Mis kasu on keeleteadusel keeletehnoloogiast

25. september 2002 - 23:45 Autor: Ülle Viks

Arvutimaailm 2002, nr 8 lk 11-14

Olen mitmel korral kogenud, et inimeste arusaamine keeletehnoloogia (KT) suhetest muude valdkondadega võib olla väga mitmesugune. Näiteks KTalaste väitekirjade kaitsmistel on ikka ja jälle kerkinud küsimus, miks on need tööd just sellesse kaitsmisnõukokku saadetud, olgu see siis informaatika või keeleteaduse nõukogu. Sama lugu on nende ametkondadega, kes raha jagavad: ükski neist ei taha keeletehnoloogiat omaks pidada.

Mis on mis

Püüan siin esitada oma arusaamise keeleteaduse ja KT vahekorrast (mis ei pruugi täielikult kattuda teiste sama valdkonna esindajate arusaamistega). Need kaks mõistet on tihedalt seotud informaatika ja arvutilingvistikaga.

Mõisted lingvistika (keeleteadus) ja informaatika (arvutiteadus) on enamvähem üheselt mõistetavad.

- keeleteadus, lingvistika, teadus keelest, selle olemusest, ehitusest, talitlemisest ja arenemisest (EE)
- informaatika, arvutil põhineva infotöötlusega tegelev teaduse ja tehnika haru (ÕS 99, AKS)

Arvutilingvistikaga on lugu segasem. Arvutilingvistikat (nimetus ka arvutuslingvistika, raallingvistika) on liigitatud nii ühele kui teisele poole. Arvutiinimesed kalduvad teda pidama rohkem keeleteaduse osaks ja vastupidi.

Arvutilingvistika koha paneb paika ÕS 99, mis määratleb seda selgesõnaliselt piiriteadusena.

- arvutilingvistika, raallingvistika, loomuliku keele automaattöötlusega tegelev keeleteaduse ja informaatika piiriala [ÕS 99 see seisukoht saab autoriteetset toetust ka Internetist: Saarlandi ülikooli arvutilingvistika professor Hans Uszkoreit nimetab arvutilingvistikat piiriteaduseks, mis tegeleb inimkeele arvutusliku küljega. (Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. (Hans Uszkoreit www.coli.uni-sb.de/~hansu/what_is_cl.html))]

Keeletehnoloogia on nii uus asi, et vähemalt Eesti sõnastikest seda sõna veel ei leia.

Ka siin võib leida väga erinevaid definitsioone, mis osalt kattuvad arvutilingvistika määratlustega. Minule tundus jällegi kõige asjakohasem Hans Uszkoreiti definitsioon, mis kõlab vabas tõlkes umbes nii: KT tegeleb meetodite, tarkvara ja seadmetega, mis on määratud tekstide ja kõne töötlemiseks. KT on arvutilingvistika tehnoloogiline haru, mis tugineb teadmistele inimkeelest. (www.dfki.de/lt/lt-general.html)

Arvutilingvistika ja KT vahele võib mahtuda veel kolmaski mõiste, nagu näiteks Sheffieldi ülikooli professori Hamish Cunninghami (1999) jaotuses: arvutilingvistika (computational linguistics), keele masintöötlus (natural language processing) ja keeletehnoloogia (language engineering). Cunningham näeb valdkonda järgmiselt. Arvutilingvistika on lingvistika haru, mis kasutab arvutit vahendina lingvistiliste probleemide lahendamisel. Keele masintöötlus on arvutiteaduse haru, mis uurib arvutisüsteeme (algoritmid, formaalsed keeled, tarkvara) ja nende kasutamist loomulike keelte töötlemisel. Keeletehnoloogia (engineering vs science) on keele masintöötluse rakendus tarbijasüsteemides (masintõlge, kõnesüntees, inimkeelne suhtlus arvutiga, jne).

Vaatepunktist sõltumata on selge, et KTs kuuluvad keel ja arvuti lahutamatult kokku, ja valdkonda seestpoolt vaadates eristatakse vastavalt sellele ka tema koostisosi:

Tarkvara: arvutiprogrammid keeleandmete töötlemiseks, nt

- teksti grammatiline analüüs ja süntees
- suulise kõne süntees ja tuvastus
- õigekeelsuse ja stiili kontroll
- masintõlge
- infotsisüsteemid
- dokumenditöötlus
- inimkeelne dialog arvutiga
- tõlkija või keeleõppija abivahendid
- jne jne

Keeleressursid: formaalsed keeleandmed tarkvarasüsteemide arendamiseks:

- elektronilised sõnastikud ja andmebaasid
- lingvistiliselt märgendatud tekstikorpused
- formaalsed grammatikakirjeldused

Mõistelise maadejagamise kokkuvõtteks võiks olla lihtsustatud skeem mis millega tegeleb.

Vastastikused ootused

Mida ootab keeletehnoloogia lingvistikalt?

See mida KT vajab, on keeleressursid: sõnastikud, märgendatud tekstid ja formaalsed grammatikad ehk lingvistiline teadmine formaliseeritud kujul. Lingvistilise uurimistöö väljundiks on aga tavaliselt teaduslikud uurimused, akadeemilised grammatikad ja traditsioonilised sõnastikud. Need kõik sisaldavad lingvistilist teadmist, mis on sõltuvalt väljaande liigist esitatud erinevate põhimõtete järgi: uurimus, grammatika ja sõnastik võivad sisaldada sama infot, aga see on vormistatud erinevalt.

KTle pakuvad huvi eelkõige suured sõnastikud: kakskeelsed tõlkesõnastikud ja rikkaliku andmestikuga ükskeelsed sõnastikud, nt ÕS või seletussõnaraamat.

Traditsioonilises sõnastikus võib olla märksõna kohta palju mitmesuguseid andmeid, nt

hääldus, sõnaliik, muuttüüp, eriala või stiilimärgend, seletus, sünonüümidantonüümid, tüüpilised väljendid, lausenäited jne. See kõik on KTle vajalik info, kuid enamasti ei saa KT seda otse kasutada, sest see pole täpselt see, mida vaja, ja mitte sel kujul, kui vaja. Uurimused, käsiraamatud ja sõnastikud on ju tehtud inimese poolt inimese jaoks, ja seetõttu tulevad mängu taustateadmised, mida inimesele pole vaja seletada, küll aga arvutile.

Nii et enne kui KT saab lingvistilisi teadmisi kasutada, tuleb need teisendada KTle sobivateks keeleressurssideks: traditsioonilistest sõnastikest peavad saama elektroonilised leksikonid ja andmebaasid ning akadeemilistest keelekirjeldustest formaalsed grammatikad.

Mida ootab lingvistika keeletehnoloogialt?

See mida lingvistika vajab, on tarkvara, mille abil keelt uurida ja töödelda ehk lingvisti töövahendid arvutis. KT lühikese ajaloo jooksul on juba loodud küllalt palju tarkvara Eesti keele jaoks (osa sellest mitmeski variandis): morfoloogiline analüüs ja süntees, morfoloogiline ühestus, süntaktiline analüüs, kõnesüntees (tekst > kõne), speller, otsisüsteemid tööks sõnastikega ja tekstidega, jne, jne. Kõik need programmid on vajalikud ka keeleteadlasele, aga probleem on analoogiline sellega, millest oli juttu eelnevas jaotises. Tarkvara on loodud nn laia tarbija jaoks, kelle vajadused on teistsugused. Lingvist ei saa tarkvara otse kasutada, sest see pole täpselt see, mida vaja, ja mitte sel kujul, kui vaja.

Selleks, et keeleteadlasele saaksid KT vahendid kättesaadavaks, tuleb need enne kohandada vastavalt lingvistilise uurimis ja arendustöö vajadustele. Alati polegi vaja olemasolevat tarkvara ümber teha piisab sobivatest liidestest (vahel on muidugi tarvis ka midagi uut).

Põhiline nõue, mille täitmist lingvist eeldab, on et keele automaattöötlus annaks lingvistiliselt usaldusväärseid tulemusi et info töö käigus ei moonduks.

Lingvisti töövahendid

Lingvist vajab töövahendeid mitmel otstarbel: selleks et hankida uusi teadmisi keelest (mida varasemad töövahendid ei võimaldanud), et kontrollidatestida teooriaid, hüpoteese ja mudeleid eksperimentaalselt, et koostada uusi sõnastikke ja grammatikaid (inimeste jaoks), et valmistada ette formaliseeritud keeleressursse (KT jaoks).

Lingvist saab küllalt palju abitöid ära teha lihtsate ja universaalsete programmidega, nagu sortimine, statistika, konkordantside koostamine, mitmesugused teisendused jne. Aga tõsisemate ülesannete puhul läheb tarvis väga komplitseeritud töövahendeid, mis sisaldavad peaaegu kogu KT arsenali.

Autorisõnastiku koostamine

Üks näide leksikograafia vallast. Vaatame, mida on vaja selleks, et koostada autorisõnastikku. See on sõnastik, mis keskendub ühe autori sõnakasutusele, võttes arvesse kõik autori poolt loodud tekstid ja kõik nendes esinevad sõnavormid.

Esmapilgul näib asi lihtne: tuleb kõik tekstid sisestada arvutisse, eristada sõnavormid, sortida need alfabeeti, loendada ja ühendada korduvad sõnavormid. Paraku ei ole sel viisil saadud tulemus mitte tõeline sõnastik, vaid alles selle lähtematerjal (sõnede sagedusloendid).

Tekst koosneb sõnavormidest (nt Tuvid ei taha kirju kanda), aga sõnastiku märksõnadeks on algvormid (lemmad), nt tuvi, ei, tahtma, kiri, kandma.

Järelikult on vaja morfoloogilist analüüsi, et jõuda sõnavormi juurest lemmani. Analüüsi väljundis on vormikood, lemma ja tüüp_sõnaliik (nt tuvid on mitmuse nimetava vorm sõnast tuvi, mis kuulub tüüpi 17 ja on substantiiv).

```
Tuvid
        PlN tuvi 17_S
ei
        Neg ei --_V
        ID ei 41_D
taha
        ID taha 41_DK
        IndPrPs_ tahtma 34_V
        ImpPrSg2 tahtma 34_V
kirju
        PlP kiri 24_S
```

SgN kirju 01_A
SgG kirju 01_A
kanda
SgP kand 22_S
SgAdt kand 22_S
Inf kandma 34_V

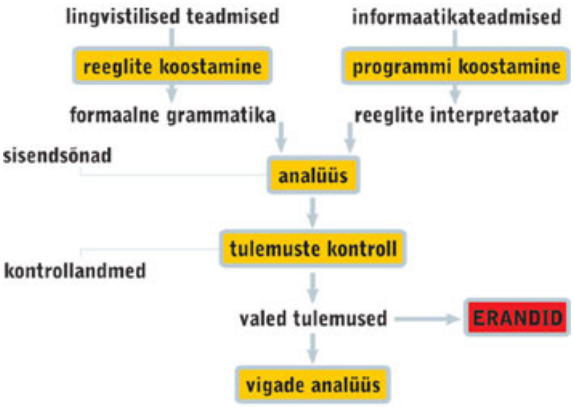
Morfoloogilise analüüsi tulemus on enamasti mitmene (Eesti keele puhul ligi pooltel juhtudel), nt taha võib olla verbivorm sõnast tahtma, või adverb või postpositsioon taha; kirju võib olla mitmuse osastav substantiivist kiri või ainsuse nimetav või omastav adjektiivist kirju; kanda võib olla dainfinitiiv verbist kandma või ainsuse osastav või lühike sisseütlev substantiivist kand.

Siin tuleb appi morfoloogiline ühestamine, mis konteksti arvestades valib igas konkreetses lauses välja õige analüüsivariandi (väljundinäites alla joonitud).

Korralikus autorisõnastikus peab olema antud ka sõna tähendus ja kasutusnäited. Tähendusi aab vaadata teistest sõnastikest, selleks on vaja sõnastikuotsingut. Aga tähendusi ei saa mujalt otse üle võtta, sest autorisõnastikus tuleb esitada ainult selle autori poolt kasutatud tähendused. Tähenduste kindlakstegemisel ja näidete valikul on leksikograafil vaja vaadata korraga kõiki sama sõna kasutusjuhtumeid. Siin aitab korpuseotsing ja selle baasil koostatud konkordants, mis paigutab järjest kõik sarnased sõnavormid ja näitab neid koos kahepoolse kontekstiga.

Kõike seda, mis juba nimetatud autorisõnastiku koostamisel, on vaja ka muude sõnastikutüüpide puhul ja kõigil sõnastiku koostamise etappidel: nii keelematerjali kogumisel, sõnaartiklite koostamisel kui sõnastiku toimetamisel. Kõik need on väga töömahukad protsessid, kus KT vahendid on suureks abiks.

Formaalse grammatika koostamine



Joonis 1: Tüübituvastuse reeglite koostamine

Teine näide on mu enda tööst automaatse reeglipõhise morfoloogiasüsteemi väljatöötamise käigus. Eesmärk oli luua formaalne grammatika muuttüübi tuvastuse mooduli jaoks. Töö käis umbes nii nagu joonisel 1.

Esialgu võtsin kokku oma teadmised muuttüüpidest ja analüüsisin kõigi tüüpide sõnu nende struktuuri aspektist. Tulemused panin kirja formaalsete reeglitena, mis arvestavad kaht tunnust:

1. silpide arv
2. sõna algvormi viimased häälikud (enamasti häälikuklassidena: V=vokaal, C=konsonant, Q=kpt)

Selles formaalses grammatikas on reeglid järjestatud nii, et konkreetsemad struktuurimallid on eespool ja üldisemad tagapool. Noole järel on tulemus: muuttüüp ja sõnaliik VVSi (VVS 1992) järgi.

Samal ajal ja koostöös minuga koostas tarkvara spetsialist (Evelin Kuusik) programmi reeglite töötlemiseks e reeglite interpretaatori (esialgse variandi).

Järgnes testimise periood sisendsõnade analüüs ja tulemuste kontroll. Oluline roll oli siin VVSil, mis täitis kaht ülesannet, olles nii sisendmaterjaliks (ilma andmeteta) kui ka kontrollmaterjaliks (koos tüübi ja sõnaliigi andmetega).

Reeglite interpretaator analüüsis läbi kõik sisendsõnad, määrates igaühele tüübinumbri ja sõnaliigi. Seejärel võrdles abiprogramm tulmusi kontrollandmetega ja väljastas vigased tulemused.

Järgnes jällegi minu töö vigade analüüs. Osa vigu võis olla tingitud programmi ebatäpsest tööst, osa aga sellest, et reeglid ei olnud küllalt head.

Pärast muudatusi mõlemas järgnes uus testimisring, mis kordus seni, kuni enam ei õnnestunud (programmi ja) reeglite parandamisega vigade hulka oluliselt vähendada.

Allesjäänud valedest tulemustest moodustati erandite loend, kus igal sõnal on kontrollsõnastikust võetud õiged andmed küljes:

aadlik 25_S
kaustik 02_S
nuustik 02_S
puuslik 25_S
päästik 02_S
ämblik 25_S

Muudes rakendustes kasutatav tüübituvastuse moodul töötab sellesama reeglite interpretaatoriga ja töö käigus tekkinud erandite loendiga. Kõik reeglipõhised moodulid töötavad ühesuguse skeemi kohaselt: kõigepealt vaadatakse, kas sisendsõna leidub erandite loendis. Kui jah, siis ongi õige tulemus käes; kui ei, siis saadakse tulemus reeglite abil.

Seda süsteemi, mida ma kasutasin tüübituvastuse reeglite ja erandite vahekorra klaarimiseks, võiks nimetada lingvisti töökeskkonnaks. Samal kombel on tehtud ka formaalsed grammatikad teiste morfoloogiamoodulite jaoks, nt silbitus ja tüvemuutused.

Kõik need moodulid kokku moodustavad reeglipõhise morfoloogiasüsteemi (Morfo), mis hõlmab nii sõnavormide analüüsi kui sünteesi ja laieneb ka sõnamoodustusele.

Reeglipõhise morfoloogia peamised tunnused:

1. Kõik, mis keeles on reeglipärane, esitatakse formaalsete reeglitena, ja ainult need üksused, mis reeglitele ei allu, esitatakse erandite sõnastikes. Erandite valik ja hulk sõltub otseselt fikseeritud reeglitest.
2. Tarkvara ja andmed on sõltumatud. Tarkvara põhiosa moodustavad reeglite interpretaatorid. Andmete põhiosa moodustavad formaalsed grammatikad (reeglid) ja nende juurde kuuluvad erandid. Kõik andmed (reeglid, erandid, juhtinfo) on antud tekstifailidena, mida saab vajaduse korral muuta.
3. Süsteem koosneb mitmest iseseisvast moodulist.

- - Igal moodulil on oma reeglite interpretaator ja oma reeglikomplekt koos vastavate eranditega.
  - Tarkvaramoodulid ja keele allsüsteemid on omavahel vastavuses.
  - Moodulid on realiseeritud iseseisvate dünaamiliste teekidena (dll dynamic link library), mis on kasutatavad ka ükshaaval muude rakenduste koosseisus.

Morfo kohta on lähemalt lugeda EKI uurimisteemade rubriigis (Viks 2000a): www.eki.ee/teemad/avatud_mrf.html

Tarkvara on antud vabasse kasutusse koos lähtekoodiga: www.eki.ee/tarkvara

Süsteemi muudetakse ja täiendatakse ka edaspidi, kui mõni moodul paremaks saab või kui andmed (reegliderandid) muutuvad. Praegu käib töö liitsõnade formaalse grammatika täiendamiseks ja liidese tegemiseks, mis võimaldaks ka teistel kasutajatel süsteemi täiendada ja oma erivajadustele kohandada.

Sõnastike generaatorid

Kolmas näide lingvisti töövahendite kohta on nn kirjegeneraator, kus rakendatakse sellesama morfoloogiasüsteemi mooduleid. See on süsteem grammatiliste andmete automaatseks lisamiseks tavalise sõnastiku sõnaartiklisse: sõna põhivormid, muuttüüp, sõnaliik, erandid jms. Kirjegeneraatori abil on seni tekitatud grammatilised andmed kolme sõnastikku: Eesti vene sõnaraamat (EVS 1997, 2000), Norra Eesti Eesti norra sõnaraamat (NEEN 1998), Eesti norra norra Eesti taskusõnaraamat (ENT 1999). Lähiajal on lisandumas Soome Eesti suursõnaraamat ja EVSi III köide.

Olgu näiteks grammatiline kirje märksõnale tuba 3 eri sõnastikus, mis esitavad grammatikat erineval määral ja moel.

EVS: tuba {tuba toa tuppa, tuba[de
tuba[sid & tub/e S 18}
NEEN: tuba [toa, adt. tuppa 18e]
ENT: tuba {toa, adt. tuppa}

Kirjegeneraatori kohta on lähemalt lugeda EKI uurimisteemade rubriigis (Viks 2000b): www.eki.ee/teemad/kirjegeneraator.html

Sõnastikugeneraator:

Eraldi tahaksin nimetada I. Heina loodud Internetipõhist masintõlke sõnastiku koostamise süsteemi koos tõlkiva brauseriga, mis on õigupoolest sõnastiku koostamise abivahend: Inglise Eesti sõnastik ja tõlkiv brauser (TÕBRAS): www.eki.ee/keeletehnoloogia/projektid/inglise-eesti

See süsteem ootab veel sidumist morfoloogiasüsteemiga, seejärel on võimalik hakata teda lähendama tegelikule masintõlkele.

Tulevikuunistus

Lõpuks tahaksin natuke unistada keeleteadlaste ja keeletehnoloogide koostööst.

Minu unistus on, et nii lingvistidel kui keeletehnoloogidel oleks kasutada:

- ühtne andmebaas (leksikaalne + grammatiline), mis oleks lingvistide kontrolli all ning täieneks ja uueneks pidevalt uurimistulemuste põhjal;
- ühtne töökeskkond, mis ühendaks keeleressursid ja tarkvara (lingvisti töövahendid).

Miks on vaja ühtset andmebaasi?

Igal KT rakenduse tegijal pole mõtet hakata endale uuesti ja otsast peale andmebaase tekitama. Teadmiste baas on rohkem alusuuringute küsimus ja vajab pidevat värskendamist, sest ühelt poolt muutub keel ise (tekivad uued sõnad ja isegi uued grammatikanähtused), teiselt poolt muutub teadmine keelest (saadakse uusi uurimistulemusi, kehtestatakse uusi norminguid).

Ja kui olemasolevad andmed on kokku koondatud, siis on ka lingvistidel kergem neid arvestada uute probleemide lahendamisel. See on koht, kus võib loota kvantiteedi üleminekut kvaliteediks.

Miks on vaja ühtset töökeskkonda?

Igal lingvistil pole mõtet hakata endale isiklikke töövahendeid tahtmategema. Paljud uurimises kasutatavad protseduurid on ühised ja sageli saab ära kasutada olemasolevaid mooduleid, neid omavahel kombineerides ja võibolla pisut täiendades.

Ja kui olemasolevad KTvahendid on koondatud ühtsesse keskkonda, saab neid paremini ära kasutada uute loomisel. Erinevaid OPsüsteeme arvestades peaks selliseid keskkondi olema ilmselt mitu (vähemalt esialgu).

Kokkuvõte

Kokkuvõtteks tahan väita, et keeletehnoloogia on üks keeleteaduse väga olulistest rakendusvaldkondadest, mille paljude erinevate tarbijate hulka kuuluvad ka keeleteadlased ise.

Keeleteadlaste töö tulemusi kasutatakse KT toodete loomiseks, kusjuures needsamad tooted on lingvistilises uurimis ja arendustöös väga vajalikud töövahendid. Lingvist vajab tegelikult kõiki neidsamu KT vahendeid, mida laiatarbekasutaja. Ainult et ta on ehk nõudlikum kasutaja. Ja tema tegevusest tõuseb omakorda tulu KTle, sest paremate vahenditega saab paremaid tulemusi, mis teevad lõpuks kvaliteetsemaks ka laiatarberakendused. Looja ja tarbija suhted keeletehnoloogia ja keeleteaduse vahel on vastastikused ja see võimendab mõlema efekti.

VIITED:

1. AKS = Vello Hanson, Arvi Tavast. Arvutikasutaja sõnastik. Tallinn 1999.
2. Cunningham 1999 = Hamish Cunningham. A definition and short history of Language Engineering. Natural Language Engineering 1999, 5: 116.
3. EE = Eesti Entsüklopeedia (Eesti Nõukogude Entsüklopeedia).
4. ENT 1999 = Turid Farbregd, Sigrid Kangur, Ülle Viks. Estisk lommeordbok. Eesti Norra Norra Eesti. Oslo 1999.
5. EVS 1997, 2000 = Eesti vene sõnaraamat I & II. Tallinn 1997 & 2000.
6. NEEN 1998 = Turid Farbregd, Hille Lepp, Ülle Viks. Norra Eesti Eesti Norra sõnaraamat. Tallinn 1998.
7. Viks 2000b = Ülle Viks. Kuidas tekib sõnastikukirjesse grammatika. Keel ja Kirjandus 2000, nr 7: 486495.
8. Viks 2000a = Ülle Viks. Eesti keele avatud morfoloogiamudel morfoloogiamudel Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu 2000, lk 936.
9. VVS 1992 = Ülle Viks. Väike vormisõnastik I: Sissejuhatus & grammatika, II: Sõnastik & lisad. Tallinn 1992.
10. ÕS99 = Eesti keele sõnaraamat ÕS1999. Tallinn 1999.

Lingid samal teemal:
www.coli.uni-sb.de/~hansu/what_is_cl.html
www.dfki.de/lt/lt-general.html
www.eki.ee/teemad/avatud_mrf.html
www.eki.ee/tarkvara
www.eki.ee/teemad/kirjegeneraator.html
www.eki.ee/keeletehnoloogia/projektid/inglise-eesti

- Lahendused
- Tarkvara