Oletaja-lemmatiseerija õppijakeele korpuse jaoks

Kairit Sirts
Tallinna Tehnikaülikool
EVKK sügisseminar 27.10.2010

Kava

- Milles seisneb probleem? Miks on vaja lemmatiseerida ja miks tavaline lemmatiseerija ei kõlba?
- Oletaja-lemmatiseerija kirjeldus
- Eksperdimendi tulemused
- Plaanid tulevikuks

Ülesanne (lingvisti vaatenurgast)

- Moodustada õppijakeele korpuse baasil õppijakeele sõnastik
  - Võimaldaks teha veaanalüüsi õppijakeeles kasutatavate vigaste vormide kohta
  - Võimaldaks automaatselt õppija sõnavara alusel kindlaks teha tema keeleoskuse taseme

Õppijakeele sõnastik

- Lemmatiseerime kõik sõnad
- Paneme kõik sõna-lemma paarid sõnastikku koos esinemissagedustega
- Iga sõna on seotud tema lemmaga
- Iga lemma on seotud kõikide oma sõnavormidega kõikides tekstides

hakkanud	7		
hakkas	345	hakka	hakkama
hakkasid	291		
hakkasime	8	hakka	hakkama
hakkasin	60	hakka	hakkama
hakkasingi	2	hakka	hakkama
hakkata	5		
hakkatakse	17		
hakkatas	1		
hakkate	9	hakka	hakkama
hakkati	69	ati	hakk+att
hakkava	1		
hakkavad	113		
hakkida	1		
hakkkasid	3		
hakkliha	3		
hakklihamasin	6	masin	hakk+liha+masin
hakklihapallid	2	palli	hakk+liha+pall
hakksid	3		
halastamatu	4		
halb	102	halb	halb
halba	30		
halbad	2		
halbade	3	halba	halb
halbadest	2	halba	halb

hakkama

Freq: 160 / 756651

hakkab hakkan hakkasin hakkas hakkama hakatakse hakati hakkama18 hakkad hakatigi

Dokumendid

VENELASTE VEAD EESTI INFINITIIVI VALIKUL (18)
Eestlaste laused (4)
Milleta tänapäeva inimene elada ei saaks (eestlaste laused) (4)
Essee (3)
Kontrolltöö (2)
Kontrolltöö (2)
Analüüs (2)
EKSAMITÖÖ (2)
Eksamitöö (2)
Essee (2)
Analüüs (2)
Kontrolltöö (2)
Eksamitöö (2)
Referaat (2)
Analüüs (2)
EKSAMITÖÖ (2)
Kontrolltöö (2)
veaanaliiusi kodutoo (2)
EKSAMITÖÖ (2)
EKSAMITÖÖ (2)
Analüüs (2)
Milleta tänapäeval elada ei saaks (eestlaste laused) (2)

Automaatne lemmatiseerimine

- Automaatne lemmatiseerimine eesti keeles on põhimõtteliselt lihtne
- ESTMORF on olemas juba üle kümne aasta

- Samas, õppijakeele tekstid on vigased
- ESTMORF ei suuda analüüsida vigaselt kirjutatud sõnu

Automaatne õigekirjakorrektor

- Eesti keele jaoks olemas vaid tekstiredaktorites kasutatavad variandid
- Me vajame automaatset õigekirjakorrektorit (mitte interaktiivset, mis pakub kandidaate, mille hulgast kasutaja peab ise valima)

Süsteemi komponendid

- Ligikaudne õigekirjakorrektor
- Morfoloogiline analüsaator ESTMORF

- Deterministlikud valikud otsustuspuude abil kirjavahemärkide, mittesõnade, pärisnimede ja ühetähenduslike sõnade jaoks
- Tõenäosuste arvutamine mitmesuste lahendamiseks

Stringide teisenduskaugus

- Kaugus võrdne operatsioonide arvuga, mis on vajalik ühe stringi teiseks teisendamiseks
- Operatsioonid: tähe lisamine, kustutamine, asendamine

Näiteks:

kantsid vs kandsid	→ ***t*** asendada ***d***-ga	→ kaugus = 1
igasugulased vs igasugused	→ kustutada ***l*** ja ***a***	→ kaugus = 2

„Sõnastik“ õigekirjakorrektori jaoks

- Eesti Ekspressi artiklitest koosnev korpus
- Ca 7 miljonit jooksvat sõna
- Eeldame, et võrreldes õppijakeeletekstidega on see korpus praktiliselt ilma vigadeta

Foneetiline algoritm Metaphone

- Moodustab igale sõnale tema kuju, mis sisaldab ainult kõige olulisemat häädusest
- Sarnaselt kõlavatel sõnadel ühesugune hääldusku ju
- Algoritm on inglise keele spetsiifiline
- Lisasime juurde eesti keele vokaalid Õ, Ä, Ö ja Ü

Ligikaudne õigekirjakorrektor

- „Sõnastik“ eeltöödeldakse foneetilise algoritmiga
- Moodustuvad hulgad sarnaselt häälduvatest sõnadest
- Vigasele sõnale leitakse tema häälduskuju
- Arvutatakse teisenduskaugused vigase sõna ja sama häälduskujuga „sõnastiku“ sõnade vahel
- Leitakse suurima tõenäosusega kandidaat nende sõnade hulgast, mille teisenduskaugus on maksimaalselt 2.

Tulemused

- Arvutame õigete lemma-sõnatüübi paaride protsendi
- Võrdleme Filosofti ühestaja tulemusega

Andmed:

- 5495 sõna õppijakeele tekste (koos kirjavahemärkidega)
- Skanneeritud automaatselt, käsitsi üle kontrollitud

Tulemused (2)

	O-L	Filosoft
Kokku	5495	5495
Õigeid	5173	4179
Õigeid %	94,1%	76,1%

O-L paremus Filosofti ees 23,8%

Tulemused (3) - ilma kirjavahemärkideta

	O-L	Filosoft
Kokku	4636	4636
Õigeid	4314	4179
Õigeid %	93,1%	90,1%

O-L paremus Filosofti ees 3,2%

Mõned tähelepanekud

- O-L-l vale tulemus, ESTMORF-il õige

  <elus elus A> vs <elus elu S>

  <ajal ajal K> vs <ajal aeg S>

- Süstemaatiline lahknevus sõnatüüpide osas ESTMORFi ja lingvisti vahel: D või K vs X

  ümber, ära, kaasa, tagasi jne

- Pärisnimede lemmatiseerimine

  Gerda ja Kai tundis ära, Lumekuningannat ja Kakukest mitte.

Kirjavigadega sõnad

Sõna	**Lemma**
Rahvusvahelinses	rahvusvaheline
Pankooke	Pannkook
Praaktikana	Praktika
Elmise	Eelmine
Praagu	Praegu
Disainega	Disain
Rejsijad	Reisija
Lõppetades	Lõpetama
Pikkal	Pikk
Interned	Internet
Prammiga	tramm

Sõna	**Lemma**
Pillede [pilvede]	Lill [pilv]
Arä [ära]	Aru [ära]
Suurestest [suurematest	Suurus [suur]
Unustumata [unustumatu]	Unustama [unustumatu]
Laendab [laiendab]	Lendab [laiendama]
Hobbiga [hobiga]	Hoop [hobi]
Muutakse [muutub]	Murdma [muutuma]
Praigu [praegu]	Praigu [pärisnimena]
Kõiged [kõik]	Kõige [kõik]

Edasine tegevus

- Integreerida oletaja-lemmatiseerija õppijakeele korpusesse, nii et see oleks kättesaadav veebiliidese abil
- Pärissõna vigade parandamine
- Grammatiliste kategooriate vigade parandamine

Lõpetuseks

- Oletaja-lemmatiseerija, mis koosneb:
  - Morfoloogilisest analüsaatorist ESTMORF
  - Ligikaudsest õigekirjakorrektorist
- Lemmatiseerib ca 24% võrra täpsemalt kui ESTMORF, kui arvestada ka kirjavahemärke
- Lemmatiseerib rohkem kui 3% võrra täpsemalt kui ESTMORF kirjavahemärke arvestamata
- Oletaja-lemmatiseerija saab tulevikus olema kättesaadav EVKK veebikeskkonna kaudu

KÜSIMUSED?

Selle töö valmimist on rahastanud projekt „VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine“