Eluraamatute arvutianalüüs – prooviuurimus

Theo Meder

Teesid: Niipea, kui andmed muutuvad suurandmeteks, võib muutuda probleemiks lähilugemise abil tehtav analüüs: sellest võib saada lõputu protsess, mida uurija aju ei suuda enam hoomata. Osaliselt võib probleemi lahendada digihumanitaaria meetoditega: suurte, näiteks etnoloogia-alaste tekstihulkade kvantitatiivseks analüüsiks on võimalik kasutada mitmesuguseid töövahendeid. Näiteks programmi AntConc abil saab uurida nii sõnade sagedust kui jaotust tekstis. LIWC2015 abi on võimalik kasutada elulugude sentimendianalüüsis, programmi Stylo tekstide stilomeetrilises analüüsis. Selliste programmide kasulikkust katsetatakse siin suhteliselt väikese nn eluraamatute korpuse juures, et selgitada välja, kui väärtuslik võiks see tööriist olla peagi korpusesse lisanduva palju suurema tekstihulga puhul.

DOI: 10.7592/methis.v21i26.16912

Võtmesõnad: stilomeetria, sentimendianalüüs, konkordants, sagedus, digihumanitaaria

Arvutuslik pilootprojekt

2013. aastal alustas fondi Humanitas Almere osakond uut projekti nimega „Levensboek" („Eluraamat"). Humanitase fondi vabatahtlikud küsitlesid vanemaid inimesi, et jäädvustada nende elulugu. Paljudel juhtudel salvestati nende lood mp3-formaadis, vabatahtlikud transkribeerisid need hiljem ja toimetasid saadud tekstid. Lõpptulemus, mida täiendasid jutustajate kogudest pärit fotod või internetist leitud ajaloolised illustratsioonid, trükiti ja avaldati väikesetiraažilise raamatukesena, nn eluraamatuna. Nagu varasematel hooldekodudes läbiviidud mälestustekogumise projektidel, oli ka seekord üheks eesmärgiks vanainimeste vaimu ärksana hoidmine ja nende hallidele ajurakkudele tegevuse pakkumine. Teiseks eesmärgiks oli aga elulugude kui möödunud aegade tunnistajate raamatu kujul levitamine neist huvitatud laste, lastelaste, perekondade ja sõprades seas.

Üks Almere Humanitase fondi algatajatest pöördus Theo Mederi poole küsimusega, kas need raamatukesed ei pakuks huvi Meertensi Instituudi arhiivile või uurijatele. Kuningliku Hollandi Kunstide ja Teaduste Akadeemia (KNAW) osaks olev Meertensi Instituut Amsterdamis on Hollandi igapäevaelu kultuuri ja keele uurimise instituut. Meertensi arhiiv sisaldab palju igapäevaelu kultuuriga seotud materjale ja näiteks päevikuid on kogutud juba palju aastaid. Eluraamatud oleksid sellele teretulnud täienduseks. Koostati lepingu tekst, kus jutustajad väljendavad nõusolekut oma lugude arhiveerimiseks ja täpsustavad, kui paljude aastate möödudes tohib raamatut uurimistöös kasutama hakata. Osa jutustajaid ei lubanud üldse oma elu-

raamatut kasutada, teistel polnud selle vastu midagi ja paljudel juhtudel lubasid nad oma eluraamatuid otsekohe uurimistöös kasutama hakata. Meertensi Instituut küsib alati ka eluraamatu digitaalset versiooni arvutuslikeks uuringuteks, kuid alati seda kaasa ei anta.

Arvutuslikes uuringutes saab rakendada automatiseeritud struktuurianalüüsi, uurida motiive, kasutada stilomeetriat[1] ja sentimendianalüüsi[2] – uurida võib kõike, alates muistenditest, rahvalauludest ja elulugudest kuni näiteks blogide ja säutsude kogumini. Võib teha ka soopõhiseid uuringuid: kas naised räägivad teistsugustest asjadest kui mehed? Meertensi Instituudis koostati küsimustik etnoloogе kõige rohkem huvitavatel teemadel: muistendid, laulud, mängud, pidustused ja tähtpäevade tähistamised, rituaalid ja meedia kasutamine. Milliseid lugusid jutustati ja milliseid laule lauldi lastele? Mida söödi hommikul, lõunaajal ja õhtul? Kuidas tähistati kuninganna päeva, lihavõtteid, Sinterklaasi päeva (nigulapäeva), jõule ja vanaaastaõhtut/aastavahetust? Kuidas korraldati eri aegadel pulmi ja matuseid? Milliseid mänge mängiti, milliseid spordialasid harrastati ja kuidas veedeti puhkust? Millised raadio- ja teleprogrammid olid lemmikud? Vabatahtlikud küsitlejad järgisid mõnikord seda küsimustikku, kuid praktikas jäi küsimustik siiski tihti kõrvale. See polnud ka koostatud arvutianalüüsi silmas pidades, vaid oli lihtsalt mõeldud küsitlejate abistamiseks intervjuude tegemisel.

Esimest eluraamatut esitleti Meertensi Instituudis pidulikult 2013. aasta mais, tegemist oli pr. Elly Ijsendijki raamatuga „Met Hart en Ziel" („Südame ja hingega"). 2018. aasta märtsis anti raehärra René Peetersile Almeres üle 50. eluraamat (kasutamiseks ka haridusalases, noorsoo- ja vabatahtlikus töös).[3] Meertensi Instituut polnud selleks ajaks veel kõiki 50 eluraamatut kätte saanud.[4] Pärast nähtavat edu Almeres hakkasid eluraamatuid valmistama ka Humanitase teised, näiteks Apeldoornis ja Zaandamis asuvad osakonnad.

4. juulil 2019 kutsuti mind Humanitase Almere ruumidesse pidama ettekannet eluraamatutega tehtavast uurimistööst. Sel ajal Meertensi Instituut peamiselt alles kogus eluraamatuid oma arhiivi tulevase uurimistöö tarbeks ja tegeles hoopis suu-

1 Stilomeetria (*stylometry*) on digihumanitaariale arendatud tekstianalüüsi meetod, millega saab võrrelda suuri tekstimassiive lingvistilise stiili (sõnakasutuse) alusel.

2 Sentimendianalüüsiga (*sentiment analysis*) leitakse tekstist psüühilistele tundmustele osutavaid indikaatoreid.

3 http://www.almeredezeweek.nl/nieuws/1453148-wethouder-krijgt-50e-humanitas-levensboek.

4 2019. aasta septembris anti Meertensi Instituudile üle 25 digitaalset eluraamatut, arhiivis on nüüd 52 paberile trükitud eluraamatut; vt http://www.meertens.knaw.nl/archieven/index.php?action=expand&querystring_b64=aW5ob3VkPWxldmVuc2JvZWsmYW1wO3NlYXJjaF9zdWJtaXR0ZWQ9Wm9law==&id=3918.

remate projektidega, uurides tänapäevaseid pidustusi ja rituaale, alternatiivseid raviviise, postkolonialismi ja rahvusvahelistel rahvauskumustel põhinevaid rahvajutte.[5] Võibolla alustame edaspidi egodokumentide projekti ja haarame sellesse ka päevikud ja eluraamatud. Eluraamatute kogu võib selleks ajaks olla juba mitmesajaköiteliseks kasvanud, nii et saaksime kohe kasutada suurandmeid. Küsimusele, mida me praegu nende eluraamatutega teeme, pean ma kahjuks vastama, et veel mitte midagi, me ainult arhiveerime neid. Sellest hoolimata otsustasin alustada pilootprojekti minu käsutuses oleva 19 digitaalse eluraamatuga, et näha, kas analüütilisi, digitaalseid vahendeid kasutades võiksid nad huvitavaks uurimisobjektiks saada.

Pean rõhutama, et olen humanitaarteadlane ja huvitun digihumanitaariast, kuid nagu paljud teised minuealised kolleegid humanitaarteadustes, ei oska ma ise programmeerida ega algoritme koostada. Olen alati kasutanud juba olemasolevaid, soovitavalt menüüpõhiseid töövahendeid. Eluraamatuid uurides otsisin struktuure, tundmusi, stiile, teemasid ja motiivide levikut, kasutades selliseid töövahendeid nagu LIWC2015 (Linguistic Inquiry and Word Count 2015), Stylo ja AntConc. Minu tegevus pole standardne etnoloogiline või antropoloogiline uurimistöö, kus kõige olulisem on igapäevase elu kultuuri kvalitatiivne uurimine (näiteks välitöödel inimesi küsitledes ja nende tegevusi vaadeldes). Projekti käigus otsisin kirjalikest tekstidest hulki ja mustreid. Lisaks olid intervjuude läbiviijad ilma etnoloogilise või antropoloogilise kogemuseta vabatahtlikud, kes enamasti jutustajatega lihtsalt juttu ajasid. Minu projekt keskendus edasistele võimalustele suurandmete kvantitatiivses uurimuses kasutamiseks (19 elulugu pole veel suurandmed).

Kõigepealt annan esmast infot eluraamatute ja nende jutustajate kohta, hiljem vaatlen mõnede arvutuslike töövahendite positiivseid ja negatiivseid külgi.

Jutustajad, toimetajad ja andmete puhastamine

Enamik elulugude jutustajaist on sündinud 1920.–1930. aastatel, väga vähesed sündisid pärast II maailmasõda. Lastena kogesid nad kriisiaastaid, enamik neist olid sõja tunnistajateks. On näha, et ilma eranditeta on need aastad jätnud jutustajatesse sügava jälje.

Digitaalsete tekstide arvutianalüüsi juures on alati üheks heidutavaks ülesandeks olnud materjali puhastamine ja arvutikõlbulikuks muutmine. Näiteks programmid Stylo ja AntConc sobivad lihtsa ASCII (või UTF-8) tekstiga töötamiseks.

5 Viimase tegevuse näiteks on „andmetesse kaevumise" projekt ISEBEL (Intelligent Search Engine for Belief Legends): http://www.isebel.eu/site/.

Eluraamatud aga alati kirjutatud MS Wordis või Adobe PDF-formaadis, nii et tekstid tuleb kõigepealt konverteerida ja puhastada, näiteks programmiga Sublime Text 3, ja salvestada lihtsasse txt-formaati. Uurijat huvitab ainult elulugu, jutustus ise, seega tuleb eemaldada ka tiitelleht, sissejuhatus ja kirjastaja andmed, aga ka fotod ja asjasse mittepuutuvad lõigud, leheküljenumbrid, kommentaarid ning toimetaja märkused. Lõpuks jääb järele „puhas" elulugu, kuid seegi on teataval määral suhteline: eluraamatutes pole kunagi audiosalvestuste täiesti töötlemata koopiad, intervjueerijad on neid ladusate ja loogiliste lugude saamiseks alati rohkem või vähem toimetanud.

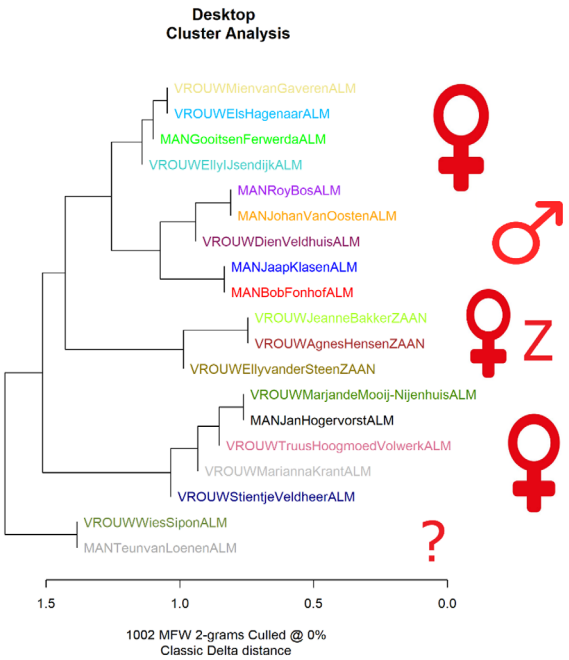
Tahtsin pilootprojektiga teada saada, kui suures ulatuses on võimalik eluraamatuid kasutada strukturaalanalüüsiks, jutustamise stiilis ja teemavalikul sugudevaheliste erinevuste leidmiseks, sentimendianalüüsiks, korduvate teemade ja motiivide leviku ning, mis võibolla kõige tähtsam, temaatiliste lünkade leidmisel: milliseid (tähtsaid) küsimusi jutustajad e i o l e puudutanud.

Pealiskaudseks strukturaalanalüüsiks pole arvutuslikke töövahendeid tarvis. Eluraamatute sisukordadest on kohe näha, et kõik lood on äärmiselt kronoloogilised. Sellise range kronoloogia puhul võib mängus olla ka intervjueerijate ja toimetajate käsi. Lugu algab tavaliselt inimese vanematega, jätkub lapsepõlvega, sõjaaja ja selle järelmõjudega, järgnevad kool, sõbrad, (meestel) sõjaväeteenistus, haridus ja elukutse, abielu, lapsed, kolimised, puhkused, haigused ja armastatud inimeste surmad. Lähilugemine näitab siiski, et hoolimata teatavate teemade rõhutamisest (näiteks võib lapsepõlv olla õnnelik, muretu, seikluslik või täidetud kannatuste ja õnnetustega) pole lood detailide tasemel kunagi sarnased. Iga kord on meie ees ainulaadne inimlik tunnistus elatud elust.

Pilootprojekti jooksul oli minu käsutuses 19 eluraamatut: seitse meesjutustajat elavad (või elasid, mõned neist on praeguseks lahkunud) Almeres. Kaheteistkümnest naisjutustajast üheksa elavad Almeres, kolm Zaandamis. Vanim jutustaja on 1909. aastal sündinud naine, noorim on 1970. aastal sündinud naine. Enamik jutustajaid on siiski sündinud 1920. ja 1930. aastatel. Mõned jutustajad sündisid väikestes linnakestes või külades, nagu Olst, Grootegast ja Scheveningen. Rohkem jutustajaid oli sündinud või vähemalt elanud suurema osa oma elust suurtes linnades: Amsterdamis, Rotterdamis, Haagis, Hilversumis, Enschedes, Nijmegenis ja Zaandamis. Lõpuks olid aga kõik jutustajad asunud elama Almeresse või Zaandami, kus Humanitase vabatahtlikud külastasid neid kodudes, vanadekodudes või hooldekeskustes.

Stilomeetria – jutustajate või intervjueerijate stiil?

Tahtsin statistikatarkvaral R[6] töötava programmi Stylo[7] abil tekitada mõned eksperimentaalsed klastrid, et vaadata, kas meeste ja naiste lugude vahel on stiililisi erinevusi. Selle eksperimendi puhul tuleb kindlasti rõhutada, et eluraamatute intervjueerijad ja toimetajad (asjaarmastajad vabatahtlikud) võisid vabalt toimida teatava filtrina: nende enda keelekasutus ja isiksused võisid analüüsis teatavat rolli mängida. Teisalt aga võib Stylo abil tehtud keeleliste väljendite analüüs põhineda (ebateadlikul) abisõnade kasutamisel ja vähem sõltuda sisu edasiandvatest (täistähenduslikest) sõnadest. Stylo suudab peamiselt vaadelda hollandi keele sõnade ja sõnaühendite, nagu *de, het, een, op, over, naast, onder* (artiklid; 'peal', 'üle', 'lähedal', 'all') jne kasutamist. Kui jutustajaid tsiteeritakse sõna-sõnalt, ei lähe sellised abisõnad kaotsi. Kuid paljudel juhtudel sõnastatakse intervjueeritavate ütlused ümber, neid ei tsiteerita sõna-sõnalt. Vaatlesin oma eksperimendis klastreid, mis põhinesid sõnade bigrammil (sõnapaaridel, kahe sõna kombinatsioonidel) ja tähtede trigrammil (kolme tähe kombinatsioonidel). Klastrite moodustumisel tekkisid alati eraldi meeste ja naiste grupid, (mõnikord ka segagrupid), nii et Zaandamist pärit naiste grupp moodustas omaette klastri ja kaks jutustajat (Sipon ja Van Loenen) jäid süstemaatiliselt klastrist välja. Allpool on üks klastri näide:



Joonis 1. 19 eluloo klasteranalüüs sõnapaaride põhjal.
Ülalt alla: esiteks valdavalt naistest koosnev grupp, järgneb valdavalt meeste grupp, siis Zaandamist pärit naiste grupp, siis veel üks valdavalt naiste grupp ja lõpuks kaks autsaiderit.

6 Vt https://www.r-project.org/.

7 Vt https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf.

Kõik analüüsid justkui näitaksid, et meestel ja naistel on erinev jutustamisstiil, kuid klasterdumise tulemusi on tõenäoliselt mõjutanud intervjueerijate sugu ja isiklik stiil. Esimeses, „naiste" klastris olid intervjueerijateks Boudewijn Vossen (M), Janneke Wiegers (N), Jacqueline Streppel (N) ja Carla Prins (N). Teises, „meeste" klastris kohtame intervjueerijat Carel de Vinki (M) kolm korda, ülejäänud teise klastri intervjueerijateks olid Lex Slager (M) ja Jolien van den Heuvel (N) koos Willem Jan Hagensiga (M) – peaaegu kõik olid mehed, mis ei pruugi olla kokkusattumus. Kõige otsustavam oli neljas, „naiste" klaster, kus Ina van der Vaart (N) oli kõigil viiel juhul nii intervjueerija kui ka toimetaja. See katse näitab selgelt, et intervjueerijate ja toimetajate mõju stiilile ei saa välistada.

Selgitamist vajavad veel kaks probleemi. Esiteks, miks erinevad Zaandamist pärit naised (kolmas klaster) teistest? Asi pole selles, et tekstidest kumab läbi Zaandami kohalik keelepruuk – nii Almere kui Zaandami tekstid on kirjutatud hollandi kirjakeeles. Leidub aga üks selge stiililine erinevus. Kõik Almere eluraamatud on kirjutatud ainsuse esimeses isikus: „Mu nimi on X, ma sündisin Y aastal ja nägin esimest korda ilmavalgust A linnas". Kõik Zaandami eluraamatud (välja arvatud mõned sõnasõnalised tsitaadid) on kirjutatud ainsuse kolmandas isikus: „Pr. X sündis Y aastal ja nägi esimest korda ilmavalgust A linnas".[8] Teine küsimus on, miks jutustajad pr. Sipon ja hr. Van Loenen seisavad kõigis visualiseeringutes alati eraldi (tihti rohkemgi, kui on näha joonisel 1). Neil ei leidu tüüpilisi sarnaseid jooni: nad on eri soost, ka nende intervjueerijad olid eri soost[9] ja nad pole ka täpselt ühevanused (esimene on sündinud 1920. a, teine 1932. a). Mõlemad on sündinud Amsterdamis, kuid seal olid sündinud ka mitmed teised, kes ei sattunud nendega ühte klastrisse. Siin tundub olevat ainult üks mõistlik seletus – nende elulood olid üldse kõige lühemad, esimeses 3867 sõna, teises 5693 sõna. Mõlemal juhul oli eluraamatus palju fotosid[10] ja vähe juttu. Stylo asetas Siponi ja Van Loeneni tekstid eraldi tõenäoliselt sellepärast, et ei leidnud piisavalt keelematerjali või klasterdas nad teistest eraldi kui „lühijutud".

Niisiis ei sobi eluraamatute stilomeetriline analüüs väga hästi jutustamisstiilide klastritesse paigutamiseks. Tekstide pikkus ja intervjueerijate mõjutused põhjustavad liiga palju müra, et korralikult analüüsida jutustajate konkreetseid, näiteks nende sooga seotud jutustamisstiile.

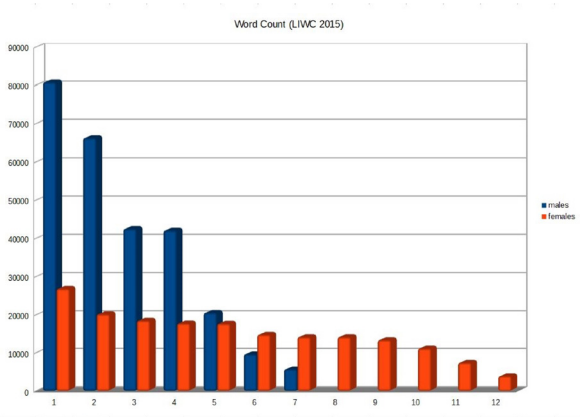
8 Zaandami naisi ei intervjueerinud üks ja sama isik: kahte intervjueeris Annemieke Blom (N), ühte Lidwien Berkhout (N). Intervjueerijate sugu võis siin rolli mängida.

9 Siponi intervjueeris Carla Prins (N), Van Loenenit intervjueeris Willem Jan Hagens (M).

10 Fotod kustutati tekstist elulugude analüüsi võimaldamiseks.

Sentimendianalüüs

Jõuame küsimuseni, kas keskmiselt jutustavad pikemaid lugusid oma elust mehed või naised. Selgub, et kui mehed juba hakkavad oma elust rääkima, siis keskmiselt teevad nad seda palju üksikasjalikumalt. Naissoost jutustajate lood on keskmiselt 14 229 sõna pikkused, meeste juttudes on aga sõnu üle kahe korra rohkem, keskmiselt 38 139. Joonis 2 näitab, et selle tulemuse taga on peamiselt neli eriti pika loo jutustanud meest.



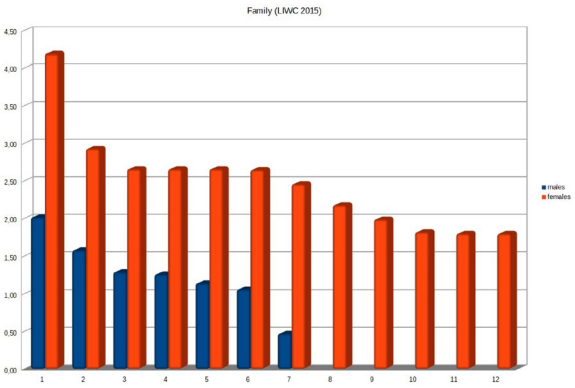
Joonis 2. Sõnade arv 19 eluloos LIWC2015 järgi, visualiseeritud MS Exceli abil. Sinine tähistab mehi, oranž naisi.

Programmi LIWC2015[11] on võimalik rakendada sentimendianalüüsis. LIWC2015 sisaldab hollandi keele sõnavara, kus eri tüüpi (peamiselt emotsionaalse laenguga) sõnad klassifitseeritakse kindlatesse kategooriatesse. Näiteks, vorm hollandi sõnast *huilen* ('nutma') sobitub järgmistesse kategooriatesse: tegusõna (kategoorias 20 korda), tundeline emotsioon (30), negatiivne emotsioon (32) ja kurbus (35). Kõik kategooriasse kuuluvad sõnad liidetakse kokku, jagatakse terve teksti sõnade koguarvuga ja tulemus korrutatakse sajaga. Tulemus 0,20 kategoorias „kurbus" on paljudel juhtudel suhteliselt madal, kuid tulemus 5,20 on üsna kõrge. Eriti tähtis on aga tulemuste väärtusi vaadata tekste võrreldes. Kokkuvõttes võib öelda, et sõnaloendur LIWC on küllaltki sobiv tendentside esiletoomiseks suurandmetes.

Analüüsisin LIWC2015 abil järgmisi kategooriaid: positiivsed emotsioonid, negatiivsed emotsioonid, ärevus, sotsiaalsed protsessid, perekond, töö, surm ja seksuaalsus. Emotsionaalsuses on naiste tulemused alati kõrgemad kui meestel, pole tähtis, kas tegemist on negatiivsete või positiivsete emotsioonidega. Naistel on sotsiaalsetest kommetest ja käitumisest, emotsioonidest ja kindlasti ka hirmudest

11 Vt http://liwc.wpengine.com/.

lihtsam rääkida. Meeste ja naiste vahel võib erinevust näha isegi siis, kui räägitakse perekonnast. Joonis 3 näitab, et perekonnast räägivad naised väga palju rohkem.



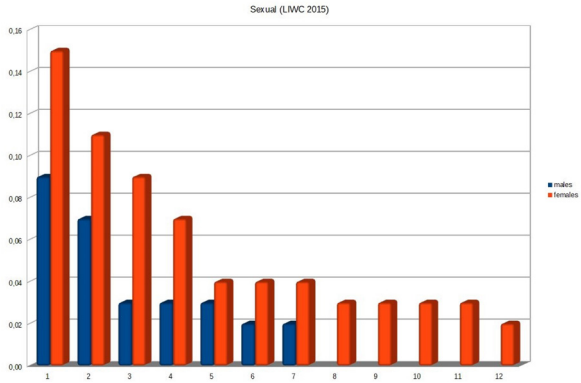
Joonis 3. Naised ja mehed räägivad „perekonnast" 19 LIWC2015 abil analüüsitud ja Exceli abil visualiseeritud eluloos. Naisi tähistab oranž, mehi sinine värv.

Võiks arvata, et vastandina naistele räägivad mehed sel juhul palju oma tööst, kuid tegelikult ei erine mehed ja naised selles osas märkimisväärselt. Pärast II maailmasõda hakkasid ka naised Hollandis palju rohkem tööl käima, kuigi pigem hoolitsemist pakkuvates elukutsetes nagu meditsiiniõed või ämmaemandad. Sotsiaalne praktika, et naised pidid pärast abiellumist töölkäimisest loobuma, on järkjärgult kadunud.[12]

Surm on naiste elulugudes palju tähtsamal kohal kui meestel. Selle mõistlikuks põhjenduseks on fakt, et keskmiselt elavad naised kauem kui mehed: ajal, kui naisi nende eluloo kohta intervjueeritakse, on nende abikaasad tihti juba surnud. Kuid naised räägivad rohkem ka lastest ja vendadest või õdedest, kes surid noorelt.

Naised räägivad ka seksuaalsusest rohkem kui mehed, kuid siin tuleb LIWCi abil saadud tulemuste tõlgendamisel ettevaatlik olla. Väärtused (vt y-telge joonisel 4) ulatuvad 0,02 kuni 0,15-ni, ülejäänud tulemustega võrreldes on need väärtused väga madalad. Tegelikult võib öelda, et ei mehed ega naised pole seksuaalsuse teemat eriti puudutanud.

12 Selliste suundumuste jälgimiseks vt Stokvis 1999 (eriti lk 134); ajalugu, mille ta visandab sekulariseerumist, sotsiokultuuriliste barjääride kadumist, töötingimusi, kasvatust, demokratiseerumist, vaba aega, puhkusi, seksuaalharidust, lahutusi ja muud sarnast käsitlevate küsitluste põhjal, on üsna heas vastavuses eluraamatutes räägitud lugudega.



Joonis 4. Naised ja mehed „seksuaalsusest" 19 LIWC2015 abil analüüsitud ja Exceli abil visualiseeritud eluloos. Naisi tähistab oranž, mehi sinine. NB: y-telje väärtused on siin märkimisväärselt madalad.

LIWCi analüüside põhjal võime niisiis öelda, et keskmiselt räägivad mehed pikemaid elulugusid kui naised. Naised aga kasutavad oma elulugudes rohkem emotsioone ja tundeid väljendavaid sõnu. Perekond on elulugudes väga tähtis teema ja seda rõhutavad üle kõige naised.[13] Oma tööst rääkides erinevad mehed ja naised üksteisest üsna vähe. Seksuaalsusest rääkimine on aga ühtviisi raske nii meestele kui ka naistele.

Seksuaalsus, sõda ja sageli kasutatavad sõnad

AntConc[14] on arvutuslik sobivuste leidmise töövahend, millega otsitakse sõnade sagedust ja varieeruvust: selle abil saab leida jutustustes lõike, mida uurijal oleks muidu raske identifitseerida. Kui kasutada AntConc'i metamärgiotsinguga **seks** (sellega leitakse ka sõna *homoseksuaalsus*), on näha, et 13 eluloos ei leidu seda üldse. Ainult kolm meest ja kolm naist on seda mõned korrad maininud. See ei tähenda aga, et kogu seksuaalsusele viitav materjal oleks elulugudest üles leitud, sest kui otsida metamärgiga **kuritarvitamine**, annab see 14 vastust. Sellest on rääkinud peamiselt naised. Üks naistest[15] jutustab, kuidas teda lapsena kuritarvitati (pärast selle sõna leidmist saame alustada vastava tekstilõigu lähilugemist):

13 Koostades nii naiste kui meeste elulugudest eraldi sõnapilvi, võib näha, et sõnad *lapsed*, *inimesed* ja *ema* ilmuvad valdavalt naiste lugudes. *Isa*, *ema* ja *lapsed* on selgelt näha ka meeste sõnapilves, kuid valdavaks on siin vormid tegusõnadest *minema*, *omama*, *tulema* ja *pidama*.

14 Vt https://www.laurenceanthony.net/software/antconc/.

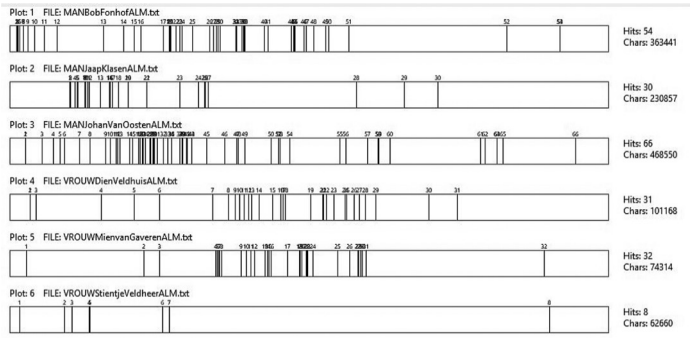
15 Privaatsuse huvides on tsitaadi allikas nii siin kui ka järgnevas anonüümseks jäetud.

> Mind kuritarvitati viiendast kuni üheksanda eluaastani. [. . .] Lisasissetuleku saamiseks lasi mu isa kodututel inimestel meie pööningul magada. Siis nad maksid majutamise ja toidu eest. Nad sõid ka siin ja „mängisid" minuga pööningul, või mul „lubati" neil süles istuda. Mulle tundus, et mind aeti öösel üles, kuigi see oli arvatavasti hilja õhtul, ettekäändel, et mul lubatakse tulla ja praekartuleid, aga mõnikord ka kana süüa. Need polnud mingil juhul kenad mehed, suurema osa ajast olid nad purjus. Ma polnud ainuke laps, kellega nad nii tegid. Ma tahtsin oma nooremaid õdesid kaitsta. Ma ei tahtnud, et neid üles aetakse. Sain selliste meeste käest raha ja andsin selle emale. Nii sai ta jälle süüa osta ja mu õed olid kaitstud.

Paljude jutustajate nooruses oli seksuaalsusega seotud veel palju naiivsust ja teadmatust. Näiteks üks teine naine (tsitaadis: X) räägib meditsiiniõeks õppimise ajast, kui ta pidi elama ühes toas koos teise naissoost õpilasega:

> Nad pidid ka üksteist medõdedeks kutsuma. Talle meeldis, et käidi kolme hoone vahet. Üks hoone magamiseks, üks töötamiseks ja kolmandas käidi söömas. Elati mitmekesi ühes toas. X oli kõige noorem, koos õde Y-ga anti talle tuba peahoones lastehaiguste osakonna kohal asuval korrusel. See oli kena vanniga tuba. Y ja X said omavahel väga hästi läbi, neist said lõpuks eluaegsed sõbrad. Nad kooskõlastasid oma vahetused ja tegid koos koduseid ülesandeid. Vabal ajal said nad ise otsustada, mida teha. Nad otsustasid pärast tööd koos vanni minna ja seal koduseid ülesandeid lahendada ja süüa. Nad arvasid, et see oleks väga meeldiv ja efektiivne. Seda kuuldes mõtles ülemõde aga teisiti. Ta kutsus tütarlapsed välja ja küsitles neid. Lõpuks tegi ta neile selgeks, seda ei lubata. X ja Y ei saanud millestki aru, aga ei pesnud ennast edaspidi enam koos. Nad polnud kunagi varem kuulnud homoseksuaalidest ega lesbilisusest. Jutustaja sai sellest teada hiljem, kui üks meessoost medõde talle avalikult teatas, et ta on homoseksuaal. Hiljem selgus, et kõik meessoost medõed olid homoseksuaalid.

Tulles tagasi varasema küsimuse juurde: kui mehed jutustavad nii pikki elulugusid, millest nad siis nii palju räägivad? Teisest maailmasõjast. Sõja üleelanud poistele tundub see tagantjärele meenutades olevat justnagu põnev poisteraamat. Mitmel mehel käsitleb pool eluloost või rohkemgi sõjaaegseid seiklusi. Joonis 5 esitab metamärgi **sõda** (hollandi k *oorlog*) jaotuse AntConci abil tehtud visualiseeringu (metamärgiotsing leiab ka sõnad *maailmasõda* (*wereldoorlog*) ja *sõjavägivald* (*oorlogsgeweld*)) algul kolme mehe loos, hiljem ka kolme naise loos. See tunnuste sarnasuse visualiseering, mis näeb välja nagu ribakood, näitab, kuidas sõna *sõda* jaotub üle kogu eluloo:



Joonis 5. Sõna **oorlog** ('sõda', koos metamärkidega) jaotus kolme mehe ja kolme naise elulugudes. AntConc'i abil tehtud visualiseering näitab leidude arvu, nende asukohta tekstis ja kogu teksti tähemärkide arvu.

Kui me visualiseerime sõna *mina* (*ik*) jaotust, siis mõnikord värvuvad terved jooniste plokid mustaks. Kui inimesed räägivad omaenda elulugusid, pole siin midagi imestada. Tabel 1 võtab kokku 19 eluloost koosneva korpuse väljapaistvad täistähenduslikud sõnad, millega inimesed viitavad iseendale, perekonnale, sugulastele ja oma lähikeskkonnale.[16]

Sõna	**Kordused**
Mina	12 495
Meie	4860
Laps(ed)	1043
Ema	1018
Isa	993
Maja	898
Perekond (+ sugulased)	483
Sõber (m/n)	388
Vanemad	366
Vend	293
Tütar	240
Poeg	152
Õde	125

Tabel 1. AntConc'i abil koostatud sagedusnimekiri kõigi 19 eluloo tähtsatest isikule, perekonnale, sugulastele, sõpradele ja lähikeskkonnale viitavatest sõnadest.

AntConc'i abil tehtud otsingud näitavad ka tähtsamaid momente elulugude jutustajate elus. Kolimistel ja abieludel, aga ka kirikul ja puhkustel on tähtis roll. Paljud jutustajad veetsid lapsepõlve kristlikus keskkonnas ja kindlasti mitte kõik ei lahkunud kirikust, hoolimata hilisemast ühiskonna ilmalikustumisest. Puhkused olid varasematel aegadel peamiselt rattaretked Hollandis. Jõukuse kasvades ja vaba aja lisandudes, kui autod muutusid kättesaadavamaks, nihkusid puhkuserei-

16 Abisõnad on välja jäetud ja kõiki täistähenduslikke sõnu pole siin ülevaates näidatud, nagu *aasta* (1050 kordust), *aeg* (826 kordust), *inimesed* (586 kordust), *kodu* (525 kordust) ja *kool* (470 kordust).

side sihtkohad kaugemale. Lennureisid viisid lõpuks inimesed puhkusele kaugetesse paikadesse, millest jäid püsivad mälestused.

Mäletamisväärsed sündmused	Jutustajate arv	Kordused
Kolimised	19	207
Abielu	18	204
Kirik	18	203
Puhkus	17	167
Armastus	17	124
Jõulud	12	79
Sinterklaas (nigulapäev)	12	39
Ristimine	11	38

Tabel 2. AntConc'i kokkusobivusgraafiku funktsiooni abil 19 eluloost koosnevas korpuses leitud meeldejäävad sündmused.

Sama meeldejäävad, kuigi negatiivsemas mõttes, on elulugudes haigused ja lähedaste ning armastatud inimeste surm. Kõik 19 jutustajat räägivad haigustest (kokku 558 kordust) ja kõik nad räägivad ka surmast (334 kordust). Surmasid mainitakse enamasti loo lõpus, kui partner sureb, kuid ka loo alguses, kui surevad vanemad või väikesed lapsed.

Välja arvatud noorim jutustaja, kes oli sündinud 1970. aastal, räägivad kõik jutustajad pikemalt või lühemalt II maailmasõjast ja jälgedest, mida see nende ellu jättis. Vanimad jutustajad räägivad ka sõjale eelnenud kriisiaastatest. Need sündmused on kirjas ka ajalooraamatutes. Pärast sõda leidsid aset põhjalikud muutused kultuuris, majanduses ja poliitikas, mis on samuti dokumenteeritud ajalooraamatutes. Üks veidi noorem jutustaja märkis, et ta oli ansambli The Beatles fänn ja et talle meeldis kuulata piraatraadiojaama Radio Veronica.

Kuna see on ainult pilootuurimus, oleks paljud tulemused olnud leitavad ka lähilugemise ja sõnade käsitsi kokkulugemisega. Niipea kui elulugude hulk hakkab suurenema, aitavad arvutitööriistad, nagu LIWC ja AntConc, suurtes korpustes mitmeid elulugudes esinevaid jooni ja detaile süstemaatilisemalt ja kiiremini leida: emotsioone, tähtsaid suhteid, olulisemaid sündmusi inimeste elus ja sõja mõjutusi.

Temaatilised lüngad: asjad, mida ei leitud

Paljut sellest, mida ajalooraamatutes nimetatakse meeldejäävateks sündmusteks, ei ole ära märgitud üheski 19 eluraamatust. AntConc'iga tehti palju päringuid. Kus on „nozemid" (1950. ja 1960. aastate mässulised noored Hollandis) ja hipid, kus on The Rolling Stones? Kus on selliste narkootikumide proovimine nagu kanep, *speed* ja LSD? Kus on president Kennedy mõrvamine või esimene laskumine Kuule? Kus on „Dolle Mina" (radikaalsed feministid), kus on rasestumisvastased tabletid ja

abort („Omaenda keha peremees")? Kus on Maagdenhuisi[17] okupeerimine (üliõpilaste demokraatiameelsed protestid Amsterdamis 1960. aastatel)? Isegi tuumaenergia ja neutronpommi vastaseid demonstratsioone ja proteste pole kusagil leida. Tundub, nagu poleks naftakriisi kunagi olnudki. Pole ka märke suurest näljahädast Bangladeshis. Elulugudes ei mainita ka kuulsaid Hollandi diplomaate, nagu Joop den Uyl, Hans Wiegel ja Dries van Agt.

Kokkupuuteid maailmas toimunud sündmustega mainitakse ainult juhuslikult. Korra ütleb üks Amsterdamist pärit naine:

> Ma pole kunagi märganud ühtegi asja, mida räägitakse kõigis neis lugudes „Provo" [18] ja üliõpilasrahutuste kohta. Märatsemist [kuninganna] Beatrixi pulmade ajal nägin televiisorist. Mul pole mingeid mälestusi külmast sõjast.

Teine naine räägib esimesest majast, mille nad ostsid:

> Oivaline maja suure eenduva aknaga, topeltklaasidega, vitraažustega, väga kaunis. Majal oli eesaed ja suur tagaaed. Maja olid üle võtnud skvotterid, see nägi alguses kole välja. Lakke oli pruuni värvi visatud. [. . .] Skvotteritel olid kassid. Maja oli kirpe täis! Kui mina ja X majja astusime, hüppasid kirbud vastu me jalgu. Me panime meelega jalga valged püksid, et neid hästi näha ja kinni püüda. Pärast seda desinfitseeris Y kõik kohad. Ma tõmbasin tolmuimejaga tühjaks ja puhtaks kõik põrandapraod. Nii saime lõpuks kirbunuhtlusest lahti. Naabrid olid väga õnnelikud, kui me selle ostsime. Samal tänaval asusid üldarst, veterinaar [. . .], samas oli ka apteek. Nii olid need inimesed äärmiselt rahul sellega, et skvotterid olid läinud.

Siin kohtame episoodi 1980. aastate ajaloost, kuid skvotterite liikumisest ei räägita siin seoses protestiga eluasemete vähesuse pärast, vaid ainult nende korrapäratu ja destruktiivse eluviisi negatiivse maine tõttu.

Üks meessoost jutustaja mainib 11. septembri rünnakut kaksiktornidele, aga mitte poliitilise ja religioosse kokkupõrke tõttu USA ja ortodokssete moslemite vahel, vaid ainult sellepärast, et rünnak rikkus tema puhkuse Los Angeleses ja San Franciscos:

> Pidime tagasi lendama 11. septembril – sellel kurikuulsal 11. septembril 2001, kui rünnati kaksiktorni – ja meie lend jäeti ära. Õhuliiklus peatus. Kui me oma hotellitoa aknast välja vaatasime,

17 Algul lastekodu, hiljem Amsterdami Ülikooli administratiivne keskus. Üliõpilased okupeerisid hoone 1969. a ja nõudsid suuremat avalikku kaasamist ja demokraatiat.

18 Provokatiivne ja autoritaarsusevastane noorteliikumine (1965–1967).

nägime, et Kuldvärava sild oli sõdureid täis. Teisel kaldal oli osariigi valitsus, ka seal olid sõdurid. X soovitas sõita tagasi Los Angelesse, sest sealt sai veel lennata. Teel sinna magasime neljakesi ühes toas. Kui jõudsime Los Angelesse, selgus, et ka sealt ei saanud enam lennata. Saime üheks ööks hotellitoa, aga pidime pärast seda lahkuma, sest hotell oli täiesti täis. Sõitsime mingisse teise hotelli ja magasime seal järgmise öö. Kokku jäime terve nädala hiljemaks. Õnneks saime edasi kasutada autot ja selle lennujaama tagasi viia. Ootasime lennujaamas kella üheksast poole neljani enne, kui saime lõpuks koju tagasi lennata.

Üks jutustaja räägib pornograafiast, kuid ta oli mõned aastad sekspoodi pidanud. Üks teine mainib lühidalt „BOM"-ema (üksikvanem omal valikul). Üks naine räägib põgusalt PvdAst (Töölisparteist) ja VVDst (Demokraatlikust Rahvaparteist). Ja üks mees nimetab raudset eesriiet. Need on ainult pisidetailid pikemates elulugudes.

On väga tõenäoline, et II maailmasõja järgsed sündmused maailmas libisesidki inimestest mööda. Ei tohi ka unustada, et me palusime neil rääkida oma isiklikku elulugu. Inimesed pole eriti huvitatud ühiskonnas toimuvate protsessidest ega nendega seotud. Need asjad toimuvad väljaspool nende „mulli". Mitmed nähtused ja sündmused pole nii suured, kui nad ajalooraamatutes paistavad – inimestel, kes polnud õiges vanuses, õiges kohas ja õige hoiakuga, võisid need kergesti märkamata jääda. Võibolla kulub ka veel üks põlvkond, enne kui näiliselt kaduma läinud lood välja ilmuvad. Võibolla inimesed isegi märkasid neid sündmusi nende toimumise ajal, kuid kui jutustajad olid saanud 75- või 80-aastaseks, olid need tähtsuse kaotanud.

Siiski – üks jutustaja osutus selgeks erandiks: tema elulugu näitab, et ta kindlasti ei elanud oma „mullis". Jutustajal on selgelt olemas oma arusaamad nii poliitikast kui ka ajaloost. Ta läks pärast II maailmasõda sõjaväkke ja saadeti Indoneesiasse osalema relvastatud mässuliste grupeeringute vastases võitluses. Jutustaja räägib realistlikult ägedatest lahingutest, kus võitsid kord ühed, kord teised, ja jõhkratest kuritegudest, mida mõlemad pooled korda saatsid. Järk-järgult sai ta aru, et Hollandi valitsuse tegevus oli seadusevastane, et tuleb loobuda kolonialistlikust tegevusest, et vastuhakkajad olid tegelikult vabadusvõitlejad ja et Indoneesia rahval oli õigus enesemääramisele ja sõltumatule riigile. Jutustaja räägib ilmekalt ka dekolonialiseerimisprotsessi järgsest ajast: KNILi sõdurite[19] saabumisest Hollandisse, Maluku saarte elanike immigratsioonist ja Hollandi valitsuse lubadusest kindlustada vaba Lõuna-Maluku vabariigi tekkimine, mis aga kunagi teoks ei saanud. Ta räägib ka rongi pantvangi võtmisest malukulaste poolt De Puntis 1977. aastal. See aga ei lõpeta veel tema osalemist maailma sündmustes. 1970. aastatel

19 Hollandi Indoneesia armee sõdurid.

töötas ta Nigeeria haridusministeeriumis. Ta räägib põhjalikult sealsest geograafilisest, poliitilisest, religioossest ja sotsiaalsest olukorrast, suguharudevahelistest konfliktidest ning moslemite ja kristlaste rivaliteedist. Lõpuks teeb ta juttu ka kodusõjast ja näljahädast endises Biafras (Lõuna-Nigeerias). Sellist tähtsate maailmaajaloo sündmuste analüüsi pole üheski ülejäänud 18 eluraamatus.

Kokkuvõte

Kas eluraamatute pilootprojekt oli edukas? Kas tulevikus on arvutianalüüsi võimalik kasutada ka palju suurema elektroonilise eluraamatute kogu puhul? Siinne eksperiment näitas, et narratiivseid struktuure on võimalik uurida ja see uurimistöö saab, vastavalt vajadusele, olla palju üksikasjalikum. Meeste ja naiste lugude stilomeetriline analüüs Stylo abil on üsna komplitseeritud, sest intervjueerijad/toimetajad võivad filtritena väga palju vahele segada. Eri stiilide kindlakstegemiseks otsib Stylo mustreid abisõnade kasutamises, kuid eluraamatud ei tsiteeri jutustajaid sõna-sõnalt, nii et mõnedel juhtudel võivad lingvistilised jooned, nagu abisõnade kasutamine, pärineda toimetajatelt, mitte jutustajatelt.

Teisalt on näiteks sooga seotud sentimendianalüüs võimalik LIWC2015 abil: see töövahend esitab üsna hästi elulugudes leiduvaid emotsioone, suhteid ja nendega seotud motiive. AntConc osutub kasulikuks töövahendiks mitmesuguste teemade esinemise ja jaotuse uurimisel. Huvitavaks võimaluseks on ka teatavate teemade ja motiivide puudumise uurimine. Lõpetuseks võib järeldada, et Humanitase eluraamatud sobivad erakordselt hästi igapäevaelu kultuuri uurimiseks, kuid väga palju üksikasjalikku infot muistendite, laulude, mängude, pidustuste ja rituaalide kohta neis pole.

Tõlkinud Marika Liivamägi

Allikad

Collectie Humanitas levensboeken, 2013–2015. Meertens Instituut, inventarinumber 496. http://www.meertens.knaw.nl/archieven/index.php?action=expand&querystring_b64=aW5ob3VkPWxldmVuc2JvZWsmYW1wO3NlYXJjaF9zdWJtaXR0ZWQ9Wm9law==&id=3918.

Stokvis, Pieter R.D. 1999. *Terugblikken op het huiselijk leven in de twintigste eeuw. Een verzameling getuigenissen over veranderingen in levensstijl sinds 1920*. Leiden.

Theo Meder – PhD, Amsterdami Meertensi Instituudi vanemteadur hollandi rahvajuttude alal ning Groningeni Ülikooli hollandi rahvajuttude ja narratiivse kultuuri professor. Tema peamised uurimisteemad on folkloor, rahvajutud, etnoloogia, filoloogia, keskaja kirjandus ja kultuur, (rahvusvahelised) andmebaasid ja digihumanitaaria.
E-post: Theo.Meder[at]Meertens.knaw.nl, T.Meder[at]Rug.nl

SUMMARY

Computational Analysis of Life Books – a Probing Study

Theo Meder

Keywords: stylometrics, sentiment analysis, concordance, frequency, computational humanities

As soon as "data" turn into "big data", analysis by "close reading" can become a problem: it can become an endless process that eventually the brain of the researcher can no longer get a grip on. Methods of computational humanities can partially solve the problem: various tools can be used to make quantitative analyses of large amounts of text, for example in the field of ethnology or folklore. Various tools may be considered for such text analysis. For example, the program AntConc can be used to study word frequencies, as well as the distribution of concepts across the text. LIWC2015 can be used for sentiment analysis of life stories and show differences between genders (or generations) in telling life stories. Stylo may be used for the stylometric analysis of texts. The usefulness of such programs is tested here on a still relatively small corpus of so-called Life Books. The Life Books ("Levensboek") is a project of Humanitas Foundation in Netherlands to publish limited edition booklets of life stories compiled from interviews with older people, recorded and edited by volunteers. This study is based on 19 digital Life Books – in fact still small enough for close reading and qualitative analysis. However, the intention here is to use the corpus as a pilot to see how valuable the tools can be for a much larger amount of texts that will be added in the near future. In this pilot I want to see to what extent the Life Books can be used for structural analysis, gender differences in narrative style and subject choice, sentiment analysis, recurring themes, distribution of motifs, and perhaps most importantly: thematic gaps. That is to say: which (important) issues are not raised by the storytellers?

The experiment shows that it is possible to do research into narrative structures, although this could be much more refined in terms of events. Stylometric analysis with Stylo of male and female repertoires is rather tricky, because interviewers/editors can (very much) interfere as a filter here. Stylo looks for patterns in the use of function words to determine different styles, but Life Books are just not quoting narrators literally all the time, so in quite some cases linguistical features, like the use of function words, may not originate from the storytellers but from the editors.

On the other hand, sentiment analysis in combination with gender, for example, is possible using LIWC2015: this tool can give a fair respresentation of emotions, relationships and related motifs in life stories. Furthermore, AntConc proves to be a useful tool to investigate the occurrence and distribution of themes and topics. Research into the lack of certain themes and motifs remains an interesting option as well.

Theo Meder – PhD, Senior researcher of Dutch Folktales at the Meertens Instituut in Amsterdam and professor of Dutch Folktales and Narrative Culture at the University of Groningen. His research interests include: folklore, folktales, ethnology, philology, medieval literature and culture, (inter)national databases and computational humanities.
E-mail: Theo.Meder[at]Meertens.knaw.nl, T.Meder[at]Rug.nl