ERINEVATE ANDMETÖÖTLUS-MEETODITE RAKENDAMINE SEOSTE LEIDMISEKS VEALIIKIDE VAHEL

Anastassia Šmõreitšik

Ülevaade

Käesolevas artiklis tutvustatakse Tallinna ülikooli üld- ja rakenduslingvistika õppetooli *Eesti vahekeele korpuse*[1] töötlemiseks sobivaid meetodeid, et esile tuua vealiikide vahelisi seoseid ja täpsustada veapuu hierarhiat.

Võtmesõnad: korpuslingvistika, lingvistiline veaklassifikatsioon, mustrid

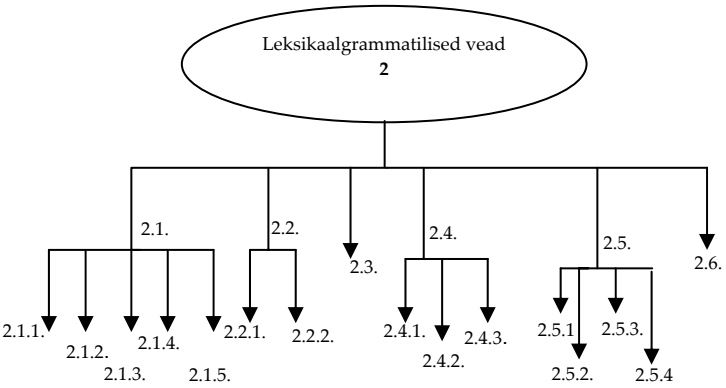
[1] http://evkk.tlu.ee, 28.09.2007. Uurimistööd on toetanud riikliku programmi "Eesti keel ja rahvuslik mälu" (2004–2008) projekt R 05/01 "Koodivahetuse, vahe- ja lastekeele korpuste töötlemine ja haldamine" ja ETFi grant nr 6151 "Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus" (2005–2008).

Eesti vahekeele korpuse lingvistilise veaklassifikatsiooni analüüsist

Mitmemõõteliseks võib nimetada objektide esitust, kui iga objektiga on seostatud kolm või enam tunnust (Tooding 2007: 195). Tavaliselt on igal objektil rida talle iseloomulikke tunnuseid, mille seostamise tulemusena saab määrata selle objekti koha teiste samalaadsete objektide hulgas. *Eesti vahekeele korpuse* (EVKK) objektiks on keelevead, mille defineerimiseks ja vigade klassifikatsiooni loomiseks on vaja valida niisugused tunnused, mis kirjeldaksid objekti (viga) erinevatest aspektidest ning erinevatel tasanditel.

EVKK veaklassifikatsioon on seotud keeletasandite hierarhiaga (grafeem, morfeem, sõna, sõnaühend, lause, tekst) ja keele uurimise aspektidega (semantika, grammatika, pragmaatika). Nii on saadud mitmemõõteline veaklassifikatsioon, mis sisaldab kaheksatteist veaklassi, igas omakorda erinev hulk paradigmaatiliselt ja süntagmaatiliselt seotud vealiike (vt Eslon 2006: 13). Niisugust vigade hierarhiat võib matemaatilises keeles kirjeldada vealiikide ja nende alamliikide vastastikuste sõltuvuste puuna, kus iga veaklass sisaldab n -1 vealiiki, mis on omavahel klassi piires seotud. Näitena toome siinkohal EVKK veapuu leksikaalgrammatilised vead (vt Joonis 1), milles number 2 tähistab veapuu teist liiki, selle all eristatakse kuut alamliiki (2.1, 2.2 … 2.6), mis omakorda jagunevad (nt 2.1.1 … 2.1.5). Kui teha korpuses päring "leksikaalgrammatilised vead", siis saab kasutaja näiteid nende alamliikide kohta, mis on antud hetkeks märgendatud[2].

[2] http://evkk.tlu.ee/Search/search_results, 28.09.2007



Joonis 1. Leksikaalgrammatilised vead

2. Leksikaalgrammatilised vead

2.1. tegevuse piiritletus / piiritlematus, nt *lülitan arvuti, vaatan elektronpost ja kontrollin oma allutatud. millal algavad tunde; kui ma õpin eesti keel siin, siis pärast ülikooli ma võin lihtselt öelda eestlasega*

2.1.1. osastavaline, omastavaline / nimetavaline objekt, nt *pane oma korvi lauale ja tule mine juurde*; *oleks paremini kui me õppesime mitte reeglid – kuidas rääkida ja kirjutada õigest*

2.1.2. ainult osastavalist võimaldavad verbid, nt *õnnitluste ja kuulutuste kujundamiseks on kasutatud erinevad ornamendid ja piltid*; *arvan, et oleks väga kasulik rohkem kuulda eesti kõne*

2.1.3. ainult täissihitist võimaldavad verbid, nt *ta neelas Punamütsikest alla*

2.1.4. kontekstivahendid, mis määravad tegevuse piiritletuse / piiritlematuse, nt *ja ainukeseks võimaluseks neid vigu vältida on õppida neid verbe pähe*; *paljud andsid õiget vastust*

2.1.5. kvantiteedisõnade kasutamine tegevuse piiritlemise vahendina, nt *kõik õppimine käib eesti keeles*

2.2. tegevuse transitiivsus / intransitiivsus, nt *kui inimene sööb see toit, siis ta haigestabt*

2.2.1. transitiivse verbi kasutamine intransitiivsena ja / või objekti ärajätmine, nt *inimene ei ole edukas, sest ta ei usu*

2.2.2. intransitiivse verbi kasutamine transitiivsena ja / või objekti lisamine – korpuse märgendamise antud etapil pole need vead välja toodud, kuid see ei anna põhjust arvata, et nad eesti vahekeeles puuduvad

2.3. tegevuslaadi väljendavad sufiksid (momentaanne, interatiivne, semelfaktiivne), nt *Mina olin üllatanud inimese arvutu hulka*

2.4. analüütilised verbid, nt *mul ei olnud inglise keele eksamit, siis ma ei astunud;*

2.4.1. ühendverbid, nt *ma ei või kujutada meie elu energia ilma*

2.4.2. väljendverbid, nt *ma saan raha rohkem, kui mul ja minu perele on vaja, et elu püsida*

2.4.3. kaksikverbid, nt *pärast ta tuleb töötab; kallis, võtma ja viima korv koogiga ja mahlaga vanaemale*

2.5. afiksaaladverbide tähendus ja kasutamine, nt *kui ma ümber vaatan, siis näen, et kui palju inimesi ei oleks*

2.5.1. perfektiivsust tähistavate afiksaaladverbide kasutamine, nt *tõmbas oma kirvese ära ja tapis hundi*

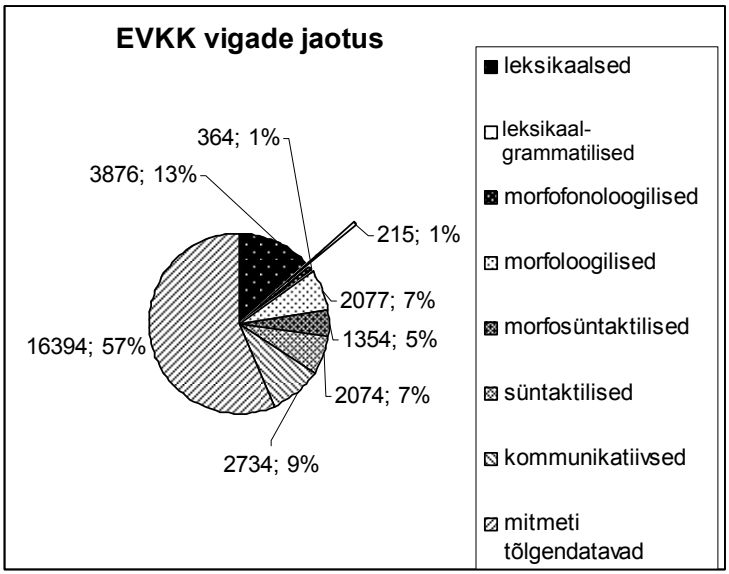
2.5.2. orientatsiooniliste afiksaaladverbide kasutamine – korpuse märgendamise antud etapil pole need vead välja toodud, kuid see ei anna põhjust arvata, et nad eesti vahekeeles puuduvad

2.5.3. seisundit väljendavate afiksaaladverbide kasutamine – korpuse märgendamise antud etapil pole need vead välja toodud, kuid see ei anna põhjust arvata, et nad eesti vahekeeles puuduvad

2.5.4. modaalafiksaaladverbide *vaja* ja *tarvis* kasutamine, nt *kui on olemas alaosakonnad, siis on vaja jagada kuulutused nende vahel*

2.6. noomeni tuletusliite tähendus, nt *talud muutusid eluvõimalitteks; Eesti on põllumajandus riik*

Vigade hierarhia alusel saab demonstreerida, missugused vea-klasside ja -liikide vahelised seosed on põhimõtteliselt võimalikud ning kas need seosed on tugevad või nõrgad. Küsimus on selles, et vigadele omast interpretatiivsust ei tohi tõlgendada määramatusena – teatud liiki vigade tekkimise vahel on kindlad põhjuslikud seosed, mida me lihtsalt ei pruugi teada. Näiteks korpuse 28758 vea hulgast tõlgendasid märgendajad mitmeti 57% juhtudest (vt Sektordiagramm 1).



Sektordiagramm 1. EVKK vigade jaotus

Sektordiagramm 1 annab ülevaate korpuses märgendatud vigade jaotumisest liikide kaupa. Kõige arvukamalt on esindatud leksikaalsed vead (13%), millele järgnevad kommunikatiivsed (9%), morfoloogilised (7%), süntaktilised (7%), morfosüntaktilised (5%), morfofonoloogilised (1%) ning leksikaalgrammati-

lised (1%). Vead, mida märgendajad polnud osanud panna ühegi liigi alla, moodustasid vigade koguarvust alla 1%, polnud seega statistiliselt olulised ning võisid oma sisu poolest kvalifitseeruda ka mitmeti tõlgendatavate vigadena, mida märgendatud vigade hulgas oli 57%. Mitmetitõlgendatavatele vigadele olid märgendajad külge pannud kaks või enam märgendit, mis tähendab ühe vea paigutamist erinevate vealiikide alla, nt *Me **andsime** koos dukumendid ühele ametile ajalooõpetaja*[3].

Verbi *andma* kasutamist on märgendaja antud kontekstis seostanud kahe vealiigiga:

1. leksikaalne viga → põhitähenduse viga → samasse mõistepesasse ja samasse sõnaliiki kuuluva sõna kasutamine (vrdl *andma* ja *sisse andma*),
2. leksikaalgrammatiline viga → tegevuse transitiivsus / intransitiivsus.

Korpuse arendamise praegusel etapil ongi eesmärk leida vigade, eriti mitmetitõlgendatavate vigade omavahelisi seoseid välja toov analüüsimudel. Sel eesmärgil on tähelepanu keskmes EVKK veaklassifikatsiooni võrdlemine empiiriliste ümberkodeeritud andmetega, mis peaks aitama leida mustreid (*patterns*), mida seejärel rakendada kui meetodit vea tekkepõhjuste selgitamisel ning vealiigituse täiustamisel. Artikli autor püüab võrdluses leksikaalgrammatiliste vigade statistilise analüüsi tulemustega näidata, kuidas see meetod põhimõtteliselt võiks töötada ning missuguseid tulemusi anda.

[3] http://evkk.tlu.ee/Documents/doc_542009824784_item, 28.09.2007

Leksikaalgrammatiliste vigade vaheliste seoste analüüs statistiliste meetodite abil

Õppijakeele vigade interpretatiivne iseloom ja selle arvutamine loovad raamistiku, millest lähtudes saab otsida veaklasside ja -liikide vahelisi seoseid ning neid vastavalt analüüsida. Selles loogilises ahelas võib vigade hierarhias aste-astmelt ülesalla liikuda, mis võimaldab omakorda veaklasse täpsustada ja näidata, missugused vead on omavahel otseses, missugused kaudses seoses. See on funktsionaalne lähenemine vahekeele uurimisele, mille alusel saab välja tuua analoogia.

Niisugune lähenemiviis on mõningas mõttes paralleelne eesti keele morfo- ja süntaksianalüsaatori väljatöötamisega formaalsete kitsenduste grammatikas (vt Kaalep & Vaino 2000: 90; Müürsepp 2000: 72). Sünteesi puhul läbib sisendsõna (lemma) kõigepealt tuvastusmooduli, mille reeglid annavad väljundisse muuttüübi ja sõnaliigi. Järgneb tüvemuutuste moodul, kus genereeritakse tüvejuhti järgides kõik vastavas tüübis ette nähtud tüvevariandid. Süntees lõpeb vormimoodustuse mooduliga, kus tüübikirjelduse reeglite järgi sobitatakse kokku iga nõutava muutevormi jaoks vajalik tüvevariant ja formatiivvariant. Ühestamise etapil eemaldab ühestaja konteksti mittesobivad tõlgendused järk-järgult, rakendades seejuures reegleid (Puolakainen 2001: 59).

Selline inimese kehtestatud reeglite ja objekti ühendamine statistilises modelleerimises loob eelduse korpuse andmete (pool)automaatseks analüüsiks: statistikal põhinevad analüüsimeetodid üritavad välja selgitada objektide vahel tõenäosuslike seoste võrgustikku. Inimeste loodud reeglitel töötavad süsteemid vajavad testimist korpuse tekstidel, et leida seoseid uute reeglite loomiseks (Muischnek 1998: 10). Seda laadi empiiriliste ja teoreetililiste andmete töötlemise ühtsus ning vastastikune sõl-

tuvus on eelduseks ka EVKK veaklassifikatsiooni hierarhiliste seoste täpsustamisel ja õppijakeele mitmekülgsel analüüsil.

EVKK veaklassifikatsiooni puhul on mitme vea ühismuutumise uurimine kindlasti viljakam kui süsteemi üksikute omaduste analüüs, samas aga ei tohi alahinnata seoste olemasolu vealiigi alamliikide vahel. Seetõttu tuleks analüüsimudeli katsetamisel arvestada seostega tunnusepaaris. Statistilise analüüsi seisukohalt räägitakse tihti korrelatsioonist tunnuste vahel, ükskõik missugust tüüpi seosest jutt ka ei käiks (Tooding 1999: 133).

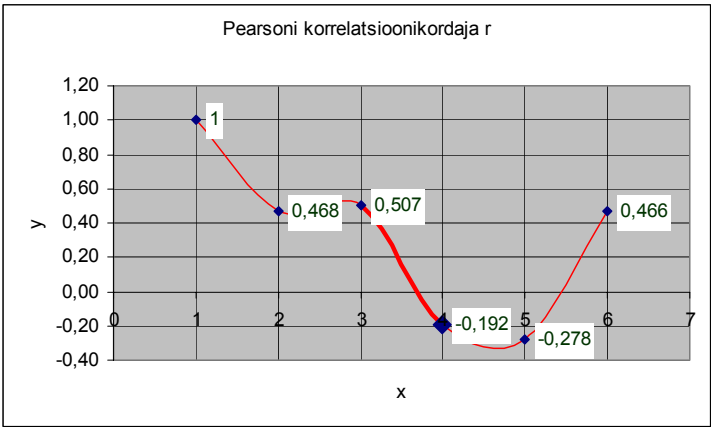
EVKK veaklassifikatsiooni tunnused on oma olemuselt mitteparameetrilised, nominaalsed ehk nimitunnused. Statistilise analüüsimeetodi üks levinumaid tunnuste ühisjaotust ja jaotuste omavahelist seost väljatoov vahend on ristabel. Kuid juba praegu, kui EVKKs on märgendatud ligi 219964 sõnet ja leitud ligi 28758 viga, osutub ristabel korpuse materjali puhul lahtrite suure arvu tõttu vähe ülevaatlikuks. Seega oleks tunnustevaheliste seoste tuvastamiseks mõttekam kasutada mitteparameetrilist testi, nt Pearsoni χ^2-testi (Tooding 1999: 133), mis annab võimaluse mõõta tunnustevahelist korrelatsiooni kindlapiirilise kordaja abil. χ^2-test on olemuselt olulisustest, mis annab vastuse küsimusele, kas erinevus tunnuste sagedusjaotustes on statistiliselt oluline või mitte. Seose olulisust ja tugevust aitab kokkuvõtlikult väljendada standardiseeritud seosekordaja ehk korrelatsioonikordaja r (Tooding 1999: 137). Kõik need kolm aspekti – seose olulisus, tugevus ja statistiline usaldusväärsus – on EVKK vealiigituse analüüsi puhul võrdväärselt tähtsad. Näitena olgu siinkohal toodud leksikaalgrammatiliste vigade alamliigis 2.1 (tegevuse piiritletus / piiritlematus) sisalduvate seoste tugevus (vt Tabel 1).

Tabel 1. Leksikaalgrammatiliste vigade seosetugevuse hindamine

	2.1.	2.1.1.	2.1.2.	2.1.3.	2.1.4.	2.1.5.
2.1. Pearson Correlation	1,000	,468(**)	,507(**)	-,192(**)	-,278(**)	,466(**)
Sig. (2-tailed)	,	,000	,000	,000	,000	,000
2.1.1. Pearson Correlation	,468(**)	1,000	,401(**)	-,072	-,203(**)	,372(**)
Sig. (2-tailed)	,000	,	,000	,061	,000	,000
2.1.2. Pearson Correlation	,507(**)	,401(**)	1,000	-,132(**)	-,250(**)	,600(**)
Sig. (2-tailed)	,000	,000	,	,001	,000	,000
2.1.3. Pearson Correlation	-,192(**)	-,072	-,132(**)	1,000	,538(**)	-,095(*)
Sig. (2-tailed)	,000	,061	,001	,	,000	,014
2.1.4. Pearson Correlation	-,278(**)	-,203(**)	-,250(**)	,538(**)	1,000	-,257(**)
Sig. (2-tailed)	,000	,000	,000	,000	,	,000
2.1.5. Pearson Correlation	,466(**)	,372(**)	,600(**)	-,095(*)	-,257(**)	1,000
Sig. (2-tailed)	,000	,000	,000	,014	,000	,

** olulisustõenäosuse näitaja usaldusnivool 0.01 (antud juhul kahepoolne)
* olulisustõenäosuse näitaja usaldusnivool 0.05 (antud juhul kahepoolne)

Tabelist 1 annab välja lugeda, et leksikaalgrammatiliste vigade 2.1 (tegevuse piiritletus / piiritlematus) alajaotuste vahel esinevad teatud seosed. Eriti kujukalt on need nähtavad korrelatsiooniseoste hajuvusdiagrammil 1, kus x-teljel on näidatud mõõtevahemik, milles lineaarset seost mõõdetakse ja y-telg näitab tunnuseid, mille vahel seost otsitakse.



Hajuvusdiagramm 1. Pearsoni korrelatsioonikordaja r

Korrelatsioonikordaja *r* kõneleb korrelatsiooniseose tugevusest (nõrk või tugev vastastikune sõltuvus). Kui seos on piisavalt tugev, siis öeldakse, et tunnused on omavahel korrelatiivsed. Korrelatsioonikordaja vähim võimalik väärtus on -1 ja suurim võimalik väärtus on 1. Väärtused -1 või 1 saavutatakse siis, kui tunnused on teineteisega täpselt lineaarselt seotud. Antud juhul mõõtis Pearsoni korrelatsioonikordaja *r* lähedusastet tegevuse piiritletuse / piiritlematuse vigade hulgas ning sai tulemuse, mida Tooding on nimetanud alaliikide täielikuks lineaarseks sõltuvuseks (Tooding 1999: 236). Hajuvusdiagrammis 1 on nii hälvete kui hälvete korrutiste seas olemas nii samamärgilisi kui erimärgilisi tulemusi. Sel juhul tuleb korrelatsioonikordajat määrav hälvete korrutiste summa nullilähedane (vt Tabel 1) ja korrelatsioonikordaja *r* näitab korrelatsiooni puudumist. Hajuvusdiagrammis kajastub kahe tunnuse vaheline seos ruutsõltuvuse kaudu, sest see seos pole lineaarne, vaid kujutatud paraboolina. Niisiis on selge, et kaks tunnust on omavahel mingil määral seotud, kuid seose tugevuse väljaselgitamiseks oleks vaja uurida tekkinud alajaotuste ruutsõltuvust.

Korrelatsiooniseose tugevuse kõrval on oluline ka korrelatsiooniseose statistiline usaldatavus. Vaadates valimi (215 leksikaalgrammatilist viga) ja üldkogumi (28758 viga) vahekorda, saab hinnata, kuivõrd tõepäraselt võiks valimi alusel leitud korrelatsioonisõltuvust kinnitada ka üldkogumi põhjal. Selleks võetakse appi *t*-test, mis antud valimi puhul näitab, et usaldusnivool 95% ei ole veel võimalust kinnitada usaldusväärse korrelatsiooniseose puudumist selle vealiigi alajaotuste vahel. Tuleb silmas pidada ka asjaolu, et seos kahe juhusliku suuruse vahel võib olla kahte tüüpi – funktsionaalne ja statistiline (korrelatiivne). Funktsionaalse seose korral vastab argumendi *X* mingile väärtusele üks ja ainult üks funktsiooni *y* väärtus. Statistilise (korrelatiivse) seose puhul võib ühe suuruse *X* mingile väärtusele vastata mitu teise suuruse *Y* väärtust, mida ei saa täpselt kindlaks määrata. Statistiline seos väljendub ühe juhusliku suuruse *Y* keskväärtuse sõltuvuses teise juhusliku suuruse *X* väärtustest. Seega korrelatiivse seose olemasolu ei tõenda veel, et suurused *X* ja *Y* on omavahel põhjuslikult seotud; kui seos statistiliselt puudub, siis tuleb hakata kahtlema klassifikatsiooni alustes.

Põhjusliku seose korral on üks nähtus põhjus ja teine tagajärg. Põhjus avaldab mõju tagajärjele ning seos nende kahe tunnuse vahel on alati kindla suunaga (Sivia 2006: 189). Huvitav oleks näiteks välja selgitada vigade X liigid ning nende seosed vigadega Y ja Z. Näiteks, kas viga Y saab olla põhjuslikus seoses veaga X? Missugustel juhtudel on Y põhjus ja X tagajärg? Mis tunnuste alusel on omavahel seotud vead X, Y ja Z? Keerukuse ja raskestimõõdetavuse tõttu ei anna õppijakeele vead lootust, et vigade erinevate aspektide vahel täheldavate seoste leidmine avaks tee süsteemi reguleerimisele. Uurimistulemuseks saab pidada seda, kui õnnestub välja selgitada, missugused seosed objektide vahel üldse on. Vigadevahelisi seoseid kui korrelatsioone nominaalsete tunnuste vahel tuleb käsitleda arvu-

liselt kirjeldatud suundumustena, mille väärtuse avab sisuline lingvistiline tõlgendus.

Mitmetitõlgendatavad vead – eeldus mustrite leidmiseks ja kasutamiseks

Kuna EVKK on mahult piisavalt suur avatud andmekogu (hetkel üle 600 000 sõne), siis annab see võimaluse täiendada või hoopiski kummutada märgendamise käigus tekkinud tunnetuslikke arusaamu veaklassifikatsiooni seoste paikapidamisest – nii või teisiti tõlgendatakse 57% ± 5% vigadest ikkagi mitmeti. Korpuse võimaluste kasutamine grammatiliste hüpoteeside tõestamiseks on juba edukalt rakendust leidnud: näiteks B. T. S. Atkinsi ja B. Levini (1995) leksikograafiaalased tööd ning L. Taylori, C. Groveri ja T. Briscoe (1989) ning G. Sampsoni (1987) morfoloogia- ja süntaksikäsitlus.

Kuna EVKK veaklassifikatsiooni loomisel on veale lähenetud lingvistiliselt, siis saab vea märgendamisel oluliseks sõnede kontekstisõltuvus, mis tihtilugu tingibki ühe ja sama näite paigutumise erinevatesse veaklassidesse. See asjaolu pärsib suuresti EVKK vigade analüüsimudeli väljatöötamisest. Samas on korpuslingvistikas toetust leidnud keeleüksuste erinevate tähenduseindikaatorite objektiivne lähenemine ehk hajusate kategooriate hüpotees. D. Mindt (1991: 188) on näidanud, et sageli kirjeldatud keeleüksuste tähendused on kindlapiirilised, mis teeb nad kasutuskontekstist sõltumatuks. See omadus võimaldabki hajutada keeleüksusi erinevate kategooriate vahel, mis EVKK veaklassifikatsiooni suhtes rakendatuna loob eelduse kasutada vigadevaheliste seoste leidmiseks mustreid (*patterns*). G. B. Davisi (2000) kohaselt on mustrid edukalt kasutatud teadmiste fragmendid, mis jagavad info üheselt mõistetavateks ja hallatavateks osadeks. Analüüsimustrid (*analysis patterns*) kir-

jeldavad mõõtmisega seotud objekte ja nedevahelisi seoseid. Mustreid kasutades võib luua ka struktuurse kirjutamise viisi, mille abil saab kirjeldada probleemi ja selle lahendust ning mis annab hulgaliselt täiendavat informatsiooni. Mustrid võivad moodustada mustrite keele (*patterns language*), mis on konkreetse valdkonna probleeme kirjeldavate mustrite kogum, milles probleemid ja nende lahendused võivad olla üksteisega seotud (Meszaros & Doble 2002)[4].

Mustrid tekivad teadmiste ja oskuste praktilise kasutamise käigus. Seetõttu võib igaüks, kes on leidnud lahendusi mingile probleemile, proovida oma teadmisi dokumenteerida mustrina. Seejärel peab muster läbima mitu tsüklit: ülevaatuse, tagasiside ja täiendamise. Kõigepealt võib mustri avaldada uurimisgrupisiseselt ja pärast retsenseerimist importida ka väljapoole. Juhul, kui autor saab tagasiside selle kohta, et samasugune muster on juba loodud, võib ta oma ideed teiste ideedega võrrelda ning mustrit täiendada (*The Software Patterns Criteria*)[5]. Mustrite haldamiseks on loodud vastav andmebaas *Portland Pattern Repository*[6], kuhu saab informatsiooni pidevalt lisada, seda muuta ja kustutada. Kataloogis olev informatsioon on omavahel hüperlinkidega seotud. Tekstilisel kujul esitatud mustrite otsimiseks saab kasutada otsingumoototit *WizDoc* (vt *WizSoft*)[7], kus kasutaja kirjeldab valdkonda, mille kohta ta infot vajab. Süsteem ei otsi tekste ainult sõnade sisaldumise järgi, vaid üritab küsimusest ja otsitavatest dokumentidest aru saada ning väljastada vaid küsimuse teemaga seotud vastuseid.

Mustrite kataloogid on eraldi tarkvaratooted või integreeritud olemasolevate tarkvarapakettidega nagu näiteks CASE

[4] http://www.hillside.net/patterns/writing/patternwritingpaper.htm, 28.09.2007
[5] http://www.antipatterns.com/whatisapattern/, 20.09.2007
[6] http://c2.com/ppr/, 20.09.2007
[7] http://www.wizsoft.com/, 20.09.2007

(*Computer-aided Software Engineering*)[8]. Need vahendid võimaldavad kasutada mustreid mudelite koostamisel. Selleks peaks mustrite analüüsimiseks kasutatav andmebaas võimaldama kirjeldada süsteemi disaini kolmel tasandil (abstraktsioon, liides, detailid). Abstraktsioonitasandil kirjeldatakse süsteemi koosnevana mustritest ja nendevahelistest sõltuvusseostest. Kõik mustrisisesed detailid on peidetud. Liideses eristatakse klasse ja operatsioone. Detailide tasandil näidatakse klasse ja nendevahelisi seoseid (Yacoub & Ammer 1999: 667). CASE-vahendite laiendus *Rational-Rose*[9] (Forbrig *et al* 2001) võimaldab kasutada objektile orienteeritud disainimudelit, milles saab luua mustrite diagramme, näidata mustreid ja nendevahelisi seoseid. Süsteem säilitab muistreid ning otsib nende kohta uut infot. Iga mustri kirjeldamiseks kasutatakse klassidiagramme ja dünaamikadiagrammi, mis näitavab mustri staatilist struktuuri, st elementide omavahelist seost jne.

Mustritel tugineva andmeanalüüsi tugev külg on selles, et mustrid üldistatakse uuritavate objektide reaalse kasutuse alusel, mis näitab, kuidas ning missugustes omavahelistes seostes objektid tegelikkuses on. Sama lähenemine oleks mõttekas ka EVKK veaklassifikatsiooni puhul. Praegune märgendamise alus – lingvistiline taksonoomia – on paratamatult abstraktsioon (nii on traditsiooniliselt kirjeldatud keelesüsteemi ja keelenormi). Vigade mustreid leitakse aga korpuspõhiselt, st selle alusel, kuidas tegelikult vead õppijatekstides esinevad. Edaspidi võib leitud mustreid kasutada ühe võimaliku meetodina vigade otsimisel tekstist, vigade identifitseerimisel ja nendevaheliste põhjuslike seoste tuvastamisel. Koos muude keele automaattöötlusvahendite kasutamisega aitaks mustrite meetod korpuse aren-

[8]http://en.wikipedia.org/wiki/Computer-aided_software_engineering, 12.08.2007
[9]http://www-306.ibm.com/software/rational/, 22.09.2007 ja http://searchsmb.techtarget.com/sDefinition/0,,sid44_gci516025,00.html, 22.09.2007

dajatel muuta õppijakeele vigade analüüs läbipaistvamaks, mis ühtlasi kiirendaks korpuse poolautomaatsele märgendamisele üleminekut. Samas täiendaks see oluliselt teadmisi õppijakeele olemusest ja annaks hea aluse pedagoogilise grammatika tarvis.

Kirjandus

Atkins, Sue & Levin, Beth 1995. Building on a Corpus: A Linguistic and Lexicographical Look at Some Near-Synonyms. – International Journal of Lexicography, 8: 2, 85–114.

Davis, Gordon B. 2000. Information Systems Conceptual Foundations: Looking Backward and Forward. – Organizational and Social Perspectives on Information Technology / Ed. by R. Baskerville, J. Stage, J. I. DeGross. Dordrecht: Kluwer Academic Publishers, 61–82.

Eslon, Pille 2006. Eesti vahekeele korpusest korrelatsioonigrammatikani. – Eesti Rakenduslingvistika Ühingu aastaraamat 2 (2005) / Toim. H. Metslang, M. Langemets. Tallinn: Eesti Keele Sihtasutus, 11–24.

Eslon, Pille 2004. Mõningatest korrelatsioonidest vene ja eesti verbisüsteemis. – Toimiv keel II. Töid rakenduslingvistika alalt / Toim. J. Lepasaar, M.-M. Sepper. Tallinn: TPÜ kirjastus, 103–122.

Forbrig *et al* = Forbrig, Peter & Laemmel, Ralf & Mannhaupt, Danko 2001. Pattern-oriented Development with Rational Rose. The Rational Edge 1, http://www.ibm.com/developerworks/rational/library/content/RationalEdge/jan01/PatternOrienteddevelopmentwithRationalRoseJan01.pdf, 18.01.2007

Kaalep, Heiki-Jaan & Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite kompleksis. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: TÜ Kirjastus, 87–99.

Meszaros, Gerard & Doble, Jim 1997. A Pattern Language for Pattern Writing. – Pattern languages of program design 3. Boston: Addison-Wesley Longman Publishing Co., Inc.

Mindt, Dieter 1991. Syntactic Evidence of Semantic Distinction in English. – English Corpus Linguistics: Studies in Hounour of Jan Svartvik / Ed. by K. Aijmer, B. Altenberg. London: Longman,182–196.

Muichnek, Kadri. 1998. Korpuslingvistika. – Keel ja Kirjandus 1, 8–12.

Müürsepp, Kaili 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu

Puolakainen, Tiina 2001. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes Mathematicae Universitatis Tartuensis 12. Tartu.

Sampson, Geoffrey 1987. Evidence against the 'grammatical'/ 'ungrammatical' distinction. – Corpus Lingvistics and Beyond / Ed. by W. Meijs. Amsterdam: Rodopi, 219 - 226.

Sivia, Devindrjit 2006. Data analysis: a Bayesian Tutorial / Ed. by D. Sivia, J. Skilling. Second Edition. Oxford: Oxford University Press.

Taylor, Lita & Grover, Claire & Briscoe, Ted 1989. The Syntactic Regularity of English Noun Phrases. – Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics. Manchester: UMIST, 256–263.

Tooding, Liina-Mai 1999. Andmeanalüüs sotsiaaltedustes. Tartu: TÜ Kirjastus.

Yacoub, Sherif M. & Ammer, Hany H. 1999. Tool Support for Developing Pattern-Oriented Architectures. – Proceedings of the 1st Symposium on Reusable Architectures and Components for Developing Distributed Information Systems (RACDIS'99), Orlando, Florida, 665–670.

Applying various data analysis methods for identifying relatioships between error types

Summary

The article introduces statistical methods of analyzing interlanguage errors on the basis of Estonian Interlanguage Corpus (EIC), which is being developed by the Chair of General and Applied Linguistics of Tallinn University. In order to find the error patterns or correlations between errors and to specify the hierarchy of the error taxonomy (or *tree*), there are used the Pearson rank correlation coefficient and t-test. The article shows an example of the statistical analysis of correlative connections of the error class *lexico-grammatical errors*.

The error classification is based on linguistic approach, so the meanings of the words in context are important. The context of usage diffuses language items between different categories and such linguistic approach to error tagging causes the effect that many errors have more than one tag which slows down the development of the analysis model of EIC.

An alternative to this method is the hypothesis of diffused categories (the objective approach of different meaning indicators) of language items. The meanings of described language items

are clear-cut and do not take into account the context of use. The method enables the use of *patterns* in finding the connections between errors.

Keywords: corpus linguistics, linguistic hierarchy of error taxonomy, error patterns