



Kristina Koppel

eesti keele instituudi vanemarvutileksikograaf

Autentsed näitelaused teoorias ja praktikas

Oled sa kunagi mõelnud, kust tulevad sõnastiku näitelaused? Kas sõnastiku koostaja mõtleb need lakke vaadates ise välja või on tal olemas mingisugune põhjatu allikas, kust näiteid ammutada? Ja kui see allikas pole salajane, siis kas sinagi võiksid sellele ligi pääseda?

Viimase paarikümne aastaga on sõnastike koostamises ehk leksikograafias toimunud ulatuslik digipööre: uurimisandmetena kasutatakse pabersedelite asemel üha mahukamaid elektroonilisi tekstikogusid ehk korpusi, sõnastikke koostatakse lihtsamate tekstitöötlusprogrammide asemel keerukamates sõnastikusüsteemides ning pabersõnastikud on asendunud elektroonilistega. Leksikograafia areng on jõudnud niikaugele, et sõnastike andmebaase luuakse automaatselt ning haruldased ei ole enam ka juhud, kus teatud tüüpi leksikograafilist infot kuvatakse sõnastiku kasutajale autentsel kujul otse korpusest.

See digipööre on jõudnud ka eesti keele instituuti. Alates 2017. aastast arendame keeleportaali Sõnaveeb¹, mis on uuenduslik mitmeski mõttes – kasutaja näeb kogu infot sõna kohta ühes päringuaknas, ta saab otsisõna ise dikteerida (selleks rakendame kõnetuvastust, mis muudab kõne tekstiks) ning lasta arvutil näitelauseid ette lugeda (selleks kasutame kõnesünteesi, mis muudab teksti kõneks). Esmakordselt eesti sõnastike ajaloos näeb sõnastiku kasutaja veebilausete näol ka leksikograafi poolt toimetamata autentset korpusinfot.

¹ www.sonaveeb.ee (03.03.2021).

Sõnastiku näitelause ja selle automaatne valik

Näitelause on sõnaartikli oluline üksus, mis täidab selget eesmärki ning annab lisaväärtust. Emakeelsele kasutajale suunatud ükskeelses sõnastikus illustreerivad näitelauseid sõnakasutust ja täiendavad sageli definitsiooni. Keeleõppijale suunatud sõnastikes peaksid näitelauseid illustreerima seda konteksti, milles otsitav sõna tavapäraselt esineb. Samuti peaks näide andma vihjeid selle kohta, kuidas sõna lauses käitub, st milliseid fraase või ühendeid sellega harilikult moodustatakse ning milliste teiste sõnadega see kõige sagedamini koos esineb, niisamuti näitlikustama tüüpilist registrit (nt ametlik keel, släng) jmt. Päritolu poolest võib näitelause olla kas autentne (pärituna korpuselt) või tehiskorpus (leksikograafi väljamõeldud) (Atkins, Rundell 2008), kuid autentseid lauseid on tehiskorpusel eelistatud juba korpusleksikograafia² algusaegadest. Kuna sõnakasutust ei saa ise välja mõelda – seda saab ainult korpuselt põhjal registreerida (Sinclair 1987) –, siis oleks kummaline ka näitelauseid ise välja mõelda, selle asemel et neid otse korpuselt võtta (Fox 1987).

Korpusel ongi leksikograafias põhiline näitelause allikas. Varem pidi leksikograaf lugema kümneid või sadu ridu, enne kui ta leidis korpuselt sobiva näitelause kandidaadi. Kõik korpuslauseid toimetamata kujul sõnastikku ei sobi, kuna need võivad olla poolikud või liiga pikad, grammatiliselt ebakorrektsed ning leksikaalselt keerulised. Viimasel aastakümnel on aga tehtud väga palju tööd selle nimel, et päring korpuselt väljastaks automaatselt ainult head näitelause kandidaadid. Selline automaatne lähenemine vähendab oluliselt leksikograafi ajakulu sõnaartikli koostamisel, samuti saab automaatselt valitud lauseid otse sõnastiku kasutajale kuvada. Seda teeme ka Sõnaveebis – kasutaja saab lugeda kuni 26 korpuslause, mille olemasoleva lihtsuse huvides nimetanud veebilauseteks.³ Veebilause (pildil 1 all paremas nurgas) kasulikkus ilmneb eriti selliste märksõnade puhul nagu *aastapalk*, kus leksikograafi näitelauseid puuduvad.

² Interdistsiplinaarne lingvistika valdkond, mille eesmärk on luua meetodeid, mis võimaldavad analüüsida suurte korpuselt põhjal sõnastike koostamiseks vajalikke andmeid.

³ Veebilauseid päritakse spetsiaalselt „Eesti keele veebilausete korpuselt 2020“, mille allikas omakorda on „Eesti keele ühendkorpus 2019“. Viimane koosneb 88% ulatuses veebist kogutud tekstidest – sellest ka termin *veebilause*. Veebilausete korpus sisaldab ainult neid ühendkorpuselt lauseid, mis vastavad hea näitelause tunnustele.





EST ▾
MENÜÜ ≡

Keel Kõik keeled ▾
Sõnakogud Kõik sõnakogud ▾

Tagasiside

et

aastapalk

EKI ÜHENDSÕNASTIK 2020

et

aasta jooksul saadav või teenitud töötasu (ühekordsena või kuupalkade summana)

ru

годовая заработная плата, годовой оклад, заработная плата за год

võrdle päevapalk, kuupalk, nädalapalk

Sõnavormid ⓘ

aastapalk

aastapalkad

aastapalka

aastapalkade

aastapalka

aastapalku ~ aa...

 Näita tabelina

Veebileauseid ⓘ

 Veebileased on automaatselt valitud ning võivad sisaldada vigu.

Töölise **aastapalk** võis olla peotäis pipart.

Teest saadav sissetulek läks linnapea **aastapalgaks**.

Paar tunnikest maksab kolmveerandi keskmise ametniku **aastapalgast**.

Töövõime täieliku kaotuse puhul maksab riik toetust viie **aastapalga** ulatuses.

Pilt 1. Veebileased Sõnaveebis

Need laused on tuvastatud reeglipõhist lähenemist rakendades. See eeldab, et programmile öeldakse ette, millistele tunnustele (nt lause ja sõnade pikkus, sõnade sagedus korpuses, teatud elemendi, nt asesõna, kirjavahemärk olemasolu või puudumine jms) hea näitelause peab vastama. Näiteks algab hea eesti keele sõnastiku näitelause suure tähega ja lõpeb kirjavahemärgiga ning sisaldab kindlasti tegusõna pöördelist vormi. Lause pikkus on 5–20 sõna ning sõna pikkus maksimaalselt 30 tähemärki. Lause ei alga sidesõnaga ega teatud anafooriliste sõnade (nt *mistõttu*, *järelikult*) ja sõnapaaridega (nt *sellest johtuvalt*, *ilmselt seetõttu*), mis viitavad tagasi eelmisele lausele, lõigule vm ning mis vajavad mõistmiseks rohkem konteksti. Samuti ei esine heas lauses teatud vulgaarseid ega rahvusi halvustavaid sõnu. (Koppel 2020)

Aga kas on üldsegi vaja kasutajale korraga nii palju näiteid kuvada? On tehtud uurimusi (Frankenberg-Garcia 2012, 2014; Lew, Adamska-Salaciak 2015), kus on selgunud, et kohati on uue sõna mõistmisel näidetest rohkem abi kui definitsioonist ning kuna näitelause jääb sageli (keeleõppija) ainsaks kokkupuuteks uue sõnavara ja grammatikaga, siis on soovitatav ühe sõna või kasutusmusteri kohta esitada mitu näitelause. Samuti on uuringud (Gavioli 2005; Aston 1997) näidanud, et korpuse ainesel põhinev õppimine

stimuleerib õpilasi, esitab neile suuremaid väljakutseid, tekitab uudishimu, mõjub motiveerivalt ning on efektiivne viis grammatikaga tutvuda ja sõnavara suurendada. Sellest johtuvalt tutvustan järgmiseks vabavaralist rakendust SkELL, mille abil saab korpuse materjali iseseisvalt analüüsida.

SkELL – rakendus keeleõppijatele ja -õpetajatele

SkELL ehk *Sketch Engine for Language Learning*⁴ on lihtne rakendus, mis aitab hõlpsasti kontrollida, kuidas mingit sõna või fraasi kasutatakse. SkELLi abil saab analüüsida inglise, saksa, itaalia, vene, tšehhi ja eesti keelt. Eesti keele info pärineb „Eesti keele õppekorpusest 2020 (SkELL)“, mis sisaldab ainult hea näitelause tunnustele vastavaid autentseid korpuslauseid.⁵

SkELLi abil näeb keeleõppija n-ö valmis tükke, mis aitavad suurendada nii passiivset kui ka aktiivset sõnavara. SkELLis saab lugeda näitelauseid (*Examples*), vaadata kollokatsioone ehk naabersõnu (*Word Sketch*) ja sarnaseid sõnu (*Similar words*) (pilt 2).



Pilt 2. SkELLi avalehel tuleb otsiakna kõrval kindlasti määrata ka keel

Näited (*Examples*) illustreerivad otsitavat sõna või fraasi selle loomulikus ümbruses. Otsida saab nii sõna algvormi (nt *päike*) kui ka kindlat sõnavormi (nt *päikesesse*). Algvormi puhul kuvatakse päringus otsisõna oma eri sõnavormides (pilt 3), sõnavormi otsides ainult selles konkreetses vormis, milles seda otsiti.

⁴ <https://skell.sketchengine.eu/> (vaadatud 03.03.2021)

⁵ Erinevus veebilausete korpusest seisneb selles, et kuna SkELL on mõeldud abistama pigem eesti keelt teise keele või võõrkeelena õppijaid, on SkELLi korpuses rohkem sõnu (vulgarismid, eufemismid jms) keelatud, st et neid korpuses üldse ei esine.

päike 142.41 hits per million

1. Tekil jalutas või võttis **päikest** hulk reisijaid.
2. Viinamarjad kasvavad kõrgel mägedes, kus on palju **päikest**.
3. Tumedad pilved katsid **päikese** ja maa värisel jalge all.
4. **Päike** vajub aeglaselt horisondi taha.
5. Rada algas ootamatult järsult ja keskpäevane **päike** ei aidanud hoogsale kõnnile samuti kaasa.
6. Kuigi õues paistis **päike**, tundus kõik nii pimedana.
7. Vaha kaitseb juuksestruktuuri ja ei lase **päikesel** liiga teha.

Pilt 3. Otsisõna *päike* näitelauseid

Otsida saab ka mitmesõnalisi üksusi (nt *toime tulema*, *nagu lepase reega*), kuid tuleb silmas pidada, et liides kuvab lauses sõnu täpselt selles järjekorras, nagu need on päringuaknasse kirjutatud.

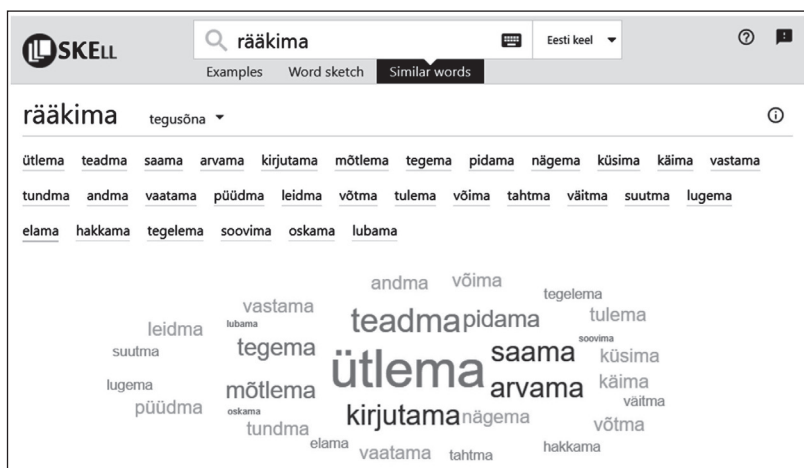
Kollokatsioone ehk naabersõnu näeb sakis *Word sketch* (sõnavisand, eesti keele sõnavisandite kohta loe lähemalt Kallas 2013). Kollokatsioonid on harilikult kahe sõna ühendid (otsisõna + selle naaber ehk kollokaat), mis kalduvad keeles sageli koos esinema ning mille tähendus on sisult läbipaistev, nt *must kohv*, *kohvi keetma* (pilt 4). Kollokatsioonide tundmine on eriti vajalik uut keelt õppides, sest sõnade naabrid ei pruugi eri keeltes kattuda. Näiteks on eesti keeles kohv kange või lahja, inglise keeles tugev või nõrk (*strong/weak coffee*). Eesti keeles me võtame kartuleid, soome keeles tõstame (*nostaa perunoita*). Eesti keeles vihma sajab, vene keeles

kohv			
nimisõna		Show context	
omadussõnad		järgnevad nimisõnad	
1. lahustuv	lahustuvat kohvi	1. kohvi_valmistamine	tee ja kohvi valmistamise võimalus
2. kange	kanget kohvi	2. kohvi_joomine	kohvi joomine
3. kofeiinivaba	kofeiinivaba kohvi	3. kohvi_hind	kohvi hind
4. kuum	kuuma kohvi	4. kohvi_maitse	kohvi maitse
5. aurav	auravat kohvi	5. kohvi_lõhn	kohvi lõhna
6. hommikune	hommikuse kohvi	6. kohvi_päev	tassi kohvi päevas
7. aromaadne	aromaatset kohvi	7. kohvi_tarbimine	kohvi tarbimine
8. must	musta kohvi	8. kohvi_tegemine	kohvi tegemiseks
9. maitsev	maitavat kohvi	9. kohvi_ekstrakt	roheline kohvi ekstrakti
10. roheline	rohelist kohvi	10. kohvi_kvaliteet	kohvi kvaliteet
11. magus	magusat kohvi	11. kohvi_keetmine	kohvi keetmiseks
12. värske	värsket kohvi	12. kohvi_tootmine	kohvi tootmine
13. kvaliteetne	kvaliteetset kohvi	13. kohvi_tass	kohvi tassi
14. tavaline	tavalist kohvi	14. kohvi_piim	kohvi piimaga
15. külm	külma kohvi	15. kohvi_nautimine	kohvi nautimiseks

Pilt 4. Otsisõna *kohv* kollokaadid

vihm hoopis läheb (*uđēm дожδь*). Kollokaadid on SkELLis reastatud sõnadevahelise seose tugevuse alusel ning jaotatud gruppidesse vastavalt sellele, mis sõnaliiki kollokaat kuulub.

Sarnaste sõnade (*Similar words*) sakis näeb automaatselt genereeritud pilve sõnadest, mis jagavad otsisõnaga samu kollokaate. Tavaliselt on nendel sõnadel ka sarnane tähendus (pilt 5). Sarnaste sõnade hulka võivad vahel sattuda ka sõnad, mis ei ole omavahel mõisteliselt sarnased, ei kuulu sarnasesse klassi või tähenduslikku välja või on isegi eri sõnaliigist. See tuleb sellest, et programm toetub sarnaste sõnade leidmisel üksnes eel- ja järelkontekstile.



Pilt 5. Otsisõna *rääkima* sarnased sõnad

SkELL peidab endas veel võimalusi keele analüüsimiseks – nüüd tuleb need vaid ise üles leida!

Kokkuvõtteks

Kuna nii Sõnaveebis kui ka SkELLis kuvatakse automaatselt korpuslau-seid, mida leksikograaf ei ole üle toimetanud, tuleb leppida tõsiasjaga, et neis esineb ka vigu. Automaatselt valitud näitelause kvaliteet sõltub aluseks olnud korpus kvaliteedist – nii selle sisust, mahust kui ka märke-dusest (loe lähemalt Koppel 2020). Sellegipoolest on korpusmaterjali analüüs efektiivne viis, et iseseisvalt enda jaoks huvitavaid leksikaalseid ja grammatilisi kombinatsioone uurida.

Viidatud kirjandus

- Aston, Guy 1997. Enriching the learning environment: Corpora in ELT. Teaching and Language Corpora. Harlow: Longman, 51–64.
- Atkins, Sue, Michael Rundell 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.
- Fox, Gwyneth 1987. The case of examples. – Looking Up. An Account of the COBUILD Project in Lexical Computing. Toim. John Sinclair. London: Collins ELT, 137–149.
- Frankenberg-Garcia, Ana 2012. Learners' use of corpus examples. – International Journal of Lexicography, 25(3), 273–296.
- Frankenberg-Garcia, Ana 2014. The use of corpus examples for language comprehension and production. – ReCall, 26(2), 128–146.
- Gavioli, Laura 2005. Exploring Corpora for ESP Learners. (Studies in Corpus Linguistics 21.) Amsterdam: John Benjamins Publishing.
- Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. (Doktoritöö, Tallinna Ülikool). Tallinn: Tallinna Ülikool.
- Koppel, Kristina 2020. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. (Doktoritöö, Tartu Ülikool). Tartu: Tartu Ülikooli Kirjastus.
- Lew, Robert, Arleta Adamska-Sałaciak 2015. A case for bilingual learners' dictionaries. – ELT Journal, 69(1), 47–57.
- Sinclair, John 1987. Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. London: Collins ELT.

OK