Fieldwork and Early Literary Texts
‘Kenttätyötä ja varhaiskirjallisia tekstejä’

Tuoda Paasonen korpiin

21.4.2017 Suomalais-Ugrilaisen Seuran kokous

Fieldwork and **E**arly **L**iterary **T**exts (FELT)
‘Kenttätyötä ja varhaiskirjallisia tekstejä’

Jack Rueter,
Nykykielten laitos,
Helsingin yliopisto

‘Kenttätyötä ja varhaiskirjallisia tekstejä’

Suomalais-Ugrilainen Seura

Yhteistyötä

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

‘Kenttätyötä ja varhaiskirjallisia tekstejä’

- Aineistot
- Työvuo
- Työkalut

Aineistot

Suomalais-Ugrilainen Seura

Varhaiskenttätyötä 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.

Disambiguaatio 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+PrsPrc+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
аштить  аштемс+V+IV+Der/ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScSg3    0.000000
аштить  аштемс+V+IV+Der/ы+Der/NomAg+N+Pl+Nom+Indef+Der/Cop+Prs+ScPl3    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
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
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
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+PrsPrc+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
аштить  аштемс+V+TV+Der/ы+Der/NomAg+N+Pl+Gen+PxSg2+Der/Cop+Prs+ScSg3    0.000000
аштить  аштемс+V+TV+Der/ы+Der/NomAg+N+Pl+Nom+Indef+Der/Cop+Prs+ScPl3    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
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
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
аштить  аштемс+V+TV+Ind+Prt1+ScSg2      0.000000
аштить  аштемс+V+TV+Imprt+ScSg2+OcPl3   0.000000
аштить  аштемс+V+TV+Ind+Prt1+ScSg2+OcPl3        0.000000
аштить  аштемс+V+TV+Ind+Prs+ScPl3       0.000000
аштить  аштемс+V+TV+Imprt+ScSg2+OcPl3   0.000000
аштить  аштемс+V+TV+Ind+Prt1+ScSg2+OcPl3        0.000000
аштить  аштемс+V+TV+Imprt+ScSg2+OcPl3   0.000000
аштить  аштемс+V+TV+Ind+Prt1+ScSg2+OcPl3        0.000000

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

Disambiguaatio 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

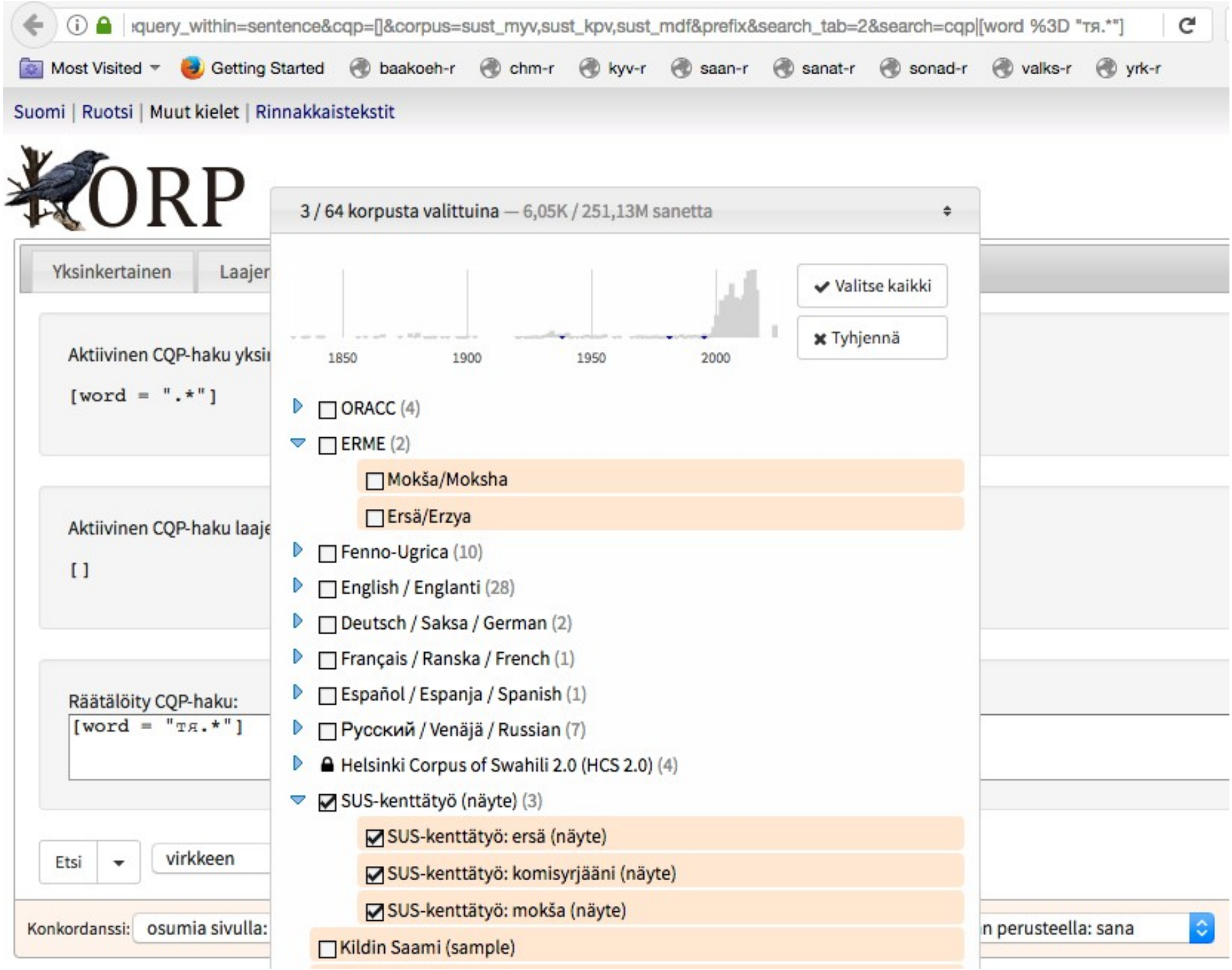
Rajoitekielioppi
(esimerkiksi Giellateknolla tai CSC:llä)

Korp

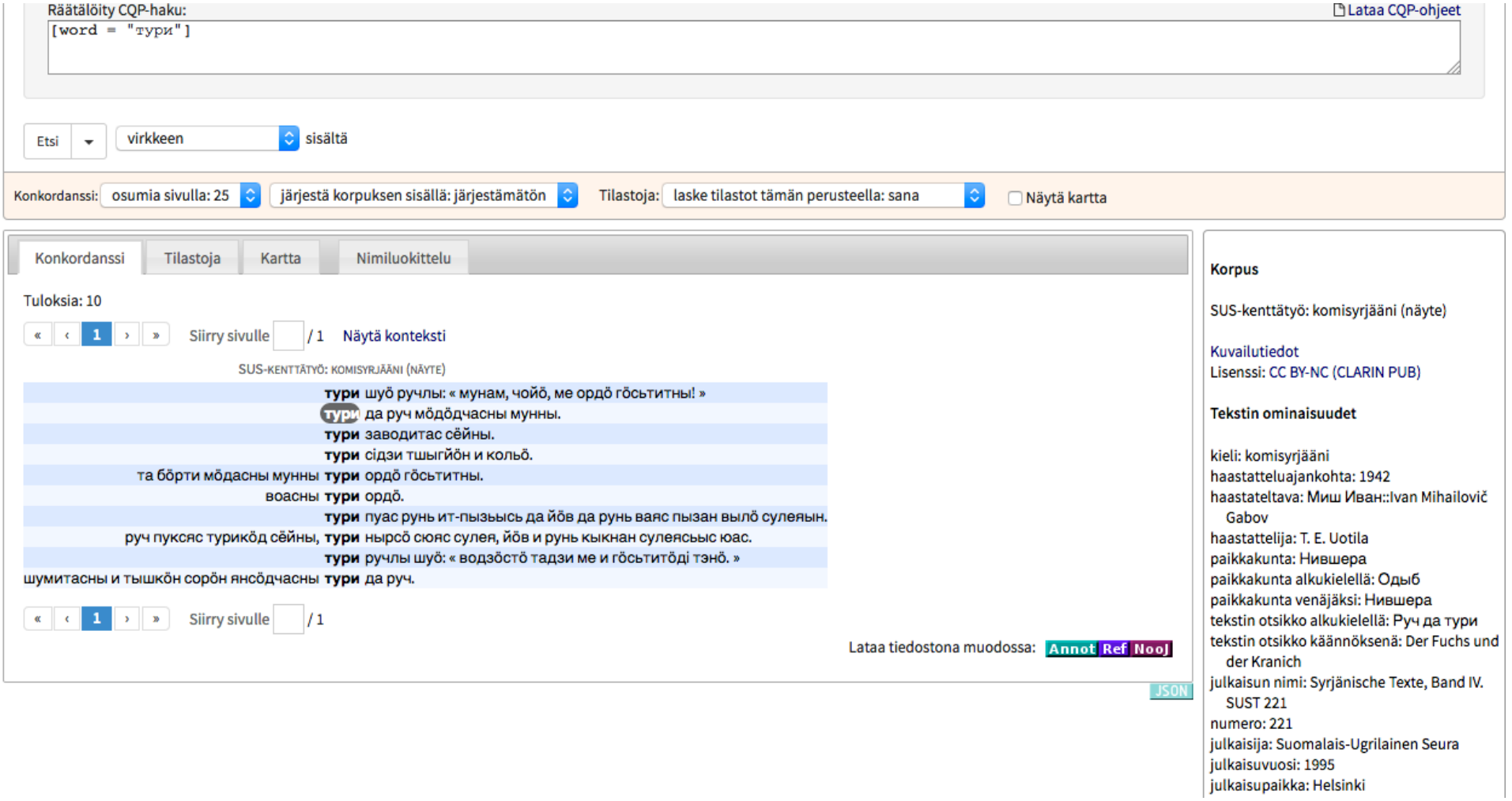
```
<sentence paragID="1" sent="3" pgNo="8" pgLi="3" orig_string="turi̯ da ruťš me̯de̯ťťšasni̯ munni̯."
          deu="Der Kranich und der Fuchs gehen los.">
<w word="тури" lemma="тури" pos="N" msd="N.Sg.Nom" sID="1" orig_string="turi̯"/>
<w word="да" lemma="да" pos="CC" msd="CC" sID="2" orig_string="da"/>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

Korp
XML: sivunumero, sivurivi, tyyppi

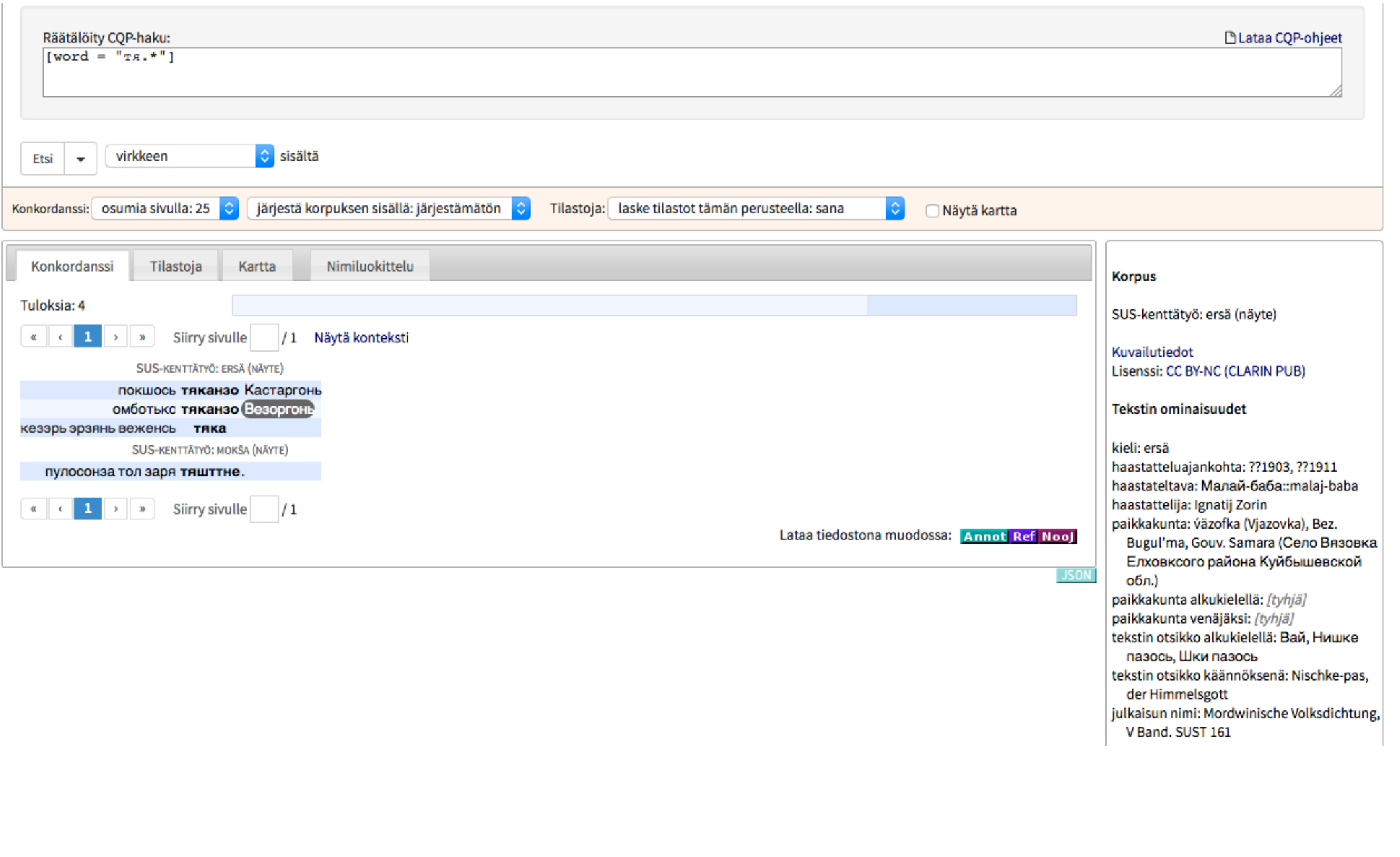
```
<sentence pgNo="0002" pgLi="15" type="li" orig_string="omboťks ťakanz
<w word="омботькс" lemma="омботькс" pos="Det" msd="Det" sID="1" orig_
<w word="тяканзо" lemma="тяка" pos="N" msd="N.Sg.Gen.PxSg3" sID="2" o
<w word="Везоргонь" lemma="Везоро" pos="N" msd="Sem/Divinity.N.Prop.
</sentence>
```


query_within=sentence&cqp=[]&corpus=sust_myv,sust_kpv,sust_mdf&prefix&search_tab=2&search=cqp|[word %3D "тя.*"]
Most Visited
Getting Started
baakoeh-r
chm-r
kyv-r
saan-r
sanat-r
sonad-r
valks-r
yrk-r
Suomi | Ruotsi | Muut kielet | Rinnakkaistekstit
KORP
3 / 64 korpusta valittuina — 6,05K / 251,13M sanetta
Yksinkertainen
Laajer
Valitse kaikki
Tyhjennä
1850
1900
1950
2000
Aktiivinen CQP-haku yksi
[word = ".*"]
Aktiivinen CQP-haku laaje
[]
Räätälöity CQP-haku:
[word = "тя.*"]
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)
Etsi
virkkeen
Konkordanssi:
osumia sivulla:
n perusteella: sana

Pari hakua korpista



Ylijumalan toinen, tytär

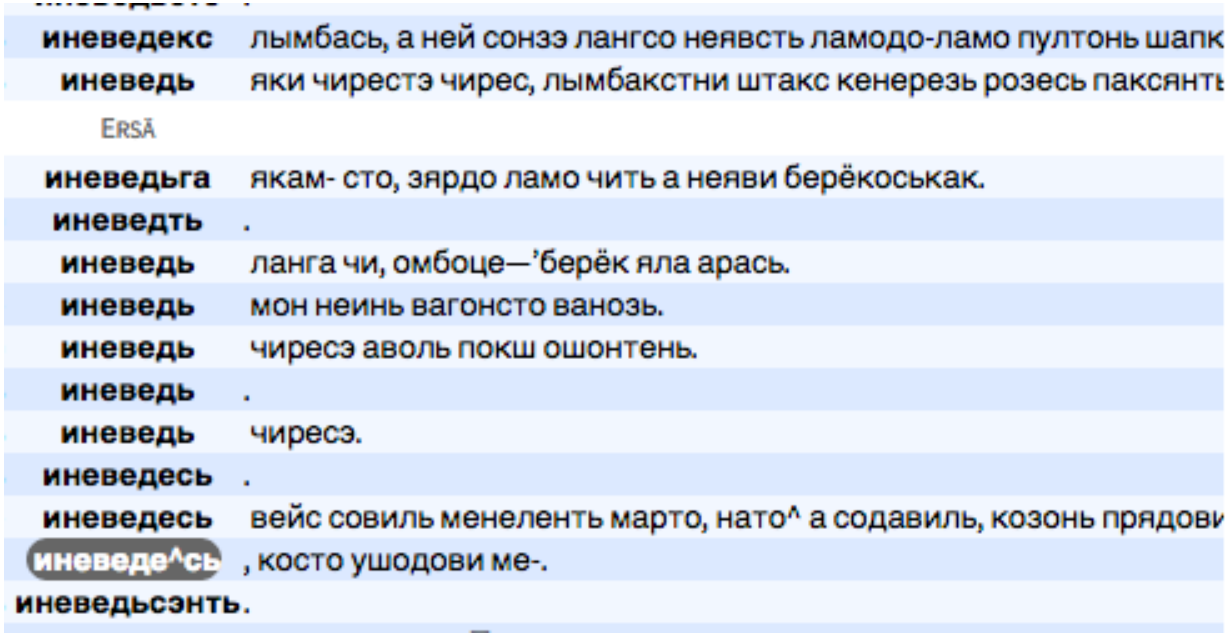


Korpissa olevaa metaa

kieli: ersä
haastatteluajankohta: ??1903, ??1911
haastateltava: Малай-баба::malaj-baba
haastattelija: Ignatij Zorin
paikkakunta: ́väzofka (Vjazovka), Bez. Bugul'ma, Gouv. Samara (Село Вязовка Елховксого 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 ksʲt akanʒo ́vezorgoń
käännös saksaksi: seine andere Tochter Vezorgo
sivunumero: 0002
rivin numero sivulla: 15

Lisää sisaruskorpuksia

Lisää työkaluja: linkki revizoriin

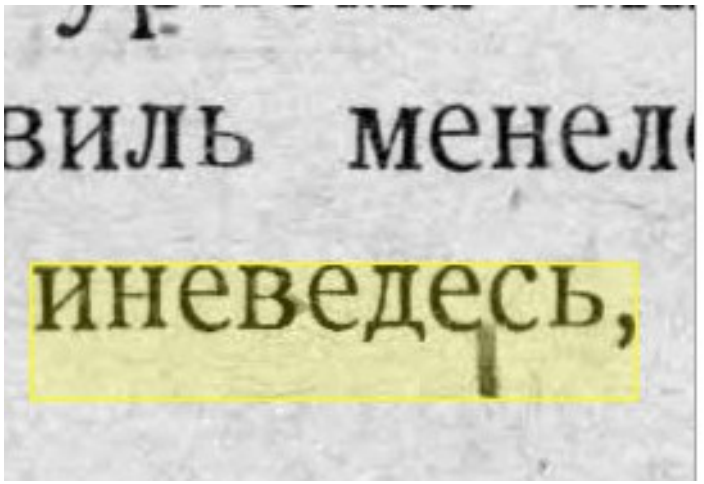


иневедекс	лымбась, а ней сонзэ лангсо неявсть ламодо-ламо пултонь шапк
иневедь	яки чиресте чирес, лымбакстни штакс кенерезь розесь паксянть
ERSÄ	
иневедьга	якам- сто, зярдо ламо чить а неяви берёкоськак.
иневедть	.
иневедь	ланга чи, омбоце—'берёк яла арась.
иневедь	мон неинь вагонсто ванозь.
иневедь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

päiväys: *[tyhjä]*
kirjailija: Èrdeli, Vladimir Georgievič
tekstin otsikko: Geografiâ : 1-ce pel'ks : vasen' školan' 3 klasso tonavtnema kniga
toimittaja: *[tyhjä]*
kieli: ersä

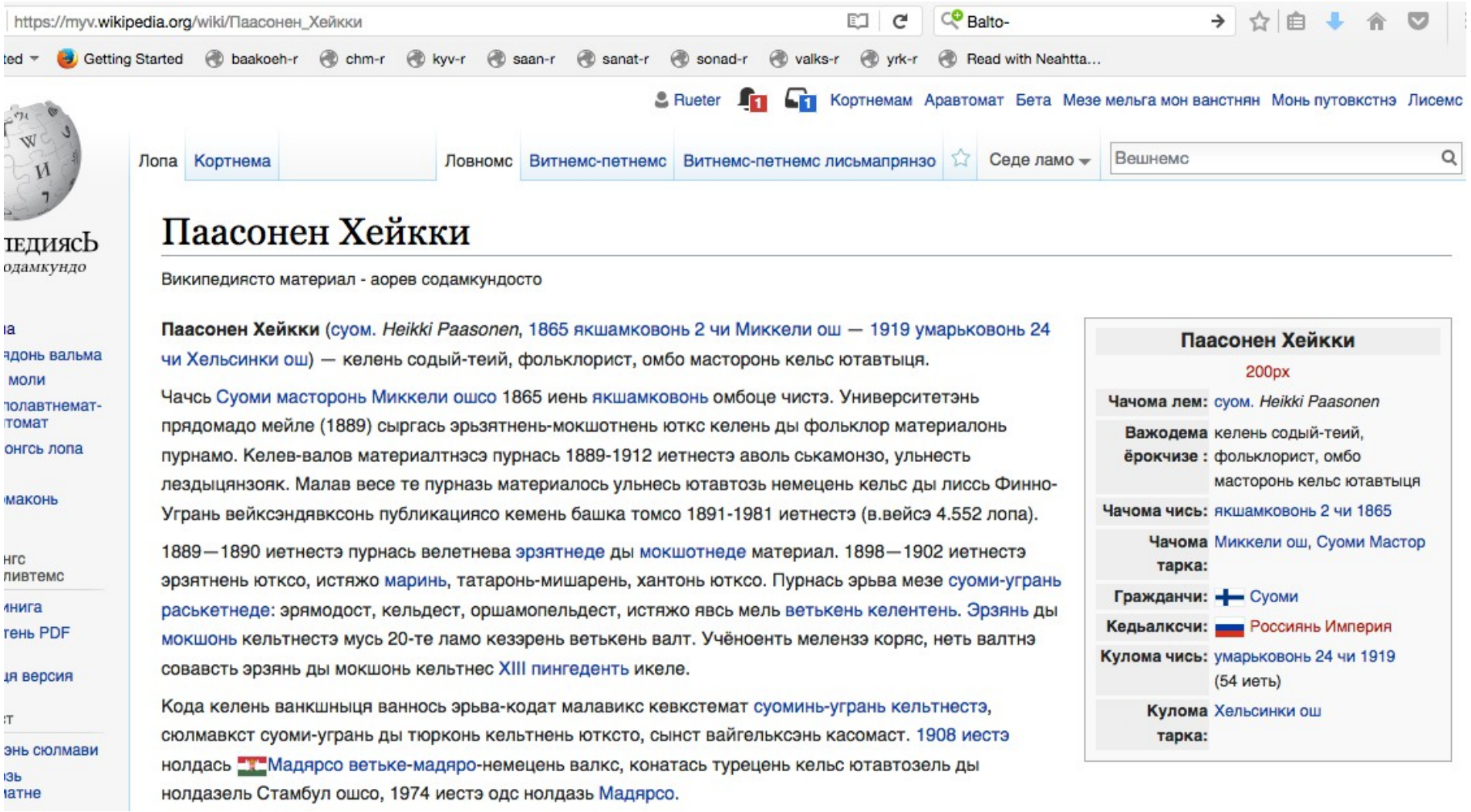
Katso asiakirja Revizorissa

Lisää työkaluja: korjataan revizorissa

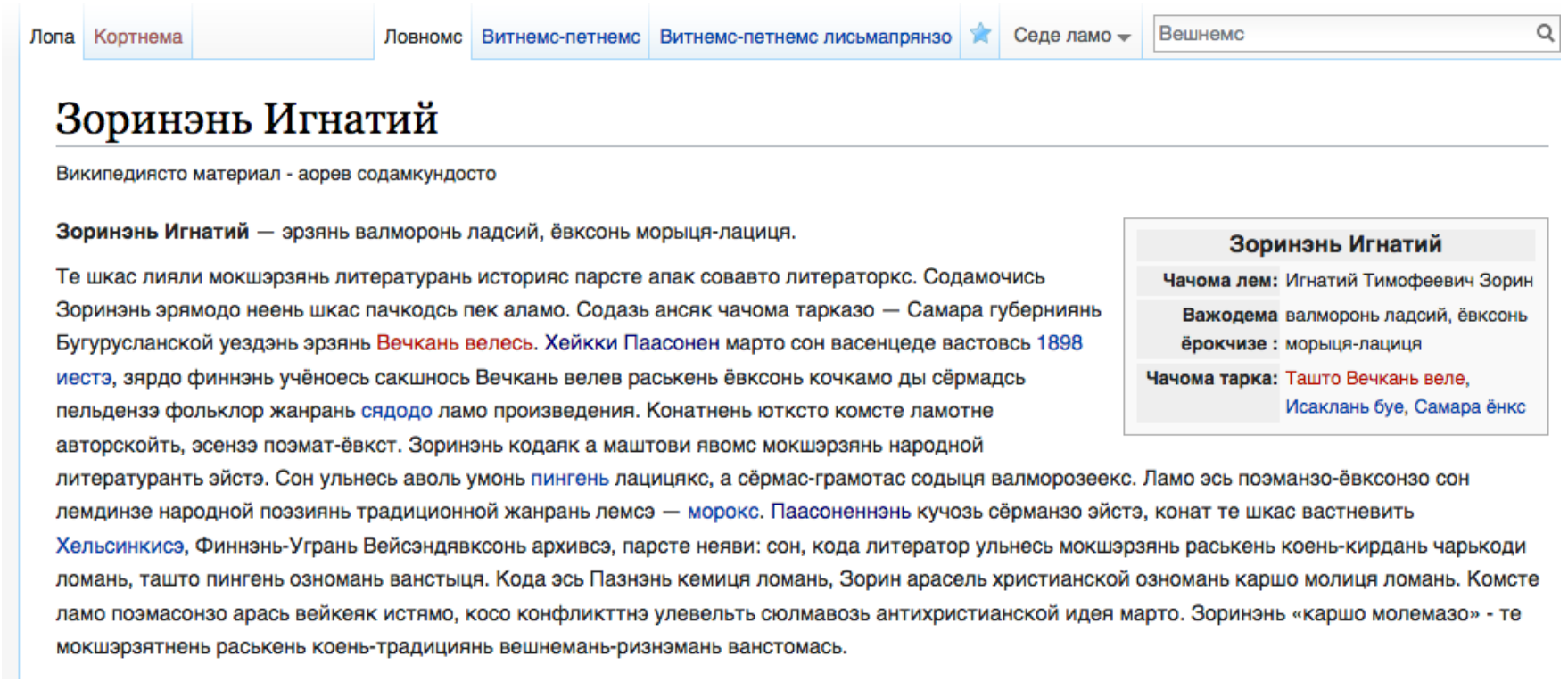


ошонтень. Мон варштынь вальмава ды талакадынь мазы-
день. Мэнь икеле сравтовсь певтеме сэнь иневедь.
Поездэсь лоткась. Мон лисинь пси вагонстонть. Иневедь
пельде пувась экше, салов варма. Аволь ламо шкань
ютазь мон ульнинь иневедь чиресэ. Монь пильгем ало
булькаесь иневедесь. Берёконтень шалнозь эцесть 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

Kenttätyöntekijöistä wikipedia



Kenttätyöntekijöistä 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-лациця.

Зоринэнь Игнатий	
Чачома лем:	Игнатий Тимофеевич Зорин
Важодема ёрокчизе :	валморонь ладсий, ёвксонь морыця-лациця
Чачома тарка:	Ташто Вечкань веле, Исаклань буе, Самара ёнкс

Те шкас лияли мокшэрзянь литературань историяс парсте апак 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, зярдо финнэнь учёноесь сакшнось Вечкань велев раськень ёвксонь кочкамо ды сёрмадсь пельдензэ фольклор жанрань сядодо ламо произведения. Конатнень юткстo комсте ламотне авторскойть, эсензэ поэмат-ёвкст. Зоринэнь кодаяк а маштови явомс мокшэрзянь народной литературантъ эйстэ. Сон ульнесь аволь умонь пингень лацицякс, 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 улевельть сюлмавозь антихристианской идея марто. Зоринэнь «каршо молемазо» - те мокшэрзятнень раськень коень-традициянь вешнемань-ризнэмань ванстомась.

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