Keeletehnoloog teoretiseerib korpuse üle

Heiki-Jaan Kaalep

Tartu Ülikool

Kava

1. Sissejuhatus: keel ja korpus
2. Isiklik ajalugu
   1. Programmeerija ... testija ... (korpus)lingvist

      (Keele)tehnoloogia on teooria rakendamine ja kontroll .... Kas ikka on?

      Probleemid keele kirjeldustes → vigane programm
   2. Korpus kui näidete või reeglite allikas
      1. NB! Reeglid olgu üldised, s.t. korpusest sõltumatud
      2. Alati ei õnnestu reegleid koostada → korpusepõhised loendid...
3. Andmete valimise kunst
4. Millist korpust saab usaldada ja mille osas?

Sissejuhatus: keel ja korpus

Korpus = andmed

Keel = reeglid e. algoritm

Keele kasutamine annab korpuse = Reeglite rakendamine annab andmed

Paradoks: reeglid pole (lingvistile) teada

Räägime, aga lingvistid ei tea, kuidas laps selle õppis

(Kui teaks, siis võiks ka arvuti rääkida/kirjutada)

Lingvisti ülesanne: leia reeglid andmete kaudu

... nad on seda alati proovinud teha...

On reeglid üldse leitavad?

Morfoloogia teooria = algoritm

1. Mida on üldse vaja moodustada?
   1. Millised sõnavormid on olemas ja mida nad grammatiliselt tähendavad?
   2. Paradigmatabeli kuju
2. Millistest osadest (morfeemidest) kokku pannakse (morfotaktika)?

   morfeemi sisu (gr. kat.) ja vorm (morf)

   nt mitmus – de/te/e/u/i

   minevik – nu/s/is/i
3. Milliste reeglitega kokku pannakse?

   M.h. mis on algvorm?
4. Milliste reeglitega morfe muudetakse (morfofonoloogia)

   astmevaheldus, vokaalide madaldumine, ...

Näited lingvistide algoritmilisest argumentatsioonist

“Tugevast astmest nõrka moodustada on kergem” (Ahrens, Wiedemann)

“...peab näitama, missugusest tüvevariandist missugune vorm moodustatakse” (Peebo)

“Ei ole võimalik ühegi algvormi puhul esitada sünteesireegleid, mille abil saaks teisi tüvevariante genereerida” (Ehala)

Algoritmi kirjeldusest programmiks e. vabamorfi tegemise etapid

1. Lihtsõnade analüüs – VVS

   Sõnavara katvus, morfoloogiline kirjeldus OK

2. Tuletised, liitsõnad

   Reeglid kirjeldatud jutuna (R. Kasik, R. Kull) + vt ka korpusi (näiteid; kriteerium arendamise lõpetamiseks – kui tundmatuid liitsõnu oli sama palju kui kirjavigu)

3. Oletamine (minu jaoks uus mõte)
   1. Ideed (Ü. Viks, K. Ross) + vt korpusi statistiliselt
   2. Sõnasagedusstatistika (hapax)

Vabamorf kui teooria ?

Sõnastikupõhine analüüs/süntees on ka teooria?

Halb (?) eeskuju ÕS-idest

Kas paralleelvormid on või pole?

seppadel/**sepil*, *Pariisi*/**Pariisisse*

“Teoreetiliselt võimalikud” sõnavormid

muuseum+*e*/*eid*/**isid*

Võimalikud ja tegelikud sõnad

hila, *ask*, *tära* (ÕSis kuni 2006)

→ tavaline leksikograafi töö (sõnastiku korrastamine)

üksiksõnade eemald/lisam

muuttüüpide modif. (**sepil*)

Õige programm, valed tulemused

"Katseklaasis ujub sirbikujulisi *<u>värde</u>*"

(VVS, ÕS: *värd*, *värra*, *värda* mitm. os. *värdi* e. *värdasid*)

→ avastasin uue muuttüübi?

käppe, *jalge*, ... ?

Ei... *värd*, *värdi*, *värdi*, mitm. os. *värde*

(*äpp*, *laik*, ...)

Vabamorf kui teooria: oletaja kui keeletunde imiteerija

Mistahes sõna puhul me teame, kuidas selle vorme moodustada; ka tundmatu sõna puhul

Oletamine: milline võib olla sõnavormi algvorm?

peele = ~~*peel*~~ ains. om. ← keel

peel mitm. os. ← steel

peele ains. nim/om ← Adeele

Keeletaju

Käänamis- ja pööramisviisi peab määrama sõna väline kuju (sest muule pole tugineda)

- haruldase ja väljamõeldud sõna muutmine on ühesugused
- see käänamine tundub loomulik, tavapärane (teistsugune oleks naljakas, nt *äpp:äpu*, *äpp:äpa*)
- kui ühesug. kujuga sõnadel on erinev muutmisviis, siis produktiivne on see viis, mis juba kehtib suuremale hulgale sõnadele

Andmete valimise kunst

Kuidas saavutada, et Vabamorfi oletaja imiteeriks keeletunnet e. pakuks just produktiivseid malle?

Mina oma keeletunnet lõpuni ei usalda

Kas keele kirjeldusi võib usaldada?

Vaata korpuse **sõnavara !**

Keeletaju (1 silp, lõpus kaash.)
näpp, rüpp, käpp, täpp

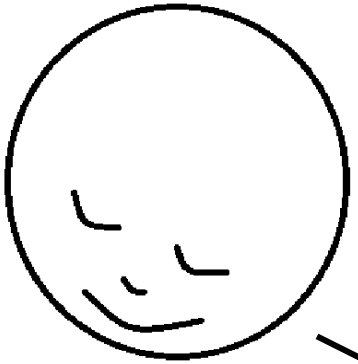
Varem nähtud tüvevokaal

Tekstis

+u	15%
+e	5%
+a	30%
+i	50%

Sõnavaras

+u	9%
+e	1%
+a	5%
+i	85%



Tundmatu sõna tüvevokaal

+u	**0%**
+e	**0%**
+a	**0%**
+i	**100%**

Keeletaju (2 silpi, II välde, lõpus *s*)

kinnas, pinnas, soodus, boonus

Varem nähtud mall

Tekstis

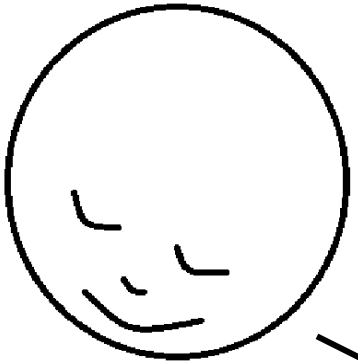
on a-v 50%

pole a-v 50%

Sõnavaras

on a-v 34%

pole a-v 66%



Tundmatu sõna mall

on a-v **0%**

pole a-v **100%**

Vahekokkuvõte

Vaata sõnavara, võrdle loendeid, ignoreeri sõna kordumisi tekstis !

Vabamorfi kood: inseneri või teadlase vaade

Loendid (lõpud, liited, tüved, erandid, ...) + kombineerimisreeglid

Näited

1. Loend, mis kajastab reeglit

   Produkt. muutmismallide loend oletamiseks:

   jakk, *paak*, *punk*, ...

2. Loend, mis kajastab lihtsalt korpust

   Ei sobi oletamisel liitsõna viimaseks komponendiks:

   biit, *bool*, *boor*, *emis*, *enne*, *foon*, ...

   Oletused: *valgus_foon*, *blama_meister*, *blamafoon*, **blama_foon*

Mida korpus esindab?

Korpus = hulk tekste

valdkond + keelekasutus

Korpus <--> “**üldkogum**”, “üleüldine keelekasutus”

sarnane / esinduslik / representatiivne

mudel <--> originaal

Ülesanne: leia parameetrid, mis korpuses ja “üldkogumis” on ühesugused

Näide: gloobus <--> Maa : vahekaugus

Tartu – Kiiev, Tartu – Istanbul

absoluutsuurus erineb, aga <u>suhe</u> sama

Kas korpus esindab

- keelekasutust, mis on keele (all)süsteemi kujunemise/muutumisega kooskõlas?

  → sel juhul on ta üldkogumiga sarnane
- kõigest keelekasutuse näiteid?

  → ikka kasulik, aga ...

Parameetri jahil

Kooskõla

inimese (1) võimete,

keeleelemendi keerukuse (≈**pikkuse**) (2)

ja **sageduse** (3) vahel

(1) omandatakse (laste poolt)	varem	hiljem
(2) kuju	lühem	pikem
(3) kordub tekstis	rohkem	vähem

Sõnamuutus: süsteemi ja korpuse kooskõla

- Sõna muutevormid on erineva kasutussagedusega
- Sagedam vorm tuleb varem ette kui harvem
- Üldiselt: paradigma liikmed on erineva kasutussagedusega

  nim > om; ains > mitmus
- Sagedam liige tuleb varem ette kui harvem liige
- Ainult varem teadaolevast saab moodustada uue vormi

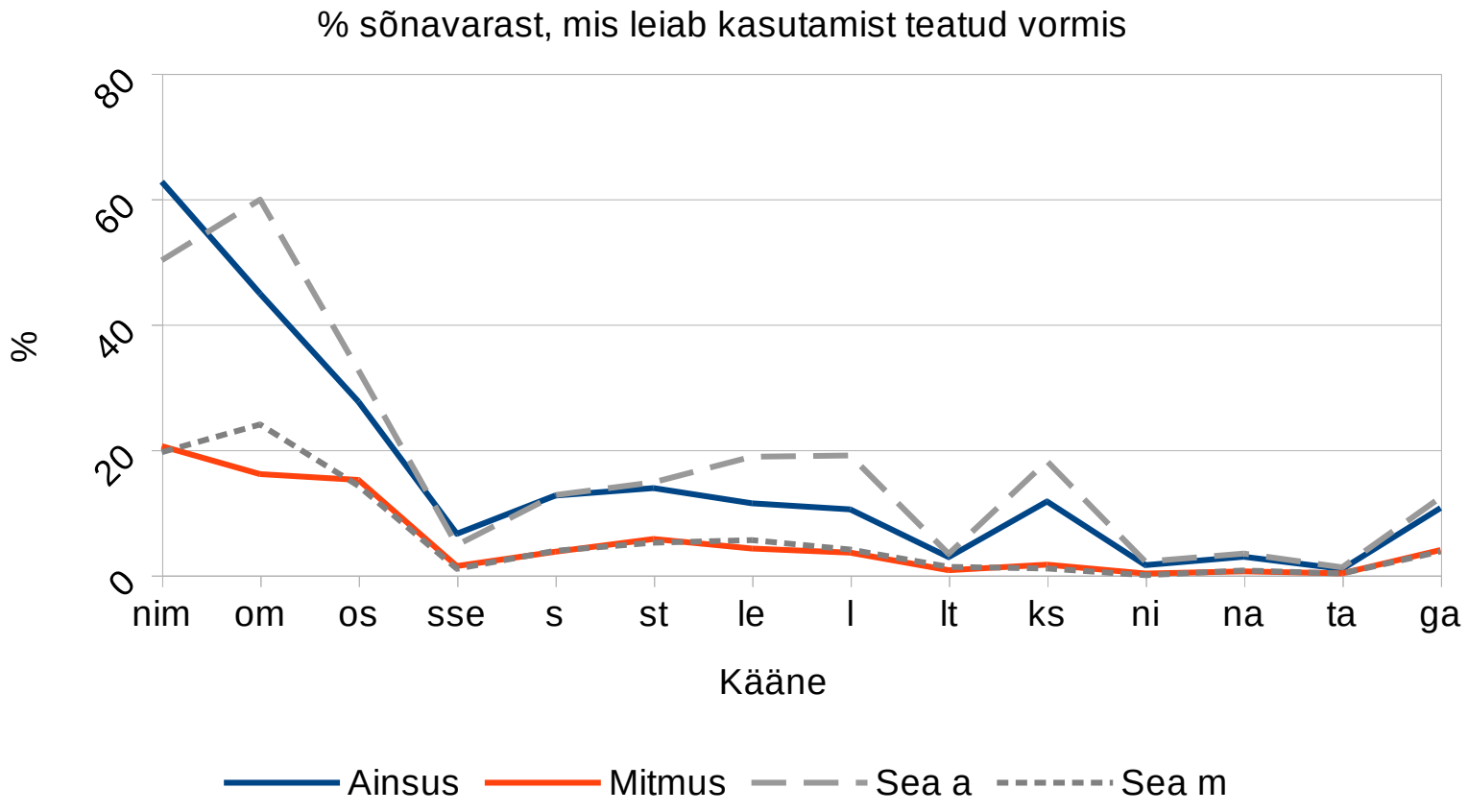
⇒ muuteparadigmadel on vormide hierarhia (Bybee)

... mis on üldkogumiga kooskõlas

Käändsõna paradigma

	ainsus	**mitmus**
nimetav	*redel*	*redelid*
omastav	*redeli*	*redelite*
osastav	*redelit*	*redeleid*
sisseütlev	*redelisse*	*redelitesse*
seesütlev	*redelis*	*redelites*
...	*...*	*...*
ilmaütlev	*redelita*	*redeliteta*

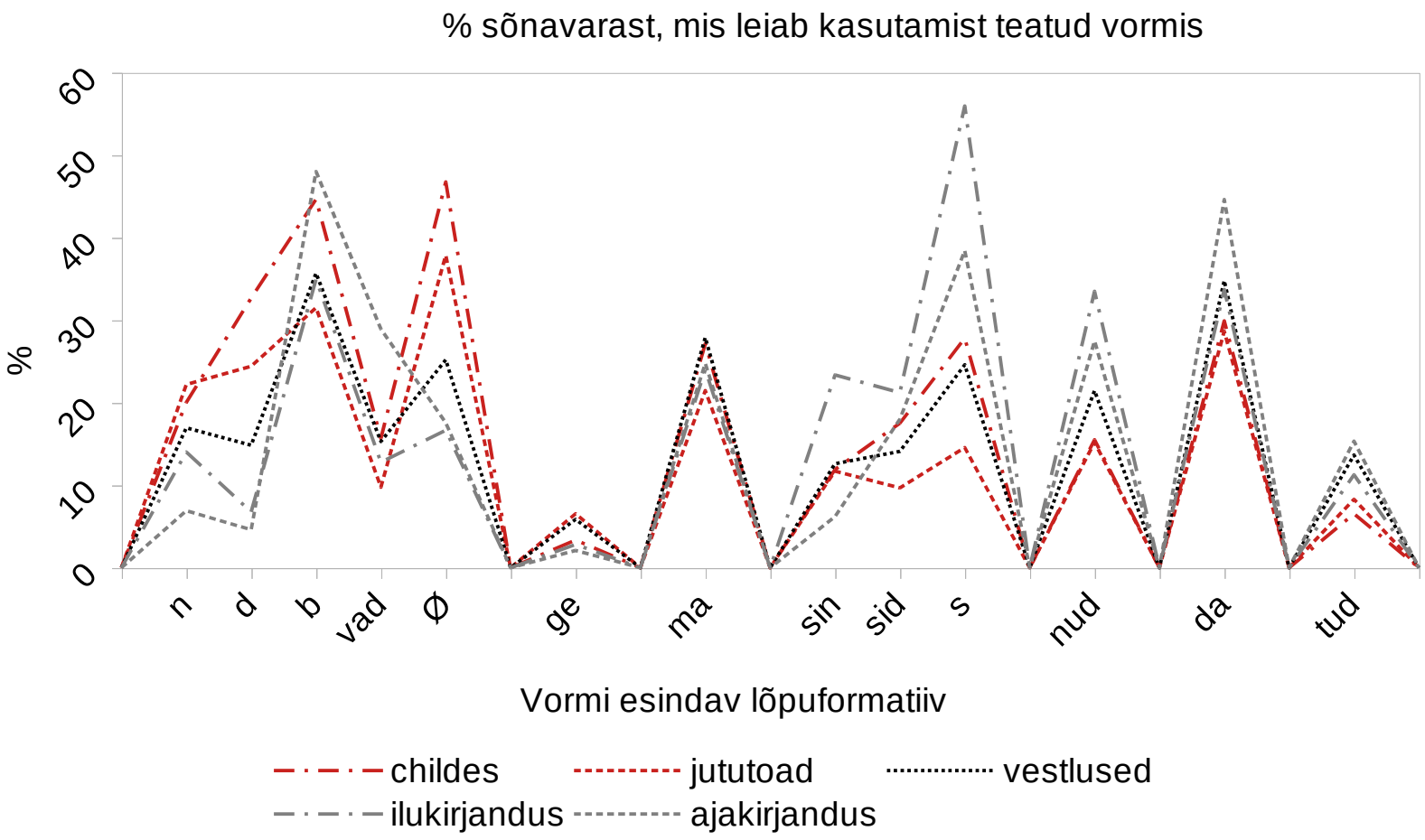
Käändsõna paradigmaliikme tavalisus



Tegusõna paradigma: analoogiarühmad

Ø	hüppa, -n, -d, -b, ... -ksin, -ks	nuta, -n, -d, -b, ... -ksin, -ks
GE	hüpa-ke, -kem, -ku	nut-ke, -kem, -ku
MA	hüppa-ma, -mas, ... -va	nut-ma, -mas, ... -vat
S	hüppa-sin, -sid, ...	nut-sin, -sid, ...
NUD	hüpa-nud, -nuksin, -nuks, ... -nuvat	nut-nud, -nuksin, -nuks, ... -nuvat
DA	hüpa-ta, -tes	nutt-a, -es
TUD	hüpa-tud, -ti, -takse, ...	nute-tud, -ti, -takse, ...

Tegusõna paradigmaliikme tavalisus



Verbi formatiivid korpustes: kokkuvõte

- Eri korpused, eri sagedusjärjestused
- Sagedus ei ole alati kooskõlas formatiivi lühidusega
- Erinevus ei tulene “suulisusest”
- Lõputa vorm sagedaim **hoidjakeeles** ja **jututubades**

  => **õiges korpuses** ei ole lood, narratiivid, vaid koos tegutsemisega seotud jutt?

  vrd majaehitus (*anna haamer*, *võta ise*, *nii ei saa*, *ära siia astu*)

Lõputa vorm: (1) omandatakse kõige varem, (2) on lühim, (3) on **õiges korpuses** sagedaim

Korpuse sobivus

Paljud korpused ei sobigi sõnamuutmissüsteemi uurimiseks !?!

Kui sõnamuutmissüsteem kujuneks nende järgi, siis oleks ta teistsugune kui see, mida me näeme ta olevat.

... Aga mõne teise aspekti uurimiseks ?

Muutuse sammud “vales” korpuses EtTenTen

naase tõuse → ela

veena, mööna naera → pööra

Ø	DA	S	MA	NUD	TUD	GE
naase 200	naasta naaseda 2000/100	naasis naases **1200/500**	naasma naasema **400/40**	naasnud naasenud 1500/20	naastud naasetud 200/0	naaske naasege 10/0
veena 500	veenda veenata 3500/1	veenis veenas **500/20**	veenma veenama **500/1**	veennud veenanud 300/1	veendud veenatud 100/0	veenge veenake 20/0
mööna 20	mööndа möönata 300/0	möönis möönas **1800/40**	möönma möönama **400/1**	möönnud möönanud 100/1	mööndud möönatud 10/0	möönge möönake 1/0

Mõte: korpus ja tiraaž

Korpus esindab keelesüsteemi toimimise tagajärge, tiraaži korpuse tegemisel ei arvestata

Aga kui korpus on keele süsteemi kujunemise aluseks, siis tiraaž on oluline

Lõpetuseks

Alustasin lihtsameelse programmeerijana, nüüdseks olen vist lingvistiks muutunud ja tunnen muret, et kust keeleuurija adekvaatseid andmeid saaks.

Alguses uskusin, mida keeleraamatud kirjutasid, nüüd ei usu alati enam korpustki.

Pean ütlema, et lingvist olla on huvitavam kui programmeerija.