

Näitelausete korpuspõhine automaattuvastus

Kristina Koppel, vanemarvutileksikograaf
Eesti Keele Instituut | EKI





Doktoritöö

Koppel, Kristina (2020). Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Doktoritöö, Tartu Ülikool). Tartu: Tartu Ülikooli Kirjastus.

Juhendajad: Jelena Kallas, PhD (EKI), Raili Pool, PhD (TÜ)



Teema sünd

„Eesti keele naabersõnade“ sõnastiku baasi täisautomaatne genereerimine 2014. aastal

- B2–C1 keeleoskustasemel eesti keele õppijale
- Kogu sõnastiku sisu tekstikorpusest sõnastikusüsteemi
 - Märksõnad
 - Kollokatsioonid
 - Sagedus
 - **Näitelauseid**

→ Järeltoimetamine

Korpusleksikograafia ja automaatne leksikograafia

Töö põhilised eesmärgid

- Välja selgitada hea näitelause formaalsed parameetrid eesti õppesõnastikele
- Luua meetod eesti keele jaoks, mis valib korpusest välja eri keeleoskustasemetele sobivaid autentseid lauseid
- Luua tasemekohaseid lauseid sisaldav õppekorpus
- Evalveerida automaatselt valitud näitelauseste kvaliteeti
- Rakendada õppekorpust keeleõppekeskkonnas [SkELL](#) ja keeleportaalis [Sõnaveeb](#)

Analüüsimaterjal

Eesti Keele Instituudis
koostatud **sõnastikud:**

- "Eesti keele põhisõnavara sõnastik" (2014)
- "Eesti keele sõnaraamat 2019"
- "Eesti keele naabersõnad 2019"

Korpused:

- "Eesti keele A1–C1 õpikute korpus 2018"

Testimiseks:

- "Eesti keele ühendkorpus 2013"
- "Eesti keele ühendkorpus 2017"

Uurimismeetod

Reeglipõhine lähenemine: GDEX ehk **Good Dictionary EXamples**

1. Funktsioon Sketch Engine'is
2. Keskmis universaalne reeglipõhine valem, mida on täiendatud keelespetsiifiliste parameetritega (lause pikkus, sõnade sagedus korpuses, märksõna asukoht, märksõna kordumine jm)
3. Töötab "sõelana": sorteerib välja sobimatud kandidaadid ja reastab ülejäänud paremuse järjekorda
4. Hindab parameetrite abil korpuslause komponente ja määrab igale lausele **skoori**

| Rank ↓ | Sentence | Score ↑ |
|--------|--|---------|
| 1 | Ettekannetega esinesid mõlema kõrgkooli üliõpilased. | 0.99 |
| 2 | Toimusid mitmed huvitavad ettekanded ja kontserdid. | 0.98 |
| 3 | Alustan oma ettekannet inimeste arvamustest. | 0.98 |
| 4 | Ma ei lahkunud enne oma ettekande ega küsimuste lõppu. | 0.97 |
| 5 | Ma tänan teid veel kord põhjalike ettekannete eest! | 0.97 |
| 6 | Ettekandeks on sõna justiitsministril. | 0.95 |
| 7 | Praktilisi ettekandeid ja töötube jätkus pikaks päevaks, kasulikke näiteid sai pea iga aine õpetaja. | 0.95 |

GDEXi poolt pakutud näitelause kandidaadid lemmale *ettekanne*.

Tulemused

6 konfiguratsioonifaili:

- GDEX 1.2
- GDEX 1.3
- **GDEX 1.4**
- etBasic-v1
- etIndependent-v1
- etProfocient-v1

```
formula: >
(50 * all(
  is_whole_sentence(),
  length > 5,
  length < 20,
  max([len(w) for w in words]) < 20,
  count_matches(tags, verb) > 0,
  blacklist(words, illegal_chars),
  not match(lemmas[0], bad_first_word),
  not match(space_separated(words), bad_first_two),
  not match(tags[0], bad_first_tag),
  match(words[0], lowercase),
  min([word_frequency(w) for w in words]) > 5,
  keyword_repetition(lemmas) == 1
)
+ 50 * optimal_interval(length, 10, 12)
* greylist(words, rare_chars, 0.05) * 1.09
* greylist(lemposs, anaphors, 0.05)
* greylist(lemma_lcs, bad_words, 0.25)
* greylist(tags, abbreviation, 0.5)
* (1 - 0.3 * (count_matches(lemmas, 'kroon') and count_matches(tags, 'N')))
* max(0, 1 - sum([0.2 for lemma in lemmas if lemma_frequency(lemma) < 200]))
* max(0, 1 - 0.5 * len([t for t in tokens if t.tag==verb and match(t.features, verb_nonfinite_suffix)]))
) / 100

frequency_reference_corpus: estonianRC
variables:
illegal_chars: ([<|\\|>|\/|\\|}{^@*~#=_~])
rare_chars: ([A-Z0-9'.,!?) (:;"'«»"…$-])
lowercase: ([a-z])
conjunction: J
abbreviation: Y
anaphors: ^(mina-p|sina-p|tema-p|meie-p|teie-p|nemad-p|ma-p|sa-p|ta-p|me-p|te-p|nad-p|see-p|too-p|siin-p|...)$
verb: V
verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
bad_first_tag: (I|Y|G|Z|J|T|X)
bad_first_word: ^(seejärel|samamoodi|ühesõnaga|niisugune|nad|see|samuti|ja|samas|aga|näiteks|kuid|seega|et|...)$
bad_first_two: ^((Ainult et)|(Ainult nii)|(Ehk siis)|(Ehk teisisõnu)|(Eriti kui)|(Eriti juhul)|(Eriti just)|...)$
bad_words: ^(loll|lollakas|lollpea|lollike|lausloll|värdjas|...)$
```

GDEXi konfiguratsioonifail

Eesti keele õppekorpus 2018 (etSkELL)

- Sisaldab GDEXi abil filtreeritud lauseid eesti keele ühendkorpusest 2017

Tulemuste evalveerimine

1. Hindajad: EKI leksikograafid ja TÜ/TLÜ üliõpilased
2. Andmestik: 40 juhuslikku märksõna, 160 juhuslikku lauset
 - juhuslik GDEX 1.4 parameetritele vastav korpuslause
 - juhuslik GDEX 1.4 parameetritele mittevastav korpuslause
 - juhuslik filtreerimata korpuslause
 - juhuslik "Eesti keele sõnaraamatu 2019" näitelause
3. Tulemused:
 1. 96% sõnastiku näitelausest hinnati sobivaks
 2. **85% GDEX 1.4 parameetritele vastavatest korpuslausetest hinnati sobivaks**
 3. 94% GDEX 1.4 parameetritele mittevastavatest korpuslausetest hinnati sobimatuks
 4. 60% filtreerimata korpuslausetest hinnati sobimatuks

Kas see lause sobib sõna **äkki** näitelauseks?

Äkki prooviks keegi teda asendada?

Jah Ei Ei oska hinnata

Lahendad praegu ülesannet number **2**. Oled lahendanud **0** ülesannet **160** -st.
Sa peaksid lahendama **40** ülesannet.
Kui sul tekib mingeid kommentaare, siis täida tagasiside [küsimustik](#).

Lause hindamine Pybossa platvormis

Rakenduslik väljund I

SkELL ehk Sketch Engine for Language Learning

SKELL

elektritõuks Eesti keel

Examples Word sketch Similar words

elektritõuks 0.33 hits per million

- Isiklik **elektritõuks** võiks kiirelt liikumise küsimuse hoobilt lahendada.
- Peale kalorete kaotamise aitab **elektritõuksiga** sõitmine vähendada stressitaset.
- Viimasel ajal on **elektritõuksid** võtnud epideemia mõõtmeid.
- Kas kihutavad **elektritõuksid** ongi kogu revolutsioon?
- Kõrvklapid ja **elektritõuks** saadetakse mängijale kahe nädala jooksul pärast kampaania lõppu.
- Takso asemel sõidame koosolekule **elektritõuksiga** ise.
- Milles erinevad naiste ja meeste arvamused **elektritõukside** omadusi hinnates?
- Elektritõuksiga** sõites soovitame alati kanda kiivrit!
- Olen esimese **elektritõuksi** otsingul ja palun veidi abi.
- Eraldi statistika me **elektritõuksidega** juhtunud õnnetuste üle ei pea.

Näitelauseid

elektritõuks nimisõna

automobiil kõnnikepp laadimis pistik jooksuking laadimisjuhe nibin šokolaadivabrik Nibin-nabin nibin-nabin
hobukaarik pisimopeed pulmakink mikker telksaun põidlaküüt lapsekäru rattakiiver lastevanker
välihospital suusapriil hiirematt msi puutepliats laadimiskarp sahtlikapp tõuks torbik surfilaud mant
surfivarustus

lastevanker rattakiiver
pisimopeed laadimisjuhe
jooksuking laadimispistik mikker
telksaun nibin automobiil šokolaadivabrik
Nibin-nabin kõnnikepp hobukaarik
nibin-nabin hiirematt pulmakink välihospital
surfivarustus lapsekäru suusapriil

Tesaurus

elektritõuks nimisõna Show context

| omadussõnad | eelnevad nimisõnad | järgnevad nimisõnad | mida tavaliselt teeb |
|---------------|---------------------------------|-------------------------------------|----------------------|
| 1. renditav | 1. aasta_elektritõuks | 1. elektritõuksi_kott | 1. mahtuma |
| 2. saatuslik | | 2. elektritõuksi_hind | 2. kaaluma |
| 3. ootav | | 3. elektritõukside_kasutaja | 3. koguma |
| 4. vinge | 2. teenuspakkujate_elektritõuks | 4. elektritõukside_buum | 4. kannatama |
| 5. müüdav | | 5. elektritõukside_tootja | 5. sõitma |
| 6. mugav | | 6. elektritõuksi_kommuun | 6. kutsuma |
| 7. kõnealune | 3. laste_elektritõuks | 7. elektritõuksi_ost | |
| 8. populaarne | | 8. elektritõuksi_säästurežiim | |
| 9. ohtlik | | 9. elektritõukside_pidur | |
| | | 10. elektritõuksi_kasutusiga | |
| | | 11. elektritõuksi_bussiootepaviljon | |
| | | 12. elektritõukside_rent | |
| | | 13. elektritõuksi_rehvirohk | |
| | | 14. elektritõukside_kojusaamine | |
| | | 15. elektritõukside_sõidukiirus | |

| mida sellega tavaliselt tehakse | ja/või |
|---------------------------------|---------------------|
| 1. jalestama | 1. kiirpildikaamera |
| | 2. elektrijalgratas |
| | 3. lehepuhur |
| | 4. rendiauto |
| | 5. nutikell |
| | 6. kõrvklapp |
| 2. saatma | 7. sülearvuti |
| | 8. jalgratas |
| | 9. ratas |

Sõnavisandid

Rakenduslik väljund II

Veebilauseid Sõnaveebis

Sõnaveeb
Eesti Keele Instituut

EST ▾ MENÜÜ ☰

Keel Kõik keeled ▾ Sõnakogud Kõik sõnakogud ▾ Ehk mul veab Tagasiside

et 3D-printer nimisõna ✎ 13.09.2019

☰ EKI ÜHENDSÕNASTIK 2022

et masin digitaalsest mudelist kolmemõõtmelise eseme tekitamiseks (printimiseks)

ru 3D-принтер

☰ TERMINIVÕRGUSTIK

Sõnavormid ⓘ

| | |
|--------------|---------------|
| 3D-printer | 3D-printerid |
| 3D-printeri | 3D-printerite |
| 3D-printerit | 3D-printereid |

↗ Näita tabelina

Veebilauseid ⓘ

▲ Veebilauseid on automaatselt valitud ning võivad sisaldada vigu.

Kool lubas **3D-printeri** olemasolul keskenduda droonide valmistamisele.

Esimene kaubanduslik **3D-printer** tuli müügile 1986. aastal.

Mängus vajalikud kujundid on **3D-printeriga** prinditud.

Esimesed ideed **3D-printerite** vallas tekkisid juba eelmise sajandi 70– 80ndatel.

Eestis on **3D-printerid** kättesaadavad olnud juba mitu aastat.

Programmi esimene pilootprojekt kingib 50 Eesti põhikoolile ja gümnaasiumile **3D-printeri**.

Pilootprojekti raames varustati 50 Eesti põhikooli ja gümnaasiumi **3D-printeriga**.

Longman Dictionary of Contemporary English

Idoceanline.com

Examples from the Corpus

performance

- There was a **performance** of "Gisele" in the San Diego State Open Air Theatre.
- The school has tried to use **technology** and writing across **subjects** to improve students' **academic performance**.
- **Targets** may be **set** for any **parameter** that can be **measured** as the **project** proceeds, such as **cost**, **time** and **performance**.
- Three **criteria** have been **chosen**, attempting to measure the most important **attributes** of company **performance** over the year.
- the **disappointing performance** of the **bond market**
- The evening **performance** will begin at 8:00 pm.
- It is the first **performance** of Berlioz's **Requiem** in this **city** in over 20 years.
- Its **performance** on **mountain roads** was **impressive**.
- Have you ever **heard** a live **performance** of Beethoven's **Seventh Symphony**?
- **Quick**, somebody **book** a **local performance**.
- Only time will tell if this is a **serious effort** at improving both **public sector accountability** and **overall performance**.
- But the Lakers were up to the **task**, despite one of the **Clippers'** better recent **performances**.
- There are no **tickets** left for this evening's **performance**.
- This evening's **performance** begins at 8:00 pm.
- The new **program** will better **evaluate** the **performance** of students and **teachers**.
- the **performance** of his **official duties**
- Some companies **link** pay to **performance**.
- **Investors** respond to **performance** and we've not been in **existence** long enough yet.

Collins Dictionary

collinsdictionary.com

performance

⚠ These examples have been automatically selected and may contain sensitive content that does not reflect the opinions or policies of Collins, or its parent company HarperCollins.

We welcome feedback: [report an example sentence to the Collins team](#). [Read more...](#)

These early mirror pieces quickly established her as a pioneer in performance.

THE GUARDIAN (2015)

The technology is proven with years of performance in the navy.

THE GUARDIAN (2015)

This premiere performance felt entirely assured.

THE GUARDIAN (2016)

So from the results and performance standpoint we were pleased.

THE GUARDIAN (2018)

All in all the bowling performance was outstanding.

THE GUARDIAN (2019)

That it has had no measurable effect on school performance, job prospects, criminality or welfare dependency.

THE SUN (2016)

This could be a drawback if his performance fails to live up to expectations.

THE SUN (2010)

Merriam-Webster

merriam-webster.com

Recent Examples on the Web

// Data literacy can also guide workers and managers toward a deeper understanding of the variables that drive *performance* measures.

— Merav Yuravlivker, *Forbes*, 17 June 2022

// Along with a potential *performance* version of the bZ4X, perhaps the Camry could be on deck for such a makeover.

— Eric Stafford, *Car and Driver*, 17 June 2022

// Some employees are also rewarded with *performance* bonuses.

— Beth Decarbo, *Washington Post*, 17 June 2022

// Features local musicians in a collaborative *performance* environment.

— Cindy Kent, *Sun Sentinel*, 17 June 2022

// UniverSoul Circus still uses *performance* animals including zebras and horses, and the animals are introduced in specific segments during the show then hurriedly corralled and wrangled out of the ring.

— Maria Morales, *Baltimore Sun*, 17 June 2022

// The sudden comedown of Three Arrows follows the firm's previously strong *performance* record.

— Serena Ng, *WSJ*, 17 June 2022

Probleemid

1. Korpuse sisu ja maht

2. Märghendamise kvaliteet

- lemmatiseerimise ja morfoloogilise märghendamise vead
- lausestamine
- mitmesõnalised üksused
- leksikon
- trükivead

3. Grammatiline mitmesus

- lekseemide homonüümia (*tamm:tamme, tamm:tammi*)
- paradigmataväline vormisisene homonüümia (*teod: tigu ja tegu*)
- paradigmataväline vormidevaheline homonüümia (*kalla: kalla ja kallama*)

4. Semantiline mitmesus

- polüseemia ehk mitmetähenduslikkus

Veebilauseid ⓘ

⚠ Veebilauseid on automaatselt valitud ning võivad sisaldada vigu.

Must vari lume valgel **tasutal** koputab uksele.

Vesi ja **tasutal** ilusad Alpid, mida rohkemat tahta?

Paslik on meelde tuletada ja kuulata tänases maailmas toimuva **tasutal** spetsialisti soovitusi käitumiseks antud olukorras.

Põõsad lisavad kodusele **tasutale** roheline ja looduslähedase tausta.

Eestisse püütakse lasta vaid kristliku **tasutaga** pagulasi, kel oleks lihtsam meie ühiskonda sulanduda.

Tegime veel kiriku **tasutal** pilti, ning lõpuks keerasime auto kodu poole.

Kokkuvõtteks, kui soovid endale sõpra, kellega rääkida aktiivselt ja huvitava **tasutaga**, siis kirjuta.

Raamat ise oli armastusest, moraalist, ihast, millegi puudumisest kõige olemasolu **tasutal**.

Kursusega püütakse vastata paljudele sisserände **tasutaga** kokku puutuvate inimeste küsimustele, mis tulenevad peamiselt religioossetest ja kultuurilistest eripäradest.

Mis tulevik toob?

CrowLL ehk Crowdsourcing for Language Learning

<http://www.crowll.org>

Rahvusvaheline koostööprojekt viie riigi vahel

- Portugal: Tanara Zingano Kuhn, Ana Luís
- Sloveenia: Špela Arhar Holdt, Iztok Kosem
- Iisrael: Rina Zviel-Girshin
- Holland: Carole Tiberius
- Eesti: Kristina Koppel



AITÄH