https://doi.org/10.7592/VK2.2006.laineste

Liisi Laineste

EESTI ANEKDOOTIDE DIGITAALNE ANDMEBAAS

Teesid: Eesti anekdootide digitaalne andmebaas lihtsustab uurija tööd selle žanri sünkroonilisel uurimisel 20. sajandi teise poole kontekstis või mõne teema lõikes ning pakub huvitavat ajaviidet anekdoodihuvilisele, kes leiab sealt muidu eri kohtades asuvad või asunud anekdoodid. Artikkel kirjeldab, millised on digitaalse kogu ja kategoriseerimise puudujäägid ja eelised. Pikemalt peatutakse materjali liigitamisel, mis peaks lähitulevikus võimaldama sooritada praegusest põhjalikumaid sisulisi otsinguid. Tutvustatakse loodud süsteemi kui üht lahendust tekstide organiseerimiseks. Edasise tegevuse põhieesmärk on kategoriseerimissüsteemi täiustamine ja senikogutud materjali kategooriatesse paigutamine, arendatakse ka verbaalse huumori teoorial põhinevat kategoriseerimissüsteemi, mis on erinev igal anekdooditeksti tasandil.

Märksõnad: huumor, kategoriseerimine, tekstiline andmebaas

Sissejuhatus

Arvutiteaduse arenguga käsikäes on kasvanud ka inimeste nõudmised (uurimis)töö lihtsustamisele. Kui varem oli arvuti abiks kvantitatiivset arvutust vajavatele reaalteadustele, siis nüüd nõuavad ka humanitaarteadlased võimalusi materjali digiteeritud säilitamiseks, klassifitseerimiseks ja analüüsiks. Ühest küljest tähendab see olemasoleva materjali digiteerimist, teisalt edasise kogumis- ja arhiveerimistöö ning analüüsi arvutipõhiseks muutmist (Fischer 1994). Eesti naljade andmebaas puutub kokku esimesega neist ülesannetest (kuigi näiteks edasine uurimustöö anekdoodirääkimise populaarsusest – vastandina anekdoodikirjutamise populaarsusele – nõuab ilmselgelt tutvumist ka teise suunaga; samuti hõlmab naljade uurimine digiteeritud ma-

terjali kvantitatiivset analüüsi, mis ei kuulu käesoleva artikli probleemistikku).

Folkloristi töö suurte tekstikorpustega kuulus nn Soome koolkonna põhitegevuste hulka, kuid vajadus töötada suurema hulga tekstidega (*versus* üksikjuhtumi analüüsiga) püsib. Peale mikrostruktuuri kirjeldamise ja kontekstuaalse analüüsi tuleb sageli uurimusele kasuks makrostruktuuri nägemine, materjali esialgne üldine kirjeldamine elementaarse statistika abil. Selleks peab digiteeritud materjal omama n-ö loendatavaid parameetreid, ehk teisisõnu: arvuti peab "mõistma" võimalikult täpselt tekstilise materjali sisu (või peab see olema liigitatud sisulisest aspektist lähtuvalt võimalikult detailselt, seda kas uurijapoolsete pingutuste tulemusena või arvuti abiga).

Praegune (Eesti) arvutiteaduse seis ei võimalda eestikeelse teksti täielikult digiteeritud sisulist kategoriseerimist, sest kuigi arvutile teksti "mõistmise" õpetamine teeb edusamme, on esialgu nõudlus nt tekstikorpuste kategoriseerimist abistava programmi järele veel peaaegu olematu. Seetõttu on kaasaegsete anekdootide praegune andmebaas vaid üks võimalik katsetus ja mitte sugugi ainuõige tee digiteeritud tekstikogu loomiseks.

Käesolev ülevaade tutvustab Eesti Kirjandusmuuseumi kaasaegsete anekdootide digitaalse andmebaasi materjali ja sellega tehtud tööd ning kirjeldab eesmärke ja arengusuundi, samuti seda, millised on antud digitaalse kogu puudujäägid ja milline näeb ideaalpildis välja kogumis-, arhiveerimis- ja toimetamisprotsess, kui selle juures on abiks arvuti. Peatun ka üldiselt digitaalse kogu võimalustel materjali kategoriseerimise vallas.

Kaasaegseks anekdoodiks või naljaks nimetatakse siinkohal folkloorset teksti, lühikest puänteeritud nalja, mis eristub vanast pikemast puänteerimata rahvanaljandist. Enamasti erinevad need kaks žanrialaliiki ka sisuliselt, kuid eristamise mõttes on esikohal siiski vormilised muutused, mis naljandiga aja jooksul on aset

leidnud. Anekdootide andmebaas sisaldab vaid uuemaid, puänteeritud nalju.

Materjal

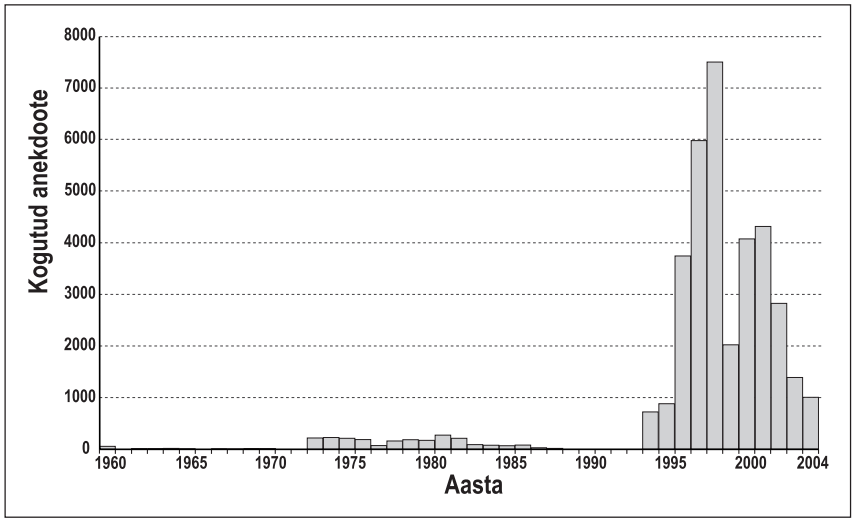
Eesti kaasaegseid anekdoote on Eesti Kirjandusmuuseumi digiteeritud kogus umbes 50 000 ja nende hulk kasvab keskmiselt paarikümne teksti võrra päevas (koos korduste ja variantidega, mis lisatakse siiski andmebaasi). Anekdoodikogu eesmärgiks oli talletada 1990. aastate teisel poolel seoses interneti levikuga hüppeliselt kasvanud naljade kirjutamise, saatmise ja ilmselt ka lugemise praktika, mis väljendus näiteks Sünerkomi *Jokebook*i *Meie Naljaraamat* (www.zzz.ee/jokes) fenomenis. See viitab anekdoodižanri tähtsale rollile internetikommunikatsioonis.

Internetikeskkonnas asuv aktiivne jututoa-tüüpi "plank", kuhu kõik võisid nalju kirjutada, teiste omi kommenteerida, naljale naljaga vastata vms, saavutas ülisuure populaarsuse ja avas seninägemata võimaluse jäädvustada ning vormistada uurimiseks sobivasse formaati midagi suulise anekdoodirääkimise sarnast. Lisaks paberkartoteegile otsustati kogu avaldada ka digitaalse andmebaasina (sh otsinguna), sest internetis leiduvate anekdootide esmane ja loomulik olek on digiteeritud olek. Pealegi on materjali sedavõrd palju, et digitaalne andmebaas oli peaaegu ainuke võimalus ülevaatlikkuse säilitamiseks.

Anekdoodiandmebaas lihtsustab uurija tööd selle žanri sünkroonilisel uurimisel 20. sajandi teise poole kontekstis või mõne teema lõikes, samuti pakub huvitavat ajaviidet lihtsalt anekdoodihuvilisele internetikasutajale, kes leiab sealt muidu eri kohtades asuvad või asunud anekdoodid. Hetkel kannab digitaalne kogu veel selle praeguse arendaja, st käesoleva artikli autori huvidest tulenevat ilmet (nt täielikult on kategoriseeritud vaid etnilised anekdoodid). Selline kallutatus on probleem, mida on teravalt tun-

netatud just tekstilise materjali digiteeritud kogudega töötamisel (Coffey & Atkinson 1996), sest kuigi kategoriseerimisüsteem püütakse luua võimalikult erinevaid huvisid arvestav ja mitmetasandiline, on see siiski välja mõeldud peamiselt ühe grupi uurijate poolt, seega dikteerib selle ülesehitus edasiste uurimisküsimuste ringi. Andmebaasi koostamise võimalusi ja kvalitatiivseid meetodeid, mis lubaksid uurijal formuleerida täpselt teda huvitavad hüpoteesid ja uurimisküsimused, mitte lastes neid dikteerida baasi struktuuril, kirjeldavad näiteks Robert G. Burgess (1995) ning Eben A. Weitzman ja Matthew B. Miles (1994).

Kaasaegsete anekdootide andmebaas koosneb peamiselt internetis leiduvatest (www.delfi.ee/jokes) või sealt leitud (koduleheküljegedel asunud anekdoodikogud) naljadest, mõned allikatest, näiteks Jüri Viikbergi kogu (Viikberg 1997) on ilmunud trükis. Andmebaas katab ajavahemikku 1960. aastatest tänase päevani. Materjal pole jaotunud kaugeltki ühtlaselt (vt joonis 1, millel on esitatud summaarselt salvestatud anekdootide hulk aastate kaupa).



Joonis 1. Materjali jaotus anekdootide andmebaasis kogumisaastate lõikes.

Kõige arvukamalt on anekdoote *Meie Naljaraamatu* algusaegadest (asus aadressil www.zzz.ee/jokes) aastatest 1996–1998, sellele järgnev tõus kuulub internetiportaali www.delfi.ee/jokes (*Delfi Naljaleht*) avamisaastasse (naljalehekülg avati anekdootide saatmiseks, hindamiseks ja kommentaarideks 2000. aasta aprillis). 1990. aastate alguses on graafikul lünk, kuigi sel ajal korraldatud koolipärimuse kogumise käigus saadi muu hulgas ka hulgaliselt anekdoote. Neid pole aga veel muust materjalist eraldatud (v.a keerdküsimused, mida võib lugeda ka anekdootide alla kuuluvaks) ega digiteeritud.

Ebaühtlane jaotus muudab tüübi või teema esinemise arvukusel põhinevad arvutustulemused kallutatuks, kuid selle kompenseerib mõnevõrra varasemate (enne 1990. aastaid kogutud) anekdootide representatiivsus – suur hulk erinevaid tüüpe, mõned paralleelvariandid, väga vähe kordusi. Enamasti näitavad graafikud siiski üsna objektiivset pilti.

Pidevalt täienev digitaalne andmebaas, mis oleks internetis kättesaadav ja käepärane nii anekdoodihuvilisele arvutikasutajale kui ka folklooriuurijale, on parim vorm anekdootide aktiivse ja elusa traditsiooni kogumiseks, säilitamiseks ja eksponeerimiseks.

Kategoriseerimine

Kõikidele sissekannetele lisatakse juba internetist alla laadides juurde info, mis puudutab nende allikat ja kogumise (või naljaportaali saatmise) aega. See võimaldab vastata küsimustele anekdootide populaarsuse, kogumise aktiivsuse ja ulatuse kohta. Kuid detailsemaks uuringuks mõne teema siseselt (nt prantslaste- või Juku-anekdoodid, nende põhiteemad, teised tegelased, vormilised eripärad vm uurijat huvitav) on vaja kategoriseerida materjal ka sisuliselt.

Enne kategoriseerimissüsteemi loomist pöördusin Ameerika, Suurbritannia ja Soome huumoriuurijate poole, et jõuda selgusele, kas on mingit levinud süsteemi, mida ka Eestis tasuks järgida. See oleks vajalik eriti juhul, kui uurijal on eesmärgiks kultuuridevaheline uuri-

mistöö. Erinevate maade anekdootide võrdlemine on palju lihtsam ja tulemuslikum, kui taksonoomiad seda soodustavad. Kahjuks selgus, et iga maa arhiivis kasutatakse erinevat süsteemi, mis enamasti lähtuvad andmebaasi looja isiklikust uurimisfookusest. Enamasti ei looda andmebaase üldiseks kasutamiseks ega pikemaajaliseks arhiveerimiseks. Huumoriuurijad, kes kasutavad uurimismaterjalina kirjutatud anekdoote, piirduvad anekdoodikogude ja/või internetiportaalide läbivaatamisega ega kogu kõike, mida leiavad. Näiteks Alan Dundese koostatud tuntud ja olemasolevatest ilmselt kõige põhjalikum (ameerika) anekdootide arhiiv jagab anekdoodid seitsmesse suuremasse alaliikideks jagunevasse kategooriasse:

1) anekdooditsüklid (*knock-knock*-naljad, anekdoodid elektripirni keeramisest jms);
2) mittenarratiivsed naljad, üherealised naljad, laused, mõistatuse vormis naljad;
3) narratiivsed naljad, mis ei kuulu anekdooditsüklitesse;
4) nn karvase koera lood (*shaggy dog stories*);
5) obstsöönsed naljad (seksuaalsed, skatoloogilised jms);
6) vembud ja vingerpussid (*practical jokes*);
7) etnilised, rassistlikud, ka poliitilised naljad (*blason populaire*).

Nimekirjast jäävad silma A. Dundese enda artiklitest tuntud huviobjektid – tema huumoriuuringud räägivad naljatsüklitest, obstsöönsetest naljadest, poliitilisest ja etnilisest huumorist (Dundes 1985, 1987; Dundes & Banc 1986 jpm). N-ö tavaliste anekdootide jaoks tundub olevat siin vaid kolmas kategooria, eraldi asetatud on seksuaalse (ning skatoloogilise vm obstsöönse) sisuga ja etnilisi rühmi pilavad naljad.

Järelpärimiste põhjal sai selgeks, et konkreetse eeskuju puudumise tõttu tuleb liigitussüsteem ise koostada. Järgnevalt kirjeldan, milliseid põhimõtteid järgides on loodud kategooriad ja kuidas toimub sisuline kategoriseerimine digiteeritud andmemassiivi sees.

Kategooriad on loodud juhuvalimisse sattunud anekdootide sisu analüüsimise tulemusel. Valim hõlmas u 3000 anekdooditeksti. Kategoriseerimissüsteemi väljatöötamiseks kasutasin paberkartoteeki, et anekdoodid oleksid reaalselt silma ees ja naljad, millele esialgu sobivat kategooriat ei leitud, võiks hiljem uuesti läbi vaadata ja ära paigutada.

Põhiliseks eristusparameetriks on võetud tegelase isik, sest see on naljas tavaliselt kõige selgemini määratletud (nt Juku; *vs* situatsioon, mis võib olla konkreetselt nimetatud või markeeritud ainult vihjeliselt, nt sissejuhatusega: *Kohtunik kohtualusele:...*). Kuid on ka situatsioonidel põhinev kategooria selleks puhuks, kui tegelasi pole otseselt nimetatud, ning liigitusvõimalus muul sisulisel alusel, kui selline kategooria tundub pakkuvat huvi uurijale või on esinduslikult ja selgelt eristatav mingi muu tunnuse poolest (nt absurdihuumor).

Liigitussüsteem laenas arenduspõhimõtteid hetkel eesrindlikemast huumoriuurimise valdkonnast, verbaalse huumori teooriast (GTVH, *General Theory of Verbal Humor*; vt Attardo & Raskin 1991; Attardo 1994; Ruch & Attardo & Raskin 1993; Hempelmann 2004; Brône & Fayaerts 2004). Selle kohaselt on verbaalseid nalju võimalik kirjeldada kuuel tasandil (nende hierarhia ja asetuse kohta üksmeel veel puudub): keel (LA, *language*), narratiivne strateegia (NS, *narrative strategy*), objekt (TA, *target*), situatsioon (SI, *situation*), loogikamehhanism (LM, *logical mechanism*), ja skriptivastandus (SO, *script opposition*). Lisaks tegelasele (GTVH sõnastuses objekt, *target*) võiksid andmebaasi lõplikus variandis olla anekdoodid eristatud ka ülejäänud mainitud tunnuste alusel (v.a keelelisel (LA) tasandil, mis tähendaks oma kategooriat peaaegu igale naljale). See on ilmselt esimene katse kasutada GTVH süsteemi naljade klassifitseerimisel, ja kui lähenemine peaks osutuma viljakaks, võib see olla aluseks ka teiste maade anekdoodibaaside loomisel.

Esimese sammuna eraldasin tegelaste jaotuse suuremad põhikategooriad, seejärel alamkategooriad ja nende alajaotused (viimased

juhul, kui tundus, et selline eristamine võib hilisemat andmebaasikasutajat huvitada või kui mingi jaotus hakkas silma arvukuse poolest). Alajaotuse vajalikkus kerkis esile näiteks abielurikkumisteemaliste lugude alamkategoorias: armukese omamisest rääkivate naljade hulgas eristuvad need, mis algavad sellega, et mees tuleb (varem) komandeeringust koju. Selline arvukas alajaotus nõuab äramärkimist ka kategoriseerimissüsteemis. Teised armukese-anekdoodid kuuluvad aga mehe-naise suhete üldisema kategooria abielurikkumise alamkategooriasse. Vajadusel lisatakse kategoriseerimise käigus uusi alamkategooriaid, kui koguneb teistest mingi tunnuse põhjal selgelt eristuvaid anekdoote, mida algsest valimist välja ei tulnud (nt *Mehe-naise suhted* > *Abielurikkumine* > *Armuke kapis*).

Loomulikult tuleb lubada anekdooditeksti kuulumist mitmesse kategooriasse samaaegselt. Tekstilised andmebaasid jagunevad kolme suuremasse rühma: ühemõõtmelised, hierarhilised ja relatsioonilised (Burnard 1987). Vaid viimane neist võimaldab kirjeldada tekstidevahelisi seoseid, lubab liigitada ühte teksti samaaegselt mitmeti ning annab kasutajale sel viisil kõige mitmekesisema ja anekdooditekstide loomulikku mitmetasandilisust arvestava ülevaate.

Andmebaasi relatsiooniline ülesehitus nõuab uurijalt ka vähem tööd ja vastutust, sest kategoriseerija ei pea oma arusaamise kohaselt ja vägivaldselt valima teksti sisulist poolt esindama vaid ühte märksõna (nt Juku-anekdoodi tegelane pole ainult Juku, vaid ka õpetaja, vanemad vm; situatsioon võib olla seks, õppimine jne). Selline tekstide mitmeti märgistatus osutub vajalikuks siis, kui uurija või huviline andmebaasist reaalselt mõnda anekdooti otsima hakkab. Et olla kindel, et soovitud anekdoot ka leidub (kui see kogus sisaldub), peab selleni jõudmiseks olema võimalikult palju erinevaid lähenemisteid – näiteks peaksid lühikeses anekdoodis *Elevant ja sipelgas olid luurel. Korraga sosistas sipelgas: "Rooma üksi edasi, mind on märgatud!"* olema eristatud kategooriad *Loomad* > *Elevant*, *Loomad* > *Sipelgas*, aga ka *Absurd*.

Mõni suurem kategooria võib olla jaotatud alajaotusteks mitmel viisi üheaegselt. Näiteks etniliste anekdootide (erinevaid rahvaid pilavate naljade) seas liigitati anekdoodid lisaks rahvuse-kategooriatele naljadeks, kus juttu tegelase ihnsusest, rumalusest, seksuaalsetest kommetest, mustusest või alkoholitarbimisest. Järgnev anekdoot näiteks kuulub kategooriatesse *Rahvused* > *Mustanahaline* ja *Rahvused* > *Mustus*).

> *Miks neegrikirstu kannavad ainult 2 meest? Sest prügikastil pole rohkem sangu* (Tanel Mägi kogu).

See, millised on etniliste anekdootide põhiteemad (kas konkreetse etnilise rühma kohta räägitakse pigem nalju nende ihnsusest kui lollusest vms), kõneleb suhtumisest naljaobjektiks olevasse rahvasse ja annab seega huvitavat teavet kultuurist, kus see anekdoot liigub (nt viimane anekdoot pole niivõrd indikaatoriks mustanahaliste ameeriklaste mustuse, kuivõrd valgete ameeriklaste hügieeniobsessiooni kohta – Davies 1990, 2002 ja mujal).

Sageli tuleb ette anekdoote, millel on erinevad tegelased, kuid iseenesest on tegu sama või sarnase skriptiga. Paralleelide toomine nende naljade vahel on tähtis, sest peamiselt ikkagi tegelaste (või mõnel harval juhul, kui tegelane puudub, situatsiooni) põhjal kategoriseeritud materjali sees otsimine ei too uurijale kõiki selle skriptiga anekdoote, kuigi need võivad lisada väärtuslikku infot, mida tasub naljade analüüsimisel arvestada.

Näiteks on etniliste anekdootide seas sageli nalju, mis võivad olla kohaldatavad mitmele erinevale rahvale. Eriti selgelt väljendub see nende naljade puhul, mille skript on üles ehitatud tegutseja rumalusele. Kui otsida rahvuse järgi nalju soomlastest, näeme tulemuste seas anekdoote, mis räägivad ajuoperatsioonist ja kirurgi eksitusest tulenevast rahvusliku kuuluvuse muutusest. Võtmefraasiks on lause, mille opereeritu pärast narkoosist ärkamist kuuldavale toob (soomlase puhul nt: *Ah perrkele, külla see käy...!* (*Delfi Naljaleht* 5.06.2003)).

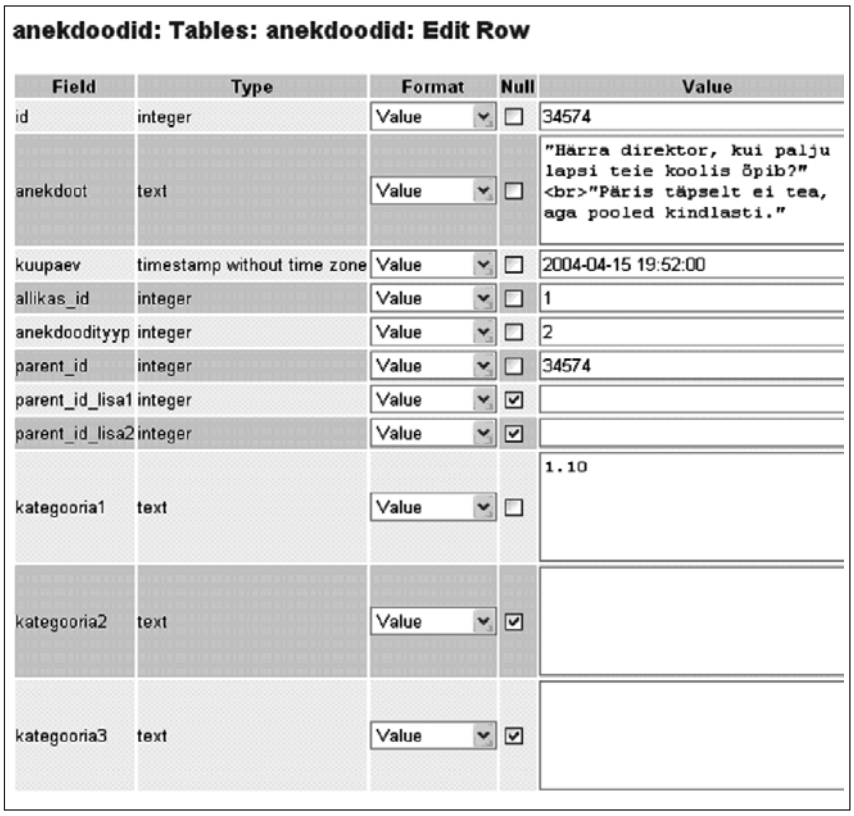
Kuid sageli tuleb ette nalju, milles inimene, kes kaotab kogemata rohkem ajust kui planeeritud, räägib ärgates mõnda muud keelt (vene, läti). Anekdooditegelase muutmise võimalus on nalja levikupotentsiaali tunnus. Siinkohal pole oluline see, mis keeles ja mis tegelasega anekdoot "alguses" võis olla, vaid mitme kultuurikonteksti ja sotsiaalse situatsiooniga sobiva skripti avastamine ja selle analüüs.

Lisaks sellistele naljadele, kus skripti ei muudeta ja vahetatakse vaid tegelast (nt sellise vastu, kes anekdoodirääkijatele on omasem ja tuntum kui tegelane algses naljas – näiteks eestlane vahetab iirlase soomlase vastu), on raskemini leitavad ja subjektiivsemalt kategoriseeritavad need, kus ühes naljad on põimitud kaks erinevat skripti või esineb sama situatsioon, kuid puänt on teine. Viimasel juhul on tulemuseks uus anekdoot (puänt kui anekdoodi kõige olulisem osa erineb), kuid nendevahelist seost on siiski huvitav märkida ja teinekord vajadusel kiiresti andmebaasist leida. Näiteks anekdoodid, kus kolme rahva esindajad püüavad kuldkala ja viimane soovija tahab erinevates variantides kas sõpru tagasi (üksikule saarele), luba tervitusi saata või lööb kala sakuska saamiseks maha jne.

Iga kategooria, kuhu anekdoot kuulub, märgitakse vastavasse veergu punktidega eraldatud numbritena (näiteks 1.11.2). Kasutades punkti kui eraldusmärki, saab need hiljem eraldi veergudesse lahutada, et kiirendada otsingumootori tööd. Esialgu pole see vajalik, sest materjali hulk on väike ja tulemused ilmuvad brauseriekraanile piisavalt kiiresti. Kategooriate numbrid ja nendele vastavad nimed on indekseeritud eraldi tabelisse.

Kategoriseerimiseks kasutatakse Postgresi andmebaasi graafilist internetipõhist kasutajaliidest PhpPgAdmin. See on andmebaasi haldamiseks loodud vahend, mis lubab administraatoril kerge vaevaga leida vajalikke sissekandeid ja muuta, lisada või kustutada andmeid.

Järgnevalt kirjeldan lühidalt olemasolevaid andmeid ja tabelit, kus need andmed asuvad – selles esinevaid veerge ja nende sisu ning veergude täitmise tingimusi. Postgresi tabelis on kategoriseerimise põhiliseks tööriistaks käsk *Edit Row*. Selle valimisel avaneb aken, mis on näha joonisel 2.



anekdoodid: Tables: anekdoodid: Edit Row

Field	Type	Format	Null	Value
id	integer	Value	☐	34574
anekdoot	text	Value	☐	"Härra direktor, kui palju lapsi teie koolis õpib?" "Päris täpselt ei tea, aga pooled kindlasti."
kuupaev	timestamp without time zone	Value	☐	2004-04-15 19:52:00
allikas_id	integer	Value	☐	1
anekdoodityyp	integer	Value	☐	2
parent_id	integer	Value	☐	34574
parent_id_lisa1	integer	Value	☑	
parent_id_lisa2	integer	Value	☑	
kategooria1	text	Value	☐	1.10
kategooria2	text	Value	☑	
kategooria3	text	Value	☑	

Joonis 2. Andmebaasi kasutajaliides PhpPgAdmin.

Kuigi tekstid on tabelisse juba paigutatud ja mõningal määral grupeeritud programmi abil, mis koondab anekdoodivariante sama nimetaja alla, vaadatakse iga tekst kategoriseerimise käigus üle, et väl-

tida sisulisi eksimusi kategoriseerimises ja korrigeerida vormilisi vigu anekdoodi tekstis. Veerud on järgmised:

Id – veerg, mis sisaldab anekdoodi unikaalset numbrit. See antakse anekdoodile kohe andmebaasi sisestamisel.

Anekdoot – nalja tekst. Vahel ka koos kommentaaridega, kui nalja saatja on lisanud naljale anekdooti puudutavat metateksti (tuleb ette nt Sünerkomi *Meie Naljaraamatust* pärit anekdootidel).

Kuupaev – enamasti täpne kuupäev ja kellaaeg, millal anekdoot on portaali või naljakogusse saadetud. Internetti riputatud anekdoodikogude puhul on kuupäevaks üks juhuslikult genereeritud päev ajavahemikust, mille jooksul konkreetne kogu koostati (sest veeru formaat nõuab ka päeva ja kuud lisaks aastale). Otsingu tulemusaknas kuvatakse ajavahemik, mil anekdoodikogu koostaja oma kogusse anekdoote lisas (nt Tanel Mägi kogu on koostatud aastatel 1994–2000).

Allikas_id – koondtabelis ülevaatlikkuse ja info kättesaadavuse kiiruse ja lihtsuse huvides väljendatud numbriga, viitega indekstabelile, kus numbrile vastav anekdoodiallikas (interneti- või trükikogu) on lahti kirjutatud.

Anekdoodityyp – nt keerdküsimus, kolmeastmeline anekdoot, anekdooditsüklisse kuuluv anekdoot vms. Anekdoodi vormiline määratlus on vajalik juhul, kui andmebaasi kasutaja soovib vastuseid ainult teatud ülesehitusega anekdootide hulgast, nt kolmeosalise ülesehitusega etnilisi anekdoote.

Parent_id – number, mis võib olla identne selle anekdoodi *Id*-numbriga (kui naljal puuduvad samasse tüüpi kuuluvad paralleelvariandid või kui nalja tekst on selle tüübi pea) või sisaldab veerg tüübi pea numbrit, kusjuures tüübi peaks on alati vastava tüübi väikseima *Id*-numbriga anekdoot. Selle veeru täidab esialgu automaatselt programm, mis hindab anekdooditekstide täht-tähelist sarnasust. Kui kokkulangevus kahe teksti vahel oli suurem kui 80%, märgitakse sinna sarnastest anekdootidest väikseima *Id*-numbriga nalja *Id*. Väiksema

kokkulangevuse puhul jäetakse veerg tühjaks ja kui ülevaatamise käigus sama tüübi anekdoote ei avastata, saab anekdoot *parent_id*-ks oma *Id*-numbriga identse arvu.

Parent_id_lisa1 – veergu märgitakse sarnase tüübi pea *Id*-number (nt juhul, kui on olemas sama anekdoot teiste tegelastega).

Parent_id_lisa2 – täidetakse juhul, kui on olemas paralleelvariant, mis erineb teistest millegi muu kui tegelaste poolest (nt juhul, kui anekdoot on sama situatsiooni ja ülesehitusega, kuid teistsuguse puändiga). Tegelikult peaks veerge, kuhu paralleelseid tüüpe märgitakse, olema veelgi rohkem (vastavalt vajadusele), praegu märgitakse siia veergu komadega eraldatult vastavate tüübipeade *Id*-numbrid.

Kategooria_1–3 – kolm veergu, mis sisaldavad kategooriasse kuuluvust. Seda väljendab punktidega eraldatud numbriline tähistus. Kui kategooriaid, mida anekdoodist võib välja lugeda, on rohkem kui kolm, üritatakse üles märkida kolm kõige selgemini välja tulevat kategooriat. Kui edasise kategoriseerimise käigus selgub, et rohkem kui kolme kategooriasse kuuluvaid anekdoote on väga arvukalt, võib kategooriaveerge juurde luua (*Kategooria_4* jne).

Näiteks järgmine anekdoot:

> *Kuldkalake on Eestis. Ta ütleb, et võib täita kolm soovi, aga ainult kolm.*
> *Eestlane ütleb, et tema sooviks, et Eestis ei oleks enam ühtegi venelast.*
> *Venelane ütleb, et tema sooviks, et Eestis ei oleks enam ühtki eestlast.*
> *Juut ütleb, et kui eelmised soovid täidetakse, siis tema sooviks pitsi konjakit*
> (Viikberg 1997).

Siin tuleb märkida kategooriaveergudesse 11.7 (“eestlane”, veerg *Kategooria_1*), 11.39 (“venelane”, veerg *Kategooria_2*) ja 11.19 (“juut”, veerg *Kategooria_3*); viite kuldkalale võib antud juhul märkida *Parent_id_lisa1* veergu, et siduda anekdoot teiste naljadega, kus kuldkala täidab kinnipüüdjate soove.

Veerge *Id*, *Kuupaev* ja *Allikas_id* ei muudeta, need on juba eelnevalt standardiseeritud ja sissekande õigsus (kuupäeva puhul) on kont-

rollitud – vales formaadis sissekannet ei lase programm sisestada, vaid teatab veast. Kui veerus, kus on anekdoodi tekst, leidub mõni nähtavalt häiriv kirjaviga, üleliigne tähemärk või anekdoodi juurde mittekuuluv tekstilõik (nt <*br*>), võib muuta selle veeru sisu. Lisatakse info anekdoodi ülesehituse kohta (veerg *Anekdoodityyp*). Kontrollitakse, parandatakse või lisatakse info selle teksti tüübilise kuuluvuse kohta (veerg *Parent_id*). Kui tegu on naljaga, millele on sarnaseid tekste (nt teise tegelasega, kuid sama ülesehituse ja puändi või isegi täiesti kattuva sõnastusega anekdoote), täidetakse ka veerg *Parent_id_lisa1* – väikseima *Parent_id*-ga anekdoot saab nende ühisnimetajaks. Veergu *Parent_id_lisa2* kirjutatakse nende anekdootide *Parent_id* numbrid, mis sarnanevad tegelaste ja/või situatsiooni osas, kuid erinevad puändi poolest. Veerud *Kategooria_1–3* täidetakse vastava kategooria numbrilise tähistusega, kuhu antud anekdoot kuulub. Esialgu kasutatakse klassifitseerimiseks punktidega eraldatud pikemat tähist (nt kategooria 6.2.2 tähistab alakategooriat *Loomad* > *Kalad* > *Lest*).

Digiteeritud materjalil on kategoriseerimisel paberkartoteegi ees eeliseid. Esiteks on sarnaste tekstide koondamiseks võimalik luua andmebaasisiseseid programme (kuigi selle tulemusel tekkinud tüübid nõuavad kindlasti uurijapoolset ülevaatamist ja korrektiive). Anekdootide grupeerimiseks loodi tekstide sarnasuse hindamisel põhinev programm. Kõiki anekdoote võrreldi omavahel ja arvutati nende tähttähelise kokkulangevuse protsentuaalne väärtus. Kui sarnasus oli suurem kui 80%, märgiti veergu automaatselt nendevaheline sarnasus (kuuluvus samasse tüüpi). Protsessi optimeerimine toimus sarnasusprotsendi märkimise piirmäära väljaselgitamise abil, vastasel juhul oleks tulemustabel tulnud u 50 000x50 000 veergu, mis kokkuvõttes oleks pigem segadust tekitanud, kui abistanud sarnasuste avastamisel.

Teiseks, palju kergem on leida ja grupeerida identseid koopiaid – see ei nõua mingit reaalset ümberpaigutamist, piisab vastava kate-

gooriamärgistuse kandmisest anekdootide koondtabeli vastavasse veergu. Sama tekst võib sel moel kuuluda mitmesse kategooriasse, ilma et seda peaks seejuures reaalselt mitu korda kopeerima.

Kolmandaks, kategoriseeritud andmemassiivis on kergem päringud sooritada, tulemusi on võimalik oma vajaduse kohaselt eelsortida (nt hilisemad sissekanded enne).

Neljanda plussina võib välja tuua selle, et digiteeritud kogu võimaldab juurdepääsu igalt poolt, selleks ei pea viibima Eesti Kirjandusmuuseumis.

Viiendaks osutub andmete digitaalne kuju kasulikuks sellele, kes soovib rakendada andmeanalüüsis statistilisi meetodeid. Andmeid saab tellida vastavalt vajadusele (nt erinevate rahvaste kohta käivaid anekdoote aastate kaupa), neid saab tõsta ümber andmetöötlusprogrammi ja seejärel analüüsida, tulemusi graafiliselt kujutada jms. Praegu nõuab see ligipääsu Postgresi tabelile andmebaasi sees, kuid esialgse üldpildi saab ka internetis kättesaadava anekdoodiotsingu abil ja lisamaterjali saamiseks või selle leidmisel abi saamiseks on võimalus pöörduda andmebaasi haldaja poole.

Digitaalsel kategoriseerimisel on ka miinuseid. Keeruline on meelde tuletada töö käigus juba varem tehtud kategoriseeringuid. Kui juhtub ette mõni redaktsioon juba varem liigitatud naljast või on vaja kategoriseerida anekdoot, milles on kahe juba kategoriseeritud nalja elemente, on nende viidete *Parent_id* leidmine aeganõudev toiming, mis nõuab sageli mitut otsingut ja eeldab naljas esinenud märksõna(de) mäletamist. Selleks tuleb kas väljuda kategoriseerimisprogrammist või teha lahti uusi otsinguaknaid, mis aeglustab tööd. Samuti ei saa anekdoote n-ö kõrvale panna ja hiljem üle vaadata, mis paberkartoteegi sorteerimisel on lihtne toiming. Hilisemaks ülevaatamiseks peab tegema uue otsingu, sest märgistatud anekdoodid paigutatakse järjest naljakogu lõppu. Otsing ajab omakorda segi märgendamist ootavate anekdootide järjestuse. Ka

brauseriakna sulgemine ajab järjestuse segamini ja uuesti alustades peab liikuma kõigepealt õigesse kohta, olles eelnevalt sorteerinud anekdoodid *Id*-numbri, tähestiku vm järgi.

Anekdoodihulga kategoriseerija kohtab ka klassikalisi probleeme, mis tulenevad kategoriseerimissüsteemi subjektiivsusest. Antud ülesande puhul üritati kategooriate loomisel vähendada subjektiivsust sellega, et üksikisikuliselt loodud süsteemi vaatasid üle üks ekspert ja kolm n-ö naiivset teadlast ehk folkloorikogumisega mitte kokku puutunud inimest, kes andsid oma hinnangu süsteemile ja esitasid parandusettepanekuid.

Kõige objektiivsem on naljade liigitamine anekdoodis selgelt väljendatud tegelas(t)e järgi, kuigi praeguste ettekirjutuste kohaselt võib tegelase järgi kategoriseerida ka juhul, kui sellele viitab ainult situatsioon (nt dialoog kahe inimese vahel toimub kohtus, millest võib järeldada, et tegelasteks on kohtunik ja kohtualune, kategooria 1.5, või nt dialoog, kus üheks tegelaseks on märgitud õpetaja, kategooria 1.10, kuid vastajat pole nimetatud – järeldatakse, et teiseks tegelaseks on laps ehk õpilane, kategooria 5.2.3).

Juhtudel, kus tegelast pole märgitud ja seda ei anna tuletada ka situatsioonist, liigitatakse anekdoot situatsiooni järgi (nt *Ühistranspordis*, kategooria 2.2, aga situatsiooninali on ka poliitiliste naljade alaliik *Nõukogudeaegne eluolu*, kategooria 10.1.3, või *Obstsöönsed naljad*, kategooria 15 ja selle alaliigitused).

Tegelaste järgi liigitamine on võimalikest variantidest objektiivsemaid, kuid selline mitmetähenduslik ja mitmekesine materjal nagu anekdoodid ei allu ühelegi kategoriseerimiskatsele ideaalselt. Seetõttu tuleb möönda, et mis tahes kategoriseering ei suuda märkida ära kõike, mis ühes anekdoodis sisaldub. Küsitavusi tuleb ette peaaegu igal sammul ja need peab lahendama kategoriseerija ise vastavalt oma nägemusele selle kohta, kuidas oleks andmebaasi kasutajal seda nalja hiljem kõige kergem leida. Võimalik, et GTVH-l põhine-

va verbaalse huumori tasandite põhjal loodud kategoriseering aitab tulevikus seda objektiivsemaks muuta.

Kategooriate arv peaks olema optimaalne – piisavalt suur, et andmehulka haaratavateks osisteks liigendada. Samas peaks kategooriaid olema võimalikult vähe, et säiliks ülevaatlikkus. Kui mõni kategooria osutub töö käigus ülearuseks, st ei sisalda piisavalt suurel hulgal ja eristuvaid nalju, võib kõik sinna kuuluvad anekdoodid kergesti ühe käsu abil ümber kategoriseerida. Suuremad kategooriad tuleb hiljem üle vaadata ja luua vajaduse korral kategooriasiseseid liigitusi.

Otsing

Nüüd vaatleme, milliseid võimalusi pakub (juba olemasolev) otsinguprogramm. Otsinguks on kaks võimalust: esiteks lihtne sõnaotsing, mis sirvib tõstutundlikkuseta läbi kogu andmebaasi anekdoodid (mitte autori, allika vm infoga veergu). Kui on soov leida mõnd anekdooti, mis osaliselt meelest läinud, või otsida nalju mõne konkreetse tegelase kohta, kellel on piiratud arv pseudonüüme (nt Juku, Illar (Allaste, aga ka Hillar või Hallaste), kuldkala vms), sobib see otsing hästi. Samas võib nt Juku-anekdootide koopiaid või variante võib leida ka kujul, kus tegelast pole mainitud (Juku asemel poiss või on tegelaseks lihtsalt õpetaja, isa vms). Tüüpilise Juku-anekdoodi tegelane võib olla mõni hoopis teise nimega poiss (nt Robert, kes tuleb ette Sõnumilehe *on-line*-väljaande anekdootides tüüpiliste Juku-anekdootide tegelasena).

Otsinguga on keeruline leida vajaminevat ka juhul, kui märksõna, mis seda anekdooti iseloomustada võiks, ei meenu või konkreetne anekdoot ei sisaldagi ühtegi "erilist", ainult sellele tekstile omast väljendit või sõna.

Lisaks sellele töötab kategoriseeritud anekdootide (esialgu *Delfi Naljalehe*) ulatuses detailsem otsing, mis leiab otsitava autori, allika, sõna

või täpse väljendiga soovitud kategooriasse kuuluva anekdoodi. Anekdootidest on kategooriate kaupa kättesaadavad *Delfi Naljalehel* leiduvad anekdoodid, mille on liigitanud Delfi portaali töötajad. See kategoriseering on väga üldine, koosnedes kuueteistkümnest suuremast naljaliigist ilma alakategooriateta. Praeguseks on lisaks kategoriseeritud ka etnilised anekdoodid, neil puudub veel vormiline liigitus, mis tuleb lisada. Etniliste naljade juures on ka kataloog, mis võimaldab kõigi praegu andmebaasi kuuluvate etniliste anekdootide seast vajalike tegelastega naljad välja otsida.

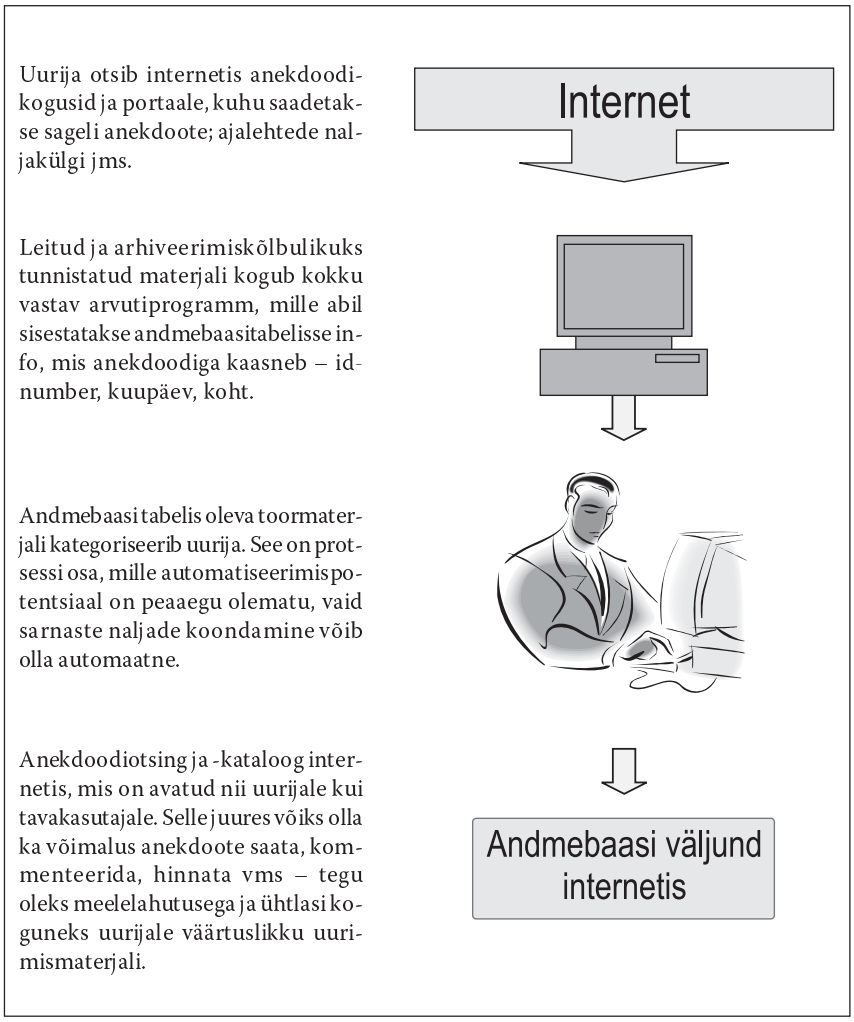
Tulevikus saab kategooriate kaupa anekdoote sirvida ka ilma otsingu abita. Selleks otstarbeks on kataloog, mis võimaldab soovi korral valida lugemiseks nt kõik naljad, mis puudutavad elukutseid, või siis vaadata alamkategooria tasemel selliseid anekdoote, mis räägivad ainult kultuuritöötajatest, või lugeda veelgi täpsemalt vaid tsirkuseartiste puudutavaid nalju, seda lihtsalt lingile klikkides.

Tulemuslehel on ka link sarnastele naljadele. Enamikul juhtudest on see link praegu alles tühi, sest suurem osa anekdoote pole vastavat tähistust paralleelsete tüüpide või variatsioonide kohta juurde saanud. Etniliste anekdootide puhul aga on süsteem lõplikult välja töötatud ja lingile vajutades jõuab andmebaasi kasutaja järgmisele tulemuslehele, kus on toodud kõik samasse tüüpi kuuluvad naljad. Lubatud on ka väikesed variatsioonid (erinevad tegelased, mõnel juhul ka erinev puänt). Pole veel otsustatud, kas lubada sarnaste naljade tulemuslehele ka nt sarnase teksti ja tegelaste, kuid erineva puändiga nalju või peaks selle kohta olema eraldi viide. Viimane oleks otstarbekas siis, kui tulemusi tuleks vastasel juhul liiga palju, praegu aga tundub, et lihtsam ja ülevaatlikum on kõik (nii sama tüübi naljad kui selle tüübi variatsioonid) koos tulemuseks tuua.

Töötab ka kategoriseerijale abiks loodud lisaotsing, mis annab tulemuseks arvuti poolt sarnasuse alusel järjestatud anekdooditekstid (rohkem kui 80% tekstidevahelist sarnasust).

Kogumis- ja arhiveerimisprotsess

Arvtuti abil on arhiveerimisprotsessi võimalik muuta kiiremaks ja paindlikumaks. Joonis 3 kirjeldab folkloori kogumise ja arhiveerimise protsessi internetis ja märgib, millised võimalused on digitaalse materjaliga töötamisel.



Joonis 3. Folkloori kogumuse ja arhiveerimise protsess internetis.

Kokkuvõte

Edaspidi tuleb jätkata kogu materjali juba alustatud ülevaatamist ja kategoriseerimist loodud süsteemi alusel. Töö edenemise põhitakistuseks on kohmakas kasutajaliides PhpPgAdmin, mille vahendusel toimub kategoriseerimine. See ei võimalda kiirelt ja paindlikult liikuda sama andmetabeli eri akende vahel – näiteks otsida sarnaseid anekdoote lisaks ka sõnaotsingu abil, kui käsil on arvuti poolt tekitatud sarnaste (sama tüübi) naljade sülemi ülevaatamine ja vajadusel korrigeerimine, – ega pöörduda tagasi mõne sülemi juurde, kui ette satub anekdoot samast tüübist või paralleelvariant, mille seost oleks otstarbekas kajastada ka kategoriseeringus. Selle asemel võiks olla spetsiaalselt programmeeritud vahend.

Miinuseks tuleb selle töö juures pidada ka suurt subjektiivsust, mis paratamatult kaasneb tõsiasjaga, et kategoriseerimisega tegeleb vaid üks inimene. Esiteks tähendab see, et kategooriad on loodud paljuski kategoriseerija enda vajadusi arvestades, teiseks seda, et anekdoodid on liigitatud lähtuvalt ühe inimese visioonist. Pealegi, nagu varemgi mainitud, on siin tegu žanriga, mis on peaaegu resistentne igasugusele lahterdamisele. Seda puudujääki saab vähendada nt sellega, kui paigutada anekdoodid võimalikult mitmesse kategooriasse, ja kui võimalik, kaasata kategoriseerimisprotsessi (vähemalt etapiti) n-ö naiivseid teadlasi või eksperte.

Küsimuseks jääb, kas andmebaasi peaks hiljem täiendama ka arhiivis olemasoleva materjaliga (näiteks puuduvad andmekogust praegu koolipärimuse kogumisel saadud anekdoodid ning enne 1960. aastaid kogutud naljad ja naljandid on samuti arhiveeritud vaid mittedigitaalsel kujul). Naljandite puhul saab komplitseerivaks asjaoluks see, et nende klassifikatsioon on hoopis midagi muud kui kaasaegsete anekdootide oma, nende teemad ja tegelased on hoopis teistest valdkondadest (nt sulase-peremehe naljad) ja nii naljandeid kui ka anekdoote katva ühise kategoriseeringu väljatöötamine on ilmselt

võimatu. Seetõttu on nende kahe ainese lahushoidmine põhjendatud, kuigi longituudsete uuringute tegemine on keerulisem, kui naljandid ja naljad asetsevad eraldi kogudes. 1990. aastate esimesel poolel eestlaste hulgas räägitud ja arhiivi jõudnud anekdootidega tuleb aga kindlasti tööd jätkata – need ülejäänud (koolipärimuse) tekstidest välja sorteerida, digitaalsesse andmebaasi sisestada ja süstematiseerida.

Tegu on projektiga, mida ei saa lõpetatuks lugeda nii kaua, kuni anekdoodid veel säilitavad oma värskuse ning traditsioon pole staatiline ja hääbuv, vaid aktiivne ja pidevalt muutuv ning kohanev. Kuigi on märke, mis osutavad anekdootide rääkimise vähenemisele, ei saa sama täheldada internetis leiduva huumoripärimuse kohta. Tundub, et anekdoodivestmine võtab lihtsalt teisi vorme, ja antud projekt ning andmebaas on üks katse seda muutust jäädvustada ning pakkuda materjali, vahendeid ja inspiratsiooni selle muutuse kirjeldamiseks.

Allikad

Delfi Naljaleht (http://www.delfi.ee/jokes – 19. aprill 2006).

Meie Naljaraamat (Sünerkomi *Jokebook*) (http://www.zzz.ee/jokes – praeguseks suletud).

Tanel Mägi anekdoodikogu (http://gabriel.ircnet.ee/jokes.html – praeguseks suletud).

Sõnumileht Online (http://www.sl.ee – praeguseks suletud).

Viikberg, Jüri (koost) 1997. *Anekdoodiraamat: Naeruga eilsest: Eesti anekdoot 1960–1990.* Tallinn: Eesti Keele Sihtasutus.

Kirjandus

Attardo, Salvatore 1994. *Linguistic Theories of Humor*. Humor research 1. Berlin: Mouton de Gruyter.

Attardo, Salvatore & Raskin, Victor 1991. Script Theory Revis(it)ed: Joke Similarity and Joke Representation Model. *Humor – International Journal of Humor Research* 4: 3/4, lk 293–348.

Brône, Geert & Fayaerts, Kurt 2004. Assessing the SSTH and GTVH: A View from Cognitive Linguistics. *Humor – International Journal of Humor Research* 17: 4, lk 361–372.

Burgess, Robert G. (toim) 1995. *Computing and Qualitative Research.* Studies in Qualitative Methodology 5. Greenwich CT: JAI Press.

Burnard, Louis D. 1987. Knowledge Base or Database? Raben, Joseph & Sugita, Shigeharu & Kubo, Masatoshi (toim). *Toward a Computer Ethnology.* Senri Ethnological Studies 20. Osaka: National Museum of Ethnology, lk 63–95.

Coffey, Amanda & Atkinson, Paul 1996. *Making Sense of Qualitative Data: Complementary Strategies.* Thousand Oaks: Sage.

Davies, Christie 1990. *Ethnic Humor Around the World: A Comparative Analysis.* Bloomington: Indiana University Press.

Davies, Christie 2002. *The Mirth of Nations.* New Brunswick (New Jersey): Transaction Publishers.

Dundes, Alan 1985. JAP and JAM in American Jokelore. *Journal of American Folklore* 98 (390), lk 456–475.

Dundes, Alan 1987. *Cracking Jokes: A Study of Sick Humor Cycles & Stereotypes.* Berkeley (California): Ten Speed Press.

Dundes, Alan & Banc, C. 1986. *First Prize, Fifteen Years!: Annotated Collection of Romanian Political Jokes.* Rutherford: Fairleigh Dickinson University Press & London: Associated University Presses.

Fischer, Michael D. 1994. *Applications in Computing for Social Anthropologists.* ASA research methods in social anthropology (Routledge (Firm)). London & New York: Routledge.

Hempelmann, Christian F. 2004. Script Opposition and Logical Mechanism in Punning. *Humor – International Journal of Humor Research* 17: 4, lk 381–392.

Ruch, Willibald & Attardo, Salvatore & Raskin, Victor 1993. Toward an Empirical Verification of the General Theory of Verbal Humor (GTVH). *Humor – International Journal of Humor Research* 6: 2, lk 123–136.

Weitzman, Eben A. & Miles, Matthew B. 1994. *Computer Programs for Qualitative Data Analysis: A Software Sourcebook.* Thousand Oaks: Sage.

VÕIM & KULTUUR 2

Koostaja ja toimetaja Mare Kõiva

http://www.folklore.ee/pubte/eraamat/voimjakultuur2/

Koostaja ja toimetaja: Mare Kõiva
Keeletoimetaja: Mare Kalda
Inglise keele toimetaja: Tiina Kirss
Makett ja kaas: Alo Paistik
Pilditöötlus: Andres Kuperjanov
HTML: Diana Kahre

ISBN 978-9949-586-83-7 (pdf)
ISBN 978-9949-418-53-4 (trükis)
DOI: 10.7592/VK2.2006
Tartu 2018

Trükis ilmunud: **Võim & kultuur 2**. Koostaja ja toimetaja Mare Kõiva. Võim ja kultuur. Tartu 2006

E-raamatu valmimist toetas: EKKM14-344 Eesti keele, kultuuri ja folkloori kasutusalade laiendamine ja tutvustamine elektroonilistel infokandjatel.