Kadri Muischnek

Tartu ülikooli eesti ja üldkeeleteaduse instituudi arvutilingvistika dotsent ja arvutiteaduse instituudi keeletehnoloogia vanemteadur. Tema peamised uurimisvaldkonnad on korpuslingvistika ja eesti keele süntaktilise struktuuri modelleerimine

Keelekorpused – sama mitmekesised kui keel ise

Tänapäeva keeleteadus ja keeletehnoloogia toetuvad suurel määral elektroonilistele tekstikogudele – korpustele. See artikkel tutvustabki eesti keele korpusi ja nende kasutusvõimalusi.

Keelekorpus – mis see on

Lühidalt: keelekorpus on elektrooniline tekstikogu, koostatud keeleteaduse ja arvutilingvistika vajadusi silmas pidades. Tavalisemad ja levinumad on kirjaliku keele korpused ja siin ülevaates nendega piirdumegi.

Ajalooliselt saab korpuste puhul eristada kolme põlvkonda. Esimese põlvkonna korpused valmisid ajal, kui arvutimälu oli veel piiratud ressurss, mistõttu tuli hoolikalt valida, milliseid tekste korpusesse võtta. Oluline oli korpuse representatiivsus, mille all mõeldi seda, et korpust koostama asudes tuleb esmalt läbi mõelda, milliseid keele kasutusvaldkondi see korpus esindama hakkab. Näiteks sobib kirjaliku eesti keele 80ndate aastate korpus[1]. Siia otsustati võtta täiskasvanutele mõeldud trükitud ja toimetatud Eestis ilmunud tekstid, mis pärinevad aastatest 1984–1987. Välja jäeti seega käsikirjalised kirjalikud tekstid, väliseesti tekstid ja lastekirjandus. Edasi jagati selle perioodi eestikeelsed trükitud tekstid tekstiklassidesse, lähtudes laias laastus raamatukogude kataloogisüsteemist. Representatiivsus tähendabki seda, et korpusesse peab iga tekstiklassi tekste võtma proportsionaalselt nende tekstiklasside osakaaluga trükitekstide üldhulgast valitud perioodil.

[1] http://www.cl.ut.ee/korpused/baaskorpus/1980/

Teise põlvkonda kuuluvad suured, sadadest või isegi tuhandetest miljonitest sõnadest koosnevad korpused, mille puhul representatiivsust enam ei taotleta, aga on ka erandeid (nt British National Corpus[2]). Eesti keele selle põlvkonna korpus on koondkorpus, mis koosneb u 250 miljonist sõnast, mille hulgas ajalehekeel on tugevalt üleesindatud. Koondkorpuse allosa – tasakaalus korpus[3] – koosneb võrdses mahus ilukirjanduse, ajakirjanduse ja teaduse tekstidest.

Korpuste kolmanda põlvkonna moodustavad automaatselt veebist n-ö korjatud korpused. Automaatselt üritatakse seda tööd teha muidugi sellepärast, et inimtöö on kallis ja aeganõudev. Sellised korpused on väga suured, aga paratamatult ka veidi n-ö sodised. Eesti keele kolmanda põlvkonna korpuseks on etTenTen[4]. Kolmanda põlvkonna korpused, ka etTenTen, sisaldavad rohkesti interneti nn kasutaja loodud sisu, st foorumite, kommentaariumide, blogide jms tekste, mis esindavad, küll erineval määral, spontaanset kirjalikku keelekasutust ja on seetõttu keeleuurija jaoks huvipakkuvad.

Kõik eelnimetatud korpused sisaldavad tänapäeva kirjalikku eesti keelt. Lisaks koostatakse ka mitmesuguseid erikorpusi, mis sisaldavad vanemat keelekasutust, murdekeelt, lastekeelt, suulist keelt jne.

Eesti vana kirjakeele korpuse[5] vanima osa moodustavad Henriku Liivimaa kroonika eestikeelseid sõnu sisaldavad laused ja korpus ulatub ajaliselt 1889. aastasse, võimaldades seega uurida eesti kirjakeele arengut umbes 650 aasta vältel. Vanemad tekstid on morfoloogiliselt märgendatud (morfoloogilisest märgendamisest tuleb lähemalt juttu allpool); kõige vanemaid tekste on võimalik veebis ka tervikuna näha. Vana kirjakeele korpuse kohta saab lähemalt lugeda Valve-Liivi Kingisepa, Külli Prillopi ja Külli Habichti artiklist[6].

Eesti murrete korpus[7] ei ole tegelikult kirjaliku keele korpus, vaid sisaldab murdesalvestiste üleskirjutusi ehk litereeringuid. Korpus on varus-

---

2 http://www.natcorp.ox.ac.uk/

3 http://www.cl.ut.ee/korpused/grammatikakorpus/

4 http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html

5 http://www.murre.ut.ee/vakkur/Korpused/korpused.htm

6 Kingisepp, Valve-Liivi, Külli Prillop, Külli Habicht 2004. Eesti vana kirjakeele korpus: mis tehtud, mis teoksil – Keel ja Kirjandus, nr 4, lk 272–280.

7 http://www.murre.ut.ee/murdekorpus/

tatud otsimootoriga. Samuti on võimalik veebis kuulata murdelindistusi. Korpust tutvustab Liina Lindströmi ja kolleegide artikkel (Lindström jt 2001)[8].

Infot teiste Tartu ülikoolis koostatavate või säilitatavate korpuste ja tekstikogude kohta saab eesti ja üldkeeleteaduse instituudi keelekogude lehelt[9]. Eesti Keeleressursside Keskuse[10] kodulehele on koondatud info enamiku eesti keele korpuste kohta.

Korpuste märgendamine

Korpus lihtsaimal kujul sisaldab ainult tekste. Enamiku korpuste tekstidesse on siiski lisatud veel mingit infot, mida seal algselt ei olnud. Sellist korpuste lisainfoga rikastamist nimetatakse korpuse märgendamiseks. Lihtsaim märgendus on teksti struktuuri esitamine – märgitakse lausepiirid, pealkirjad, autorinimed, tabelid jms.

Eesti keele puhul on väga oluline morfoloogiline märgendamine, mille puhul igale tekstisõnale lisatakse tema algvorm ning info sõnaliigi ja grammatiliste kategooriate kohta. Näiteks on tekstisõna *kätega* kohta märgitud, et see on mitmuse kaasaütleva vorm sõnast *käsi*. Selliselt märgendatud korpusest saab juba otsida näiteks lauseid, mis sisaldavad sõna *käsi* ükskõik milliseid vorme või lauseid, mis sisaldavad ükskõik millist sõna kaasaütlevas käändes.

Järjest enam levib ka süntaktiline märgendamine, mille puhul igale tekstisõnale lisatakse info tema funktsiooni kohta selles lauses ning sageli ka süvasüntaktilist infot. Näiteks lauses *Tegin selle oma kätega* on sõnavormi *kätega* juures kirjas, et see on määrus. Tüüpiliselt on süntaktiliselt märgendatud korpused ka morfoloogiliselt märgendatud, nii et neist saab otsida näiteks lauseid, mis sisaldavad kaasaütlevas käändes määrusi.

Samuti on olemas semantiline märgendamine, mis lisab tekstile infot sõnade või fraaside tähenduse või pigem küll tähenduse mingi aspekti kohta.

---

8 Lindström, Liina, Varje Lonn, Mari Mets, Karl Pajusalu, Pire Teras, Ann Veismann, Eva Velsker, Jüri Viikberg 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. TÜ eesti keele õppetooli toimetised 17. Toim. R. Kasik, Tartu, lk 186–211.

9 http://www.keel.ut.ee/et/keelekogud

10 http://keeleressursid.ee/et/keeleressursid/tekstikorpused

Korpuste kasutamine

Korpuste kasutamiseks on kaks võimalust – laadida alla korpus tervikuna ja tegelda sellega siis ise kirjutatud või mujalt hangitud (palju on vastavat vabavara) tarkvara abil. Selline viis on valdavalt keeletehnoloogide ja informaatikute pärusmaa. Tartu ülikooli tänapäeva eesti kirjaliku keele korpused on enamuses allalaaditavad[11].

Teine võimalus on esitada korpusele päringuid kasutajaliideste kaudu. See, milliseid päringuid saab esitada, sõltub kasutajaliidese funktsionaalsusest ja sellest, kuidas korpused on märgendatud. Tartu Ülikooli arvutilingvistika uurimisrühma kodulehel oleva kasutajaliidese[12] kaudu saab sealsetele korpustele esitada lihtpäringuid – otsida tekstis esinevaid märgijadasid (sõnavorme, sõnaosi, sõnavormide järgnevusi), kasutades selleks ka regulaaravaldisi[13]. Sealsed korpused, peale ühe erandi, ei ole morfoloogiliselt märgendatud, kuid on lausestatud, st päringule vastuseks saab terviklaused.

Keeleveebi portaali[14] on koondatud suured eesti tänapäeva kirjaliku keele korpused ja neile saab seal esitada ühispäringuid. Kõik sealsed korpused on märgendatud morfoloogiliselt, tasakaalus korpus ka süntaktiliselt (süntaktilised funktsioonid) ja semantiliselt (ajaväljendid, nimeüksused).

Keeleveebi korpusepäring on kompleksne, st saab otsida sõnavormide / sõnade / sõnaliikide / grammatiliste kategooriate koosesinemisi ja järgnevusi. Lisaks lausepiiridele on märgendatud ka osalausepiirid, mis tähendab, et keelendite koosesinemisi saab piirata osalausega. Keeleveebi korpusepäringu koostamise kohta on valminud Tartu ülikoolis aine „Multimeedia" raames ka õppevideo[15].

Korpusele Keeleveebis päringu esitamiseks tuleb korpuste loendis esmalt klõpsata vajaliku korpuse nimel, korraga võib päringuks valida ka mitu korpust. Valitud korpus (või korpused) ilmub halli klahvina Keeleveebi avalehe ülaossa, klahvi „Sõnastikud" kõrvale. Sisestades otsitava

---

11 http://www.cl.ut.ee/korpused/

12 http://www.cl.ut.ee/korpused/kasutajaliides/

13 Regulaaravaldist võiks kirjeldada kui otsingul kasutatavat mustrit. Lihtsaim regulaaravaldis on sõna ise, aga regulaaravaldise teeb võimsaks just metamärkide kasutamise võimalus, täpsemalt vt http://www.cl.ut.ee/korpused/kasutajaliides/erispikker#reg

14 www.keeleveeb.ee

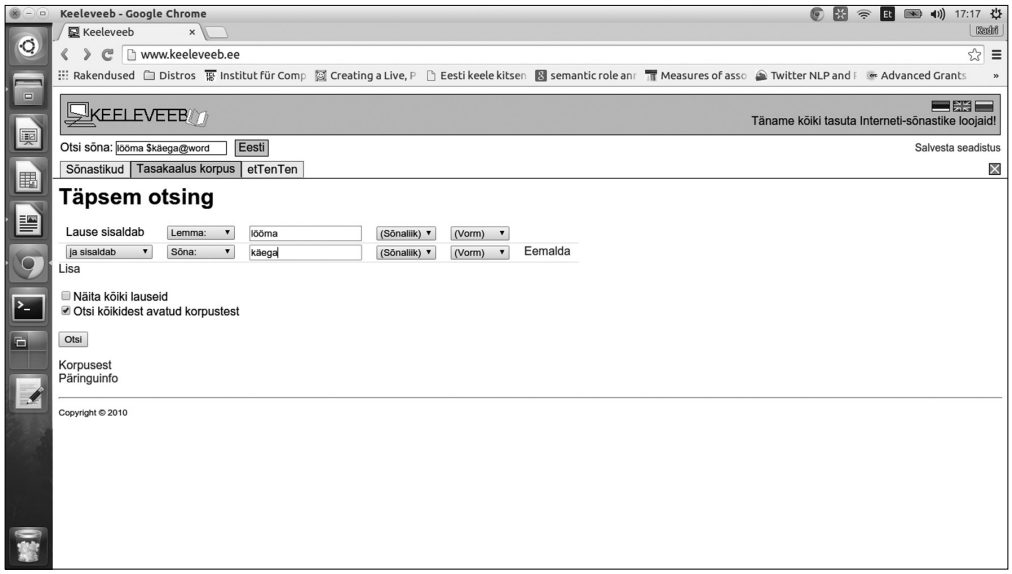
15 http://www.uttv.ee/naita?id=21284

sõna (algvormina) sealsamas olevasse aknasse, saame esitada selle ühe sõna kohta ühispäringu sõnastikesse ja valitud korpus(t)esse.

Komplekspäringu esitamiseks tuleb esmalt klõpsata ühe valitud korpuse klahvile ja avanenud lehel lingile „Täpsem otsing". Kui enne on valitud mitu korpust ja soovitakse korraga esitada päring kõigile neist, tuleb nüüd teha linnuke lahtrisse „otsi kõikidest avatud korpustest". Avanenud komplekspäringu lehel saab otsitavaid koosesinevaid keelendeid lisada või täpsustada, klõpsates viidal „Lisa", mis avab uue päringurea. Päring võib sisaldada metamärke ? (üks suvaline märk) ja * (suvaline hulk suvalisi märke, st ka mitte ühtegi märki). Näiteks otsides sõna *tegemat*a*, saame vastuseks nii tegemata kui ka *tegematta*; otsides *tegemat?a*, aga ainult *tegematta*.

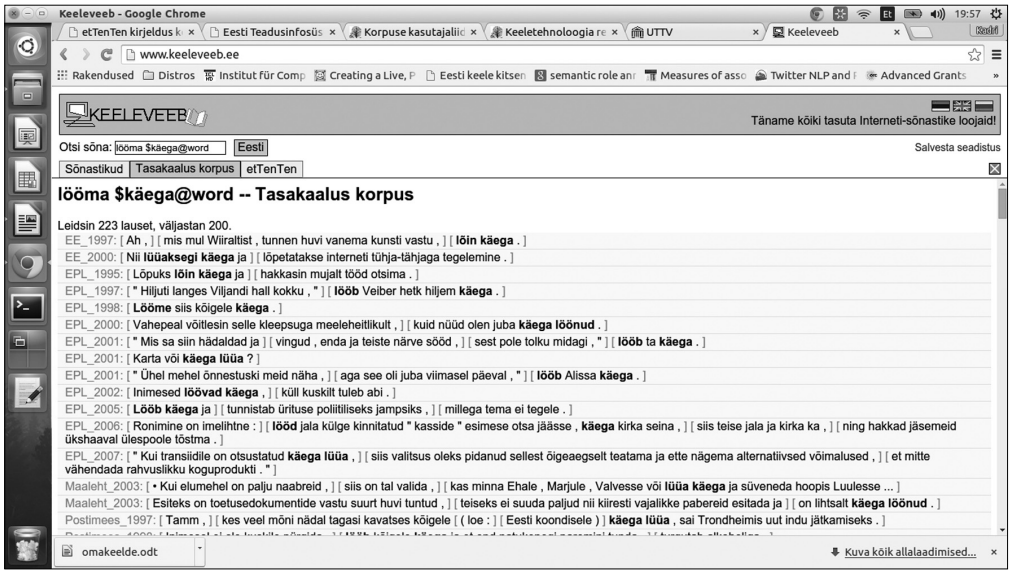
Rippmenüüdes kasutatavate lühendite seletuste juurde pääseb, klõpsates lingile „Päringuinfo".

Joonisel 1 on näidatud päring, mis otsib Tasakaalus korpusest ja etTenTenist lemma (st algvormi) *lööma* ja sõna (st sõnavormi, tekstisõna) *käega* koosesinemist ühes osalauses, st otsib väljendverbi *käega lööma*.



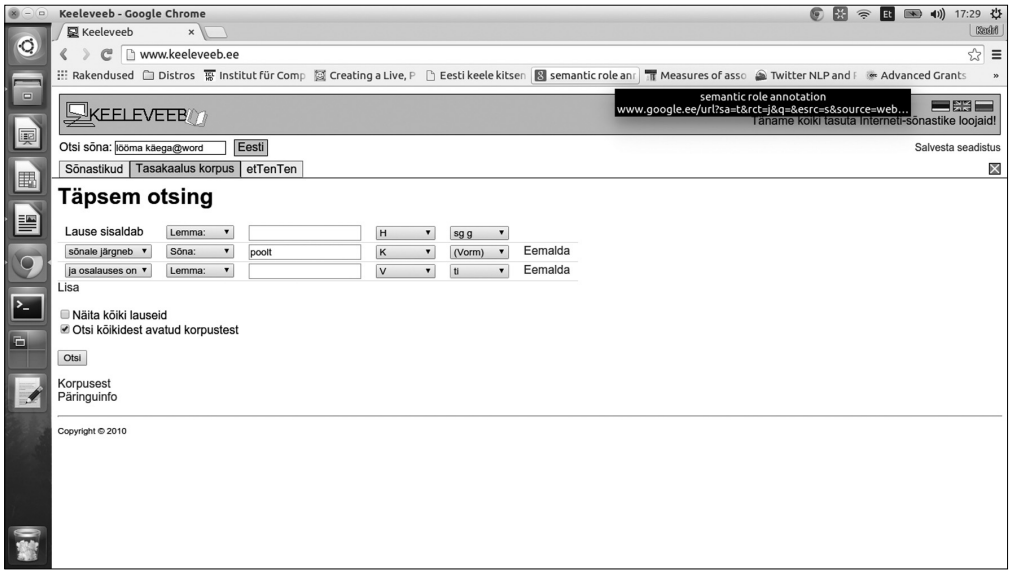
Joonis 1. Päring, mis otsib tasakaalus korpusest ja etTenTenist lemma

Joonisel 2 on näha selle päringu vastus. Klõpsates vastuseks väljastatud lausete alguses oleval lühendil (nt EE_1997), see n-ö avaneb ja näitab lause päritoluviite. Klõpsates ükskõik millisel tekstisõnal, avaneb see ja näha on morfoloogiline, tasakaalus korpuse puhul ka süntaktiline ja semantiline märgendus.



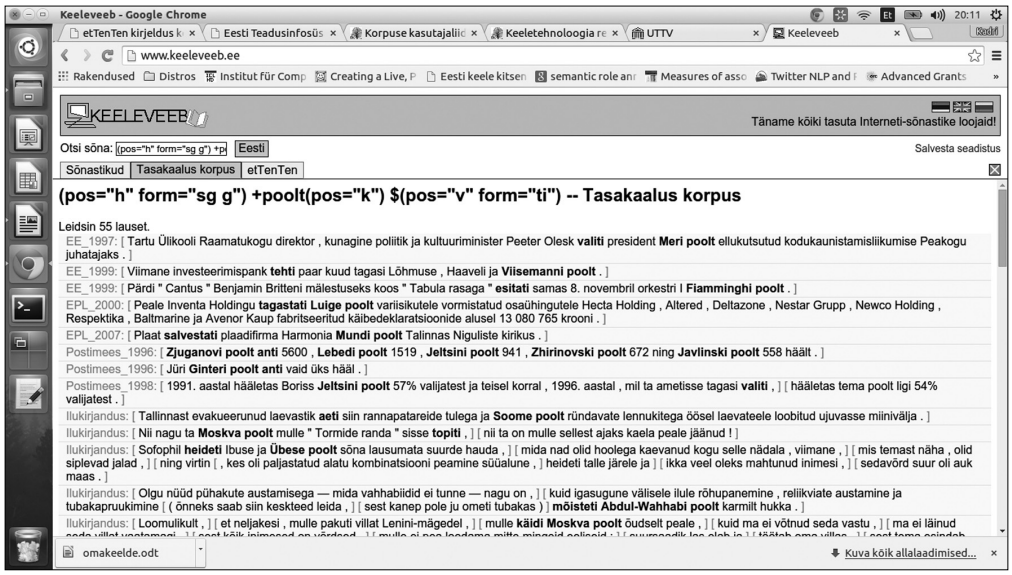
Joonis 2. Tasakaalus korpuse päringu vastus

Päringuid ei pea esitama ainult kindla lemma või sõnavormi kohta, kasutada saab ka ainult grammatilisi kategooriaid. Joonisel 3 on näidatud päring, mis otsib tasakaalus korpusest ja etTenTenist lauseid, kus esineb pärisnimi (sõnaliigi lühend H) ainsuse omastavas e genitiivis (sg g), millele järgneb sõna poolt kaassõnana (K), ning samas osalauses peab leiduma ka ükskõik milline tegusõna ehk verb (V) mineviku impersonaalis ehk umbisikulises tegumoes (lühend *ti*), st otsitakse impersonaalilauseid, milles tegija on väljendatud *poolt*-konstruktsiooni abil.



Joonis 3. Päring, mis otsib tasakaalus korpusest ja etTenTenist lauseid, kus esineb pärisnimi

Selle päringu vastus on näha joonisel 4. Nagu näidetel näha, ei ole päringu koostamisel vaja täita kõiki lahtreid ega teha rippmenüüdes kõiki võimalikke valikuid.



Joonise 4. Päringu vastus

Kõige mitmekesisemat infot suudab eesti keele korpustest hetkel välja sõeluda SketchEngine'i nimeline tarkvara, millest saab lugeda näiteks Jelena Kallase, Maria Tuuliku ja Madis Jürviste artiklist (2012)[16].

Korpuste baasil loodud sagedusloendid

Kasutajaliides annab päringule vastuseks otsitavat keelendit sisaldava lause ja tavaliselt ütleb ka ära nende lausete arvu. Selle kaudu ei saa aga vastust küsimusele, mis sõna (sõnaliik, kääne jne) on selles korpuses kõige sagedasem. Eesti keele tasakaalus korpuse baasil on koostatud mitmesuguseid sagedusloendeid[17], kust saab teada, mis on selle korpuse ja tema allkorpuste sagedasimad sõnad, sõnavormid, sõnaliigid või käänded ja kas näiteks ilukirjanduse sagedasim kääne on seda ka teaduse keeles. Need loendid on ka heaks võrdlus- või taustamaterjaliks – kui keeleuurija on koostanud näiteks eesti muinasjuttude korpuse ja teinud selle baasil sagedusloendi, saab ta oma loendit ja eelnimetatud loendeid võrreldes

16 Kallas, Jelena, Maria Tuulik, Madis Jürviste 2012. Leksikograafilise tarkvara Sketch Engine eesti keele moodul. – Eesti ja soome-ugri keeleteaduse ajakiri 3–2, lk 57–77.

17 http://www.cl.ut.ee/ressursid/

teada, kuidas erinevad muinasjuttude ja tänapäeva kirjalik üldkeel oma sõnavara ja selle sagedusomaduste poolest.

Kokkuvõtteks

Keeleteadus on viimase paarikümne aastaga muutunud silmatorkavamalt empiiriliseks teaduseks, järeldusi keele kohta tehakse keeleandmete põhjal ja keeleandmete (üheks) allikaks on keelekorpused. Kahtlemata on tekstikorpusel põhinev keelekirjeldus adekvaatsem kui ainult keeleuurija intuitsioonil põhinev.

Eesti keeleteadlased ja arvutilingvistid on ära teinud suure töö, koostades ja tehes kasutajaliideste kaudu kättesaadavaks mahukad keelekorpused, nende kallal töötades jätkub avastamisrõõmu paljudele ja pikaks ajaks. Omal kohal on ka paar hoiatavat sõna – keelekorpusi kasutades peab silmas pidama, et nende põhjal saab teha üldistusi ainult nende keelevaldkondade kohta, mida need korpused esindavad. Siinses ülevaates räägiti ainult kirjaliku keele korpustest ja suured korpused ongi kõik kirjaliku keele korpused. Kuigi keeleuurijad on ühel meelel, et kõige ehedamalt avaldub keel spontaanses suulises vestluses, on seda sisaldavate korpuste koostamine aeganõudev ja kallis töö, nii et suulise kõne korpused on tunduvalt väiksemad kirjaliku keele omadest.