KOMAVIGADE TUVASTAJA

Krista Liin

Ülevaade

Artiklis tutvustatakse eesti keelele komavigade kontrolliks loodud grammatikakorrektori prototüüpi. Kitsenduste grammatika formalismis kirjutatud reeglipõhine programm tegeleb tekstis üleliigsete ning puuduvate komade tuvastamisega, kuid ei tee veel parandusi. Märgendatakse sõnaliike ja -vorme, mille ees sageli komadega eksitakse: sidesõnu, küsisõnu ja verbide pöördelisi vorme. Reeglid põhinevad grammatikaõpikuis leiduvatel õigekirjareeglitel, lisaks on reeglite loomisel ja testimisel kasutatud kolme eri tekstiliiki kuuluvaid korpustest leitud tekstinäiteid. Peamine korpus koosneb internetiportaalide kommentaaridest kogutud väära komakasutusega lausetest, lisaks on komavigade tuvastajat testitud ka ajalehe- ja teadustekstidel. Artiklis antakse ülevaade vigade tuvastamisel tekkinud probleemidest ning nende võimalikest põhjustest. Saavutatud tulemused on võrreldavad Põhjamaade teiste grammatikakorrektoritega.

Võtmesõnad: automaatne veatuvastus, grammatikakontroll, komavead[1]

[1] Komavigade tuvastaja loomist on toetanud riikliku programmi "Eesti keele keeletehnoloogiline tugi" (2006–2010) projekt "Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid" (2006–2008).

1. Grammatikakorrektori ülesanded

Kirjutaja abivahenditest on eesti keele kättesaadav speller ehk õigekirjakorrektor, mis tuvastab kirjutamise käigus tekkivad ortograafiavead. Samas toetub speller piiratud lingvistilisele infole, vaadates reeglina korraga üksnes üht sõna ning vastava sõnavormi esinemist keeles. Seega jäävad spelleri poolt märkimata need juhud, kus on küll tehtud õigekirjaviga, kuid saadud mõni teine, konteksti sobimatu, kuid eesti keeles võimalik sõnavorm (näiteks *kindlasti* asemel on kirjutatud *kindalasti*). Abivahendit, mis arvestaks õigekeelsuse üle rohkema, morfoloogilise ja süntaktilise info põhjal kasutaks otsustamisel suuremat konteksti kui üksainus sõna ning kontrolliks siis ühe lause või mitmest lausest koosneva tekstiosa sobivust, nimetatakse grammatikakorrektoriks.

Grammatikakorrektor kasutab oma otsustustes nii morfoloogilisi kui ka süntaktilisi teadmisi, võimaldades tuvastada õigekirja-, ühilduvus-, kokku-lahkukirjutamise ja kirjavahemärgivigu, parandada sõnavalikut ja sõnajärge ning muudki. Kirjutaja seisukohalt ei piisa tekstiredaktorist, mis üksnes vigased kohad ära märgib. Õigekirjakontroll peaks võimaldama ka (poolautomaatset) parandust: pakkuma iga vea kohta vähemalt üht varianti, kuidas korrektne lause välja näeks. Samas on mõttekas hoida parandused vigade tuvastusest eraldi, kuna alati ei pruugi kohene parandamine soovitav olla. Näiteks on võimalik kontrollida lausete õigekirja keeleõppeprogrammides, kus õpilasele ei näidata kohe kogu infot, vaid antakse kõigepealt teada vea liigist ja kohast lauses ning alles teatud tingimustel näidatakse ülesande lahendust – korrektset lauset. Teisest küljest võib osutuda vajalikuks ka kohene automaatne korrektuur. Seda näiteks juhul, kui grammatikakorrektor on osa pikemast automaatsest protsessist ning järgmine samm, olgu selleks siis süntaksianalüüs, vajab vigadeta sisen-

dit. Niisiis oleks kasulik ehitada grammatikakorrektor kaheosalisena: vigade märgendamise osa, mida saab soovi korral kasutada koos teise, vigade korrigeerimist võimaldava osaga.

2. Komavigade tuvastamine

Sageli alustatakse grammatikakorrektori loomist kas ühilduvus- või interpunktuatsioonivigade märgendamisest. Enamasti valitakse välja üks nimetatud vealiikidest ning jäetakse teise veatüübi kontroll hilisemaks. Seda tehakse tihti arvestusega, et tolle sisendis on esimest liiki vead juba parandatud. Arvamused selle kohta, kas enne tuleks tegelda kirjavahemärkide või ühilduvusega, lähevad lahku. Eckhard Bick (2006: 9) argumenteerib taani grammatikakorrektori kirjelduses, et komavigu tuleks tuvastada grammatilisemas kontekstis, kus lause on juba muus osas keeleliselt korrektne. Teisest küljest toovad baski keele grammatikakorrektori loojad välja, et muid ühildumisvigu on tunduvalt kergem tuvastada ja parandada, kui komad ja seega ka osalausepiirid on korrektselt määratud (vt Aldezabal jt 2003: 1). Niisiis pole selget üksmeelt, millist veatüüpi tuleks kõigepealt korrigeerida, kuna iga vea vähendamine aitab pea alati kaasa järgmiste paremale tuvastamisele.

Eesti keele grammatikakorrektori puhul on valik langenud komavigade tuvastamisele. Eksimused kirjavahemärkide kasutuses on muudest vealiikidest võrdlemisi kergesti eristatavad, samuti võiks eeldada, et kuna kirjavahemärke kasutatakse üksnes tekstis, mitte suulises kõnes, siis sõltuvad selle valdkonna vead suhteliselt vähe muus osas keeleliselt korrektse lause moodustamisest. Teisisõnu, komavigu teevad kirjas ka need, kes on suulise keele suurepäraselt omandanud. Teine põhjus komavigade valikuks oli asjaolu, et teiste keeltega võrreldes tundub komakasutus eesti keeles olevat suhte-

liselt kindlalt reguleeritud, samas aga suhteliselt keeruline ja raskesti omandatav. Seega on olemas praktiline vajadus komavigade automaatse tuvastamise järele.

Komakasutuses on võimalik eristada kaht tüüpi vigu: koma ärajätmine seal, kus see olema peaks (puuduv koma), ja koma asetamine sinna, kus seda vaja pole (üleliigne koma). Vaadates üldist keelekasutust, võib öelda, et pigem jätame koma lausest välja, olgu selle põhjuseks siis hooletus või eeldus, et paus lauses mõne muu vahendiga esile tuuakse. Nii on juhtunud ka näites (1), kus paksus kirjas olevate sõnade ees peaks koma olema.

(1) *Müüsin vana auto maha hetkel **mil** oleks pidanud riskigrupp langema 0.77-0.60-le **aga** uut autot kindlustama minnes tõusis hoopis 0.87.*

Kuigi valitseb tendents koma lausest välja jätta, on ka juhtumeid, kui kas segadusseajavatest õigekirjareeglitest juhindudes või mõnel muul põhjusel kirjutatakse koma sinna, kus see reeglite kohaselt olema ei peaks ega olla tohiks. Sel juhul eksitakse pigem kontekstis, kus sarnaste sõnade juures on võimalik nii koma kasutus kui ka koma ärajätmine. Nii on näites (2) eksitud koma kasutamisega sidesõna *kui* ees: koma pannakse vaid siis, kui võrdluse teises pooles esineb verb.

(2) *Ei väike hiinlanna ei olegi parem**, kui** väike venelanna Kohtla-Järvelt.*

Eri keelte grammatikakorrektorites on valdavalt püütud lisada puuduolevaid komasid, vähematel juhtudel ka eemaldada üleliigseid. Nii Bick (taani keel) kui ka Izaskun Aldezabal jt (baski keel) pigem lisavad puudu olevaid kirjavahemärke (Bick 2006; Aldezabal jt 2003). Daniel Hardt (2001), kes püüdis masinõppemeetodil luua taani keele grammatikakorrektorit, tuvastas seevastu vaid üleliigseid komasid, kasutades õppe-

alusena ajalehetekste, kuhu oli suvaliselt komasid lisatud. Baski keele teisel, masinõppemeetodil loodud grammatikakorrektor tuvastab mõlemat tüüpi komavigu, kusjuures tulemused osutavad, et puuduolevate komade leidmine on märgatavalt keerulisem (Alegria jt 2006). Eesti keele puhul võtsin arvesse nii puuduolevaid kui ka üleliigseid komasid, kuid mõlemat tüüpi vaid teatud kontekstis. Teisisõnu kontrollisin komade esinemist seal, kus sagedamini eksitakse, ega arvestanud võimalike üleliigsete komadega muude sõnade ümbruses.

Üleliigsete komade korral piisab vea parandamiseks selle tuvastamisest: kui on teada, et koma on tekstis sobimatu, siis tuleb see vaid eemaldada. Ent kui mõni koma, mis tekstis olema peaks, on tegelikult ära jäetud, ei piisa parandamiseks teadmisest, et koma tuleb lausesse lisada, vaid peab ka teadma, kuhu see täpselt lisada. Mõnel juhul, näiteks koma nõudvate sidesõnade korral on seegi ülesanne küllaltki lihtsalt lahendatav. Teistel juhtudel võib osalausepiiri leidmine osutuda keerulisemaks. Näiteks on raske määrata, kus täpselt peaks kahe finiitse verbivormi vahel koma olema ja kas vahepealsed sõnad, iseäranis määrused, kuuluvad pigem esimesse või teise osalausesse. Sellest tulenevalt tegelen esialgu vaid komavigade tuvastamise ja vastavate märgendite lisamisega, parandus-te väljapakkumiseks läheb vaja täiendavat analüüsi. Samas on mõnel juhul küllaltki kerge juba vastavalt olemasolevale veamärgendile otsustada, kuidas lause õigekeelsust parandada.

Komavigade tuvastamisel on aluseks võetud grammatikakäsiraamatus (Erelt 2006) komakasutust käsitlevad õigekirjareeglid, tuvastaja otsib lausetest nii üleliigseid kui ka puuduolevaid komasid. Tuvastajasse on valitud need reeglid, mis määravad komakasutuse üheselt ära, s.t välja on jäetud olukorrad, kus on põhimõtteliselt õige nii komaga kui ka komata kasutus, millest ühte vaid hea stiili huvides soovitatakse. Nagu maini-

tud, ei ole sisse võetud kõik vastavad käsiraamatus esinevad reeglid, vaid üksnes need, mis normeerivad koma olemasolu või puudumist side- ja küsisõnade ning osalausete öeldiste vahel.

3. Lähenemine

Nagu eespool öeldud, kasutatakse grammatikakorrektorite juures ja keeletehnoloogias üldse nii reeglipõhiseid, statistilisi kui ka hübriidmeetodeid. Sealjuures on sama keele jaoks proovitud eri tüüpi lahendusi, kuigi enamasti mitte samade vealiikide puhul.

Eesti keele grammatikakorrektori loomisel on aluseks võetud reeglipõhine lähenemine ja kitsenduste grammatika formalism. Kitsenduste grammatika formalismi on välja töötanud Fred Karlsson Helsingi Ülikoolist, mõeldes sealjuures eelkõige süntaktilisele analüüsile. Hiljem on seda formalismi kasutatud ka paljudes muudes valdkondades, sealhulgas grammatikavigade märgendamisel. Reeglid tegelevad kas märgendite lisamise (lubades ühe sõna juures mitu eri märgendit) või õige märgendi määramisega, s.t sobiva valiku määramise või ebasobivate valikute eemaldamisega. Lisaks märgendatavale sõnale saab reeglis määrata kontekstimustri nii sõnade kui ka märgendite kujul, millal reeglit rakendada. Sealjuures on konteksti ulatuseks vaikimisi küll lause, kuid selle piire on võimalik vajadusel muuta. Kitsenduste grammatika reeglid on küllaltki arusaadavad ning hõlpsasti muudetavad, mis on üks reeglipõhise lähenemise eeliseid statistiliste või masinõppesüsteemide ees (Karlsson jt 1995: 42).

Üheks põhjuseks formalismi valikul oli asjaolu, et eesti keele automaatne morfoloogiline ning süntaktiline analüüs kasutavad sama meetodit, mis muudab eri tööriistade integreerimise

lihtsamaks. Et kitsenduste grammatikas on reeglid jaotatud eraldi moodulitesse, mida järjest rekursiivselt rakendatakse, siis on formalism iseäranis sobiv. See võimaldab muuhulgas vigade leidmise ja korrigeerimise jagamist eri moodulitesse või ka ühilduvus- ja komavigade järkjärgulist kontrolli, kasutades igal järgmisel analüüsiringil juba suurema kindlusega sooritatud parandusi. Kitsenduste grammatika puudusena võib välja tuua, et ehkki sõnade märgendeid on võimalik muuta ja sel moel näiteks ühilduvusvigu parandada, on formalism mõeldud eelkõige siiski analüüsiks ega võimalda sõnade (või kirjavahemärkide) lisamist ega eemaldamist. Selle lähenemise olen võtnud aluseks vigade leidmisel ja märgendamisel.

Kuna märgendada saab vaid lauses esinevaid sõnu, siis pole võimalik puuduolevale komale märgendit lisada. Seega on märgendamisele võetud pigem need sõnad, mille läheduses on oht komakasutuses eksida: sidesõnad ja küsisõnad ning samuti verbide pöördelised vormid, mille vahel peab alati leiduma kas sidesõna või koma. Märgendamise algul lisatakse kõigile kaks märgendit 'õige' ja 'väär', järgnevalt valitakse kitsendusreeglitega välja korrektne märgend. Siinkohal tähistab 'väär' olukorda, kus sõna ees esineb komaviga. Kuna lauses võib olla mitu märgendatavat sõna, võib osa neist saada ühe, teised teise märgendi, mis aitab selgitada vea asukohta. Arvestades, et eelistatum on olukord, kus grammatikakorrektoril jääb mõni keerulises lauses esinenud viga märkamata, kui liig sagedased valeteated, on reeglite koostamisel seatud eesmärgiks pigem võimalikult täpsed kui laiad reeglid ning võimalikult vähene valealarmide hulk.

See tähendab, et iga reegli puhul tuli vastavalt võimalikele eranditele täpsustada konteksti ning kaheldavas olukorras jätta alles pigem korrektset kasutust tähistav märgend. Nii

näiteks on ühendi *nii et* kasutuses lubatud koma panna nii ühendi ette kui ka vahele, kuid üldjuhul peab ühend lauses esinema koos komaga. Kuna aga erandjuhuna on lubatud püsiühendite (näiteks "*Valetab nii et suu suitseb.*") korral koma panemata jätta, siis tuli tuvastaja reeglites lubada ka komata kasutus, kuigi sel juhul võib osa puuduolevaid komasid leidmata jääda.

Reeglite koostamise aluseks olid grammatikaõpik, millest saab lause õigekeelsuse reegleid arvutiformalismi ümber kirjutada, ja tekstikorpused, millel reegleid testida ning mille põhjal katmata jäänud juhtumeid lisada. Kõik õigekirjareeglid kasutatud korpusel rakendust ei leidnud, kuid samas tuli välja mitu juhtumit, kus tegelik keelekasutus tingis vähemalt olemasolevate reeglite täpsustamise või neile erandjuhtumite lisamise, kuna automaatne analüüs vajas detailsemaid juhendeid.

4. Korpused

Reeglite koostamisel kasutasin kolme liiki tekstikorpusi, mis olid automaatselt morfoloogiliselt ja süntaktiliselt analüüsitud ning grammatikavead käsitsi märgendatud. Peamine korpus, mille põhjal uusi reegleid konstrueerida, koosneb Delfi internetiportaali foorumite kommentaaridest kogutud grammatiliselt vigastest lausetest, milles on kokku üle 9000 sõna. Tegu on nii teemade kui ka autorite poolest varieeruva tekstiga, mis võimaldab vaadelda erinevate inimeste keelekasutust ning eri eksimistasemeid õigekirjas. Nii on eespool toodud näites (1) kas tahtlikult või tahtmatult jäetud välja kõik komad, samal ajal kui näites (3) on autor osa komasid välja kirjutanud ning vaid kaks ära jätnud.

(3) *Praegu ongi nii **et** palu jumalat, et juhul kui satud süütult õnnetusse, siis süüdioleval osapoolele oleks norm kindlustaja (mitte Salva või Inges näiteks) – muidu jäädki uste vahet jooksma ja aega raiskama, õnnetuse pärast **milles** sina üldse süüdi pole.*

Kuna tegu on võrdlemisi spontaanse tekstiga, mida eelnevalt tõenäoliselt kuigivõrd ei kontrollita, siis on internetikommentaarides õigekirjavigade osakaal suhteliselt suur. Samas on tegu loomulike eksimustega, mitte konstrueeritud vigadega, nagu näiteks rootsi grammatikakorrektori treenimisel tehti (vt Hardt 2001). Grammatikakorrektori koostamisel on mõttekam kasutada võrdlusallikana tegelikku teksti, millega sarnasele hiljem tööriista rakendama hakatakse, kui tehislikult vigast teksti, mida korrektor hiljem tõenäoliselt ei kohta. Sarnastel põhjustel kasutati ka taani keelele grammatikakorrektori loomisel düsgraafikute kirjutatud suurte vigade osakaaluga tekste (vt Bick 2006).

Kuna üks oluline kriteerium reeglite koostamisel oli valealarmide vältimine, siis kontrollisin komavigade tuvastaja tööd ka ligi sajalauselisel Eesti keele koondkorpuse[2] osal. Eeldatavasti on publitseeritud tekstide õigekeelsus kontrollitud ning nii võis lihtsalt, vähese märgendamise vaevaga kasutada küllaltki suurt korpust. Kuigi uute reeglite koostamisel polnud kirjakeele korpusest suurt kasu, sai sellel kontrollida loodud reeglite töökindlust.

Võib arvata, et internetikommentaarid ei esinda just kõige tüüpilisemat kirjakeelt – laused on küllaltki pikad, sisaldavad palju osalauseid, ei pruugi olla struktureeritud samal moel kui tüüpiline tekstiredaktorisse kirjutatav tekst. Seetõttu kontrolliti komavigade tuvastaja reegleid ühel magistritööl (Liin 2008). Selgus, et teiste tekstiliikide põhjal koostatud reeglid toimivad

[2] http://www.cl.ut.ee/korpused/segakorpus/ (20.03.2009).

hästi ka tüüpilise arvutil kirjutatud teksti puhul. Programm leidis üles isegi ilmselt meediumipõhise tekkega vead nagu tekstiosa kopeerimise tõttu lausesse jäänud korduvad verbivormid, vt näide (4).

(4) *Muul juhul **peab** koma nõudva sidesõna ees **peab** olema koma.*

5. Tulemused

Komavigade tuvastaja koostamisel valmis 98 kitsendusreeglit, mida on küllaltki palju üksnes komavigade määramiseks. Sealjuures 4,5% juhtudest ei õnnestunud korrektset märgendit valida. Probleeme tekkis oodatult seal, kus komakasutus sõltub rohkemast kui morfoloogilisest infost ja arvesse tuleb võta ka kõrvallause sisu. Küllaltki sage on siinkohal selline juht, kus tuleb otsustada, kas osalause kuulub eelneva kõrvallause või pealause juurde – näites (5) tuleks *ja* ette koma panna just seetõttu, et sellele järgnev osalause on rinnastatud pealausega, viimast on aga automaatselt raske tuvastada.

*(5) Täna siis üks õnnetu opeli juht ei leidnud oma suunatule kangi üles kui reastus **ja** ühe kaubiku juhil puudus ka lisavarustuses suunatule kang.*

Teine suurem raskuste tekkepõhjus on asjaolu, et morfoloogilise analüüsi ja ühestamise tööriistad on mõeldud kasutamiseks keeleliselt korrektsetel lausetel ning seetõttu eksisid need vigaste lausete sõnadele analüüsil. Sääraseid analüüsivigu on võimalik mingil määral parandada, kui arvestada süntaktilise infoga, mis ongi üks grammatikakorrektori eesmärkidest. Lihtsaim juhtum on trükiviga, mis on tuvastatav kui keeles mitteesinev vorm; sõnade kokkukirjutamise või mõne muu sõnavormiga kattuvuse korral võib see aga probleemsemaks osutuda. Grammatikakorrektori segadusse ajamiseks piisab sellestki, kui pärast koma on tühik ära jäänud

(vt näide 6) – eelneva parandamiseta ei suuda märgendajad teist osalauset korrektselt märgendada ja eitussõna *pole* saab komavea märgendi.

(6) *Alguses öeldi kohe et väljaload unustage **ära,kui** just midagi pakilist **pole**.*

Teine vigase analüüsi põhjus on viga morfoloogilisel ühestamisel. Kuna morfoloogiline ühestaja arvestab otsuste tegemisel korrektse kontekstiga, siis võib see mõne ärajäänud või vale sõnavormi läheduses valida võimalikest analüüsidest tegelikult sobimatu. Nii on näites (7) analüüsitud sõnavormi *saate* kui nimisõna ilmselt just seetõttu, et sellele paistab samas osalauses eelnevat teinegi finiitne verbivorm.

(7) *Ise nad ju ütlesid et otsige süüa sealt **kust saate***

- *saate("saade+0" // _S_ com sg gen ADVL*

Lisaks võimalikule valele valikule võib morfoloogiline ühestaja põhjustada raskusi ka siis, kui liiga vähese info tõttu jäetakse sõnale alles mitmene analüüs, millest osa analüüse osutab lause ebakorrektsusele, teised aga sobivad olemasolevasse lausesse probleemideta. Näites (8) on lause analüüs raskendatud, kuna sõnal *täis* on alles jäetud ka verbi märgend.

(8) *Üks tüüp kes oli pidev hüppeskäia oli juba paar kuud teenistuse lõpetanud, kui juua **täis** peaga üle väeosa aia ronis.*

- "täis+0" // _A_ pos AN> PRD
- "täis+0" // _D_ ADVL
- "täis+0" // _S_ com sg nom SUBJ
- "täi+s" // _S_ com sg in ADVL NN>
- **"täi+s" // _V_ main indic impf ps3 sg ps af #FinV #InfP +FMV**

Määratud sõnade puhul oli komavigade tuvastaja täpsus 95% ja saagis 93%. Sellega oli täidetud üks algselt püstitatud eesmärk – viia ekslike veamärgendite hulk miinimumini. Lisaks õnnestus leida üllatavalt suur osa tekstides leidunud komavigadest. Lausetes tervikuna on tulemused veelgi paremad, kuna ühe sõna märgendamisel tehtud vea võib kompenseerida teise sõna märgend. Nii on näites (9) *vaid* märgitud ekslikult sidesõnaks, mille tõttu *mis* kohta käivad reeglid lubavad seal ekslikult komata kasutuse. Samas aga rakendub sidesõna *vaid* ette koma nõudev reegel ning lauses leitakse siiski komaviga. Samuti annaks alarmi sõna *on*, kuna reeglite kohaselt peab kahte pöördelist tegusõna eraldama kas koma mittenõudev sidesõna või koma.

(9) *Ta ju küsis **vaid mis** vahe **on** automaat- ja poolautomaatkäigukastiga autol?*

Nii õnnestuski vältida valealarme, kuigi korrektselt märkimata lauseid oli ligi 5%, mis tähendab, et 150 kontrollitud lausest, millest pooled sisaldasid süntaksivigu, jäi leidmata 4 vigast lauset.

Kokkuvõtteks

Komavigade tuvastamisel saavutatud täpsus näitab, et loodud tööriist on võrreldav muude keelte grammatikakorrektoritega: kitsenduste grammatikas kirjutatud Põhjamaade grammatikakorrektorite töötäpsused jäävad vahemikku 70–95% (vt Hagen jt 2001; Hagen jt 2002: 4). Samas tuleks arvestada, et teiste vealiikide lisamisel ning pärast vea tuvastamist sellele korrektse paranduse väljapakkumisel võib täpsus väheneda. Sellest hoolimata võib kindlalt väita, et antud lähenemine on eest keele grammatikakorrektori loomisel tulemuslikuks osutunud.

Nagu reeglite koostamisel selgus, oli korpuste põhjal uuritud tegelikest keelenäidetest grammatikakorrektori töö parandamisel palju kasu. Ka testimisel tuvastamata jäänud vigade hulgas oli selliseid, milles suurema hulga näidete kasutamisel oleks saanud eksimise välistada. Seega tuleks reeglite loomisel ja täpsustamisel edaspidi kasutada suuremat korpust ja kindlasti mitte piirduda üksnes üht tüüpi keelekasutusega, et vältida reeglite ülepassitamist ja sellest tulenevat sobimatust teistele keelekasutustele. Grammatikakorrektori loomisel on plaanis abiks võtta Tartu Ülikoolis kirjutatud bakalaureusetööde mustandversioone, milles kõik vead veel parandatud pole, ja keeleõppijate kirjutatud tekste, milles on samuti ohtrasti vigu.

Kui tekstis on komavead parandatud ja osalausepiirid paigas, siis saab grammatikakorrektorile lisada muude kirjavahemärkide kasutuse, ühildumise, rektsiooni, kokku-lahkukirjutamise, konteksti sobimatute sõnade (näiteks kirjavea tõttu vale tähenduse saanud sõnavormide) leidmise ja asendamise ning muudegi vealiikide määramise. Veatuvastusreeglite loomise käigus tuleb lisaks mõelda parandusettepanekute koostamisele, mida läheb vaja grammatikakorrektori automaatsel rakendamisel ning kasutaja töö kergendamiseks. Tekstiredaktoris tasub kirjutaja abivahendina mõelda stiilikorrektorile, mis aitaks vältida mitte üksnes otseseid vigu, vaid ka kordusi ning sobimatut sõnakasutust.

Kirjandus

Aldezabal jt 2003 = Aldezabal, Izaskun & Aranzabe Maxux & Arrieta Bertol & Maritxalar Montse & Oronoz Maite 2003. Toward a punctuation checker for Basque. ATALA workshop of punctuation (Paris). http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1069080468/publikoak/Toward-a-punctuation-checker-for-Basque.pdf (20.03.2009).

Alegria jt 2006 = Alegria, Iñaki & Arrieta, Bertol & Díaz de Ilarraza, Arantza & Izagirre, Eli & Maritxalar, Montse 2006. Using Machine Learning Techniques to Build a Comma Checker for Basque. Coling-ACL. Sydney. Australia, 1–8. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1150185248/publikoak/komak-ML.pdf (20.03.2009).

Bick, Eckhard 2006. A Constraint Grammar Based Spellchecker for Danish with a Special Focus on Dyslexics. – SKY Journal of Linguistics. Vol. 19. http://www.ling.helsinki.fi/sky/julkaisut/SKY2006_1/1.6.1.%20BICK.pdf (20.03.2009).

Erelt, Mati 2006. Lause õigekeelsus. Juhatused ja harjutused. Bookmill.

Hagen jt 2001 = Hagen, Kristin & Lane, Pia 2001. "Det er fort gjort og skrive feil." En presentasjon av en automatisk grammatikkontroll for bokmål. Foredrag på Mons, Oslo. http://www.hf.uio.no/tekstlab/prosjekter/Mons.gr-sjekker.htm (20.03.2009).

Hagen jt 2002 = Hagen, Kristin & Johannessen, Janne Bondi & Lane, Pia 2002. The performance of a grammar checker with deviant language input [Proceedings of the 19th International Conference on Computational Linguistics.]. Vol. 2. Taipei, Taiwan. http://portal.acm.org/citation.cfm?id=1071884.1071894 (20.03.2009).

Hardt, Daniel 2001. Transformation-Based Learning of Danish Grammar Correction [Proceedings of RANLP 2001]. http://www.id.cbs.dk/~dh/papers/ranlp.pdf (20.03.2009).

Karlsson jt 1995 = Karlsson, Fred & Voutilainen, Atro & Heikkila, Juha & Anttila, Arto 1995. Constraint Grammar: A Language-independent System for Parsing. Berlin and New York: Walter de Gruyter. http://books.google.com/books?hl=en&lr=&id=70IvVPlH63cC&oi=fnd&pg=PP10&dq=Constraint+Grammar:+A+Language-independent+System+for+Parsing&ots=mA5pxnqGGb&-sig=ajX5aV_JUpdNp5FmhqpscP9UhIY (20.03.2009).

Liin, Krista 2008. Reeglipõhine komavigade tuvastaja eestikeelsetele tekstidele. Magistritöö. Juhendaja Kaili Müürisep. Tartu Ülikool, matemaatika-informaatikateaduskond, arvutiteaduse instituut. Tartu.

Grammar checker for detecting comma mistakes

Krista Liin

Summary

The aim of this grammar checker was to detect comma mistakes in written Estonian. As of yet, the checker does not suggest corrections, but that function could be added to the existing system later. The grammar checker rules are based on the Constraint Grammar Formalism framework.

The corpus used for rule development and testing consists of grammatically incorrect sentences gathered from the user postings on an Internet site. More than 9000 words were first morphologically and syntactically analyzed, and then manually tagged for comma error detection. Finite verb forms, interrogative words and conjugations were tagged and marked as correct or incorrect, depending on whether there was a comma mistake before those words.

The 98 constraint rules were tested on a 150-sentence test corpus of both incorrect and correct sentences. A precision of 95% and recall of 93% was achieved on tagged words. There were no sentence-level false alarms. The problems in detecting mistakes were mainly caused by incorrect spelling, previous tagging, or situations where the usage of comma depends on semantic information.

The results achieved are comparable to other grammar checkers. In the future, the grammar checker for Estonian will be further developed using larger corpora and targeting also other error types, such as agreement mistakes. The aim is to

test it also on the texts written by language learners and on other text types.

Autor

MSc Krista Liin, Tartu Ülikooli doktorant, Tartu Ülikooli arvutiteaduse instituudi projektijuht, krista.liin@ut.ee