Morfoloogiline analüüs, morfoloogiline mitmesus ehk kuidas tehakse tekstist lemmad ja mis sealjuures valesti saab minna

Kaks protsessi: analüüs ja ühestamine

Morfoloogiline analüüs lisab tekstisõnale kõik tema võimalikud tõlgendused, konteksti ei arvestata:

demo http://www.filosoft.ee/html_morf_et/html_morf.cgi

Kana sai ära söödud.

Masin ei ole kõikvõimas

ehk mis saab valesti minna **analüüsil**:

tundmatud sõnad st

- sõnad, mis puuduvad analüsaatori leksikonis ≈ pole ÕS-is
- ja mis pole analüüsitavad liitsõna või tuletisena

nt *selfi, äpp, mannoos-6-fosfaat*

nt *Ka täna ommikul olli Konrad tema ainokene mõtte*

Mida ei tea, seda oletab: demo

Morfoloogiline ühestamine

on sõnavormi **õige analüüsi väljavalimine** konteksti põhjal.

ei suuda saavutada 100% ühest tulemust

sagedasemate lahendamata jäävate mitmesuste näited:

on ole+0 //_V_ b, // ole+0 //_V_ vad, //

läinud läinud+0 //_A_ // läinud+0 //_A_ sg n, // läinud+d //_A_ pl n, // mine+nud //_V_ nud, //

arvatud arva+tud //_V_ tud, // arva=tud+0 //_A_ // arva=tud+0 //_A_ sg n, // arva=tud+d //_A_ pl n, //

Taali Taal+0 //_H_ sg g, // Taali+0 //_H_ sg g, // (vokaaliga lõppeva pärisnimevormi või üldse oletatud sõna lemma)

Vigade mõju lemmadele

Oletamine: vokaallõpulise tundmatu sõnavormi puhul ei ole sageli võimalik öelda, kas nimetav kääne on vokaal- või konsonantlõpuline. *Liisi* (*Liis* vs *Liisi*), *Jõekalda* (*Jõekallas* vs *Jõekalda*), *selfi* (*self* vs *selfi*)

Mitmeseks jäänud ühestaja väljundi puhul võetakse lemmatiseerimisel tüüpiliselt lihtsalt mitmest alternatiivsest lemmast esimene, aga järjestus on paraku juhuslik

nt sisend: *See töö on meil korda läinud. Tehtud tööga on kõik korras.*

Väljund: *see töö olema mina kord läinud . tegema töö olema kõik korras .*