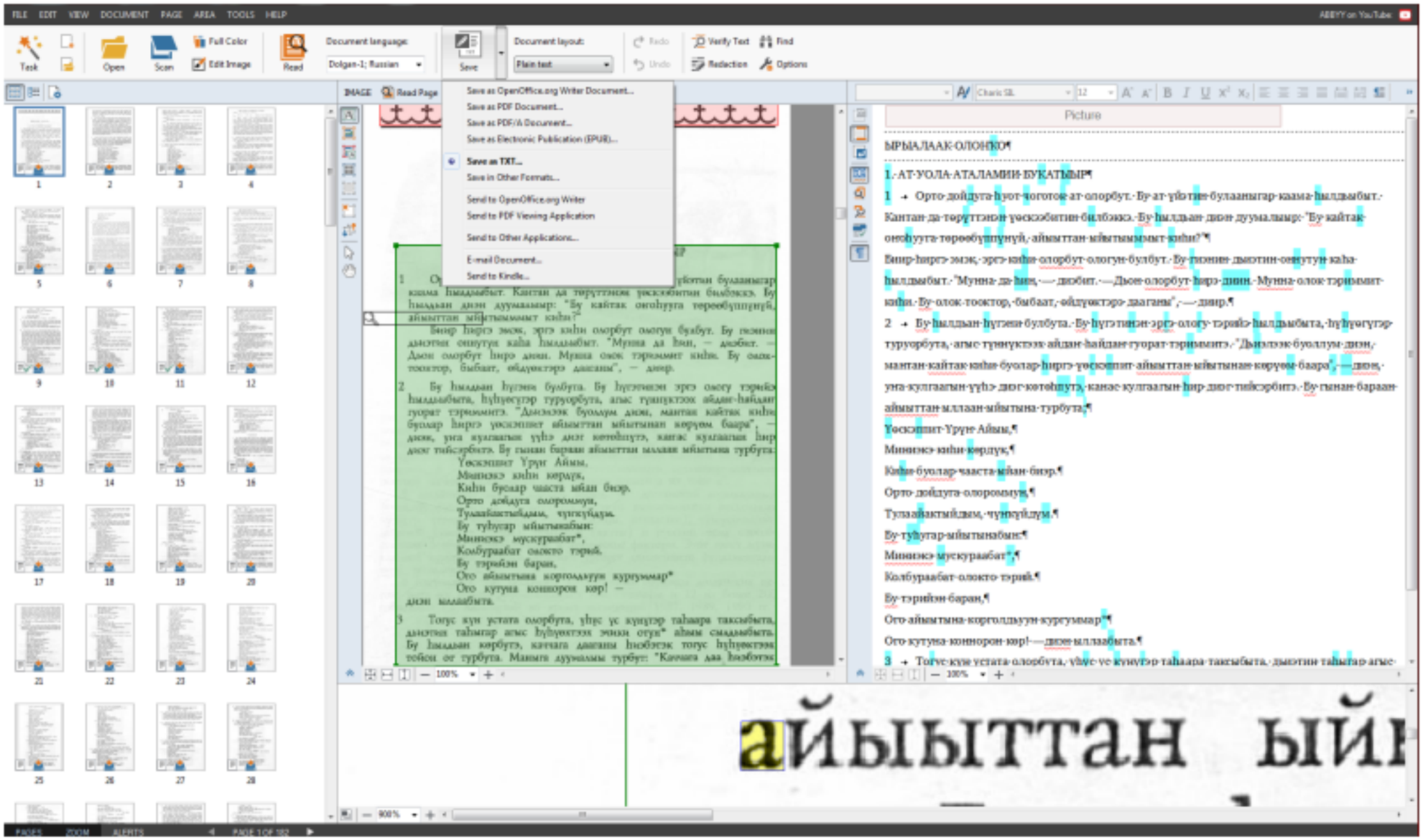
Paradox with ABBYY

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


FILE EDIT VIEW DOCUMENT PAGE AREA TOOLS HELP
ABBYY on YouTube
Task
Open
Scan
Full Color
Edit Image
Read
Document language:
Dolgan-1; Russian
Save
Document layout:
Plain text
Redo
Undo
Verify Text
Find
Redaction
Options
Save as OpenOffice.org Writer Document...
Save as PDF Document...
Save as PDF/A Document...
Save as Electronic Publication (EPUB)...
Save as TXT...
Save in Other Formats...
Send to OpenOffice.org Writer
Send to PDF Viewing Application
Send to Other Applications...
E-mail Document...
Send to Kindle...
IMAGE
Read Page
Picture
ЫРЫАЛААК-ОЛОҢКО¶
1.·АТ-УОЛА-АТАЛАМИИ·БУКАТЫЫР¶
1 → Орто·дойдуга·ыуот·чоготок·ат·олорбут.·Бу·ат·үйэтин·булааныгар·каама·ыллдьыбыт.·Кантан·да·төрүттэнэн·үөскээбитин·билбэкэ.·Бу·ыллдьан·диэн·дуумалыыр:·"Бу·кайтак·оноһуура·төрөөбүппүнүй,·аймыттан·мйытыммэмэт·киһи?"¶
Биир·ыирэ·эмэк,·эрэ·киһи·олорбут·олоҕун·булбут.·Бу·гэнэнэ·дьэтэн·онкутун·каһа·ыллдьыбыт.·"Мунна·да·ыын,·—·диэбит.·—Дьоо·олорбут·ыирэ·диин.·Мунна·олок·тэриммит·киһи.·Бу·олок·тооктор,·быбаат,·өйдүөктэрэ·даганы",·—·диир.¶
2 → Бу·ыллдьан·ыүгэнэ·булбута.·Бу·ыүгэнэн·эрэ·ологу·тэрийэ·ыллдыбыта,·ыүһүөгүтэр·туруорбута,·агыс·түнүктээх·айдан·ыайдан·туорат·тэриммитэ.·"Дьэлээк·буоллум·диэн,·мантан·кайтак·киһи·буолар·ыирэ·үөскээпит·аймыттан·мйытынан·көрүөм·баара",·—диэн,·уна·кулгаагын·үүһэ·диэг·көтөһүтэ,·канас·кулгаагын·ыир·диэг·тийсэрбитэ.·Бу·гынан·бараан·аймыттан·мллаан·мйытына·турбута.¶
Үөскээпит·Үрүҥ·Айыы,¶
Миниэкэ·киһи·көрдүк,¶
Киһи·буолар·чаастаабайын·биэр.¶
Орто·дойдуга·олоромыыбын,¶
Тулаайактыйдьым,·чүмчүйдүм.¶
Бу·түбүтар·мйытынабын:¶
Миниэкэ·мүскураабат",¶
Колбураабат·олокто·тэрий.¶
Бу·тэрийэн·баран,¶
Ого·аймытына·көрголдьуун·кургуммар"¶
Ого·кутуна·коннорон·көр!—диэн·ыллаабыта.¶
3 → Тоҕус·күн·устата·олорбута,·үһүс·үс·күнүгэр·таһаара·тахсыбыта,·дьэтэн·таһыгар·агыс·
айыыттан ыйы

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?

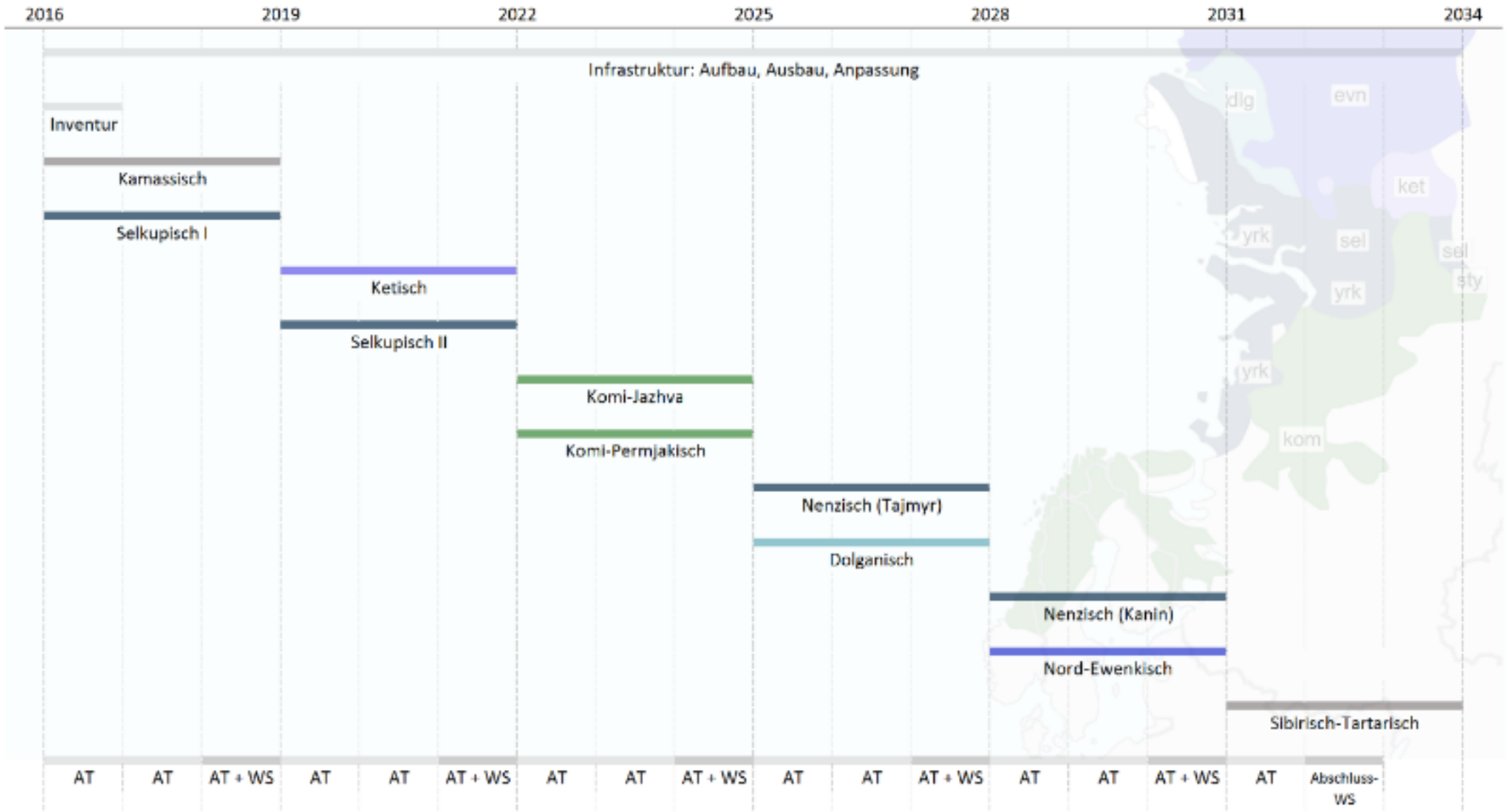
Combining corpora for research purposes

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

Аттьö! **Thank you!** Спасибо!

More information and publication:

https://github.com/nikopartanen/syktyvkar2017


2016
2019
2022
2025
2028
2031
2034
Infrastruktur: Aufbau, Ausbau, Anpassung
Inventur
Karnassisch
Selkupisch I
Ketisch
Selkupisch II
Komi-Jazhva
Komi-Permjakisch
Nenzisch (Tajmyr)
Dolganisch
Nenzisch (Kanin)
Nord-Ewenkisch
Sibirisch-Tartarisch
AT
AT
AT + WS
AT
AT
AT + WS
AT
AT
AT + WS
AT
AT
AT + WS
AT
AT
AT + WS
AT
Abschluss-WS
dlg
evn
ket
yrk
sel
sel
sty
yrk
yrk
kom

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

Pecularities of Dolgan case

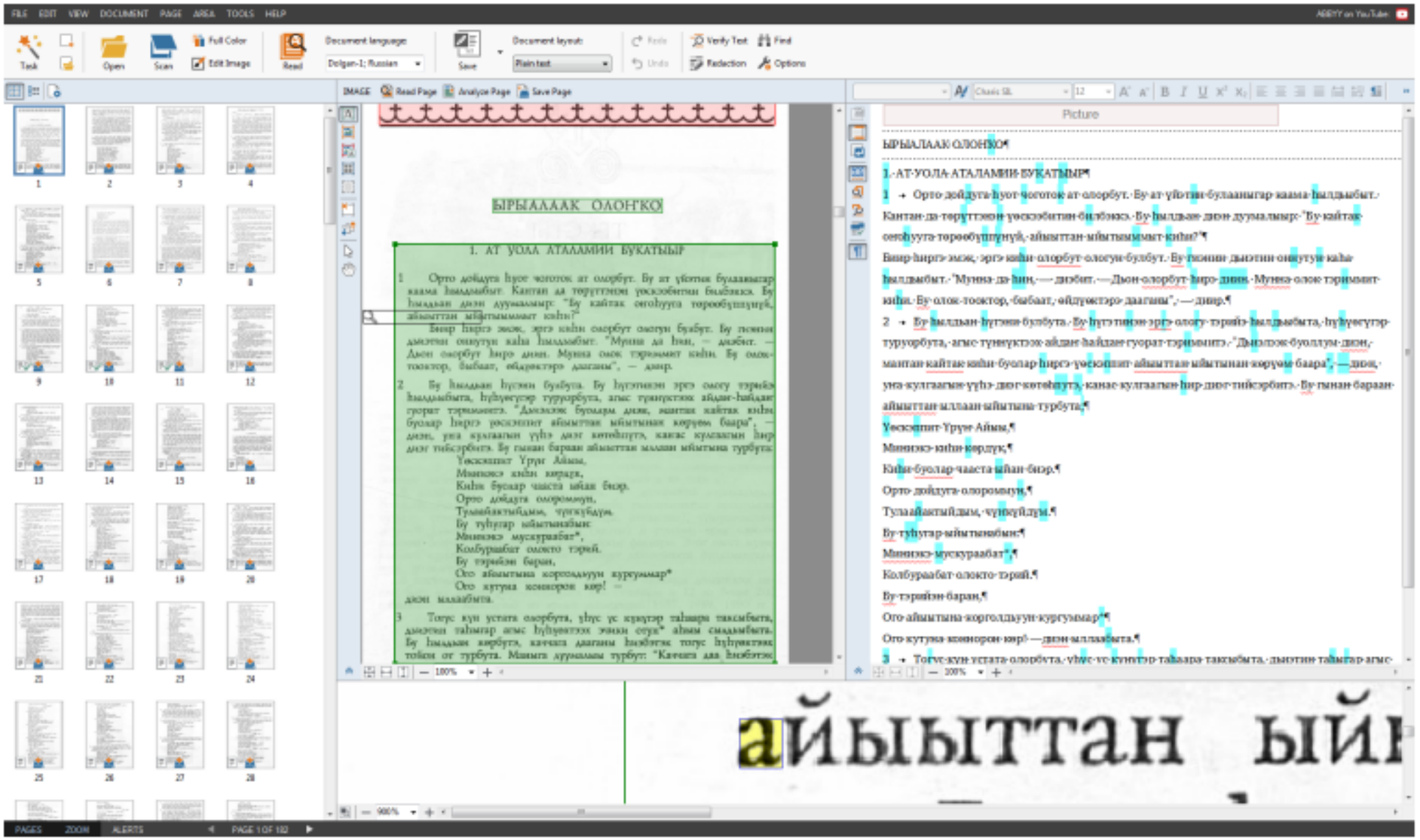
- Variety of transcription systems
  - Transliterated to INEL conventions
- Some printed publications rather new
  - High print quality, easy OCR
- Sentence-level translation to Russian
  - Need for aligning
  - Russian translations often in internet already

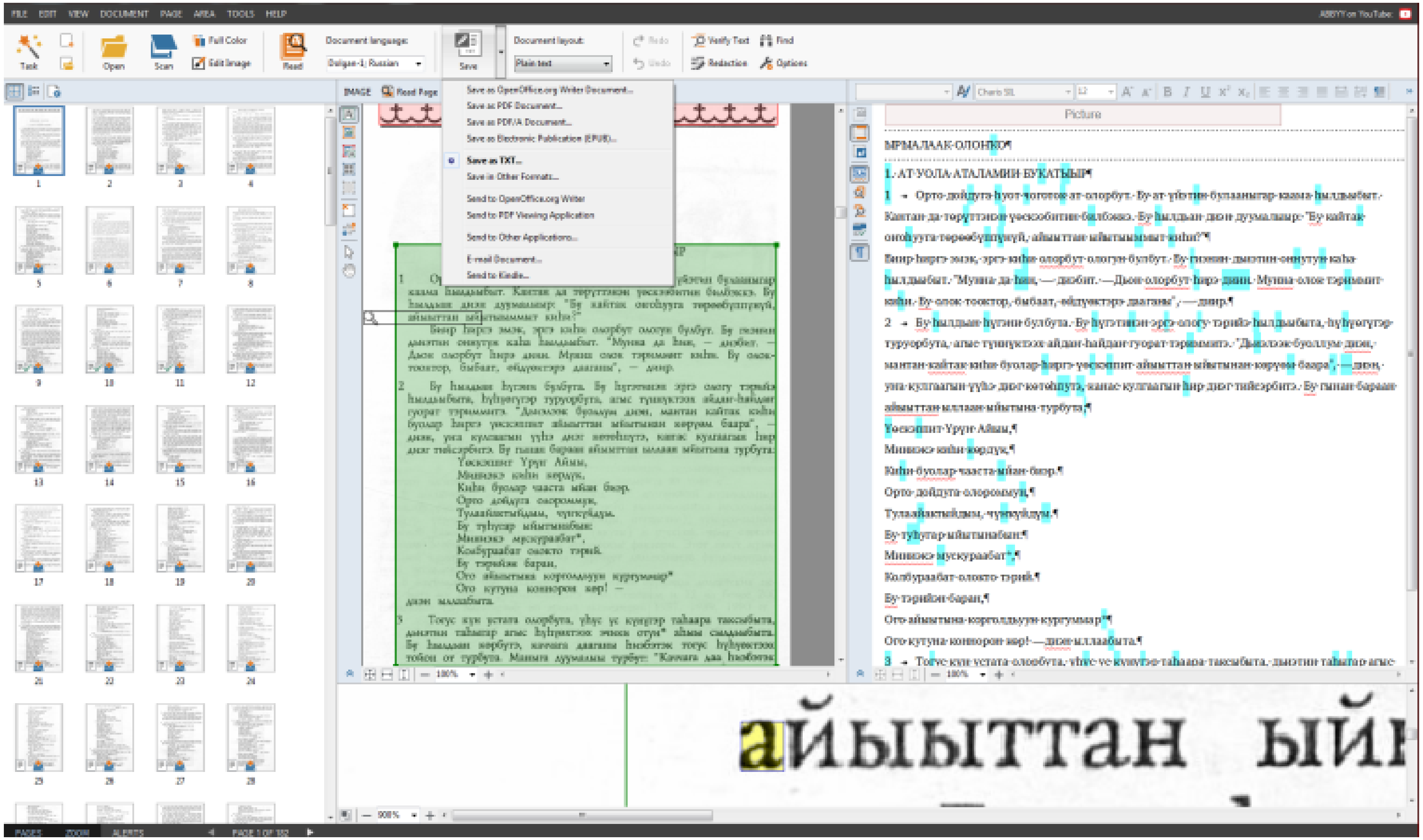
Basics of INEL OCR workflow

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


ЫРЫАЛААК ОЛОҢКО
1. АТ УОЛА АТАЛАМИИ БУКАТЫЫР
Picture
ЫРЫАЛААК ОЛОҢКО¶
1.·АТ·УОЛА·АТАЛАМИИ·БУКАТЫЫР¶


FILE EDIT VIEW DOCUMENT PAGE AREA TOOLS HELP
Task
Open
Scan
Full Color
Edit Image
Read
Document language:
Save
Document layout:
Plain text
Verify Text
Find
Redaction
Options
Save as OpenOffice.org Writer Document...
Save as PDF Document...
Save as PDF/A Document...
Save as Electronic Publication (EPUB)...
Save as TXT...
Save in Other Formats...
Send to OpenOffice.org Writer
Send to PDF Viewing Application
Send to Other Applications...
E-mail Document...
Send to Kindle...
Picture
айыттан