

# **VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)**

Pille Eslon

EKKTT kolmas konverents

Tartu, 25.-26. november 2010



# Riikliku programmi "Eesti keele keeletehnoloogiline tugi (2006–2010)" toetus projektile VAKO aastatel 2008–2010

2008 – 350 000

2009 – 494 000

2010 – 400 000

Kokku: 1 244 025

# Põhitäitjad ja täitjad

Vastutav täitja: PhD Pille Eslon

Põhitäitjad: PhD Erika Matsak, doktorant Kairit Sirts, MSc  
Jaagup Kippar, doktorant Helena Metslang, magistrant  
Hanna Sinijärv, magistrant Anne Kostenko, magistrant  
Anni Muru

Täitjad: PhD Annekatrin Kaivapalu, MA (teaduskraad) Ellen  
Dovgan, BA Vahur Rebas



# VAKO-projekti põhieesmärgid 1

- ▶ Esmane: olemasolevat keeletarkvara arendades luua EVKK automaatseks töötlemiseks sobivad tarkvararakendused, mis võimaldavad korpuse tekstide käsitsimärgendamisel üle minna poolautomaatsele
  - ▶ Luua **vealeidja prototüüp**, mis sisaldaks
    - ▶ morfo- ja süntaksianalüsaatorit
    - ▶ oleks seotud vealiigi määramisega korpuse lingvistilise veataksnoomia alusel



# VAKO-projekti põhieesmärgid 2

- ▶ Teine eesmärk – EVKK funktsionaalsuste laiendamine
  - ▶ õppijakeele elektroonne sõnastik
  - ▶ õppijakeele sagedussõnastik
  - ▶ uus vealiigi märgendusmoodul
- ▶ Kolmas eesmärk – EVKK kasutajaliidese täiustamine, võimaluse loomine uute alamkorpuste tekitamiseks
  - ▶ Seotud EVKK olemasoleva keeletehnoloogilise ressursi suurendamisega (REKK-i kogud; akadeemilise eesti keele kui K1 ja K2 allkorpus)



# Tulemused 1

- ▶ Esimene eesmärk täidetud
  - ▶ loodud EVKK sõnajärjevealeidja prototüüp, mis implementeeritud EVKK-sse
    - ▶ prototüübi graafiline liides on valmimisjärgus
  - ▶ Sõnajärjevealeidja prototüübi programmeerimisel võeti aluseks leitud õigete sõnajärjemallide kogum. Realiseerimiseks on kasutatud Zope andmebaasi ning programmeerimiskeelt Python.
    - ▶ statistikapõhine programm



# Tulemused 1: järg

- ▶ programm testib lauseliikmete järgnevusi ehk järjendeid esimeses osalauses ja lihtlauses
- ▶ korduvad lauseliikmete järjendid moodustavad sõnajärjemustreid
- ▶ sõnajärjemustrid on paigutatud andmepuudesse, kust prototüüp otsib õige algusmärgendiga puu, siis sagedasema sõnajärjemustri, mida eesti keeles selle algusmärgendiga kasutatakse



# Näide 1

- ▶ sõnajarje seisukohalt on olulised finiitne ja infiniitne verb

@FMV – finiitne verb

@IMV – infiniitne verb

@FCV – olema liitaegades ning modaalverbid  
ahelverbides, finiitne vorm

@ICV – olema liitaegades ning modaalverbid ahelverbides,  
infiniitne vorm

@NEG – verbi eitus

# järg

## ▶ lauseliikmed

@SUBJ – alus ehk subjekt

@OBJ – sihitis ehk objekt

@PRD – öeldistäide ehk predikatiiv

@ADV L – määrus ehk adverbiaal, sh fraasiadverbiaal

Nt:

## lauseliikmete järjend

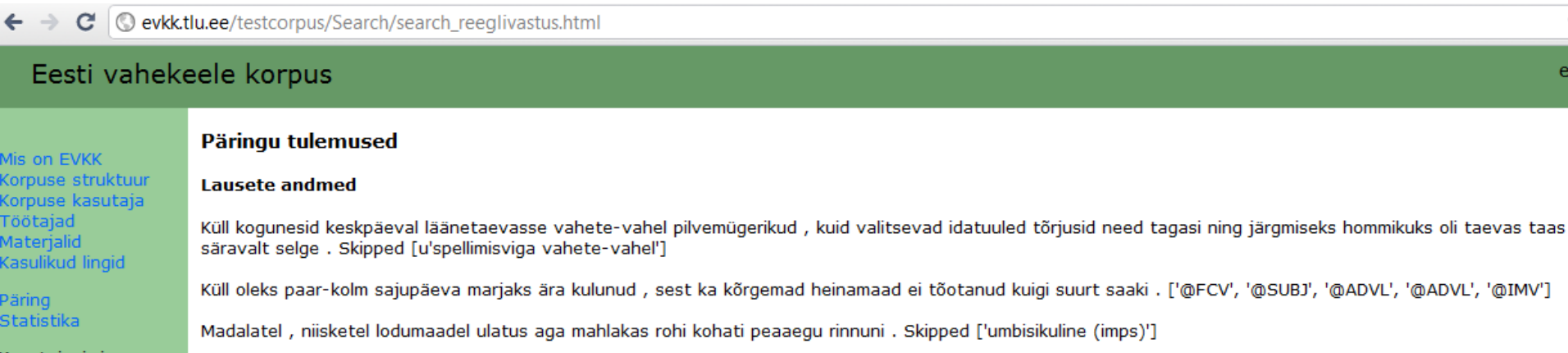
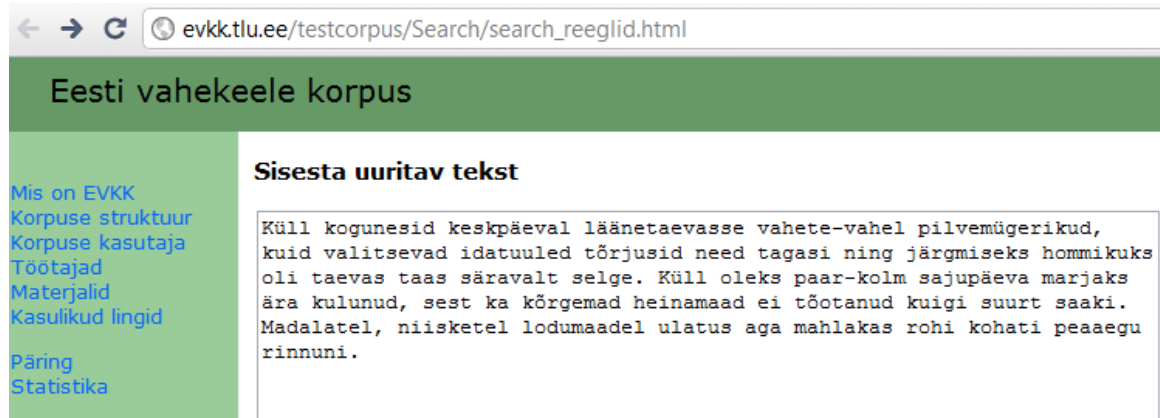
Internetis (@ADV L) on (@FMV) võimalik (@PRD) kasutada (@SUBJ) mitmeid (@NN>) teenuseid (@OBJ)

**sõnajärjemall** ['@ADV L', '@FMV', '@PRD', '@SUBJ', '@OBJ']



# Andmepuud: süntaktilised märgendid ja sõnajärjemustrid

- ▶ Analüüsitud 20000 lauset ilukirjanduse korpusest
- ▶ Andmepuud genereeriti rohkem kui 10000 lause alusel (prototüüp jättis vahele 8590 lauset)



# Näide 2: eitusega algava andmepuu sõnajärjemustrid (kõige harvem)

2

['@NEG', '@FMV', '@SUBJ', '@ADVL']

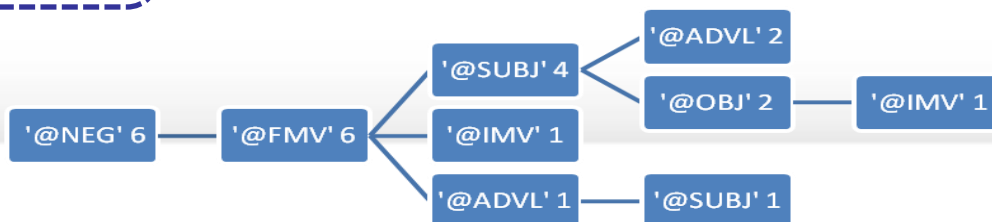
['@NEG', '@FMV', '@SUBJ', '@ADVL']

['@NEG', '@FMV', '@SUBJ', '@OBJ', '@IMV']

['@NEG', '@FMV', '@SUBJ', '@OBJ']

['@NEG', '@FMV', '@IMV']

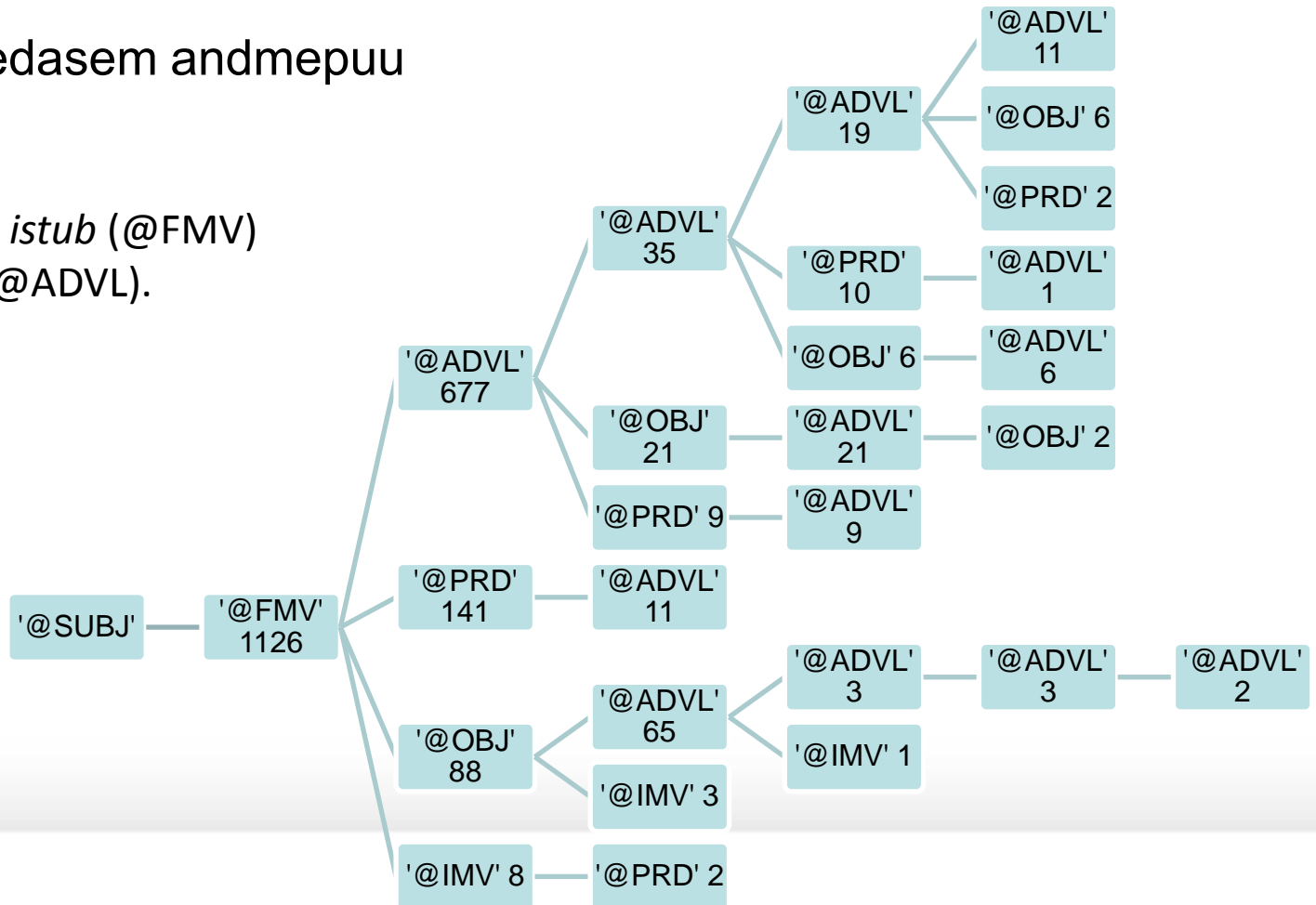
['@NEG', '@FMV', '@ADVL', '@SUBJ']



# Näide 3: subjektiga algava andmepuu sõnajärjemustrid

- ▶ Kõige sagedasem andmepuu

*Mu mõrsja (@SUBJ) istub (@FMV)  
salongi laua ääres (@ADVL).*



# Prototüübi testimistulemused

## EVKK B-taseme tekstidest juhuvaliku alusel 300 lauset

suhtarvuliselt on õigeks hinnatud, veakahtlustusega ja analüüsist väljajäetavate lausete hulk sama

- 300st lausest peab prototüüp vigaseks 143 lause sõnajärge, korrektseks 72 ja väljajätmisele 85 lauset
- lingvisti hinnangul vastavalt 146, 75 ja 79
- prototüübi töö ja lingvisti hinnangu kokkulangevus on 87,82%



# Tulemused 2

- ▶ EVKK funktsionaalsuste laiendamine
  - ▶ Korpusel on tähestikuline ja üldine sagedussõnastik
  - ▶ Korpuse sõnastiku jaoks on loodud eesti õppijakeele lemmatiseerija-oletaja (veakindel lemmatiseerija), mida testiti 60000-sõnase korpusvalimi alusel
    - ▶ õppijakeele vormimoodustus- ja ortograafiavigade analüüs
    - ▶ õppijakeele sõnavara ja morfoloogiat iseloomustavate joonte automaatne esiletoomine – annab teavet õppija keeleoskustasemest



# Lemmatiseerija-oletaja süsteemi komponendid

- ▶ On ligikaudne õigekirjakorrektor - baseerub stringide teisenduskaugusel ja foneetilisel algoritmil Metaphone
- ▶ Aluseks ESTMORF (lemmade leidmiseks)
- ▶ Deterministlikud valikud otsustuspuude abil kirjavahemärkide, mittesõnade, pärisnimede ja ühetähenduslike sõnade jaoks
- ▶ Tõenäosuste arvutamine mitmesuste lahendamiseks



# Ligikaudne õigekirjakorrektor

- ▶ Referentssõnastik koostati EE korpuse baasil (7,2 mlj)
- ▶ Referentssõnastik eeltöödeldi foneetilise algoritmiga
- ▶ Moodustuvad hulgad sarnaselt häälduvatest sõnadest
- ▶ Vigasele sõnale leitakse tema häälduskuju
- ▶ Arvutatakse teisenduskaugused vigase sõna ja sama häälduskujuga sõnade vahel referentssõnastikus
- ▶ Võimalikud kandidaadid leitakse teisenduskauguse arvutamise abil

# Tulemused 2

Andmed:

- ▶ 5495 sõna õppijakeele tekste (koos kirjavahemärkidega)
- ▶ Arvutame õigete lemma-sõnatüübi paaride protsendi
- ▶ Võrdleme Filosofti ühestaja tulemusega



# Tulemused 2: järg

	Oletaja- lemmatiseerija	Filosoft
Kokku	5495	5495
Õigeid	5173	4179
Õigeid %	94,1%	76,1%

O-L paremus Filosofti ees 23,8%

# Tulemused 2: järg

- ▶ Tulemused kirjavahemärke arvestamata

Kokku	4636	4636
Õigeid	4314	4179
Õigeid %	93,1%	90,1%

**O-L paremus Filosoofi ees 3,2%**

# Näiteid 1

- ▶ O-L-I vale tulemus, ESTMORF-il õige  
<elus elus A> vs <elus elu S>  
<ajal ajal K> vs <ajal aeg S>
- ▶ Süstemaatiline lahknevus sõnaliikide osas ESTMORFi ja lingvisti vahel: D või K vs X  
üumber, ära, kaasa, tagasi jne
- ▶ Pärinimede lemmatiseerimine  
*Gerda ja Kai* tundis ära, *Lumekuningannat* ja *Kakukest* kui pärisnimesid mitte

# Näiteid 2

## Kirjavigadega sõnade äratundmine

Rahvusvahelises            rahvusvaheline

Pankooke                    Pannkook

Praaktikana                Praktika

Elmise                      Eelmine

Praagu                      Praegu

Disainega                  Disain

Rejsijad                    Reisija

Pillede [pilvede]

Lill [pilv]

Hobbiga [hobiga] Hoop [hobi]

Arä [ära]

Aru [ära]

Kõiged [kõik] Kõige [kõik]

# Tulemused 3

- ▶ Kolmas eesmärk – EVKK kasutajaliidese täiustamine, võimaluse loomine uute alamkorpuste ja allkorpuste tekitamiseks
  - ▶ Töö käigus selgus, et korpuse tekstid on vaja üle vaadata ja puhastada: liigsed tühikud, skanneeritud tekstide puhul sümbolite asendamine õigete tähtedega, eksikombel sisse jäänud isikunimede äramuutmine, tekstide kordumine jm
  - ▶ kõik veamärgendust mittesisaldavad tekstid on läbi vaadatud ja parandatud
  - ▶ märgendatud tekstide töötlemist võimaldab **uus märgendusliides**



# Uus vealiigi märgendusmoodul

- ▶ vealiigi märgenduse on üle 500000 sõne suurusel korpuse osal teinud üks märgendaja
- ▶ teine, kolmas jne märgendaja peavad saama lisada uusi märgendeid, parandama teiste märgendajate tehtud tööd ja teksti sisestamisel tehtud apse, tühikuid jm
  - ▶ tulemus: mitmemõõtmeline veamäärang
    - ▶ oluline eesti keele kui teise keele omandamise aspektist: õpiraskuste lingvistiline sisu
    - ▶ pedagoogiline rakendus
    - ▶ õpikute ja õppematerjalide koostamiseks oluline materjal



# Edasised tegevused

- ▶ oletaja-lemmatiseerija tulemuslikkus; programmi integreerimine õppijakeele korpusesse ja veebiliidese abil kättesaadavaks tegemine
- ▶ oletaja-lemmatiseerija integreerimine sõnajärje vealeidjaga, mille tulemusel prototüüp analüüsib kõiki õigekirja- ja vormivigu sisaldavaid lauseid ja avarduvad oluliselt prototüübi rakendusvõimalused



# Edasised tegevused (järg)

- ▶ nt aitab tuvastada õppijakeele morfoloogia- ja sõnajärjevigu, määrab kindlaks keeleoskustasemeid iseloomustavad lingvistilised mustrid
- ▶ Uue veamärgendusmooduli arendamine keeleõppija kirjutatud tekstide automaatanalüüsi vahendiks
  - ▶ veebiliidest saab kasutada nii keeleõppija vealiikide esiletoomiseks (pedagoogiline aspekt) kui ka õigekirjakorrektorina (laiem kasutusvaldkond)
  - ▶ laiendada korpuse kasutajaliidese funktsionaalsusi uutele allkorpustele



# Edasised tegevused (järg)

- ▶ EVKK 500000 sõnest koosnev tuumkorpus (eesti keel kui K2)
  - ▶ EVKK allkorpused (eesti keel kui võõrkeel)
- ▶ REKK-i tekstikogud
- ▶ Akadeemilise eesti õppijakeele korpused
  - ▶ K1 ja K2