Arvutimorfoloogia 2022: eesti morfoloogia kirjeldus lõplike muundurite abil

Isiklik vaade

1. VVS + insener → estmorf, vabamorf
2. estmorf + korpused → oletamine
3. oletamine, paralleelvormid → produktiivsus
4. produktiivsus → teooria, K&K artiklid
5. teooria + insener → formaalne kontroll

Morfoloogia – sõnade muutmine

Näide:

põtrades ->

põder; nimisõna, mitmuse seesütlev (sisu: kategooriad)

põtra+de+s (vorm: morfid)

Milleks morfoloogiat vaja?

idee -> ...->

leksikaalne üksus + grammatiline funktsioon -> sõnavorm ->

... -> valmis lause

Lingvisti probleemid

- Paradigma on sõnavormina väljenduvate grammatiliste kategooriate komplektide korrastatud hulk
  - Millised on kategooriad?
  - Kuidas kategooriad väljenduvad?
  - Kombinatoorika e. kes kellega käib (nii kategooriate kui morfide osas) ja käis?
  - Mis süsteemi üleval hoiab ja mis muudab?

Eesti traditsioon vs arvutianalüüs

Morfoloogia = vormiõpetus e. vormimoodustus e. sõnamuutmine, s.t. pööramine ja käänamine

Tuletamine (maja+ke) ja liitmine (elu+maja) on hoopis sõnamoodustus (EKG I 1995)

Arvutianalüüsis:

Morfoloogia = sõnamuutmine + sõnamoodustus

Keele ajalugu ja arvutimorfoloogia

Kas „No history lessons! No bullshit!“ (nagu ütles NATO läbirääkija R. Holbrooke Bosnia ja Hertsegoviina tuleviku üle vaidlevate poolte korralekutsumiseks)?

Ajalooliselt eri perioodidel keelde lisandunud samakujulised sõnad võivad muutuda erinevalt, nt:

- susi soe sutt
- kusi kuse kust
- musi musi musi
- Uzi Uzi Uzit

(kuid keele ühtlustumistendents suunab neid muutuma ühesuguselt, nt https://kodu.ut.ee/~hkaalep/arvutimorf_16/sonade_kaanamine_ajas.mp4

ommatiidis

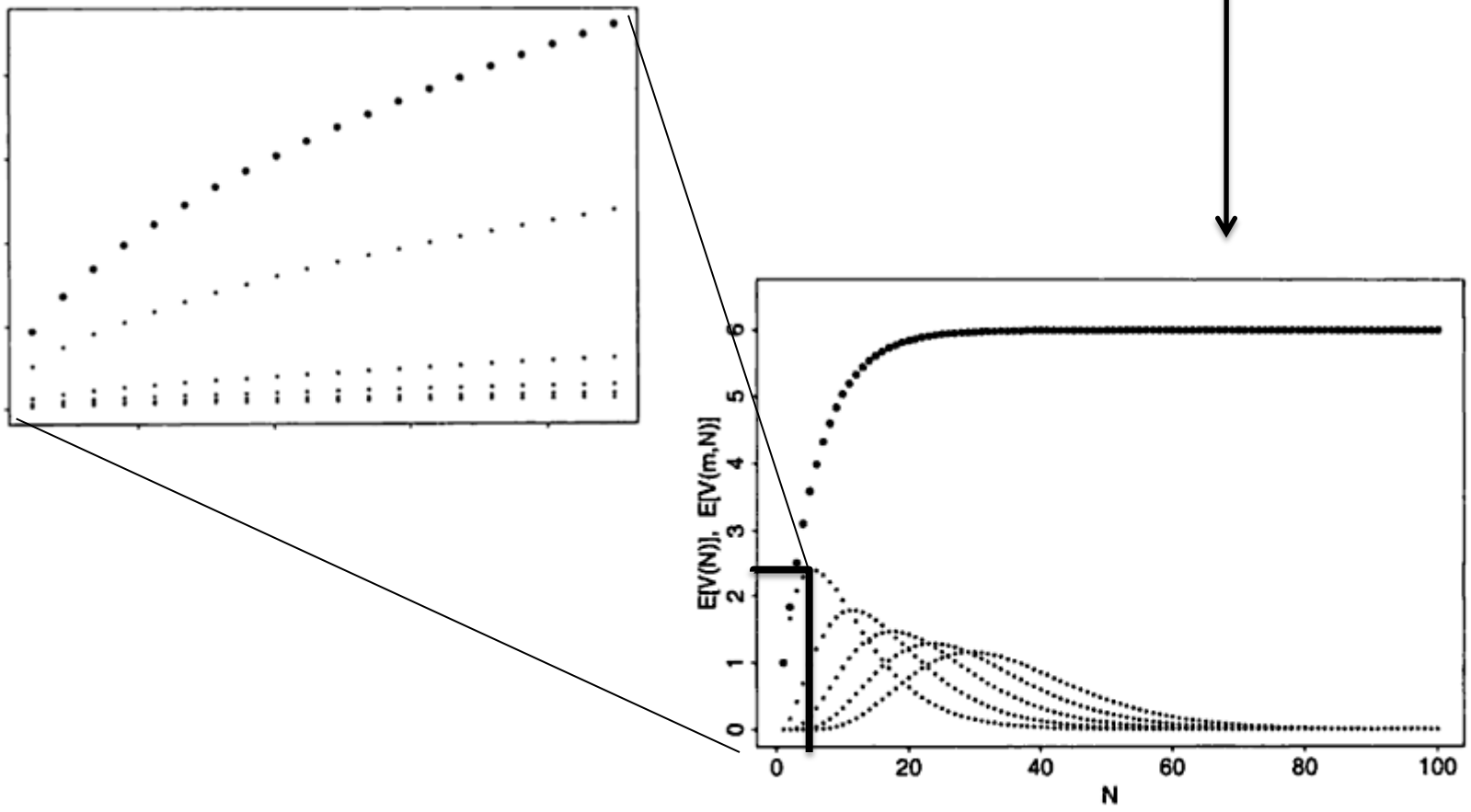
- (mis on selle sõna algvorm? kääne? pööre?)

- etTenTen www.keeleveeb.ee

Sõnavara ja LNRE

LNRE = Large Number of Rare Events

Erinevate sõnade arv: keel vs täring (H. Baayeni järgi)

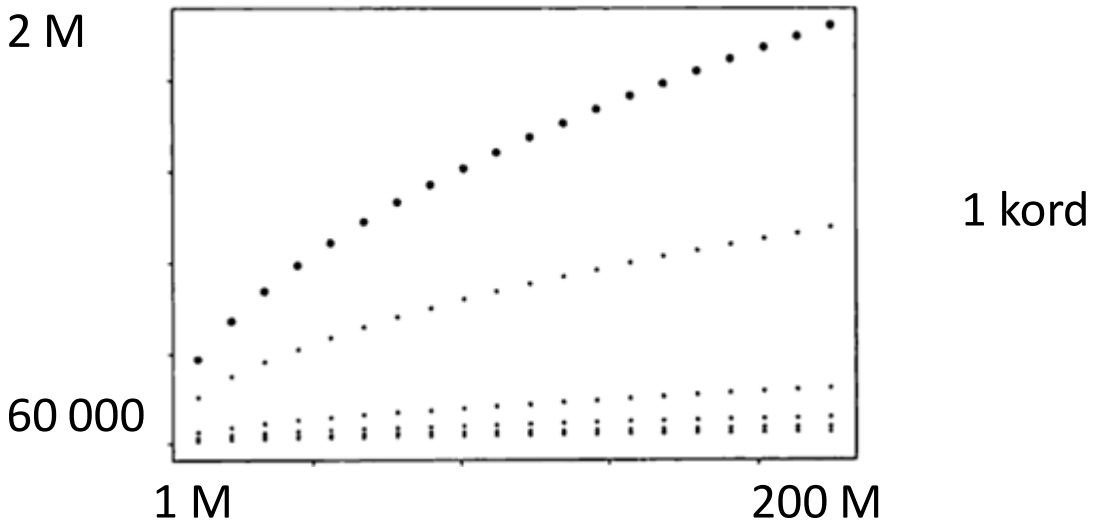


Kui palju on eesti keeles sõnu?

Just nii palju kui vaja!

1 M sõna teksti - vaja 60 000 sõna (pooled 1 kord)

200 M - vaja 2 M sõna (pooled 1 kord)



Sõnastikust pole abi

- kõigi sõnade käänamist pole võimalik ära õppida

  TÜ koondkorpuses on 100 000 erinevat lihtsõna, ÕSis ja EKSSis 2 korda vähem (*jahta*, *jorss*)

- uusi sõnu oskavad kõik automaatselt käänata

  äpp, *meem*, *võõruk*

- ... ja neid käänatakse ühte moodi, „tunde järgi“ (telepaatia?)

Harvaesinevad sõnad on tähtsad

- nende kaudu avaldub keelesüsteem
- neid on kokku palju

- koolis neid ei drillita
- nad ei jää meelde, nad on “nähtamatud”

Hea morfoloogiakirjeldus

- Võimaldab moodustada sõnavorme nagu inimene
  - Just tundmatute sõnade puhul!

Arvuti(morfoloogia) puhtteaduslik väärtus

Marx: tõe kriteerium on praktika,

- alles siis ‘saame asjast aru’, kui oskame seda ise teha/esile kutsuda
- alles siis saame väita, et oleme “mõistnud’ morfoloogiasüsteemi, kui eksisteerib programm, mis sõnu analüüsib/sünteesib nagu inimene

Muuttüübid ja produktiivsus

Muuttüübid

- Muuttüüp – ühel moel käänatavate/pööratavate sõnade hulk
- Aglutinatiivses keeles on vähe muuttüüpe, flektiivses palju (W. Wurzel)
- Sõna liigitamine muuttüüpi peab olema lihtne
- ... ja toetuma morfoloogia-välistele tunnustele
- Aga need ei määra üheselt?

→ Muuttüübid pole võrdsed

Wurzel, Wolfgang Ullrich 1987. System-dependent Morphological Naturalness in Inflection. – Leitmotifs in Natural Morphology. Studies in Language Companion Series, vol 10. Toim. Wolfgang U. Dressler, Willi Mayerthaler, Oswald Panagl, Wolfgang U. Wurzel.

Muuttüübid eestis

- Väga palju variante
- Eri liigitusalused
- Paralleelvormid

Ü. Viks. Muuttüübid eesti sõnastikes https://www.eki.ee/teemad/tyybijutt.html

Pärisnimede tüübid

1	Ago
2	Alfred, Aksel
7	Toomas Tooma
9	Joonas
10	Vanemuine, Jullinen
11	Kivikas, Pilatus
12	Merike
16	Aldo, Alina
17	(Meri)
19	Laidoner, Lebedev
22	Epp, Bill
25	Reykjavik
26	Aleksei

Muuttüüp võib olla:

- Aktiivne, stabiilne, produktiivne (kõne)
- Passiivne, ebastabiilne, mitteproduktiivne

(kese, tase, rase) -- lihtsõna, 2 silpi, esimene välde, -e

Morfoloogiateooria küsimus

Kuidas teame, kuidas käänata?

(raskusi pole, aga miks?)

Nt. milline tüvevokaal valida:

1K: *näpp*:*näpu*, *käpp*:*käpa*, *täpp*:*täpi*

2uk: *tüdruk*:*tüdruku*, *nooruk*:*nooruki*

või kas astmevaheldus või mitte:

kinnas:*kinda*, *pinnas*:*pinnase*

Sõnastikust pole abi

- kõigi sõnade käänamist pole võimalik ära õppida

  TÜ koondkorpuses on 100 000 erinevat lihtsõna, ÕSis ja EKSSis 2 korda vähem (*jahta*, *jorss*)

- uusi sõnu oskavad kõik automaatselt käänata

  äpp, *meem*, *võõruk*

- ... ja neid käänatakse ühte moodi, „tunde järgi“ (telepaatia?)

Keeletaju

- käänamisviisi peab määrama sõna väline kuju

  (sest muule pole tugineda)

  – haruldase ja väljamõeldud sõna käänamine on ühesugused

  – see käänamine tundub loomulik, tavapärane (teistsugune oleks naljakas, nt *äpp:äpu*, *äpp:äpa*)

Tekstikorpus ja produktiivsus

Kui sõna on vaja, siis ta luuakse

- to google -> guugeldama (laenamine)
- pub -> pubi (laenamine)
- linnavalitsus -> linnaviletsus (keelemängus)
- Vikerraadio -> vikerraadiolik (uue tähenduse väljendamiseks)

Uus sõna on ...

- lühiealine
- kergesti moodustatav
- kergesti mõistetav

Produktiivsus

... on keelekasutajate võime luua uusi sõnu

- sõnavorme
- tuletisi
- liitsõnu

NB!

häälduslikud piirangud; sobivus muutmissüsteemi; vormi ennustatavus

Produktiivsuse hindamine

Alternatiivsed loometeed

- Millist valitakse? Valiti?

Vaata korpust!

- uued sõnad
- hapax

Nähtu kujundab arvamuse

- Tähtis on malli sõnastikusagedus (mitte tekstisagedus)
- Malli raskus pole tähtis

  (J. Bybee katse vormimoodustuse õppimisel)

Milline käänamisviis on produktiivne?

- Näited eesti keele koondkorpusest
- Võitja võtab kõik ?

Tüvevokaal (*sepp*-tüüp)

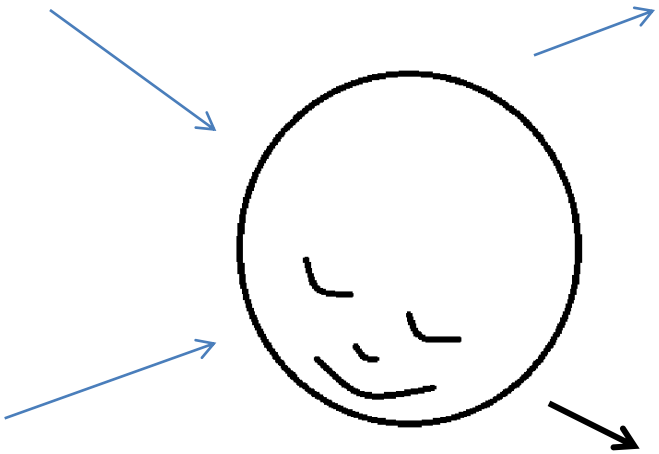
Mis kujundab ?
Varem nähtud tüvevokaal

Tekstis

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

Sõnastikus

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



Mis on tulemus ?
Tundmatu sõna tüvevokaal

Võiks olla

+u	? %
+e	? %
+a	? %
+i	? %

Korpuse põhjal on

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

Tüvevokaal (2 silpi, lõpus *k*)

tulnuk, ürik, tüdruk, uluk

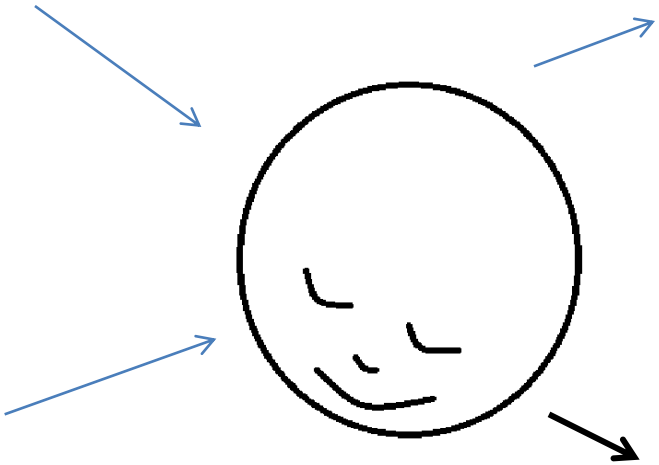
Mis kujundab ?
Varem nähtud tüvevokaal

Tekstis

ik+u	100%
uk+a	0,3%
uk+i	60%
uk+u	40%

Sõnastikus

ik+u	100%
uk+a	5%
uk+i	85%
uk+u	10%



Mis on tulemus ?
Tundmatu sõna tüvevokaal

Vällik, Baaruk

ik+u	? %
uk+a	? %
uk+i	? %
uk+u	? %

Korpuse põhjal

ik+u	**100%**
uk+a	**0%**
uk+i	**100%**
uk+u	**0%**

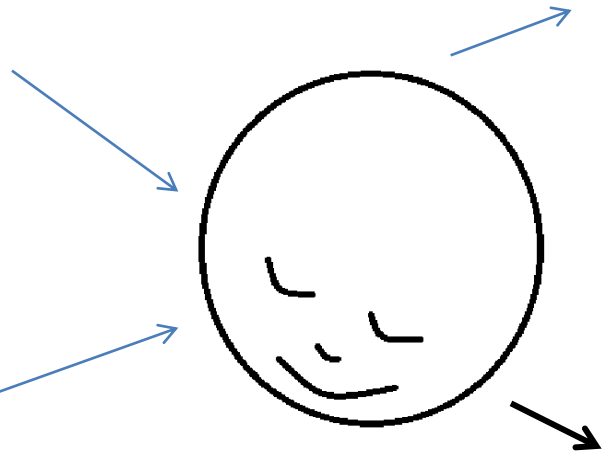
Astmevaheldus (2 silpi, II välde, lõpus *s*)

kinnas, *pinnas*, *soodus*, *boonus*

Mis kujundab ?
Varem nähtud mall

Tekstis
on a-v 50%
pole a-v 50%

Sõnastikus
on a-v 34%
pole a-v 66%



Mis on tulemus ?
Tundmatu sõna mall

Ninnas, voodus

on a-v ? %
pole a-v ? %

Korpuse põhjal
on a-v **0%**
pole a-v **100%**

Ainsuse sisseütlev

Kas valida lõpuks

sse (pesasse)

0 (pessa)

Ainsuse sisseütleva lõppude valik eri sõnade poolt (morf. ühest. korpus, 500 tuh sõna)

Lõpp	**Sõnavara**	**Sagedamad**
sse	468	registrisse 86, ametisse 84, Eestisse 56
0	448	toime 74, pähe 70, tuppa 52, korda 47
se	137	teenistusse 29, asutusse 12, vastavusse 8
0, sse	22	koju-kodusse 80-8, teise-teisesse 51-1, külla-külasse 22-1
de	14	meelde 97, läände 8, keelde 8
se, sse	12	teadvusse-teadvusesse 13-1, valdusse-valdusesse 5-1
tte	2	kätte 90, vette 14
de, sse	1	uude-uuesse 12-1

Mitmuse osastav

Kas valida lõpuks

sid	(pesasid, laikusid, paatisid, sabasid)
i	(pesi)
e	(laike, paate)
u	(sabu)

Mitmuse osastava lõppude valik eri sõnade poolt

VVS tüüp	tüüpsõna	kokku	sid	i	e	u	sid, i	sid, e	sid, u
17a	saba	30	16	2	1	6	3	0	2
17i	ribi	11	9		1			1	
18a	sõda	12	2		2	5		2	1
24	padi	17	1	1	4	10	0	0	1
22(a)	sepp	96	7	38	1	44	2	0	4
22(u)	laik	86	15		60			11	
22(e)	hing	27	2	25			0		
22kond	piirkond	10	0	10			0		
22(i)	siil	411	1		409			1	
19(i)	seminar	28	0		28			0	
25(u)	õnnelik	84	0		84			0	

V:sid vahekord tüübis 17a (saba)

sõna (53:0), vana (24:0), ala (0:18), maja (8:1), tera (8:0) , tava (0:6), osa (5:1), kala (5:0), häda (0:4), vaba (3:1), lina (3:0), vara (0:3), püha (2:1), keha (2:1), muna (2:0), kava (0:2), küla (0:2), kena (1:0), kana (1:0), saba (1:0), võsa (0:1), tara (0:1), tala (0:1), reha (0:1), raha (0:1), pala (0:1), oja (0:1), nina (0:1), lava (0:1), kaja (0:1)

id:sid vahekord tüübis *idee*

puu (28:0), töö (22:0), hea (20:0), luu (17:0), tee (16:0), maa (10:0), kuu (6:0), idee (5:1), pea (5:0), intervjuu (0:3), suu (2:0), essee (2:0), varietee (0:1), truu (1:0), soo (1:0), süžee (0:1), portree (0:1), palee (0:1)

Järeldus

- Sagedusandmed võimaldavad ennustada, mis suunas eesti keele käänamine areneb

Paralleelvormid

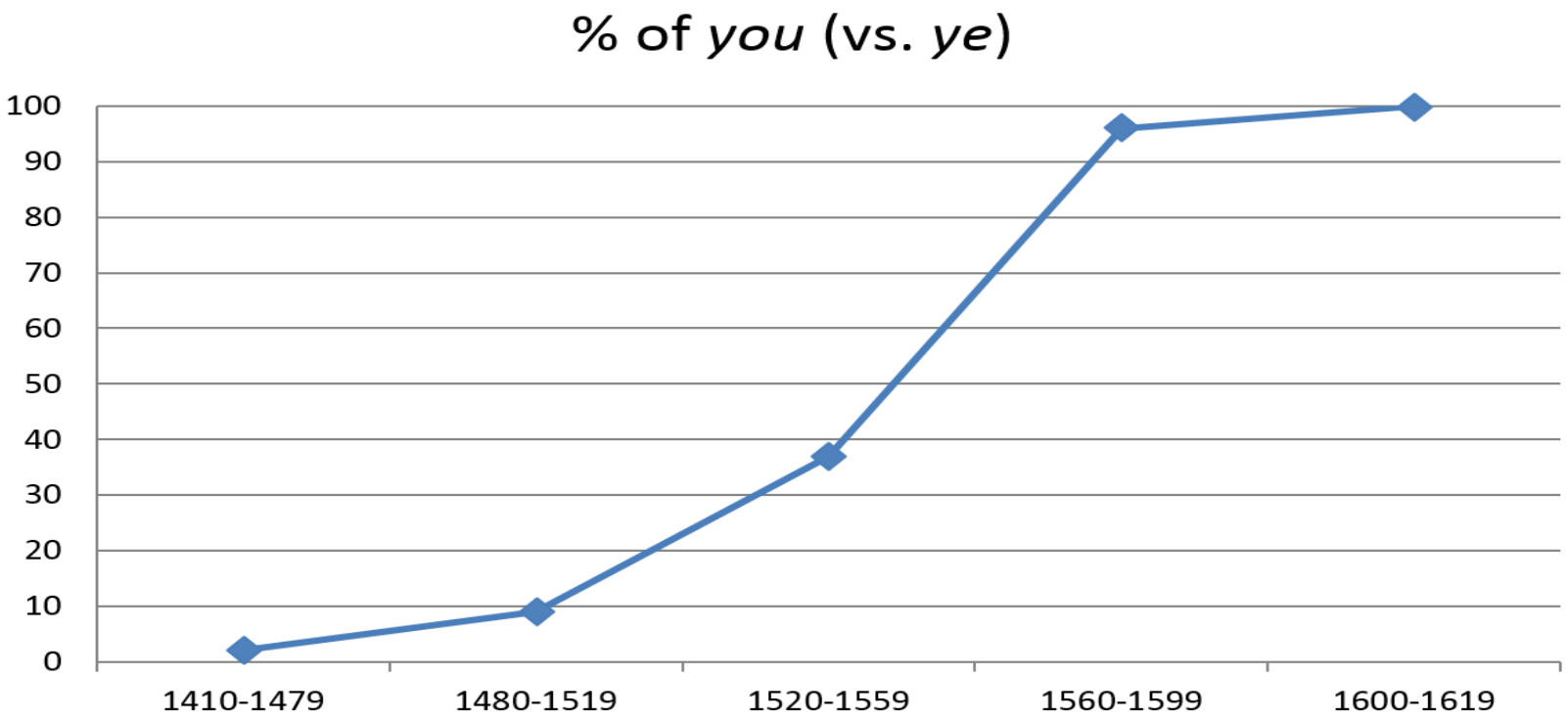
- Neid ei tohiks olla...

1 määratlus – mitu sõnavormi idee+N+Pl+Par → ideid, ideesid

- Intuitsioon petab ?
  - murded ?
  - liiga hea mälu ?
- Keelekirjeldus petab?
  - riigesse, külmisse, leivusse (ÕS 2013)
  - sõnade vale rühmitamine ?

    puu ←→ raguu

S-kõver



T. Nevalainen 2015

Paralleelsuse allikad

- Fonoloogilis-derivatiivne alus on mitmeti tõlgendatav
  - ümbrik, muuseum, ...
- Alus on üks, aga vanad sõnad ei järgi seda
  - siga (giga)

  Vana sõna reeglipärastub
  - hani – hane – hand (Wied) (lumi – lume – lund)
  - lumi – lumen – lumea

Kasutussagedus ja norm

- Hyp
  - C1 – kotisse, Fiatisse
- CommonNotNorm
  - peen – peeneid (kuni 2013)
  - köömen – kööment
- Rare
  - C1 – presidentisid; kuubi
- NotNorm
  - ämblikuid

Suundumused suurtes muuttüüpides

- Mitmuse osastav
- Sisseütlev
- Osa paradigmast (fonoloogilis-derivatiivne tõlgendus muutub)

Mitmuse osastav

1C

sid ↓ e ↑	+i	taud (Fiat, kabinet)
sid ↓ e ↑	+u	elanik
sid ↑ vokaal ↓	+a, +u, +e	piim, koon, eit

2V

sid ↑ vokaal ↓	kava

Ainsuse sisseütlev

1C

sse ↓ Ø↑	+i	taud (Fiat kabinet)
sse ↓ Ø ↑	+u	elanik
sse ↓ Ø ↑	+a, +u, +e	piim, koon, eit

2V

sse ↑ Ø ↓		kava, pere

Pool paradigmat

- ümbrik & ämblik
- VIRSIK (Rare) / ELANIK
  - ümbrikut / ümbrikku
  - ümbrikute / ümbrike
  - ümbrikuid / ümbrikke

  + ämblikut ämblikute ämblikuid (NotNorm)

- muuseum
- TAUD / REDEL
  - muuseumi / muuseumit (Rare)
  - muuseumi (Rare) / muuseumisse
  - muuseumide / muuseumite (Rare)
  - muuseume / muuseumeid (Rare)

Väikesed muuttüübid

- hoolas (hoolsat/hoolast; hoolsate/hoolaste)
- piprate (NotNorm) /piparde