Võistlus, valik ja vajadus keele muutumises: uuring korpusandmete, arvutusliku modelleerimise ja eksperimentide põhjal

Populaarteaduslikus võtmes ülevaade doktoritööst

Andres Karjus

1 Sissejuhatus: miks kõik keeled kogu aeg muutuvad?

Pidev muutumine on universaalne omadus, mis iseloomustab kõiki elavaid ehk aktiivselt kasutusel olevaid keeli kõikjal maailmas. See on ühelt poolt midagi väga loomulikku, ent kui korraks peatuda ja järele mõelda, siis ka midagi väga imelikku. Keelte muutumine ja hargnemine uuteks keelteks takistab suhtlust. Palju lihtsam oleks suhelda üle-lahe-naabritega, kui läänemeresoome algkeel poleks lahknenud eesti ja soome keeleks. Palju lihtsam oleks lugeda mõnesaja aasta vanuseid tekste, kui keele kirjaviis, sõnade tähendus ja lauseehitus selle ajaga ei muutuks. Erinevatel põlvkondadel oleks lihtsam omavahel suhelda, kui nad ei peaks vastastikku mõistatama uute laensõnade, ja teistpidi, vanade arhaismide tähenduste üle. Ometi inimkeeled nii ei toimi. Miks nad siis kogu aeg muutuvad?

Käesolev uurimus keskendub just sõnavara muutumisele. Mind huvitab võistlus uute ja vanade sõnade vahel, vaadatuna evolutsioonilisest perspektiivist. Erinevate keelte ajaloos leidub küllaldaselt juhtumeid, kus laenatud või loodud uus sõna ja selle levik põhjustab teise, sarnase või samatähendusliku vana sõna kadumise või vähemalt kasutuse olulise vähenemise. *aerodroomist* saab *lennuväli* ja *levimuusikaansambli* asemel öeldakse *bänd*. Siin võib leida paralleeli loodusliku valikuga: tugevam (keeles: parema kõlaga, efektiivsem või moekam) jääb tõenäolisemalt ellu. Samas ei vii kõikide uute sõnade keelde tulemine kaugeltki alati vanade sünonüümide välja suremiseni. *drink* elab kõrvuti *joogiga* ja *šoppamine poes käimisega*. Mind huvitab küsimus, miks osade uute sõnade tulemine keelde tekitab võistlust uue ja vana vormi vahel — samas mõnel teisel juhul võivad jääda kasutusse kõrvuti nii uus kui vana sarnase tähendusega sõna.

Käesoleva lühikese kokkuvõte eesmärk on anda ülevaade doktoritööst "Competition, selection and communicative need in language change: an investigation using corpora, computational modelling and experimentation", mis püüab muuhulgas leida vastust just sellele küsimusele. Hüpoteesin, et oluline tegur sellistes protsessides — lisaks paljudele muudele teguritele — on keelekasutajate väljendus- ja suhtlusvajadused. Mida olulisem on mingi teema ja selle nüansside kommunikeerimine kõnelejatele, seda tõenäolisem on, et nad hoiavad kasutuses mitut sarnast sõna. Kui teema muutub igapäevases suhtluses vähem oluliseks, jäävad välja ka detailsemad terminid — keel kui süsteem pürgib alati efektiivsuse suunas. Seda ideed — et keelte kuju ja sõnavara võiks peegeldada kõnelejate eelistusi ja vajadusi — on arutletud keeleteaduses üle sajandi (vt. Boas 1911; Sapir 1921; Martinet 1952; Coulmas

1989; Lupyan et al. 2010; Christensen et al. 2016).

Selle arutluskäigu paikapidavuse süstemaatiliseks uurimiseks kasutan kolme erinevat, kuid üksteist toetavat metoodilist lähenemist: arvutuslikud simulatsioonid, kommunikatsiooni-eksperimendid, kus katseisikutel tuleb lühikese aja jooksul omandada lihtsaid tehiskeeli ja nende abil omavahel suhelda, ning tehisintellektile toetuv tekstikaeve ajaloolistes keelekorpustes ja selle kaudu korjatud andmete statistiline analüüs, Korpusteks nimetatakse suuri teksti-andmebaase, mis võivad hõlmata nii mingisse žanri kuuluvaid spetsiifilisi tekste lühikesest ajaperioodist, või ka laia valikut tekste kümnete või lausa sadade aastate lõikes — seda viimast tüüpi andmebaase, suurusjärgus tuhanded tekstid ja sajad miljonid sõnad, kasutan laialdaselt käesolevas töös.

Peale sissejuhatuse ja kokkuvõtte koosneb töö teadusajakirjades avaldatud või avaldamisprotsessis olevatest artiklitest, millest annan järgnevalt ülevaate. Kõik artiklid on kirjutatud koostöös doktoritöö juhendajatega (kaks keeleteadlast ja üks füüsik), ning kõigi puhul on tegemist avatud ligipääsuga (*open access*) publikatsioonidega.

2 Evolutsiooniliste protsesside tuvastamise keerukusest ajaloolistes keelekorpustes

Esimene artikkel (avaldatud ajakirjas Glossa: a journal of general linguistics) võtab ette hiljuti geneetikute poolt keeleteadusele välja pakutud meetodi, millega peaks saama analüüsida evolutsioonilisi protsesse ajaloolistest keelekorpustest tuletatud sõnasageduste aegridades. Autorid toovad näiteks inglise keele mineviku, kus võistlevad regulaarne (*walk-walked*) ja ebaregulaarne paradigma (*run-ran*). Aja jooksul on mõnede tegusõnade minevikuvormide kasutus muutunud regulaarsemaks, teiste oma aga liikunud ebaregulaarse kasutuse poole. Ajaloolised korpuses saab mõõta vormide sagedust erinevatel ajaperioodidel, mis omakorda võimaldab hinnata nende kasutussagedust keele kasutajate hulgas.

Välja pakutud meetodi (*Fitness increment test* ehk FIT) eesmärk on testida, kas huvipakkuva sõnavormi sageduste aegrida kasvab või kahaneb valiku (*selection*) tõttu, ehk kõnelejad eelistavad teadlikult ühte varianti teisele, näiteks ebaregulaarset vormi regulaarsele — või juhusliku muutumise (*drift*) tõttu, ehk keelemuutus tuleneb puhtjuhuslike väikeste teisenduste ajas kuhjumisest, mida samuti keeleajaloos kindlasti ette tuleb. See on oluline erinevus: kui mingi paradigmaatiline muutus on suure tõenäosusega juhuslik, siis pole mõtet otsida paralleelvormi, millega uus sõna võinuks võistelda ja mida kaduma sundida.

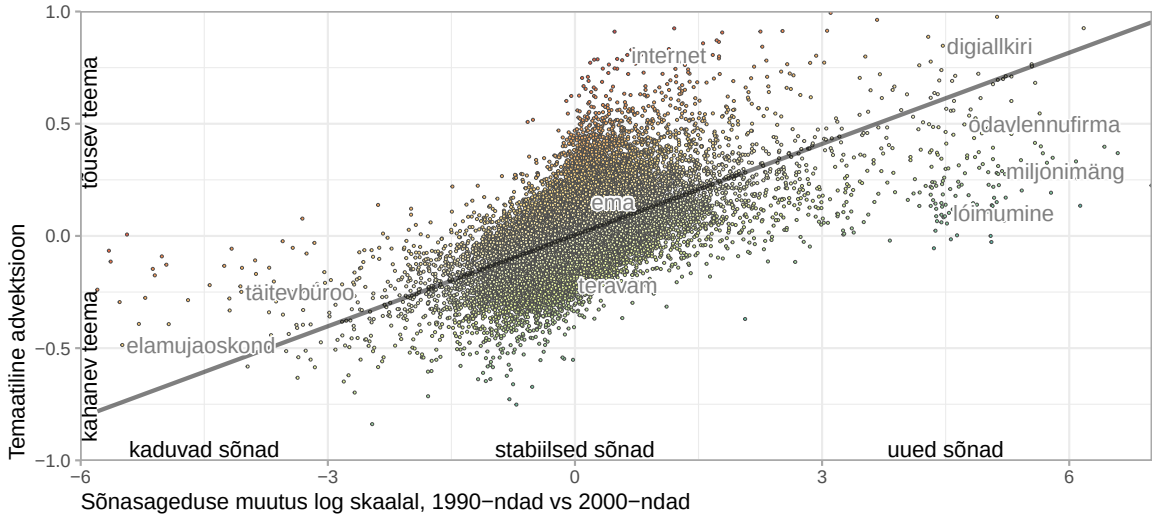
Analüüsin seda meetodit kahel viisil, alustades esialgse uurimistöö replitseerimisega laiendatud parameetriruumis, ja jätkates selle testimist samuti geneetikateadusest laenatud simulatsiooniparadigma (Wright-Fisher'i mudeli) abil. Lühidalt kokku võttes, leian, et FIT'i puhul on tegemist on paljulubava lähenemisega, ent ajalooliste keelekorpuste spetsiifika (sh keele- ja geneetiliste andmete erinevused) ja FIT'i tundlikkus aegridade konstrueerimisel tehtud valikutele — piiravad selle kasutusvõimalusi. Seetõttu ülejäänud doktoritöös ma seda meetodit siiski ei kasuta.

3 Temaatiliste lainetuste dünaamika mõõtmine keeleandmetes

Teises artiklis (avaldatud ajakirjas Language Dynamics and Change) töötan välja meetodi, mis võimaldab mõõta teemade (*topics*) kasvamist ja kahanemist keele ajaloos korpusandmete põhjal. Teemade all on siin mõeldud sisuliselt jututeemasid: mis on asjad, millest inimesed mingil ajastul kõige rohkem räägivad.

Näitan, et neid protsesse on võimalik piisavalt suurte ja esinduslike korpuste abil kvantifitseerida, kasutades keeletehnoloogiast laenatud lahendusi. Meetodi valiidsuse hindamiseks kasutan taas simulatsioone, seekord lähenemist, kus päris korpustesse viiakse sisse süstemaatilisi tehislikke muutusi.

Temaatiliste lainetuste kvantifitseerimine võimaldab paremini analüüsida mitmeid muid keeles toimuvaid muutusi. Üks rakendus on sõnasageduste muutumise uurimine. Sõnade kasutuse sagedust keelekogukonnas mõjutavad mitmed tegurid, sealhulgas näiteks sotsiolingvistilised aspektid nagu sõna moekus või kultuurilised konnotatsioonid, aga ka sõna effektiivsus (kõik sagedasemad sõnad on kõigis keeltes lühikesed!). Näitan käesolevas artiklis, et teemasageduste muutus (ehk temaatiline advektsioon) on üks sellistest teguritest, ja sobib hästi baasmudeliks, seletades keskmiselt 20-40% variatsioonist sõnasageduste muutustes kümnendite lõikes (vt joonis 1). Teisisõnu, teemad tõmbavad sõnu kaasa: näiteks sõja ajal kirjutavad ajalehed ja romaanid lahingutest, tankidest ja rahulepingutest; pandeemia ajal räägitakse viirustest ja tervishoiust, samas kui mingid muud teemad — ja seega ka mõned sõnad — muutuvad vähemolulisteks, mis pikema aja jooksul võib viia sõnade kadumiseni keelest.



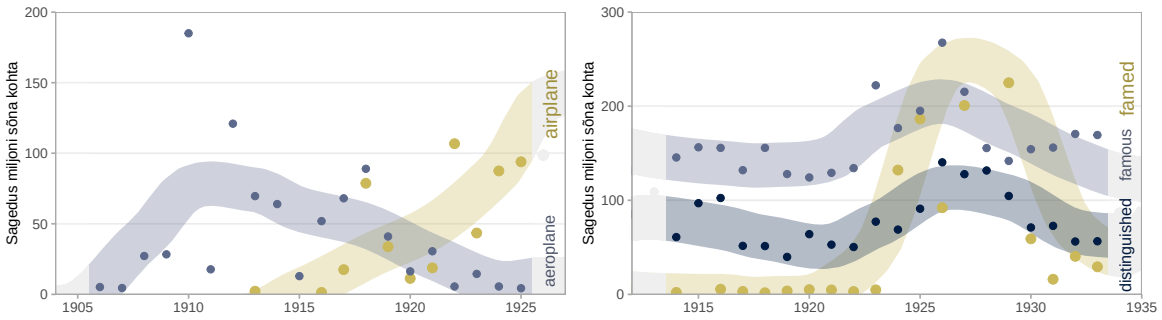
Joonis 1. Temaatiline advektsioon kui lähtepunkt sõnasageduste muutuste seletamiseks; siin eesti kirjakeele korpusandmete näitel. Iga täpp esindab ühte sõna (kokku üle 36000). Paljude sõnade kasutussagedus muutub kümnendite vältel vähe (joonisel horisontaalteljel keskel, 0 ümbruses asetsevad punktid). Sõnad, millega seotud teema kasvab, levivad ka ise suurema tõenäosusega (joonisel parem ülanurk, positiivne advektsioon vertikaalsel teljel), ja vastupidi. Sõnatäppide värv kajastab regressioonimudeli (hall diagonaal) jääkliikmeid, ehk tumedamad punktid on sõnad, mille muutust mudel nii hästi ei seleta. Temaatiline advektsioon kirjeldab siin ca. 33% sõnasageduste muutuste variatsioonist.

Teine rakendus, millele toetub ka järgmine artikkel, on kommunikatiivsete vajaduste kvantifitseerimine teemasageduste muutuste kaudu. Näitena kommunikatiivsest (ehk suhtlus-, väljendus-)vajadusest võib tuua olukorra, kus keeles eksisteerivad koos kaks või rohkem kontseptuaalselt väga sarnast sõna. Näiteks võiks kõiki arvuteid nimetada *arvutiteks*, aga ometi eristame me *(laua)arvutit* ja *sülearvutit*, mida omakorda nimetatakse olenevalt registrist ka *läptopiks*, *läpakaks* või hoopis *rüperaaliks*. Võib järeldada, et antud keelekogukonnas on arvutitehnikat puudutava sõnavara alamhulgaga (teemaga) seonduv kommunikatiivne vajadus piisavalt kõrge, et hoida kasutusel kõiki neid termineid, selle asemel et öelda alati lihtsalt *arvuti*. Teine hea näide on sugulussuhteid kirjeldav sõnavara, mille keerulisus varieerub maailma keeltes üpris palju, peegeldades seda, kui oluline või mitteoluline on erinevates ühiskondades sugulastele täpne viitamine (vt lähemalt Kemp et al. 2012). Näiteks on eesti keeles varem eristatud *onu* ja *lell*'e, kuid viimase kasutus on üldkeeles vähenenud.

Vajaduse hindamine temaatilise advektsiooni kaudu järgib järgnevat loogikat: kui inimesed räägivad ja kirjutavad mingist teemast rohkem kui varem, näiteks sõja ajal lahingutest — mis omakorda kajastub ajaloolistes korpustes, mis koosnevad valikust tekstidest igast kajastatud ajajärgust — siis järelikult on tegemist teemaga, mis on kõnelejatele kasvava olulisusega. Teisisõnu, antud teemaga assotsieerub kõrge kommunikatiivse vajaduse tase. Kui mingist muust teemast räägitakse vähem, on sellel ilmselt madalam vajaduse tase. Selle illustratsiooniks näitan artiklis inglise keele 20. sajandi uudissõnade valimi põhjal, et uued sõnad kipuvad keelde ilmuma just siis, kui nendega seotud teema on varasemaga võrreldes laiemalt kasutatav.

4 Kommunikatiivsed vajadused suunavad keelte muutumist

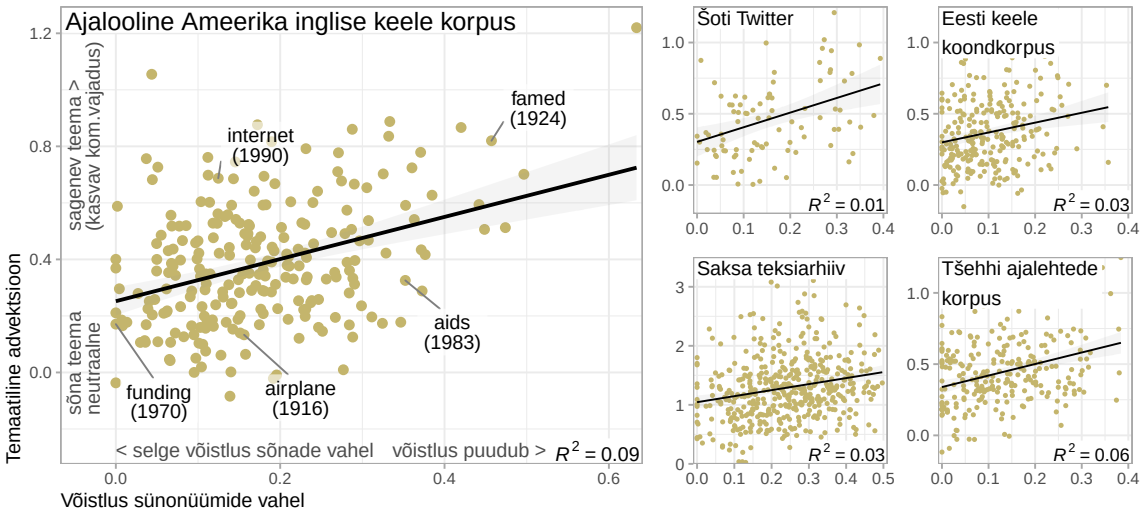
Kolmandas (hetkel retsenseerimisel olevas) artiklis käsitlen hüpoteesi, et muutused sõnavaras (ja keelemuutused laiemalt) toimuvad vähemalt osaliselt keelekogukonna kommunikatiivsete vajaduste muutumise tõttu. Täpsemalt uurin võistlust sõnade vahel: mõne uue sõna levik sunnib vanema sünonüümi taanduma, samas kui mõnel teisel juhul jäävad nii uus kui vana koos kasutusse (vt joonis 2). See on variatsioon, mida vajadus võiks selgitada: kui keele kasutajatel on vaja tihti eristada mingi kontseptsiooni tähendusvarjundeid — näiteks *poodlemise*, *ostlemise* ja *šoppamise* erinevust — siis on suurem tõenäosus, et mitu varianti jäävad kasutusse. Kui mingi teema või semantilise väljaga seonduv kommunikatiivne vajadus on madal, siis on tõenäolisem, et osa lähisünonüümidest ja täpsustavatest terminitest ära kaob: näiteks kui suurem osa keele kasutajatest enam *vikatiga* ei niida, siis pole imestada, et *rauts* ja *lüsi* vaikselt käibelt kaovad.



Joonis 2. Sarnaste sõnade võistlus ja võistluse puudumine Ameerika inglise keele korpuses COHA. Kui kahest viisist väljendada "lennukit" jääb järele vaid üks (vasakpoolne joonis), siis kaks erinevat "kuulus" tähendusega sõna eksisteerivad rahus koos pikema aja vältel. Näitan käesolevas artiklis, et kommunikatiivsed vajadused on tegur, mis seletab sellist variatsiooni (statistilises mudelis on kontrollitud kõikvõimalikud muud leksikostatistilised tegurid, sh asjaolu, et kaks sarnast sõnavormi nagu *aeroplane* ja *airplane* võivad olla lihtsalt kaks erinevat kirjapilti).

Selle hüpoteesi testimiseks ajalooliste andmete peal on kõigepealt vaja leida korpustest valim sõnu, mis võiks mõne muu tähenduselt sarnaselt sõnaga võistelda. Kasutan selleks tehisintellekti-põhist lahendust, mis tuletab konteksti põhjal välja kõikide sõnade omavahelised tähendus-sarnasused igal korpusperioodil, ja edasi, selle uurimistöö jaoks välja töötatud statistilist lahendust, mis hindab võistluse tõenäosust iga huvipakkuva sõna ja tema tähendus-naabrite vahel. Rakendan seda metoodikat andmetele viiest erinevast korpusest: Ameerika inglise, saksa, tšehhi ja eesti kirjakeele korpused, suuruses kümnendid kuni sajandid, ja aastase perioodi vältel kaevandatud šoti-inglise Twitteri-korpus (vt joonis 3). Analüüs näitab, et eespool kirjeldatud advektsiooni-meetrik korrelleerub võistlust hindava meetrikuga, kinnitades hüpoteesi: kõrgem kommunikatiivse vajaduse tase ennustab sarnaste sõnade kooseksisteerimist, madalama vajaduse puhul aga on suurem tõenäosus, et uus sõna võistleb vanema

sarnase sõnaga ja võib viimase keelest välja süüa.



Joonis 3. Valim sageduses kiiresti kasvanud sõnu viiest korpusest. Horisontaalsel teljel on võistluse "kaugust" hindav leksikostatistiline tunnus: nulli-lähedase väärtusega sõnad põhjustavad oma kiire levikuga mõne lähedase sõna kadumise; suurema väärtusega sõnade puhul selline selge võistlus puudub. Verikaalsel teljel on temaatiline advektsioon, mis kaudselt hindab kommunikatiivsete vajaduste muutust keelekogukonnas. Need kaks tunnust korreleeruvad: kõrgendatud vajaduste korral on tõenäolisem, et uus kiiresti kasutuses leviv sõna ja juba keeles olemas olevad sarnased sõnad eksisteerivad kõrvuti edasi; madalam vajadus ennustab, et uue sõna kasutuselevõtt viib mõne sünonüümi kadumisele (R^2-väärtus joonise nurgas näitab, palju advektsioon seletab võistlustunnuse varieeruvusest, skaalal 0 kuni 1, pärast kõikide muude kontrolltunnuste arvesse võtmist).

5 Kommunikatiivsete vajaduste roll koleksifikatsiooni mustrite dünaamikas

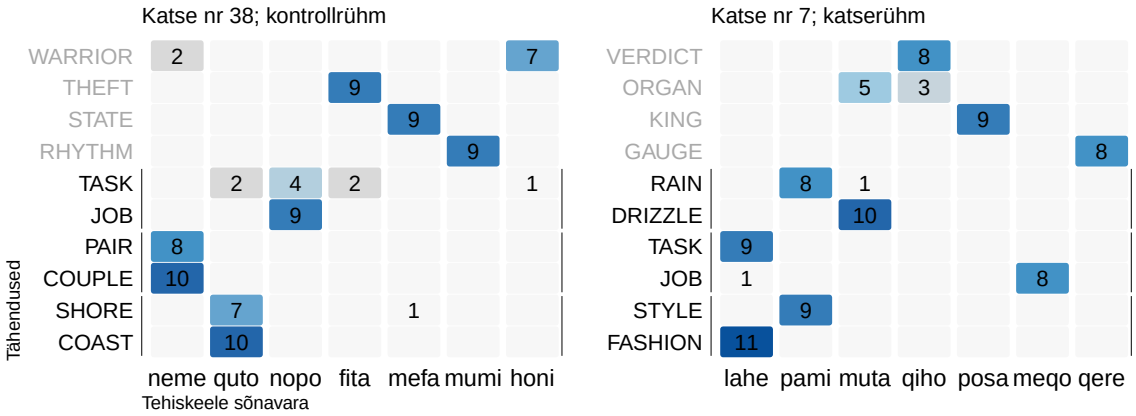
Neljas artikkel (hetkel veel avaldamata) võtab ette sõnavara muutumisdünaamika koleksifikatsiooni nurga alt, täiendades kolmanda artikli tulemusi. Koleksifikatsioon on käepärane termin, mis kirjeldab olukorda, kus mingi sõna tähendus katab kahte või rohkemat kontseptsiooni, mida mõnes teises keeles väljendatakse eraldiseisvate sõnadega. Näiteks eesti keeles on mitu sõna erineval kujul külmunud vee kohta: *lumi*, *jää*, *lörts* jne. On aga palju keeli, kus lumele ja jääle viidatakse ühe ja sama sõnaga — lisaks kipuvad need olema keeled, mida kõneldakse troopilise kliimaga kohtades (olles taas näide sellest, kuidas nii väljendusvajadused kui nende puudumine suunavad sõnavara arengut; vt Regier et al. 2016). Näiteks Nigeerias kõneldava hausa keele sõna *ƙàn ƙáráa* "koleksifitseerib" jää ja lume kontseptsiooni.

Hiljuti avaldatud artiklis käsitlevad Xu et al. (2020) koleksifikatsiooni-mustreid sadades maailma keeltes ja näitavad, et tüpoloogilised tendentsid on paljuski ennustatavad kontseptuaalsete tähendussarnasuste kaudu. Lisaks pakuvad autorid, et kõrgendatud kommunikatiivsed vajadused võivad samas neid mustreid mõjutada: kui mingeid isegi väga sarnaseid tähendusvarjundeid on vaja suhtluses tihti eristada, siis pole praktiline neid mitte koleksifitseerida, vaid viidata eraldi sõnadega. Tegemist on loogilise arutluskäigiga, ent süstemaatiliselt seda hüpoteesi testitud ei ole.

Minu artikli eesmärk on seda teha, kasutades virtuaalkeskkonnas toimuvaid mängustatud tehiskeele-katseid, mida on kognitiivteadustes laialdaselt kasutatud inimeste mõtlemist ja suhtlemist suunavate protsesside uurimiseks (vt. Kirby et al. 2008; Winters et al. 2015). Selles eksperimentaalses

paradigmas antakse katsealuste paaridele ette väike katse jaoks tehislikult loodud keel — siinjuhul lihtsalt sõnavara ilma grammatika ja süntaksita — ja juhis selles keeles üritada üksteisele sõnumeid saata ja nende tähendus ära arvata.

Katses osalejatel tuli suhtluse käigus siduda ise sõnavormid ette antud tähendustega. Mida katsealused ei teadnud, oli see, et tähendusi oli rohkem kui sõnavorme — sellisel juhul eeldab edukas suhtlemine, et mõnda sõna tuleb kasutada rohkem kui ühele tähendusele viitamiseks (vt joonis 4). Katsete tulemused näitavad, et kui ühtegi tähendus-paari, mida tuleb eristada, ei tule ette sagedamini kui ühtegi teist, siis eelistatakse sarnaste kontseptide (nt *töö* ja *ülesanne*) koleksifitseerimist (kinnitades Xu et al. (2020) tüpoloogilisi tulemusi; vasak paneel joonisel 4). Kui aga mingite tähendustega seonduv kommunikatiivne vajadus on kõrge, ehk suhtluses on pidevalt vaja eristada kahte muidu sarnast kontsepti — seda vajadust saab stiimuli manipuleerimisega tekitada — siis kujuneb välja sõnavara, kus sellistel tähendustel on kas ühemõttelised eraldi sõnavormid (*meqo* märgib paari number 7 jaoks alati tähendust JOB; vt parem paneel joonisel 4), või koleksifitseeringud, mis ei tekita arusaamatusi (*pami* viitab tähendustele FASHION ja RAIN, mis pea kunagi koos ei esine ja eristamist ei vaja).



Joonis 4. Kahe katse tulemused. Tähendused ehk võimalikud sõnumid, mida osalejad ekraanil nägid ja said üksteisele saata, on inglise keeles, kuna katsealused olid Edinburghi Ülikooli tudengid. Tehiskeeleks on eksperimentide jaoks algoritmiliselt genereeritud "sõnavara", mis on ehitatud nii, et see oleks inglise keelest võimalikult erinev, ent siiski koosneks loomuliku kõlaga silpidest. Ruutudes olevad numbrid näitavad, mitu korda osalejate paar kasutas (pärast esialgset treenimisfaasi) mingit vormi mingi tähenduse väljendamiseks. Näiteks paari number 38 (vasakpoolne paneel) mõlemad liikmed kasutasid alati vormi *mumi* ainult tähenduse RYTHM ("rütm") kommunikeerimiseks. Samas *neme* väljendab kahte sarnast tähendust, PAIR ja COUPLE ("paar"). Katserühma kuulunud paar (7, paremal) on aga koleksifitseerinud tähendused nii, et tähendused, mida tihti tuli eristada (nt FASHION ja STYLE) ei oleks väljendatud sama sõnavormiga.

Lisaks katsele testisin analoogsetet diakroonilist hüpoteesi ajaloolise keelekorpuse andmetel, näidates et kõrgem kommunikatiivne vajadus (siin taas tuletatud temaatilise advektsiooni kaudu) ennustab sõnade poolest rikkamate tähendusväljade teket. Ent nii psühholingvistilised katsed kui korpused annavad vaid osalise pildi keele olemusest. Katsed võimaldavad kontrollida keele ja selle kasutust mõjutavaid variaableid, aga on paratamatult tehislikud situatsioonid, kus kõnelejad ei pruugi käituda nagu tavasuhtluses. Korpused annavad ülevaatliku pildi keele kasutusest ajas ja ruumis, aga kajastavad ainult piiratud hõredat valimit keele rikkusest ja varieeruvusest; lisaks on statistika ja masinõppe põhjal leitud tulemused paratamatult ainult ebatäpsed hinnangud tegelikest protsessidest. Kui aga mõlemad lähenemised, katsed ja korpusuuringud, viivad võrreldavatele tulemustele — nagu selles uurimuses näitan — lisab see kindlust, et leitud on midagi käegakatsutavat keele olemuse ja ehituse kohta.

6 Kokkuvõte: sest neil on vaja muutuda

Selles doktoritöös uurisin, kuidas keelekasutajate kommunikatiivsed vajadused — nii tuletatuna korpusandmetest kui testituna labori-eksperimentides — suunavad keelte muutumist. Töötasin selleks välja mitu arvutuslikku meetodit nende protsesside modelleerimiseks, ja kasutasin varianti tehiskeelset suhtlust hõlmavast eksperimentaalsest paradigmast. Püüdsin leida (vähemalt osalist) vastust ühele keeleteaduse fundamentaalsemale küsimusele: miks kõik keeled kogu aeg muutuvad? Põhjusi on selleks palju. Oma ma uurimistöös kontrollin süstemaatiliselt leksikostatistilisi tegureid nagu sõnavormide kuju ja sarnasus, sagedus jms; lisaks mõjutavad seda kindlasti paljud sotsio- ja psühholingvistilist laadi põhjused, mida see töö ei hõlma. Minu eesmärk on aga näidata, et üks põhjus, miks keeled muutuvad, on see, et neil on vaja muutuda — et püsida oma kõnelejatele efektiivsete kommunikatsiooni-tööriistadena muutuvas maailmas. Teistpidi, keeled, mis ei muutu, muutuvad kasutuks. Siit tuleneb muuhulgas ka võimalik lahendus ühele ühiskonnas tihti arutelu leidvale küsimusele, kas uued põlvkonnad mitte ei “riku" keelt, näiteks laenates uusi sõnu, muutes keele häädlust või vormide kasutust. Käesoleva töö tulemuste põhjal võib öelda, et on tõenäolisem, et nad lihtsalt lasevad keelel muutuda — vastavalt oma kommunikatiivsetele vajadustele.

Viidatud kirjandus

Boas, Franz (1911). *The Mind of Primitive Man*. The Macmillan Company.

Christensen, Peer, Riccardo Fusaroli, and Kristian Tylén (2016). "Environmental Constraints Shaping Constituent Order in Emerging Communication Systems: Structural Iconicity, Interactive Alignment and Conventionalization". *Cognition* 146, pp. 67–80. DOI: 10.1016/j.cognition.2015.09.004.

Coulmas, Florian (1989). "Language Adaptation". *Language Adaptation*. Ed. by Florian Coulmas. Cambridge University Press, pp. 1–25.

Kemp, Charles and Terry Regier (2012). "Kinship Categories across Languages Reflect General Communicative Principles". *Science (New York, N.Y.)* 336.6084, pp. 1049–1054. DOI: 10.1126/science.1218811.

Kirby, Simon, Hannah Cornish, and Kenny Smith (2008). "Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language". *Proceedings of the National Academy of Sciences* 105.31, pp. 10681–10686. DOI: 10.1073/pnas.0707835105.

Lupyan, Gary and Rick Dale (2010). "Language Structure Is Partly Determined by Social Structure". *PLOS ONE* 5.1, pp. 1–10. DOI: 10.1371/journal.pone.0008559.

Martinet, André (1952). "Function, Structure, and Sound Change". *WORD* 8.1, pp. 1–32. DOI: 10.1080/00437956.1952.11659416.

Regier, Terry, Alexandra Carstensen, and Charles Kemp (2016). "Languages Support Efficient Communication about the Environment: Words for Snow Revisited". *PLOS ONE* 11.4, pp. 1–17. DOI: 10.1371/journal.pone.0151138.

Sapir, Edward (1921). *Language. An Introduction to the Study of Speech*. New York: Harcourt, Brace and Company.

Winters, James, Simon Kirby, and Kenny Smith (2015). "Languages Adapt to Their Contextual Niche". *Language and Cognition* 7.3, pp. 415–449. DOI: 10.1017/langcog.2014.35.

Xu, Yang, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan (2020). "Conceptual Relations Predict Colexification across Languages". *Cognition* 201, p. 104280. DOI: 10.1016/j.cognition.2020.104280.