

Fieldwork and Early Literary Texts 'Kenttättyötä ja varhaiskirjallisia tekstejä'

Tuoda Paasonen korpiin

21.4.2017 Suomalais-Ugrilaisen Seuran kokous

Fieldwork and Early Literary Texts (FELT)
‘Kenttättyötä ja varhaiskirjallisia tekstejä’

Jack Rueter,
Nykykielten laitos,
Helsingin yliopisto

‘Kenttätöitä ja varhaiskirjallisia tekstejä’

Suomalais-Ugrilainen Seura

Yhteistyötä

- Suomalais-Ugrilainen Seura
- Suomen kansalliskirjasto...
- Yliopistoja/Instituutteja... (Helsinki, Tromsø, Syktyvkar)

'Kenttätöitä ja varhaiskirjallisia tekstejä'

- Aineistot
- Työvuot
- Työkalut

Aineistot

Suomalais-Ugrilainen Seura

Varhaiskenttätöitä 1800-luvun loppupuolelta
vuoteen 1917

Päätös avata vuoteen 1947 asti julkaistut
tutkimustyöt.

Poikkeuksia on tehty, esimerkiksi laajennuksia
Paasosen mordvalaismateriaalien osalta

Vaiheet

- OCR eli tekstintunnistus
- Oikoluku
- Normalisointi
- Morfologinen analyysi
- Disambiguaatio eli yksiselitteistäminen
- Korp

OCR

OCR on suoritettu yhteistyönä Suomalais-Ugrilaisen Seuran ja Suomen kansalliskirjaston kanssa

Oikoluku

Oikolukua on teetetty kielellisesti valveutuneella äidinkielisellä

Kolmen kuukauden oikoluvulla on käyty läpi kolme osaa kahdeksasta, noin 800 sivua.

Normalisointi

Tarkekirjoitus >> nykykirjakielen tapainen muoto

Ratkaisu: tehdä uutta työtä

Kieli ja murrepiirteitä huomioidaan

Vokaaleja, konsonantteja, tavuja

(ersä ja mokša)

Analogisissa projekteissa, esimerkiksi molempien komien osalta voidaan hyödyntää Syktyvkarin FU-Labin Molodtsov-konversiota. Vastaavia ratkaisuja uumoillaan Joškar-Olan Marnii:lta.

Analyysi

Avoimen lähdekoodin morfologisia jäsentimiä:
(Giella)

Uralilaisia vähemmistökieliä:

ersä, mokša;

komi-syrjääni, komi-permjakki, udmurtti;

niittymari, vuorimari;

mansi;

tundranenetsi;

aunuksenkarjala, liivi, kveeni, meänkieli, yms.

Disambiguation 1

ersä 'seisoa. valvoa. olla'

```
<w word="," lemma="," pos="CLB" msd="CLB" sID="8" orig_string=","/>
```

```
<w word="аштить" lemma="" pos="" msd="" sID="9" orig_string="аштить"/>
```

аштить	аштемс+V+IV+Der/Ы+Act+PrsPrс+Pl	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Nom+Indef+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Nom+Indef	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Nom+PxSg2	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Pl+Nom+PxSg2+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Sg+Gen+PxSg2	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Sg+Gen+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Sg+Nom+PxSg2	0.000000	
аштить	аштемс+V+IV+Der/Ы+Der/NomAg+N+Sg+Nom+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+IV+Ind+Prt1+ScSg2	0.000000	
аштить	аштемс+V+IV+Ind+Prs+ScPl3	0.000000	
аштить	аштемс+V+TV+Der/Ы+Act+PrsPrс+Pl	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Nom+Indef+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Nom+Indef	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Nom+PxSg2	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Pl+Nom+PxSg2+Der/Cop+Prs+ScPl3	0.000000	0.000000
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Sg+Gen+PxSg2	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Sg+Gen+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Sg+Nom+PxSg2	0.000000	
аштить	аштемс+V+TV+Der/Ы+Der/NomAg+N+Sg+Nom+PxSg2+Der/Cop+Prs+ScSg3	0.000000	0.000000
аштить	аштемс+V+TV+Ind+Prt1+ScSg2	0.000000	
аштить	аштемс+V+TV+Imprt+ScSg2+0cPl3	0.000000	
аштить	аштемс+V+TV+Ind+Prt1+ScSg2+0cPl3	0.000000	0.000000
аштить	аштемс+V+TV+Ind+Prs+ScPl3	0.000000	
аштить	аштемс+V+TV+Imprt+ScSg2+0cPl3	0.000000	
аштить	аштемс+V+TV+Ind+Prt1+ScSg2+0cPl3	0.000000	0.000000
аштить	аштемс+V+TV+Imprt+ScSg2+0cPl3	0.000000	
аштить	аштемс+V+TV+Ind+Prt1+ScSg2+0cPl3	0.000000	

```
<w word="кудосо" lemma="кудо" pos="N" msd="Sem/Build+N+SP+Ine+Indef" sID="10" orig_string="кудосо"/>
```

Disambiguation 2

Komi-syrjääni 'kurki' vs 'kiirehtiä'

```
> тури
тури      тури+N+Sg+Nom      0.000000
тури      турны+V+Ind+Prt1+Sg1  0.000000
тури      турны+V+Ind+Prt1+Sg3  0.000000
```

□

Disambiguaatio 3

Työtä kielitieteilijöille ja kielenoppijolle

Syntaksi

Rajoitekielioppi

(esimerkiksi Giellateknolla tai CSC:llä)

Korp

```
<sentence paragID="1" sent="3" pgNo="8" pgLi="3" orig_string="тури да рутš меџеттšасни мунни."  
  deu="Der Kranich und der Fuchs gehen los.">  
<w word="тури" lemma="тури" pos="N" msd="N.Sg.Nom" sID="1" orig_string="тури"/>  
<w word="да" lemma="да" pos="CC" msd="CC" sID="2" orig_string="да"/>  
<w word="руч" lemma="руч" pos="N" msd="N.Sg.Nom" sID="3" orig_string="рутš"/>  
<w word="мөдөдчасны" lemma="мөдөдчыны" pos="V" msd="V.Ind.Fut.Pl3" sID="4" orig_string="меџеттšасни"/>  
<w word="мунны" lemma="мунны" pos="V" msd="V.Inf" sID="5" orig_string="мунни"/>  
<w word="." lemma="." pos="CLB" msd="CLB" sID="6" orig_string="."/>  
</sentence>
```

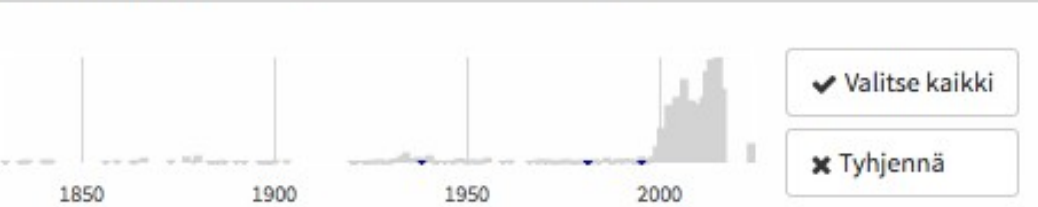
Korp

XML: sivunumero, sivurivi, tyyppi

```
<----->
<sentence pgNo="0002" pgLi="15" type="li" orig_string="ombotks takanz
<w word="омботькс" lemma="омботькс" pos="Det" msd="Det" sID="1" orig_
<w word="тяканзо" lemma="тяка" pos="N" msd="N.Sg.Gen.PxSg3" sID="2" o
<w word="Везоргонь" lemma="Везорго" pos="N" msd="Sem/Divinity.N.Prop.
</sentence>
<sentence pgNo="0002" pgLi="16" type="li" orig_string="... kibeid bi
```




3 / 64 korpusta valittuina — 6,05K / 251,13M sanetta



- ORACC (4)
- ERME (2)
 - Mokša/Moksha
 - Ersä/Erzya
- Fenno-Ugrica (10)
- English / Englanti (28)
- Deutsch / Saksa / German (2)
- Français / Ranska / French (1)
- Español / Espanja / Spanish (1)
- Русский / Venäjä / Russian (7)
- Helsinki Corpus of Swahili 2.0 (HCS 2.0) (4)
- SUS-kenttätyö (näyte) (3)
 - SUS-kenttätyö: ersä (näyte)
 - SUS-kenttätyö: komisyrjääni (näyte)
 - SUS-kenttätyö: mokša (näyte)
- Kildin Saami (sample)

Yksinkertainen Laajen

Aktiivinen CQP-haku yksinkertainen
[word = ".*"]

Aktiivinen CQP-haku laajennettu
[]

Räätälöity CQP-haku:
[word = "тя.*"]

Etsi ▼ virkkeen

Konkordanssi: osumia sivulla:

Results list area with search results and a dropdown menu at the bottom set to 'n perusteella: sana'.

Pari hakua korpista

Räätälöity CQP-haku: [word = "тури"] [Lataa CQP-ohjeet](#)

Etsi virkkeen sisältä

Konkordanssi: osumia sivulla: 25 järjestä korpuksen sisällä: järjestämätön Tilastoja: laske tilastot tämän perusteella: sana Näytä kartta

Konkordanssi Tilastoja Kartta Nimiluokittelu

Tuloksia: 10

« < 1 > » Siirry sivulle / 1 Näytä konteksti

SUS-KENTTÄTYÖ: KOMISYRJÄÄNI (NÄYTE)

тури	шуд ручлы: « мунам, чойд, ме ордõ госьтитны! »
тури	да руч мөддочасны мунны.
тури	заводитас сёйны.
тури	сідзи тшыгйбн и кольдõ.
та бõрти мөдасны мунны	тури ордõ госьтитны.
воасны	тури ордõ.
тури	пуас рунь ит-пызыысь да йбв да рунь ваяс пызан вылõ сулеяын.
руч пуксяс турикөд сёйны,	тури нырсõ сюяс сулея, йбв и рунь кыкнан сулеясыс юас.
тури	ручлы шуõ: « водзõстõ тадзи ме и госьтитõдõ тәнõ. »
шумитасны и тышкõн сорõн янсõдчасны	тури да руч.

« < 1 > » Siirry sivulle / 1

Lataa tiedostona muodossa: [Annot](#) [Ref](#) [Nooj](#)

[JSON](#)

Korpus

SUS-kenttätyö: komisyrjääni (näyte)

Kuvailutiedot
Lisenssi: CC BY-NC (CLARIN PUB)

Tekstin ominaisuudet

kieli: komisyrjääni
haastattelujankoha: 1942
haastateltava: Миш Иван::Ivan Mihailovič Gabov
haastattelijä: T. E. Uotila
paikkakunta: Нившера
paikkakunta alkukielellä: Одыб
paikkakunta venäjäksi: Нившера
tekstin otsikko alkukielellä: Руч да тури
tekstin otsikko käännöksenä: Der Fuchs und der Kranich
julkaisun nimi: Syrjänische Texte, Band IV.
SUST 221
numero: 221
julkaisija: Suomalais-Ugrilainen Seura
julkaisuvuosi: 1995
julkaisupaikka: Helsinki

Ylijumalan toinen, tytär

Räättälöity CQP-haku:

[word = "тя.*"]

[Lataa CQP-ohjeet](#)

Etsi

virkkeen

sisältä

Konkordanssi:

osumia sivulla: 25

järjestä korpuksen sisällä: järjestämätön

Tilastoja:

laske tilastot tämän perusteella: sana

Näytä kartta

Konkordanssi

Tilastoja

Kartta

Nimiluokittelu

Tuloksia: 4

« < 1 > » Siirry sivulle / 1 Näytä konteksti

SUS-KENTTÄTYÖ: ERSÄ (NÄYTE)

покшось **тяканзо** Кастаргонь

омботькс **тяканзо** Везоргонь

кезерь эрзянь веженсь **тяка**

SUS-KENTTÄTYÖ: MOKŠA (NÄYTE)

пулосонза тол заря **тяштне**.

« < 1 > » Siirry sivulle / 1

Lataa tiedostona muodossa: [Annot](#) [Ref](#) [Nooj](#)

[JSON](#)

Korpus

SUS-kenttätyö: ersä (näyte)

Kuvailutiedot

Lisenssi: CC BY-NC (CLARIN PUB)

Tekstin ominaisuudet

kieli: ersä

haastatteluajankohta: ??1903, ??1911

haastateltava: Малай-баба::malaj-baba

haastattelija: Ignatij Zorin

paikkakunta: väzofka (Vjazovka), Bez.

Bugul'ma, Gouv. Samara (Село Вязовка
Елховского района Куйбышевской
обл.)

paikkakunta alkukielellä: [tyhjä]

paikkakunta venäjäksi: [tyhjä]

tekstin otsikko alkukielellä: Вай, Нишке
пазось, Шки пазось

tekstin otsikko käännöksenä: Nischke-pas,
der Himmelsgott

julkaisun nimi: Mordwinische Volksdichtung,
V Band. SUST 161

Korpissa olevaa metaa

kieli: ersä
haastatteluajankohta: ??1903, ??1911
haastateltava: Малай-баба::malaj-baba
haastattelija: Ignatij Zorin
paikkakunta: vāzofka (Vjazovka), Bez.
Bugul'ma, Gouv. Samara (Село Вязовка
Елховского района Куйбышевской
обл.)
paikkakunta alkukielellä: [tyhjä]
paikkakunta venäjäksi: [tyhjä]
tekstin otsikko alkukielellä: Вай, Нишке
пазось, Шки пазось
tekstin otsikko käännöksenä: Nischke-pas,
der Himmelsgott
julkaisun nimi: Mordwinische Volksdichtung,
V Band. SUST 161
numero: 161
julkaisija: Suomalais-Ugrilainen Seura
julkaisuvuosi: 1977
julkaisupaikka: Helsinki
editointivuosi: 2016
editoija: Olga Erina
tekstilaji: Ersänische Lieder verschiedenen
Inhalts
kommentti saksaksi: Die göttliche Beratung,
die Erschaffung der Ersänen und die
Errichtung der ersänischen
Lebensordnung auf Erden
tekstin sivunumeroväli: 2-12
käyttöoikeus: NC
tekstin tila: scientificUse
tekstin tyyppi: fieldwork
tekstin numero: 1
virkkeen tyyppi: li
transkriptio: ombot ks t akanzo vezorgoñ
käännös saksaksi: seine andere Tochter
Vezorgo
sivunumero: 0002
rivin numero sivulla: 15

Lisää sisaruskorpuksia

▼ ERME (2)

Mokša/Moksha

Ersä/Erzya

▼ Fenno-Ugrica (10)

Ersä

Hanti

Inkeroinen

Itämeri

Länsimari

Mansi

Mokša

Selkupp

Tundranenetsi

Vepsä

▶ English / Englanti (28)

▶ Deutsch / Saksa / German (2)

▶ Français / Ranska / French (1)

▶ Español / Espanja / Spanish (1)

▶ Русский / Venäjä / Russian (7)

▶ Helsinki Corpus of Swahili 2.0 (HCS 2.0) (4)

▼ SUS-kenttätyö (näyte) (3)

SUS-kenttätyö: ersä (näyte)

Lisää työkaluja: linkki revizoriin

иневедекс	лымбась, а ней сонзэ лангсо неявисть ламодо-ламо пултонь шапк
иневедь	яки чирестэ чирес, лымбакстни штакс кенерезь розесь паксянтэ
ERSÄ	
иневедьга	якам- сто, зярдэ ламо чить а неяви берёкоськак.
иневедть	.
иневедь	ланга чи, омбоце—'берёк яла арась.
иневедь	мон неинь вагонсто ванозь.
иневедь	чиресэ аволь покш ошонтень.
иневедь	.
иневедь	чиресэ.
иневедесь	.
иневедесь	вейс совиль менеленть марто, нато^ а содавить, козонь прядови
иневеде^сь	, косто ушодови ме-.
иневедьсэнтъ	.

päiväys: *[tyhjä]*

kirjailija: Ėrdeli, Vladimir Georgievič

tekstin otsikko: Географиâ : 1-ce pel'ks :

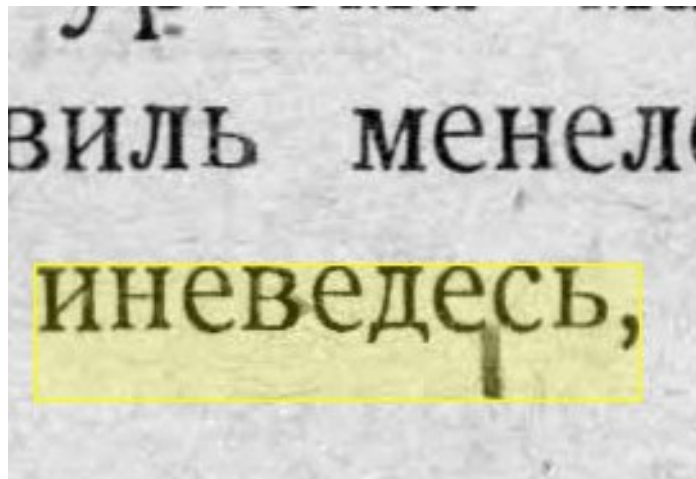
vasen' školan' 3 klasso tonavtnema kniga

toimittaja: *[tyhjä]*

kieli: ersä

[Katso asiakirja Revizorissa](#)

Lisää työkaluja: korjataan revizorissa



ошонтень. мон варштывь вальмава ды талакадрывь мазыг-
денть. Мэнь икеле сраговсь певтеме сэнь иневедь.
Поездэсь лоткась. Мон лисинь пси вагонстонть. **Иневедь**
пельде пувась экше, салов варма. Аволь ламо шкань
ютазь мон ульнинь иневедь чиресэ. Монь пильгем ало
булькаесь иневедесь. Берёконтень шалнозь эцесть пиже-
голубой волнатне, сэтьме урнома марто туильть мекев.
Ве пеле иневедесь вейс совиль менеленть марто, **натой** а
содавиль, козонь прядови **иневеде^сь**, косто ушодови **менелесь**.
Мон кузинь покш кев лангс, кона аштесь иневедьсэнть.
Ведесь перька ульнесь пек ванькс, пачканзо неят парсте.
4* 51

Kenttäyöntekijöistä wikipedia

https://myv.wikipedia.org/wiki/Паасонен_Хейкки

Getting Started baakoeh-r chm-r kyv-r saan-r sanat-r sonad-r valks-r yrk-r Read with Neahta...

Rueter Кортнемам Аравтомат Бета Мезе мельга мон ванстнян Монь путовкстнэ Лисемс

Лопа Кортнема Ловномс Витнемс-петнемс Витнемс-петнемс лисьмапрянзо Седе ламо Вешнемс

Паасонен Хейкки

Википедиясто материал - аорев содамкундосто

Паасонен Хейкки (суом. *Heikki Paasonen*, 1865 якшамковонь 2 чи Миккели ош — 1919 умарьковонь 24 чи Хельсинки ош) — келень содый-тейй, фольклорист, омбо масторонь кельс ютавтыця.

Чачсь Суоми масторонь Миккели ошсо 1865 иень якшамковонь омбоце чистэ. Университетэнь прядомадо мейле (1889) сыргась эрзятнень-мокшотнень юткс келень ды фольклор материалонь пурнамо. Келев-валов материалтнэсэ пурнась 1889-1912 иетнестэ аволь ськамонзо, ульнестэ лездыцянозяк. Малав весе те пурназь материалось ульнестэ ютавтозь немецень кельс ды лиссь Финно-Угрань вейксэндявксонь публикациясо кемень башка томсо 1891-1981 иетнестэ (в.вейсэ 4.552 лопа).

1889—1890 иетнестэ пурнась велетнева эрзятнеде ды мокшотнеде материал. 1898—1902 иетнестэ эрзятнень ютксо, истяжо **маринь**, татаронь-мишарень, хантонь ютксо. Пурнась эрва мезе **суоми-угрань раськетнеде**: эрямодост, кельдест, оршамопельдест, истяжо явсь мель **ветькень келентень**. Эрзянь ды мокшонь кельтнестэ мусь 20-те ламо кезэрень ветькень валт. Учёноенть мелензэ коряс, неть валтнэ совавсть эрзянь ды мокшонь кельтнес **XIII пингеденть** икеле.

Кода келень ванкшныця ваннось эрва-кодат малавикс кевкстемат **суоми-угрань кельтнестэ**, сюлмавкст суоми-угрань ды тюрконь кельтнень ютксто, сынст вайгельксэнь касомаст. 1908 иестэ нолдась **Мадярсо ветьке-мадярсо**-немецень валкс, конатась турецень кельс ютавтозель ды нолдазель Стамбул ошсо, 1974 иестэ одс нолдазь **Мадярсо**.

Паасонен Хейкки
 <div>200px</div>
Чачома лем: <i>суом. Heikki Paasonen</i>
Важодема ёрокчизе : келень содый-тейй, фольклорист, омбо масторонь кельс ютавтыця
Чачома чись : якшамковонь 2 чи 1865
Чачома тарка : Миккели ош, Суоми Мастор
Гражданчи : Суоми
Кедьалксчи : Россиянь Империя
Кулома чись : умарьковонь 24 чи 1919 (54 иеть)
Кулома тарка : Хельсинки ош

Kenttätöntehtijöistä wikipedia: Zorin

Лопа

Кортнема

Ловномс

Витнемс-петнемс

Витнемс-петнемс лисьмапрянзо



Седе ламо

Вешнемс



Зоринэнь Игнатий

Википедиясто материал - аорев содамкундосто

Зоринэнь Игнатий — эрзянь валморонь ладсий, ёвксонь морыця-лаиця.

Те шкас лияли мокшэрзянь литературань историйс парсте атак совавто литераторкс. Содамочись Зоринэнь эрямодо неень шкас пачкодсь пек аламо. Содазь ансяк чачома тарказо — Самара губерниянь Бугурусланской уездэнь эрзянь **Вечкань велесь**. Хейкки Паасонен марто сон васенцеде вастовсь **1898 иестэ**, зярдэ финнэнь учёноесь сакшнось Вечкань велев раськень ёвксонь кочкамо ды сёрмадсь пельдензэ фольклор жанрань **сядодо** ламо произведения. Конатнень ютксто комсте ламотне авторскойть, эсензэ поэмат-ёвкст. Зоринэнь кодаяк а маштови явомс мокшэрзянь народной литературанть эйстэ. Сон ульнесь аволь умонь **пингень** лаицяк, а сёрмас-грамотас содыця валморозеекс. Ламо эсь поэманзо-ёвксонзо сон лемдинзе народной поэзиянь традиционной жанрань лемсэ — **морокс**. Паасоненнэнь кучозь сёрманзо эйстэ, конат те шкас вастневить **Хельсинкисэ**, Финнэнь-Угрань Вейсэндявксонь архивсэ, парсте неяви: сон, кода литератор ульнесь мокшэрзянь раськень коень-кирдань чарькоди ломань, ташто пингень озномань ванстыця. Кода эсь Пазнэнь кемиця ломань, Зорин арасель христианской озномань каршо молиця ломань. Комсте ламо поэмасонзо арась вейкеяк истямо, косо конфликттнэ улелельть сюлмавозь антихристианской идея марто. Зоринэнь «каршо молемазо» - те мокшэрзянтнень раськень коень-традициянь вешнемань-ризнэмань ванстомась.

Зоринэнь Игнатий

Чачома лем: Игнатий Тимофеевич Зорин

Важдема валморонь ладсий, ёвксонь
ёрокчизе : морыця-лаиця

Чачома тарка: **Ташто Вечкань веле,**
Исаклань бую, Самара ёнкс

Työkaluja ja kokemusta jatkoon

- Morfologisia jäsentimiä
- Oikolukusovelluksia (aikuisten kirjoittaman kohennusta)
- Sanakirjoja
- Revizori (OCR korjausta)
- Mainontaa/aktivointia (wikipedia)

- Yhteistyötä
- Työpajoja

Kiitos!

Hyödyllisiä osoitteita:

<https://korp.csc.fi>

<http://ocrui.lib.helsinki.fi/>

<http://giellatekno.uit.no/index.fin.html>