Keeletehnoloog teoretiseerib korpuse üle

Heiki-Jaan Kaalep (Tartu Ülikool, Eesti)

Ettekanne käsitleb teoreetilisi ja ehk isegi filosoofilisi küsimusi, millele eesti keeletehnoloogia praktikuna olen pidanud oma töös vastama.

Praktikuna oli minu algne eesmärk luua programm (praeguseks on sellest kujunenud vabavaraline Vabamorf), mis oskaks suvalise sõnavormi puhul öelda, milline on selle algvorm, kääne, pööre jm grammatilised tunnused, ja muidugi kas see on üldse mingi eesti keeles oleva sõna vorm. Ideaaljuhul oleks tegu programmiga, mis käitub nagu keeletunnetusega inimene, või vähemalt käitub nii küllalt sageli ja teeb etteaimatavaid vigu, mida päris inimesel on kerge ette näha ja parandada.

Rakendusena on tegu õigekirja kontrollija e. spelleriga ja morfoloogilise analüsaatoriga; viimast kasutatakse m.h selleks, et tekstis olev sõnavorm algvormiks teisendada ja seeläbi teksti sõnavara kergemini käsitleda.

Programmi loomise etapid olid: 1) lihtsõnade analüüs, tuginedes Ü. Viksi „Väikesele vormisõnastikule“, 2) liitsõnade ja tuletiste analüüs, tuginedes mitmete Eesti lingvistide teoreetilistele töödele, ja 3) sõnastikust puuduvate tüvedega sõnade analüüs (e. oletamine), mille jaoks sai tugineda lingvistide ideedele ja tekstikorpustele. Etapi kaupa muutus töö programmeerimisest järjest rohkem lingvistiliseks uurimistööks.

Kas Vabamorf on kõige tõenäolisem kandidaat „Eesti keele morfoloogia teooria“ nimetusele? Programmil on olemas teadmised selle kohta, millised on morfeemid, tuletusliited, liitsõnade moodustamise mehhanism jpm lingvistilisi teadmisi. Tõsi, nad on esitatud programmeerimiskeeles, mitte lingvistile mõeldud proosatekstina, kuid see on ju ainult esitusviisi valik. Tal on ka üks oluline omadus: ta oskab öelda, kuidas mistahes sõnast vorme moodustatakse, s.t. ta on ennustusvõimeline. Milline keelekirjeldus oleks veel sama põhjalik?

Kahjuks tuleb tunnistada, et kui mõni lingvist tõesti loeks Vabamorfi lähtekoodi, siis oleks ta vist vapustatud, et kas tõesti selline ongi „Eesti keele morfoloogia teooria“ tase... Nimelt on seal palju loendeid ja reegleid, mille ainus õigustus on, et „nii saab palju ettetulevaid sõnu analüüsitud“, s.t programmi rumalus saab ära varjatud, kuid sügavam põhjendus puudub. Programm kannab tema arendamisel kasutatud tekstikorpuse pitserit. See viib meid küsimuse juurde: kas keele kui süsteemi seaduspärasused on üldse keelekorpusest kui kasutusnäidete hulgast tuletatavad? Või peame piirduma tõdemusega, et korpuse põhjal tehtud järeldus kehtibki kindlalt ainult sama korpuse raames?

Tekstikorpus on elektroonilisel kujul olevate tekstide kogu, mis esindab mingit valdkonda ning seal kasutatud keelt (nt ametikirjad, lastehoid) ja mida uurides saab uusi teadmisi nii valdkonna enese kui seal kasutatud keele kohta. Heal juhul võimaldab tekstikorpuse uurimine

anda teadmisi, mis kehtivad laiemalt kui korpuse tekstidel; mis kehtivad võib-olla isegi terves keeles ehk kõigi selles keeles loodud juttude-tekstide puhul.

Viimane on võimalik juhul, kui korpus esindab üldkogumit – olles viimasest palju väiksem, on ta sellega mingis olulises mõttes sarnane e. üldkogumi suhtes esinduslik e. representatiivne.

Kumbki aspekt – mis on tekstide ja/või keele „üldkogum“ ja kuidas määratleda „sarnasus” – ei ole lihtsalt defineeritav. „Mittesarnase” korpuse peal tehtav statistika oleks aga eksitav.

Üks lahendus oleks vaadata, kas korpuse põhjal leitavad arvulised parameetrid on kooskõlas mujalt pärit andmetega, milleks võib olla nt. keele omandamisprotsess laste poolt, keele muutumine ajaloos, Zipfi seadus vms. Kui on kooskõlas, siis korpus on teatavas mõttes esinduslik.

Mõttekäigu aluseks on eeldus, et keele (all)süsteem – olgu selleks morfoloogia, sõnavara koosseis vms – kujuneb keelekasutuse alusel. Mitte igasugune korpus ei ole selline, et selle põhjal saaks kujuneda keele süsteem sellisena, nagu me seda tunneme.

Illustreerimiseks vaadeldakse korpuse esinduslikkust eesti verbimorfoloogia näitel.

Sõna kõik muutevormid ei ole inimesel tervikuna mälus, vaid osa neist moodustatakse sama sõna mõne teise vormi alusel. Alusvormiks sobib ainult selline, mis on juba varem teada, s.t. mida on varem kohatud. Tekstikorpuse sõnavarale rakendatuna tähendab see, et alusvormi leksikaalne kasutusulatus – vormi levimus sõnavara lõikes – ei saa olla väiksem kui moodustatava vormi oma.

Käändsõnade puhul on kõige suurema kasutusulatusega vormiks ainsuse nimetav (s.t. kõige rohkem erinevaid sõnu esineb ainsuse nimetava kujulisena), sellele järgnevad ainsuse omastav, osastav ja seejärel muud käänded. Mitmuse vormid on harvemad kui ainsuse omad.

Vormide kasutusulatuse hierarhia on kooskõlas ka Zipfi seadusega, mille kohaselt sõnavormi sagedus ja pikkus on pöördvõrdelises seoses, s.t. mida sagedasem vorm, seda lühem see on. Nimetav kääne on ilma lõputa, omastava käände puhul lisandub tüvevokaal, osastava puhul veel lõpp d/t jne.

Pöördsõna vormide kasutusulatuse hierarhia on aga eri korpustes erinev. Osutub, et verbiparadigmade puhul on esinduslikuks korpuseks ainult väikelaste hoidjakeele ja interneti jututubade korpus.

Näib, et mitmed korpused sisaldavad liiga palju lugusid, narratiive, võrreldes selle päris loomuliku keelekasutusega, mis morfoloogia süsteemi kujundab ja mille iseloomule keelesüsteem kohandub. Selles loomulikus keelekasutuses räägitakse pigem soovidest, käskudest ja keeldumistest. Need on iseloomulikud olukordadele, kus inimesed teevad midagi koos, on need siis ema ja laps või ehitusmehed koos maja ehitamas („anna haamer“, „võta ise“, „nii ei saa“, „ära siia astu“).

Keelesüsteemi toimimise teatud aspekte, nt. üleminekukiirust sõna vanalt muutmisviisilt uuele saab jälgida siiski ka mujal kui ainult ülalnimetatud korpustel.

Ettekujutus tekstikorpusest kui keelesüsteemi kujundavast jõust tõstab esile ka küsimuse tiražeerimisest kui olulisest tegurist korpuse koostamisel ja hindamisel. Joh. Aavik ei piirdunud oma uute sõnade puhul nende avaldamisega leksikonides, vaid kasutas neid just laialt leviva meelelahutuskirjanduse tõlgetes.