**Eesti Rahvusraamatukogu**

**Teadusnõukogu koosolek**

**6. septembril 2017**

**Milleri salongis V. korrusel**

**kl 13.30 - 15.30**

Päevakava:

1. Suured tekstiandmed humanitaarteadustes – kellele ja kuidas?

   Tartu Ülikooli eesti ja üldkeeleteaduse instituudi vanemteadur dotsent **Liina Lindström**

2. Keeleseire ja leksikograafilise analüüsi tööriistad.

   Eesti Keele Instituudi teadussekretär-arvutileksikograaf **Jelena Kallas**

3. Arutelu.

**Kohal viibisid:**

Janne Andresoo, Indrek Ibrus, Juhan Kreem, Aija Sakova, Urmas Sutrop, Tõnu Tender, Piret Lotman, Raivo Ruusalepp, Kai Välbe. Kristel Veimann, Jane Makke, Urmas Sinisalu.

**J. Andresoo** avas koosoleku tõdemusega, et RRil on üha suurem digitaalsete tekstide kollektsioon, mida teadlastele uurimiseks pakkuda. Vaja oleks ühiselt mõelda, kuidas saaks sisu teadlastele nii pakkuda, et sellest võimalikult palju kasu oleks.
Ka **R. Ruusalepa** sõnul on tähtis teadlaste vaade, mida raamatukogu saaks selles vallas teha. Tekstikaeve teemaga on RRis tegeletud ca 9 kuud. Moodustatud on asutustevaheline töörühm, mis on uurinud nii juriidilisi piiranguid kui tehnilisi võimalusi. On peetud seminare praktikutega. Käesolev kokkusaamine on järgmine samm selles suunas, et saada aru, milliseid teenuseid tekstikaevega tegelev teadlane raamatukogult sooviks. Ühiselt peaks teadusnõukogus arutama, kas ja kuidas raamatukogu saaks osaleda teadusprojektides.

## 1. Suured tekstiandmed humanitaarteadustes – kellele ja kuidas?

**Liina Lindström** eristas oma ettekandes kahte suuremat tekstiandmete sihtrühma.

Esiteks humantitaar- ja sotsiaalteadlased, üliõpilased ning asjaarmastajad. Seda rühma iseloomustavad vähesed tehnilised oskused ja eeltöödeldud materjalide vajadus.

Teiseks keeletehnoloogid ja arvutilingvistid, st andme- ja tekstikaeve spetsialistid. See on potentsiaalne edasiarendajate rühm, kes teeb asju oma tehnilistest huvidest lähtuvalt. Ka nemad

ootavad andmeid võimalikult lihtsal kujul, kuid korrastatult. Oluline on kiire ja lihtne ligipääs, et ei peaks aega kulutama andmete kokku otsimisele.

Kummalgi rühmal on oma vajadused ja digitaalhumanitaaridele tuleb andmed esitada erineval viisil. Ettekandja arvates oleks vaja, et tekstiandmed oleksid esitatud kõigest puhastatult lihttekstina, nt txt-formaadis, mitte pdf-ina. Tasub panustada OCR kvaliteedi tõstmisesse, veebitekstide ja trükiste digitaalsetesse versioonidesse. Metaandmed on kusjuures peaaegu sama tähtsad kui sisu.

Analüüsi märksõnad oleksid: tuvastusvigade automaatne parandamine; lemmatiseerimine; visualiseerimine (näiteks ajateljed, et millal mõni märksõna tekkis ja kadus); vektorsemantika rakendamine päringute tegemisel; tulemuste allalaadimise võimalus (txt+csv); kvantitatiivsete andmete kasutamise võimalus. Piltlikult – kui ei ole veebis, siis pole (eriti) vajagi.
Spetsialistile on oluline, et ta saaks raamatukogu ressurssidega töötada oma töökohal. Järelikult oleksid vajalikud vastavad vahendeid, millega tekstidele raamatukogu väliselt ligi saada ja tulemusi veebipõhiselt edastada.
Kokkuvõtteks oleks vaja Digarit edasi arendada tekstipõhiselt otsimiseks. Tekstianalüüsi tulemused peaksid olema allalaetavad (txt, csv).

Sihtgruppe on palju ja nende vajadused on erinevad. Üldine on vajadus autoriseeritud töökoha järele, kus oleks võimalik interaktiivselt töötada ja tulemusi alla laadida.

J. Kreem: pilt ei asenda küll teksti, siiski pole digiteeritud piltide olemasolu mõttetu, vaid vahel ainuvõimalik variant.

R. Ruusalepp: Digaris pole midagi, mis lihtsalt pildina skaneeritud ja poleks tekstina tuvastatav. Probleem on, et moodsate keeletehnoloogia reeglite järgi ei tunta vahel tekste ära ja vanemaid tekste peaks oskama selles mõttes toimetada.

L. Lindstöm: keeletehnoloogid oleksid tänulikud, kui saaksid selliseid vanade tekstide korpusi.

U. Sutrop: teadlane ei saa tegeleda, kui materjal ei ole veebis kättesaadav. Siin algavad probleemid autorikaitsega.

I. Ibrus: veebis võib ju olla, aga peab olema ka allalaetav.

U. Sinisalu: üksikutel juhtudel võib uurimise eesmärgil tekste anda, aga veebi üles riputamine on võimatu. Pole ju võimalik vahet teha, kas on teadlane (kasutab teadusliku uurimise eesmärgil) või nö asjaarmastaja.

U. Sutrop: teadlased on ikka institutsionaalselt määratletud.

R. Ruusalepp: kogunenud on juba üle 1400 faili (RRis), ainult üksikud kirjastajad neid ei anna, paberil säilituseksemplarid skaneerime siiski kohe ise otsitavasse tekstiformaati. Seis on hetkel rahuldav, kuid käib pidev tasakaalu otsimine RRi kui usaldusväärse säilitaja ja kättesaadavaks tegija vahel – mõlemad funktsioonid ei tohiks kannatada.

## 2. Keeleseire ja leksikograafilise analüüsi tööriistad

**Jelena Kallase** ettekanne käsitles tekstiandmete (korpuste) kasutamist sõnaraamatute koostamisel Eesti Keele Instituudi näitel. Ta esitas mitmeid ettepanekuid autoriõiguse seaduse muutmiseks, millest radikaalseim oli see, et kõiki andmeid võiks praeguse teadustöö erandi kõrval vabalt kasutada ka ärilistel eesmärkidel. Uute veebikorpuste loomine toimub koostöös ettevõttega Lexical Computing Ltd, veebikorpused on kättesaadavad korpuspäringusüsteemi *SketchEngine* kaudu. Eesti keele korpused on kättesaadavad samuti korpuspäringusüsteemi KORP ja portaali Keeleveeb (www.keeleveeb.ee) kaudu. Peamine probleem, et puuduvad keeleseireks vajalikud ressursid, viimane veebikorpus loodi aastal 2013, korpuste uuendamine ei ole süstemaatiline. Peaaegu olematu on näit ilukirjanduse korpus.

Korpuspäringusüsteemides on alati viide algallikale olemas. Neologismide analüüsi näide: iga kolme kuu tagant luuakse vahepeal laekunud tekstidest korpus ja analüüsitakse varasematega, et tuvastada uusi sõnu, mis on kasutusse tulnud. Seni eesti keele jaoks sellised vahendid puuduvad (*neologism tracker*).

Näide paralleelkorpuste põhjal tõlkevastete tuvastamisest on http://context.reverso.net/translation/

Korpuste loomise allikad: veeb kui korpus (*Webcorp*); korpuse loomine veebist (*WebBootCat*); korpuse loomine tekstist (*Corpus architect*).

EKI-l oleks huvi kasutada RR trükifaile jooksvalt korpuse loomiseks/täiendamiseks.

EKI kogemus ütleb, et sageli autor ei tea, kus tema trükifailid tegelikult asuvad. RRis lastakse ka digiteerida, aga tekstiks muutmine ja märgendamine on väga vaevaline.

EKI huvi puudutab kõne all olevate tekstide säilitamise ja taaskasutuse problemaatikat ning vajalik on tugi õigusaktide ja regulatsioonide vallas, et luua uusi korpusi.

Andmekaeve ei ole tuletatud teose loomine! Arvi Tavasti meelest ei tohiks olla vahet, kas teksti õpib inimene või masin. Praegu ei saa ka teaduslikuks kasutamiseks tasuta kuigi lihtsalt tekste kätte.

Keeleseireks vajalik taristu koosneb vahenditest: 1. keeleandmete kogumiseks; 2. töötlemiseks; 3. korpuspäringusüsteemist.

Kasutuslugu võiks välja näha järgmine: kasutaja märgib ära teda huvitavad teosed (nt. Luts, Tammsaare); märgitud failid läbivad töötluse ja sisenevad korpuspäringu süsteemi, kus sellega saab juba analüüsi edasi teha (tasuks arendada KORPi, kuna see on kohalik vahend ja tasuta saadaval); korpus on loodud ja seda on võimalik kasutada.

**3. Arutelus** püüti leida lahendusi tõstatatud probleemidele.

J. Kallas: korpuste tekitamise olukord pole kiita, kuigi autoriõigus ei ole korpustes rikutud. RR on potentsiaalne andmeandja korpuste loomise jaoks, näit uute sõnade automaattuvastamisel (EKI jaoks). Praegu lootus SäeSi peal, et seaduse toel hakkaks saama hästi toimetatud keelt korpustesse. Andmekaeve ei ole koguteose kasutus.

U. Sutrop: Rahvusraamatukogu on muutumas suureks digiandmete (korpuste) pakkujaks humanitaar- ja sotsiaalteadlastele (k.a keeletehnoloogia ja arvutilingvistika). Olemas on suur veebiarhiiv, jõudsalt kasvab säilituseksemplari seadusest tulenev kõikide eesti raamatute käsikirjade digiarhiiv jm. Selle andmestiku kasutamise lõppeesmärk on (näit sõnastike tegemisel) siiski äriline.

Ettekandes kõlanud ettepanek rahuldaks vaid keeleteadlasi. Kirjandusteadlane vajab terviktekste, talle lausepõhisest korpusest ei piisa. Seni on korpuste loomisel kõige rohkem puudu ilukirjanduslikest tekstidest.

RR: RRi digihoidla tarkvara ei võimalda veel vaba juurdepääsu. Iga korpuse puhul lepingu sõlmimine teadlasega ei ole väga praktiline stsenaarium.

J. Kallas: võtmesõna on taaskasutus!

U. Sinisalu: tegime uurijatele eraldi võimaluse tekste kasutada, aga olime sunnitud lõpetama, sest andmemaht läks liiga suureks. Vahel peakski olema KORP`i tarkvara.

J.Kallas: võimalus mitte mälupulgale salvestada, vaid kasutada osana süsteemist.

R. Ruusalepp: erinevatele kasutajatele tuleks luua erinevad instrumendid. Kas oleks huvi või vajadust liita Digari külge see võimalus ja kas perioodika või ilukirjandus?

J.Kallas: kuna juurde tuleb vähe kroolimiseks kõlbulikku materjali, siis huvi kahtlemata oleks.

A. Sakova: veebikoguga on pigem hästi. Huvi on rohkem vanema perioodika ja uuema ilukirjanduse vastu.

J. Makke: Digarist on juba praegu kätte saadavad 150 ilukirjanduslikku teost.

U. Sutrop: toetan ettekandes kõlanud ettepanekut, et võiks olla projekt, milles partneriteks RR, EKRK ja/või erasektor. Eelduseks õiguslike aspektide lahendamine.

J. Andresoo: see on hea mõte koostööks. Tuleks teha Haridus- ja teadusministeeriumile taotlus.

L. Lindström: selle koostöö raames võib laiendada ja samas täpsustada huviliste ringi.

R. Ruusalepp: miks mitte tekstide parandamisse vabatahtlikke kaasata, nn *crowdsourcing (rahvahange)*?

I. Ibrus: midagi on selles vallas juba tehtud, näiteks vast avatavas Filmimuuseumis.

L. Lindström: pigem näen siin väljundit üliõpilaste praktikatööna.

**Koosolek otsustas:**

1. Volitada RRi juhtkonda läbi viima konsultatsioone HTMi töörühmaga, mis valmistab ette järgmist keeletehnoloogia programmi sisu, et anda omapoolne sisend kavandatavasse uude riiklikku programmi.
2. Järgmine teadusnõukogu koosolek toimub 2018. aastal säilituseksemplari seaduse teemal.

Koosolekut juhatas Janne Andresoo

Kokkuvõtte tegid:

Raivo Ruusalepp

Urmas Sutrop

Kai Välbe

19.09.2017