SUUNDUMUSI INIMSUHTLUSE KEELELISES ANALÜÜSIS JA MODELLEERIMISES (I)

1. Vyvyan Evans. The Crucible of Language. Cambridge: Cambridge University Press, 2015. 359 lk.

2. Arefeh Farzindar, Diana Inkpen. Natural Language Processing for Social Media. (Synthesis Lectures on Human Language Technologies 30.) San Rafael: Morgan & Claypool Publishers, 2015. 146 lk.

3. Bing Liu. Sentiment Analysis. Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press, 2015. 367 lk.

4. Natural Language Generation in Interactive Systems. Toimetanud Amanda Stent, Srinivas Bangalore. Cambridge: Cambridge University Press, 2014. 363 lk.

5. Rainer Schulze, Hanna Pishwa. The Exercise of Power in Communication. Devices, Reception and Reaction. Basingstoke–New York: Palgrave Macmillan, 2015. 337 lk.

Järgnevas ülevaateartiklis kirjeldame viit hiljuti ilmunud (ja TÜ raamatukokku tellitud-saabunud) raamatut. Põhjus, miks neid koos käsitleme, on neid ühendav teema. Kõige lihtsamini võib seda iseloomustada väljendiga „inimsuhtlus ja selle analüüs", kuid vahetu põhjus on see, et need raamatud (täpsemini 2–4)

esindavat üht suhteliselt uut, kuid seejuures selgelt piiritletavat suunda selles valdkonnas. Selleks on suhtlusviisid, kus suhtlusvahendiks on keel, suhtluskeskkond on sotsiaalmeedia ning selle analüüsi- ja uurimisvahendina kasutatakse tekstide arvutitöötluse meetodeid. Nende raamatute kaudu püüame anda ülevaate selle temaatika käsitlemisest, probleemidest ja praegusest seisust. Ühtlasi loodame selgitada, miks see temaatika KK lugejale oluline on.

Kuivõrd inimloomus on eeldatavasti jäänud samaks, nagu ta aastatuhandeid on olnud, on vaadeldav probleem selline, mida ei lahenda suhtlusuurijad ega keeleteadlased omaette. Keel ja selle kasutus tuleb siduda kehtivate suhtlusvajadustega. Kunagi (nt kiviajal) olid selleks omad igapäevategevused, nüüd hakkamasaamine tänapäevase tsivilisatsiooni keskkonnas. Miski siiski seob neid keelekasutusi. Ja kui neid kasutusi tahetakse uurida, siis nüüd on lisandunud informaatikud-andmetöötlejad, seejuures võtmetegijatena. Siit ka siinse ülevaate algtõuge ja raamatute valik. Ülevaate esitame küll raamatute kaupa, kuid iga raamatu üldraamist nopime välja meid huvitava temaatikaga haakuva materjali ja keskendume sellele. Ning lisaks – ja see on ülevaate üks omaette eesmärke – kirjeldame, mida on sama problemaatika käsitlemisel Eestis ja konkreetselt eestikeelse andmestiku käsitlemisel tehtud.

1. Keel ja vaim kui tähenduse loojad

Vyvyan Evansi raamatu „Keele sulatusnõu" valisime ülevaate sissejuhatavaks teoseks mitte niivõrd sellepärast, et see on seni üks viimaseid ülevaateraamatuid kognitiivsest keeleteadusest, kuivõrd seetõttu, et selles on inimkeele ja eriti selle tähendussüsteemi käsitluses rõhk mitte inimese individuaalse maailmatunnetuse rollil selle süsteemi arengus selliseks, nagu see on, vaid inimliigi sotsiaalsel toimimisel ja sellega seotud eripäraste isikutevaheliste suhtlemisviiside rollil selles toimimises (mis muuhulgas päädivad ka tänapäevases sotsiaalmeedias).

Raamat on järg autori eelmisele raamatule „Keele müüt" (alapealkirjaga „Miks keel ei ole instinkt").[1] Viimane on peaaegu täielikult pühendatud (Chomskyst lähtunud) ideede kritiseerimisele, mille järgi inimese keelevõime geneetilis-pärilik tuum on universaalne grammatika ning konkreetselt selle eripärane süntaks (laused, nende struktuuri formaalsed iseärasused), kusjuures sellise universaalse grammatika kinnistumise „evolutsiooniliseks tõukejõuks" ei pruukinud üldse olla keele suhtlusfunktsioon. Siinses raamatus esitab Evans oma käsituse sellest, mis moodustab inimkeele tegeliku tuuma, tuginedes eelkõige andmetele, mis viimaste aastakümnete jooksul on lisandunud inimliigi evolutsiooni kohta tervikuna. Et meid siinses ülevaates huvitavad seisukohad ja andmed, mis seostuvad keele kui sotsiaalse suhtluse vahendiga, siis keskendume nendele.

Evansi keskne idee sisaldub juba raamatu pealkirjas: keele kui inimeste suhtlusvahendi sündi evolutsioonis ei saa kujutada mingi omaette sündmuse või protsessina, keel on mitme evolutsioonis koos kulgenud protsessi ühistulemus, sulam. Hästi lihtsalt võib seda kontseptsiooni kirjeldada järgmiselt (märgitagu, et kontseptsiooni kui terviku autor ei ole Evans, seda on pakkunud ja arutanud ka mitmed evolutsiooniteoreetikud): 1) tänapäevainimese eellaste sotsiaalses evolutsioonis toimunud muutused tekitasid vajaduse muutunud keskkonnale vastava suhtlusvahendi järele ja määrasid ühtlasi nõuded, mida see pidi rahuldama; 2) bioloogilise evolutsiooni tulemusel kujunesid vajalik neurobioloogiline aparaat ja anatoomilised struktuurid ning 3) kultuuriline evolutsioon mõjutas seda, missuguse konkreetse vormi uus suhtlusvahend – inimkeel – erinevates populatsioonides omandas.

[1] V. Evans, The Language Myth: Why Language is not an Instinkt. Cambridge University Press, 2014. http://www.vyvevans.net/Books

Algtõuge tuli niisiis sootsiumi liikmete vahelistes suhetes ja suhtluses toimunud evolutsioonilisest muutusest. Ja siinses kontekstis ongi esmatähtis, milles see muutus väidetavalt seisnes. See seisnes eelkõige sootsiumi liikmete vaheliste suhete ja neile tugineva koostöö iseloomus, mis ühelt poolt avaldus üksteise individuaalsete soovide-taotluste-eelistuste tundmises ja arvestamises mingi ühisettevõtmise (nt jahilkäigu) kavandamisel ning teostamisel ja teiselt poolt ka tulemuste (saagi) jagamisel (ka nt paljud kiskjad tegutsevad saagi hankimisel koordineeritult, kuid tabatud saagi jagamisel osalenute „panused" ei mängi rolli, kehtib tugevama õigus). Uue liigi eripära esiletoomiseks kasutatav tüüptermin on *cooperative species* ehk siis bioloogiline liik, kelle määratlevaks tunnuseks on koostöö-, ühistegevusvõime. Ühistegevus ei avaldunud mitte ainult ühistes ettevõtmistes, nagu toidu hankimine, vaid ka kogemuste, teadmiste teadlikus jagamises, ja seda indiviidide tasemel.

Siin tulevadki sisse Evansi käsitluse kesksed mõisted: tähendus (ingl *meaning*) ja vaim (*mind*); ja alles nende kaudu keel kui uue liigi suhtlusvahend. Raamatu 11 peatükki on jagatud kolme ossa ning iga osa keskne mõiste on *tähendus*: tähenduse välistamatus (*ineffability of meaning*), tähenduse roll vaimus (*meaning in mind*) ja tähenduse roll keeles (*meaning in language*). Tõsi, eesti keele *tähendus* ei kata siin kaugeltki kõiki inglise mõiste *meaning* tähendusi. Näiteks tegevuste vastena on eesti keeles kasutusel pigem *mõte* (*mis on selle* (*tegevuse*) *mõte?*, *mis mõttega ta seda tegi?*, *mida sa sellega* (*öelduga*) *mõtled?*). Kui jätta detailid kõrvale, siis võib raamatus esitatud arutluskäigu kokku võtta järgmiselt.

- Kõnealuse sotsiaalse süsteemi muutuse aluseks pidi olema muutus selle liikmete mõttesüsteemis, vaimus (*mind*). Sootsiumi liikmeid kui indiviide ja nende tegevusi hakati tajuma või mõtestama teistmoodi (*mind* > *meaning*).
- Milleks sellises uues sotsiaalses keskkonnas tegutsemiseks oli vaja teistsugust suhtlusvahendit? Sest uus mõtte- ja mõistesüsteem liigendas nii maailma kui ka kaaslaste tegevusi teistmoodi. Episoodiline mälu suutis salvestada eri aegadel ja kohtades toimunud sündmused ja näha nende seoseid. Tekkis vajadus mõtestada need seosed, luua maailma uut tüüpi kontseptuaalne mudel, omistades episoodidele mingi terviktähenduse (*meaning-in-mind*).
- Sellise tervikliku teadmise arendamiseks ja kasutamiseks oli vaja sellest kaaslastega rääkida. Kontseptuaalne mudel oli tervikstruktuur. Et teadmisi-arvamusi selle suvalistest osadest teistega arutada (seda eeldas kooperatiivsus), oli vaja oskust ja võimalust tervikstruktuuri tükeldada ning väljavalitud osi vastavalt vajadusele omavahel kombineerida. Ja siin tuligi mängu keel, pakkudes nii mõisteloomisvõimalusi (sõnad) kui ka teadete edastamise vahendeid – lauseid, grammatikat (*meaning-in-language*).
- Ja lõpuks (aga ajalises mõttes kindlasti mitte viimases järjekorras) võttis uus suhtlusvahend paljuski üle ka algsemate suhtlusvormide rolli, sulandudes nendegagi – pakkudes võimalust väljendada keeleliselt ka tundeid, suhtumisi, lihtsalt lobiseda. Et see valdkond kuulub inimloomuse ühte kõige ürgsemasse kihti, siis pole ime, et vastavad väljendusvahendid moodustavad ka kõigi keelte universaalse osa (*mind* > *meaning*, *meaning-in-language*).

Võib ilmselt öelda, et ka sotsiaalmeedias saavad kõik need aspektid kokku.

2. Sotsiaalmeedia analüüs

Kaks raamatut (Farzindar ja Inkpen 2015; Liu 2015) käsitlevad otseselt sotsiaalmeedia analüüsi. Viimastel aastatel on sotsiaalne suhtlus veebis plahvatuslikult suurendanud nn kasutajate

genereeritud sisu mahtu ja variatiivsust. Andmete allikateks on sotsiaalvõrgustikud (Facebook, MySpace), blogid ja mikroblogid (Twitter), foorumid, netikommentaariumid jms. See keelematerjal avab uued võimalused ka sotsiaalse käitumise uurimiseks, mille eesmärk on luua sotsiaalselt intelligentseid suhtlussüsteeme. Sotsiaalmeedia analüüsil on mitmesuguseid praktilisi rakendusi: ettevõtluses (toodete ja teenuste soovitamine, meelelahutustööstus), poliitikas (nt valimistulemuste ennustamine), tervishoius (haiguste avastamine ja ennetamine), kaitses ja julgeolekus (kriiside ennetamine) jm. Tähtsal kohal on kasutaja modelleerimine soo, rahvuse, poliitilise orientatsiooni, terviseseisundi jms alusel.

2.1. Automaatne keeletöötlus sotsiaalmeedia analüüsis

Arefeh Farzindar ja Diana Inkpen rõhutavad oma raamatus „Loomuliku keele automaattöötlus sotsiaalmeedia jaoks", et see valdkond nõuab traditsiooniliste meetodite kohandamist ja uute meetodite arendamist, sest erinevalt traditsioonilisest tekstist on sotsiaalmeedia tekst reaalajas postitatud sotsiaalne vestlus. Tekstid on struktureerimata, kirjutatud erinevate inimeste poolt erinevates formaatides, mitmes keeles ja erinevas stiilis, sageli esinevad keelevead ja släng, mistõttu on enne automaatset sisuanalüüsi vaja tekstide eeltöötlust.

Töö algab tüüpiliselt korpuse kogumisest ja märgendamisest, sest sageli kasutatakse analüüsis nn juhendatud masinõpet (tehisõpet), kus arvuti õpib teksti sisu analüüsima märgendatud treeningandmete alusel. Korpuse tekstide eeltöötlusel eraldatakse sõnavormid, morfoloogilisel analüüsil märgendatakse sõnaliigid jms, leitakse analüsaatorile tundmatud sõnad ja parandatakse õigekirjavead. Seejärel määratakse lausepiirid ja viiakse läbi lausete süntaktiline analüüs, märgendatakse nimeüksused, identifitseeritakse keeled, kui tegu on mitmekeelse tekstiga. Kõigi nende ülesannete täitmiseks saab küll kasutada olemasolevaid loomuliku keele automaattöötluse vahendeid, aga neid tuleb kohandada, nagu rõhutavad raamatu autorid.

Eeltöödeldud andmetest saab leida mitmesugust kasulikku infot: kasutajate geograafilisi asukohti, emotsioone ja hoiakuid, siduda üksteisega nimeüksusi jms. Teksti autori geograafilist asukohta, mis on vajalik näiteks turunduseesmärgil või võimaliku julgeolekuohu tuvastamiseks, saab määrata sotsiaalvõrgustiku teenuse tarbimiseks registreeritud kasutajaprofiili alusel. Vahel aga annavad kasutajad enda kohta valeinfot, mistõttu tuleb asukoha määramiseks teha järeldusi sotsiaalvõrgustiku taristu ja/või kasutaja tekstide sisu põhjal. Tekstides esinevate nimeüksuste sidumise tulemusel tekivad ühendatud andmed (*linked data*) – struktureeritud andmeressursid, mis luuakse semantilise veebi standardite kohaselt ning mida saavad kasutada automaatsed süsteemid.

Emotsioonide, arvamuste ja hoiakute tuvastamisega tegeleb uudne valdkond, mille eestikeelne nimetus võiks olla „hoiakute analüüs", „tundmusanalüüs" või „meelestatuse analüüs" (ingl *sentiment analysis*, ka *opinion mining*[2]); ka

[2] Bing Liu kasutab neid kahte mõistet sünonüümidena, sest tänapäevane hoiakute analüüs rakendab enamasti statistilisi andmekaeve meetodeid, B. Liu, Sentiment Analysis; vt ka E. Vainik, Kuidas õpetada kõnesüntesaatorile empaatiat? Emotsiooni automaatse tuvastuse võimalustest eestikeelses kirjalikus lauses sisalduva info põhjal. – Eesti Rakenduslingvistika Ühingu aastaraamat 2010, nr 6, lk 327–347; B. Ojamaa, Tartu Ülikooli üliõpilaste tagasiside hoiakute analüüs. Bakalaureusetöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut, 2014. http://hdl.handle.net/10062/44256; M. Koit, H. Õim, Modelling communicative space. From human communication to conversational agents. – INTELLI 2015. The Fourth International Conference on Intelligent Systems and Applications, St. Julians, Malta, October 11–16. Toim Ingo Schwab, Leo van Moergestel, Gil Gonçalves. International Academy, Research, and Industry Association (IARIA) XPS Press, 2015, lk 1–5; M. Koit, H. Õim, Suhtlusruum ja selle

„arvamus- ja hinnangu-uuringud"[3]. See on ala, mille eesmärgiks on arvamuste ja tundmuste ekstraheerimine loomuliku keele tekstidest, kasutades arvutuslikke meetodeid.[4] Paljud hoiakute analüüsi meetodid põhinevad statistikal ja masinõppel ning kasutavad mitmesuguseid tunnuseid, nt sõnade või tähtede n-grammid, emotikonid, suurtähe info, sõnaliigi märgend, eitus (lk 50). Populaarsed statistilised meetodid on otsustuspuud (*decision trees*), naiivne Bayes (*naïve Bayes* ehk lühendina NB), tugivektormasinad (*support vector machines*, SVM), maksimaalne entroopia (*maximal entropy*), tingimuslik juhuslik väli (*conditional random field*), lineaarne regressioon (*linear regression*).

Omaette probleemina käsitlevad Farzindar ja Inkpen raamatus sarkasmi ja iroonia avastamist (nt *see on küll hea seade: pärast esimest kasutamist läks rikki*). Esmalt märgendatakse postitused (iroonia/mitte) ja kasutatakse liigitamisel mingit juhendatud masinõppemeetodit. Sellega sarnane on ka rämpsposti (spämmi) tuvastamine.

Sotsiaalmeedia analüüsi perspektiivikatest uurimisküsimustest on raamatus esile tõstetud järgmisi: millest inimesed räägivad sotsiaalmeedias, miks nad teevad postitusi, kuidas nad ennast väljendavad, kuidas on seotud keel ja sotsiaalvõrgustiku omadused; semantiline veeb, ontoloogiad ja valdkonnamudelid kui sotsiaalmeedia mõistmise vahendid; keel läbi vertikaalide (nt suhtlus ametiasutusega *vs.* sõpradega; sellega seostub ka võimu väljendamine, millest tuleb juttu tagapool), osalejate iseloomustus, käitumine sotsiaalmeedias, keele automaattöötluse tehnikad.

2.2. Hoiakute analüüs

Bing Liu käsitleb oma raamatus „Hoiakute analüüs: arvamuste, hoiakute ja emotsioonide kaeve" põhjalikult ühte valdkonda – hoiakute analüüsi: selle olemust, ülesandeid, meetodeid ja rakendusi.

Igasuguse tekstiinfo võib liigitada kaheks: faktid (kindlad teadmised) ja arvamused. Arvamusi on suhteliselt vähe uuritud, sest enne veebi tulekut nappis selleks materjali. Veeb on oluliselt muutnud inimeste vaadete ja arvamuste väljendamise viise, mistõttu on vaja luua automaatseid süsteeme arvamuste tuvastamiseks ja neist koondarvamuse loomiseks. Sellest vajadusest ongi välja kasvanud hoiakute analüüs (ehk hoiakute kaeve). Hoiakute analüüs hõlmab erinevaid uurimissuundi ja uurimisprobleeme. Üks oluline uurimisteema on toodete ja teenuste kohta väljendatud arvamuste kasulikkus.

Raamatu autor väidab, et hoiakute analüüs ei ole loomuliku keele automaattöötluse alamvaldkond, vaid on pigem omaette iseseisev loomuliku keele „minitöötlus", sest siin on tegu semantilise analüüsi probleemiga, kus ei ole vaja teksti täielikku mõistmist (lk 14). Eesmärk on tuletada struktureerimata tekstidest struktureeritud andmed.

Hoiakud ja arvamused erinevad faktidest selle poolest, et nad on subjektiivsed. Subjektiivsus tuleneb mitmest asjaolust. Kõigepealt on inimestel erinevad kogemused. Näiteks keegi ostis teatud tüüpi kaamera ja on sellega väga rahul; tal on positiivne hoiak, meelestatus selle kaamera suhtes. Teine isik aga, kes ostis sama tüüpi kaamera, sai paraku defektiga toote ja tal on negatiivne kogemus ning seetõttu negatiivne hoiak. Teiseks võivad erinevad inimesed näha sama objekti erinevalt. Näiteks isik, kes ostis kaamera enne hinnalangust, on rahulolematu, samal ajal teine, kes ostis samasuguse kaamera pärast hinnalangust, on väga rahul.

Emotsioonid on meie subjektiivsed tundmused ja mõtted. Kui objektiivne lause väljendab faktilist infot, siis sub-

modelleerimine. – Eesti Rakenduslingvistika Ühingu aastaraamat 2016, nr 12, lk 113–124.

[3] H. Pajupuu, Kõne ja teksti emotsionaalsuse statistilised mudelid. – Eesti keeletehnoloogia kolmas konverents. Teesid. Tartu, 2015, lk 42.

[4] B. Liu, Sentiment Analysis, lk xi.

jektiivne lause väljendab mingeid personaalseid tundmusi või arvamusi, eksplitsiitset või implitsiitset positiivset või negatiivset hoiakut.

Kõige enam on uuritud hoiakute ja subjektiivsuse liigitamist. Hoiakute analüüsi käsitletakse kui teksti klassifitseerimise probleemi: 1) hoiakut sisaldav dokument liigitatakse kas positiivset või negatiivset hoiakut väljendavaks (nn dokumendi hoiaku analüüs); 2) lause liigitatakse kas subjektiivseks, objektiivseks või neutraalseks (nn subjektiivsuse liigitamine) ning kui tegu on subjektiivse lausega, mis väljendab hoiakut, siis määratakse, kas see hoiak on positiivne, negatiivne või neutraalne (s.o lause hoiaku liigitamine).

Dokumentide liigitamisel on kasutatud nii juhendatud kui ka juhendamata masinõpet. Juhendatud masinõppe meetoditest on populaarsemad (eespool mainitud) NB ja SVM. Seejuures kasutatakse tunnuseid, nagu näiteks terminid ja nende esinemissagedus, sõnaliikide märgendid, hoiakusõnad ja -fraasid, süntaktilised sõltuvused, eituse esinemine. Juhendamata masinõppel (mis erinevalt juhendatud masinõppest ei vaja eelnevalt märgendatud korpusi) on domineerivateks indikaatoriteks hoiakusõnad ja -fraasid.

Lause subjektiivsuse liigitamisel ja hoiaku määramisel on kasutatud enamasti juhendatud masinõpet. Kitsaskohaks on siin käsitsitöö maht vajalike korpuste märgendamisel, mistõttu püütakse korpusi märgendada automaatselt. Seejuures on leidnud rakendamist „iseenda ülestöötamise" meetod (ingl *bootstrapping*), kus olemasoleva väikese korpuse põhjal moodustatakse juhuslikkuse alusel väga suur hulk uusi valimeid ja leitakse nende statistilised parameetrid, mis peaksid iseloomustama suvalisi (ka korpuses mitteesinevaid) lauseid.

Hoiakute analüüsis kasutatakse sageli märksõnu. Positiivset või negatiivset arvamust väljendavad sõnad (hoiakusõnad ehk polaarsed sõnad) koondatakse leksikoni. Positiivset hoiakut väljendavad näiteks sõnad *ilus*, *hea*; negatiivset hoiakut *inetu*, *halb*, *vilets*. Lisaks sõnadele võib leksikon sisaldada ka fraase ja idioome. Eesti keeles on kasutatud ka termineid „polaarsusleksikon"[5], „emotsioonileksikon", „tundeleksikon"[6].

Leksikoni võib koostada käsitsi või genereerida automaatselt, kasutades selleks kas iseenda ülestöötamise meetodit, mis alustab väikesest hoiakusõnade hulgast („külvist") ja kasutab seejärel selle hulga laiendamiseks mingit olemasolevat sõnastikku (nt Wordnet), või korpusepõhist meetodit, mis vajab algul samuti hoiakusõnade loendit ja leiab siis uusi hoiakusõnu suurest korpusest.

Tunnustepõhine hoiakute analüüs leiab esmalt sihtmärgid, mille kohta lauses väljendatakse arvamust, ja määrab siis, kas arvamus on positiivne, negatiivne või neutraalne. Sihtmärkideks võivad olla objektid, nende komponendid, atribuudid või tunnused. Objektiks võib olla mingi toode, teenus, institutsioon, sündmus vms. Näiteks tootearvustuse lauses identifitseerib analüüs toote tunnused, mida arvustaja kommenteerib, ja määrab, kas kommentaar on positiivne või negatiivne. Nt lauses *selle kaamera patarei kestus on väike* on objektiks kaamera, tunnuseks patarei kestus ja hoiak on negatiivne. Paljud praktilised rakendused vajavad just sellise detailsusega analüüsi, sest toote täiustamiseks peab tootja teadma, milliseid tema komponente või tunnuseid tuleb parandada. Ainult hoiakute ja subjektiivsuse liigitus ei avasta seda infot. Positiivne hoiak objekti suhtes ei tähenda veel, et hindajal on positiivne hoiak objekti kõigi aspektide ja tunnuste suhtes.

Mõnes rakenduses on kasulik identifitseerida ka hoiakukandja, st seda hoiakut väljendanud isik või organisatsioon.

Omaette uurimisprobleem on võrdlevate arvamuste analüüs (nt

[5] S-M. Marran, Sentimentaalne analüüs eestikeelse peavoolumeedia veebiartiklite kommentaaride baasil. Bakalaureusetöö. Tartu Ülikool, arvutiteaduse instituut, 2012.

[6] B. Ojamaa, Tartu Ülikooli üliõpilaste tagasiside hoiakute analüüs, lk 6, 27.

selle kaamera pildi kvaliteet on parem kui *kaameral X* tähendab, et seda kaamerat tuleb eelistada kaamerale X).

Debatid ja kommentaarid esindavad sotsiaalmeedia arvustustest erinevat sisu. Need on dialoogid, milles osalejad vahetavad suhtlusaktide kaudu oma arvamusi ja argumente. Sellised sotsiaalmeedia vormid toovad sisse teist tüüpi hoiakuid, sh nõustumise või mittenõustumise varem väljendatud hoiakutega. Debattide ja diskussioonide analüüs viib ka sotsiaalteaduste sellistesse valdkondadesse nagu suhtlus ja poliitikateadus ning osalejate käitumine. Seisukoha (ingl *stance*) tuvastamiseks debatis (st mitte ainult seda, kas kommentaar on positiivne või negatiivne, vaid missugusel seisukohal on kommentaari autor: kas kommenteeritava hoiaku toetaja või vastane) on kasutatud kahealuselist graafi, mille sõlmedeks on osalejad, ja SVM klassifitseerijat. Selles valdkonnas on töö alles alguses, enamasti tegelevad sellega arvutiteadlased, aga raamatu autori arvates oleks vaja kaasata ka sotsiaalteadlasi, näiteks selleks, et uurida, missugustel seisukohtadel asuti, hoiakute sisu, kirjeldada probleeme, uskumusi poliitika mõjust jne.

Kavatsuste kaeve on teine sotsiaalmeedia analüüsi ülesanne, mis on tihedalt seotud hoiakute kaevega, aga millele pole seni eriti palju tähelepanu pööratud. Ometi on sel suur potentsiaal kommertsrakendusteks. Nii näiteks on kommertshuvid, mida väljendatakse sotsiaalmeedias, kasulikud reklaami- ja soovitussüsteemides (*kavatsen osta uue auto – soovitage*).

Hoiakute analüüsil tuleb arvestada ka rämpspostiga. Pahatahtlikud kasutajad püüavad teisi kasutajaid desorienteerida, postitades nt halva restorani kohta positiivseid arvamusi. Et arvamustest oleks kasu, tuleb rakendada meetodeid rämpsposti filtreerimiseks. Arvamusspämmi uurimine annab lisavõimaluse valetamise uurijatele. Ebaeetilised postitused on veebis kahjuks laialt levinud ja nende avastamine parandaks sotsiaalset kliimat. See uurimissuund on aga alles algusjärgus.

Hoiakute analüüs pakub suuri tehnilisi väljakutseid. Olgugi et juba on käsitletud paljusid alamprobleeme ja pakutud neile erinevaid lahendusviise, pole seni ükski alamprobleem rahuldavalt lahendatud. Pole isegi standardseid lähenemisi ühelegi alamprobleemile, sest meie teadmised on veel piiratud. Selle põhipõhjus on raamatu autori arvates, et siin on tegu loomuliku keele automaattöötluse ülesandega, kus ei eksisteerigi lihtsaid küsimusi. Teine põhjus on, et seni on valdav lahendusviis olnud masinõpe. Efektiivsed masinõppe algoritmid, nt juba mainitud SVM ja NB, töötavad kui „mustad kastid" ja produtseerivad tulemusi, mis ei ole inimesele vahetult arusaadavad. Kuigi nende meetoditega on saavutatud häid tulemusi, teatakse seni veel vähe selle kohta, kas ja kuidas need algoritmid jäljendavad inimese kasutatavaid tunnuste töötlemise protsesse ja miks nad seda teevad just nii.

Olukorra parandamise võimalusena on selles raamatus rõhutatud keeleteadmuse olulist rolli, mis ütleb, missuguseid väljendeid inimesed sageli kasutavad selleks, et oma arvamusi kuuldavale tuua, ja kuidas saab neid väljendeid automaatselt ära tunda. Raamatu autor julgustab lingviste ühinema selle uurimissuunaga. Hoiak on loomuliku keele semantika oluline aspekt, aga samuti praktilise tähtsusega. See peaks pakkuma piisavat motivatsiooni arendada arvutilingvistilist teooriat hoiaku, arvamuse ja nendega seotud mõistete, nagu emotsioon, mood, afekt, kohta. Neid mõisteid on põhjalikult uurinud psühholoogid, neuropsühholoogid ja sotsioloogid, aga nemad keskenduvad inimeste vaimu psühholoogilistele seisunditele, mitte keelelistele konstruktsioonidele, mida kasutatakse selliste seisundite ja tundmuste väljendamiseks. Lisaks rakendustele hoiakute kaeves võib arvutilingvistiline teooria leida rakendust inimese vaimuseisundi analüüsis sotsiaalmeedia postituste alusel ja olla seega kasulik ühiskonnale. Nt koolid

soovivad teada saada oma õpilaste vaimuseisundit, et avastada depressiooni, suitsiidikalduvusi. Kohtunikud soovivad ära tunda potentsiaalseid kriminaale, kes on ohtlikud ühiskonnale.

Sotsiaalmeedia andmete analüüs võimaldab uurijatel mõista paremini inimesi ja ühiskonda. Enamik varasemaid tulemusi sotsiaalteadustes on saadud väikeste laborieksperimentide põhjal. Suurandmete alusel saab aga läbi viia massiivseid uuringuid ja seeläbi paremini mõista inimloomust, aga ka iga inimese individuaalsust. Näiteks Facebooki ja Twitteri postituste analüüs võimaldab varakult avastada ohumärke.

Viimasel kümnendil on saavutatud edu nii uurimistöös kui ka rakendustes, on tekkinud palju firmasid, mis pakuvad hoiakute analüüsi teenuseid, mille järele on ühiskonnas suur vajadus. Ettevõtted on huvitatud sellest, kuidas nende tooteid vastu võetakse; tarbijad soovivad saada parimaid tooteid.

Paljulubavatena on raamatus esile toodud kaks suunda. 1) Arendada uusi masinõppe algoritme, mis õpivad suurandmetest nii üldisi kui ka valdkonnaspetsiifilisi teadmisi ja mida saab kasutada hoiakute analüüsis. 2) Koos uurimistööga tuleb luua uusi praktilisi süsteeme. Ei saa loota, et täiesti automaatsed ja korrektsed lahendused oleksid lähitulevik. Pigem tuleks arendada poolautomaatseid süsteeme. Siin arvatakse olevat vaja arvuti- ja sotsiaalteadlaste koostööd: ühed oskavad töödelda suurandmeid, teised tunnevad sisu. Kuna andmeteks on keeleliselt vormistatud tekstid, siis seostuvad mõlemad probleemid vältimatult ka loomuliku keele automaattöötlusega.

2.3. Mis on Eestis tehtud?

On kogutud mitmeid korpusi, mis sisaldavad sotsiaalmeedia tekste, sh uue meedia korpus[7] (jututubade, uudisegruppide, foorumite ja netikommentaaride tekstid) ja netiallkeelte korpus Tartu Ülikoolis ning eestikeelsete veebilehtede korpus etTenTen[8] Eesti Keele Instituudis. Korpuste abil on uuritud internetikeele iseärasusi[9], kirjakeele jaoks loodud morfoloogilise analüsaatori kohandamist[10] ning internetis publitseeritud arvamusartiklite kohta avaldatud kommentaaride liigitamist ja suhtlusruumi kui erilise mentaalse ruumi mõõtmete väärtuste liigitamist positiivseteks, negatiivseteks ja neutraalseteks, eesmärgiga edaspidi kasutada hoiakute analüüsi[11].

2006. aastal algas Eestis kõneemotsioonide süstemaatiline uurimine, kui riikliku keeletehnoloogia programmi raames alustati Eesti Keele Instituudis emotsionaalse kõne korpuse loomist.[12]

[7] https://keeleressursid.ee/et/keeleressursid-cl-ut/korpused/83-article/cluteelehed/212-koondkorpus-uus-meedia

[8] https://www.sketchengine.co.uk/etTenTen-corpus/

[9] A. Oja, Eesti keel internetis. – Keel ja arvuti. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6.) Tartu, 2010, lk 259–267; K. Soodla, Morfoloogilisi, morfosüntaktilisi ja sõnamoodustuslikke erijooni eesti internetikeeles. Magistritöö. Tartu Ülikool, filosoofiateaduskond, eesti keele osakond. Tartu, 2010. http://hdl.handle.net/10062/15263; K. Kerge, Veebikommentaariumi mitmetahuline maailm. – Tekstid ja taustad III. Lingvistiline tekstianalüüs. Tartu: TÜ Kirjastus, 2004, lk 51–73; S. Salla, Jututuba kui võrgusuhtlusvorm. – Tekstid ja taustad. Artikleid tekstianalüüsist. Tartu: TÜ Kirjastus, 2002, lk 128–156.

[10] K. Muischnek, H-J. Kaalep, R. Sirel, Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. – Eesti Rakenduslingvistika Ühingu aastaraamat 2011, nr 7, lk 111–127.

[11] M. Koit, Debate formed by Internet comments: Towards the automatic analysis. – Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD), Lissabon. Toim A. Fred, J. Dietz, D. Aveiro, K. Liu, J. Filipe. Portugal: SCITEPRESS – Science and Technology Publications, LDA, 2015, lk 328–333; M. Koit, H. Õim, Suhtlusruum ja selle modelleerimine.

[12] H. Pajupuu, Emotsioonid – kõnetehnoloogia olevik ja tulevik. – Keel ja Kirjandus 2012, nr 8–9, lk 629–643; vt ka R. Altrov, The Creation of the Estonian Emotional

Eesmärgiks on nii kirjaliku teksti emotsionaalsuse tuvastamine, mis võimaldaks kõnesünteesi puhul kindlaks teha loetava teksti emotsionaalsuse ja seda sünteeskõnes arvesse võtta, kui ka kõne emotsionaalsuse tuvastamine, mis võimaldaks arvutil sellele adekvaatse emotsiooniga vastata.[13]

Emotsiooni automaatset tuvastamist eestikeelses kirjalikus lauses sisalduva info põhjal uurinud Ene Vainik tõdeb, et „automaatse emotsioonituvastaja loomisel ei ole võimalik otse üle võtta välismaiseid eeskujusid, kuna emotsioonide väljendamine on mõningal määral kultuurisidus ja nende väljendus kirjalikus tekstis ripub ära konkreetse keele eripärast".[14]

Keeletehnoloogia riikliku programmi käigus on loodud veebis kasutatav kirjaliku teksti polaarsuse määraja,[15] mis märgendab eestikeelses tekstis positiivset ja negatiivset hoiakut väljendavad sõnad ning leiab ühtlasi, kas tekst tervikuna kannab positiivset või negatiivset hoiakut. Emotsioonidetektor annab teksti emotsioonile nii leksikonipõhise kui ka statistilise hinnangu. Leksikonipõhises meetodis kasutab detektor reegleid ja valentsileksikoni, mis sisaldab sagedasemaid eesti keele emotsioonisõnu.[16] Statistilise klassifitseerija treenimiseks on kasutatud juhendatud masinõpet (NB-meetod) ja valentsikorpust – märgendatud tekstilõigukorpust, mida inimesed on hinnanud positiivseks, negatiivseks, neutraalseks või vastuoluliseks.[17]

Emotsioonidetektori rakendus – veebis toimiv meediainspektor[18] – kogub erinevate Eesti uudisteportaalide artikleid ja uurib nende poliitilist sisu. Poliitiliste artiklite pealkirjadele ja juhtlõikudele tehakse emotsioonidetektori abil hoiakute analüüs, mille tulemusena liigitatakse artikkel positiivseks, negatiivseks või neutraalseks.

Hoiakute analüüsi on rakendatud mitmes üliõpilastöös. Siim-Toomas Marran on koostanud programmi uudisetekstide kommentaaride analüüsimiseks sõnastiku ning statistilise klassifitseerija põhjal. Rakenduses kasutatakse inglise keele jaoks loodud vabavara, mis põhineb Bayesi klassifitseerijal.[19] Birgitta Ojamaa on loonud Tartu Ülikooli õppeinfosüsteemist saadud üliõpilaste tagasiside korpuses väljendatud hoiakuid analüüsiva programmi. Programm kasutab eestikeelsete positiivsete ja negatiivsete sõnade leksikoni ja regulaaravaldistena esitatud reegleid.[20] Sama lähenemisega on hiljem uuritud hoiakuid esmakohtujate kõnelustes.[21]

Speech Corpus and the Perception of Emotions. (Dissertationes linguisticae Universitatis Tartuensis 18.) Tartu: University of Tartu Press, 2014.

[13] H. Pajupuu, Kõne ja teksti emotsionaalsuse statistilised mudelid.

[14] E. Vainik, Kuidas õpetada kõnesüntesaatorile empaatiat?; vt ka E. Vainik, Kuidas määrata eesti keele sõnavara tundetoone? – Eesti Rakenduslingvistika Ühingu aastaraamat 2012, nr 8, lk 257–274; E. Vainik, T. Kirt, H. Orav, Conceptual co-presence of motion and emotion in the Estonian terms of personality. – Eesti Rakenduslingvistika Ühingu aastaraamat 2010, nr 6, lk 349–368; E. Vainik, T. Kirt, The structure of Estonian concepts of emotion: A self-organizational approach. – Trames: Journal of the Humanities and Social Sciences 2008, nr 4, lk 382–399.

[15] H. Pajupuu, R. Altrov, J. Pajupuu, Identifying polarity in different text types. – Folklore. Electronic Journal of Folklore 2016, nr 64, lk 25–42. http://www.folklore.ee/folklore/vol64/polarity.pdf; H. Pajupuu, Kõne ja teksti emotsionaalsuse statistilised mudelid; Emotsioonidetektor. http://193.40.113.56:5000/valence

[16] H. Pajupuu, K. Kerge, R. Altrov, Lexicon-based detection of emotion in different types of texts: Preliminary remarks. – Eesti Rakenduslingvistika Ühingu aastaraamat 2012, nr 8, lk 109–122.

[17] H. Pajupuu, Kõne ja teksti emotsionaalsuse statistilised mudelid.

[18] Meediainspektor. http://www.meediainspektor.ee/

[19] S-M. Marran, Sentimentaalne analüüs.

[20] B. Ojamaa, Tartu Ülikooli üliõpilaste tagasiside hoiakute analüüs.

[21] B. Ojamaa, P. K. Jokinen, K. Muischnek, Sentiment analysis on conversational texts. – Proceedings of NODALIDA, 2015, lk 233–237. http://www.aclweb.org/anthology/W/W15/W15-1829.pdf

Tanel Pärnamaa uurib ühe probleemina lausete meelsuse (viha või rõõmu) tuvastamist, kasutades eesti emotsionaalse kõne korpust ja tehisnärvivõrke. Pärnamaa rakendab keeltevahelist siirdeõpet (ingl *transfer learning*), üldistades olemasolevat ingliskeelsete tekstide liigitajat eesti keelele. Aluseks võetakse ajaleheartiklite pealkirjade ingliskeelne andmestik, kus on märgendatud viha ja rõõm. Keeltevaheline siirdeõpe võib töö autori arvates olla perspektiivikas hoiakute analüüsis, sest sageli ei ole vajalikku eestikeelset korpust olemas, samas võib leiduda mitu sobivat võõrkeelset andmestikku.[22]

(järgneb)

HALDUR ÕIM, MARE KOIT

[22] T. P ä r n a m a a, Piltide automaatne kirjeldamine eesti keeles – visuaalse ja semantilise ühisesituse õppimine neurovõrkudega. Magistritöö. Tartu Ülikool, matemaatilise statistika instituut, 2015. http://dspace.ut.ee/handle/10062/47568