Eesti keeletehnoloogia: kellele ja milleks

Ettekanne LI J. V. Veski päeval 27. juunil 2018

Kadri Vider
Eesti Keeleressursside Keskus

EKRK

Küsimus

Minu arvuti või nutiseade on...

A. Ingliskeelne – ca 50%
B. Eestikeelne – ca 18%
C. Muukeelne – ca 2%
D. Ei mäleta – ca 30%

Keel ja arvuti (laiemalt IT)

- Inimkeel on suhtlusvahend
  - Mis keeles suhtleb sinuga arvuti?
  - Või suhtled sina arvutiga?

  Spontaansed vestlused navi-seadmetega ei ole veel keeletehnoloogia

  Ka lokaliseerimine ja tarkvara eestindamine ei ole veel keeletehnoloogia

=> Kuid iga inimkeel vajab tehnoloogilist tuge

EKRK

„Ega inimene massina vasta saa!“

…ehk see, miks arvutilingvistika ja keeletehnoloogia on minu meelest lahe:

- Iga päev on kopp ees

+ Iga päev teen karjääri

…või ka poeetilisemalt:

„…kus kasteheinas põlvini me lapsed jooksime“

(L. Koidula „Meil aiaäärne tänavas“)

=> kas JOOKS+i+me või JOO+ksi+me

EKRK

Riiklik programm “Eesti keeletehnoloogia (2011-2017)” www.keeletehnoloogia.ee

Eesti keele arengukava (2011-2017) meede 3: **keele tehnoloogilise toe arendamine**

> **Eesmärk**: eesti keele keeletehnoloogiline tugi on võrdsel tasemel arenenud keeletehnoloogiaga riikide (nt Põhjamaad) keeltega suundades, mida nõuavad eesti keelele orienteeritud tarkvara arendused ja rakendused

- Osalevad: Eesti keeletehnoloogia ja arvutilingvistika kogukond, peamiselt TÜ, EKI, TTÜ kõnetehnoloogid
- Riigi rahaga tehtud projektide **tulemused on vabalt kasutatavad** (kui isikuandmete kaitse vms takistusi pole)

EKRK

EKT olulisemad tulemused

Tegevused ja õnnestumised

- On paranenud keeletehnoloogia olulisemate baastehnoloogiate (kõnetuvastus, kõnesüntees, masintõlge) kvaliteet ja kasvanud baasressursside (korpused) maht.
- On suurenenud keeletehnoloogia baastehnoloogiate ja –ressurssidest integreerimine erinevatesse süsteemidesse ja teenustesse.
- On kinnitatud eesti keeletehnoloogia jätkuprogramm aastateks 2018-2027.

Probleemid ja takistused

- Teadlikkus keeletehnoloogia pakutavatest võimalustest ei ole laialdaselt levinud; võimalikud kasutajad ei tea, mida, kus ja kuidas rakendada; sõnastus pole üldarusaadav.
- Keeletehnoloogia komponendid ei ole üldjuhul universaalsed ja koheselt kasutusele võetavad, vaid rakendamisel tuleb arvestada valdkondliku spetsiifikaga ning tehnilise valmisolekuga.

EKRK

Suured eesmärgid – kuidas neid saavutada?

- Eesti inimeste toomine infotehnoloogia juurde. Liidesed eestikeelseks, tulevikus kõnepõhised; seega vaja: masintõlge, kõnetuvastus, kõnesüntees. Sarnane asi puuetega inimeste puhul.
- Eesti keele õpetamine muukeelsetele - vaja toredaid süntaksi- ja morfoloogiamänge, eesti keele tehnoloogilise toe lõimimist keeleõppesüsteemidesse
- Rahvusliku mälu kasutatavus - eesti keelest teadlikud otsisüsteemid arhiividele, nt raadio- ja telearhiiv, üldisemalt kogu (kuuldav) internet.
- Rahvuslik uhkus (Google teab eesti keelest meist rohkem, aga meile ei ütle; mingi naviseadmega peab inglise keeles rääkima)

EKRK

Puutepunkte teiste valdkondadega

- Arvutiteadustega, laiemalt eestikeelsete suurandmete ja nende analüüsiga
- Meediauuringutega: **eestikeelse** teksti vormi-, lause- ja sisuanalüüs; sisukokkuvõtete tegemine; nime- ja ajaüksuste tuvastamine
- Õigusteadusega: keeleandmestik kui võimalik autoriõiguse objekt; isikuandmete kaitse (nt keele- ja rahvaluuleandmestike kogumisel ja taaskasutamisel), teksti- ja andmekaeve õiguslikud alused
- Geograafiaga: onomastika ehk nimeteadus, s.h toponüümika; eestikeelsete nimeüksuste tuvastamine

EKRK

Puutepunkte infoühiskonna kavaga

- **Paremad avalikud teenused**: EKT tulemused, eelkõige kõnetehnoloogiad, eesti keele vormitundlikku otsingut ja mitmekeelsust toetavad komponendid lõimida avalike teenuste kasutajaliidestesse
- **Nutikas riigivalitsemises**: teksti- ja suurandmete analüütikas saaks (palju enam) kasutada eesti keelt toetavat vormi- ja lauseanalüüsi, kõnetuvastust, emotsioonituvastust, valdkondlikku sisukokkuvõtete tegemist, tõlkeabivahendeid, terminite ja dokumentide linkimist jne

EKRK

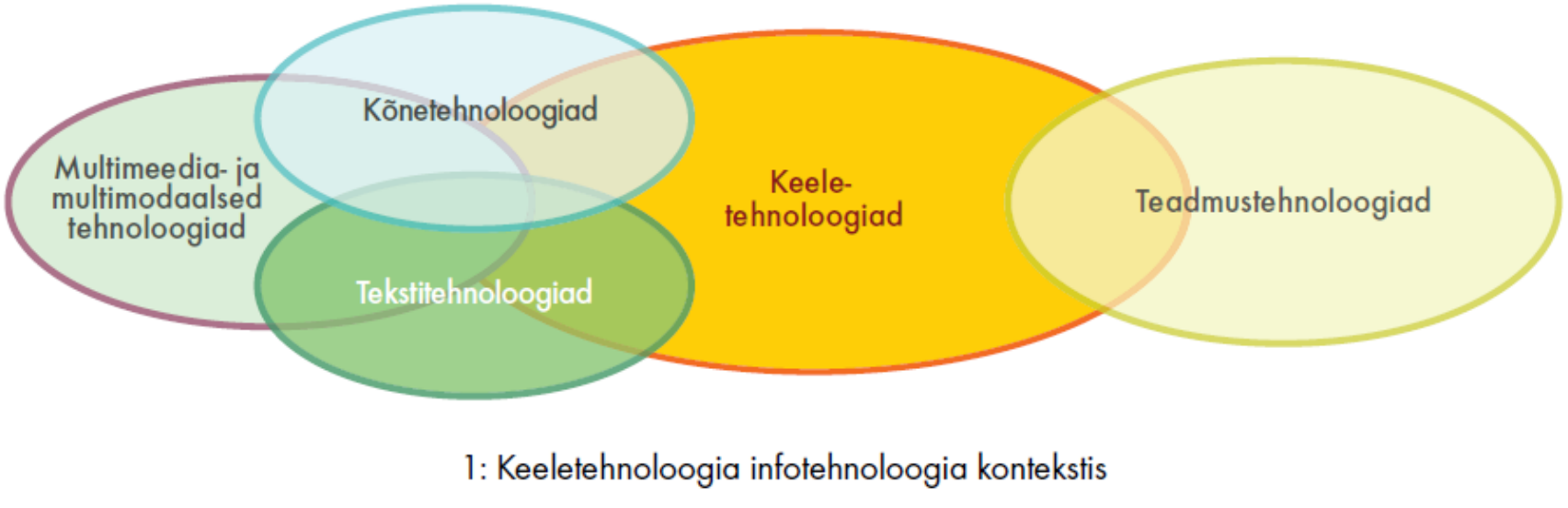
Tänan tähelepanu eest!

Küsimused, mõtted, ideed...

Võib kirjutada ka kadri.vider@ut.ee

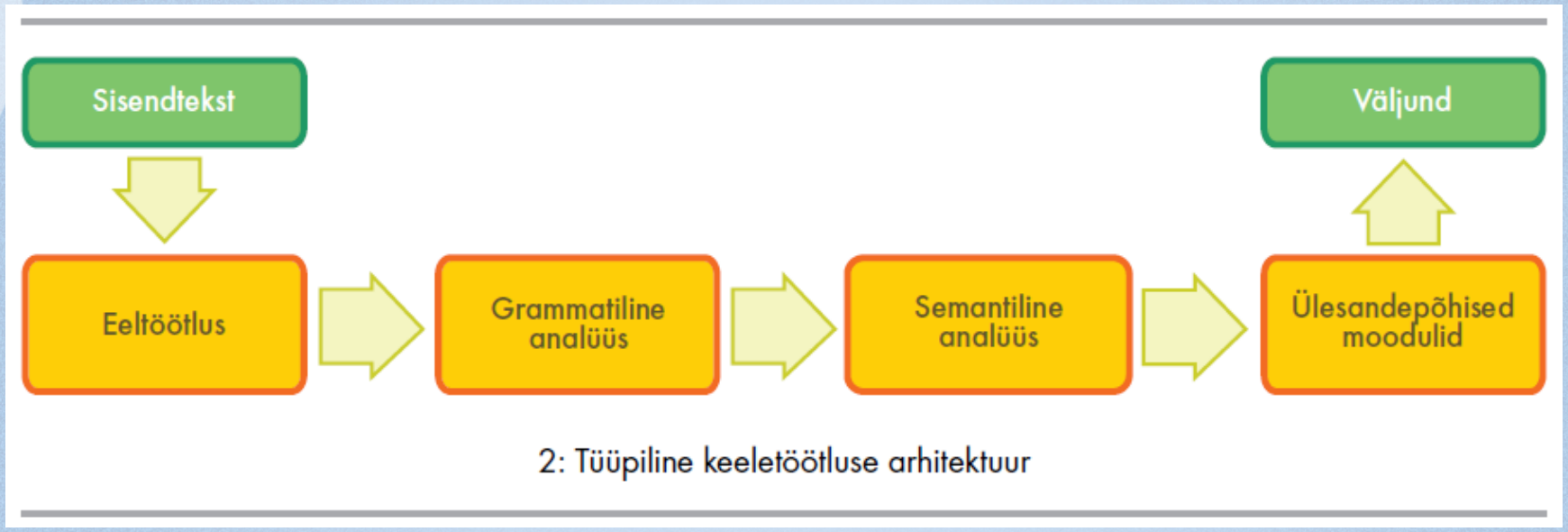
EKRK

Keeletehnoloogia infotehnoloogia kontekstis



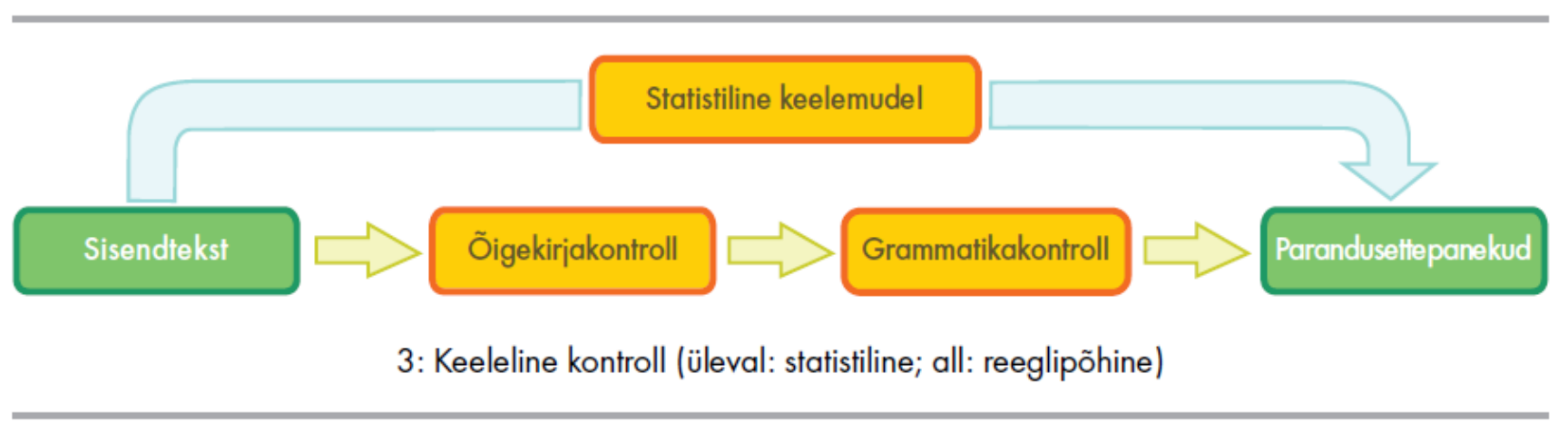
1: Keeletehnoloogia infotehnoloogia kontekstis

Keeletöötluse arhitektuur



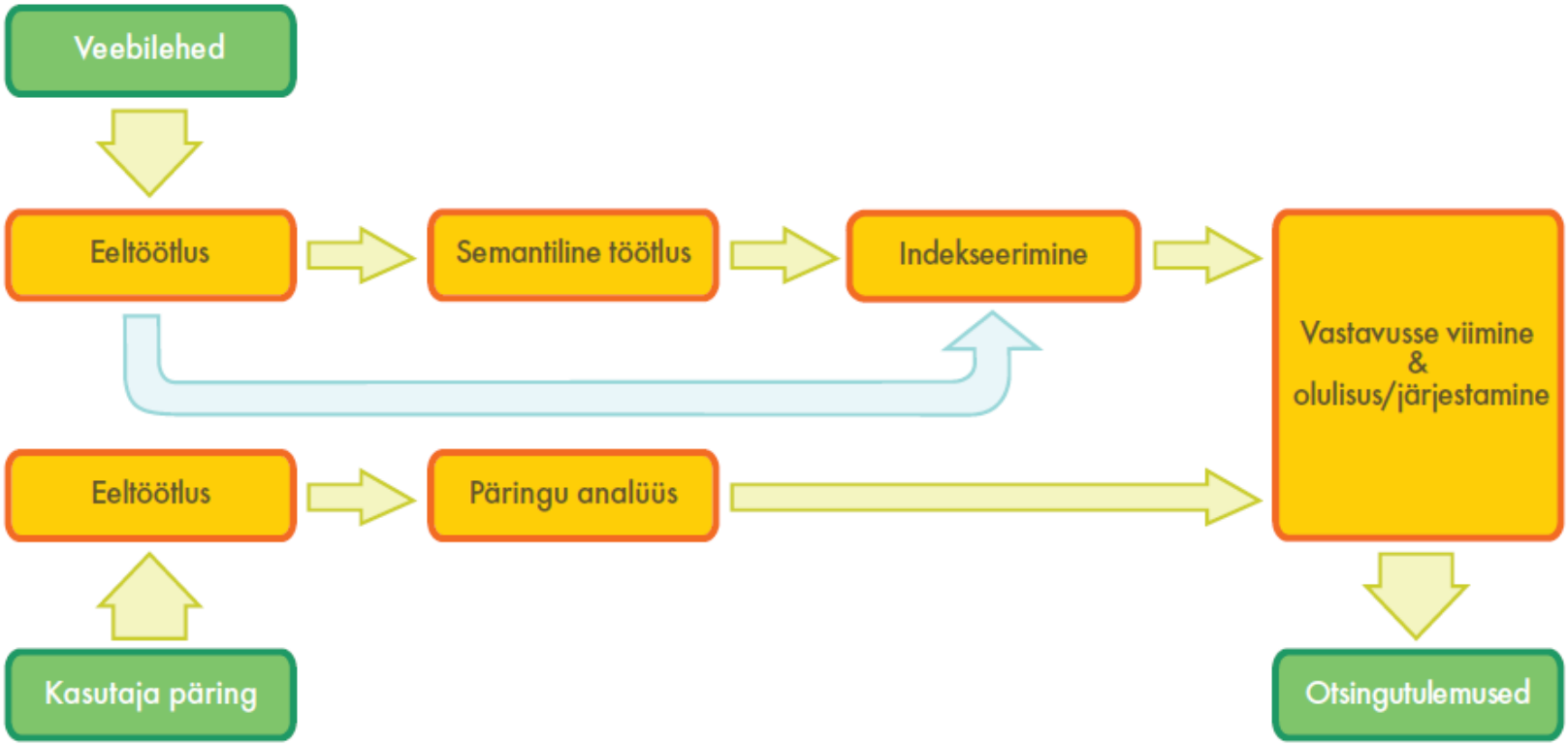
2: Tüüpiline keeletöötluse arhitektuur

Keeleline kontroll



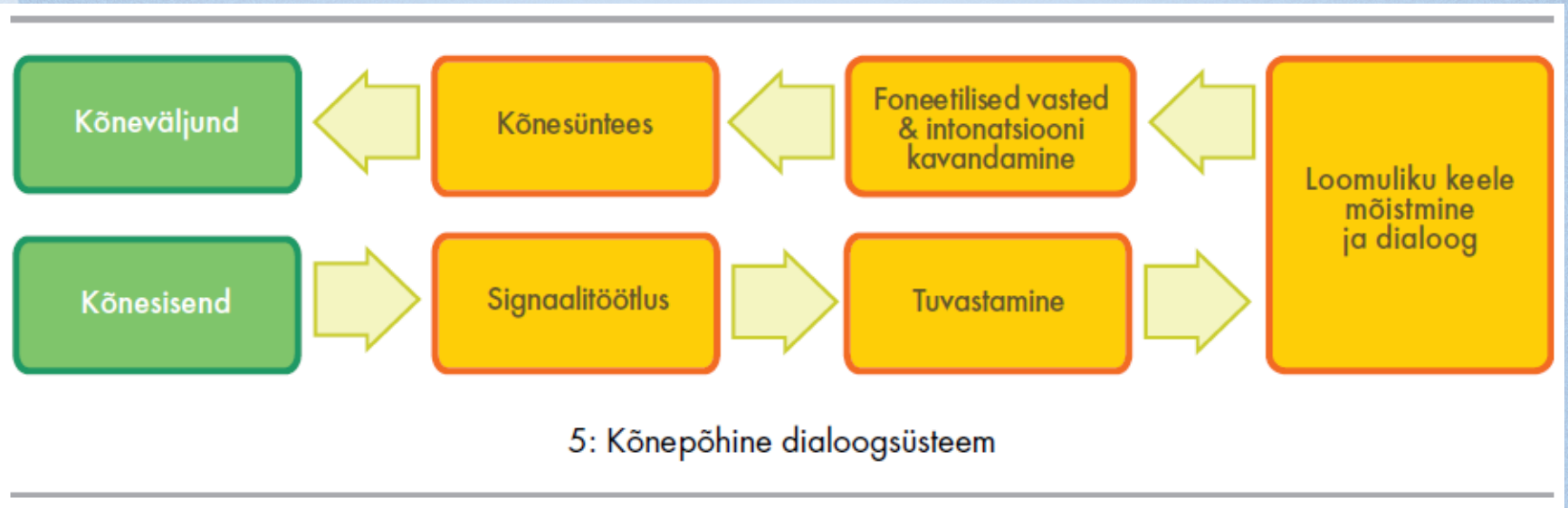
3: Keeleline kontroll (üleval: statistiline; all: reeglipõhine)

Veebiotsing



4: Veebiotsing

Kõnepõhine dialoog



5: Kõnepõhine dialoogsüsteem

Masintõlge

6: Masintõlge (vasakul: statistiline; paremal: reeglipõhine)

EKRK

Eesti keel digiajastul. Estonian Langugage in the Digital Age http://www.meta-net.eu/whitepapers/e-book/estonian.pdf

Eesti keele arengukava (EKA 2011-2017)

Meede 2: Eesti keele uurimine ja keelekogud

- Keeleuurimise tulemuslikkus sõltub andmestikust ehk kõigile uurimisülesannetele vastavatest keelekogudest

- Riiklik programm “Eesti keel ja kultuurimälu”

Mis on keeleressurss?

- Digitaalne
- Keeleandmestik:
  - Sõnavarakogum ehk leksikaalne andmestik
  - Tekstikogum ehk korpus
  - Heli kujul andmestik ehk kõneandmebaasid
- Keeletöötlusvahendid:
  - Tekstitöötlusvahendid, nt speller, vormianalüüs, lauseanalüüs...
  - Kõnetöötlusvahendid, nt kõnetuvastus
  - NB! Ka masintõlge, vt

EKRK

Sõnastikud jm sõnavararessursid

- Eesti Keele Instituudi sõnastike lehekülg: http://portaal.eki.ee/sonaraamatud.html
  - Ükskeelsed sõnastikud: ÕS, seletav sõnaraamat, etümoloogia, murdesõnastik, slängisõnastik…
  - Mitmekeelsed sõnastikud: soome, inglise, vene, läti, ungari, norra...
- Vaata ka keeleveeb.ee -> Tõlkesõnastikud
  -> Erialasõnastikud
  -> Koolisõnastikud
  -> Viited

EKRK

Wordnet

- Tesauruse tüüpi sõnavarakogum, mille aatomiteks on sünohulgad = ühte mõistet väljendavad sünonüümsed sõnad
- Paljude keelte wordnetid: http://globalwordnet.org/wordnets-in-the-world/
- Eesti wordnet ja selle päring: http://www.cl.ut.ee/ressursid/teksaurus/
- Vaata ka Keeleveeb.ee: ühispäringu valikutes on ka eesti wordnet

EKRK

Terminibaasid

- https://term.eki.ee/termbases/index/ - avalikud terminibaasid on eriala ekspertide vabatahtliku terminitöö tulemus
  1) milline on hetkel suurim terminibaas?
  2) Leia akadeemilise väljendusoskuse terminite seast mõni tundmatu termin
- Vaata ka keeleveeb.ee -> Erialasõnastikud
  -> Koolisõnastikud
  -> Viited
- Vaata ka kohanimed ja maailma maade nimed EKI lehelt: http://www.eki.ee/knab/knab.htm

EKRK

Kasulikke veebikodusid

- Keeleveeb.ee – erinevate allikate sõnastikud, korpused, erialasõnastikud, mitmekeelsed sõnastikud, wordnet – kõik ühe päringu abil!
- http://kn.eki.ee/ - EKI keelevarad koos
- KORP (korp.keeleressursid.ee) - paindlik korpuspäring suurimate eesti keele korpuste pealt (kuni pool miljardit sõna eestikeelset teksti)

- Proovige otsinguna näiteks sõnu: **pime**, **taodelda**, **naine**

EKRK

Least, but not last – let’s play!

- Keelemängud: https://keeleressursid.ee/et/keelemangud