

Reeglipõhine komavigade tuvastaja eestikeelsetele tekstidele

Krista Liini magistritööst

Põhijooned

- Reeglite loomisel oli eesmärgiks luua võimalikult täpsed reeglid ehk püüda vältida valealarmide hulka nii palju kui võimalik.
- Eesmärgiks polnud koostada mitte niivõrd kontrollija, vaid tuvastaja, mis leiab küll vead üles, kuid ei soovita parandatud, korrektset lauset.
- Keskendub komavigadele.
- Kasutab morfoloogianalüsaatori ja süntaksianalüsaatori väljundit.

Töö käik

- Esimese sammuna lisatakse kaks märgendit, @ERR ja @OK, vastavalt vigase ning korrektse koha märgendamiseks, kõigile sidesõnadele ja verbi pöördelistele vormidele.
- Alternatiivina võiks lisada märgendid kõigile sõnadele sisendtekstis, ent see raskendaks otsustamist, millise sõna juures osutada puuduvale komale.
- Õige märgendi valik toimub kas Select- või Remove-tüüpi reeglite abil, kuid mitmese sisendi puhul Select on parem.

Reeglite allikad

- Mati Erelt. Lause õigekeelsus. 2006, valides välja toodud komakasutuse reeglitest need, mille automaatne kontroll teostatav tundus.
- Reegleid katsetati korpusel.
- Korpuse lausetest.

Ei tuvastata

- teiste kirjavahemärkide puudumist
- semantiliselt ja pragmaatiliselt põhjendatud komasid (eriliigilised vrs samaliigilised määrused)
- kiilu komakasutus

Keskendutakse

- sidesõnade komareeglitele, sest nende eelnev analüüs on suhteliselt ühene ja veatu
- pigem morfoloogiline märgendus kui süntaktiline

Reeglite rühmad

- Pöördelised verbivormid
- Sidesõnad
 - kui ja nagu
 - siis ja mitte
 - ja, ning, ega, ehk, või
 - ülejäänud

Reeglite näited

- Üldised reeglid
- Kui märgendatav sõna paiknes lauses esimesel positsioonil, määrasin ta korrektseks.
- Kui kaks sidesõna asuvad järjest, siis nende vahele koma ei panda. Nt Tuleks tõrelda, kuid et ta on nii tundlik, siis piisab ehk märkusest.
- Kui küsisõnale eelneb sidesõna, siis nende vahele koma ei panda. Nt Ja mis see sinu asi on?

Näited

- Kui sõnale *siis* järgneb vahetult sõna *kui*, siis nõuda nende vahele koma. Nt *Tulen siis, kui tahan.*
- Kui sõnale *siis* eelneb lauses sõna *kui* ning järgneb võimalik öeldis, siis juhul, kui sõnade *kui* ja *siis* vahel leidub võimalik öeldis, nõuda *siis* ette koma. Nt *Kui teha, siis teha hästi.* Vastasel juhul koma mitte panna. Nt *Ja kui siis hakkas sadama.*
- Kui järjest esinevad sõnad *kui*, *siis* ja küsisõna, siis nõuda *siis* ette koma. Nt *Ja kui, siis mida?*

Korpusest

- Antud magistritöös kasutatud korpuseks on Heli Uibo poolt Delfi internetiportaali kommentaaride hulgast kogutud grammatiliselt ebakorrektsed laused (kättesaadavad veebis aadressil http://www.ut.ee/~heli_u/vigadekorpus/). Kasutatud 298 lauses on pea igaühes üks või mitu koma puudu.

Vead

- Näide 1: Verbita lauselühendist tulenev viga:
Meil nägu hämmingus peas aga süda rahul kui leitnant 90 kraadi kannaka tegi ja jooksega sinna pani.
- Näide 2: Morfoloogilisest ühestamisest tulenev viga: *Ise nad ju ütlesid et otsige süüa sealt kust saate("saade+0" // _S_ com sg gen ADVL).*
- Näide 3: Liiga mitmesest analüüsist tulenev viga: *Üks tüüp kes oli pidev hüppeskäia oli juba paar kuud teenistuse lõpetanud, kui juua **täis** peaga üle väeosa aia **ronis**.*

Vead

- Näide 6: Koondlause eraldajana esinevast sidesõnast tulenev viga: *Kui keegi möödub ohutult **ja** sind segamata mis selles halba **on** kui sa ise takistav tegur olid.*
- Näide 9: Osalause taseme valesti määramisest tulenev viga: *Täna siis üks õnnetu opeli juht ei leidnud oma suunatule kangi üles kui reastus ja ühe kaubiku juhil puudus ka lisavarustuses suunatule kang.*

Tulemused

- Grammatikakorrektor tegi testkorpusel märgendite valimisel vea kokku üheksas kohas ehk 7,3% juhtudest, lisaks eemaldas kahel korral mitmese analüüsiga sõnadelt määrsõna analüüsid. Kuuel juhul jäi komaviga vastavas sõnas tuvastamata, kolmel korral aga määrati viga sinna, kus seda tegelikult ei esinenud. Vead tekkisid kuuel juhul verbide märgendites, lisaks sõnade kas, siis ja nagu märgendamisel.

Tulemused

- Korrektselt tuvastatud komavigade osakaal kõigist grammatikakorrektori poolt antud komavigade märgenditest ehk grammatikakorrektori täpsus testkorpusel on 93,8%.
- Saagis ehk leitud komavigade suhe korpuses leidunud (ja käsitsi märgendatud) komavigadesse on 94,1%.
- 30 lause hulgas oli 27 komaveaga, neist leiti 24.