CLARA

Sjálfvirk textagreining við mat á persónuleika

Guðmundur Gunnlaugsson

Lokaverkefni til BS-gráðu
Sálfræðideild
Heilbrigðisvísindasvið

HÁSKÓLI ÍSLANDS

CLARA

Sjálfvirk textagreining við mat á persónuleika

Guðmundur Gunnlaugsson

Lokaverkefni til BS-gráðu í sálfræði

Leiðbeinandi: Jakob Smári

Sálfræðideild

Heilbrigðisvísindasvið Háskóla Íslands

Júní 2010

Prentun: Háskólaprent

Reykjavík, Ísland 2010

Efnisyfirlit

Útdráttur

Greining á tungumáli fólks getur gefið vísbendingar um einstaklingsmun. Noktun textagreiningartóla hefur aukist yfir árin og hafa rannsóknir sýnt fram á tengsl milli orðanotkunar og persónuleikaþátta. Þessar rannsóknir hafa leitt í ljós mikilvægi smáorða, þá sérstaklega persónufornafna, og áhrifa mismunandi persónuleikaþátta á notkun þeirra. Nýlegur íslenskur hugbúnaður að nafni CLARA (Collective Large-scale Affect Research and Analysis) hefur tæknilega sérstöðu sem gæti bætt þekkingu á tungumálagreiningu. Þessi sérstaða felst í n-gram (*n-grams*) nálgun að textagreiningu, sem leyfir hugbúnaðinum að greina orð saman í stað þess að telja einstök orð. Þessi aðgerð opnar möguleikann á samanburðargreiningu, samhengisgreiningu og einræðingu (*word sense disambiguation*). Með því að beita samanburðargreiningu er hægt að sjá einstök mynstur í orðanotkun mismunandi hópa. Samhengisgreiningu er hægt að nota til þess að greina orðanotkunarmynstur innan ákveðins hóps. Einræðing er síðan notuð til þess að þyrpa (*cluster*) saman fólk í kringum ákveðin mynstur. Þessi tæknilega sérstaða opnar nýja möguleika í því að tengja saman textagreiningu við þekkingu á persónuleikaþáttum. Stefnan er að nýta CLARA til þess að veita notendum betri þekkingu á viðskiptavinum þeirra. Til þess að það sé hægt eru margar hindranir sem þarf að komast yfir ásamt því að framundan þarf að prófa kerfið og sýna fram á áreiðanleika og réttmæti CLARA. Lagt er til að fyrst verði sýnt fram á próffræðilega eiginleika kerfisins, þar næst verða rannsóknir annarra endurteknar til þess að staðfesta sambærilega getu og að lokum verður hafist handa við að prófa nýja þekkingu.

Inngangur

Það eru engir tveir einstaklingar eins en sem betur fer berum við öll ákveðin merki sem gefa í skyn hver við erum. Fingraför eru eitt slíkt merki. Í dag eru fingraför notuð til þess að leysa sakamál eða opna læstar dyr og það er hægt vegna þess að engin tvö okkar eru með eins fingrafar. Það eru fleiri merki sem greina okkur frá öðrum og er persónuleiki áreiðanlega eitt flóknasta samansafn slíkra merkja. Persónuleiki er eitthvað sem við myndum okkur skoðun um hjá einstaklingum strax og við hittum þá í fyrsta skipti. Í sumum tilvikum byggist skoðunin á fordómum einum saman, í öðrum tilvikum á frekar nákvæmu mati. Fólk hópar svipaða einstaklinga saman og ber saman við annan hóp. Þetta ferli að flokka einstaklinga í hópa eftir ætluðum persónuleika þeirra er nánast ósjálfrátt og við gerum þetta oft á dag.

Hugmyndin um að fólk sé mismunandi er ekki ný af nálinni. Áhuga manna á persónuleika og einstaklingsmun má auðveldlega rekja aftur um rúmlega 2000 ár, meðal annars til Hippocratesar og Galen. Það var með þessum þekktu Forngrikkjum sem eins konar flokkunarfræði með tilliti til einstaklingsmunar hófst á Vesturlöndum (McAdams, 1997). Gríski heimspekingurinn Theophrastus lýsti í bókinni *Manngerðir* sinni sýn á einstaklingsmun. Bókin, sem var skrifuð á fjórðu öld fyrir Krist, innihélt teikningar af mismunandi persónuleikagerðum, svo sem *heigullinn, smjaðrarinn, ruddinn* og fleiri (Gleitman, Fridlund og Reisberg, 2004). Fyrstu leikritshöfundarnir áttu líka stóran þátt í mótun skilnings mannsins á einstaklingsmun. Sem dæmi þá er orðið *persónuleiki (personality)* myndað úr orðinu *persóna (persona)*, sem var nafnið á grímunum sem grískir og rómanskir leikarar notuðu til þess að sýna hvaða persónur þeir voru að leika (Allport, 1937).

Þrátt fyrir þennan áhuga gerðist lítið í persónuleikasálfræði þar til á 20. öldinni og áttu margir fræðimenn þátt í því að móta hana í barnæsku hennar. Comte (1852) sá persónuleikasálfræðina, sem hann kallaði *la morale*, sem vísindagrein sem átti að líta á einstaklinginn í heild sinni. Allport (1954) benti á að það sem Comte sá fyrir sér var í raun persónuleikasálfræðin (Í McAdams, 1997). Comte var ekki einn um að leggja áherslu á heildræna sýn á fólk, en sú áhersla sést sömuleiðis í skrifum James (1980) og Freud (1921/1961), þar sem þeir skrifuðu um einstaklinginn (Í McAdams, 1997). Í kringum 1940 varð persónuleikasálfræði skýrt

afmarkaður hluti af félagsvísindum. Sviðið var þá leitt af fræðimönnum á borð við Gordon Allport (1937), Henry Murray (1938) og Kurt Lewin (1935). Allport var einn af þeim sem mótaði framtíð persónuleikasálfræði. Hann taldi persónuleika vera breytilegan þátt sem ákvarðar hvernig einstaklingurinn hagar sér, hugsar og bregst við umhverfi sínu (Í McAdams, 1997).

Á þeim árum sem hafa liðið frá því að þessir menn settu fram sínar hugmyndir um persónuleika hefur fjöldinn allur af kenningum verið settur fram, fjölmörg persónuleikapróf hönnuð og fjölmargar bækur skrifaðar um persónuleikasálfræði. Ein þeirra hugmynda sem hefur náð nokkurri fótfestu fjallar um tengsl tungumáls við persónuleika. Það sem rannsakendur hafa haft áhuga á að kanna er hvernig tungumálið – myndun þess, uppbygging og notkun – er tengt persónuleika. Geðlæknirinn Southard (1916) tók eftir og benti á tengsl milli notkunar einstaklinga á málfræðiháttunum (boðhætti, framsöguhætti, viðtengingarhætti og óskhætti) og hefðbundnu skapgerðanna (bráðlyndi, rólyndi, þunglyndi og léttlyndi). Piaget (1928; 1932) tók eftir að myndun tungumáls hjá börnum getur endurspeglað innri hvatir og tilfinningastig þeirra (Í Sargent, 1945). Í gegnum árin hafa rannsakendur á borð við Gottschalk og Gleser (1969), Rosenberg og Tucker (1978) og Stiles (1992) fundið vísbendingar sem benda til þess að líkamleg heilsa og geðheilsa fólks er tengd orðavali þeirra (Í Pennebaker, Chung, Ireland, Gonzales og Booth, 2007).

Fram á síðustu ár hefur greining á texta og tungumáli þótt vera hægvirk, flókin og of kostnaðarsöm aðferð til rannsókna í sálfræði. Oftar en ekki þurfti að beita mjög nákvæmum og flóknum aðferðum við að handmerkja texta (Chung og Pennebaker, 2007). Það kom ekki í veg fyrir að textagreining væri nýtt innan sálfræðinnar. Það þótti lengi vel, og þykir enn, gagnleg aðferð að beita textagreiningu til að greina sjúkdóma byggt á tungumálanotkun einstaklinga. Gottschalk og Gleser (1969) og Walter Wintraub (1989) hafa verið leiðandi á því sviði og aðferðir þeirra miða að því að hópa saman orð í ákveðna flokka sem gefa vísbendingar um hvaða sjúkdómar hrjá einstaklinga (Í Pennebaker og King, 1999). Önnur nálgun textagreiningar sem hefur verið vinsæl er kerfisbundin orðatalning. Báðar þessar aðferðir byggjast á þeirri hugmynd að ef maður er að reyna að tjá gleði með orðum, þá er hann líklegri til þess að nota orð eins og glaður, hamingja, árangur eða bros. Það hafa nokkur tölvuforrit verið þróuð með það að markmiði að

greina texta kerfisbundið. Þessi forrit á borð við *The General Inquirer, DICTION, LSA (Latent Semantic Analysis)* og fleiri, nýta mismunandi aðferðir til þess að svara spurningunni um hvort hægt sé að greina einstaklingsmun úr texta (Chung og Pennebaker, 2008; Pennebaker og King, 1999). Þessi forrit hafa náð ágætum árangri til dæmis við að greina raskanir á borð við líkömnunarraskanir, geðklofa, sjálfsvígshvatir og þunglyndi. Aukin þekking á textagreiningu með tölvum hefur haft í för með sér þróun á öflugri forritum sem lofa góðu við rannsóknir á einstaklingsmun. Eitt þeirra er LIWC (*Lingustic Inquiry and Word Count*; Pennebaker, Booth o.fl., 2001; Pennebaker, Chung o.fl., 2007) sem byggist á orðtíðnigreiningu. LIWC hefur verið notað til þess að sýna fram á tengsl milli orðavals við batnandi líkamlegrar heilsu, við persónuleikaþætti svo sem úthverfu og finna mun á tungumálastíl stjórnmálamanna og tengsl þeirra við tungumálastíl þunglyndra (Argamon, Dhawle, Koppel og Pennebaker, 2005; Pennebaker, Mayne og Francis, 1997; Slatcher, Chung, Pennebaker og Stone, 2006) svo eitthvað sé nefnt. Íslenskur hugbúnaður að nafni CLARA hefur verið í þróun síðastliðin tvö ár og lofar tæknileg sérstaða hans góðu hvað varðar greiningu á flóknari mynstrum í tungumálanotkun og tengsl þeirra við persónuleika.

Þetta verkefni miðar að því að útskýra ólíkar hugmyndir sem menn hafa haft um hvernig eigi að nálgast persónuleika og greiningu á honum. Þessar ólíku aðferðir eru kynntar og hvaða tól hafa verið mótuð með þeim til þess að meta persónuleika og hvaða árangur hefur hlotist þar af. Lögð er sérstök áhersla á þær aðferðir, tól og niðurstöður sem snerta greiningu tungumáls. Þegar umfjölluninni um þessar ólíku aðferðir er lokið mun hugbúnaðurinn CLARA fá sérstaka umfjöllun þar sem rætt er um nýja möguleika í greiningu texta og hvaða þýðingu það gæti haft fyrir persónuleikarannsóknir sem beita textagreiningu.

1 Rannsóknir á persónuleika

Það hafa margar hugmyndir, kenningar og tilgátur verið settar fram um hvað persónuleiki er, hvernig eigi að meta persónuleika og hver tengsl persónuleika við hegðun séu. Í áranna rás hafa nokkrar mismunandi stefnur orðið til þar sem mismunandi aðferðafræði er beitt. Þessar stefnur hafa ólíkan hugmyndafræðilegan bakgrunn, en oft er leitast við að svara sömu spurningunum. Þegar kemur að því að meta persónuleika með aðstoð textagreiningar byggist það aðallega áþví sem hefur

verið kallað orðasafnsnálgunin. Frávarpsnálgunin að persónuleika byggir að mörgu leyti á svipuðum hugmyndum, það er að segja að orðaval okkar gefi vísbendingar um persónuleika. Markmiðið með þessum kafla er að kynna þá hugmyndir sem þessar aðferðir byggjast á, þau tól sem hafa upp úr þeim sprottið við mat á persónuleika og þá gagnrýni sem þær hafa orðið fyrir. Sérstök áhersla er lögð á orðasafnsnálgunina og farið dýpra í þær niðurstöður sem hafa fengist með aðstoð textagreiningatóla. Að lokum verður síðan dregin upp mynd af því hvernig textagreiningatól gætu nýst til þess að koma til móts við mikið af þeirri gagnrýni sem orðasafnsnálgunin og frávarpsnálgunin hafa orðið fyrir.

1.1 Orðasafnsnálgun

Ein meginhugmynd um persónuleika og mat á persónuleika fjallar um það hvernig einstaklingsmunur kemur fram í orðum. Goldberg (1981) skilgreindi orðasafnstilgátuna (*lexical hypothesis*) sem „Sá einstaklingsmunur sem skiptir hvað mestu máli í daglegum samskiptum fólks mun á endanum verða táknbundinn í tungumáli þess. Því mikilvægari sem þessi munur er, því líklegra er fólk til þess að taka eftir honum og vilja tala um hann, sem veldur því að á endanum vill það búa til orð fyrir hann.“ (Í Ashton og Lee, 2005).

Orðasafnsnálgunin á upphaf sitt í rannsóknum Allport og Odbert (1936) sem héldu því fram að tungumál væri mikilvægt fyrir lýsingu persónuleika (Í Chung og Pennebaker, 2008). Allport og Odbert leituðu uppi öll þau orð sem áttu við einstaklingsmun. Þessi listi þeirra innihélt 17.953 orð og í framhaldi tóku þeir út öll þau orð semvoru margræðin, óskýr, úrelt, útlitslýsingar sem og orð sem lýstu tímabundnu ástandi. Samkvæmt Saucier og Goldberg (1996) var niðurstaðan af þeirri vinnu, listi af persónulýsingum sem var byggður upp af lýsingarorðum (Í Chung og Pennebaker, 2008).

Vinna Allport og Odbert mótaði hvernig rannsakendur líta á mögulega notkun orðasafna í rannsóknum á persónuleika. Þessi vinna þeirra leiddi af sér persónuleikakenningu sem var þróuð af Cattell (1957) og byggist á orðum sem fólk notar til að lýsa öðru fólki. Líkt og Allport og Odbert byrjaði Cattell með mikinn fjölda orða en stytti listann með því að taka út samheiti, slangur og óþjál orð. Að því loknu hafði hann útilokað stærsta hluta orðanna og eftir stóðu 171 þáttanöfn. Hann bað hópa af einstaklingum að meta fólk í ljósi þeirra þátta. Þannig fékk Cattell

(1966) fram það sem hann leit á sem 16 persónuleikaþætti sem náðu utan um öll þessi 171 þáttanöfn (Í Gleitman o.fl., 2004). Tupes og Cristal (1961), Norman (1963) og Goldberg (1993) áttu sinn þátt í því að vinna með 16 persónuleikaþætti Cattells og mynda það sem í dag er þekkt sem *hinir stóru fimm (Big Five*) (Í Gleitman o.fl., 2004). Þessi vinna Cattells og þeirra sem fylgdu í kjölfarið er oftar en ekki kölluð þáttanálgunin (*trait approach*). Sú nálgun byggir á því að hægt er að lýsa persónuleika einstaklinga með þáttum, eins og hinum fimm stóru, og að samspil þessara þátta geti myndað alla þá persónuleika sem við sjáum hjá fólki.

Orðasafnsnálgunin snérist ekki bara um þætti heldur tengist einnig meiri greiningu á sjálfu tungumálinu. Skilgreiningin á orðasafnstilgátunni gaf einnig möguleikann á að skoða sjálft orðavalið og sjá hvaða vísbendingar væri að finna þar. Náttúrulegt tungumál hefur ekki verið skoðað mikið innan sálfræðinnar og í gegnum árin hefur notkun tungumáls ekki verið talin spegla einstaklingsmun. Framfarir í rannsóknum hafa leitt okkur nær þeim skilningi að það hvernig fólk tjáir sig er ótrúlega stöðugt yfir tíma og aðstæður (Pennebaker og King, 1999). Það gerir rannsóknir á einstaklingsmun með nýtingu textagreiningar að raunhæfum möguleika. Í dag er litið á tengsl tungumáls við persónuleika á þann veg að notkun tungumáls og orðaval endurspegla sálræn ferli og gefa vísbendingu um hugarástand, tilfinningar og persónuleika einstaklinga (Argamon, Dhawle, Koppel og Pennebaker, 2005). Rannsakendur nýta sér framfarir í textagreiningu, gervigreind og þekkingu á tungumálasálfræði til þess að greina málvísindaleg merki sálræns ástands.

Þættir og greining á texta eru ólíkar nálganir, sem byggjast á hugmyndinni að einstaklingsmunur endurspeglist í tungumáli fólks. Þáttanálgunin hefur fengið meiri athygli og verið meira rannsökuð. Það gerir það að verkum að niðurstöður persónuleikaprófa, sem byggjast á þáttanálguninni, eru notuð til samanburðar við niðurstöður textagreiningar. Næst verður ítarlegar greint frá þessum tveimur leiðum í rannsóknum.

1.1.1 Þættir

Þáttanálgunin að persónuleika er einn armur orðasafnsnálgunarinnar. Í raun hefur þessum hugtökum, þáttanálgun og orðasafnsnálgun, verið víxlað í gegnum árin og þau notuð til að lýsa sömu hugmyndafræðinni. Eins og komið hefur fram byggir

hún á þeirri forsendu að helstu víddir persónuleika ættu að vera táknaðar með lýsingarorðum í tungumáli einstaklinga. Fjölbreytileiki persónuleika býr til vandamál sem snúa að því að skilja mikilvæga eiginleika persónuleika, svo sem myndun hans, þróun og afleiðingar (Asthon og Lee, 2005). Þessi vandamál leiddu til þess að rannsakendur tóku til við að skoða fáar víddir persónuleika, sem gætu náð að útskýra þessa mikilvægu eiginleika persónuleika. Til þess að ná því fram var farið í þá vinnu að þróa próf sem áttu að meta þessar víddir.

Persónuleikaprófin sem hafa sprottið upp úr þáttanálguninni eru frekar þekkt. Þau byggjast öll á því að hægt sé að lýsa fjölbreytileika persónuleika með því að lýsa fólki út frá stöðu þeirra á víddum. Þessar víddir eru á samfellu og samspil þeirra er það sem býr til þá ótrúlegu flóru persónuleika sem við sjáum í daglegu lífi. Þessi próf miða að því að spyrja próftakann að spurningum, sem oftar en ekki byggjast á staðhæfingum um daglegt líf. Dæmi um spurningu væri til dæmis að athuga hversu sammála próftakinn væri eftirfarandi fullyrðingu: Mér líður oft óþæginlega þegar ég hitti nýtt fólk. Cattell þróaði eitt af fyrstu prófunum, sem átti að meta alla 16 persónuleikaþættina sem hann lagði til. Það próf er kallað *Sixteen Personality Factor Questionnaire* og samanstendur af spurningum sem byggja á raunhegðun einstaklinga (Í Cattell og Cattell, 1993). Þáttakenningin um hina fimm stóru hafði líka mikil áhrif á þróun persónuleikaprófa. Þau próf miða að því að mæla þættina úthverfa (*extroversion*), taugaveiklun (*neuroticism*), samvinnuþýði (*agreeableness*), samviskusemi (*conscientiousness*) og víðsýni (*open to experience*, Goldberg, 1993). Eitt algengasta prófið sem metur hina fimm stóru er *NEO PI-R* sem kom út árið 1992 og var hannað af þeim Costa og McCrea (Í Costa og McCrea, 2002). Hans og Sybil Eysenck (1975) lýstu í kenningu sinni að persónuleiki væri samspil þriggja þátta. Þau þróuðu *persónuleikaspurningalista Eysenck* (*Eysenck Personality Questionnaire*) til þess að meta úthverfu, taugaveiklun og harðlyndi (*psychoticism*) einstaklinga. Fjölmörg önnur próf hafa verið þróuð til þess að meta persónuleika fólks. Sömuleiðis hafa komið út fleiri útgáfur af ofangreindum prófum, sum miðuð að börnum en önnur koma til móts við ókosti fyrri prófa.

Þáttanálgunin og persónuleikaprófin sem voru þróuð í framhaldi af þeirri kenningueru mjög mikilvæg. Þau gefa gagnlega sýn á persónuleika einstaklinga og hægt er að nota prófin til samanburðar við aðrar mælingar. Hægt er að sýna fram á réttmæti annarra mælinga með því að bera saman við viðurkenndari

persónuleikapróf. Það felur í sér að þau persónuleikapróf sem borið er saman við þurfa að vera áreiðanleg og réttmæt. Mikið af rannsóknum hafa snúist um það og eru margir þeirrar skoðunar að þáttaprófin séu áreiðanleg og réttmæt (Ashton og Lee, 2005; Goldberg, 1993). Þrátt fyrir það eru ekki allir sammála um það og gagnrýnt orðasafnsnálgunina (sjá betur kaflann *Gagnrýni á orðasafnsnálgunina* hér að neðan).

1.1.2 Tungumálagreining

Þáttanálgunin er þess eðlis að það þarf að búa til spurningalista til þess að meta persónuleika. Það er einmitt þessi þörf sem aðgreinir þáttanálgunina frá hinum armi orðasafnsnálgunarinnar, sem er tungumálagreining. Grunnhugmynd tungumálagreiningar er sú að orðin sem fólk notar endurspegli hvaða mann það hefur að geyma. Þessi hreyfing sálfræðinnar byrjaði um 1950 meðal lítils hóps af rannsakendum sem komust að því að hvernig fólk talaði væri tengt líkamlegum og hugrænum heilsuvandamálum (Argamon o.fl., 2005).

Tungumálagreining byggist á hugmyndum tungumálasálfræðinga og því sem hefur verið kallað tölvunarstílfræði (*computational stylistics*). Tölvunarstílfræði lítur á heildar merkingu texta sem mun meira en viðfangsefnið sem textinn lýsir. Merking textans finnst einnig með því að horfa á tilfinningavægi hans, til hvers textinn talar, hvert markmiðið með textanum er og hvernig einstaklingur skrifaði textann (Argamon o.fl., 2005). Það eru þessir þættir merkingar textans sem koma fram í *stíl* textans. Við getum þannig borið saman muninn á *hvernig* textinn er skrifaður (stíll) og *hvað* kemur fram (viðfangsefni).

Það er þessi munur á stíl og viðfangsefni sem er ótrúlega mikilvægur í því ferli að skilja einstaklingsmun. Til þess að sýna fram á hvernig stíll og viðfangsefni geta gefið góða vísbendingu um einstakingsmun gáfu Pennebaker, Mehl og Niederhoffer (2002) eftirfarandi dæmi:

> Munurinn á milli tungumálastíls og merkingar sést vel í hvernig tveir einstaklingar koma með einfalda bón. „Væri möguleiki á því að þú gæti rétt mér saltið?“ og „Réttu mér saltið,“ tjá bæði ósk ræðumannsins um salt og stýra hegðun áheyrandans. Hinsvegar, gefa þessir frasar upp mismunandi þætti í sambandi ræðumannsins og áheyrandans,

persónuleika ræðumannsins og jafnvel hvernig ræðumaðurinn lítur á sjálfan sig (bls. 548).

Tveir menn eða fleiri geta verið að fjalla um sama efnið en hvernig þeir nálgast það (stíllinn) sýnir muninn á milli þeirra. Með því að skoða stíl þarf að hætta að horfa á orðin sjálf einvörðungu og spá í orðflokkunum og notkun þeirra.

1.1.2.1 Aðferðir

Núverandi aðferðum við greiningu texta er hægt að skipta í þrennt. Greining á efni byggt á mati dómenda (*judge-based thematic content analysis*) er ein þeirra. Samkvæmt Smith (1992) gengur þessi aðferð út á það að dómendur meta efni textans og nota til þess sérstakt kóðakerfi (Í Pennebaker, Mehl o.fl., 2007). Þessi aðferð hefur verið notuð til þess að skoða fjöldan allan af sálfræðilegum fyrirbærum svo sem myndmál hvatningar, skýringarstíla, margbreytileika hugsunar, geðræn heilkenni og stig hugsunar.

Nýlegri aðferð sem spratt upp úr gervigreindarsamfélaginu er orðamynstursgreining (*word pattern analysis*). Þessi aðferð greinir orðamynstur í texta með því að beita stærðfræðilegum nálgunum til þess að reikna út hvernig orð breytast saman í stóru úrtaki texta. Það er ekki unnið með fyrirfram skilgreindar orðabækur eða orðflokka heldur er unnið frá rótinni og upp (*bottom*-up). Það er að segja, þessir stærðfræðilegu útreikningar finna þessi orðamynstur án aðstoðar. Markmiðið er að sýna fram á þáttabyggingu orðanotkunar innan úrtaksins. Þegar búið er að finna orðamynstur sem móta þessa þáttabyggingu innan úrtaksins er hægt að bera nýtt úrtak af textum við eldra úrtakið. Ef sambærileg orðamynstur finnast í nýja úrtakinu þá er áætlað að það úrtak hafi sömu þáttabyggingu og eldra úrtakið (Pennebaker, Mehl o.fl., 2007). Í flestum tilvikum er þessi aðferð nýtt til þess að kanna hversu sambærilegt innihald tveggja úrtaka af texta er.

Þriðja aðferðin beitir orðatalningarnálguninni (*word count strategy*) sem felur í sér talningu á orðum tengdu efni og stíl textans. Í sumum tilvikum eru þróaðar orðabækur sem flokka orð eftir sálrænum, hugrænum og félagslegum tengslum. Sú flokkun krefst mannlegs mats og tekur oft mjög langan tíma að þróa slíkar orðabækur. Algengast er að nota einfalda talningu á orðum, til dæmis á setningarfræðilegum orðum eða orðflokkum (Pennebaker, Mehl o.fl., 2007).

Þessar aðferðir eru afar ólíkar og hafa leitt af sér ólík textagreiningatól. Öll þeirra hafa sína kosti og galla sem mótast af þessum aðferðum. Hér verður stuttlega farið í þau helstu en sérstök áhersla verður sett á LIWC textagreiningartólið, sem er talið eitt það fremsta í dag.

1.1.2.1.1 LIWC

James W. Pennebaker hefur verið mjög framarlega á sviði rannsókna á tengslum tungumáls við tilfinningalega og líkamlega heilsu (sem dæmi, Cohn, Mehl og Pennebaker, 2004; Pennebaker, 1997; Pennebaker og Beall, 1986). Rannsóknir Pennebaker leiddu til þróunar LIWC árið 1993 og síðan þá hefur LIWC farið gegnum tvær stórar umbreytingar og endurbætur, fyrst árið 2001 (Pennebaker, Booth o.fl., 2001) og svo aftur árið 2007 (Pennebaker og fleiri, 2007). Í grunnatriðum tekur LIWC inn textaskrár og telur tíðni orða í textanum. Það býr til fingraför innan textans með því að flokka orð í fyrirfram skilgreinda flokka og gefur vísbendingar um notkun höfundarins á tungumáli (Pennebaker og Graybeal, 2001). Fyrir hverja textaskrá eru um það bil 80 textabreytur fundnar. Kjarninn í textagreiningunni er svo LIWC2007 orðabókin sem samanstendur af næstum 4.500 orðum og orðstofnum. Hvert orð tilheyrir nokkrum flokkum, sem dæmi tilheyrir orðið *grét* flokkunum: sorg, neikvæð tilfinning, tilfinning, sögn og sögn í þátíð. Í hvert skipti sem LIWC finnur orðið grét í textaskrá hækka allir þessir flokkar um einn. Þegar búið er að fara í gegnum allan textann fyrir ákveðinn höfund sést hvaða flokka hann notar mest (sjá Pennebaker og fleiri, 2007, fyrir nákvæma útlistun á virkni LIWC og þróun LIWC2007 orðabókarinnar). Þessa flokka er að finna í sérhönnuðum orðabókum, sem metnar hafa verið af dómendum. Hver flokkur hefur í flestum tilvikum farið í gegnum nokkrar umferðir þar sem dómendur meta hvort ákveðin orð tilheyra flokknum og voru dómarar sammála í 93-100% tilvika í síðustu umferðinni (Tausczik og Pennebaker, 2009).

Einn kostur LIWC er getan til að flokka orð í flokka sem hafa verið metnir af dómendum sem sálfræðilega viðeigandi. Tveir flokkar, sem má segja að séu sálfræðilega viðeigandi eru tilfinningavægi (*affect*) og hugræn ferli (*cognitive mechanism*). Flokkurinn tilfinningavægi inniheldur orð sem talin eru neikvæð eða

jákvæð en flokkurinn hugræn ferli inniheldur orð sem talin eru lýsa innsýn eða öryggi sem dæmi (Pennebaker, Chung o.fl., 2007). Þetta gefur mun skýrari mynd af uppbyggingu tungumáls hjá fólki, en þessi flokkun yrði ekki möguleg án aðstoðar dómenda. Helsti ókostur LIWC er hversu takmörkuð samhengisgreining þeirra er. Kerfið byggir algjörlega á að telja einstök orð og þar af leiðandi getur það ekki áttað sig á samhengi textans. Þetta er kallað að greina eingröm (*unigram*) en mun betra væri að greina svokölluð n-gröm (*n-grams*), þar sem fleiri en eitt orð eru tekin saman í streng. Með því að greina n-gröm er hægt að sjá út flóknari mynstur í textanum sem hafa mikið að segja um í hvaða samhengi er verið að tala og gæti gefið betri vísbendingar um einstaklingsmun. Annar ókostur LIWC er takmörkuð samanburðargreining. Forritið er ófært um að taka tvo texta og sjá hvaða munur er þar á milli, hvað er sameiginlegt og hvað er ólíkt.

1.1.2.1.2 Önnur textagreiningartól

The General Inquierer. Eitt af fyrstu textagreiningatólunum, sem náði þokkalegum árangri, var *The General Inquirer*, sem er orðtalningar forrit og var þróað árið 1966 af þeim Stone, Dunphy, Smith og Ogilvy (Í Pennebaker og King, 1999). Í áranna rás hafa þrjár orðabækur verið þróaðar fyrir forritið, en ein þeirra þörf-árangur orðabókin (*need-achievement dictionary*) var upphaflega þróuð til þess að koma í staðinn fyrir flókna skorun dómenda á árangurs myndmáli sem kemur fram í TAT-prófinu (*Thematic Apperception Test*, Pennebaker, Mehl o.fl., 2007). Markmiðið var að nota forritið til þess að lesa sögurnar sem komu úr TAT-prófinu og þannig ná betri túlkun með forriti heldur en mannlegum skorunar aðferðum. The General Inquirer er einstaklega sveigjanlegt og getur verið notað til þess að skoða nánast hvaða viðfangsefni sem er. Helsti kostur og ókostur forritsins er að getan til að gera samhengisbundna orðatalningu. Forritið nýtir orðabækur, sem hafa verið metnar af dómurum, til þess að framkvæma þessa samhengisbundnu orðatalningu. Það virkar þannig að orðabækurnar eru hannaðar til þess að finna tvíræð orð og athuga orðin í kring og meta út frá því samhengi til hvaða merkingar orðsins er verið að vísa (Pennebaker, Mehl o.fl., 2007). Það er ótrúlega mikið verk að viðhalda þessum orðabókum, sem eru nýtar í þessari samhengisbundnu orðatalningu, þar sem þær byggja á mati dómara.

The General Inquirer var þróað í áhugaverðum tilgangi. Eins og kom fram hér að ofan þá voru þeir Stone og félagar með það í huga að forritið gæti komið í stað dómara við túlkun á frávarpsprófinu TAT. Með því að nota textagreiningatól til þess að túlka eða greina sögurnar sem koma frá próftökum í TAT er möguleiki á að koma til móts við þá gagnrýni sem það próf hefur fengið.

DICTION. Annað svipað forrit er DICTION, sem einfaldlega telur orð tengd ákveðnu þema og var þróað af Roderick Hart árið 2001 (sjá einnig, Chung og Pennebaker, 2007; Hart, Jarvis, Jennings og Smith-Howell, 2005). DICTION var hannað til þess að greina yfirbragð stjórnmálastaðhæfinga með því að flokka texta eftir fimm breytum (virkni, bjartsýni, vissu, raunsæi og sameiginlegum eiginleikum) sem eru tölfræðilega óháðar (Pennebaker, Mehl o.fl., 2007). Helsti eiginleiki DICTION er geta forritsins til þess að læra, það er að segja, að uppfæra gagnagrunninn sinn með hverjum texta sem er greindur. Tölfræðileg viktun á tvíræðum orðum (*homographs*) er einnig einn af helstu kostum DICTION. Það er almennt talið vera þægilegt í notkun og auðvelt til að byrja greiningu sem er stór kostur.

DICTION er frekar takmarkað forrit. Í fyrsta lagi þá er aðeins hægt að greina 500 orð í einu (Sjá leiðarvísirinn DICTION 5.0, Hart og Carroll, 2000). Þetta gerir það að verkum að fyrir stóra textaskrá þarf að lesa úr nokkrum greiningum til þess að fá heildarmyndina. Í öðru lagi þá var það hannað með mjög afmarkað markmið fyrir augum, að greina yfirbragð stjórnmálastaðhæfinga, og nýtist þá illa við greiningu á einstaklingsmun við aðrar aðstæður.

LSA og TAS/C. Áhugaverð aðferð er LSA (*Latent Semantic Analysis*), sem meðal annars hefur verið notuð af Foltz (1996) við textagreiningu og beitir gervigreind til þess að átta sig á því hvað er líkt með tveimur textum. LSA byggir á orðamynstursgreiningu og beitir þáttagreiningu á einstaka orð og orðamynstur til þess að sjá hversu líkir tveir textar eru. Þann samanburð er hægt að nota til að meta hvort nýr texti hafi sömu uppbyggingu og eldri texti í gagnagrunni og þannig yfirfæra eiginleika eldri textans á nýja textan (Pennebaker, Mehl o.fl., 2007). LSA hefur aðallega verið notað til þess að skoða hversu vel fólk nær upplýsingum úr texta sem það les með því að greina útdrátt sem það skrifar og meta gæði ritgerðaskrifa (Foltz, 1996). Fræðilega væri hægt að nota LSA til þess að spá fyrir um betri heilsu fólks með því að greina hversu lík ritgerðaskrif um áfall eru, yfir

þrjá til fjóra daga (Pennebaker og Graybeal, 2001). LSA greining á þremur slíkum verkefnum náði ekki að sýna fram á tengsl milli innihalds texta og heilsu (Pennebaker, Mehl o.fl., 2007). Til þess að geta nýtt LSA til þess að nálgast upplýsingar um einstaklingsmun þyrfti að vera nægilega öruggur grunnur fyrir því að innihald texta gæti gefið í skyn þennan mun. Samkvæmt Pennebaker og fleirum (2002) hefur ekki verið hægt að sýna fram á þann grunn og að stíll textans virki betur til þess að greina einstaklingsmun. Þetta er mikill ókostur fyrir LSA, sem skoðar ekki smáorð og byggist aðeins á orðum sem gefa í skyn viðfangsefni. Þegar hugsað er út í upplýsingarnar sem fást með stíl annars vegar og viðfangsefni hins vegar þá er þetta stór annmarki á þessari aðferð. Það þarf þess vegna að gæta þess að aðferðin sé nýtt í rannsóknum þar sem viðeigandi er að sleppa smáorðum í greiningu.

TAS/C er textagreiningarforrit sem skoðar aðeins tvær tungumálavíddir, tilfinningavægi og óhlutstæð hugtök. Forritið var þróað af Mergenthaler árið 1996 og er mjög kenningabundið og takmarkað við mjög þröngt svið tungumálagreiningar (Pennebaker, Mehl o.fl., 2007).

Þessi forrit og fleiri hafa verið þróuð til þess að mæta afmörkuðum verkefnum (sjá góða yfirlitsgrein yfir textagreiningatól í Mehl, 2005). Ekkert þeirra var sérstaklega þróað með persónuleikamat í huga. Það þýðir að ef það á að nota þau með mat á persónuleika í huga, þarf að aðlaga greininguna og orðabækurnar. Fæstir hafa lagt út í þá vinnu og nýta þeir sér sína nálgun til þess að meta persónuleika með sem bestu móti. Þetta býr til tækifæri fyrir forrit í þróun. Það er hægt að taka rannsóknir á persónuleika og einstaklingsmun og nýta þá þekkingu sem þar hefur myndast við þá þróun. Þessi þekking ætti að gefa til kynna hvaða geta þarf að vera við greiningu og birtingu gagna. Sömuleiðis er mikilvægt fyrir slík forrit að taka mark á þeirri gagnrýni sem orðasafnsnálgunin hefur hlotið yfir árin. Það þarf að vera skýrt hvernig textagreiningartól nær að bægja þeirri gagnrýni frá sér. Til þess þarf að vera góður skilningur á þessari gagnrýni og er hún umfjöllunarefni næsta undirkafla.

1.1.3 Gagnrýni á orðasafnsnálgunina

Mikið af gagnrýninni sem fallið hefur á orðasafnsnálgunina á við þáttahluta hennar. Oftar en ekki á sú gagnrýni einnig við hinn hluta orðasafnsnálgunarinnar;

tungumálagreininguna. Það er þess vegna mikilvægt að skilja hver þessi gagnrýni er þannig hægt sé að meta hvort textagreiningatól falli undir hana. Þegar svo er þá þarf að svara því hvort sú gagnrýni sé óréttlát, röng eða ef hún reynist rétt, hvað þurfi að gera til að bæta upp fyrir það.

Gróflega er hægt að flokka gagnrýnina í 4 flokka; gagnrýni á notkun lýsingarorða sem persónuleikabreytur, gagnrýni á notkun leikmanna sem athugendur persónuleika, gagnrýni á útskýringargetu persónuleikavídda og gagnrýni byggð á samanburði við aðrar vísindagreinar (Asthon og Lee, 2005). Hér er farið í helstu gagnrýni sem fallið hefur á þáttanálgun og mótrök tekin fram.

Westen (1996) og fleiri hafa gagnrýnt það að lýsingarorðin sem leikmenn nota til að lýsa persónuleika séu of einföld til að útskýra flóknari þætti sem vekja áhuga sálfræðinga (Í Ashton og Lee, 2005; Block, 1995). Þeir nefndu að það að raða saman lýsingarorðum undir persónuleikaþátt væri ekki nægilega skýrt. Með þessu er átt við að lýsingarorðin sem fólk notar í daglegu tali til þess að vísa til ákveðins persónuleikaþáttar eða eiginleika, séu í raun of einföld og ná ekki yfir margbreytileika persónuleika. Það hefur í för með sér að ekki er rétt að nota þessi sömu lýsingarorð til þess að búa til þá þætti sem stuðningsmenn þáttanálgunarinnar hafa gert. Í svari sínu nefndu Ashton og Lee (2005) að það væri frekar auðvelt að búa til skýra mynd af persónuleikaþætti með einföldum röðum lýsingarorða. Sömuleiðis hafa þessi orð komið til vegna þess að þau hafa langa sögu af því að vera notuð og valist úr hópi margra annarra sem þóttu ekki lýsa sömu einkennum eins vel. Ef það á að yfirfæra þessa gagnrýni á tungumálagreiningu, má alveg eins segja að orðaval einstaklinga sé ekki góð mælistika á einstaklingsmun. Þessi gagnrýni myndi sem sagt snúa að innihaldi orðanna. Talning á lýsingarorðum, nafnorðum og öðrum mikilvægum orðum sem gefa upp merkingu og innihald texta væri ekki réttmæt mæling á persónuleika. Þetta undirstrikar mikilvægi þess að skoða stíl samhliða viðfangsefni. Ef ekki er hægt að meta persónuleika eftir orðavali, vegna þess að ákveðin orð endurspegli ekki mikilvæga flókna þætti sem sálfræðingar hafa áhuga á, þá er kannski betra að líta til uppbyggingu (stíl) textans.

Block (1995) gagnrýndi sömuleiðis notkun lýsingarorða sem væru tvíræð í merkingu. Hann nefndi að þessi tvíræðni dregur úr notagildi lýsingarorðanna sem lýsing á sérstökum hugtökum. Ashton og Lee (2005) svöruðu því að þrátt fyrir að þetta sé rétt, þá hefur þetta ekki sérstök áhrif á notagildi þáttanálgunar. Ef

lýsingarorð eru notuð til þess að lýsa tveimur eða fleiri persónuleikavíddum, munu niðurstöður þáttagreiningar sýna flóknari hleðslumynstur fyrir þau lýsingarorð. Fyrir tungumálagreiningu, sérstaklega þá sem byggir eingöngu á sjálfvirkum leiðum, er þetta einstaklega mikilvæg gagnrýni. Tvíræð merking orða getur skipt sköpum þegar beita á orðatalningu. Til þess að komast hjá þessu þarf textagreiningatól að vera fært um greina samhengi og út frá samhenginu áætla merkingu orðsins. Ef litið eingöngu á stíl textans þá verður þessi gagnrýni ekki eins viðeigandi.

Westen (1996) gagnrýndi að leikmenn skuli vera þeir sem meta persónuleikann. Í flestum tilvikum eru það leikmennirnir sjálfir eða einhver sem þekkir þá (sem sjálfur er leikmaður) sem metur persónuleikaeinkenni einstaklingsins. Samkvæmt Westen ætti þetta að vera í höndum sérfræðinga sem þekkja orsök breytileika í persónuleika (Í Ashton og Lee, 2005). Þessi gagnrýni tekur á áhugaverðum punkti. Fyrir tungumálagreiningu hefur þetta þá þýðingu að ef einstaklingar eru ekki nógu færir um að meta sinn eigin persónuleika, þá gætu þær aðferðir sem tungumálagreiningartólin nota til að sýna fram á réttmæti þeirra skilað röngum niðurstöðum. Í sumum tilvikum gæti þetta útskýrt lága fylgni milli mælinga sem fundnar eru með greiningu á tungumáli og greiningu með þáttaprófum. Ashton og Lee (2005) benda á að mikið af rannsóknum hafa sýnt að sjálfsmat og mat annarra á persónuleikabreytum eru þokkalega nákvæmar. Samkvæmt Kolar, Funder og Colvin (1996) er mat annarra á persónuleika einstaklinga nákvæmari heldur en sjálfsmat einstaklingsins (Í Ashton og Lee, 2005). Það bendir til þess að gott væri að afla gagna um mat annarra á persónuleika fólks sem viðmið fyrir tungumálagreiningu. Þetta á betur við mælingar á persónuleikaþáttum sem eru sýnilegar öðrum, svo sem úhverfa eða taugaveiklun. Mat annarra á þessum persónuleikaþáttum hefur ekki verið notað sem viðmið fyrir tungumálagreiningu, sem best sem höfundur veit, og væri áhugavert að skoða.

Bandura (1999) og Shoda og Mischel (2000) gagnrýndu það að þáttanálgunin nær ekki yfir kerfisbundinn mismun einstaklinga sem kemur fram við mismunandi aðstæður (Í Ashton og Lee, 2005). Það er að segja að hegðun einstaklinga er breytileg eftir aðstæðum og einföld notkun lýsingaorða til þess að komast að persónuleika nær ekki yfir svo breytilegar aðstæður sem manneskjan getur lent í. Þessi gagnrýni Bandura annars vegar og Shoda og Mischel hins vegar

er mikilvæg fyrir tungumálagreiningu. Ef tungumálanotkun einstaklinga er breytileg eftir aðstæðum gerir það einstaklega erfitt að beita tungumálagreiningu til þess að komast að mun milli einstaklinga. Pennebaker og King (1999) vildu athuga hvort að orðaval einstaklinga væri stöðugt yfir tíma og milli aðstæðna. Niðurstöður þeirra bentu til þess að svo væri og að margar ólíkar hliðar tungumálanotkunar væru áreiðanlegar. Allt frá notkun smáorða til notkunar á viðhorfsorðum (jákvæðum og neikvæðum) og tíðir sagna. Þessar niðurstöður hafa verið endurteknar (Pennebaker, Chung o.fl., 2007).

1.2 Frávarpsnálgun

Freud var einna fyrstur til að skoða duldar hvatir og þýðingu þeirra í uppruna sálarlífs (Sargent, 1945). Hann var einnig með þeim fyrstu til að taka eftir því að það sem við segjum getur gefið upp duldar hvatir með því sem hann kallaði mismæli (*slip of the tongue*,Tausczik og Pennebaker, 2009). Frávarpsnálgunin byggist á hugmyndum Freuds um að við höfum öll duldar hvatir, óskir og togstreitu. Þessi ferli eru ómeðvituð og við þurfum að beita ákveðnum aðferðum til þess að komast að því hver þau eru. Þessar aðferðir eru kallaðar frávarpsaðferðir eða frávarpspróf þar sem próftaki tekst á við verkefni sem eru þess eðlis að vera opin og á ákveðinn hátt óformbundin, eins og að segja sögu út frá mynd eða túlka blekklessu. Svör próftakans eru skráð niður af rannsakanda sem metur svarið eftir fyrirfram ákveðnum stöðlum og eiga þau að gefa innsýn í *innri* persónuleika próftakans sem ekki er hægt að nálgast með hefðbundnum leiðum. Það má hugsa um þessi próf sem skjá þar sem próftakinn *varpar* sínum innri tilfinningum, óskum, átökum og hugmyndum á (Gleitman o.fl., 2004).

1.2.1 Forsaga frávarpsaðferða

Frávarpsaðferðir eiga rætur að rekja meðal annars aftur til *Gestalt* kenningarinnar, sem hófst með tilraunum Wertheimers á skynjun í kringum 1910, til rannsókna Brittain (1907) á ímyndun og notkun Bartletts (1916) á blekklessum (Í Sargent, 1945). Bartlett nýtti blekklessurnar til þess að skoða einstaklingsmun á greind, bakgrunni, starfsáhuga og fleira. Hin ýmsu form af frávarpsprófum höfðu verið í notkun löngu áður en þau voru notuð í persónuleikarannsóknum. Sem dæmi rannsakaði Galton (1883) ímyndun og Binet og Simon (1905) notuðu blekklessur í fyrri prófum sínum (Í Sargent, 1945).

Á fimmta og sjötta áratug 20. aldar var svo mikil aukning í vinsældum frávarpsprófa. Rorschach blekklessuprófið og TAT-prófið voru hvað vinsælust. Manneskjuteikniprófið (*Draw-a-person test*) og setningalokaprófið (*Sentence Completion Test*) nutu sömuleiðis mikilla vinsælda (Heiða María Sigurðardóttir, 2005).

Rorschach prófið var þróað af Hermann Rorschach árið 1921, sem var svissneskur geðlæknir og notaði skynjun á óskipulögðun (*unstructured*) formum sem greiningartól (Gleitman, Friedlund og Reisberg, 2004). Eins og komið hefur fram var hann ekki sá fyrsti til að nota blekklessur, en hann setti fyrstur manna fram aðferð til þess að meðhöndla flókin svarmynstur einstaklinga (Sargent, 1945). Í Rorschach prófinu er próftakanum sýnd tíu spjöld sem á eru samhverfar blekklessur, sumar í lit og aðrar svart-hvítar. Rorschach lagði ekki áherslu á innihald svara heldur á eiginleika áreitisins sem birtust á spjaldinu og hvernig próftakinn nýtti sér þá (Gleitman o.fl., 2004).

Eftir dauða Rorschach fylgdi mikill fjöldi af rannsóknum á prófinu. Eitt af þekktari ritum sem komu út eftir lát Rorschachs um prófið var bók Becks (1937) sem hét *Introduction to the Rorschach Method* (Í Sargent, 1945). Þetta var fyrsti kerfisbundni leiðarvísirinn á því hvernig ætti að leggja prófið fyrir, gefa einkunnir og túlka. Það var á svipuðum tíma sem að Morgan og Murray (1935) kynntu aðferð til rannsókna á ímyndun en þar lýstu þeir grunntækni TAT-prófsins í fyrsta sinn (Í Sargent, 1945). Prófið samanstendur af myndum sem einstaklingar eru beðnir um að nota sem útskýringar á söguþræði sem þeir búa sjálfir til. TAT-prófið leggur aðaláherslu á innihaldið og miðar að því að komast að helstu hvötum og athöfnum einstaklinga, varnarháttum, átökum og túlkunum á veröldinni (Gleitman o.fl., 2004). Þessi tvö próf, Rorschach og TAT, áttu stærstan þátt í mótun frávarpsaðferða eins og þær þekkjast í dag og hafa mest verið rannsökuð.

Tilgangur frávarpsaðferða er þríþættur. Þær voru upphaflega þróaðar til notkunar í klínískum aðstæðum, sem tól til þess að skoða ímyndunarafl, taka saman meðvitaða og ómeðvitaða hugsun og til skilnings á því hvernig einstaklingar skipuleggja skynjun. Frávarpsaðferðirnar hafa einnig verið notaðar sem meðferðarúrræði, til þess að létta á ómeðvitaðari togstreitu eða komast að duldum hvötum og óskum. Síðast en ekki síst hafa frávarpsaðferðir verið nýttar í tilraunaskyni (Sargent, 1945).

Notkun frávarpsaðferða hefur mikið verið rannsökuð og hafa þær oftar en ekki beinst að réttmæti prófanna, en samkvæmt Holt (1978), Kleinmuntz (1982), Kline (1995) og Rorer (1990) þá hafa þessar aðferðir takmarkað eða lítið sem ekkert réttmæti (Gleitman o.fl., 2004).

1.2.2 TAT

TAT prófið var þróað af Christina Morgan og Henry Murray árið 1935 og miðar að því að kanna hvatir, ásetning og væntingar próftaka. Prófið samanstendur af 31 svart-hvítum myndum og innihalda flestar fólk en eitt þeirra er autt (Hood og Johnson, 1997). Samkvæmt Hood og Johnson (1997) eru 20 spjöld sýnd hverju sinni og próftakar eru beðnir um að segja sögu um hverja mynd í eins miklum smáatriðum og þeir geta. Prófdómarar geta túlkað niðurstöðurnar á ýmsan hátt. Ein leið við túlkun sem lofar góðu var hönnuð af Drew Westen við Boston háskóla. Hún nýtir sérstakt einkunnarkerfi til þess að meta skynjun fólks á öðrum (*object relations*). Þrátt fyrir að kerfisbundnar aðferðir eru betri þá hafa rannsóknir sýnt að fæstir nýta slíkar aðferðir við túlkun (Lilienfeld, Wood og Garb, 2001).

TAT-prófið hefur orðið fyrir mikilli gagnrýni hvað varðar áreiðanleika og réttmæti. Helsta gagnrýnin á áreiðanleika snýr að endurprófunaráreiðanleika (*test-retest reliability*) sem felur í sér að einkunnagjöf eða túlkun ætti að vera svipuð eða sú sama ef prófið er lagt fyrir tvisvar á mismunandi tímum. Réttmæti TAT er mjög vafasamt, þar sem sum einkunnakerfi hafa til dæmis ekki náð að greina á milli heilbrigðra og sjúklinga sem þjást af þunglyndi. Kerfisbundin einkunnakerfi, líkt og það sem Drew Westen þróaði, standa sig ágætlega við greiningu á ákveðnum þáttum persónuleika, svo sem þörfin fyrir árangur og skynjun á öðrum. Hins vegar hefur þeim mælingum ekki tekist að spá vel fyrir um raunhegðun einstaklinga (Lilienfeld o.fl., 2001). Það er verkefni út af fyrir sig að fjalla um gagnrýni á frávarpsaðferðir og hafa nokkur slík yfirlit verið birt (sjá til dæmis, Hood og Johnson, 1997; Lilienfeld, Wood og Garb, 2000; Lilienfeld o.fl., 2001; Wood, Nezworski og Stejskal, 1996).

Til þess að taka saman má segja að gagnrýnin á TAT hafi snúist um a) skort á kerfisbundunum reglum við fyrirlögn og einkunnagjöf, b) lélegum áreiðanleika, bæði samkvæmni matsmanna og við endurprófun og c) lítið réttmæti. Við þetta vakna spurningar hvort hægt sé að koma til móts við þessa gagnrýni með því að

beita nýjum nálgunum við greiningu frávarpsprófa, þá sérstaklega TAT. Með því að nýta textagreiningatól til þess að skoða sögurnar sem koma út úr TAT-prófinu er hægt að stórbæta endurprófunar áreiðanleika og koma á kerfisbundnum reglum. Tölvuforrit hafa það fram yfir manneskjur að þau vinna eftir sömu reglunum aftur og aftur. The General Inquirer var þróað einmitt í þessum tilgangi, að bæta greiningu á niðurstöðum TAT. Önnur forrit hafa einnig verið nýtt til þess að athuga gagnsemi þess að beita textagreiningu á niðurstöður TAT. Það var einmitt það sem Pennebaker og King (1999) nýttu LIWC í með því að greina sögur úr TAT-prófum. Þau komust að því að textagreining spáði betur fyrir um TAT-tengda hegðun en TAT aðferðafræðin gerði ráð fyrir. Þetta vekur upp þá hugmynd að hægt sé að auka réttmæti prófsins með því að beita textagreiningu við greiningu á svörum próftaka.

Hugmyndin á bakvið frávarpspróf er ekki ósvipuð þeirri sem liggur að baki orðasafnsnálguninni. Í báðum tilvikum er gert ráð fyrir því að hegðun einstaklingsins – orðaval, uppbygging setninga eða hvernig túlkun myndaog blekklessa – gefi innsýn í innri sálræna þætti eins og persónuleika. Með öðrum orðum er tungumálið grundvallarforsenda í báðum tilvikum. Það er þess vegna ekki fjarstæðukennd hugmynd að þessar nálganir gætu virkað saman undir einni stefnu, sem miðar að því að samþætta kosti þeirra og vega upp galla.

1.3 Ein sameiginleg nálgun

Það er augljóst að þessar tvær nálganir eða aðferðir sem hafa fengið umfjöllun hér byggjast á svipuðum hugmyndum. Sálfræðingar eru stöðugt að leita leiða til þess að komast að leynda eða falda *raun* sjálfinu. Freud byggði sínar kenningar á þessum hugmyndum og frávarpsprófin miða að því að svara spurningum um hið falda raun sjálf. Greining á uppbyggingu máls með því að horfa á smáorð er í raun svipað ferli. Þetta er önnur nálgun að því að komast að sálrænum ferlum sem eru annars falin og koma fram í orðavali og setningarmyndun einstaklinga (Tausczik og Pennebaker, 2009).

Ein sameiginleg nálgun myndi beita textagreiningu til þess að komast að þessu falda sjálfi. Þrátt fyrir ítrekaða gagnrýni á frávarpsaðferðir eru þær ennþá vinsælar. Árið 1988 töldu 49% af yfirmönnum framhaldsnáms í klíníska sálfræði og 65% af yfirmönnum lærlingsnáms í klíníska sálfræði að þekking á frávarpsaðferðum væri mikilvæg (Durand, Blanchard og Mindell, 1988). Árið 1995

voru Rorschach og TAT-prófið á meðal 10 vinsælustu matstækja klínískra sálfræðinga. Sem dæmi sögðust 82% nota þau stundum og 43% nota þau oft eða alltaf við mat á sjúklingum (Lilienfeld og fleiri, 2001; Watkins, Campbell, Nieberding og Hallmark, 1995). Árið 2000 var Rorschach fjórða mest notaða prófið af klínískum sálfræðingum og TAT sjötta mest notaða prófið (Camara, Nathan og Puente, 2000). Þetta bendir til þess að vinsældir frávarpsprófa séu ekki að dala þrátt fyrir gagnrýni. Það virðist vera eitthvað við frávarpsaðferðirnar sem heillar sálfræðinga. Þær gefa einstaklega ríka mynd af hugarheimi fólks. Þetta þýðir að leita þurfi annarra leiða við að koma til móts við gagnrýni og vinsældir. Textagreining hefur sýnt að geta spáði betur fyrir um TAT-tengda hegðun heldur en tilgátur TAT matsins gerðu ráð fyrir að TAT-prófið ætti að gera (Pennebaker og King, 1999). Það vekur upp spurninguna um ef markvisst er farið að vinna með textagreiningatól við greiningu svara er þá hægt að auka áreiðanleika og sýna fram á betra réttmæti?

2 Niðurstöður textagreininga

2.1 Próffræðilegir eiginleikar

Til þess að hægt sé að treysta niðurstöðum greiningar á tungumáli þarf fyrst að mæla próffræðilega eiginleika orða. Mynstur í orðanotkun einstaklinga þarf að fullnægja próffræðilegum kröfum um stöðugleika fyrir tíma og milli aðstæðna (Pennebaker, Mehl o.fl., 2007). Með öðrum orðum þarf að sýna fram á áreiðanleika mælinga. Þar að auki þarf að sýna fram á að textagreiningatól sýni nægilegt réttmæti til þess að engin vafi sé á að þau mæli það sem þau eiga að mæla, persónuleika eða annan einstaklingsmun.

Ólíkt þróun hefðbundinna mælitækja þá er frekar flókið ferli að meta áreiðanleika og réttmæti orðanotkunar. Í grófum dráttum er hefðbundna aðferðin sú að sálfræðingar þróa spurningalista og meta áreiðanleika spurninganna með því að reikna út viðeigandi áreiðanleikastuðul, prófa svo endurprófunaráreiðanleika (*test-retest reliability*) prófsins með að leggja það fyrir og að lokum finna réttmæti þess með að sjá hversu vel það spáir fyrir eða tengist annarri hegðun (Tausczik og Pennebaker, 2009). Orðflokkar eru ólíkir spurningalistum. Það er ekki hægt að gera sömu próffræðilegu kröfur til orðflokka, þar sem orðtíðni er sjaldan normaldreifð og flestir flokkar mynda hlutfallslega lítið magn af heildartextanum. Þar af leiðandi eru

hefðbundnu áreiðanleika prófin ekki alltaf viðeigandi (Tausczik og Pennebaker, 2009). Sem dæmi er ekki hægt að biðja fólk um að lýsa sér tvisvar og ætlast til þess að lýsingarnar séu ekki mismunandi. Þetta setur aukna pressu á rannsakendur, sem vilja beita textagreiningu til að sýna fram á réttmæti og áreiðanleika mælinga og rökstyðja þá aðferðafræði sem þeir beita. Þessar áhyggjur hafa ekki stöðvað fylgjendur tungumálagreiningar í að athuga próffræðilega eiginleika orðanotkunar.

Walter Mischel (1968) benti á mikilvægi þess að sýna fram á stöðugleika persónuleika yfir tíma og milli aðstæðna til þess að geta rannsakað hlutverk persónuleika í forspá hegðunar (Í Pennebaker og King, 1999). Hann tók eftir að hegðun einstaklinga er mjög breytileg eftir aðstæðum og persónuleikaþættir ná ekki að spá fyrir um hegðun með stöðugum hætti (Gleitman o.fl., 2004). Gleser, Gottschalk og Watkins (1959) létu fólk tala í nokkrar fimm mínútna lotur um áhugaverða lífsreynslu og sýndu fram á stöðugleika 21 tungumálaflokks með því að reikna helmingunaráreiðanleika (*split-half reliability*) milli tveggja lota. Schnurr, Rosenberg, Oxman og Tucker (1986) notuðu The General Inquirer nálgunina til þessa að styðja frekar stöðugleika tungumálanotkunar yfir tíma (Í Pennebaker, Mehl o.fl., 2007). Pennebaker og King (1999) sýndu fram á að orðanotkun í skrifuðu máli er áreiðanleg yfir tíma, viðfangsefni og tegund texta. Tekið saman benda rannsóknir á próffræðilegum eiginleikum þess að nota greiningu á tungumáli sem mælingu á einstaklingsmun til þess að hún sé viðeigandi. Þetta á bæði við greiningu á orðaflokkum og sálfræðilega byggðum tungumálavíddum (Pennebaker, Mehl o.fl., 2007).

2.2 Einstaklingsmunur

Það er munur á því hvernig einstaklingar tala og skrifa. Þessi munur getur verið kerfisbundinn og á við dýpri eða innri þætti eins og persónuleikaþætti. Í öðrum tilvikum felst munurinn í breytum eins og aldri eða kyni. Ein besta leiðin til þess að athuga kerfisbundinn mun í orðanotkun einstaklinga hefur reynst verið að skoða smáorðin sem þeir nota (Argamon o.fl., 2005; Chung og Pennebaker, 2007; Cohn o.fl., 2004; Mehl og Pennebaker 2003; Tausczik og Pennebaker, 2009).

2.2.1 Persónuleiki

Persónuleiki og þá sérstaklega persónuleikaþættirnir, úthverfa (*extraversion*) og taugaveiklun (*neuroticism*), hafa á undanförnum árum fengið mikla athygli

rannsakenda sem beita textagreiningu. Úthverfa er persónuleikaþáttur sem tengist mikið samskiptum milli persóna og félagslyndi, á meðan taugaveiklun eða tilfinningastöðugleiki (*emotional stability*) er tengdur innra ástandi tilfinningalífs. Rannsóknir hafa sýnt að þessir persónuleikaþættir geta haft þýðingarmikil áhrif á hvernig fólk notar tungumál við hinar ýmsu aðstæður (Dewaele og Furnham, 1999; Pennebaker og King, 1999).

Pennebaker og King (1999) gerðu eina af fyrstu rannsóknunum sem miðuðu að því að tengja orðanotkun við persónuleikaþætti. Þar nýttu þau sér LIWC til þess að gera orðatalningu á texta og fundu hófleg en áreiðanleg áhrif persónuleikaþátta á orðanotkun, en fylgnistuðlarnir voru á bilinu 0,10 og 0,16. Helstu niðurstöðurnar voru þær að taugaveiklun hafði jákvæða fylgni við notkun neikvæðra viðhorfsorða og neikvæða fylgni við notkun jákvæðra viðhorfsorða. Taugaveiklun var líka tengd við tíðari notkun fyrstu persónufornafna í eintölu. Úthverfa og samvinnuþýði (*aggreeableness*) höfðu jákvæða fylgni við jákvæð viðhorfsorð. Enn frekar var jákvæð fylgni milli úthverfu og notkunar orða sem tjá félagslega ferla (sem notuð eru við félagslegar aðstæður) og neikvæð fylgni milli samvinnuþýði og neikvæðra viðhorfsorða. Það má gagnrýna lága fylgni í þessari rannsókn. Í öllu falli væri ákjósanlegt að sýna fram á hærri fylgni ef vilji er fyrir því að fara að flokka fólk byggt á texta sem það skrifar, sem hlýtur að vera endanlega takmarkið. Á móti kemur að þetta var ein fyrsta rannsóknin og hafa nýlegri rannsóknir gefið haldbærari upplýsingar um tengsl orðanotkunar og persónuleikaþátta.

Furnham (1990) setti fram einstaklega áhugaverða sýn á eiginleika úthverfs og innhverfs (*introvert*) tungumáls. Úthverfir einstaklingar nota óformlegra og takmarkaðra tungumál. Einnig eiga þeir það til að vera með lauslegri orðaforða í þeim skilningi hversu rétt orð eru notuð og hversu óvenjuleg þau eru. Enn frekar þá nota þeir meira af sögnum, atviksorðum og fornöfnum. Dewaele og Furnham (2000) kölluðu þetta óbeint tungumál (*implicit language*), sem er ólíkt beinu tungumáli (*explicit language*) þar sem einstaklingar nota meira af nafnorðum, lýsingarorðum og forsetningum (Oberlander og Gill, 2004). Oberlander og Gill (2004) vildu athuga hvort þetta ætti við þegar greina átti texta sem skrifaður er í tölvupóstum. Þeir greindu tölvupóstsamskipti einstaklinga, þar sem persónuleiki þeirra var þekktur, og gátu ekki staðfest niðurstöður Dewaele og Furnham (2000). Oberlander og Gill fundu að úthverfir einstaklingar notuðu meira af samtengingum

og innhverfir meira af sögnum í lýsingarhætti þátíðar. Þeir vildu einnig skoða hvort hugmyndir Dewaele og Furnham ættu frekar við taugaveiklunarvíddina. Sú tilgáta þeirra var studd að hluta, þar sem þeir sem einkenndust af mikilli taugaveiklun notuðu meira af fornöfnum (og samtengingum) en þeir sem einkenndust af lítilli taugaveiklun notuðu meira af nafnorðum og lýsingarorðum.

Argamon og fleiri (2005) nýttu ritgerðir sem skrifaðar voru af nemendum við Texas háskólann í Austin á árunum 1997-2003 til greiningar. Nemendur á þessu tímabili höfðu verið beðnir um að skrifa tvö verkefni, eitt þeirra um hugleiðingar líðandi stundar (HLS) og eina greiningu á sjálfum sér (SG). Niðurstöðurnar voru á þá leið að orð sem lýstu viðhorfi (neikvæðu og jákvæðu) gátu spáð fyrir um hvort einstaklingur væri taugaveiklaður (neuroticism) í 58,2% (HLS) og 58% (SG) tilvika. Það verður að teljast nokkuð góður árangur þar sem greiningin byggði aðeins á einu verkefni frá hverjum einstaklingi og oftast þarf ítarlega spurningalista eða mikið magn af texta til að ná að meta persónuleikaeinkenni á réttmætan hátt. Þessar niðurstöður gefa vísbendingu um að munurinn á þeim sem skora hátt og lágt á taugaveiklun sé notkun þeirra á viðhorfsorðum til þess að lýsa hlutum og persónum (sjálfum sér og öðrum) í umhverfi sínu. Þessi rannsókn styrkir enn betur hugmyndina að hægt sé að nota orðflokka til þess að greina einstaklingsmun. Orðflokkar voru notaðir í mismunandi tilgangi og það er þessi tilgangur sem gefur vísbendingu um einstaklinginn.

2.2.2 Mikilvægi smáorða

Smáorð er sá flokkur orða sem undir falla fornöfn, forsetningar, greinar, samtengingar og hjálparsagnir. Ekki er vitað með vissu hver fjöldi orða sem hinn almenni enskumælandi maður hefur í orðaforða sínum. Goulden, Nation og Read (1990) telja að það séu um 20.000 orð, en samkvæmt Baayen, Piepenbrock og Gulikers (1995) þá eru það færri en 400 smáorð (Í Chung og Pennebaker, 2007). Þessi örsmái hluti af orðaforðanum mynda samkvæmt Rochon, Saffran, Berndt og Schwartz (2000) yfir helming þeirra orða sem notuð eru í daglegu tali (Í Chung og Pennebaker, 2007). Þrátt fyrir að taka ekki mikið eftir smáorðum þá hafa þau mikil áhrif á hlustandann og þau geta gefið góðar vísbendingar um tungumálastíl einstaklinga.

Smáorð gefa ekki bara vísbendingar um tungumálastíl heldur geta þau gefið innsýn í sálrænt ástand fólks (Argamon o.fl., 2005). Sem dæmi sýndu Pennebaker og Lay (2002) fram á að þegar einstaklingar eru þunglyndir eða í tilfinningalega viðkvæmu ástandi nota þeir meira af fornöfnum (sérstaklega í fyrstu persónu eintölu), nota færri greina og auka notkun á hjálparsögnum í nútíð. Sömuleiðir hafa rannsóknir Cohn og fleiri (2004) á orðanotkun bandarískra bloggara í kringum hryðjuverkaárasirnar í New York leitt í ljós að einstaklingar nota meira af fyrstu persónu fleirtölu fornöfnum þegar tekist er á við sameiginlegt vandamál eða umrót. Það er að segja að fólk hætti að tala um sjálft sig og fór að tala um fólkið í kringum sig; vinina, fjölskylduna og aðra í hópnum þeirra. Þessar rannsóknir og fleiri (til dæmis, Chung og Pennebaker, 2007; Mehl og Pennebaker 2003) hafa fundið tengsl milli smáorða og hina ýmsu persónuleikaþátta eins og taugaveiklunar, úthverfu, víðsýni (openness to experience), sjálfsálits og félagslegrar stöðu.

Smáorð hafa verið skoðuð við ýmsar aðstæður. Í rannsóknum á heiðarleika og svikum sýndu Newman, Pennebaker, Berry og Richards (2003) fram á að fólk sem segir satt á það til að eigna sér söguna með því að nota meira af fyrstu persónufornöfnum. Newman og fleiri sýndu einnig fram á að konur eiga það til með að nota fyrstu persónufornöfn oftar en karlar. Það þýðir þó ekki að konur eru séu heiðarlegri heldur en karlar. Líkleg útskýring á þessu er sú að konur eru sjálfsmiðaðri en karlar og líklegri til að þjást af þunglyndi (Chung og Pennebaker, 2007). Aldur hefur einnig verið tengdur við smáorð. Með auknum aldri minnkar notkun á fyrstu persónu eintölu orðum og fleirtölu orð aukast. Sömuleiðis hefur greining á hjálparsögnum gefið í skyn að fólk notar meira af framtíð og minna af fortíð (Pennebaker og Stone, 2003).

Sýnt hefur verið fram á tengsl smáorða við athygli. Smáorðin gefa góða innsýn í það hvernig einstaklingar beita athygli og að hverju athyglin beinist. Sem dæmi þá er munur á notkun fornafna þegar einstaklingar lýsa því hvort þeim hafi verið strítt eða þeir hafi strítt öðrum. Einnig nota jákvæðar auglýsingar meira af sjálfsvísunum ("ég" og "við") á meðan neikvæðar auglýsingar notuðu meira af vísunum í aðra ("hann", "hún" og "þau",Tausczik og Pennebaker, 2009). Ef athygli beinist að ákveðnu sálrænu ástandi, t.d. sorg, þá er einstaklingur líklegur til að nota orð tengd sorg. Skilningur á því hvernig sá einstaklingur notar smáorð til að lýsa

sorginni gefur vísbendingu um hvort hann sé sá sem þjáist af sorg (núna eða í fortíðinni) eða hvort hann er að lýsa sorg annarra.

2.2.3 Lýðfræðilegar breytur

Pennebaker og Stone (2002) skoðuðu tengsl tungumálanotkunar og aldurs. Þeir komust að því að fólk sýnir stöðugar breytingar á tungumálastíl með aldri. Það sem þeir tengdu við hækkandi aldur var hærri tíðni jákvæðra viðhorfsorða, færri neikvæð viðhorfsorð, færri fyrstu persónufornöfn og vísanir í sjálfan sig, tíðari notkun sagna í framtíð og færri í þátíð (Í Pennebaker, Mehl o.fl., 2007). Þessar niðurstöður benda til þess að með auknum aldri verður fólk jákvæðara í skrifum og kýs að tala um aðra eða sjálfa sig sem hluta af hópi. Þetta bendir einnig til þess að með aldri horfir fólk frekar fram á við en er lítið að spá í fortíðinni.

Mehl og Pennebaker (2002) skoðuðu daglegt mál 52 háskólanemenda með því að taka upp samræður þeirra með EAR (*EAR technology*). Þeir vildu með þessari rannsókn athuga tungumálanotkun kynjana. Þær niðurstöður sýndu að það voru nokkrir þættir í tungumálanotkun kvenna og karla sem voru mismunandi en engin skýr mynd hefur komið í ljós (Í Pennebaker, Mehl o.fl., 2007).

2.2.4 Hugræn heilsa

The General Inquireer hefur með greiningu á textabútum sjúklinga náð á áreiðanlegan og árangursríkan hátt að flokka þá eftir greiningu, svo sem geðklofa, þunglyndi, ofsóknarkennd eða líkömunarraskanir. Tölvuleg greining á texta spáði betur fyrir um greiningu heldur en geðlæknir sem las útprentun af máli sjúklinga (Pennebaker, Mehl o.fl., 2007).

Stirman og Pennebaker (2001) nýttu textagreiningu til þess að skoða sjálfsvígstilhneigingu. Þetta gerðu þeir með því að skoða verk eftir ljóðskáld, sem höfðu framið sjálfsvíg, og báru þau saman við verk eftir ljóðskáld sem gerðu það ekki. Niðurstöðurnar bentu til þess að aukin notkun fyrstu persónufornöfn í eintölu og minni notkun fyrstu persónufornöfn í fleirtölu væri einkennandi fyrir ljóðskáld sem tóku sitt eigið líf. Sjálfsvígs ljóðskáldin notuðu einnig færri vísanir í annað fólk og notuðu meira af orðum tengd dauða (Í Pennebaker, Mehl o.fl., 2007).

Rannsóknir á þunglyndi og oflæti hafa sýnt fram á svipuð tengsl við persónufornöfn, sérstaklega notkun á fyrstu persónufornöfn eintölu. Þetta bendir til þess að fólk með þessar raskanir séu upptekið af sjálfu sér, sem endurspeglast í

orðavali þeirra (Pennebaker, Mehl o.fl., 2007). Til þess að geta með áreiðanlegum hætti greint þessar raskanir út úr texta sem fólk skrifar þarf að setja fram skýr viðmið um í hvaða samhengi notkun persónufornafna þarf að vera til að einstaklingur teljist þunglyndur eða með oflæti.

Þær niðurstöður sem hafa fengið umfjöllun hér sýna hvaða körfur þarf að gera til þess hugbúnaðar sem ætlað er að nota til þess að feta ámóta leiðir. Það er augljóst að mikið af rannsóknum hafa sýnt fram á getu textagreiningar til þess að greina einstaklingsmun. Það er þess vegna mikilvægt að vinna að því að ná sömu niðurstöðum. Aðeins þá er raunhæft að fara að vinna að því að ýta þröskuldi þekkingar áfram. CLARA er þróað með það í huga. Ekki aðeins er markmiðið að þróa tækni sem hefur sérstöðu á markaði heldur einnig að varpa nýju ljósi á fræðilega þekkingu á tengslum einstaklingsmunar og tungumáls. Hér í kjölfarið verður ítarleg umræða um CLARA og sérstöðu hugbúnaðarins.

3 CLARA

CLARA (*Collective Large-scale Affect Research and Analysis*) er íslenskur textagreiningar hugbúnaður þróaður af sprotafyrirtækinu CLARA (hér eftir nefnt *fyrirtækið* til einföldunar). Hafist var handa við þróun á CLARA fyrir tveimur árum. Síðan þá hefur hugbúnaðurinn verið notaður til þess að greina íslenskt og enskt tungumál. CLARA hefur hingað til ekki verið notaður í fræðilegum tilgangi, heldur er þetta þjónusta sem viðskiptavinir fyrirtækisins kaupa reglulega. Í dag er þjónustan, *Vaktarinn*, keyrð á CLARA kerfinu og býður upp á greiningu á umræðu um vörumerki og fyrirtæki. CLARA hlaut styrk frá Tækniþróunarsjóð Rannís (Rannís, 2009) árið 2008 en þar að auki fékk fyrirtækið á dögunum fjárfestingu að upphæð einni milljón Bandaríkjadollara.

Þróun CLARA er ennþá í fullum gangi og er verið að vinna í að stórbæta greiningu á enskum texta. Markmiðið er að bjóða upp á lausn fyrir Bandaríkjamarkað, þar sem hugbúnaðurinn er nýttur af tölvuleikjafyrirtækjum til þess að skilja betur viðskiptavini sína. Þessi skilningur mun fela í sér dýpri þekkingu en bara að vita hvað þeir eru að skrifa. CLARA leitast við að tengja saman fræðilega þekkingu og markaðslegt gildi. Þar af leiðandi er hluti af tæknilegri þróun hugbúnaðarins að tengja saman textagreiningu og sérkenni einstaklinga. Tæknileg sérstaða CLARA er það sem aðstandendur fyrirtækisins

leitast við að nýta til þess að ná betri árangri en núverandi textagreiningalausnir. Hér verður gerð grein fyrir sérstöðu CLARA, hvaða virði sú nálgun skapar á klínísku, markaðslegu og fræðilegu sviði. Að því loknu verður fjallað um hvernig best er að prófa kerfið þannig að það í fyrsta lagi standist þær kröfur sem önnur tól hafa staðið undir og í öðru lagi auki þekkingu á sviði tengsla textagreiningar og persónuleika. Að lokum verður rætt um hver næstu skref eru hvað varðar þróun og prófanir.

3.1 Sérstaðan

Tæknileg sérstaða er eitthvað sem nauðsynlegt er að sýna fram á ef kynna á til sögunnar nýja lausn. Það er ekki nóg að herma eftir fyrri lausnum, heldur þarf að koma með eitthvað nýtt að borðinu, sem hefur ekki sést áður. Það er markmiðið með CLARA. Hér að ofan var nokkrum lausnum lýst og farið í þá kosti og þá galla sem þeim fylgja. Í þessari umræðu um sérstöðu CLARA verður greint frá eiginleikum sem eru sambærilegir og hvernig CLARA mætir þeim göllum sem fylgja öðrum lausnum.

3.1.1 Lýsing á tækni

Tækninni á bakvið CLARA er hægt að skipta í þrennt; gagnaöflun, greining gagna og birting niðurstaðna. Gagnaöflun á við um það hvernig gögnin eru fengin og meðhöndluð. Greining gagna er síðan kjarninn í kerfinu þar sem gögnum er breytt í upplýsingar. Þessar upplýsingar eru svo birtar notandanum. Það er á öllum þessum þremur sviðum þar sem sérstaða CLARA kemur fram.

3.1.1.1 Öflun gagna

CLARA getur aflað gagna með þremur leiðum. Í fyrsta lagi eru það svokallaðar vefköngulær (*web crawlers*), sem fara á tiltekin vefsvæði og sækja öll viðeigandi gögn sem þar er að finna. Vefköngulærnar sækja allan texta sem er skrifaður á vefsíðunum, dagsetningar og tímastimpla, upplýsingar um höfundinn þegar þær er að finna og getur áttað sig á grófri staðsetningu með því að lesa *IP-tölu* síðunnar. Í öðru lagi getur CLARA nýtt gagnasendingar frá vefmiðlum en þá senda vefsvæðin sem áhugi er fyrir að sækja allar sínar upplýsingar í gegnum ákveðið viðmót. Þannig er hægt að fá allar upplýsingar um textann sem er skrifaður, tíma- og dagsetningar og ítarlegri upplýsingar um höfundinn. Þriðja leiðin sem CLARA nýtir

er að lesa inn skjöl sem eru á textaskrárformi. Þannig er hægt að koma með safn af skrám og lesa allar upplýsingar sem í þeim eru að finna.

Sérstaða CLARA felst í fyrstu tveimur aðferðunum. Aðrar lausnir hafa stuðst eingöngu við lestur á textaskrám, sem eru mataðar inn í forritin. Með því að beita vefköngulóm og geta tekið við gagnasendingum er CLARA að opna fyrir möguleikann að ná í gífurlegt magn af texta og greina hann. Sem dæmi væri hægt að sækja allar færslur af stærstu bloggsvæðum, sem samtals telja yfir 126 milljónir blogga samkvæmt *BlogPulse* sem er þjónusta í boði rannsóknafyrirtækisins Nielsen[1]. Ekki aðeins eru það bloggin sem heilla heldur einnig samfélagsvefir á borð við Twitter og Facebook[2], sem hafa yfir hálfan milljarð notenda.

Ef litið er til möguleika þess að skoða allt netið og nýta upplýsingar þaðan til þess að keyra rannsóknir er það við fyrstu sýn mjög heillandi. Það er hins vegar löng leið framundan áður en hægt verður að beita textagreiningu til þess að greina einstaka höfunda byggt á textanum sem þeir skrifa. Þörf er á frekari rannsóknum, sem sýna að slík greining er áreiðanleg og réttmæt þegar gögnin koma frá mismunandi vefsvæðum. Einnig þarf að sýna fram á að hægt sé að greina einstaklingsmun án þess að nokkrar upplýsingar um höfundinn sé að finna og það gert með áreiðanlegum og réttmætum hætti. Svo að lokum þarf að sýna fram á að slík greining sé viðeigandi þegar textamagnið er ekki mikið meira en 50-100 orð. Í tilviki örsamskiptavefsins Twitter, þar sem aðeins er leyfilegt að nota 140 slög, þarf að beita sérstakri greiningu til þess að ná að greina einstaklingsmun.

Augljóslega er mikið sem þarf að rannsaka áður en hægt verður að beita CLARA á nýjum vettvangi og geta með þokkalega öruggum hætti áætlað að greiningin sé rétt. Þar af leiðandi hefur fyrirtækið ákveðið að byrja á mjög afmörkuðum stað. Fyrirtækið mun einbeita sér að nota CLARA til þess að greina texta á spjallsvæði tölvuleikjaframleiðanda. Þar er að finna mikið magn af texta,

[1] Sótt 6.5.2010 af http://www.blogpulse.com/

[2] Sótt 6.5.2010 af http://www.facebook.com/press/info.php?statistics

sem dæmi hafa rétt rúmlega hálf milljón spjallnotenda framleitt yfir 288 milljónir þráða á spjallsvæði tölvuleiksins Diablo[3]. Þarna er hægt að vinna með framleiðendum tölvuleikja sem myndu gefa CLARA aðgang að gögnum um spjallnotendurna. Það myndi fela í sér að hægt væri að rekja hvern þráð og hvert svar til ákveðins höfundar og hægt væri að nota upplýsingarnar um höfundinn til þess að framkvæma viðeigandi greiningu.

3.1.1.2 Greining gagna

Við greiningu á gögnum er markmiðið að sækja upplýsingar. Það er að segja að breyta textanum í upplýsingar. Fyrst þarf að skipta textanum upp í orð og tákn og því næst að greina tungumál textans. Það er gert til að geta svo greint textann niður í setningar og greinar og einnig til að geta greint orðin í orðmyndir svo sem kyn, tölu, stig, mynd, fall, tíð og hverja þá aðgreiningu sem orðin hafa til að mynda sérnöfn eða staðanöfn. Þegar búið er að forvinna textann svona er mun fljótlegra að greina mikið magn af honum og hann tekur minna rými í geymslu. Einn þáttur í þeirri greiningu er hefðbundin orðflokkatalning. Sú talning svipar til þeirra sem LIWC og fleiri forrit beita. En þá greiningu má taka lengra og beitir CLARA svokallaðri n-gram nálgun á textagreiningu. Sú nálgun felur í sér að litið er á textann sem samansafn orða sem hópast saman. Það er að segja að fyrir setninguna „ég fór út í búð“, þá er litið á þetta sem fimm orða hóp (fimmgram) eða þrjá hópa af þriggja orða hópum (þrjú þrígröm: ég fór út; fór út í; út í búð) og svo framvegis. Slík nálgun leyfir CLARA að sjá ákveðin mynstur n-grama sem koma fram í texta. Það býður svo upp á möguleikann að fara að athuga hvort þessi mynstur séu einkennandi fyrir ákveðna hópa. Í hvert skipti sem einstaklingur nefnir „ég fór út“, sama í hvaða tilgangi, þá fær það þrígram hækkun um einn í talningunni. Það opnast tveir möguleikar við það að skoða mynstur með n-grömum. Fyrst er það að skoða hversu oft ákveðið n-gram kemur fyrir, til dæmis „ég fór út“ og þar með að skoða mynstur orðflokka sem n-gröm. Seinni möguleikinn þýðir að litið er fram hjá orðunum

[3] Sótt 6.5.2010 af http://forums.d2jsp.org/

sjálfum og „ég fór út“ verður að þrígrami sem táknar fornafn, sögn og atviksorð eða jafnvel ennþá nákvæmar (persónufornafn 1. persónu eintölu nefnifall, sögn og atviksorð). Það þýðir að hvert skipti sem það mynstur kemur fram, myndi kerfið telja það.

Að blanda saman þessum aðferðum við greiningu texta opnar möguleikann fyrir ítarlegri greiningu. Í þróun eru eiginleikar í CLARA kerfinu sem munu koma að góðum notum við að sýna fram á tengslin milli greiningu tungumáls og einstaklingsmun. Þessir eiginleikar eru samanburðargreining, samhengisgreining og hugtakaskilningur. Hér verður farið stuttlega í þessa eiginleika.

3.1.1.2.1 Samanburðargreining

Það er mikilvægt að geta séð muninn á milli tveggja einstaklinga, tveggja hópa af einstaklingum eða tveggja umræðuefna. Til þess að geta gert það hefur verið í þróun samanburðagreining sem miðar að því að sjá hvað er einstakt í texta fyrir tvö viðfangsefni (einstaklingar, hópar eða umræðuefni). Þetta er gert með því að bera saman tvo texta, kasta burtu öllum n-grömum sem eru sameiginleg með þessum textum og raða svo með ákveðnum hætti þannig að þau n-gröm sem eru mest afgerandi fyrir hvorn texta fyrir sig komi fram. Með því er hægt sjá muninn á innihaldi eða stíl tveggja texta. Það er hægt að bera saman sama einstaklinginn á tveimur tímabilum og sjá hvort stíllinn breytist. Sömuleiðis væri hægt að sjá hvað væri einstakt í skrifum þunglyndra samanborið við heilbrigða.

Þessi samanburðargreining fæst ekki með hefðbundinni orðatalningu. Þar af leiðandi er þetta sérstaklega öflug sérstaða fyrir CLARA kerfið. Rannsóknir, sem beita n-gram nálgun á tungumálanotkun úthverfa og innhverfa einstaklinga hafa sýnt að notkun eingrama gefur ágæta yfirborðssýn á texta en mynstur sem fást með tví- og þrígrömum gefa skýrari mynd af tungumálanotkuninni (Gill og Oberlander, 2003; Oberlander og Gill, 2004). Í báðum rannsóknunum þurfti að nota nokkur kerfi við greiningu gagna. Það er því öflug sérstaða CLARA að geta hópað slíka greiningu undir eitt kerfi.

3.1.1.2.2 Samhengisgreining

Samanburðargreiningin gekk út á það að taka allt sem væri sameiginlegt með tveimur textum og kasta því í burtu til að sjá hvað væri einstakt.

Samhengisgreiningin gengur út á það að halda því sem er sameiginlegt með þessum tveimur textum sem verið er að skoða. Með samhengisgreiningunni er markmiðið að sjá fljótlega hvað er verið að tala um. Það er hægt að nota þessa samhengisgreiningu til þess að sjá hvað er sameiginlegt með ákveðnum einstaklingum eða ákveðnum hópum einstaklinga. Þannig væri hægt að greina mismunandi texta tveggja einstaklinga með þunglyndi og sjá hvað væri sameiginlegt með þeim. Ef viðfangsefnið er ólíkt, er líklegt að það sem er sameiginlegt milli þeirra er ákveðið orðanoktunarmynstur.

Samhengisgreiningin og samanburðargreiningin eru í raun tvær ólíkar aðferðir til þess að finna einstaklingsmun í texta. Ef nota á þessar tvær greiningar í rannsókn mætti segja að samhengisgreiningin væri með innanhópasniði (*within subjects design*) og samanburðargreiningin væri með millihópasniði (*between subjects design*).

Samanborið við önnur forrit þá nálgast CLARA samhengisgreininguna á einstakan hátt. The General Inquirer beitir frekar frumstæðri leið til þess að finna samhengi í texta. Í raun er það orðatalning með samhengi, þar sem orðabækurnar sem stuðst er við telja orð eftir því í hvaða samhengi þau eru notuð. Það þarf að viðhalda þessum orðabókum, sem krefst mikillar vinnu. LSA er þokkalega flókin aðferðafræði við greiningu á samhengi. Hún virkar mjög vel þegar markmiðið er að átta sig á samhengi mikils magns af ólíkum texta. Þar sem engar orðabækur eru notaðar og ekki er stuðst við orðflokka, þá er erfitt að meta hvort LSA myndi nýtast til þess að skoða einstaklingsmun byggt á samhenginu. CLARA nýtir sér orðabækur, orðflokkagreiningu og n-gram greiningu til þess að finna samhengi í texta.

3.1.1.2.3 Einræðing

Samhengisgreiningin býr til fleiri tækifæri heldur en að greina hvað er líkt með tveimur textaskrám. Tvíræðni hugtaka hefur oft verið nefnt sem ein af helstu hindrunum textagreiningatóla (Pennebaker, Mehl o.fl., 2007; Pennebaker og King, 1999). Með tvíræðni er átt við að orðið *Síminn* getur átt við tækið og fyrirtækið. Þessi eiginleiki CLARA er í prófun í dag og lofar góðu. Einræðni (*word sense disambiguation*) felst í samhenginu. Með því er átt að þegar það er verið að tala um *tækið*, þá eru einstaklingar líklegri til þess að nota orð sem tengjast tækinu (iphone,

sms, hringja, tala og fleira). Þegar einstaklingar eru að vísa í *fyrirtækið*, þá eru þeir líklegri til að nota orð sem tengjast því (Nova, Vodafone, ljósleiðari, inneign og fleira). Ef sérstakur áhugi er að skoða umfjöllun sem tengist *fyrirtækinu* þá væri sérstaklega hægt að taka það fram og öll umræða um *tækið* er ekki tekin með í greiningu. Það sem er að gerast er að fyrir orðið *síminn* er að það myndast ákveðnar þyrpingar (*clusters*) af orðum. Þessar þyrpingar byggja á því í hvaða samhengi er orðið er notað.

Einræðning er enn á þróunarstigi og þekkist ekki meðal annarra forrita. Það er þess vegna óskýrt hversu mikið hann myndi nýtast við greiningu á einstaklingsmun en ótvírætt er að hann mun nýtast í allri greiningunni hátt og lágt. Það er í raun aðferðafræðin á bakvið einræðninguna sem lofar hvað mestu í tengslum við skilning á einstaklingsmun.

Þyrpingarnar þurfa eitthvað viðfangsefni, sem þurfa ekki að vera orð. Það er vel hægt að sjá þyrpingar af einstaklingum. Það sem þarf þá er um hvað þessar þyrpingar myndast. Þær geta myndast í kringum ákveðin orð, orðflokka, orðflokkamynstur (n-gröm) og fleira áhugavert. Þetta þýðir að CLARA ætti að geta tekið þekkt orðflokkamynstur, sem til dæmis er algengt að þunglyndir nota, og séð hvaða höfundar hópast í kringum það. Þannig áætlað hvort ákveðnir höfundar séu þunglyndir eða ekki. Þetta er í raun hópun einstaklinga eftir að mynstur einstaklingsmunar eru þekkt.

3.1.1.3 Birting gagna

Það er augljóst að greining á gögnum getur verið mjög flókin og ítarleg. Það gerir birtingu gagna, nánar tiltekið niðurstaðna, eftir greiningu að mjög mikilvægum þætti. CLARA leggur mikið upp úr því að gera viðmótið mjög notendavænt. Endanotandinn á að geta séð niðurstöðurnar á einfaldan máta og geta skilið hvaða þýðingu þær hafa fyrir sig og sitt fyrirtæki. Þetta þýðir að þarf að leggja mikinn metnað í að búa til gröf og töflur sem eru auðskiljanlegar og krefast ekki mikillar þjálfunar í túlkun gagna eða tölfræði.

Notendur CLARA þurfa ekki að sækja forrit af netinu eða setja það upp á tölvu. Aðgangur er veittur í gegnum vefviðmót, sem þýðir að öll vinna og greining á sér stað á netþjónum CLARA. Þetta er ólíkt öðrum aðferðum sem krefjast þess að notandinn setji upp forrit á tölvunni sinni.

Kosturinn við að veita aðgang í gegnum netið er að allar breytingar sem verða á viðmóti uppfærast sjálfkrafa hjá öllum notendum. Þetta er einstaklega mikilvægt fyrir hugbúnað í þróun, þar sem hægt er að byrja að fá notendur til þess að nota hann áður en hann er í raun tilbúinn. Það kemur einnig í veg fyrir að auglýsa þurfi sérstaklega uppfærslur á hugbúnaðinum. Þetta er sérstaklega mikilvægt fyrir CLARA þar sem aðgangur að hugbúnaðinum er seldur í þjónustuformi, það er að segja viðskipavinir borga mánaðarlega áskrift.

Ókosturinn við að vefviðmót er sá að hraði kerfisins takmarkast að hluta við hraða nettengingar sem notandinn er með. Þegar á að gera mjög þungar greiningar þá getur það tekið nokkurn tíma áður en niðurstöðurnar birtast. Í besta falli er það ekki nema nokkrar sekúndur eða í versta falli ein til tvær mínútur.

3.1.2 Helstu ókostir

Þegar litið er á sérstöðu CLARA felst mikið af henni í nálguninni að greiningu gagna. Í dag er stór ókostur CLARAað geta ekki flokkað orð sérstaklega eftir flóknari víddum heldur en orðflokkum. LIWC og fleiri forrit nota orðabækur, sem eru mjög fágaðar og ítarlegar. Þessar orðabækur gera þessum forritum kleift að greina orð eins og grét í fleiri flokka heldur en bara orðflokkinn *sögn*. Það að geta merkt hvort tíð sagna, tilfinningavægi þeirra og annað í þá átt, gefur möguleikann á finna nákvæmari mynstur. Því ítarlegri sem upplýsingarnar eru því betur gengur að finna nákvæmari mynstur sem skilar sér í betri skilningi á mun milli einstaklinga.

Enskugreining CLARA er enn í þróun og er kerfið þróað þannig að það getur tekið við hvaða orðabók sem er og nýtt hana við greiningu. Þannig er ekkert í fyrirstöðu fyrir því að CLARA muni nýta orðabækur sem leyfa ítarlegri greiningu. Það er óraunhæft að CLARA muni þróa sína eigin orðabækur líkt og orðabækurnar á bakvið LIWC, þar sem það er áralangt ferli (sjá betur Pennebaker, Chung o.fl., 2007). Þess í stað verður að leita til opins hugbúnaðs (*open source*), sem er opinn öllum þeim sem hafa áhuga og hugbúnaðs til sölu. Opinn hugbúnaður er í þessu tilviki orðabækur og forrit, sem hafa verið búin til af ákveðnu samfélagi

áhugamanna um tungumálagreiningu. Öll vinna þeirra er opin þeim sem vilja nýta þá þekkingu, gegn því að þeir skili til baka í samfélagið aukinni þekkingu. CLARA nýtir í dag breska landsritsafnið (*British National Corpus*[4]) við enskugreiningu. Ritsafnið var keypt og gerir CLARA kleift að merkja til dæmis tíð sagna. Fleiri ritsöfn, orðabækur og forrit verða keypt og nýtt í framtíðinni til þess að bæta enska greiningu. Það er þó ágætar líkur á því að hvergi verðir hægt að nálgast orðabækur jafn sálfræðilega fágaðar og þær sem LIWC styðst við til dæmis. Þetta er ókostur sem þarf að bæta upp með öðrum leiðum.

Annar ókostur er val á gögnum sem á að skoða. CLARA mun einbeita sér að skoða spjallvefi hjá tölvuleikjaframleiðendum. Málfars-, stafsetningar- og prentvillur eru mjög algengar á vefnum. Það veldur því að greiningartólin merkja ranglega stafsett orð sem eitthvað sem það er ekki eða nær ekki að merkja það. Þetta gæti valdið því að CLARA muni vanmeta tíðni ákveðinna orða eða orðflokkamynstra.

Að skoða spjallvefi þýðir að það þurfi að sýna fram á fyrrgreindar niðurstöður eigi við á þeim vettvangi. Þetta er í raun ekki ókostur en frekar almenn staðreynd sem þarf að taka alvarlega. Það þarf að athuga sérstaklega hvort texti sem er skrifaður á spjallvefum sé jafn áreiðanlegur yfir tíma og aðstæður (til dæmis á milli þráða). Það er ekki hægt að prófa hugmyndir annarra fyrr en það er búið að sýna fram á áreiðanleikann. Jafnvel þá er ekki víst að hugmyndir annarra séu viðeigandi á spjallþráðum. Þetta þarf að rannsaka sérstaklega.

3.2 Hvernig á að prófa kerfið

Það er eðlilegt að við prófun á CLARA kerfinu að það fylgi sömu röð af prófunum og önnur kerfi hafa fengið. Það þýðir að fyrst þarf þarf að sýna fram á próffræðilega eiginleika greiningar, næst verði farið að skoða hvort hægt sé að

[4] http://www.natcorp.ox.ac.uk/getting/index.xml.ID=cost

greina einstaklingsmun og að lokum fleiri hugmyndir og fleiri svið prófuð. Hér verður farið í hvernig prófa eigi þessa þætti.

3.2.1 Staðfesting próffræðilegra eiginleika

3.2.1.1 Áreiðanleiki

Til þess að prófa próffræðilega eiginleika CLARA er hægt að líta til þeirrar aðferðafræði sem var nýtt til þess að staðfesta próffræðilega eiginleika LIWC (Pennebaker, Booth o.fl., 2001; Pennebaker, Chung o.fl., 2007; Pennebaker og King, 1999). Það þarf þó að hafa í huga að ólíklegt er að spjallþræðir hafi sömu eiginleika og þau verkefni sem athuguð voru fyrir LIWC. Þar af leiðandi er eftirfarandi aðferð lögð til.

Fyrst verður hluti af Pennebaker og King (1999) rannsókninni endurtekinn. Þar nýttu þau sér skólaverkefni nemenda sem áttu að skrifa um hin ýmsu málefni, svo sem atburði líðandi stundar og mat á áfanga. Miðað var við að hver nemandi eyddi 20 mínútum 10 daga í röð (fyrir utan helgar) til þess að skrifa um þessu ólíku málefni. Þessi verkefni yrðu lesin inn af CLARA og borin saman með tilliti til stöðugleika yfir tíma og milli aðstæðna (ólík málefni). Þar sem CLARA hefur ekki sambærilegar orðabækur og LIWC þá er lagt til aðeins verður stuðst við greiningu á samsetningu tungumáls (*language composition*) til að byrja með. Í því felst að hlutföll fyrir hvern orðflokk af heildinni eru reiknuð fyrir hvert verkefni. Að því loknu er fylgni milli þessara hlutfalla reiknuð fyrir hvern orðflokk. Þannig er hægt að sjá hversu stöðugt einstaklingar nota orðflokka yfir tíma og á milli aðstæðna (sjá Pennebaker og King, 1999, fyrir nákvæmari útlistun á aðferðafræði og niðurstöður). Markmiðið hér verður að sýna fram á að CLARA getur ná sambærilegum árangri og aðrar lausnir.

Þegar búið er að sýna fram á og ef sambærilegar niðurstöður fást, þá er hægt að hefjast handa við að sýna fram á próffræðilega eiginleika tungumáls spjallnotenda. Það verður gert með því að safna gögnum um spjallnotendur tölvuleiksins *Eve Online*. Þessi tölvuleikur er framleiddur af CCP, sem er íslenskt

tölvuleikjafyrirtæki, og hefur verið í góðu samstarfi við CLARA hingað til. Spjallvefurinn þeirra telur rétt rúmlega 1,2 milljónir þráða[5]. Vefköngulær CLARA verða nýttar til þess að sækja allan textann, tíma- og dagsetningar og upplýsingar um höfunda þráða og svara. Til þess að tryggja að hægt sé að sjá áreiðanleika í tungumálanotkun yfir tíma og milli aðstæðna verður aðeins stuðst við texta frá þeim höfundum sem mæta eftirfarandi kröfum: 1) Hafa skrifað fleiri en einn þráð eða fleiri en eitt svar, 2) heildar orðafjöldi nær yfir 100 orð og 3) þræðir og svör þurfa að hafa verið skrifaðir á fleiri en einu spjallsvæði innan spjallvefsins. Ekki er hægt að fá gögn um hversu margir þátttakendur falla undir þessar kröfur að svo stöddu. Þar sem yfir 300.000 einstaklingar spila *Eve Online* má fastlega reikna með að ekki ætti að skorta þátttakendur eftir að þessum kröfum hefur verið framfylgt. Þar næst verður hafist handa við að endurtaka greinguna sem skilgreind var hér að ofan. Markmiðið hér væri að sjá hvort tungumálanoktun væri sambærilega áreiðanleg og í fyrra verkefni.

Að þessu loknu ættu niðurstöður að benda til þess hvort hægt sé að reikna með því að CLARA nái að greina stöðugleika með sambærilegum hætti og aðrar lausnir (fyrra verkefnið). Einnig hvort tungumálanotkun á spjallvefjum sé áreiðanleg.

3.2.1.2 Réttmæti

Það þarf að sýna fram á að CLARA geti með réttmætum hætti greint einstaklingsmun. Til þess er þörf að endurtaka sambærilegar prófanir á öðrum lausnum. Aftur er leitað til rannsóknar Pennebaker og King (1999) sem sýndu fram á réttmæti LIWC með því að finna fylgni við viðurkenndar persónuleikamælingar. Þetta er tveggja þrepa rannsókn. Fyrst þarf að sýna fram á að hægt sé að þáttagreina byggt á tungumálanotkun og svo þarf að bera þáttagreiningu saman við viðurkenndar persónuleikamælingar.

[5] http://www.eveonline.com/ingameboard.asp

Við þáttagreiningu er aðeins hægt að styðjast við þá flokka sem fengu nægilega háan áreiðanleikastuðul í áreiðanleika prófuninni. Pennebaker og King (1999) miða við 0,60 eða hærra og er ágætt að miða einnig við það hér. Sömuleiðis er miðað við að flokkur þarf að hafa að minnsta kosti 1% grunnnotkun í texta til þess að vera tekinn með í þáttagreiningu.

Fyrir CLARA er mikilvægt að þáttabygging tungumáls komi fram á spjallvefum. Þar af leiðandi verður stuðst við sömu gögn og voru tilgreind fyrir *Eve Online* spjallvefinn hér að ofan. Þannig er hægt að sjá hvort ákveðin þáttabygging myndast og hvaða orðflokkar og orðflokkamynstur hópast saman undir hvaða þáttum.

Ef hægt er að sýna fram á þáttabyggingu tungumáls á spjallvefum þá er hægt að fara í afmarkaðari rannsókn til þess að kanna réttmæti betur. Það er hér sem að málin flækjast. Til þess að geta sýnt fram á réttmæti þarf að bera saman þáttabygginguna sem kemur fram hjá fólki við niðurstöður annarra persónuleikamælinga. Það krefst þess að það þarf að afla þeirra mælinga fyrir þátttakendur. Ólíklegt er að hægt sé að nálgast þátttakendur í tölvuleiknum *Eve Online*, þannig leita þarf annarra leiða til þess að kanna réttmæti.

Lagt er til að fyrst verðir gerð forrannsókn. Sú forrannsókn mun hafa fáa þátttakendur og ekki miða við texta sem skrifaður er á spjallvefi. Í staðinn er miðað við svipuð verkefni og fást í fyrra verkefninu við áreiðanleika prófunina. Sem dæmi væri hægt að fá hluta af nemendunum til þess að taka persónuleikapróf á borð við *NEO PI-R*. Þessi forrannsókn myndi leiða í ljós hvort CLARA yfir höfuð gæti sýnt réttmæti við mat á einstaklingsmun. Ef sú forrannsókn hefur jákvæða niðurstöður fyrir CLARA væri hægt að fara í ítarlegri rannsókn sem myndi skoða fleiri þátttakendur og miða að því að sýna fram á réttmæti á greiningu einstaklingsmun á spjallvefum. Einnig væri hægt að gefa sér að niðurstöður eigi við sama á hvaða vettvangi textinn er skrifaður.

3.2.2 Rannsóknir

Tæknileg sérstaða CLARA, sem var lýst hér að ofan, opnar möguleikann að fara að gera nýjar rannsóknir. Helsta sérstaða CLARA felst í greiningu á gögnunum. Þessi sérstaða fram yfir aðrar lausnir er í stuttu máli n-gram nálgunin, samanburðargreining, samhengisgreining og hugtakaskilningur. Það er hægt að nýta

þessa sérstöðu til þess að endurtaka rannsóknir með nýjum aðferðum eða að koma með nýjar og áhugaverðar rannsóknir.

3.2.2.1 Endurtekning rannsókna

Þegar búið er að sýna fram á próffræðilega eiginleika tungumálagreiningar með CLARA þá er hægt að hefjast handa við að endurtaka rannsóknir. Með því að endurtaka rannsóknir á nýjum vettvangi, nánar tiltekið spjallvefjum, er hægt að staðfesta hvort tengsl séu á milli orðavals og persónuleikaþátta á þeim vettvangi. Einnig er hægt að endurtaka rannsóknir með sambærilegri aðferðafræði en beita annars konar greiningu. Sem dæmi kemur til greina að endurtaka þær rannsóknir sem greint er frá í kaflanum um *niðurstöður textagreininga*.

Ein rannsókn sem væri áhugavert að endurtaka er rannsókn Oberlander og Gill (2004). Þeir nýttu n-gram nálgun við textagreiningu sem leiddi í ljós tengsl milli ákveðinna orðaflokka og úthverfs- og innhverfs tungumáls. Niðurstöðurnar voru samt ekki eins sannfærandi og áætla mætti í n-gram greiningu. Þeir útskýrðu niðurstöðurnar þannig að þótt að það sé ekki munur á hlutfallslegri tíðni orðflokka milli úthverfu og innhverfu, þá gæti annar hópurinn verið að nota orðflokka í mismunandi samhengi.

Það væri hægt að endurtaka þessa rannsókn, með svipuðum gögnum og þátttakendum, og athuga hvort samhengisgreining CLARA gæti sýnt fram á meira sannfærandi tengsl milli tungumáls og úthverfu og innhverfu. Í þeirri rannsókn myndi CLARA einangra n-gram mynstur einstaklinga bundið við ákveðið samhengi og þannig greina sérstaklega hvort hægt sé að tengja þessi samhengisbundnu n-gram mynstur við tungumál úthverfu og innhverfu. Ef samhengið er stöðugt ættu vandamálin sem Oberlander og Gill lentu í, ekki að hafa áhrif. Það tryggir þó ekki að einhver einstaklingsmunur komi fram.

3.2.2.2 Nýjar rannsóknir

Eitt helsta verkefni sem stendur frammi fyrir CLARA er að geta veitt sínum viðskiptavinum ítarlegri greiningu á hegðun viðskiptavina þeirra. Þar sem CLARA mun einbeita sér að tölvuleikjaiðnaðinum liggur beint við að skoða eitt stærsta vandamálið sem hrjáir þann iðnað. Það vandamál er brottfall einstaklinga frá tölvuleikjum. Rannsóknarspurningin gæti þar af leiðandi verið: Getur textagreining

á samskiptum einstaklinga á spjallvefum tölvuleikja gefið vísbendingar um hvort þeir séu líklegri til þess að hætta að spila tölvuleikinn.

Til þess að skoða þessa rannsóknarspurningu þyrfti CLARA að skoða spjallhegðun þeirra sem vitað er að hafa hætt. Það krefst þess að öllum upplýsingum um spjallhegðun einstaklingsins er safnað; textinn sjálfur, hversu oft hann skrifar, hvenær dags hann skrifar, hver er meðallengd texta og fleira í þá áttina. Það þyrfti einnig að fá upplýsingar um sjálfa höfundana; aldur, kyn, hversu lengi þeir spiluðu og hvenær þeir hættu.

Fyrsta verkefnið væri að athuga hvort samanborið við þá sem hafa ekki hætt hvort það séu einhver einstök mynstur í textanum sem þeir skrifa. Til þess þyrfti að nýta samanburðargreiningu CLARA til þess að kasta burtu öllum mynstrum sem væru sameiginleg og skoða eins það sem aðgreinir þessa hópa. Ef þessi greining gæti aðgreint sérstök mynstur meðal þeirra sem hætta þá væri hægt að nota eiginleikann til þess að finna þyrpingar til að greina sambærileg mynstur hjá notendum sem enn hafa ekki hætt.

4 Framtíðarnálgun

4.1 Vilji rannsakenda

Pennebaker og fleiri (2002) lýstu því sem þeir sáu sem framtíðina í notkun á textagreiningu við rannsóknir á einstaklingsmun. Þar bentu þeir á að skoða smáorð væri áreiðanlega besta leiðin til þess að nálgast skilning á einstaklingsmun. Þeir tóku það samt fram að sérstaklega þyrfti að athuga hlutverk hvers og eins þeirra. Þar koma fyrstu persónufornöfn efst upp í huga. Í rannsóknum hafa komið í ljós tengsl fyrstu persónufornafna við þunglyndi (Pennebaker og Lay, 2002), áhrifa áfalla og vandamála sem fólk upplifir saman (sem dæmi hryðjuverkárásin í New York, Cohn o.fl., 2004), persónuleikaþætti á borð við taugaveiklun, úthverfu og víðsýni, sjálfsálit og félagslega stöðu (Chung og Pennebaker, 2007; Mehl og Pennebaker, 2003) og jafnvel heiðarleika (Newman o.fl., 2003). Þessar ólíku niðurstöður benda til þess að greina þurfi hvað skilur þarna á milli. Í hvaða samhengi nota þunglyndir fyrstu persónufornöfn og í hvaða samhengi gera heiðarlegir einstaklingar það? Það er þarna sem CLARA hefur mikla möguleika að bæta fræðilega þekkingu. Með því að beita ítarlegri og betri greiningum er hægt að sjá samhengi í umræðunni og byggja niðurstöður á því.

4.2 Næstu skref

Til þess að geta gert þær rannsóknir sem hafa fengið umfjöllun hér þarf fyrst að ná þróun CLARA á það stig að svo sér hægt. Það er ákveðinn kostur að tæknileg þróun er ekki komin eins langt og aðrar lausnir. Það gefur meira rúm fyrir að nýta þá fræðilegu þekkingu sem hlotist hefur á hvað virkar og hvað virkar ekki þegar nýta á textagreiningu til þess að finna einstaklingsmun.

Næstu skref CLARA verða að ná tæknilegri þróun á það stig að hægt sé að framkvæma þessar rannsóknir. Þegar það er orðið að veruleika geta einstaklingar sem stunda grunn- og framhaldsnám í sálfræði, markaðsfræði eða málvísindum nýtt CLARA til þess að prófa þær hugmyndir sem skilgreindar hafa verið hér eða komið með nýjar tilgátur sem þeir vilja prófa.

Á næstu mánuðum mun CLARA vinna að verkefni með CCP, sem framleiðir *Eve Online* tölvuleikinn og PatternVision, sem sérhæfir sig í mynsturgreiningu í flóknum gagnasöfnum. Þetta verkefni miðar að því að sjá hvort hægt sé að greina mynstur í texta sem spjallnotendur *Eve Online* sýna. CLARA mun sjá um greiningu á textanum en PatternVision mun aðstoða við mynsturgreininguna. CCP sér um að veita upplýsingar um spjallnotendur.

Áður en þetta verkefni getur farið að stað þarf að sýna fram á próffræðilega eiginleika CLARA eins og tilgreint var hér að ofan. Það verkefni mun fara af stað um leið og ensk tungumálagreining er orðin nægilega fullkomin til þess að það sé raunhæft að nota hana.

5 Samantekt og lokaorð

Persónuleiki er einstaklega heillandi viðfangsefni. Á 20. öldinni hófust rannsóknir á persónuleika fyrir alvöru innan sálfræðinnar. Allport og Odbert (1936) voru upphafsmenn svonefndrar orðasafnsnálgunar (Í Chung og Pennebaker, 2008) sem svo var skilgreind með varanlegri hætti af Goldberg (1981) sem orðasafntilgátan (Í Ashton og Lee, 2005). Samkvæmt þeirri hugmynd býr fólk, yfir tíma, til orð fyrir þau sérkenni sem eru lýsandi fyrir einstaklinga. Það er þessi hugmynd sem hvatti áfram rannsakendur sem vildu athuga hvort hægt væri að nýta þessi orð, sem einstaklingar hafa búið til fyrir einstaklingsmun, til þess að rannsaka persónuleikavíddir. Sumir rannsakendur hófu að þróa matstæki til þess að geta rannsakað þessar víddir. Þessi matstæki ganga út á að það að spyrja fólk spurninga

um skoðanir og hegðanir þess í þeirri von að þau svör sem það gefur veita innsýn í persónuleika fólksins. Aðrir rannsakendur fóru aðra leið og leituðu að vísbendingum í tungumálinu sjálfu. Þeir þróuðu flóknar aðferðir við að greina tungumál, sumar þeirra voru handvirkar en aðrar byggðust á sjálfvirkum forritum. Með tæknilegum framförum hafa þessi textagreiningartól orðið öflugri og hafa leitt í ljós áhugaverðar niðurstöður sem gefa vísbendingar um tengsl persónuleika og tungumáls. Það eru þó einhverjar takmarkanir á getu þessara tóla, þar sem flest þeirra hafa verið notuð í mjög afmörkuðum tilgangi. Þar af leiðandi er mjög mikilvægt að við þróun á nýjum hugbúnaði, sem á að geta metið persónuleika, að þessar takmarkanir séu teknar með í reikninginn.

CLARA er íslenskur hugbúnaður sem hefur verið í stöðugri þróun í tvö ár. Tæknileg sérstaða hans getur haft mikil áhrif á framþróun á sviði persónuleikamats með tungumálagreiningu. Með því að beita n-gram nálgun við textagreiningu opnast fyrir möguleikinn að greina samhengi í texta. Þetta samhengi er ekki skilningur á um hvað er verið að tala, heldur frekar í hvaða samhengi ákveðnum orðum og orðflokkum er beitt. Þessi samhengisgreining býður upp á að hægt sé að byrja að greina mjög flókin mynstur í texta. Sem dæmi hafa niðurstöður fyrri rannsókna (sem dæmi, Chung og Pennebaker, 2007; Mehl og Pennebaker, 2003; Pennebaker og Lay, 2002) sýnt að persónufornöfn spila mikilvægan þátt í tungumáli þunglyndra, sjálfsálits og hina ýmsu persónuleikaþátta, svo sem taugaveiklunar og úthverfu. Með samhengisgreiningu er möguleiki á að athuga í hvaða samhengi persónufornöfn eru mikilvæg fyrir tungumál þunglyndra og í hvaða samhengi þau eru í úthverfu tungumáli. Samanburðargreining opnar síðan möguleikann að skoða hver munurinn er á milli þeirra sem eru innhverfir og úthverfir. Þessi samanburðargreining getur stutt við samhengisgreininguna við það að túlka þessi flóknu tungumálamynstur sem fólk sýnir. Einræðing og notkun þyrpinga myndi síðan nýta þá þekkingu sem myndast hefur um þessi ólíku mynstur til þess að greina út frá nýjum texta hvort ákveðið fólk sýnir sömu tungumálamynstur. Þetta fólk myndi þá þyrpast í kringum þetta mynstur og hægt væri að áætla að þeir tilheyrðu sama hópi.

Það er búið að mála fallega mynd af framtíð persónuleikamats en til þess að hún verði að veruleika þarf að komast yfir þó nokkuð margar hindranir. Orðasafnsnálgunin hefur verið gagnrýnd og textagreiningatól verða aldrei jafn öflug

og mannlegt mat. Sjálfsmatsmælitæki eins og þau sem meta hina fimm stóru eða persónuleikavíddir Eysencks eru umdeild. Þau hafa meðal annars verið gagnrýnd fyrir að útskýra illa mismun milli einstaklinga við mismunandi aðstæður og að sjálfsþekking leikmanna sé ekki nægilega góð til að útskýra hina flóknu þætti persónuleika (Ashton og Lee, 2005). Þessi gagnrýni hefur mikla þýðingu fyrir textagreiningatól, sem eiga rætur sínar að rekja til orðasafnsnálgunarinnar að persónuleika. Ef einstaklingar eru ólíklegir til þess að haga sér eins við mismunandi aðstæður eru allar líkur á því að þeir muni ekki tala eins eða beita tungumálinu á stöðugan hátt. Þetta býr til vandamál, sem er mikilvægt þar sem tungumálagreining gerir kröfu um stöðugleika ef niðurstaðan eigi að vera nothæf. Sömuleiðis þarf að sýna fram á réttmæti textagreiningartóla með því að bera niðurstöður saman við niðurstöður annarra persónuleikaprófa. Ef sjálfsþekking fólks er vandamál þá verður réttmæti textagreiningartóla einnig skekkt. Þessa gagnrýni er hollt að hafa í huga þegar verið er að þróa nýjan hugbúnað, sem þarf að sýna fram á að tungumál sé áreiðanleg mæling og jafnvel leita annarra leiða til þess að sýna fram á réttmæti.

Það liggur fyrir að CLARA þarf að komast yfir þessar hindranir. Þessi hugbúnaður er hugsaður sem þjónusta, sem á að selja til viðskiptavina, og þar af leiðandi er ósættanlegt að það sé óvissa um fræðilegt gildi hans. Með því að gera viðeigandi prófanir og sýna fram á sambærilegar niðurstöður og aðrir hafa fengið, þá er hægt að halda áfram með ákveðinni vissu. CLARA og önnur textagreiningartól verða verðmætust þegar þau geta farið að spá fyrir um hegðun. Tungumálagreining og frávarpsaðferðir deila því að þar er verið að skoða undirliggjandi ferla, sem oftast eru ekki á valdi einstaklingsins að hafa áhrif á. Suma gagnrýni á orðasafnsnálgunina er hægt að nota sem vörn fyrir frávarpsaðferðir, þar sem ekki eru um sjálfsmat að ræða. Sem dæmi þá er sjálfsþekking einstaklingsins ekki vandamál, þar sem þjálfaður sérfræðingur sér um að túlka frávarpsprófin. Frávarpsaðferðir eru hins vegar gagnrýndar fyrir lágan áreiðanleika og réttmæti, sem og að spá illa fyrir um hegðun (Lilienfeld o.fl., 2001). Stóra spurningin er hvort CLARA kemst yfir þá hindrun og getur farið að spá fyrir um hegðun einstaklinga með góðum árangri.

Það er ljóst að tæknin verður aldrei fullkomin, en hún lofar góðu. Rétt áhersla á réttu þættina, sem koma í ljós með rannsóknum í persónuleikasálfræði, við þróun á hugbúnaði sem greinir tungumál er mjög mikilvæg. Hér hefur slíkur

hugbúnaður verið kynntur. CLARA hefur burðina til þess að vera leiðandi á sviði tungumálagreiningar. Með réttri stefnu og réttum áherslum eru möguleikarnir margir.

Heimildaskrá

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt

Argamon, S., Dhawle, S., Koppel, M. og Pennebaker, J. W. (2005). Lexical predictors of personality type. Í *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*.

Ashton, M. C. og Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, 19, 5-24.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*(2), 187-215.

Camara, W. J., Nathan, J. S. og Puente, A. E. (2000). Pychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*(2), 141-154.

Cattell, R. B., Cattell, A. K., og Cattell, H. E. P. (1993). *16PF fifth edition questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Chung, C. K, og Pennebaker, J. W. (2007). The psychological functions of function words. Í K. Fiedler (ritstjóri), *Social communication* (bls. 343-359). New York: Psychological Press.

Chung, C. K. og Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality, 42*(1), 96-132.

Cohn, M. A., Mehl, M. R. og Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*(10), 687-693

Costa, P. T. Jr., og McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Dewaele, J. M. og Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49*, 509-544.

Durand, V. M., Blanchard, E. B. og Mindell, J. A. (1988). Training in projective testing: Survey of clinical training directors and internship directors. *Professional Psychology: Research and Practice, 19*(2), 236-238.

Eysenck, H. J. og Eysenck, S. B. G. (1975). *Manual of the Eysenck personality questionnaire*. London: Hodder and Stoughton.

Foltz, P. W. (1996) Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers. 28*(2), 197-202.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology: Interdisciplinary and Applied, 8*, 389-413.

Furnham, A. (1990). Language and personality. Í H. Giles og W. Robinson (ritstjórar), *Handbook of language and social psychology* (bls. 73-95). Chichester: Wiley.

Gill, A. J. og Oberlander, J. (2003). Looking forward to more extraversion with n-grams. Í Lagerwerf, L., Spooren, W. og Degand, L. (ristjórar), *Determination of information and tenor in texts* (bls. 125-137). Amsterdam: Stichting Neederlandistiek.

Gleitman, H., Fridlund, A. J. og Reisberg, D. (2004). *Psychology* (3. útgáfa). New York: W. W. Norton

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Goulden, R., Nation, P. Og Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*(4), 341-363

Hart, R. P., og Carroll, C. (2000). *DICTION 5.0 manual: The text analysis program* [tölvuforrit]. Austin, TX: Digitext, inc.

Hart, R. P., Jarvis, S. E., Jennings, W. P. og Smith-Howell, D. (2005). *Political keywords: Using language that uses us*. New York: Oxford University Press.

Hood, A.B., og Johnson, R.W. (1997). *Assessment in counseling: A guide to the use of psychological assessment procedures* (2. útgáfa). Alexandria, VA: American Counseling Association.

Heiða María Sigurðardóttir. „*Af hverju nota sálfræðingar svartar klessumyndir og spyrja sjúklingana út í þær?*“. Vísindavefurinn 7.6.2005. http://visindavefur.is/?id=5039. (Skoðað 22.4.2010).

Lilienfeld, S. O., Wood, J. M. og Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 23*(5), 32-39.

Lilienfeld, S. O., Wood, J. M. og Garb, H. N. (2001). What's wrong with this picture? *Scientific American, 284*(5), 80-87.

Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49*, 141-149.

McAdams, D. P. (1997). A conceptual history of personality psychology. In R. Hogan, J. Johnson & S. Briggs (ritstjórar), *Handbook of personality psychology* (bls. 3-39). San Diego: Academic Press.

Mehl, M. R. (2005). Quantitative text analysis. Í M. Eid & E. Diener (ritstjórar), *Handbook of multimethod measurement in psychology* (bls. 141-156). Washington, DC: American Psychological Association.

Mehl, M. R. og Pennebaker, J. W. (2003). The social dynamics of a cultural upheaval: Social interactions surrounding September 11, 2001. *Psychological Science, 14*(6), 579-585.

Newman, M. L., Pennebaker, J. W., Berry, D. S. og Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Oberlander, J. og Gill, A. (2004). Individual differences and implicit language: Personality, parts-of-speech and pervasiveness. *Í Proceedings of the 26th annual conference of the Cognitive Science Society*, 1035-1040.

Pennebaker, J. W. (1997). *Opening up: The healing power of expressing emotions*. New York: Guilford Press.

Pennebaker, J. W., Booth, R. J., og Francis, M. E. (2001). *Linguistic Inquiry and Word Count: A text analysis program* [tölvuforrit]. Austin, TX: LIWC.net.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. og Booth, R. J. (2007). *LIWC2007 Manual: The development and psychometric properties of LIWC2007* [tölvuforrit]. Austin, TX: LIWC.net.

Pennebaker, J. W. og Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*, 90-93.

Pennebaker, J. W. og King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Pennebaker, J. W. og Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 271-281.

Pennebaker, J. W., Mayne, T. J. og Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*, 863-871.

Pennebaker, J. W., Mehl, M. R. og Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

Pennebaker, J. W. og Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85*(2), 291-301.

Rannís. (2009, febrúar). *CLARA*. Sótt 1. maí 2010 af http://www.rannis.is/sjodir/taeknithrounarsjodur/verkefnalisti/nr/1751/

Ritzler, B. (1995). Putting your eggs in the content analysis basket: A response to Aronow, Reznikoff, and Moreland. *Journal of Personality Assessment, 64*, 229-234.

Sargent, H. (1945). Projective methods: their origins, theory, and application in personality research. *Psychological Bulletin, 42*(5), 257-293.

Slatcher, R. B., Chung, C. K., Pennebaker, J. W. og Stone, L. D. (2006). Winning words: Individual differences in linguistic styles among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*, 63-75.

Tausczik, Y. R. og Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods [rafræn útgáfa]. *Journal of Language and Social Psychology*. Sótt 1. maí 2010 af http://jls.sagepub.com/cgi/content/abstract/29/1/24.

The Huffington Post. (2010, apríl). *Twitter user statistics revealed*. Sótt 5. maí 2010 af http://www.huffingtonpost.com/2010/04/14/twitter-user-statistics-r_n_537992.html.

Watkins Jr., C. E., Vicki, L. C., Nieberding, R. og Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*(1), 54-60.

Wood, J. M., Nezworski, M. T. og Stejskal, W. J. (1996). The comprehensive system for the Rorschach: *A critical examination. Psychological Science, 7*(1), 3-10.