

Legacy data, text recognition and transliteration

Course: NLP for Endangered Languages of the Amazon.
From a Uralic perspective. Lecture 7.

Jack Rueter, Niko Partanen,
Mika Hämäläinen & Khalid Alnajjar

Today's lecture

Two parts:

- We discuss the current state of text recognition tools and their application
- We discuss some general methods and prerequisites in processing this kind of non-standard text, especially normalization and transliteration
- One starting point is that for our research purposes we often need different texts in one consistent transcription system
 - We can always save additional systems on different ELAN tiers!

13755_2az...			
	00:09:28.000	00:09:29.000	00:09:30.000
comment [0]			
ref@UAK-F-189	13755_2az.016	13755_2az.017	13755_2az.018
note(ref)@UA		Niko: Here is also variation какен ~скакен	
orth@UAK-F-1	харей выйтыв тай чеччис,	какен тай нин чеччис дай,	ылэ кодь тай нин чеччис тай,
ft-eng-vaszol	and jumped over the goad	he leapt up at full tilt	and jumped a long way away
orth-vaszolyi	xarej výtiys taj čećcis,	kaken taj nín čećcis da i,	yle kod' taj nín čećcis taj,
orth-vaszolyi	xarej výtiys taj čećcis	kaken taj nín čećcis da i	yle kod' taj nín čećcis taj
ft-eng-vaszol	and jumped over the goad	he leapt up at full tilt	and jumped a long way away
affricate@U	xarej viitvys taj {tc}e{tc:}is,	kaken taj nín {tc}e{tc:}is daj,	ile koj taj nín {tc}e{tc:}is taj,
orth-mikusev	Харей выйтыв тай чеччис,	Скаќон тай нин чеччис дай и,	Ылъодъ тай нин чеччис тай,

25. Небыд юрсна ныв

Важён вёвл яран гозъя,
Яран гозъялён нёль бур пи нин,
Яран гозъялён ётик бур ным,
Сылён вёл нимыс Небыд юрсна нин.
5 Яран гозъя тай кулісны,
Нёль бур пыыс тай коли нин,
Отик бур нымыс коли нин.
Ныя тай кучисны мунны нин,
Мореладорас вёрзбони нин,
10 Меддэоля вокыс тай сылён
Кёр тай вётлә тай нин, шуас,
Меддэоля вокыс тай мыльк вылд кайис,
Вылд нёль сывъя харейс сувтодис тай,
Вылд тай нин чечис дай и,
15 Харей вьвтис тай чечис,
Сакаён тай нин чечис дай и,
Ылдокъ тай нин чечис тай,
Дадъс босътис ді воддö войдис,
Бёррас тай нин бергдочис-а,
20 Куим даддя тай вётдич,
Сыа тай харей-вожжисо
Пуктис дай и вичысыштис.
«Сакаён вайб ордъбдамол»
Сакаён тай ордъбдисны,
25 Куимнансо ордъбдис нин.
Харейнисо сувтодисны,
Некод эз и вермы чечыны,
Сыа тай и чечыштис нин.
«Вайб бертчомон вермасям!»
30 Бертчыны тай кучисны да,
Отикос тай му бердас сетис —
Нырсымс-вомсымс и вир чечыштис,
Моддис тай нин сетыштис да —
Нырсымс-вомсымс и вир чечыштис,
35 И коймодс тай сетыштис да —
Нырсымс-вомсымс вир чечыштис.
«Медым, шуас, тэ кё вына,
Тая, шуас, казявлан нин!»
Даддяяс тай бёрр войдисны,
40 Кёр бёрсянныс тай муннин,

2. Небыд jursi coj pomlaš

(1)

važen vøli jaran gozja
jaran gozjalen nol' bur pi nín
jaran gozjalen øtik bur ny
sylen vøli nímys ød'd'erči
ød'd'erči taj nín vøli
jaran gozja taj kuliny
nol' bur piys taj kol'i nín
øtik bur nylys kol'i nín
nya taj kućis munny nín
more lazdore vørzyny nín
medžol'a vokys taj sylen
kør taj vøtle taj nín šuas
medžol'a vokys taj myl'k vyle kajis
vyle nol'-sýja xarejse sütedis taj
vyle taj nín čećcis da i
Xarej výtiys taj čećcis
kaken taj nín čećcis da i
yle kod' taj nín čećcis taj
dad'se boštis di vod'e vojedis
børas taj nín bergećis a
kujim dad'd'a taj vøtećce
kujim dad'd'a taj vøtećce
sya taj xarej, vøžžise
punktis daj i viććyşyštis

Legacy materials

It's extremely common that earlier materials exist in recorded, printed and handwritten forms.

The problems we have are rather universal:

- Orthography / transcription system varies widely
- Transcription level varies
- Printed and handwritten materials can be in poor condition, or single copies
- It is not always clear which audio connects to which recording

Value questions

There are always more old materials! (Where are our materials in 50 years?)

... and we also have to work with people today.

Prioritization is a challenge!

Some thoughts:

- Large transcription collections already contain plenty of work put into them
- Old materials, if context is known, are important for studying language change
- Still, some materials are certainly more work than what we can get from them
- This connects closely to **the amount of work needed**

Legacy digital materials

Existing texts in other formats and writing systems

Includes often old text files, PDF's, Word documents – often recoverable

- Scenario according to when and who
 - Legacy digital format
 - Legacy orthography
 - Mixture of both
- Conversions may include character-to-character transformation
 - Latin to Cyrillic conversion
 - Phonetic alphabet to literal language
 - Literary norm to new literary norm
- Means
 - Scripts: perl, python, any programming language
 - transducer

Experience with Parallel Bible corpus

Test translation work was begun in the early 1990s

Before standardization of character representation, i.e. Unicode

Font distinction that are readily lost in document conversions

Conversion of legacy document to UTF-8 document

```
# :- coding : utf-8 :-
```

At top of .txt file

Python script to convert

```
# sent_id = Facundes.2000:350:4a.  
# text = hătako-ro umaka-nanu-ta  
# gloss_en = youth-F sleep-PROG-VBLZ  
# text_en = The girl is sleeping.  
  
# sent_id = Facundes.2000:350:4b.  
# text = hătako-ro unuro iri-pe  
# gloss_en = youth-F mother fall-PFTV  
# text_en = The girl's mother has fallen down.
```

```
#####
s/u/y/g;
s/U/Y/g;
#
s/ũ/ÿ/g;
s/Ū/Ŷ/g;
#
s/o/u/g;
s/O/U/g;
#
s/õ/ũ/g;
s/Ō/Ū/g;
#####
#
```

```
# sent_id = Facyndes.2000:350:4a.  
# text = h̄taku-ru ymaka-nany-ta  
# gluss_en = yuyth-F sleep-PRUG-VBLZ  
# text_en = The girl is sleeping.  
  
# sent_id = Facyndes.2000:350:4b.  
# text = h̄taku-ru ynyru iri-pe  
# gluss_en = yuyth-F muther fall-PFTV  
# text_en = The girl's muther has fallen down.
```

```
# separating text_orig
# and text_orig
s/(\#\text)(\ =\ [^\n]*)/$1_orig$2\n$1$2/g;
# remove hyphen in glossing
s/(\#\text\ =\ [^\n]*)\-$1/g;
```

```
# sent_id = Facundes.2000:350:4a.  
# text_orig = h̄tako-ro umaka-nanu-ta  
# text = h̄takuru ymakananya  
# gloss_en = youth-F sleep-PROG-VBLZ  
# text_en = The girl is sleeping.  
h̄takuru ymakananya
```

```
# sent_id = Facundes.2000:350:4b.  
# text_orig = h̄tako-ro unuro iri-pe  
# text = h̄takuru ynyru iripe  
# gloss_en = youth-F mother fall-PFTV  
# text_en = The girl's mother has fallen down.  
h̄takuru ynyru iripe
```

We should be able to retain the info we have

- Pay attention to final result before starting conversion
- Many scrambled looking files may still be salvaged
- Remember context
- Consider possibilities for reuse

Modern text recognition

Since ~2015 text recognition was mainly done through fonts

This was lots of work and the accuracy was not always good

Modern systems work extremely well with relatively small amounts of data!

Consequences:

- Extracting a text from document can be fairly easy and fast
- Some documents still need so much other work that this doesn't solve it all

Jäärhaada *Toumair*
Neäkälgaada *Poraasyur / depnyur*
Tede tih *Bomr velen oreret*
Tanja nuungaa *tyyr emostor.*
Myud adsealma *oputnoba te budo emana*
Tarem muunaandit *taaker emosoro*
Juruunum biruuna *okalo comte*
Teä too *ott oreretaa toruudar*
Tjukun turuypaä *mu gifa bud (tunngaa)*
Seäi fjundieü *te moore - ans som fil (Bom) te kora koraader*
Härpanda waaniid *torni eä ilere*
Kunnat malgana *kolda eur*
Myäli jaa wackana *Ha Jayrou emponn demu*
Jien méngrpaah *Kor bigfrangy qides*
Tanjaa poodertor *thorda satipera*
Seäi fjundiem *Brury ko esoddy*
Taaved haara *Myymr emarsch*
Taathaua hajeh *(Vier) Toumerr*
Suijunda taeviit *Bo tseemey domuir*
Juru meän *600 Comte*
Teä rjuotaa *oleh, utyne*
Suijunda aedaada *Melissa enymaur*
Meädaa taevii *do kymuy domuir*
Parolieneje nje *4. f.*
Meädal naanda *Ha tyymay*
Kacjuviree *ressaasor*
Juruqunidama *Cno cassen'*
Myärdinda pylemida *Dryys sooroberry*
Nikkajietta *Paropbara*

189
Mjanta milva eny onsara *Gofens namn*
frieli nätnie/mekanda taeviit. Herrnrl mader to rym
ku fjordas 10/7.9.
Ju kündlioda *to die*
Raoden to die *Katidöö viretib / no otsvungoellim*
Mjamaa *Toumair*
Klaader i dox *Yhun*
Jefi mtnie *4. N.*
Ranjuda *Harmemur*
(Harmemur tier avar, Ha caseret hygistic
Tielola jeans *Orinen matrem channus*
Mjallada haispöödela *Cerebel to komismito*
Häppala *taas*
Tjukki jaanaando *Ha mairi unent*
Jugumbroos jiletih *to rodro tseligmr*
Jerkonietihi *4. N.*
Jepptifift *Dakuer oleciu deportuua dom*
Salie häradaa *Daberto yksar*
Ju jibid jaamban *4. 10. rodro*
Tee jibigädo *olema paskesodumer syuer*
Ulo oksama *torano, entida cattaro.*
Jiunq yatonzana *(Nordi tooro) Rataq tropy.*
Pari tie näte *Mr. Mönson hore*
Pin dargiuse *Ha yemuy boimesir*
Jipindla *Bozuy bzauro*
Tennya jaadaltiis *ko oderenur pomiser*
Jari pooderla *Fischbuu koinorolle (020)*
Kirmala *Domuas*
Mekanda taade *Hi ryas orubessi*
Kacjuviree *Zapkaa*
Janda vyn dijü *Ha caseret enur*
Kareada seä fundaada *Nenotnaam omeralo kaas*

+ Nykhetie, te 32.ayur.
172 Seä fundaada *m* Tih. k.
Njekatla satmaa *tu asyimb see austemr*
Mihula nige jadaa *Ha kolihaxr ulemor/pele*
"Tjukki siie to *tu monu oreret*
Mjana weewonuh *Topamo kyle*
Tubii Yena *Ha emosor oreret*
Lungal Nambria *Korda depozu doudy/sandar,*
Talugum talwinqudm. *Bo besu canuy*
Wl lue Lippida *Ha cunuk dohitey/val*
Jinjeq po pihquab *Botson yd arquise*
Tafinda jaamna *(r myzy (nestifran), reper*
Hjär jlaada *med hanig/stjus) todkello/leffor*
Seä fjundieü *To komaerbi*
Sienda häesi *(Elyu vaerc ed lepde qido)*
Tjäun taarrada *To komaerbi choerur*
Paludie paniil *Br te aankori sova*
Tideel wipnepäda *3.2.0 (nja, kles, por) Dokku jum*
Karaasla *Eduu "deptumiau" (nir reen koppa*
Pi jaamthana *app. Jan paka upp on haag häiset*
Numda jaalumax *medikatora konten vid paa of hels*
Seäi fjundieü *To besu soove.*
T. k.
Ar mudi domuir *Med. Cnnit o (dell) omait*
Ha seurlo mair *To (huskyto) Preverto*
Tafjäama *Wahr endemr*
Taremy & clatik *To seurta*
Tafjäama jauna *44 Dopolta Sura*
Pie dassuhi salvi *Oresek intolo/palay*
Tielola orkaseerit *Qura*
Sarpcä madahy *Meroni vprndasnel.*
Maro deata *Ek. var. nobr*
Nieda raha *Kahr sydmo uor dopola*
Juud H. Qaana *Za 10 rodro*
Häppala *Uppini*

R - 2
Komi - коми-зырянский язык

Informator Ye.S. Gulajev,
urozenec der. Czernjansk

Tak, me C'uzi d'ererjyannoy s'iktin.
taje s'ikty's su'lale kulemdins'on'
r'iz' rovajt verst, loas seco'ez, ve.
tym'in so krajtym'in kuim verst nene
suene yergem'yo s'ern'yo. noce Guljaev.
Islam ole'mis', menam ole'min u'lini'
una imtere sne'j'e volemlento'ras. Pozas
ristalni' lacemtor. Et'oi'd mumin rok-
negot c'eri nijni' pieten kulem... d'e-
revya. nne'j'an primorne kuim verst
vylam natim. Ses's'a prodelenjas
risollini' sile' rugirjas vilg. sis'-
s'a dimer c'ertalim mijec' apodol-
mijas, ric'c'is'am kor c'eri zede,
o zithis' nñn, rit lope. Ses's'a pamdis
nñn, minem er' kut. Tidalni, par'lecc'os
ni' ne'sjas ni'. sis's'a vor menam
r'iz' nijin ajlas leg'c'etema, kultim
nijni' pyrolo. en'irkse. et'u rugirjin
hinem abu sedema, med rugirje
s'oy' ye' minem abu sedema mijne
mis'a taje, taje mi'an nijsem
staris' torks'e tan, ses's'a me le-
zakis' minem eg' ay'zi, a sile

R - 1.
Komi - коми-зырянский язык.

Информатор Е.С. Гуляев
Уроженец деревни Деревенск

Так, я родился в деревне Деревенск.
Это село расположено (выс. синим) на Курган
гина на реке Сылва. Язык коми-зырянин
шумеров. До myga бывшем село шумеров при
Борисе. Имя звучит Борисово Гуляев.
На своей пасхе. Всю мою жизнь я интересовался
историей своего народа. Моя семья занималась о
таком. Однажды нашли с братом избушку
под забором. Всю свою жизнь я занималась
наши письмена погибли. Тогда
надоело это на уборку. Тогда
дядя занес в избушку чмо-чмо письмена
Баран, когда росла настурция. И унес в ба-
рчу, кирпичную. Там они и остались
тысячи. Умерло все снаружи.

Также моя бабушка в деревне звали
Люси айлас (?) письмена (сделала). Наша
левушка письмена. На огне угорну
широко не попало, но другую угорну широкое
широко не попало. Что все, забор, эмо.
Эта мама любила ее рассматривать. Так
как она в деревне среди широких заборов не было
а она

186.

medis i v a n velettščini p e t š o r a e štšetegednej kurs vīlē. sija sen olis velettščinj kik-suda demjn. sen sija tevarjš- jesky pondis berittščinj. i sija tevarjšess bjdngs lebedlis. vēdžē pondisnī koščni loken. i sija pondisnī tevarjšesis vīnī. i v a n tšetšts kik-suda demiš muč. vēdžē pišjies jage. sen oli sutki. vēdžē seš sije vajedisnī jagiš jejen. si berin vētšinī gortas. gortas vois jejen. i vēdžē pondis sija mamse nešavnī. mamis unaš ker-kašis pišjalis sušedjes orde. et-pir vētlis mamse ker-kašis ulištē. a mamislen veli potanij ponī tatei, tuplalema roten. i - v a n boštis tateirot niņnijs i lebedis bija paštē. no mamis kašalis, tateirot boštēm, i pišjis ker-kae tatei ding. no tateis potanas iz vev. vidližtis bija paštē, tateis veli lebedema paštēas. mamis kiskis kruken berge i iz uđit tateis sotščinj.

vēdžē dīk i v a n ojīn dumaitis mednj j e r t ī m tuje. sija vētlis kīž-vit vers vit lun nāgiteg. seki veli sī tuj kuža turun kiskalgnj. i v a n tuj vīlēn pukalis kāzdej das šag vīvti i sīlen sitan-ulas ni-nem iz vev puktema. i sīvdis sitanjs līmse kāzdej das šag vīvti (-mišti) guen. berge loktigenis i v a n siž-ž pukalis bid mestejn, ken munigenis pukalis. i sija berge iz vo gor te džis kujim vers. sija mesteseg ūueni ū e m e v -jegireg. i set- tē dīk i v a n kulis.

187.

važen olis l o p i - d i n j n kreštanin. sije šuisnī t š o v - j e s - m i š agn. olis ar kīž-džis i boštisnī sije germa-nsek vējnač. vējnač vētlem berin gortas oligen sija jejmīs. jejalig māznijs bit-to-reč ker-ka pītškesšis vāž torjessē torklis. boštas lok niž tser da sijen keravlas stav batislič ker-kaas karem torjessē. ešin betjessē keravlas, džodž-plakjessē lebedlas da vīles teftšas. i ūue dīk m i š : "iatīš ešin-betjessē da džodž-plakjessē kile lok bate-vskej dukjesis."

vēdžē et-pir dīk m i š pazeļdis ešin-lisjessē da medis ražnī kirpiča-patēs. ešin pirjijs medis lebedlinj. vēdžē ker-kaad loji onij dīk m i šli keždžd. si berin m i š k a međettšis tšuščiš ūikjesin vētlinj. a si vēdžē ūikjesas polinj iždžid lūd i ponī lūd. et-pir v e n - d i n j n vētligenis lebetišis loken vētnī eti mušikēs. si berin muni as ūikas berge da vētšis as ker-kaas kert-poni-patēs. a ūemja-is dīkaligenis veli as ūikas sušedjes ordin džepščma. vēdžē m i š olis ne-ki-ym lun as ker-kaas kert-pa-tēšen ūontemēn.

si berin dumaitisnī m i šes međednj sumošetšej deme. voisni si ding kujim as ūiksa mort da kik militische-r. m i š ni-nem dumaitteg boštis kosa da zīredis sījies vīlē. sijeza stavnīs povžisnī da pondisnī pišjīnī jaglač. m i š vētšis vētšis da pīris vartan riniše. da ībēse ūiptis, mōz ne-kod adžāi. ībēs ūiptem beras voisni settēs vit mortiđ da gorisnī: "m i š k a, pet! kosatē koř!" no m i š k a is pet. vēdžē kik militische-r iđ vētšinī ponī zadviškaa ešin da sešnī lijismī m i š k alič mōrgeas. m i š k a Jonas gorettšēmen settēs i kulis. vēdžē m i š k aeg božni lokis kojmed na tšačnik mili-tsii. i vašnī dīk m i šes gortas. kik lun mišti m i šes džebisnī v e n - d i n pleštšād vīlē. a kik militische-r tēs nu edisnī da pukšēdīnī raijo-nnej turmač. kik lun pukalem berin sužtisnī kīkna-nmīse kujim voen. tajen delonas i kētšišsīs t š o v j e s - m i š .

Diacritics are not necessarily a large challenge, at least technically.

They can still be difficult to edit!

This representation is useful for some tasks, but usually we want something phonemic.

How difficult the phonemic level is to retrieve varies from case to case.

šul̥gasas šul̥ga-viv šut̥škas, a veškiðnas veš-d šut̥škalis kik ijdžid grad.

Ixo-z kesjis vētšni ziska-nnēz. no pravke-nneijs nīžen. vejdžę me si berin sutki miš sije ošse

čāñkgsckə ton xūp ą̄-li tōuppeersobgs. čāñkgsckə l̥á'tetj̥i: *ski'tþorðotor
ą̄-pr̥yom, þði'tv̥áþtöör lez k̥uol̥tþps: 'sa·man n̥x̥ræk̥, īj̥-j̥en, sa-
man p̥y'þtálka, þa·l īj̥-j̥en!» ki'tþorðotor l̥á'tetj̥i: *þámpriðánörna p̥yþ
čāñkgsckə, n̥asor t̥enkh̥on t̥oð'þtsens t̥er?» ki'tþorðotor k̥yððat̥ t̥ari l̥á'tetj̥i:
*þáñkæt̥q̥en þártj̥iþt̥aþol t̥o'st̥q̥lm, aðil'ð k̥ðamol p̥óðli áv̥mðlþt̥l̥n!»
čāñkgsckə olm̥p̥ n̥omt̥obgs, keelþp̥ k̥þaq̥n.áðr̥cñýt̥obgs. þið s̥enl̥áyl̥ keelþ
ki'tþorðotor k̥yððat̥ juþðista, m̥ot s̥enl̥áyl̥ keelþma þasta, k̥eppð'ssamt p̥y-
þðlþp̥n t̥oymasta: p̥yþþat̥ juþðista, s̥elmk̥a.izs keelþ áðino, s̥omq̥n n̥y-
räý j̥amti. k̥ontauðnasmēna īzgl̥na, palnøþþr̥ ńá't j̥amti n̥en pjátl̥q̥nt.»
m̥anazt̥ j̥azt̥ teerst̥omannu īzgl̥st̥. þojet̥q̥m n̥árl̥osta: p̥uræp̥þ
n̥aðlþþt̥q̥z̥u k̥ðar̥z̥. p̥uræp̥þ n̥aðlþq̥as. »n̥ár k̥ðar̥z̥, ki'tþorðotor!»
ki'tþorðotor l̥á'tetj̥i: »n̥ár k̥ðar̥z̥?! k̥ontauðz̥u j̥at īj̥-j̥en!» p̥uræp̥þ
l̥á'tetj̥i: »n̥orsomk̥, m̥anawm, ńá'l n̥orsomk̥, ńá'l m̥anawm; om ńáðom
gá-ti.» ki'tþorðotor l̥á'tetj̥i: »sk̥ormel ńá't juþðan? j̥az̥u k̥ðar̥z̥. n̥áy

bunden. Die Alte sagt: »Zwei-Bergrücken-Fürst, mein Junge, der Voitavter Fürst hat eine Botschaft zurückgelassen: 'Wenn dein Herz stark ist, komm, wenn dein Herz schwach ist, komm nicht!' Der Zwei-Bergrücken-Fürst sagt: »Verfluchte Hündin³⁵, Alte, was für eine Botschaft hast du bis jetzt (bei dir) bewahrt³⁶?» Der Zwei-Bergrücken-Fürst sagt zu seinem Bruder: »Schlage sie mit dem Ruder aus hartem Holz auf den Kopf (und) spalte sie bis mitten zwischen ihre Beine entzweij³⁷. Die Alte wurde entzweigerissen, ihr Blut strömte hervor. Ein kleines Birkenrindenkörbchen voll Blut trank der Bruder des Zwei-Bergrücken-Fürsten, er [der Zwei-Bergrücken-Fürst] nahm ein zweites Birkenrindenkörbchen voll Blut, reichte es seinem Sohne in dem Hinterteil des Bootes: sein Sohn trank es. »(Wenn) ihr Menschenblut trinkt, wird euer Herz stark. (Wenn) ihr auf den Kampfplatz kommt, werdet ihr keine Furcht haben³⁸.»

Sie gingen, fuhren, kamen zu dem Tuman Teriz. Er [der Fürst] liess seinen Vogel schreien³⁹: man muss den Püre-Sohn⁴⁰ ans Ufer rufen. Der Püre-Sohn kam ans Ufer. »Was ist (dir) nötig, Zwei-Bergrücken-Fürst?» Der Zwei-Bergrücken-Fürst sagt: »Was (mir) nötig ist?! Komm mit zum Kämpfen!» Der Püre-Sohn sagt: »Wenn ich will, gehe ich, wenn ich nicht will, gehe ich nicht; ich habe nichts

me t̄sužli tiša t̄sa ek-mis-šo das-ńołęd vojn p o j o l v̄elęštijn,
 kreššańin pi. kękjà.-mis aręss̄ań muni veleſt̄s̄inj školaę. veleſt̄s̄i
 kik vo. b̄ernas uđzali gort̄in mu-vidž dorin tiša t̄sa ek-mis-šo ko-
 m̄in ſižimęd voęd̄z̄. seki mijanęs kula t̄sitisn̄. menjm ſetisn̄ tu-
 remnež zak lu t̄še n̄ne vit vo. seni oli kujim vo da džin. setiš leđzem
 berin uđzali s i k t i v - k a r i n gružšikin das-ętik teliš. sišań
 muni gort̄. p̄iri kolkoz̄. kolkozas uđzali traktori st̄in n̄ełà.-m̄in

168 Texte von Šiškin

94.

me t̄sužli tiša t̄sa ek-mis-šo das-ńołęd vojn p o j o l v̄elęštijn
 kreššańin pi. kękjà.-mis aręss̄ań muni veleſt̄s̄inj školaę. veleſt̄s̄i
 kik vo. b̄ernas uđzali gort̄in mu-vidž dorin tiša t̄sa ek-mis-šo ko-
 m̄in ſižimęd voęd̄z̄. seki mijanęs kula t̄sitisn̄. menjm ſetisn̄ tu-
 remnež zak lu t̄še n̄ne vit vo. seni oli kujim vo da džin. setiš leđzem
 berin uđzali s i k t i v - k a r i n gružšikin das-ętik teliš sišań
 muni gort̄. p̄iri kolkoz̄. kolkozas uđzali traktori st̄in n̄ełà.-m̄in
 ętikęd voęd̄z̄ oktab telišęd̄z̄.
 oktab telišęd̄z̄ n̄ołęd lunę boštisn̄ armijaę. armijaas služiti dékab
 telišęd̄z̄. dékad telišin kvałęd lunę šuri plenę. plenaś uđzali

Example

(Niko shows in this point some of his work in Transkribus)

Basic workflow (Transkribus as example)

Register as a Transkribus user

Install the software

Upload the document

Write email and ask for model training rights

Transcribe few pages manually

Create the first model, transcribe more.

Repeat until quality is good! (Ask for more credits from administrators)

How does it work?

Each line is arbitrarily mapped into string of characters.

The system doesn't "know" which character is "correct".

= it doesn't really matter which diacritics we use, or leave out, as long as the representation is systematic.

Actual text recognition is usually done after the document layout is detected.

3.

sumka

olis vīlis gōzja. nūlen nī-efik tšelqād' iz nē. muži k kežis prosa. se tše m peti prosa:js bur! sešsa zēi vīna tē kiskis da i bīdsen prosas žugedis. muži k berde i babajsli rištale: »eni los rōktag ūnjl a baba sīli i rištale: »mun perij tē ords da kōr sliš sud, med tēnjd mintas prosa:js vīle.» muži k i muni perij tē orde, mīkirtsis sīli da i rištale: »mē tē ordad vīgi dela:en: tejad piijd mēntšum prosaes žugedis. vāi minti!» tē rištale: »mē og tēd, mījen tēnjd mintini!, a sešsa kuťsis rištōni: »mē šēta tēnjd sumka!» muži k rištale: »mīj me sijen kuťsa karni?» tē rištale: »kūz tēnjd ūoijnī kōšas, tol'ko rištō: 'sumkae, vāi ūoijnī!', sia tēnjd mīmē mīj kōls šēlemidli sije i vajas; kōr pētan, rištō: 'sumkae, sumkae, vōd!', sia i vōdas.» »no bur tāj!» muži k babajsli kuťsis ūoijišni: »rot, mīj rāji mel!». babajs

3.

Der ranzen

Es war einmal ein ehepaar. Sie hatten gar kein kind. Der mann säte hirse. Eine so grossartige hirse wuchs auf! Darauf erhob sich ein sehr heftiger wind und vernichtete die hirse mit stumpf und stiel. Der mann weint und sagt zu seiner frau: »Jetzt müssen wir ohne brei leben!» Aber da sagt das weib zu ihm: »Geh zum alten Wind und fordere von ihm gerechtigkeit, dass er dir für deine hirse ersatz bezahle.» Der mann ging denn auch zum alten Wind, verbeugte sich vor ihm und sagt: »Ich komme mit einem anliegen zu dir: dein sohn hat mir meine hirse ver-

juale: »mīj vajin sumka?» »a rot mīj, baba! etaja se tše m sumka:js, rištalan-ke: 'sumkae, vāi mē ūoijnī?' mīmē i bīdsen los pīzan vīlin mīj šēlemidli kōls.» baba boštis vedra da i muni vala i kuťsis ūoijišni bidenli. kumis sīli rištale: »mē vāa voskrešenīa ebēdajtnī!» kum i vōjīs nī orde gosti. pukšedisnīs pīzan saje, a pīzan vīlin nī-nēm abu. kum dīvūttīs: »kuťsem etaja prītša! mījen kuťšasnis verdni?» a zača in boštis sumka da i puktis pīzan vīle i rištalis: »sumkae, sumkae, vāi ūoijnī!» drug pīzan vīle bīdsen loji mīj kōli šēlemenisli: dōna vinajas i bīd-sama ūojanjs. ūojsnis, ūojsnis da iz vermjīs bīredni. muni kum gōrte i rištale babajsl: »rot, baba!, kuťsem sumka:js!» »a mīj sīja se tše m sumka:js?» »mīj?! da rot mīj: pukšedisnīs mēns ūoijnī pīzan saje, boštis sumka da i rištale: 'sumkae, vāi ūoijnī?' i bīd-sama burjs loji pīzan vīle. ūon vērmī bīredni! kīj sije boštis! muna juala!» muni da i jualis kumljs: »kijs tē vajin

len: »Schau, was ich mitgebracht habe!» Die frau fragt: »Warum hast du einen ranzen gebracht?» »Pass auf, weshalb, mein frau-chen! Dies ist solch ein ranzen, dass, wenn du sagst: 'Mein ranzen, gib mir essen!' sofort auf deinen tisch kommt, was dein herz begehrt.» Die frau nahm einen eimer und ging hin, um was-ser zu holen, und begann vor allen zu prahlen. Ihr gevatter sagt zu ihr: »Ich komme am sonntag zum mittagessen!» Der gevatter kam denn auch zu ihnen auf besuch. Sie setzen ihn hinter den tisch, auf dem tisch aber war nichts. Der gevatter verwundert sich: »Was für eine seltsame geschichte ist das doch! Womit gedenken sie mich zu bewirten?» Der hausherr aber nahm den ranzen und stellte ihn auf den tisch und sagte: »Mein ranzen, mein ranzen, gib luns! essen!» Plötzlich kam auf den tisch, was

*uśis vaas. i kijkna·nīs kutisniż bērdniż. seśsa mamiś istas koimed
nivse¹: »vetli, miila naje dır oz lokniż. i t'set niv lett'sas t'sojojisjas
dineż i juależ: »miila že ti bērdad?» naje viştaleniż: »kičdži že oge
berde! mijan vęd vedranim uši». i kujimna·nīs kutisniż bērdniż. seśsa
loktasniż sett'se mamiś i baťis. naje jualeniż: »mijiś nę ti bērdad?»
naje viştaleniż: »vedranim uśis vaas». i mamiś baťis kutasniż bērdniż.*

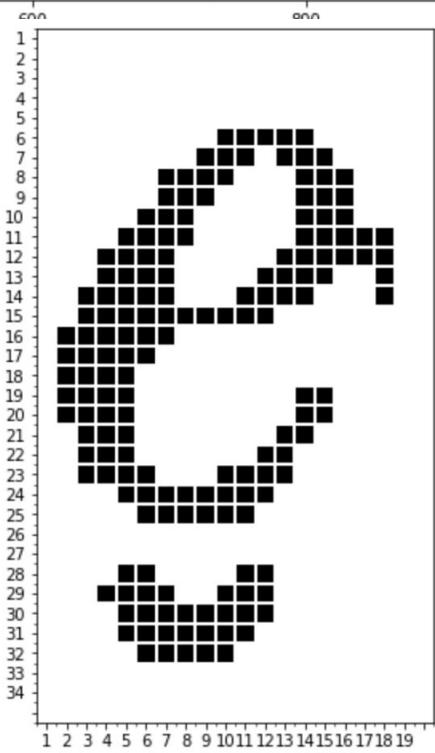
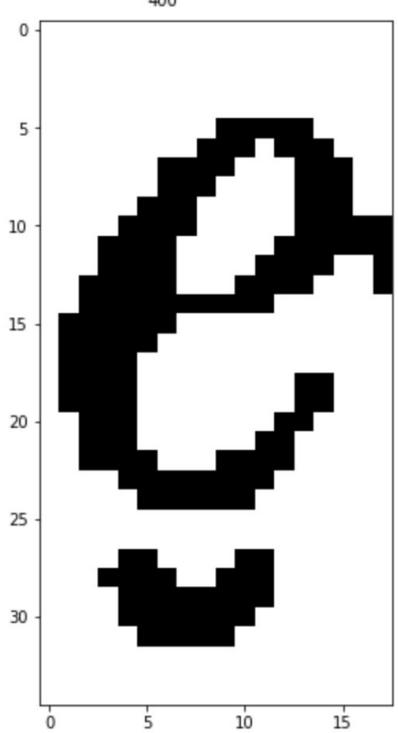
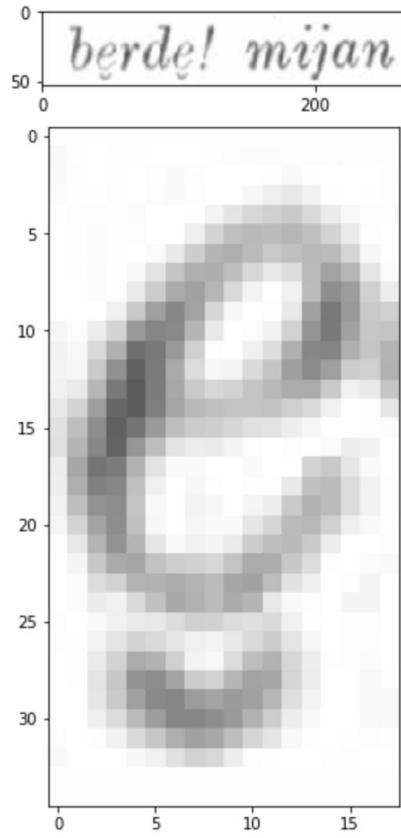
¹ auch: *koime·d nivse*.

27.

Die dummköpfe.

Es war einmal ein mann und eine frau. Sie hatten drei
töchter und einen sohn. Sie schickten die älteste tochter zum
flussufer um wasser zu holen. Sie ging. Das wasser war sehr
reisigend. Sie begann zu schwören und dann einem fiel ihr wagon

berde! mijan vəd vedranim uši». i kujimna·niş kutisniј berdnj. sešsa



bərde! mijan vəd vedranım uşı», i kujimna·nis kutisni bərdni. sessé

Sequence to sequence model (when lots of materials already exist)

or

Transliteration rules (works always – complex variation hard to account)

bərde! mijan vəd vedranım uşı», i kujimnanis kutisni bərdni. seç:a

Compare to: бөрдө! Миян вөд ведраным уси», И куимнаныс кутісны бөрдны. Сәсся



Calamari OCR

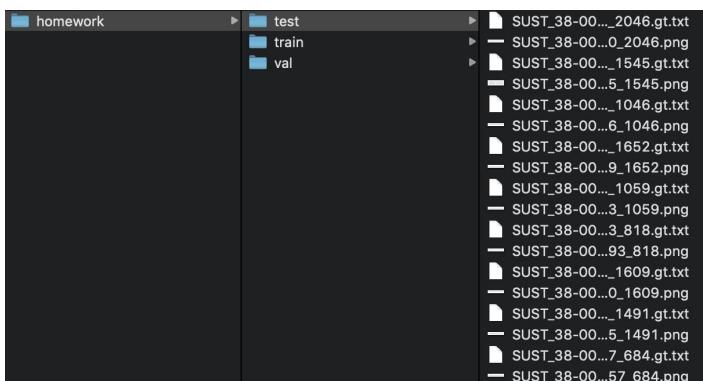
OCR Engine based on OCROpy and Kraken using python3. It is designed to both be easy to use from the command line but also be modular to be integrated and customized from other python scripts.

In der furstlygher ordenügen

Jn der furstlygher ordenügen

The accelerated weathering tests were

The accelerated weathering tests were



Pretrained model repository

Pretrained models are available at (https://github.com/Calamari-OCR/calamari_models). The current release can be accessed [here](#) (336 MB).

Installing

Installation using Pip

The suggested method is to install calamari into a virtual environment using pip:

```
virtualenv -p python3 PATH_TO_VENV_DIR (e. g. virtualenv calamari_venv)
source PATH_TO_VENV_DIR/bin/activate
pip install calamari_ocr
```


Current results

Printed text usually starts to work well after ~10 pages are proofread

- In this point rare characters and upper case letters may still cause problems
- The accuracy with printed documents may easily be in 99.99%
- In some point the mistakes in the training data start to appear in result

With handwritten documents we usually need more than 40 pages in one hand

- Dataset should be larger than i.e. 50 pages
- Under good conditions the accuracy may get above 95%
- Ideally the collection would have similar layout and structure
- Many writers, difficult (or inconsistent / bad) writing and multiple languages are still a large problem, and the accuracy may remain below 90%

Normalization

We trained a model with large Finnish dialect corpus

More than 700,000 tokens.

The accuracy is extremely good, but the conditions also very unique.

Conceptually not far from text recognition:

dialectal string > normalized string



The amazing Murre (*genitive Murren* 🐕) will normalize non-standard Finnish (kirjakieli). This repository is maintained by [Mika Hämäläinen](#).

Installation

This library is designed for Python 3 and it may not work on Python 2.

```
pip3 install murre  
python3 -m murre.download
```

Normalize

To normalize Finnish, all you need to do is to run:

```
from murre import normalize_sentence  
  
normalize_sentence("mä syön paljo karkkii")  
>> minä syön paljon karkkia
```

Blue Livonian Cow

