Esseetehtävien tietokoneavusteinen arviointi

Tuomo Kakkonen

13.1.2003

Joensuun Yliopisto
Tietojenkäsittelytiede
Pro gradu -tutkielma

Tiivistelmä

Tietokoneita käytetään arvioinnin apuvälineenä monin eri tavoin. Eräs automaattisen arvioinnin sovellus on esseiden arviointi, johon on kehitetty menetelmiä jo 1960-luvulta lähtien. Arvioinnin automatisoinnilla pyritään saavuttamaan erityisesti kustannussäästöjä. Lisäksi se auttaa objektiivisen arvioinnin saavuttamisessa. Esseiden arvioinnin automatisointi on kehittyvä tutkimusala, johon sisältyy edelleen myös ratkaisemattomia ongelmia. Koska esseiden arviointia pidetään yleisesti inhimillistä käsityskykyä vaativana tehtävänä, on pyrkimys sen automatisointiin aiheuttanut myös epäilyjä ja voimakastakin kritiikkiä.

Suomen kieli aiheuttaa tiedonhaun sovelluksille eräitä erityisvaatimuksia, johtuen mm. sijamuotojen suuresta määrästä. Nämä erityispiirteet vaikuttavat myös automaattisen arvioinnin soveltamiseen suomen kielisten esseiden arviointiin. Osana tutkielmaa toteutettiin suomen kielisten esseiden arviointijärjestelmä. Järjestelmä käyttää morfologista jäsennintä suomen kielen käsittelyyn liittyvien ongelmien ratkaisemiseksi.

Avainsanat: tietokoneavusteinen arviointi, automaattinen arviointi, esseiden arviointi, Latent Semantic Analysis, Project Essay Grade, Naiivi Bayes -luokitin, Text Categorization Technique

Sisällysluettelo

1. JOHDANTO

Tietokoneita käytetään opetuksen apuvälineenä yhä laajemmin (CAA Centre 2002). Myös tietokoneavusteisen kuulustelun sovelluksia on ollut markkinoilla jo jonkin aikaa. Suurin hyöty niistä saavutetaan kustannussäästöinä, jos tietokonetta voidaan käyttää hyväksi myös vastausten arvioinnissa. Esimerkiksi monivalintatehtävien pisteytys tietokoneella on varsin yksinkertainen tehtävä ja paperille tehdyt vastaukset on muunnettavissa tietokoneen ymmärtämään muotoon automaattisesti mm. optisten lukijoiden avulla. Tietokoneita on käytetty myös erilaisten lyhyiden vastausten tarkastamiseen, esimerkiksi puuttuvan sanan täydennystehtävissä. Automaattista arviointia on sovellettu onnistuneesti myös ohjelmointitehtävien arvioinnissa (Benford et al. 1993, Joy ja Luck 1998, Foxley et al. 2001, Malmi 2002).

On selvää, ettei ainoastaan monivalinta- tai sanantäydennystehtävien käyttö ole riittävä tapa arvioida opiskelijan tietämystä. Tästä syystä on pyritty kehittämään menetelmiä vapaata tekstiä sisältävien vastatusten, esseiden, arvioinnin automatisoimiseen (mm. Page 1966, Page ja Petersen 1995). Tietokoneen käyttö esseiden arvioinnin apuvälineenä tai avustajana, jonka antamia arvosanoja arvioija voi käyttää tukenaan, voi kustannussäästöjen lisäksi auttaa oikeudenmukaisen ja objektiivisen arvioinnin saavuttamisessa.

Esseevastausten arviointi on vaativa ja aikaa vievää työtä (Hopkins et al. 1990, Thorndike 1971). Siten myös vastausten arvioinnissa aiheutuvat kustannukset ovat merkittäviä. Tietokoneiden yleistyessä ja laskentakapasiteetin kasvaessa on niiden käytöstä myös esseevastausten arvioinnin apuvälineenä kehittymässä varteenotettava ja käytännönläheinen mahdollisuus (mm. ETS Technologies 2002). Jotta esseiden arviointi automaattisesti olisi mahdollista, on vastaukset saatava tietokoneen ymmärtämässä muodossa. Helpoimmin tämä tapahtuu tietenkin siten, että opiskelijat kirjoittavat vastauksensa suoraan tietokoneella.

Aloitan tutkimuksen esittelemällä esseetehtäviä arvioinnin välineenä sekä tarkastelemalla ihmistä esseen arvioijana. Lisäksi esittelen esseiden arvioinnin automatisointiin liittyviä periaatteita sekä arvioinnin automatisointiin kohdistettua kritiikkiä ja sen perusteita. Tarkoituksena on luoda katsaus aihepiirin käsitteisiin sekä esitellä niitä lähtökohtia, joiden valossa arvioinnin automatisointiin pyrkiminen on mielekästä.

Esittelen tutkielmassa viisi olemassa olevaa automaattista esseevastausten arviointimenetelmää. Esiteltävät menetelmät eroavat lähestymistavaltaan sekä toteutukseltaan toisistaan osittain merkittävästikin. Yhteistä kaikille menetelmille on niiden avulla tutkimuksissa saavutetut hyvät tulokset. Esitellyt menetelmät luovat varsin kattavan kuvan alan tutkimuksen merkittävimmistä kehityssuunnista. Tarkastelen tutkielmassa menetelmien eroja ja yhtäläisyyksiä useasta eri näkökulmasta.

Osana tutkielmaa on toteutettu Java-sovellus, jolla voidaan arvioida suomen kielisiä esseevastauksia. Arviointi perustuu latentin semanttisen analyysin käyttöön, joka on tiedonhakua varten kehitetty, sisällön vertailuun perustuva menetelmä. Suomen kielien erityispiirteiden ratkaisemiseksi järjestelmä käyttää apunaan suomen kielen morfologista jäsennintä.

Tutkielman tarkoituksena on etsiä vastausta seuraaviin kysymyksiin:

- Onko automaattinen esseiden arviointi todellinen vaihtoehto ihmisten suorittamalle arvioinnille?
- Minkä tyyppisten esseetehtävien arviointiin automaattista arviointia voidaan soveltaa?
- Mitkä ovat automaattisen arvioinnin keskeiset ongelmat ja kehityssuunnat?

Termillä *arvioija* tarkoitetaan tässä tutkielmassa nimenomaan ihmistä, joka arvioi esseetä. Puhuttaessa automaattisesta arvioinnista mainitaan joko *järjestelmä* tai *menetelmä*. Näin vältytään toistamasta hieman keinotekoista termiä ihmisarvioija.

2. ESSEETEHTÄVIEN ARVIOINTI

Esseetehtävät ovat yleisesti käytetty oppimisen arvioinnin menetelmä. Esseiden arviointiin liittyy monia erityispiirteitä. Se on huomattavasti vaativampi tehtävä kuin esimerkiksi vaihtoehtoisen arviointimenetelmän, monivalintatehtävien, pisteytys. Esseetehtäviä käytetään laajasti, koska niillä on kiistattomia etuja muihin tehtävätyyppeihin verrattuna. Esseiden arvioinnin automatisointiin liittyy seikkoja, jotka tulisi huomioida automaattista arviointimenetelmää kehitettäessä. Tutkijat eivät ole yksimielisiä esseiden arvioinnin automatisoinnin eduista. Jotkut näkevät sen avulla saavutettavan suuriakin hyötyjä kustannussäästöjen sekä arvioinnin puolueettomuuden ja yhdenmukaisuuden osalta.

2.1 Esseetehtävien luonne

Arviointi on erottamaton osa opetus-oppimis -prosessia (Hopkins et al. 1990). Arviointi on prosessi, jossa määritellään asioiden ansiokkuutta ja arvoa (Takala 1994). Se edellyttää laadun ulottuvuuksien ja tasojen määrittelyä. Arvioinnilla yleensä on tärkeitä vaikutuksia ja seurauksia. Se asettaa korkeat yhdenmukaisuuden ja tasapuolisuuden vaatimukset. Kaikkeen arviointiin sisältyy virhettä ainakin jossain määrin. Myös esseiden arviointi on siis virheille altis tehtävä.

Esseetehtävä voidaan määritellä koetehtäväksi, jonka yksittäistä vastausta ei voida määritellä ehdottomasti joko oikeaksi tai vääräksi (Thorndike 1971). Vastauksen tarkkuuden ja laadun voi määritellä henkilö, joka tuntee tehtävän käsittelemän aihealueen. Arvioijan muodostama käsitys esseen laadusta on tietenkin vain yhden henkilön subjektiivinen näkemys. Toinen merkittävä ominaisuus esseetehtävissä on niiden vastaajalle suoma vapaus. Esseetehtävät voivat olla laajuudeltaan hyvinkin erilaisia; toiset vaativat asian yksityiskohtaista tarkastelua, toisissa riittää lyhyehkö aiheen käsittely. Esseiden voidaan ajatella myös jakautuvan kaunokirjallisiin ja asiaesseihin. Kaunokirjallinen essee vaatii omaperäisempää tyylin hallintaa kuin asiaessee. Yhteistä esseelajeille on, että kirjoittaja keskustelee tekstin kanssa ja samalla itsensä ja lukijoiden kanssa.

Esseen sisällön ja kirjoitustyylin arvioiminen eroavat arvioinnin tavoitteiden osalta (Hopkins et al. 1990). Tyylin ja virheettömyyden painottaminen on varsin perusteltua missä tahansa

kirjoitustaidon testissä, esimerkiksi äidinkielen kokeessa. Jos esseen arvioinnin tavoitteena on tutkia esseenkirjoittajan tietämystä jollakin aihealueella, tulisi arvioinnissa keskittyä sisällön tarkasteluun. Tyyliin liittyvistä virheistä tulisi huomioida vain tekstin selkeyttä ja ymmärtämistä vaikeuttavat seikat, kuten lauserakenteisiin tai kappalejakoon liittyvät ongelmat. Arvioitaessa esseen sisältämää tietoa, ei esittämistyylille tulisi siis antaa liian suurta painoarvoa.

Esseevastauksiin voi sisältyä monen tyyppisiä virheitä (Foltz et al. 2000). Virheet voivat olla esim. oikeinkirjoituksen, kieliopin tai sisällön alueella. Kielioppiin ja oikeinkirjoitukseen liittyvät virheet ovat selvästi helpommin havaittavissa kuin sisältöön liittyvät ongelmat. Esseen sisältämä tieto voi olla puutteellista, virheellistä tai väärin ymmärrettyä.

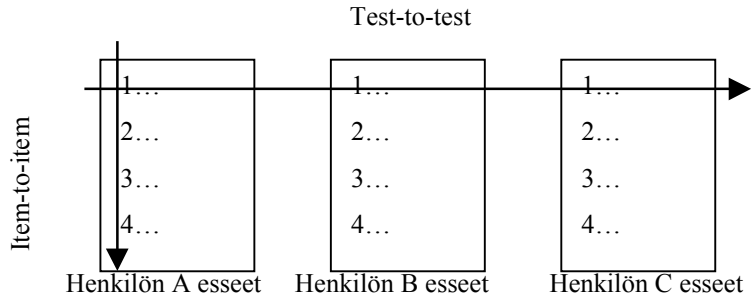
Arvioinnin yhdenmukaisuuden ja luotettavuuden korostaminen on erityisesti Yhdysvalloissa johtanut esseille vaihtoehtoisten, objektiivisempien testaustapojen tutkimiseen (Takala 1994, Hopkins 1990). Tästä on seurannut monivalinta- ja erilaisten täydennystehtävien käytön voimakas lisääntyminen. Niiden pisteytys on vakioitua, koska vastaus on aina yksiselitteisesti oikein tai väärin. Tällainen testaaminen johtaa kuitenkin helposti siihen, että pistemäärän luotettavuus muuttuu tärkeämmäksi kuin muut vaatimukset, kuten varsinainen oppimistulosten arvioiminen.

2.2 Ihminen esseen arvioijana

Esseiden arviointi asettaa arvioijalle monenlaisia vaatimuksia (Hopkins et al. 1990). Arvioijan on tunnettava tehtävän käsittelemä aihealue hyvin, eivätkä esimerkiksi väsymys, ennakkoasenteet esseen kirjoittajaa kohtaan tai muut seikat saisi vaikuttaa arvioihin. Koska arviointi on inhimillistä toimintaa, on selvää, että eri ihmisten arviot samasta esseestä eivät ole samanlaisia ja arvioinnissa tapahtuu myös virheitä. Myös sama arvioija antaa eri arviointikerroilla samalle esseelle usein eri arvosanoja. Tutkimuksissa on arvioijien antamien arvosanojen välisen korrelaation havaittu vaihtelevan koeasetteluista riippuen varsin laajalla vaihteluvälillä 0,35:sta aina 0,95:n korrelaatioihin (mm. Thorndike 1971, Hearst et al. 2000). Arvioijien välisen korrelaation voidaan todeta olevan tyypillisesti luokkaa 0,50…0,75 (mm. Page ja Petersen 1995, Landauer et al. 1997, Shermis et al. 2001, Hopkins et al. 1990). Arvioijien omien, eri ajankohtina samalle esseelle antamien arvosanojen korrelaatioiden on havaittu olevan keskimääriin 0,70 luokkaa tulosten vaihdellessa tässäkin lähteestä riippuen.

Eri arvioijat siis arvostavat esseitä eri tavoin ja arvioinnissa tapahtuu myös virheitä. Arviointivirheiden syitä voidaan jakaa eri ryhmiin (Rudner 1992, Hopkins et al. 1990). *Sädekehävaikutuksella* (halo effect) tarkoitetaan tilannetta, jossa arvioijan käsitys esseen kirjoittajan henkilökohtaisista ominaisuuksista vaikuttaa hänen käsitykseensä esseen laadusta. Vaikutus voi muuttaa arviota joko negatiiviseen tai positiiviseen suuntaan. *Stereotyypitys* (stereotyping) vaikuttaa siten, että arvioijan käsitykseen esseestä vaikuttaa hänen käsityksensä henkilöryhmästä johon esseen kirjoittaja kuuluu. Jos arvioijan tiedot arvioitavan esseen käsittelemästä aihealueesta ovat puutteelliset objektiivisen arvion tekemiseen, saattaa hän kompensoida tiedollisia puutteitaan antamalla systemaattisesti joko liian alhaisia tai korkeita arvosanoja. Arvioijat eroavat toisistaan myös arvoasteikon laajuuden käytön suhteen. Osa arvioijista välttää asteikon ääripäiden käyttöä ja sijoittaa suurimman osan esseistä keskiarvon tuntumaan.

Arvioitaessa useita esseitä peräkkäin on edeltäville esseille annettujen arvioiden todettu vaikuttavan myös seuraavana arvioitavien esseiden arvosanoihin (Hopkins et al. 1990). Tämä vaikutus on havaittu saman henkilön kirjoittamia esseitä peräkkäin arvioitaessa (item-to-item carryover) sekä eri henkilöiden samaan esseetehtävään antamia vastauksia arvioitaessa (test-to-test carryover). Kuva 1 havainnollistaa carryover-vaikutuksia.



Kuva 1. Carryover-vaikutukset. Edellisenä arvioitujen esseiden taso vaikuttaa arvioijiin. Vaikutus on havaittavissa sekä saman henkilön eri tehtäviin että eri henkilöiden samaan tehtävään antamien vastausten arvioinnissa.

Eräs syy arvosanojen vaihtelulle on se, että eri arvioijat painottavat arvioinnissa eri kriteereitä (Rudner 1992). Joihinkin arvioijiin esimerkiksi kirjoitusvirheet tai esseen rakenteeseen liittyvät seikat vaikuttavat enemmän kuin toisiin. Jopa kirjoittajan käsialan on havaittu vaikuttavan arvosanoihin käsin kirjoitettuja esseitä arvioitaessa. Arvosanojen yhdenmukaisuutta on todettu

voitavan parantaa merkittävästikin yhteisesti sovittujen arviointikriteerien ja koulutuksen avulla. Hyväkään koulutus ei kuitenkaan estä ihmisten erilaisista taustoista ja kokemuspiireistä sekä puhtaasti mielipide-eroista johtuvien arviointierojen syntymistä.

Arvioinnin avulla tuotettuja arvosanoja on käytetty myös esimerkiksi oppilaitosten vertailuun (Takala 1994). Vertailtaessa kouluja tai esimerkiksi tietyllä luokkatasolla olevien opiskelijoiden tason kehittymistä vuodesta toiseen ongelmaksi muodostuu arvioinnin yhdenmukaisuuden saavuttaminen (Page & Petersen 1995). Arvioijien vaihtuessa myös arviointiperusteet pyrkivät muuttumaan. Myös kullakin arviointikerralla arvioitavina olevien esseiden yleisen tason on todettu vaikuttavan arvioinnin lähtökohtiin. Tämä vaikeuttaa luotettavien vertailujen toteuttamista.

Edellä esitetyt seikat huomioiden yksittäistä ihmisen esseelle antamaa arvosanaa ei missään nimessä voida pitää ehdottoman objektiivisena ja virheettömänä arvioina esseen laadusta. Ihminen on altis ennakkoasenteilleen ja monille muille virhelähteille. Käyttämällä useampaa arvioijaa ja ennalta sovittuja arviointikriteerejä arvioinnin luotettavuutta voidaan parantaa merkittävästi (Rudner 1992). Useista arvioijista koostuvan paneelin käyttö ei kuitenkaan yleensä ole resurssisyistä mahdollista.

2.3 Esseiden arvioinnin automatisointi

Lähtökohtana esseiden arvioinnin automatisoinnille on se, että ihmiset määrittelevät esseiden todellisen laadun ja automaattisten järjestelmien tulee pyrkiä jäljittelemään ihmisten antamia arvosanoja (Page & Petersen 1995, Chung ja O'Neil 1997). Tietokone ei siis voi määritellä esseen laatua täysin itsenäisesti. Lisäksi yleisesti hyväksytään, että ihmisten antamat arvosanat muodostavat esseen "todellisen" arvon. Esseen arviointiin kehitetyn järjestelmän tarkoituksena on pyrkiä arvioimaan, minkä arvosanan useamman arvioijan ryhmä antaisi arvioitavalle esseelle. Edellisessä luvussa kuvatun kaltaiset virheet, kuten stereotyypitys tai sädekehävaikutus, joille yksittäinen arviointia suorittava ihminen on altis, eivät vaikuta suoraan automaattisiin arviointijärjestelmiin. Niiden toiminta perustuu siihen, että esseitä kerätään tilastollisin sekä *kieliteknologian* (natural language processing, NLP) menetelmin *pintasyntaktista* (surface features) ja asiasisältöön liittyvää tietoa. Pintasyntaksilla tarkoitetaan sisältöön kuulumattomia ominaisuuksia, joita ovat esim. lauseiden pituus, sanojen pituus ja kir-

joitusvirheiden määrä (Kaplan et al. 1998). Kun ihmisen arvioimista esseistä muodostuvan *valmennusmateriaalin* avulla on luotu tilastollisia menetelmiä käyttäen tehtäväkohtainen arviointimalli, voidaan sitä käyttäen arvioida muita samaan esseetehtävään annettuja vastauksia.

Nykyiset automaattiset menetelmät arvioivat esseetä *holistisesti* (mm. Burstein ja Chodorow 2002). Holistisessa arvioinnissa esseestä muodostetaan kokonaisvaltainen käsitys tutkimatta arvioon vaikuttaneita yksityiskohtia. Arviointi rajoittuu siis lähinnä yleisvaikutelman luomiseen ja yksittäisen arvosanan antamiseen. Uusimissa tutkimuksissa on kehitetty järjestelmiä, jotka kykenevät suorittamaan *analyyttistä* arviointia. Näiden menetelmien avulla esseen eri osa-alueita, kuten sisältöä, tyyliä ja rakennetta, voidaan tutkia toisistaan erillään ja hienojakoisemmin. Tämä mahdollistaa myös yksityiskohtaisemman palautteen antamisen kirjoittajalle. Tarkemaan palautteen avulla kirjoittajan on helpompi tiedostaa esseestään osa-alueet, joihin hänen tulee kiinnittää enemmän huomiota (Shermis et al. 2002).

Analyyttisen arvioinnin toteuttaminen on myös ihmiselle holistista arviointia hankalampaa ja hitaampaa (Foltz et al. 2000). On hyvin yleistä, että esseevastauksista saatava palaute rajoittuu holistisen arvioinnin tuloksena tuotettuun arvosanaan. Arvioija ei tavallisesti ehdi kirjoittaa pitkiä oikeita vastauksia tai korjausehdotuksia. Palautteen antamiseen kykenevän automaattisen järjestelmän avulla esseen kirjoittaja voisi saada tarkasteltavakseen esimerkiksi mallivastauksen sekä tietoa siitä, mitä asioita omasta esseevastauksesta puuttui tai mitkä tiedot olivat virheellisiä.

Ihmisen käyttämä kieli on suhteiltaan ja merkityksiltään niin monimutkainen, että sen täydelliseen ymmärtämiseen retorisine ja kulttuurillisine konteksteineen sekä merkityksen muuntaminen raa'aksi informaatioksi ei nykyisellä tietämyksellä ole mahdollista (Herrington ja Moran 2001). Tietokone ei siis pysty lukemaan ja eikä ymmärtämään tekstiä. Luonnollisen kielen tulkintaan perustuvissa sovelluksissa, kuten esseiden automaattisessa arvioinissa, on siis keskityttävä lukuprosessin simuloinnin sijasta tarkastelemaan lukemisen tuloksia jossakin rajatussa yhteydessä. Esseen arvioinnin automatisoinnissa pyritään esseen sisällön täydellisen ymmärtämisen sijasta mallintamaan ihmisen suorittaman arviointiprosessin tuloksena syntyneitä arvosanoja sekä palautetta.

Kaplan et al. (1998) määrittelevät automaattiselle esseen arviointijärjestelmälle neljä hyväksyttävyyskriteeriä. Nämä ovat *tarkkuus*, *puolustettavuus*, *valmennettavuus* ja *kustannustehokkuus*. Jotta menetelmä olisi hyväksyttävissä, on kaikkien näiden kriteerien täytyttävä.

Ollakseen tarkka arviointimenetelmän täytyy olla kykenevä tuottamaan luotettavia arvosanoja (Kaplan et al. 1998). Vertailukohtana tarkkuudelle käytetään ihmisten samoille esseille antamia arvosanoja. Tarkkuuden vaatimus on ilmeinen ja onkin luonnollista, että alan tutkimuksessa on kiinnitetty tähän seikkaan erityistä huomiota.

Puolustettavuudella tarkoitetaan, että menetelmän avulla tuotettujen arvosanojen antoperusteet on pystyttävä jäljittämään (Kaplan et al. 1998). Arviointimenetelmän toimintatavan ja arvioinnin perusteiden tulee olla rationaalisesti selitettävissä ja perustua järkeviin ja koulutuksellisesti hyväksyttäviin periaatteisiin.

Toisaalta menetelmän tulee olla laadittu siten etteivät sen pisteytykseen käyttämät menetelmät ole niin läpinäkyviä, että esseitä laativat henkilöt voidaan valmentaa tuottamaan esseitä, joille järjestelmä antaa ansaittua parempia arvosanoja (Kaplan et al. 1998). Tätä tarkoitetaan valmennettavuudella. Esimerkkinä valmennettavuudesta voidaan käyttää esseen sisältämien sanojen lukumäärää, joka varhaisimmissa arviointijärjestelmissä oli yksi tärkeimmistä arvosanan ennustamiseen käytetyistä ominaisuuksista. Näin yksinkertaisia ominaisuuksia käytettäessä koetta tekevät henkilöt oletettavasti ryhtyisivät käyttämään tietoa hyväkseen kirjoittamalla pitkiä, asiasisällöltään vajaita esseitä.

Lisäksi järjestelmän tulee olla kustannustehokas (Kaplan et al. 1998). Yksi tärkeimmistä syistä automaattisten arviointimenetelmien kehittämiselle ja käytölle on pyrkimys kustannussäästöihin, joten syntyvät kustannukset eivät luonnollisestikaan saa ylittää ihmisten suorittaman arvioinnin aiheuttamia kustannuksia. Automaattisen arviointijärjestelmän kustannuksiin on laskettava arviointiprosessista suoranaisesti koostuvien kustannusten lisäksi järjestelmän käyttöönotosta ja ylläpidosta koituvat kulut.

Arvioinnin automatisoinnista koituviksi eduiksi on arvioitu kustannussäästöt, arvioinnin lisääntynyt objektiivisuus, palautteen nopeus sekä testiryhmien välisen vertailun tarkkuuden parantuminen arviointikriteerien yhdenmukaisuuden ja pysyvyyden lisääntyessä (Page ja Petersen 1995, Malmi 2002).

Kustannussäästöt tulevat luonnollisesti sitä ilmeisimmiksi, mitä suuremmasta arvioitavien esseiden määrästä on kyse. Jotta automaattinen arviointijärjestelmä voisi ylipäätään toimia, on sen käytettävissä oltava riittävän suuri valmennusmateriaali, jonka avulla järjestelmä kalibroidaan toimimaan kunkin esseetehtävän arvioinnissa. Automaattista arviointia ei siten yleensä ole järkevää käyttää pieniä vastausmääriä arvioitaessa.

On hyvin tavallista, että esseen arvioinnin suorittaa vain yksi tai kaksi henkilöä. Automaattisten arviointijärjestelmien käyttö mahdollistaa useamman arvioijan yhtäaikaisen käytön ilman, että kustannukset kasvavat moninkertaisiksi. Käytettäessä useampaa arvioijaa arvioinnin objektiivisuus paranee, koska yksittäisen arvioijan mahdollisten arviointivirheiden vaikutus arvosanaan pienenee (Hopkins et al. 1990).

Automatisoinnin mahdollistama arvioinnin nopeus on hyödyllinen ominaisuus, jota voidaan käyttää hyväksi esim. verkkokurssien yhteydessä erilaisissa harjoitustehtävissä (Hearst et al. 2000, Foltz et al. 2000). Opiskelija voi kirjoitettuaan esseen lähettää sen automaattisen järjestelmän arvioitavaksi ja saada välittömästi palautetta. Välittömän palautteen voidaan ajatella auttavan oppimisprosessia (Malmi 2002). Välitöntä palautetta voidaan hyödyntää myös siten, että opiskelija voi lähettää saamansa palautteen perusteella muokkaamansa vastauksen uudelleen arvioitavaksi. Näin opiskelija joutuu aidosti miettimään, mikä alkuperäisessä vastauksessa oli puutteellista.

Bursteinin ja Chodorowin (2002) mukaan automaattisen arvioinnin validiuteen liittyy kaksi erillistä osa-aluetta: tilastollinen ja käsitteellinen validiteetti. Tilastollinen validiteetti liittyy tarkkuusvaatimukseen. Järjestelmän ja arvioijien antamien arvosanojen välisen korrelaation mittaaminen kertoo järjestelmän kyvystä mallintaa ihmisten antamia arvosanoja. Burstein ja Chodorow näkevät validiuden mitaksi toisaalta sen, kuinka hyvin menetelmän käyttämät arviointimetodit vastaavat ihmisten käyttämiä arviointimenetelmiä. Vaikka kielen tulkintaan perustuvia lingvistisiä ominaisuuksia lisätään arviointijärjestelmään, tulee järjestelmän tarkkuuden pysyä ennallaan. Arviointijärjestelmien on pyrittävä säilyttämään tasapaino näiden validiteettikäsitteiden välillä.

2.5 Automaattisen arvioinnin kritiikki

Kykyä käyttää luonnollista kieltä pidetään yleisesti inhimillisen älykkyyden määritelevänä ominaispiirteenä (Hearst et al. 2000). Erityisesti kyvyn ilmaista ajatuksia kirjoittamalla ajatellaan olevan tämän ominaisuuden huipentuma. Ajatus tietokoneiden käyttämisestä niin vaativan ja yleisen käsityksen mukaan inhimillistä päättelykykyä ja kokemusta edellyttävän tehtävän suorittamisessa kuin esseiden arviointi on luonnollisesti aiheuttanut myös suuria epäilyksiä.

Ensimmäisten tutkimusten ilmestyessä 1960-luvulla oli kritiikki osittain varsin jyrkkää (Herrington ja Moran 2001, Wresch 1993). Kritiikki on kohdistunut erityisesti arvioinnin perusteisiin ja rakenteeseen (Hearst et al. 2000). Arvosteluun on pyritty vastaamaan kehittämällä entistä tarkempiin tuloksiin kykeneviä menetelmiä ja toisaalta pyrkimällä käyttämään yhä lähemmin ihmisten käyttämiä arviointiperusteita vastaavia arviointimenetelmiä (mm. Powers et al. 2000).

Page ja Petersen (1995) jakavat esseiden automaattiseen arviointiin kohdistuvan kritiikin kolmeen ryhmään: *humanistiseen*, *puolustelevaan* ja *rakenteeseen kohdistuvaan* kritiikkiin. Humanistisen kritiikin lähtökohtana on tietokoneiden kehittämisen alkuajoista asti viljelty ajatus siitä, että on olemassa tehtäviä joiden suorittamiseen tarvitaan inhimillistä taustaa sekä tietämystä, ja että tietokone pystyy suorittamaan vain tehtäviä, joita se on ohjelmoitu suorittamaan. Koska tämän käsityksen mukaan tietokoneelta puuttuu tarvittava tietämys, se ei siis voi ymmärtää tai arvostaa esseetä ja tästä syystä tietokoneella ei olisi mahdollista mitata esseestä samoja asioita, joita ihmiset niissä arvioivat. Tästä syystä automaattinen arviointi olisi hylättävä mahdottomana. Kumotakseen tämän perustelun Page ja Petersen (1995) esittivät vertauksen 1950-luvulla kehitettyyn kuuluisaan Turingin testiin, jossa ihminen yrittää päätellä onko hänen kanssaan kommunikoiva osapuoli tietokone vai toinen ihminen. Page ja Petersenin esittivät muunnelman Turingin testistä, jossa ihmisen tulee tunnistaa kuuden ihmisen ja yhden automaattisen järjestelmän antamia arvosanoja tutkimalla, mikä arvosanajoukko on tietokoneen antama. Tämä on Pagen ja Petersenin mukaan todettavissa helposti, koska tietokoneen antamat arvosanat erottuvat selkeästi siten, että ne ovat lähempänä kaikkien arvioijien arvosanojen keskiarvoa kuin ihmisten antamat arvosanat. Tietokoneen antamat arvosanat erottuvat siis, koska ne ovat ”tarkempia”. Tämä tosiasia jättää Pagen ja Petersenin mukaan humanistisen vastustuksen varsin heikolle pohjalle.

Puolusteleva vastustus lähtee ajatuksesta että automaattisen arvioinnin menetelmiä tuntevan kirjoittajan olisi mahdollista huijata järjestelmää antamaan ansaittua parempia arvosanoja. Page ja Petersen (1995) myöntävät tämän jonkinasteiseksi ongelmaksi nykyisten järjestelmien osalta. Kysymys on eräs automaattisen arvioinnin keskeisimpiä tutkimuskohteita tällä hetkellä. Eräs ratkaisu ongelmaan on erikoisten ja jollakin tavalla ”epäilyttävien” vastausten tunnistaminen ja merkitseminen. Automaattinen järjestelmä ei pyrikään arvioimaan tällaisia esseitä, vaan jäättä ne ihmisen arvioitavaksi. Pagen ja Petersenin mukaan nykyisillä menetelmillä ihmisen käyttäminen rinnan automaattisen menetelmän kanssa on ainoa toimiva ja luotettava ratkaisu arvioinnissa, johon liittyy korkea tarkkuusvaatimus. Osa alan tutkijoista, mm. Powers et al. (2000) pitävät täysin automaattisen, ilman ihmisarvioijien tukea tapahtuvan arvioinnin toteutumista varsin epätodennäköisenä lähitulevaisuudessa.

Rakenteeseen kohdistuva kritiikki painottaa arvioinnin suorittamiseen liittyvää rakennetta (Page ja Petersen 1995). Tietokoneella olisi kyettävä mittaamaan esseestä samoja ominaisuuksia kuin ihmiset niistä mittaavat. Lisäksi arviointimenetelmien tulisi vastata ihmisten käyttämiä menetelmiä. Nämä seikat nousivat keskeisiksi kysymyksiksi erityisesti automaattisen arvioinnin ensimmäisten sovellusten osalta, jotka mittasivat esseistä varsin yksinkertaisia tunnuslukuja, kuten esseen pituus tai pilkkujen määrä. Tutkimuksen edetessä nykyisiä järjestelmiä on pyritty kehittämään entistä enemmän siten, että käytetyt menetelmät vastaisivat ihmisten käyttämiä arviointimenetelmiä (mm. Burstein ja Chodorow 2002).

Jotkut kritisoijat ovat epäilleet automaattisten järjestelmien kykyä tunnistaa ja arvostaa luovia ja omaperäisiä esseitä. Herrington ja Moran (2001) ovat esittäneet, että automaattisten järjestelmien käyttö ja erityisesti niiden antama palaute johtaa siihen, että opiskelijat kirjoittavat mielikuvituksettomia ja yksitoikkoisia esseitä. Heidän mukaansa se, että kirjoittaja suuntaa tekstinsä tietokoneen eikä ihmisen luettavaksi, vaikuttaa tuotettuun tekstiin.

Varsin hyvistä tuloksista huolimatta yleiseen hyväksyttävyyteen ja menetelmien laajaan käyttöönottoon on vielä matkaa, vaikkakin automaattisen esseenarvioinnin kaupallisia sovelluksia on myös tutkimuksen ulkopuolisessa, käytännönläheisessä käytössä. Esimerkiksi maailman suurin yksityinen arviointipalveluita myyvä organisaatio, Educational Testing Service, kehittää ja käyttää automaattista esseenarviointijärjestelmää. Järjestelmä otettiin käyttöön vuoden 1999 alussa (Burstein ja Chodorow 2002) ja vuoden 2002 huhtikuuhun mennessä sitä

oli käytetty toisena arvioijana ihmisen rinnalla kahden arvioijan raadissa yli miljoonan esseen arvioinnissa (ETS Technologies 2002).

3. PERUSMENETELMIEN KUVAUS

Automaattisen esseen arvioinnin menetelmiä on kehitety 1960-luvun puolivälistä saakka. Tässä tutkielmassa esitetyt menetelmät on valittu niillä saavutettujen hyvien tulosten vuoksi sekä siksi että ne käyttävät arvioinnin suorittamiseen toisiinsa nähden hieman erilaisia lähestymistapoja. Menetelmien ja laitteiden kehittyminen on mahdollistanut yhä parempien tulosten saavuttamisen. Kehitys on johtanut 1990-luvun lopulla ensimmäisten automaattisen esseen arvioinnin kaupallisten sovellusten julkaisuun ja käyttöönottoon.

3.1 PEG-järjestelmä

PEG (Project Essay Grade) -järjestelmän kehittäminen aloitettiin 1960-luvulla Ellis Batten Pagen johdolla (1966). Automaattisen esseiden arvioinnin tutkimuksen voidaan katsoa alkaneen näistä Pagen tutkimuksista. PEG:n kehitystyö jatkuu edelleen (mm. Shermis et al. 2001).

3.1.1 Idea

Menetelmän perustuu olettamukseen, että esseet sisältävät tiettyjä ominaispiirteitä, joiden perusteella ihmiset arvioivat niitä (Page ja Petersen 1995). Näitä ovat sisältö, tyyli, organisointi, luovuus ja oikeinkirjoitus (mechanics). Oikeinkirjoitukseen kuuluvat sanojen oikeinkirjoituksen lisäksi välimerkkien sekä suurten alkukirjaimien käytön kaltaiset ominaisuudet. Koska esseiden *todelliset ominaisuudet* (trins, intrinsic variables) eivät ole tietokoneella suoraan mitattavissa, on niitä kuvattava *arvioiden* (proxes, approximations) avulla.

Esimerkkejä esseiden todellisista piirteistä ovat sanavalintojen osuvuus, ilmaisutyyli, sujuvuus sekä lauserakenteen monimutkaisuus (Page 1966). Koska esimerkiksi ilmaisutyyli on käsitteenä sellainen, ettei sen suora mittaaminen tietokoneen avulla ole mahdollista, käytetään sen mittaamiseen yleisten ja harvinaisten sanojen suhdelukua. Tämä perustuu olettamukseen, että esseet joissa on käytetty suhteellisen vähän yleisimpiä sanoja, ovat ilmaisutyyliltään kehittyneitä. Käyttämällä apuna yleisimpien sanojen listaa voidaan tietokoneen avulla helposti muodostaa suhdeluku esseen sisältämille yleisimpien sanojen listalla ilmeneville sanoille ja muille sanoille. Lauserakenteiden monimutkaisuutta oletetaan voitavan arvioida esimerkiksi

prepositioiden tai alistuskonjunktioiden suhteellisen määrien avulla. Sujuvuuden arvioimiseen voidaan käyttää esseen sisältämien sanojen kokonaismäärää ja ilmaisutyylin mittaamiseen sanojen pituuksien vaihtelua (Page 1994). Jälkimmäinen mitoista perustuu olettamukseen, että harvinaiset sanat ovat yleensä usein esiintyviä sanoja pidempiä.

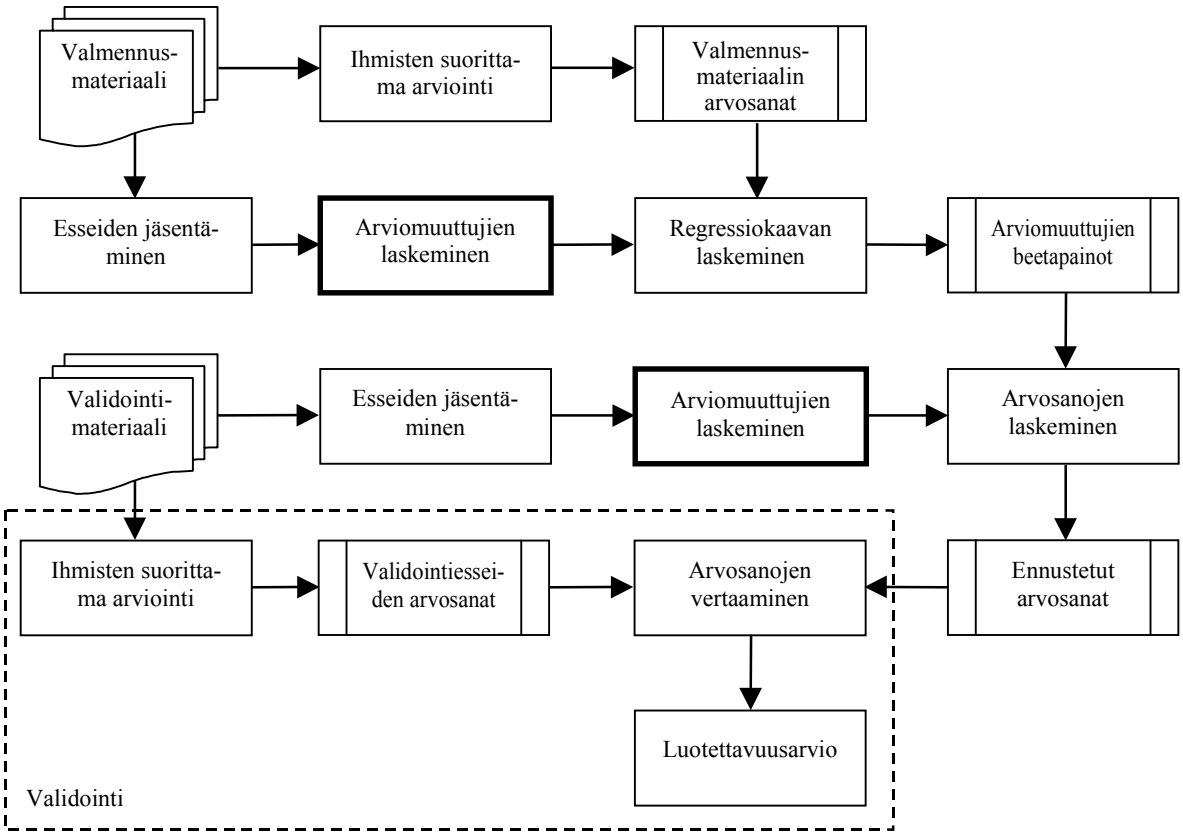
Eräs tärkeimmistä arvioita kuvaavista muuttujista on tutkimusten alkuvaiheista saakka ollut esseen pituus (Page ja Petersen 1995). Pituuden mittana on käytetty sekä sanojen lukumäärää että myöhemmin sanojen lukumäärän neljättä juurta. Sanojen määrän neljännen juuren käyttö perustuu havaintoon, että arvioijat vaativat yleisesti esseevastaukselta tiettyä vähimmäispituutta, mutta kun tämä raja on saavutettu, esseen pituuden kasvaminen edelleen ei vaikuta arvosanaan merkittävästi. Esseen pituuden käyttö arviointikriteerinä ei suinkaan tarkoita, että esseen pituus olisi välttämättä ihmisten käyttämä arviointiperuste. Useat esseen todellisista ominaisuuksista kuitenkin korreloivat esseen pituuden kanssa. Sama ilmiö pätee myös muiden arvioita kuvaavien muuttujien kohdalla. Vaikka käytetyt muuttujat varsinkin tutkimuksen alkuaikoina olivat melko yksinkertaisia, sisältyy niiden käyttöön se tärkeä ominaisuus, että ne mittaavat asioita jotka korreloivat useiden esseen laadun todellisten mittojen kanssa.

Pagen (1966) ensimmäisissä PEG:n versiossa käyttämät muuttujat olivat edellä kuvatun tyylisiä ja siis varsin yksinkertaisia. Tutkimuksen myötä muuttujat ovat kehittyneet, mutta niistä ei ole julkaistu yksityiskohtaisia tietoja. Page ja Petersen toteavat (1995) nykyisen sovelluksen mm. tutkivan lauseiden rakenteen eheyttä. Perusajatus ominaispiirteiden tutkimisesta arvioiden kautta on kuitenkin säilynyt muuttumattomana. Jatkuvana tavoitteena on ollut siirtyminen arvioista lähemmäs todellisten ominaisuuksien mittaamista (Page 1994).

3.1.2 Toteutus

PEG:n toimita perustuu pintasyntaktisten ominaisuuksien tutkimiseen, eikä luonnollisen kielen prosessointimenetelmiä käytetä. Kuva 2 esittää PEG:n toiminnan kaaviona (Chung ja O'Neil 1997). Ensimmäisessä vaiheessa arvioijat suorittavat esseiden arvioinnin. Elleivät esseevastaukset ole jo valmiiksi tietokoneen ymmärtämässä muodossa, ne täytyy muuntaa jollakin menetelmällä. Valmennusmateriaalin kustakin esseestä mitataan arviomuuttujia vastaavat tunnusluvut. Arviomuuttujien ja ihmisten antamien arvosanojen perustella suoritetaan *usean muuttujan regressioanalyysi* (multiple regression analysis), jossa selittävinä muuttujina

ovat esseistä mitatut arviota kuvaavat muuttujat ja selitettävänä muuttujana arvioijien antamien arvosanojen keskiarvo. Kun arvioitavasta esseestä mitataan samat arvot kuin valmennusmateriaalin esseistä ja sijoitetaan ne luotuun regressiomalliin, saadaan tulokseksi esseen arvosana. Osaa arvioitavista esseitä voidaan käyttää validointimateriaalina, jonka avulla voidaan mitata luotettavuusarvio järjestelmälle.



Kuva 2. PEG:n toiminta (Chung ja O'Neil 1997). Tavalliset nelikulmiot esittävät suoritettuja prosesseja. Nelikulmiot, joiden reunat on jaettu pienemmillä nelikulmioilla, esittävät niiden tuloksia. Katkoviivalla rajattu alue esittää tulosten validoinnin vaiheet.

Taulukossa 1 on esitetty PEG:n ensimmäisessä versiossa käytetyt arviomuuttujat ja niiden väliset korrelaatiot arvioijien arvosanaan sekä regressioanalyysissä muodostetut beetapainot (Page 1968). Taulukko osoittaa, että voimakkainta positiivista korrelaatiota arvosanan ja muuttujien välillä oli sanojen pituuksien keskihajonnalla ($r = 0.53$), sanojen keskipituudella ($r = 0.51$), pilkkujen määrällä ($r = 0.34$) sekä esseen sanamäärällä ($r = 0.32$). Nämä muuttujat toimivat siis regressiomallissa tärkeimpinä muuttujina arvosanojen määrittelyssä. Voimakkain negatiivinen korrelaatio oli yleisimpien sanojen määrällä ($r = -0.48$), heittomerkkien määrällä

(r = -0.23) sekä oikeinkirjoitusvirheillä (r = -0.21). Wreschin (1993) mukaan tämä on tulkittavissa siten, että arvioijat arvostavat esseessä kokonaisia, hyvin esitettyjä teemoja ja kehittynyttä sanaston käyttöä, mutta eivät lauserakenteiden monimutkaisuutta.

Taulukko 1. Arviomuuttujat ensimmäisessä PEG-järjestelmässä (Page 1968).

Muuttuja	**Korre-laatio**	**Beeta-paino**
Sanojen pituuksien keskihajonta	0.53	0.30
Sanojen keskipituus merkkeinä	0.51	0.12
Pilkkujen määrä	0.34	0.09
Pituus sanoina	0.32	0.32
Prepositioiden määrä	0.25	0.10
Yhdysmerkkien määrä	0.22	0.10
Sidesanojen määrä	0.18	-0.02
Tavuviivojen määrä	0.18	0.07
Lainausmerkkien määrä	0.11	0.04
Relatiivipronominien määrä	0.11	0.11
Puolipisteiden määrä	0.08	0.06
Kappaleiden määrä	0.06	-0.11
Otsikon läsnäolo	0.04	0.09
Lauseiden keskipituus	0.04	-0.13
Sulkujen määrä	0.04	-0.01
Kaksoispisteiden määrä	0.02	-0.03
Alleviivattujen sanojen määrä	0.01	0.00
Lauseen loppumerkkien määrä	-0.01	-0.08
Pisteiden määrä	-0.05	-0.05
Huutomerkkien määrä	-0.05	0.09
Kauttaviivojen määrä	-0.07	-0.02
Lauseiden pituuksien keskihajonta	-0.07	0.03
Alistuskonjunktioiden määrä	-0.12	0.06
Kysymysmerkkien määrä	-0.14	0.01
Subjekti-verbi yhdistelmällä alkavien lauseiden määrä	-0.16	-0.01
Oikeinkirjoitusvirheiden määrä	-0.21	-0.13
Heittomerkkien määrä	-0.23	-0.06
1000 yleisimmän englannin kielisen sanan määrä	-0.48	-0.07

Uusimmat PEG:n versiot keräävät tietoa yli 200 muuttujasta, mutta niiden yksityiskohtia ei ole julkaistu (Shermis et al. 2002). Uusia arviomuuttujia ovat mm. toistuvien lauseiden määrä, kirjoitusvirheiden määrä, *minä*-sanojen lukumäärä sekä yksinkertaisten lauseiden määrä.

Kun arviota kuvaavat muuttujat ja niitä vastaavat beetapainot ovat selvillä, voidaan arvioitavalle esseelle laskea arvosana (kuten kuvassa 2 on esitetty) käyttämällä regressioanalyysiä. Arvioijien antamien arvosanojen keskiarvolle saadaan ennuste käyttämällä usean muuttujan regressiota kaavasta (Page 1994):

$$J = a + \sum_{i=1}^{k} b_i P_i \quad (1)$$

missä a on vakio ja P_1, P_2,…, P_k ovat arviota kuvaavia muuttujia ja b_1, b_2, …, b_k niitä vastaavat regressiokertoimet.

3.2 Naiivi Bayes -tekstinluokittelumenetelmät

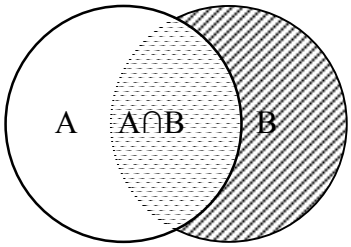
Naiivi Bayes -luokitteluun perustuvaa esseiden arviointia ovat tutkineet mm. Larkey (1998) sekä Rudner ja Liang (2002). TCT-tekstinluokittelutekniikka (Text Categorization Technique, TCT) perustuu Larkeyn (1998) tutkimukseen, jossa tarkasteltiin erilaisten tekstinluokittelumenetelmien käyttöä esseiden arvioinnissa. Rudnerin tutkimusten tuloksena on syntynyt tutkimuskäyttöön vapaasti käytettävissä oleva sovellus, BETSY (Bayesian essay testing system) (Rudner 2002).

3.2.1 Idea

Tekstinluokittelussa on yleisesti käytössä kaksi Naiivi Bayes-mallia: *usean muuttujan Bernoulli-malli* (multivariate Bernoulli model, binary independence model, BIM) ja *multinominaalinen malli* (multinominal model) (McCallum ja Nigam 1998), joita myös tässä esiteltävät Larkeyn (1998) sekä Rudenrin ja Liangin (2002) menetelmät soveltavat. Naiivi Bayes -mallien pohjautuvat *Bayesin teoreemaan* (Bayes' theorem) (Bayes 1764) ja *Naiivi Bayes -oletukseen* (Naive Bayes assumption). Bayesin teoreema saadaan kaavasta 2:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B \mid A)}{P(B)} \quad (2)$$

missä, esitettynä esseiden luokittelun yhteydessä, *P(A|B)* on todennäköisyys, että essee kuuluu luokkaan *B* ja sisältää termin *A*. *P(B|A)* on todennäköisyys, että arvosanakategoriaan *A* kuuluva essee sisältää termin *B*. *P(A)* on todennäköisyys, että essee kuuluu arvosanaluokkaan *A* ja *P(B)* todennäköisyys, että essee sisältää termin *B*. Graafisesti Bayesin teoreema voidaan esittää kuvan 3 mukaisesti.



Kuva 3. Bayesin teoreema Venn-diagrammiesityksenä.

Naiivi Bayes -oletuksen mukaan kukin luokittimen piirre vaikuttaa luokitukseen toisista piirteistä riippumattomasti (McCallum ja Nigam 1998). Oletuksen nojalla todennäköisyys, että dokumentti joka sisältää termit $B_1 \ldots B_n$, kuuluu kategoriaan Aj saadaan kaavasta:

$$P(A_j \mid B_1 \ldots B_n) = \frac{P(A_j) * P(B_1 \mid A_j) * P(B_2 \mid A_j) * \ldots * P(B_n \mid A_j)}{P(B_1)} \quad (3)$$

missä $P(A_j)$ on todennäköisyys, että mikä tahansa kokoelman dokumenteista kuuluu kategoriaan A_j ja $P(B_1)$ että dokumentti sisältää termin B_1. $P(B_1|A_j) \ldots P(B_n|A_j)$ ovat todennäköisyyksiä, että kategoriaan A_j kuuluva dokumentti sisältää termin $B_1 \ldots B_n$.

Vaikka monissa reaalimaailman tilanteissa Naiivi Bayes -riippumattomuusoletus on virheellinen, luokitin toimii erittäin hyvin esimerkiksi tekstin luokittelussa (McCallum ja Nigam 1998). Toimivuutensa vuoksi Naiivi Bayes -malleja käytetään tekstinluokittelun lisäksi monilla muilla sovellusalueilla, esimerkiksi ns. *älykkäissä agenteissa*. Tunnetuin esimerkki Naiivi Bayes -malliin perustuvasta älykkäästä agentista lienee Microsoft Office -sovelluksen avustaja (Office Assistant) (Myllymäki ja Tirri 1998).

(1) Usean muuttujan Bernoulli -mallin ja (2) multinominaalisen mallin lähestymistavat eroavat toisistaan seuraavalla tavalla (McCallum ja Nigam 1998):

(1) Usean muuttujan Bernoulli -mallissa dokumentin sisältö kuvataan binääriarvoina, jotka osoittavat nollan tai ykkösen avulla sisältyykö tietty termi dokumenttiin (McCallum ja Nigam 1998). Laskettaessa todennäköisyyttä, että dokumentti kuuluu johonkin kategoriaan, kerrotaan kaikkien sanojen ilmenemistodennäköisyydet keskenään. Mukaan sisällytetään myös niiden sanojen todennäköisyydet, jotka eivät ilmene dokumentissa.

(2) Multinominaalisessa mallissa dokumentin sisältö kuvataan sanojen ilmentymäkerrat huomioiden. Dokumentin kuuluminen tiettyyn kategoriaan lasketaan huomioimalla vain dokumentissa ilmennevät sanat.

Bernoulli-mallissa dokumentin voidaan ajatella olevan *”ilmiö”* (event), joka koostuu sanojen ilmenemistä tai ilmenemättömyyttä kuvaavista ominaisuuksista. Multinominaalisessa mallissa sanat ovat ”ilmiöitä” ja dokumentti on näistä muodostuva kokoelma.

Larkeyn tutkimuksen (1998) tavoitteena oli verrata bayesilaista luokitinta kahteen muuhun tekstin luokittelutapaan. Rudner ja Liang (2002) vertasivat tutkimuksessaan multinominaalisen ja usean muuttujan Bernoulli-mallien toimivuutta esseiden arvioinnissa.

3.2.2 Toteutus

Rudnerin tutkimuksissa toteutettu BETSY-järjestelmä on saatavilla ilmaiseksi tutkimuskäyttöön. Larkeyn kokeissaan käyttämää järjestelmää ei ole julkaistu, eikä menetelmiä käyttävää kaupallista sovellusta ole ainakaan toistaiseksi olemassa. Larkeyn ja Rudnerin järjestelmät ovat niin toteutuksen kuin ideankin tasolla monilta osin varsin samankaltaisia.

3.2.2.1 TCT-tekstinluokittelutekniikka

TCT-tekstinluokittelutekniikka käyttää esseiden luokittelemiseen kolmea menetelmää: (1) usean muuttujan Bernoulli -malliin perustuvaa luokitinta, (2) k:n lähimmän naapurin luokitinta ja (3) tekstin monimutkaisuusominaisuuksia (Larkey 1998).

(1) Larkeyn käyttämä Bayesin riippumattomuusluokitin vastaa Lewisin (1992) kehittämää usean muuttujan Bernoulli -malliin pohjautuvaa binääristä luokittelumenetelmää. Naiivi Bayes -mallin mukaisesti esseiden arviointi suoritetaan tutkimalla sen sisältämiä termejä ja laskemalla tämän perusteella, mihin arvosanaluokkaan se todennäköisimmin kuuluu. Termeillä TCT:ssä tarkoitetaan yksittäisiä sanoja. Luokittelutermien valinta suoritetaan kunkin arvosanakategorian osalta itsenäisesti etsimällä niitä parhaiten kuvaavat termit. Sanat, jotka

ilmenevät vähintään kolmessa valmennusmateriaalin esseessä, ovat ehdokkaita valittaviksi termeiksi.

Termien valinnan ensimmäisessä vaiheessa tekstistä poistetaan *sulkusanalistalla* (stopword list) esiintyvät sanat (Larkey 1998). Sulkusanat ovat tekstissä yleisimmin esiintyviä sanoja ja siten luokittelun kannalta merkityksettömiä. TCT käyttää hieman yli 400 sanan kokoista sulkusanalistaa. Lisäksi kaikista sanoista etsitään *vartalot* (stem) poistamalla niistä päätteet.

Lopullisen termijoukon valinta luodusta ehdokasjoukosta suoritetaan kullekin binääriselle luokittelijalle itsenäisesti. Ensimmäisessä vaiheessa ehdokkaat asetetaan järjestykseen käyttäen *odotetun yhteisinformaation mittaa* (expected mutual information measure, EMIM) (van Rijsbergen 1979). EMIM-arvo kuvaa kahden sanan välisen suhteen sisältämää informaation määrää. Sanojen *i* ja *j* välinen EMIM-arvo saadaan kaavan 4 mukaisesti.

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{4}$$

Kaavassa 4 $P(x_i, x_j)$ on todennäköisyys sille, että sanat x_i j x_j ilmenevät peräkkäin ja $P(x_i)P(x_j)$ sanojen x_i j x_j esiintymistodennäköisyydet. Optimaalinen termien määrä valitaan testaamalla valmennusmateriaalia eri kokoisilla termijoukoilla (Larkey 1998). Lopulliseksi termien määräksi valitaan se, jolla järjestelmän antamat arvosanat tuottavat korkeimman korrelaation arvioijien arvosanojen kanssa.

Kun luokittimessa käytettävät termit on valittu, lasketaan usean muuttujan Bernoulli -mallia käyttäen todennäköisyys sille, että arvioitava essee kuuluu kuhunkin arvosanaluokkaan (Larkey 1998). Esseen arvosana määräytyy sen arvosanaluokan mukaisesti, johon se kuuluu suurimmalla todennäköisyydellä.

(2) k:n lähimmän naapurin luokittimen avulla määritellään valmennusmateriaalista *k* kappaletta esseitä, jotka vastaavat lähimmin arvioitavaa esseevastausta. Arvioitavan esseen arvosana määräytyy näiden k:n esseen arvosanojen keskiarvon mukaisesti. Samankaltaisuus arvioitavan ja vertailuaineiston esseiden välillä määritellään *tf-idf-painotusta* (term frequency - inverse document frequency weighting) käyttävän Inquery-järjestelmän avulla (Callan et al. 1995).

Tf-idf -painotuksen tarkoituksena on dokumenttien välisen samankaltaisuuden tunnistamisen helpottaminen (Baeza-Yates ja Ribeiro-Neto 1999). Painotuksen avulla sanojen painoarvoja muutetaan siten, että tekstissä harvemmin esiintyvien sanojen merkitys kasvaa ja usein ilmenevien sanojen merkitys pienenee. Tf-idf -painotuksessa *tf* (term frequency) mittaa termien frekvenssiä dokumentissa ja *idf* (inverse term frequency) käänteistä dokumenttifrekvenssiä. Tf-idf painotus siis yhdistää sanan merkittävyyden tutkittavassa dokumentissa sekä sanan esiintymistiheyden koko aineistossa. Se voidaan laskea kaavan 5 mukaisesti.

$$w_{i,j} = \frac{freq_{i,j}}{max_l(freq_{l,j})} \times \log \frac{N}{n_i} \tag{5}$$

$W_{i,j}$ on sanan k_i tf-idf -painotettu arvo dokumentissa d_j. Lausekkeen alkuosa kuvaa sanan tf-arvoa ja loppuosa sanan idf-arvoa. $Freq_{i,j}$ on sanan k_i ilmentymien määrä dokumentissa d_j ja max_l dokumentissa d_j useimmin ilmenevän sanan frekvenssi. Idf -lausekkeessa N on dokumenttien yhteismäärä ja n_i niiden dokumenttien määrä, jossa sana k_i ilmenee.

(3) Esseestä tutkitaan TCT:ssä lisäksi yhtätoista tekstin monimutkaisuusominaisuutta (Larkey 1998). Nämä ovat merkkien kokonaismäärä, sanojen yhteismäärä, eri sanojen määrä, sanojen määrän neljäs juuri, lauseiden lukumäärä, sanojen keskipituus, lauseiden keskipituus sekä yli 5, 6, 7 ja 8-merkkisten sanojen lukumäärät.

Arvosanan muodostamiseksi Bayesin- ja k:n lähimmän naapurin luokittimista sekä tekstin monimutkaisuusominaisuuksista valitaan *askeltavaa lineaariregressiota* (stepwise linear regression) käyttäen ominaisuudet, jotka ovat arvosanan määrittelyssä merkitseviä (Larkey 1998). Arvioitavalle esseelle voidaan asettaa arvosana muodostettua regressiokaava käyttäen. Taulukossa 2 on esimerkki arvosanan muodostumisesta erään koeaineiston osalta. Sarake r kuvaa kunkin muuttujan korrelaatiota arvioijien antamiin arvosanoihin. Osatekijät-sarake esittää mitkä osat kustakin muuttujasta valikoituivat lineaariregressiossa mukaan luokitteluun. Bayes-muuttujat kuvaavat kutakin arvosana-asteikon kategoriaa kohti luotuja Naiivi Bayes -luokittimia. Sulkujen sisällä olevat luvut ilmaisevat Bayes-muuttujissa luokittimen perustaksi valittujen termien määrää ja k:n lähimmän naapurin kohdalla k:n arvoa luokittimessa. Bayes yhteensä -muuttuja on muodostettu valitsemalla regressioanalyysin avulla kaikista Bayes-

muuttujista merkitsevimmät. Kaikki yhteensä -muuttuja on muodostettu sisällyttämällä kaikki muuttujat regressioanalyysiin.

Taulukko 2. Esimerkki eri muuttujien vaikutuksesta arvosanan muodostumiseen (Larkey 1998).

Muuttuja	**r**	**Osatekijät**
Tekstin monimutkaisuus	0.86	Eri sanojen, lauseiden ja yli 6-merkkisten sanojen määrät
K:n lähim. Naapurin luokitin (220)	0.75	
Bayes 1 (300)	0.69	
Bayes 2 (320)	0.84	
Bayes 3 (300)	0.84	
Bayes 4 (280)	0.83	
Bayes 5 (380)	0.82	
Bayes 6 (600)	0.86	
Bayes yhteensä	0.86	Bayes 1, 2, 5 ja 6
Kaikki yhteensä	0.88	Bayes 1, 5, 6, yli 5 ja 6-merkkisten sanojen määrät, lauseiden määrä ja k:n lähimmän naapurin luokitin

Larkey (1998) totesi tutkimuksensa tuloksena, että Naiivi Bayes -luokitin on esseiden luokittelusta tekstin monimutkaisuusominaisuuksia ja k:n lähimmän naapurin luokitinta tarkempi. Joillakin testiaineistoilla, kuten taulukon 2 esimerkissä, kaikkien luokittelumenetelmien yhdistäminen osoittautui parhaaksi menetelmäksi. Lisäksi Larkey totesi, että TCT:ssa esseen pituudella ei yleensä ollut luokittelussa merkitystä. Hän olettaa sen johtuvan siitä, että Naiivi Bayes -luokitin tavoittaa esseistä tehtävän kannalta merkittävät sanat, eikä esseen kokonaispituudella siten ole merkitystä luokittelun kannalta. Sen, että esseen laatija on kirjoittanut pitkän esseen ilman aiheeseen liittyvää asiasisältöä, ei tietenkään tulekaan vaikuttaa arvosanan määräytymiseen.

3.2.2.2 BETSY

Rudner ja Liang (2002) vertasivat tutkimuksessaan Naiivi Bayes -mallien, usean muuttujan Bernoullin ja multinominaalisen mallin, toimivuutta esseiden luokittelussa. Heidän kehittämänsä järjestelmä on monilta osin hyvin samankaltainen kuin edellä esitelty TCT-

tekstinluokittelumenetelmä. BETSY-järjestelmä käyttää kolmen arvosanan asteikkoa, jossa essee luokitellaan joko *asianmukaiseksi* (appropriate), *osittain asianmukaiseksi* (partial) tai *epäasianmukaiseksi* (inappropriate) (Rudner ja Liang 2002).

Bayesilaisen luokittimen termeinä käytetään yksittäisten sanojen lisäksi *kahdesta sanasta muodostuvia fraaseja* (two-word phrases) sekä argumentteja (Rudner ja Liang 2002). Argumentilla tarkoitetaan tässä yhteydessä kahdesta sanasta koostuvaa järjestettyä paria, joista toinen ilmene tekstissä ennen toista. Esimerkiksi ekologiaa käsittelevässä esseessä argumentin voisi muodostaa sanapari myrkky / kala. Esseen tulkitaan sisältävän näistä sanoista muodostuvan argumentin mikäli siinä ilmenee sana myrkky ennen sanaa kala. Argumenttien valinta tapahtuu automaattisesti.

Termien valinta suoritetaan valmennusmateriaalin perusteella (Rudner ja Liang 2002). Se voidaan toteuttaa kahdella eri menetelmällä. Toisessa termien valinta perustuu niiden mahdollistaman *informaation lisäyksen* (information gain) mittaamiseen ja toisessa termien ilmentymien määrään valmennusmateriaalissa. Frekvenssiin perustuvassa mallissa termien ilmentymistiheydelle asetetaan tietty raja-arvo, jonka ylittävät termit valitaan luokittimeen. Informaation lisäys on *entropiaan* (entropy) (Shannon 1948) perustuva malli, joka avulla voidaan mitata luokittimen kunkin sanan kykyä erotella dokumentteja toisistaan (Yang ja Pedersen 1997). Tätä mallia sovellettaessa suurimman informaation lisäyksen mahdollistavat termit valitaan luokittimeen.

Rudner ja Liang (2002) testasivat Naiivi Bayes -mallien toimivuutta sekä ilman sulkulistaa että sulkulistan kanssa. Luokittelua testattiin sekä alkuperäisillä, päätteet sisältävillä sanamuodoilla että päätteettömillä sanavartaloilla. Usean muuttujan Bernoulli -mallin todettiin olevan multinominaalista mallia tarkempi. Parhaat tulokset Bernoulli-mallin mukaisessa luokitellussa saavutettiin, kun sulkulistaa tai sanojen perusmuotoistamista ei käytetty. Lisäksi argumentit osoittautuivat yksittäisiä termejä ja kaksisanaisia fraaseja tarkemmiksi luokittimiksi. Ilmenemistiheyteen perustuva ominaisuuksien valinta tuotti paremman tuloksen kuin informaation lisäykseen perustuva menetelmä.

3.3 Latentti semanttinen analyysi

Latentti semanttinen analyysi (Latent Semantic Analysis, LSA) kehitettiin tekstin indeksointia ja tiedon hakua varten (Deerwester et al. 1990). Myöhemmin sitä on sovellettu myös esseetehtävien automaattiseen arvioimiseen mm. Thomas Landauerin johdolla kehitetyssä Intelligent Essay Assessor -järjestelmässä (mm. Hearst el al. 2000).

3.3.1 Idea

Latentti semanttinen analyysi on menetelmä, jonka avulla pyritään tutkimaan dokumentin sisältöä esittämällä se vektorikuvauksen avulla (Landauer et al. 1998a). Vektoriesityksessä kuvataan dokumentissa ilmenevät sanat ja sanojen esiintymiskerrat. Käyttämällä matriisilaskennan menetelmiä tutkittavista dokumenteista luodaan matemaattinen kuvaus, jota kutsutaan *semanttiseksi avaruudeksi* (semantic space). Vertaamalla dokumentista luotua kuvausta muihin semanttisen avaruuden dokumentteihin, voidaan päätellä mikä tai mitkä niistä vastaavat sisällöltään lähimmin toisiaan. Käytetty menetelmä mahdollistaa samaa aihetta käsittelevien dokumenttien tunnistamisen samankaltaisiksi, vaikka ne sisältäisivät hyvin vähän tai ei lainkaan samoja sanoja. Menetelmän kykyä tunnistaa dokumenttien samankaltaisuuksia voidaan soveltaa myös esseiden automaattisessa arvioinnissa.

Ajatus siitä, että dokumentin sisältö on esitettävissä matemaattisessa muodossa perustuu olettamukseen, että kukin sana on osa semanttista avaruutta, jossa sanan merkitys määräytyy sen suhteesta kaikkiin muihin semanttisen avaruuden sisältämiin sanoihin (Landauer ja Psotka 2000). Latentti semanttinen analyysi pyrkii nimensä mukaisesti selvittämään nämä latentit, eli piilevät sanojen väliset suhteet. Oletuksena on, että yhtäläisyydet ja erot sanojen merkitysten välillä voidaan selvittää tutkimalla kontekseja, joissa sanat ilmenevät tai eivät ilmene. Toisaalta tekstin osien yhtäläisyydet on pääteltävissä niiden sisältämien sanojen kombinaatioiden perusteella. Kunkin sanan merkitys siis määräytyy ikään kuin kaikkien niiden tekstinosien merkitysten keskiarvona joihin se sisältyy, samalla ottaen huomioon myös mihin tekstin osiin kyseinen sana ei sisälly. Tekstinosan merkitys taas määräytyy kaikkien sen sisältämien sanojen merkitysten keskiarvona.

Keskeinen osa LSA:n toimivuudessa on *dimension reduktiolla* (dimension reduction), jonka avulla sanojen ja tekstinosien välisiä suhteita kuvaavia parametrejä matriisiesityksessä vähennetään (Landauer et al. 1998a). Tämä saa aikaan syvempien suhteiden löytymisen sanojen frekvenssiluvuista. Deerwester et al. (1990) esittävät, että dimension reduktiovaiheen teho perustuu sen kykyyn vähentää asiaankuulumatonta informaatiota sekä *hälyä* (noise) ja mahdollistaa sanojen välisten piilevien suhteiden paljastumisen. Dimensioiden vähentäminen muuttaa dokumentin esitystä siten, etteivät yksittäiset sanat ole edustettuina itsenäisinä, vaan ne ilmenevät jatkuvina arvoina jäljelle jääneissä dimensioissa (Foltz et al. 1996). Jos jotkin sanat ilmenevät samankaltaisissa konteksteissa, ovat niiden vektoriesitykset alennetun ulottuvuuksisessa esityksessä samankaltaiset. Tämä mahdollistaa sisällöltään samankaltaisten tekstinosien tunnistamisen, vaikka ne eivät sisältäisi samoja sanoja vaan esimerkiksi synonyymejä. Optimaalisen dimension löytäminen on erittäin tärkeää luotettavien tulosten saavuttamiseksi. Landauer et al. (1997) havaitsivat LSA:n tarkkuuden kolminkertaistuvan käyttämällä optimaalista dimensiota verrattuna redusoimattoman dokumentti-sana -matriisin käyttöön.

On siis selvää ettei LSA:n tuottama informaatio perustu pelkästään sanojen esiintymien tilastointiin. LSA:ta on sovellettu esseiden arvioinnin lisäksi monilla muilla alueilla pyrkimyksenä mallintaa ihmisen käsitteellisen tiedon ja kielen ymmärtämiseen liittyviä kognitiivisia prosesseja (Landauer et al. 1997, Landauer et al. 1998a, Foltz et al. 1998, Landauer 2002). Testaamalla LSA:ta mm. tiedonhaussa, lauseiden koherenssin tutkimisessa ja synonyymikokeeseen vastaamisessa, sen on todettu pystyvän tavoittamaan merkittävän osan sanojen, lauseiden ja tekstinosien merkityksestä. Vaikka LSA ei siis ole lähellekään täydellinen kielen tai kognition teoria, se on kuitenkin osoittautunut menetelmäksi, jonka avulla voidaan menestyksekkäästi mallintaa erilaisia inhimillisen kielellisen ymmärtämisen osa-alueita (Landauer 2002).

Huomionarvoinen seikka LSA:n toiminnassa on, että sanojen järjestys tekstissä jää sen avulla huomioimatta (Landauer ja Psotka 2000). Edellä esitetty olettamus tekstinosien merkityksen määräytymisestä sen sisältämien sanojen kombinaatioina kuitenkin tarkoittaa, että sanavalinnat sekä sanojen yhdistelmien muodostuminen ilmauksiksi ovat hallitsevia ominaisuuksia kielellisessä merkityksen määräytymisessä. Tästä johtuen tekstinosien merkitysten vertailu on useimmissa tapauksissa toteutettavissa luotettavasti ilman että sanajärjestys huomioidaan.

Sanajärjestyksen merkitystä ihmisen tekstin ymmärtämisessä ei tunneta riittävästi. Landauer et al. (1997) tutkivat sanajärjestyksen vaikutusta sovellettaessa LSA:ta esseiden arviointiin.

Tutkimuksessa todettiin että LSA:n antamat arvosanat olivat yhtä tarkkoja kuin arvioijien arvosanat ja että ainakin tietyissä tehtävissä, kuten esseen arviointi, voidaan saavuttaa hyviä tuloksia vaikka sanajärjestystä ei huomioida.

3.3.2 Toteutus

Käytännössä sanojen välisten piilevien suhteiden selvittäminen sanojen frekvenssiluvuista johtamalla tapahtuu LSA:ssa soveltamalla *singulaariarvohajotelmaksi* (Singular Value Decomposition) kutsuttua matriisilaskennan menetelmää (Landauer et al. 1998a). Menetelmän ensimmäisessä vaiheessa esseiden sisältämistä sanoista muodostetaan matriisiesitys, ns. *dokumentti-sana -matriisi* (word-by-document matrix). Matriisin kukin solu esittää tietyn sanan ilmentymien lukumäärän tietyssä *kontekstissa* (Landauer et al. 1998a). Konteksti voi olla esimerkiksi tekstikappale, lause tai kokonainen dokumentti. Matriisin sisällytetään vain sanat, jotka ilmenevät vähintään kahdessa eri kontekstissa. Esitykseen ei sisällytetä tiettyjä *sulkusanoja* (stopword). Ne ovat tekstissä yleisimmin ilmeneviä sanoja ja täten tiedonhaun ja luokittelun kannalta merkityksettömiä. Sulkusanat on määritelty ennalta sulkusanalistalle. Dokumentti-sana -matriisia muodostettaessa ei myöskään huomioida numeroita tai välimerkkejä, vaan ne jätetään matriisista pois. Osassa LSA:n sovelluksista dokumentti-sana -matriisi muodostetaan käyttäen sanojen vartaloita, osassa taivutettuja sanamuotoja. Sulkusanalistan käyttö, numeraalien poisto sekä sanojen perusmuotoistaminen sekä vaikuttavat menetelmällä saavutettaviin tuloksiin että lisäävät tehokkuutta vähentämällä käsiteltävän datan määrää. Kuvassa 4 on esitetty esimerkki dokumentti-sana -matriisista, joka on muodostettu ihmisen ja koneen vuorovaikutusta sekä verkkoja käsittelevän muistion otsikosta.

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement
m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

{*X* }=

	c1	c2	c3	c4	c5	m1	m2	m3	m4
Human	1	0	0	1	0	0	0	0	0
Interface	1	0	1	0	0	0	0	0	0
Computer	1	1	0	0	0	0	0	0	0
User	0	1	1	0	1	0	0	0	0
System	0	1	1	2	0	0	0	0	0
Response	0	1	0	0	1	0	0	0	0
Time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
Survey	0	1	0	0	0	0	0	0	1
Trees	0	0	0	0	0	1	1	1	0
Graph	0	0	0	0	0	0	1	1	1
Minors	0	0	0	0	0	0	0	1	1

r (human.user) = -.38

r (human.minors) = -.29

Kuva 4. Yhdeksästä teknisen muistion otsikosta (*c1…c5, m1…m4*) muodostettu dokumentti-sana -matriisi (Deerwester 1990). Matriisiin sisällytetyt sanat on esitetty otsikoissa kursiivilla. Korrelaatiokertoimet kuvaavat sanaparien *human/user* sekä *human/minors* välisiä suhteita.

Ennen singulaariarvohajotelman muodostamista dokumentti-sana -matriisia esikäsitellään kaksiosaisella menetelmällä, jonka tarkoituksena on painottaa sanojen ilmentymäarvoja siten, että ne vastaavat paremmin sanojen merkitystä dokumentin sisällön suhteen (Landauer et al. 1998a). Prosessissa dokumentti-sana -matriisin solujen arvot muunnetaan sanojen frekvens-seistä sanojen merkitystä paremmin kuvaaviksi. Yleisimmin LSA:n sovelluksissa käytetään entropiaan perustuvia painotusmalleja. Eräs menetelmä on Landauerin et al. (1998b) esittämä kaavan 6 mukainen tapa. Olkoon *X* dokumentti-sana -matriisi, joka sisältää *m* sanaa ja *n* kon-tekstia. Tällöin solun *(i, j)* painoarvo *Mij* saadaan kaavan

$$M_{ij} = \frac{\log(freq_{ij} + 1)}{-\sum_{j=1}^{n}(p_{ij} * \log(p_{ij}))} \quad (6)$$

mukaisesti, missä $freq_{ij}$ on solun *(i, j)* arvo ja p_{ij} sanan suhteellinen frekvenssi dokumentti-sana -matriisissa *X*. Se saadaan kaavan 7 mukaisesti.

$$p_{ij} = \frac{freq_{ij}}{\sum_{l=1}^{n} freq_{il}} \quad (7)$$

Kaavassa 7 jakajan summalauseke on rivin *i* sanan ilmentymien yhteismäärä dokumentti-sana -matriisissa.

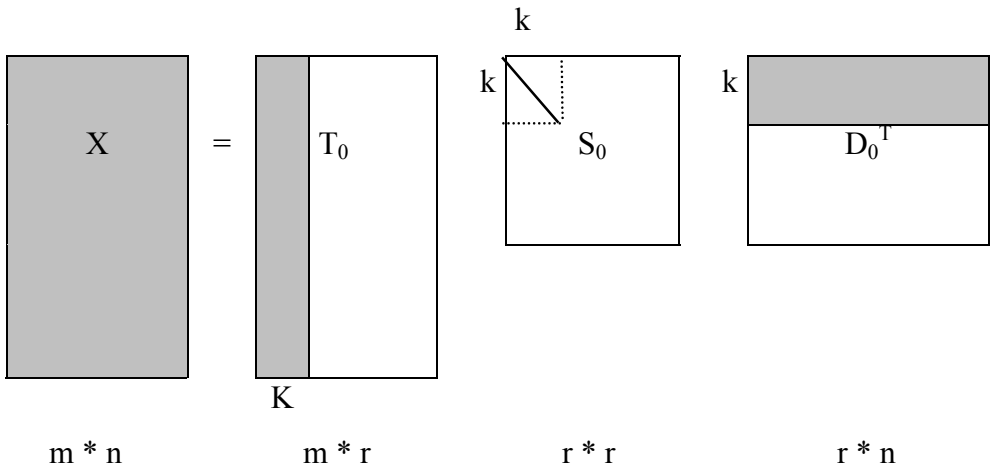
Landauer et al. (1998b) arvioivat painotuksen tärkeyden korostuvan tekstinkappaleen esittämisessä sanojen kombinaationa, koska se korostaa käsitellyllä aihealueella erityistä merkitystä sisältävien sanojen arvoa samalla pienentäen vähemmän tärkeiden sanojen merkitystä analyysissä. Kaavassa 6 jaettava arvo kuvaa sanan *paikallista painoa* (local weight) ja jakaja sanan *kokonaispainoarvoa* (global weight). Kokonaispainoarvon mittana käytetään sanan entropiaa, joka mitataan tutkimalla kaikkia konteksteja, joihin sana sisältyy. Merkityksen määräytymiseen vaikuttavat siis sanan tärkeys tutkitussa kontekstissa sekä sanan arvo koko dokumenttikokoelmassa.

Kun dokumentti-sana -matriisin solujen arvot on muunnettu edellä esitetyn painotuksen mukaisiksi, muodostetaan matriisista singulaariarvohajotelma (Landauer et al. 1997). Singulaariarvohajotelmassa matriisi jakautuu kolmen matriisin tuloksi kaavan:

$$X = T_0 S_0 D_0^T \quad (8)$$

mukaan, missä X on painotukset sisältävä matriisi ja T_0 ja D_0 ovat ortogonaalimatriiseja ja S_0 diagonaalimatriisi. T_0 sisältää sanojen ja D_0 kontekstien esittämiseen tarvittavat singulaarivektorit. S_0 esittää singulaariset skaalauskertoimet. Matriisien T_0, D_0^T ja S_0 kertominen keskenään tuottaa tulokseksi alkuperäisen dokumentti-sana -matriisin *X*.

Singulaariarvohajotelman muodostamisen jälkeisessä dimension reduktiovaiheessa diagonaalimatriisista S_0 poistetaan singulaariarvoja aloittaen pienimmästä arvosta (Landauer et al. 1998a). Matriisin S_0 arvojen vähentämisen jälkeen matriisien T_0, S_0 ja D_0^T kertominen keskenään tuottaa tulokseksi matriisin, joka sisältää likimääräisen arvion matriisin X sisällöstä. Dokumentti-sana -matriisin tapauksessa yksittäiset sanat korvaantuvat tällöin niitä kuvaavilla kertoimilla. Tämä mahdollistaa dokumenttien samankaltaisuuden toteamisen, vaikka ne sisältäisivät hyvin vähän yhteisiä sanoja (Deerwester et al. 1990). Sanojen frekvenssiluvut siis muuntuvat kuvauksiksi k-ulotteisessa avaruudessa. Kuvassa 5 singulaariarvohajotelman muodostaminen on esitetty graafisessa muodossa. Kuvassa m on termien, n dokumenttien ja k säilytettävien singulaariarvojen määrä sekä r dokumentti-sana -matriisin aste.



Kuva 5. Matriisin singulaariarvohajotelma graafisena esityksenä (Deerwester et al. 1990).

Dimensioiden määrän valinnassa on tavoitteena löytää taso, joka on riittävän korkea jotta dokumentti-sana -matriisin todellinen rakenne säilyy, mutta niin alhainen että irrelevantit yksityiskohdat ja ”melu” katoavat (Deerwester et al. 1990). Tavallisesti tiedonhakusovelluksissa suurilla, noin 20 000 - 60 000 sanan ja 1 000 - 70 000 dokumentin kokoisilla, dokumenttikokoelmilla dimensioita säilytetään 50 - 1500, yleisimmin noin 100 - 300 (Landauer et al. 1998b, Landauer ja Psotka 2000). Dimensioiden valinta on suoritettava kokeilemalla (Landauer et al. 1998a). Dimensioiden määrä on optimaalinen, kun löydetään taso, jolla saavutetaan paras luokittelutarkkuus johonkin ulkoiseen kriteeriin verrattaessa. Esseiden arvioinnin yhteydessä kriteerinä toimivat arvioijien arvosanat. Kuva 6 havainnollistaa dimensionreduktion vaikutusta sanojen ilmentymisfrekvenssit sisältävään matriisin. Kuvassa 6 on esitetty

kuvan 4 dokumentti-sana -matriisi singulaariarvohajotelman avulla muunnettuna. Singulaariarvohajotelma on muodostettu säilyttäen kaksi suurinta singulaariarvoa.

X{ }=

	c1	c2	c3	c4	c5	m1	m2	m3	m4
Human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
Interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
Computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
User	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
System	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
Response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
Time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
Survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
Trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
Graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
Minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

r (human.user) = 0.94

r (human.minors) = -0.83

Kuva 6. Kuvan 4 dokumentti-sana -matriisi kaksiulotteisen singulaariarvohajotelman muodostamisen jälkeen (Landauer et al 1998a). Sanaparien *human* ja *user* sekä *human* ja *minors* korrelaatiokertoimet havainnollistavat sanojen suhteiden muuntumista.

Kuvasta 6 on havaittavissa, että singulaariarvohajotelma on aiheuttanut matriisissa varsin merkittäviä muutoksia (Landauer et al 1998a). Näitä havainnollistavat korostettuna esitettyjen sanojen, *human*, *user* ja *minors* väliset korrelaatiot. Sana *human* ei ilmennyt alkuperäisessä kuvan 4 mukaisessa dokumentti-sana -matriisissa samassa kontekstissa kummankaan sanoista *user* tai *minors* kanssa. *Human/user* -sanaparin Spearman-korrelaatio on muuntunut kaksiulotteisen singulaariarvohajotelman avulla alkuperäisen matriisin -0.38:sta 0.94:n. Sanaparin *human/minors* välinen korrelaatio oli alkuperäisessä matriisissa -0.29, kun se singulaariarvohajotelman jälkeen on -0.84. Menetelmä on muuntanut sanojen arvoja perustuen tietoon, että sanat *human* ja *user* ilmenevät merkitykseltään samankaltaisissa konteksteissa. Sitä vastoin sanat *human* ja *minors* eivät esiinny merkitykseltään samankaltaisissa konteksteissa. Kontekstien, esimerkin tapauksessa siis muistioiden otsikoiden, kuvaukset muodostuvat vastaavalla tavalla, ei pelkästään sanojen frekvenssien vaan niiden ilmenemiskontekstien pohjalta.

Dokumenttien semanttista samankaltaisuutta voidaan verrata tutkimalla niistä muodostettuja vektoreita, siis niiden semanttisia kuvauksia. LSA:n kehittäjät ovat käyttäneet vertailuun useita eri menetelmiä, mm. dokumenttivektoreiden välistä pistetuloa ja vektorien euklidista etäisyyttä (Rehder et al. 1998). Aihealueen tietämyksen mittana käytetään lisäksi joissain LSA:n versioissa esseestä muodostetun matriisin dimensionreduktion jälkeistä pituutta (Landauer et al. 1998b). Yleisimmin käytetty menetelmä on dokumenttien vertailu vektorien välisen kulman kosiniarvon perusteella. Kun oletetaan, että *X* ja *Y* ovat *n*-ulotteisen avaruuden vektoreita

$$\mathbf{X} = (x_1, x_2, ..., x_n) \text{ ja } \mathbf{Y} = (y_1, y_2, ..., y_n) \tag{9}$$

niiden välinen kulma lasketaan kaavan 10 (Rehder et al. 1998) mukaisesti

$$\cos(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \bullet \mathbf{Y}}{\|\mathbf{X}\| * \|\mathbf{Y}\|} \tag{10}$$

ja || **X** || ja || **Y** || ovat vertailtavien vektoreiden pituudet, jotka lasketaan kaavasta:

$$\|\mathbf{X}\| = (\mathbf{X} \bullet \mathbf{X})^{1/2} = \left(\sum_{i=1..n} x_i^2\right)^{1/2} \tag{11}$$

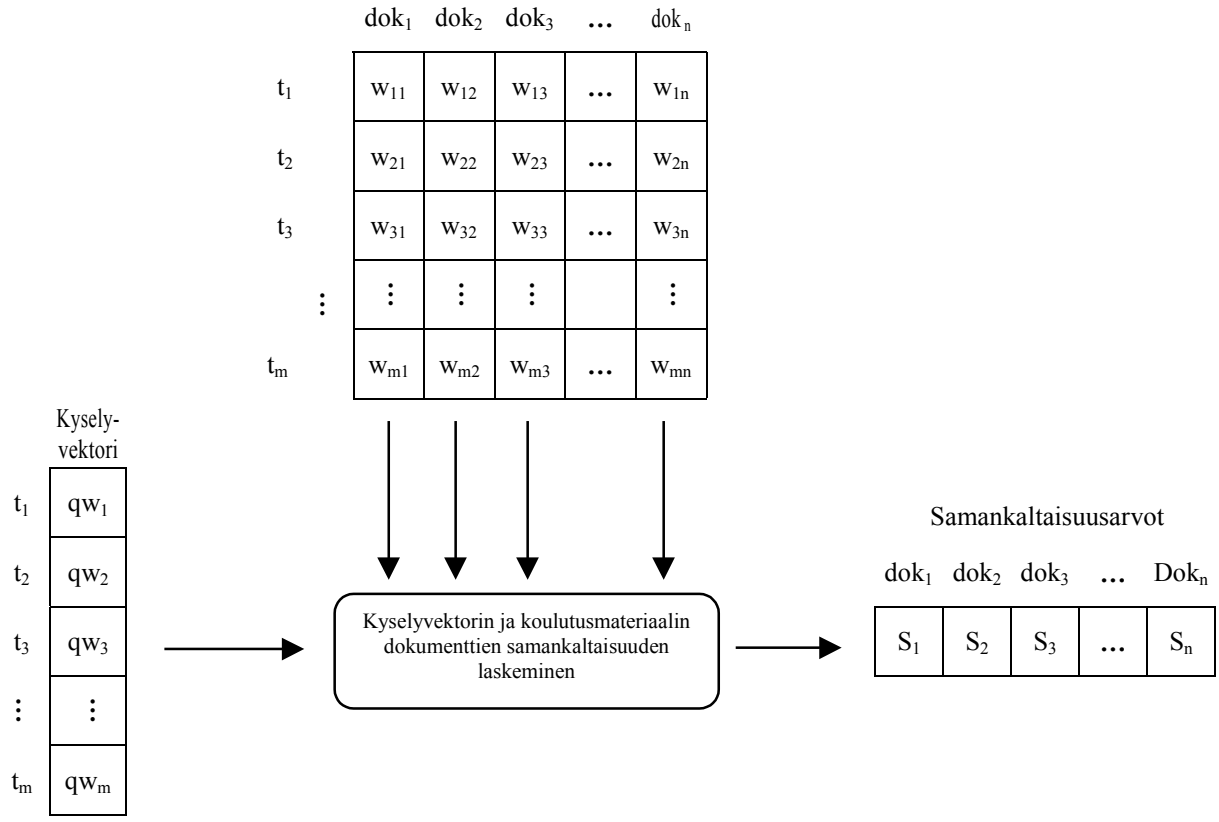
ja $u \bullet v$ pistetulo (sisätulo eli skalaaritulo), joka saadaan kaavasta:

$$\mathbf{X} \bullet \mathbf{Y} = x_1 y_1 + x_2 y_2 + ... + x_n y_n \tag{12}$$

Sovellettaessa LSA:ta esseiden automaattiseen arviointiin, esseen laatua voidaan tutkia vertaamalla sitä valmennusmateriaalin dokumentteihin edellä kuvatulla menetelmällä. Valmennusmateriaali voi sisältää sekä ihmisten arvioimia esseitä, asiantuntijan kirjoittamia esimerkkivastauksia että katkelmia oppikirjoista (Hearst et al. 2000). LSA:han perustuvassa Intelligent Essay Assessor -järjestelmässä optimaaliseksi valmennusmateriaalin kooksi arvioituja esseitä käytettäessä on osoittautunut noin 100 esseetä. Oppimateriaalin tai mallisvastausten käyttö koulutusmateriaalina ei tuota yhtä luotettavia tuloksia kuin arvioitujen esseiden käyttö. Se kuitenkin mahdollistaa LSA:n soveltamisen myös tilanteissa, joissa valmiiksi arvioituja esseitä ei ole saatavilla. Oppikirjan katkelmia voidaan käyttää esim. verkkokurssiin liittyvässä tehtävässä, jossa opiskelijat kirjoittavat esseen tai tiivistelmän jostakin kirjan kappaleesta.

LSA:n avulla opiskelijan vastausta voidaan verrata kirjan luvun sisältöön ja antaa arvosana samankaltaisuuden perusteella.

Kuva 7 esittää dokumenttien vertailun vaiheet (Chung ja O'Neil 1997). Kyselyvektori on arvioitavasta esseestä muodostettu vektori. Ennen kyselyn suorittamista kyselyvektorin arvot painotetaan käyttäen samaa painotusmenetelmää kuin valmennusmateriaalia muodostettaessa. Dokumentti-sana -matriisi sisältää valmennusmateriaalin esseistä muodostetut vastaavanlaiset vektoriesitykset (*dok$_1$…dok$_n$*). Vertaamalla kutakin matriisin sisältämää saraketta (eli valmennusmateriaalin yksittäisen dokumentin vektoria) kyselyvektoriin saadaan tulos, joka sisältää arvioitavan esseen ja kunkin valmennusmateriaalin dokumentin väliset samankaltaisuusarvot.

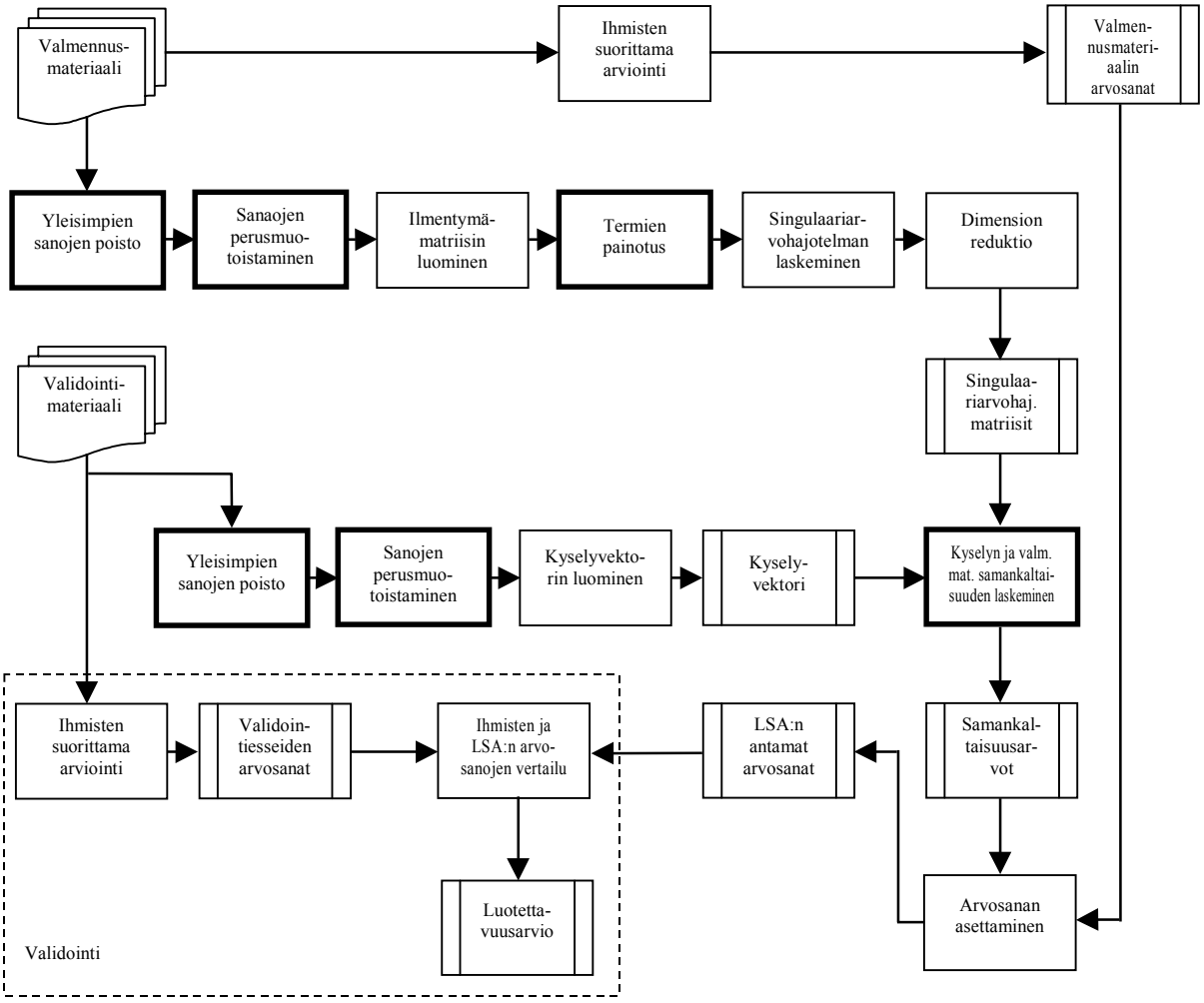


Kuva 7. Dokumentti-sana -matriisin ja kyselyvektorin samankaltaisuuden laskeminen. $W_{11}…w_{ij}$ ovat valmennusmateriaalin dokumenttien kutakin sanaa vastaavat arvot sekä $Qw_1…qw_m$ arvioitavasta esseestä muodostetun sisältövektorin kunkin sanan arvot (Chung ja O'Neil 1997).

Arvioitavan esseen arvosana siis määräytyy vertaamalla sitä valmennusmateriaaliin (Landauer et al. 1997). Yleisimmin käytetyssä menetelmässä esseen arvosanaksi määräytyy koulutus-

materiaalin *k*:n sitä lähimmin vastaavan esseen arvosanojen keskiarvo. *K*:n arvo voidaan asettaa vakioksi, esim. 10 tai se voidaan määritellä ottamalla huomioon kaikki tietyn samankaltaisuusarvon ylittävät esseet. Käytettäessä vertailuaineistona oppikirjan katkelmia, arvosana voidaan määrittää esimerkiksi vertaamalla arvioitavan esseen jokaista lausetta vertailuaineiston kuhunkin lauseeseen ja arvioijien valitsemiin tärkeimpiin lauseisiin (Foltz et al. 1996).

Kuva 8 esittää edellä kuvatun LSA:n perustuvan esseidenarviointijärjestelmän toiminnan kaaviona (Chung ja O'Neil 1997). Tulosten validointi tapahtuu tavallisesti vertaamalla arvioijien arvosanojen keskinäistä korrelaatiota järjestelmän ja arvioijien keskiarvosanan korrelaatioon.



Kuva 8. LSA:n toiminta. Tavalliset nelikulmiot esittävät suoritettuja prosesseja. Nelikulmiot, joiden reunat on jaettu pienemmillä nelikulmioilla, esittävät niiden tuloksia. Tulosten validoinnin vaiheet on rajattu katkoviivalla (Chung ja O'Neil 1997).

3.4 Elektroninen esseen arvioija (E-Rater)

Elektroninen esseen arvioija (Electronic Essay Rater, E-Rater) -järjestelmä on kehitetty Jill Bursteinin johdolla Educational Testing Service (ETS) -yhtiössä. 90-luvun lopulla alkaneiden tutkimusten tavoitteena on erityisesti ollut luoda järjestelmä, joka mittaa esseiden ominaisuuksia aiemmin kehitettyjä järjestelmiä suoremmin ja ihmisten käyttämiä menetelmiä lähemmin vastaavin menetelmin. ETS on käyttänyt E-Rateria yli miljoonan esseen arvioinnissa vuoden 2002 huhtikuuhun mennessä (ETS Technologies 2002).

3.4.1 Idea

E-Raterin perusajatus on, että automaattisen esseiden arviointijärjestelmän on pyrittävä arvioimaan esseistä samoja ominaisuuksia kuin ihmiset niistä arvioivat (Burstein et al. 1998b). Tästä syystä esim. esseen pituuden käyttö arvioinnin kriteerinä, kuten edellä esitellyssä PEG:ssa, ei tule kyseeseen. Arviointiperusteiden tiukka rajaaminen vain ihmisten käyttämiin perusteisiin estää esseen pituuden tyyppisten ominaisuuksien käytön E-Raterissa.

Menetelmän kehittämisen perustana on ollut Graduate Management Admissions Testin (GMAT) analyyttisen kirjoittamisen osion (Analytical Writing Assessment, AWA) arvioimiseksi laadittu ohjeisto (Burstein ja Marcu 2000b). GMAT-koe on kaupallisen alan jatko-opintoihin erityisesti Yhdysvalloissa valintaperusteena käytetty koe. GMAT:n AWA-osassa on kahden tyyppisiä esseetehtäviä, joista toisessa kirjoittajan tulee esittää mielipiteensä tehtävässä määrätystä asiasta ja toisessa analysoida annettua väittämää.

GMAT AWA -kokeen arviointiohjeiden pohjalta johdetut ominaisuudet jakautuvat lauseopillisen ja retorisen rakenteen arviointiin sekä aiheen käsittelyn analysointiin (Burstein et al. 1998a). Järjestelmä laskee osa-alueita kuvaaville muuttujille tunnuslukuja. Ihmisten arvioimien esseiden perusteella muodostetaan esseetehtäväkohtainen malli, jonka avulla järjestelmä suoritta arvioitavan esseen pisteytyksen.

3.4.2 Toteutus

E-Rater jakautuu myös toteutuksellisesti kolmeen pääosaan: lauseopillisen rakenteen, diskurssin sekä aiheen käsittelyn analysointiin (Burstein ja Marcu 2000b). Valmennusmateriaalin esseiden ominaisuuksista kerätyt tiedot yhdistetään malliksi, jota käytetään arvioitavien esseiden pisteyttämiseen. Optimaalinen valmennusmateriaali sisältää 270 kahden ihmisen arvioimaa esseetä, jotka jakautuvat tasaisesti eri arvosanakategorioiden kesken (Burstein et al. 1998a). E-Rater käyttää kuuden arvosanan asteikkoa, jolloin valmennusmateriaalin tulee sisältää 5 arvosanan nolla ja 15 arvosanan yksi saaneita esseitä. Lisäksi kutakin arvosanan kahdesta kuuteen saaneita esseitä tulee valmennusmateriaalissa olla 50 kappaletta.

3.4.2.1 Lauseopillinen rakenne ja monipuolisuus

Lauseopillisen rakenteen arvioimiseksi esseetä tutkitaan käyttäen *luonnollisen kielen käsittelymenetelmiä* (natural language processing, NLP). E-Rater käyttää tähän tarkoitukseen versiosta riippuen Abneyn (1996) kehittämää CASS-järjestelmää (Burstein ja Chodorow 1999) tai Microsoftin Natural Language Processing -välinettä (Burstein et al. 1998a). Esseen sisältämien sanojen sanaluokat tunnistetaan ja lauseet merkataan. Tuotettuja jäsennyspuuta tutkimalla voidaan arvioida esseen lauseopillista monipuolisuutta mm. käytettyjen verbi- ja lausetyyppien perusteella. Järjestelmä laskee predikatiivi-, sivu-, infinitiivi- ja relatiivilauseiden sekä modaalisten eli vaillinaisten apuverbien (kuten would, could, should ja may) määrät ja muodostaa näistä erilaisia suhdelukuja.

3.4.2.2 Retorinen rakenne ja argumenttien organisointi

Diskurssin analysoinnissa pyritään selvittämään esseen retorisia rakenteita ja paikantamaan siihen sisältyvät kokonaisuudet (Powers et al. 2000). Rakenteen selvittäminen mahdollistaa esseessä esitettyjen väitteiden löytämisen ja niiden jäsentelyn tutkimisen. Näiden tietojen avulla pyritään selvittämään, onko kirjoittaja pystynyt tuottamaan analyyttisen ja hyvin jäsennellyn kokonaisuuden. Esseen rakennetta tutkimalla pyritään löytämään sen sisältämät *argumentit*. Argumentilla tarkoitetaan järkevää esityskokonaisuutta, jonka tarkoituksena on perusteluja tai esimerkkejä hyväksi käyttäen vakuuttaa lukija esitetyn asian paikkansapitävyy-

destä. Argumenttien rakenne voi noudattaa kappalejakoa tai olla siitä riippumaton (Burstein et al. 1998b). Argumenttejä voidaan esseen rakenteen tutkimisen lisäksi hyödyntää esseen asia-sisällön tarkastelussa.

Argumenttien tunnistamiseksi tutkitaan tekstin osien suhteita (Burstein et al. 1998b). Niitä ovat esimerkiksi *rinnakkaisuus* (parallelism) ja *vastakohtaisuus* (contrast). Argumenttien ja niiden välisten suhteiden tunnistamiseksi tekstistä etsitään ns. *merkkisanat* (cue word). Nämä ovat termejä, jotka ilmaisevat argumenttien mahdollisia alkukohtia sekä argumenttien kehittelyä. Termit voivat olla yksi- tai useampisanaisia. Esimerkiksi termi ”lopuksi” (in summary, in conclusion), voidaan tulkita aloittavan yhteenvedon tai termin ”luultavasti” (perhaps) osoittavan kirjoittajan jatkavan edellisen argumentin kehittelyä.

3.4.2.3 Aiheen käsittely ja sanaston käyttö

Käytetty sanasto on eräs GMAT AWA:n arviointiohjeissa mainituista kriteereistä (Burstein ja Chodorow 1999). Hyvät esseet luonnollisesti käsittelevät annettua aihetta. Yleensä hyvin kirjoitetuissa esseissä käytetään myös enemmän tarkkaa ja tehtävän käsittelemän aihealueen erikoissanastoa kuin heikompitasoisissa esseissä. Voidaan siis olettaa, että hyvätasoiset esseet muistuttavat sanavalinnoiltaan toisia hyvätasoisia esseitä. Toisaalta heikkotasoiset esseet sisältävät samankaltaisia sanoja kuin toiset heikkotasoiset esseet. Esseen sisällön laatu voidaan siten määritellä vertaamalla esseessä käytettyä sanastoa valmennusmateriaalin esseiden sisältöön.

E-Rater tutkii käytettyä sanastoa *vektoriavaruusmalliin* (vector space model) (Salton 1989) perustuvan *sisältövektorianalyysin* (content vector analysis) avulla. Analyysi suoritetaan kahdella eri menetelmällä, joista toisessa tutkitaan (1) esseen sanastoa kokonaisuutena ja toisessa (2) kunkin argumentin osalta erikseen (Burstein ja Chodorow 1999).

Esseen sanaston analysointi aloitetaan muodostamalla sen sisällöstä vektoriesitys, jonka kukin elementti edustaa yksittäisen sanan esiintymismääriä esseessä (Burstein ja Chodorow 1999). Vektoria muodostettaessa jätetään pois sulkusanat. Sanojen frekvenssimatriisi muunnetaan korvaamalla kunkin solun arvo sen tf-idf -painolla (ks. kohta 3.2.2.1). Sanan *w* paino arvosanakategoriassa *j* määräytyy kaavan 13 mukaisesti.

$$weight_{wj} = \frac{freq_{wj}}{maxfreq_j} * \log \frac{nessays}{essays_w} \quad (13)$$

Kaavassa 13 $freq_{wj}$ on sanan frekvenssi esseessä j, $maxfreq_j$ on kategoriassa j useimmin esiintyvän sanan frekvenssi, *nessays* on valmennusmateriaalin esseiden lukumäärä ja $essays_w$ sellaisten valmennusmateriaalin esseiden määrä, jotka sisältävät sanan w (Burstein ja Chodorow 1999).

(1) Esseetä kokonaisuutena tarkastelevassa menetelmässä arvioitava essee ja kaikki valmennusmateriaaliin kuuluvat esseet muunnetaan edellä kuvatulla menetelmällä vektoreiksi (Burstein ja Chodorow 1999). Esseetä verrataan valmennusmateriaaliin vektorien välisten kosiniarvojen avulla. Arvosanaksi asetetaan kuuden arvioitavaa esseetä sisällöltään lähimmin vastaavan esseen arvosana kaavan:

$$score_x = \frac{\sum_{i=1}^{6} (\cos(X, Y_i) * score_{Y_i})}{\sum_{i=1}^{6} \cos(X, Y_i)} \quad (14)$$

mukaisesti, missä $score_{yi}$ on arvioijien antaman arvosanan keskiarvo esseelle y_i ja *cos(X,Yi)* on arvioitavan esseen X ja valmennusmateriaalin esseen Y_i sisältövektorien välinen kosini.

(2) Argumenttitasolla tapahtuvaa analyysiä varten kustakin arvosanakategoriasta muodostetaan yksi vektori, joka sisältää kaikkien siihen kuuluvien esseiden sisältämät sanat (Burstein ja Chodorow 1999). Esseen argumenteille, jotka on selvitetty kohdassa 3.4.2.2 kuvatulla menetelmällä, muodostetaan vektoriesitys kaavassa 12 esitettyä vastaavalla tavalla. Vertaamalla vektoriesityksiä voidaan tutkia arvosanakategorioiden ja arvioitavan esseen välisiä yhtäläisyyksiä. Kutakin argumenttia verrataan yksitellen valmennusmateriaalin arvosanakategorioittain muodostettuihin vektoreihin käyttäen mittana vektorien välistä kosinia.

Tutkiessaan argumenttitasolla annettuja arvosanoja Burstein et al. (1998a) havaitsivat, että esseen sisältämien argumenttien lukumäärä vaikuttaa arvioijien antamiin arvosanoihin siten, että vain muutaman argumentin sisältävälle esseet saavat tavallisesti argumenttien pistekeskiarvoa alhaisemman arvosanan. Vastaavasti essee, jossa on esitetty useita argumentteja, saa

argumenttien arvosanojen keskiarvoa hieman korkeamman arvosanan. E-Rater huomioi tämän siten, että argumenttikohtaisen analyysin tuloksista ei muodosteta arvosanaa suoraan niiden keskiarvon perusteella, vaan argumenttien lukumäärää huomioiden. Arvosana määräytyy kaavan 15 mukaisesti (Burstein ja Chodorow 1999):

$$score_t = \frac{\sum_{j=1}^{n} argscore_j + nargs_t}{nargs_t + 1} \tag{15}$$

missä $\sum_{j=1}^{n} argscore_j$ on esseen t sisältämien argumenttien arvosanojen summa ja $nargs_t$ on argumenttien määrä esseessä t.

3.4.2.4 Mallin muodostaminen ja esseen pisteytys

Syntaktisen rakenteen, retorisen rakenteen ja aiheen käsittelyn analysoinnin tuloksena esseille muodostetaan 57 ominaisuutta (Burstein et al. 1998a). Valmennusmateriaalista muodostetaan ominaisuuksille arvioitavan esseekysymyksen suhteen optimaaliset painotukset *askeltavaa lineaariregressiota* (stepwise linear regression) käyttäen. Laskettujen regressiokertoimien perusteella valitaan ennustavuudeltaan parhaat ominaisuudet. Yleensä niitä on 8 - 12 kappaletta (Powers et al. 2002). Taulukossa 3 on esitetty Bunstein et al. (1998a) kokeissa havaitsemat yleisimmin regressiomalliin valikoituneet ominaisuudet.

Taulukko 3. Useimmin käytetyt ominaisuudet 15 esseetehtävän joukossa (Burstein et al. 1998a)

Ominaisuus	Luokka	Käyttökerrat
Sisältö argumenttitasolla	Aihe / diskurssi	15/15
Sisältö koko esseen tasolla	Aihe	14/15
Argumentin kehittelysanojen lukumäärä	Diskurssi	14/15
Modaalisten apuverbien lukumäärä	Syntaksi	12/15
Argumenttien aloitus: predikatiivilauseet	Diskurssi	7/15
Argumentin kehittely: retoriset kysymyssanat	Diskurssi	6/15
Argumentin kehittely: merkkisanat (evidence words)	Diskurssi	6/15
Sivulauseet	Syntaksi	4/15
Relatiivilauseet	Syntaksi	4/15

Muodostettua mallia käytetään arviointivaiheessa suoraan arvosanojen asettamiseksi arvioitaville esseille. Kun arvioitavasta esseestä mitataan regressiomalliin valittuja ominaisuuksia vastaavat tunnusluvut ja sijoitetaan ne regressiokaavaan, tulokseksi saadaan esseen arvosana (Burstein ja Marcu 2000b).

4. Menetelmien vertailu

Edellä esitellyillä automaattisen esseenarvioinnin menettelemillä on saavutettu kokeissa varsin hyviä tuloksia. Esseiden arvioinnin automatisointiin liittyy myös yleisiä ongelmia, joita mikään nykyisistä järjestelmistä ei ratkaise. Olemassa olevilla menetelmilläkin saavutetaan kuitenkin tutkimustulosten valossa tarkkuus, joka on täysin vertailukelpoinen ihmisten suorittaman arvioinnin kanssa.

4.1 Mallit

Olen koonnut taulukkoon 4 tässä tutkielmassa esitellyt arviointimenetelmät Pagen (1966) esittämällä jaottelutavalla. Pagen mallissa esseiden automaattisen arvioinnin lähestymistavat jaetaan neljään kategoriaan. Malli jakaa menetelmät simuloiviin ja analysoiviin sekä toisaalta sisällön ja kirjoitustyylin arviointia painottaviin menetelmiin. Tyylillä Page tarkoittaa tapaa, jolla asiat on esitetty ja sisällöllä esitettyä asiasisältöä. Tyyliin kuuluvat mm. lauseopin oikeellisuus, oikeinkirjoitus sekä ilmaisutyyli. Sisältöön sitä vastoin kuuluu käytetty sanasto ja esitetyt argumentit.

Analysoivat menetelmät pyrkivät mittaamaan esseestä sen todellisia ominaisuuksia eli samoja ominaisuuksia, joilla ihmiset muodostavat käsityksensä esseen laadusta (Page 1966). Simuloivat menetelmät pyrkivät mallintamaan useasta arvioijasta koostuvan ryhmän antamia arvosanoja, ilman että esseestä välttämättä mitataan vastaavia ominaisuuksia kuin ihmiset mittaavat. Arviointi tapahtuu käyttäen todellisten ominaisuuksien korrelaatteja ja tavoitteena on tuottaa mahdollisimman hyvä korrelaatio arvioijien antamiin arvosanoihin nähden.

Taulukko 4. Arviointimenetelmien jaottelu Pagen (1996) esittämän mallin mukaan.

	Sisältö	**Tyyli**
Simulointi	LSA, TCT, BETSY	PEG, TCT
Analyysi	E-RATER	E-RATER

Menetelmien jaottelussa PEG sijoittuu selvimmin asetettuihin kategorioihin. Se on selkeästi simuloiva järjestelmä, joka painottaa tyylillisiä seikkoja. PEG perustuu ajatukseen esseiden

todellisten ominaisuuksien arvioinnin mahdollistavista arvioista. PEG ei myöskään pyri tutkimaan esseen asiasisältöä, vaan keskittyy pintaominaisuuksien tarkastellun. PEG:n etu muihin menetelmiin nähden on sen laskennallinen helppous (Chung ja O'Neil 1997). PEG on rakenteeltaan suhteellisen yksinkertainen ja helposti ymmärrettävissä. Lisäksi sen kyky arvioida tyylillisiä seikkoja, kuten välimerkkien käyttöä ja oikeinkirjoitusta, on muihin menetelmiin verrattuna hyvä. Merkittävin puute PEG:ssa on sen keskittyminen pintasyntaktisiin ominaisuuksiin sisällön tarkastelun sijasta. Puutteeksi on erityisesti tutkimuksen kannalta laskettavissa myös se, että PEG on suljettu menetelmä, eikä tarkkoja tietoja käytetyistä arviomuuttujista tai järjestelmän tarkasta rakenteesta ole julkaistu 60-luvun jälkeen.

LSA on menetelmistä selvimmin sisällön tarkastelua painottava. Se perustuu valmennusmateriaalista ja arvioitavista esseistä luotaviin dokumentti-sana -matriiseihin. LSA:n lähestymistapaa voidaan pitää simuloivana itse pisteytysprosessin osalta. Onhan dokumenttivektorien välisten kulmien laskeminen varsin kaukana ihmisen suorittaman arvioinnin prosesseista. Toisaalta itse tekstin sisällön tutkimista voidaan pitää analysoivana. Kuten jo aiemmin tässä tutkielmassa on todettu, LSA:n tapa tekstin merkityksen tulkitsemisessa vastaa läheisesti ihmisen luetun ymmärtämisessä käyttämiä prosesseja. LSA:n vahvuus muihin menetelmiin verrattuna on, että se on menetelmistä ainoa, jonka avulla voidaan suorittaa arviointia ilman, että valmennusmateriaalina käytetään ihmisten arvioimia esseitä.

LSA:n suurimmaksi heikkoudeksi Landauer et al. (1998a) mainitsevat sen, ettei menetelmä huomioi sanajärjestystä. Tästä johtuen myös lauseopilliset suhteet ja johdonmukaisuus jäävät menetelmältä huomioitta. Puutteistaan huolimatta LSA pystyy erottelemaan tekstinosien ja sanojen merkityksiä varsin hyvin. Toinen merkittävä heikkous LSA:ssa on sen laskennallinen vaativuus (Chung & O'Neill 1997). Esseistä ja lähdemateriaaleista muodostetut matriisit kasvavat varsin suuriksi ja niillä suoritettavat laskutoimitukset siten aika- ja tilavaativiksi.

Kummankin Naiivi Bayes -malliin perustuvan menetelmän, TCT:n ja BETSY:n, voidaan katsoa kuuluvan simuloiviin esseen sisältöä painottaviin menetelmiin. BETSY ei huomioi kirjoitustyyliin liittyviä seikkoja lainkaan. Myös TCT:ssa nämä ominaisuudet ovat varsin rajoittuneita. Kumpikaan Naiivi Bayes -menetelmissä ei myöskään kykene esseen rakenteen tarkasteluun. Itse arviointiprosessi on Naiivi Bayes -malleissa lähtökohdiltaan puhtaan matemaattinen dokumenttien luokittelu eri kategorioihin. Naiivi Bayes -mallien etuna, esim. LSA:han verrattuna, voidaan nähdä niiden laskennallinen helppous. Järjestelmät ovat myös

rakenteeltaan suhteellisen yksinkertaisia ja helposti ymmärrettäviä. Lisäksi ne ovat täysin avoimia menetelmiä. Käytetyistä menetelmistä on saatavilla yksityiskohtaiset kuvaukset.

E-Rater on esitellyistä menetelmistä ainoa, jonka lähtökohta arvioinnin suorittamiseen on analyyttinen. Siinä kiinnitetään huomiota sekä tyyliin että sisältöön liittyviin ominaisuuksiin ja kehittämisen lähtökohtana on käytetty arvioijia varten luotuja ohjeita. Näitä lähtökohtia voidaan pitää sen vahvoina ominaisuuksia. E-Rater on järjestelmänä varsin monimutkainen ja suhteellisen hankalasti lähestyttävä. E-Rater on kaupallisuudestaan huolimatta pidetty varsin avoimena järjestelmänä ja sen toiminnasta on saatavilla varsin yksityiskohtaista tietoa.

Taulukkoon 5 olen kerännyt menetelmät tutkielman luvussa 1 esitettyjen automaattisen esseenarvioinnin vaatimusten (Kaplan et al. 1998) täyttymisen osalta.

Taulukko 5. Automaattisten arviointimenetelmille asetettujen vaatimusten toteutuminen esitellyillä menetelmillä.

Menetelmä	Puolustettavuus	Tarkkuus	Valmennettavuus	Kustannukset
PEG	Kritisoitu juuri tästä syystä. Välimerkkien ja sanojen ym. määrien laskeminen on kritisoijien mielestä heikko perusta arvioinnille. Toimintatapa sinänsä on helposti jäljitettävissä.	Erinomainen. Koetulosten valossa PEG:llä saavutetaan esitetyistä menetelmistä tarkimpia tuloksia.	Käytetyt arviointimenetelmät suhteellisen yksinkertaisia, joten voi aiheuttaa ongelmia.	Vaaditun suuren valmennusmateriaalin vuoksi ei sovellu pienten vastausmäärien arviointiin. Laskennallisesti yksinkertaisempi kuin esim. LSA, joten ei vaadi kalliita laitteistoja.
TCT	Tutkii sekä tyyliä, että sanastoa, joten puolustettavuus on vahvempi kuin esim. PEG:ssa.	Hyväksyttävä. Koetulokset osoittavat tarkkuuden olevan samaa luokkaa kuin ihmisillä arvosanojen välistä korrelaatiota mittana käytettäessä.	Tutkii esseen sisältöä ja tyyliä, mutta ei rakennetta, joten valmennettavuus on ongelma.	Kuten PEG.
BETSY	Tutkii vain esseen sisältöä ja jättää tyyliseikat huomiotta.	Kuten TCT	Tutkii esseen sisältöä, mutta ei rakennetta tai tyyliä, joten valmennettavuus on ongelma.	Kuten PEG.
LSA	Kuten BETSY, perustuu vain sisällön arviointiin.	Kuten TCT	Kuten BETSY	Mahdollistaa oppikirjojen ja mallivastauksien käytön valmennusmateriaalina. Käytettäessä esseitä suhteellisen pieni valmennusm. on riittävä. Laskennallinen vaativuus asettaa vaatimuksia laitteistolle.
E-RATER	Huomioitu jo lähtökohdissa. Tässä suhteessa vahvin järjestelmistä.	Kuten TCT	Tässä suhteessa vahvin menetelmistä. Tutkii myös esseen rakennetta ja väitteiden organisointia.	Vaadittu valmennusmateriaali on 270 esseen kokoinen, joten mahdollinen ja taloudellisesti järkevä vain suurehkojen esseemäärien arvioinnissa.

Selvimmin kaikkien menetelmien osalta täyttyy tarkkuuden vaatimus. Jo Pagen (1966) ensimmäinen PEG-versio 1960-luvulla oli koetulosten valossa kykenevä tuottamaan tarkkuudeltaan arvioijien arvosanoihin verrattavia arvosanoja. Kustannustehokkuuden toteutuminen on riippuvainen arvioitavien esseiden määrästä. Valmennusmateriaalin kokoaminen järjestel-

män käytettäväksi ei tietenkään ole mielekästä, mikäli arvioitavia esseitä on hyvin vähän. Parhaiten tästä vaatimuksesta suoriutuu LSA, joka mahdollistaa lähdekirjallisuuden tai asiantuntijoiden luomien mallivastausten käytön osana valmennusmateriaalia. Kaikki järjestelmät edellyttävät kysymyskohtaisen mallin luomista, joka on huomioitava kustannuksia laskettaessa. Ehkä suurimmat ongelmat kaikkien menetelmien osalta liittyvät valmennettavuuteen. Tuntemalla järjestelmän toimintatapoja, kirjoittajan voi olla mahdollista tarkoituksellisesti harhauttaa järjestelmältä ansaittua parempia arvosanoja. Tähän seikkaan on uusimmissa tutkimuksissa kiinnitetty erityistä huomiota (mm. Powers et al. 2002). Parhaaksi menetelmäksi on osoittautunut ”epäilyttävien” esseiden merkkaaminen ja arvioinnin jättäminen ihmisen tehtäväksi. Puolustettavuuden vaatimus täyttyy parhaiten E-Raterin osalta, joka jo lähtökodiltaan painottaa siihen liittyviä ominaisuuksia.

Järjestelmien ominaisuuksien ja mahdollisten virhelähteiden analysoinnissa olen käyttänyt Chungin ja O’Neilin (1997) esittämää tapaa. Tutkimuksessaan he jaottelivat PEG:n ja LSA:n mahdolliset ongelmakohdat. Olen koonnut taulukon 6 Chungin ja O’Neilin esittämien lisäksi Naiivi Bayes -menetelmiä ja E-Rateria koskevat tiedot.

Taulukko 6. Järjestelmien ominaisuudet pisteytyksen eri osa-aluilla sekä mahdolliset virheitä aiheuttavat tekijät esitellyissä menetelmissä Chungin ja O'Neilin (1997) pohjalta.

Muuttuja	PEG	Naiivi Bayes	LSA	E-Rater
Ihmisten suorittama arviointi	Merkittävä. Regressiomalli perustuu suoraan valmennusmateriaalin arvosanoihin. Useamman arvioijan käyttö pienentää virhemahdollisuutta.	Vaikutus kuten PEG:ssa. Esseiden kategorisointi perustuu arvioijien antamiin arvosanoihin.	Käytettäessä esseitä valmennusmateriaalina vaikutus sama kuin PEG:ssa. Jos valmenusmateriaalina on oppkirjan katkelmat, ei vaikutusta.	Vaikutus kuten PEG:ssa.
Esseen ominaisuudet				
Sisältö	Ei suoraa vaikutusta malliin.	Sekä TCT että BETSY käyttävät sanojen ilmentymiin perustuvaa luokitinta. Sisältö vaikuttaa tätä kautta.	Huomattava vaikutus. Esseiden arviointi perustuu nimenomaan sisällön samankaltaisuuden vertailuun.	Huomattava vaikutus. Yksi arvosanaan vaikuttavista osatekijöistä on sisältö, jota tutkitaan sisältövektorianalyysilla.
Pintasyntaksi	Merkittävä. Ovat mukana yksittäisinä, erillisinä ominaisuuksia regressiomallissa.	TCT:ssä vaikutus vaihtelee sen mukaan, mitkä muuttujat valitaan malliin regressioanalyysissä. BETSY:ssä ei vaikutusta.	Ei vaikutusta. LSA ei tutki pintasyntaktisia ominaisuuksia.	Merkittävä vaikutus. Pintasyntaktisia seikat huomioidaan tutkitaan. Vaikutus ei kuitenkaan yhtä merkittävä kuin PEG:ssa.
Esseen pituus	Merkittävä. Eräs tärkeimmistä ennustavista muuttujista.	TCT:ssä merkitys yleensä vähäinen. Vaihtelee sen mukaan, mitkä muuttujat valikoituvat regressioanalyysissä malliin. BETSY:ssä ei vaikutusta.	Ei suoraa vaikutusta malliin.	Ei vaikutusta. Esseen pituuden käyttö on suljettu pois jo peruslähtökohdissa.
Välimerkkien käyttö	Merkittävä. Eri välimerkkien määrät ovat merkittävässä osassa pisteytysmallissa.	Ei vaikutusta TCT:ssä tai BETSY:ssä.	Ei vaikutusta. LSA ei tutki välimerkkien käyttöä.	Ei vaikuta pisteytysmalliin suoraan.
Dokumentin esikäsittely				
Sanojen muuntaminen perus muotoon	Vaikutusta toimintaan ei tunneta, mutta oletettavasti varsin merkittävä.	Käytössä TCT:ssä. BETSY:n mallissa on kokeiltu sekä perusmuotoisia että alkuperäisiä, taivutettuja sanoja.	Vaikutusta ei tarkkaan tunneta, mutta jotkut tutkijat olettavat sen parantavan menetelmän suorituskykyä.	Käytetään, mutta tarkkoja vaikutuksia ei tunneta.
Sulkusanojen poisto	Vaikutusta ei tunneta.	TCT käyttää sulkulistaa. BETSY:ä voidaan käyttää sulkulistalla ja ilman sulkulistaa.	Käyttö tärkeää, koska yleisimmät sanat eivät vaikuta vertailuun, mutta kasvattavat vektorien kokoa tarpeettomasti.	Sisällön arviointia suorittava osa käyttää sulkulistaa, mutta tarkkaa vaikutusta ei tunneta.
Termien painotus	Ei käytössä.	TCT:ssä k:n lähimmän naapurin luokitin käyttää termien tf*idf-painotusta. BETSY:ssä termien painotusta ei käytetä.	Merkittävä. LSA:ssa on sovellettu monia erilaisia painotusmalleja ja ne vaikuttavat lopputulokseen merkittävästi.	Merkittävä vaikutus. Sisällön tarkastelussa käytetään tf*idf-painotuksen kaltaista menetelmää.
Valmennus				
Materiaalin koko	Merkittävä. Mallin toiminta heikkenee, mikäli valmennusmateriaali ei ole riittävän kokoinen.	Merkittävä. Luokittimien toimintatarkkuus on riippuvainen valmennusmateriaalin koosta.	Merkittävä. LSA:n on saatava riittävästi tietoa sanojen merkityksistä toimiakseen luotettavasti.	Ei vaikutusta. Käytetään vakiokokoista valmennusmateriaalia.
Luokittelu				
Samankaltaisuusmitta	Ei käytössä.	TCT:ssä vaikutus merkittävä, mikäli k:n lähimmän naapurin luokitin valikoituu malliin. BETSY:ssä ei vaikutusta.	Kokeissa on testattu erilaisten mittojen toimivuutta. Eri mitat aiheuttavat merkittäviä muutoksia tuloksissa.	Merkittävä. Sisällön samankaltaisuutta verrataan kahdella eri menetelmällä, koko esseen sekä argumenttitasolla.

Taulukosta 6 on havaittavissa, että kaikissa malleissa on tiettyjä ominaisuuksia, jotka ovat herkkiä virheiden syntymiselle. Käytettäessä ihmisten arvioimista esseistä koostuvaa valmennusmateriaalia, on ilmiselvänä virhelähteenä arvioijien antamien arvosanojen epätarkkuus. Riskiä voidaan pienentää käyttämällä valmennusmateriaalina useamman kuin yhden henkilön arvioimia esseitä, jolloin yksittäisen arvioijan arvosanojen vaihtelu ei saa niin suurta merkitystä. Valmennusmateriaaliin liittyvä mahdollinen virhelähde on myös valmennusmateriaalin koko. Jos järjestelmä ei saa riittävää vertailuaineistoa arvioinnin pohjaksi, sen tarkkuus kärsii merkittävästi.

Mahdollisina virhelähteinä voidaan pitää myös valmennusmateriaalin ja arvioitavan esseen vertailussa käytettyjä yhtäläisyysmittoja sekä sanojen painotusmalleja. Useissa menetelmissä on käytetty useita erilaisia vertailu- ja painotuskaavoja ja usein niiden käyttö perustuu kokeilemalla saavutettuihin tuloksiin. Sulkulistoja ja sanojen muuttamista perusmuotoon käytetään menetelmissä yleisesti ja niiden on useimmissa malleissa todettu parantavan luokittelutarkkuutta. Niiden tarkkoja vaikutuksia ei kuitenkaan ole tutkittu kovinkaan laajasti tai tuloksia ei ainakaan ole esitetty julkisesti.

4.2 Tulokset

Taulukossa 7 on esitetty tässä tutkielmassa tarkastellun neljän arviointimenetelmän avulla tutkimuksissa saavutettuja tuloksia. Tulokset on pyritty keräämään mahdollisimman uusista tutkimuksista, jotta ne edustaisivat hyvin menetelmien nykyistä tilaa.

Taulukko 7. Menetelmillä kokeissa saavutetut tuloksia (Shermis et al. 2001, Larkey 1998, Landauer et al. 1997, Burstein et al. 1998c).

Menetelmä	Valmennus-materiaalin koko	Arvioitujen esseiden määrä	Arvosanano-ja asteikossa	Arvioijien lukumäärä	Arvioijien välinen r	Arvioijien ja järjestelmän r
PEG	1293	617	22	6	0.62	0.71
TCT	383	225	6	2	0.88	0.86
LSA	oppikirja	273	5	2	0.65	0.64
	94	94	5	2	0.77	0.77
E-RATER	270	634	6	2	0.79	0.74

Shermis et al. (2001) kokeessa PEG:n avulla arvioitiin high school- ja collegeopiskelijoiden kirjoittamia tasokoe-esseitä. Esseet pisteytettiin asteikolla 1 - 22, jonka mukaan opiskelijat jaetaan eri tasoisille kirjoittamiskursseille. Kokeessa esseet oli arvioitu kuuden henkilön toimesta. Tulosten perusteella todettiin, että PEG pystyi tuottamaan kuuteen arvioijaan verrattuna keskimäärin tarkempia arvosanoja kuin nämä toisiinsa verrattuna.

Larkeyn (1998) tutkimuksessa eräänä koemateriaalina käytettiin vastauksia esseetehtävään, jossa yliopiston jatko-opintoihin pyrkivän opiskelijan tuli arvioida kysymyksessä esitettyä väittämää. Larkey totesi kaikissa viidessä testiaineistossaan TCT:n ja arvioijien antamien arvosanojen korrelaatioiden olevan samaa luokkaa.

Landauer et al. (1997) testasivat LSA:ta käyttäen valmennusmateriaalina arvioituja esseitä sekä oppikirjaa. Esseitä valmennusmateriaalina käyttäneessä kokeessa arvioitiin 94 yliopisto-opiskelijan kirjoittamaa, ihmissydämen toimintaa ja rakennetta käsittelevää esseetä. Toisessa kokeessa, jossa valmennusmateriaalina käytettiin oppikirjaa, psykologian opiskelijat kirjoittivat esseen jostakin annetusta kolmesta aiheesta oman valintansa mukaan. Kummankin kokeen tulokset osoittivat, että LSA:n antamat arvosanat vastasivat arvioijien antamia arvosanoja yhtä tarkasti kuin arvioijien arvosanat toisiaan. Lisäksi tutkimuksessa verrattiin sekä LSA:n että arvioijien antamia arvosanoja ulkoiseen kriteeriin. Ulkoisena kriteerinä käytettiin tuloksia kokeesta, jossa samat kirjoittajat vastasivat esseen aihealuetta koskeviin, lyhyen vastauksen vaativin tehtäviin (short-answer test). LSA:n arvosanat korreloivat myös ulkoiseen kriteeriin verrattuna vähintään yhtä voimakkaasti kuin arvioijien antamat arvosanat.

Burstein et al (1998c) kokeessa eräs koemateriaali oli GMAT AWA-kokeen (ks. luku 3.4.1) väitteen analysointitehtävä. Suurin osa esseevastauksista oli sellaisten henkilöiden kirjoittamia, joiden äidinkieli ei ollut englanti. Tulokset olivat kaikilla koeaineistoilla taulukossa 7 esitetyn kaltaisia ja Burstein et al. totesivat, että E-Raterin arvosanat ovat täysin vertailukelpoisia arvioijien arvosanojen kanssa.

4.4 Haasteita ja tulevaisuuden kehityssuuntia

Sekä Burstein ja Chodorow (2002), Hearst et al (2000) että Shermis et al. (2002) näkevät esseiden arvioinnin automatisoinnin tärkeimmäksi kehityssuunnaksi luoda järjestelmiä, jotka kykenevät antamaan kirjoittajalle arvosanan lisäksi yksityiskohtaisempaa palautetta sekä ohjeita esseen korjaamiseksi. Hearst et al. 2000 toteavat, että hyväksi kirjoittajaksi oppimisen eräistä edellytyksistä on yksityiskohtaisen palautteen saanti. Heidän mukaansa palaute rajautuu kirjoitusharjoitustehtävissä usein pelkästään pilkku- ja kirjoitusvirheiden merkitsemiseen. Syitä tähän ovat opettajien työpaineista johtuva kiire ja toisaalta asenteet. Jotta automaattisten järjestelmien avulla olisi mahdollista helpottaa tehtävää, olisi niissä holistisen arvioinnista kohti analyyttisempia, yksityiskohtaisemman palautteen ja ohjeiden antamista tukevia menetelmiä (Burstein ja Chodorow 2002).

Eräs tapa yksityiskohtaisemman palautteen tuottamiseksi asiakeskeisissä esseissä on luoda esseestä automaattista menetelmää käyttäen tiivistelmä ja tutkia sen pääkohtia (Burstein ja Marcu 2000a). Kun hyviä arvosanoja saaneiden esseiden pohjalta on luotu käsitys niiden pääkohdista, voidaan heikompitasoisista esseistä osoittaa puuttuvat asiat vertaamalla sitä luotuun malliin. Marcun (1999) kehittämä, *retoristen rakenteiden teoriaan* (rhetorical structure theory, RST) perustuva tapa tiivistelmien generointiin ja tekstiyksiköiden tärkeyden pisteytykseen on osoittautunut varsin toimivaksi ratkaisuksi.

Shermis et al. (2002) ovat kehittäneet PEG:n pohjautuvan järjestelmän, joka antaa holistisen arvosanan lisäksi erillisen arvion kustakin esseen viidestä ominaisuudesta. Nämä ovat sisältö, luovuus, tyyli, oikeinkirjoitus ja organisointi. He näkevät eri osa-alueilta saatavan palautteen kirjoittajan kannalta erittäin tärkeäksi ja toteavat etteivät arvioijien ja opettajien ajalliset resurssit nykyisin yleensä mahdollista tämän tyyppistä erittelevää arviointia.

Yksityiskohtaisemman palauutteen lisäksi Burstein ja Chodorow (2002) näkevät tärkeäksi automaattisten menetelmien kehityksessä sen, että yhteys koulutuksellisiin tavoitteisiin säilytetään jatkuvasti ja että järjestelmät noudattavat samoja arvioinnin periaatteita kuin ihmiset. He toteavat joidenkin nykyisten järjestelmien kuten, E-Raterin ja LSA:han perustuvan Intelligent Essay Assessorin, noudattavan tätä periaatetta siten, että niiden pisteytysmallit on luotu arvioijia varten kehitettyihin arviointiohjeisiin perustuen.

Tämän tutkielman ensimmäisessä luvussa esitelty valmennettavuuteen liittyvä vaatimus on eräs toistaiseksi ratkaisemattomista ongelmista. Powers et al. (2002) tutkivat E-Raterin kykyä havaita esseistä tarkoituksellisia yrityksiä huijata järjestelmää. Tutkimuksen tuloksena todettiin, että E-Rater antoi helpommin liian korkeita kuin alhaisia arvosanoja. Tutkimuksessa päädyttiin samaan johtopäätökseen kuin Page ja Petersen (1995) PEG:n osalta. Heidän mukaansa nykyisiä järjestelmiä ei tulisi käyttää ainoana arvioijana ainakaan arviointitehtävissä, joihin liittyy korkea oikeellisuuden vaatimus, kuten tenteissä tai pääsykokeissa. Powers et al. (2002) esittävät ratkaisuksi ongelmaan suodattimia (off-topic filters), joiden avulla esseen relevanssia esitettyyn kysymykseen voitaisiin tutkia. Powers et al. näkevät erääksi ratkaisuksi esseen ja tehtävänannon leksikaalisen sisällön vertaamisen. Jos niiden päällekkäisyys on vähäinen, voidaan olettaa, ettei essee käsittele annettua tehtävää. LSA:han perustuvassa Intelligent Essay Assessor:ssa pyritään havaitsemaan mahdollista plagiointia vertaamalla arvioitavaa esseetä valmennusmateriaaliin (Hearst et al. 2000). Jos essee on ”liian” samankaltainen kuin jokin valmennusmateriaalin esseistä, voidaan sen epäillä olevan plagioitu. Toisaalta, jos essee ei ole samankaltainen minkään valmennusmateriaalin esseen kanssa, voidaan olettaa että se ei käsittele annettua aihetta. Vaikka tutkimuksissa on todettu saavutetun varsin hyviä tuloksia poikkeavien esseiden havaitsemisessa, ongelman tyydyttävä ratkaiseminen vaatii selvästikin vielä lisää tutkimusta.

Miltsakaki ja Kukich (2000) tutkivat *keskitysteoriaan* (centering theory) perustuvaa menetelmää esseiden johdonmukaisuuden arvioimiseen. Menetelmän tavoitteena on löytää esseestä äkillisiä aiheen vaihdoksia etsimällä *karkeita siirtymiä* (rough-shifts) aiheesta toiseen. Karkeat siirtymät ilmaisevat lyhytkestoisia aiheen käsittelyjä ja osoittavat kirjoittajan käyttäneen heikkoa teeman kehittelyä. Tutkimuksessa todettiin, että karkeiden siirtymien suhteelliset määrät olivat esseiden arvosanoihin merkittävästi vaikuttava osatekijä. Lisäksi kyky paikallistaa aiheen vaihdoksia antaisi mahdollisuuden osoittaa kirjoittajalle, mitkä esseen kohdat kaipaisivat hiomista.

Myös sanaston käyttöön ja kielioppiin liittyvien virheiden havaitsemiseen on kehitteillä uusia menetelmiä. Chodorowin ja Leacockin (2000) ALEK (Assesssing of Lexical Knowledge) -tekniikan avulla on tarkoitus pystyä havaitsemaan tiettyjen sanojen käytössä virheitä tutkimalla konteksteja, joihin kyseinen sana sisältyy. Perinteiset oikeinkirjoituksen tarkastamisen välineethän yleensä tutkivat vain puhtaasti kirjoitusvirheitä. ALEK pyrkii havaitsemaan vir-

heitä sanojen käyttötavoissa. Esimerkki sanan käyttövirheestä on vaikkapa ”a desks”, joka selvästikin rikkoo englannin kielioppisääntöjä. Tutkittavan sanan käyttöä esseessä verrataan englannin kielen *korpuksesta* (corpus) automaattisesti muodostettuun malliin. Kokeissa ALEK:n on todettu yltävän noin 80 % tarkkuuteen sanojen käyttövirheiden havaitsemisessa.

Hearst et al. 2000 mainitsevat luonnollisten kielten käsittelyn pitemmän aikavälin tavoitteeksi luoda järjestelmä, joka ”ymmärtää” lukemaansa tekstiä. Tutkimuksissa on kehitetty järjestelmiä, jotka pystyvät tutkimansa tekstin perusteella vastaamaan sekä monivalintatehtäviin että tekstimuotoisiin kysymyksiin lyhyin, muutaman lauseen mittaisin vastauksin. Jotta automaattisten kysymyksiin vastaavien järjestelmien luominen olisi mahdollista, käytössä on oltava myös automaattisia menetelmiä arvioida generoitujen vastausten laatua.

5. JÄRJESTELMÄ SUOMEN KIELISTEN ESSEIDEN ARVIOINTIIN

Osana tätä tutkielmaa toteutettiin kokeellinen automaattisen esseen arvioinnin järjestelmä. Tutkielmassa esitellyt viisi esseen arvioinnin automaattista järjestelmää on kehitetty englanninkielteisten esseiden arviointiin. Eräs lähtökohta järjestelmän kehittämiselle oli siis automaattisen esseen arviointimenetelmin soveltuvuuden testaaminen suomen kielisten esseiden osalta.

5.1 Lähtökohdat

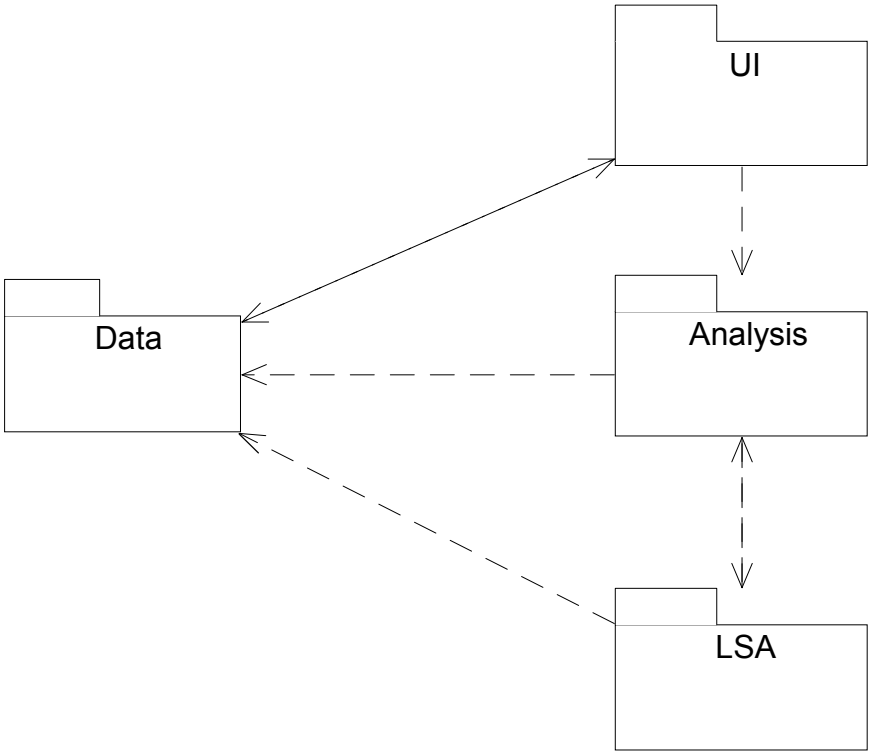
Järjestelmän suorittama esseiden arviointi perustuu tutkielman luvussa 3.3 esiteltyyn latenttiin semanttiseen analyysiin. Perusteena latentin semanttisen analyysin käytölle oli se, että se on arviointimenetelmistä ainoa, joka mahdollistaa esseiden arvioinnin myös oppimateriaaliin perustuen. Tämä mahdollistaa arvioinnin myös tilanteissa, joissa käytettävänä ei ole suurta määrää valmiiksi arvioituja esseitä.

Suomen kielen erityispiirteet moneen muuhun kieleen verrattuna aiheuttavat ongelmia sen käyttämisessä tiedonhaun sovelluksissa (Alkula 2000). Ongelmat koskevat siis myös automaattista esseiden arviointia. Ongelmiksi suomen kielen kohdalla muodostuvat mm. sanojen sijamuodot. Nomineilla on 15 sijamuotoa. Erilaisia sanamuotoja muodostetaan myös käyttämällä tunnuksia, johtimia ja liitteitä sekä näiden yhdistelmiä. Yhdyssanojen runsaus on toinen suomen kielen erikoisominaisuus. Esimerkiksi Nykysuomen sanakirjan sanoista noin kaksi kolmasosaa muodostuu yhdyssanoista. Suomen kielen sanojen vartalot eivät ole löydettävissä riittävällä tarkkuudella esim. englannin kielisen tekstin yhteydessä käytetyllä menetelmällä, jossa sanojen päätteet poistetaan katkaisemalla. Tämän vuoksi järjestelmä käyttää tukenaan suomen kielen jäsennintä.

5.2 Toteutus

Järjestelmä sisältää toiminnot tehtävien ja esseiden tallentamiseen, poistoon, hallinnointiin sekä arviointiin. Tehtävä- ja esseetiedot säilytetään tietokannassa. Suomen kielisen tekstin käsittelyyn järjestelmä käyttää suomen kielen jäsennintä. Sen avulla tuotetaan mm. sanoista

niiden perusmuodot. Järjestelmä on toteutettu Javalla ja sen rakenne on pyritty laatimaan siten, että jatkokehittäminen olisi mahdollisimman vaivatonta. Järjestelmän rakenne on esitetty kuvassa 9 UML-komponenttikaaviona.



Kuva 9. Komponenttikaavio järjestelmän rakenteesta.

UI-paketti sisältää käyttöliittymäkomponentit ja Data-paketti tietokantayhteyden ylläpitoon liittyvän luokan sekä tehtävä-, essee- ja sanasto-luokat. Lisäksi se huolehtii yhteydestä suomen kielen jäsennintä suorittavaan palvelinkoneeseen. Analysis-paketti vastaa esseiden arviointiin liittyvistä toiminnoista ja LSA-paketti latentin semanttisen analyysin suorittamisesta. Singulaariarvohajotelman laskeminen tapahtuu käyttäen C-kielellä toteutettua SVDPACKC-kirjaston (Berry et al. 1993) bls2-algoritmin muokattua versiota. Bls2 käyttää singulaariarvohajotelman laskemiseen Block Lanczos -menetelmää.

Suomen kielen jäsentimenä järjestelmässä käytetään Conexor Functional Dependency Grammar 3.7 for Finnish -sovellusta (Conexor Oy 2002). Jäsennin tuottaa funktionaaliseen dependenssikieliopin mukaisen jäsennyksen luonnollista kieltä sisältävästä syötteestä (Tapanainen 1999). Taulukossa 8 on esimerkki Conexor FDG:n tuottamasta jäsennyksestä esimerkkilauseelle. Jäsennyksen tuloksena syöte jaetaan alkioiksi, joita ovat esim. sanat, monisanaiset

yksiköt ja numeroita sisältävät ilmaisut. Jäsennin tuottaa kullekin alkiolle sen perusmuodon sekä tiedot sen funktionaalisesta riippuvuudesta ja morfosyntaktisista ominaisuuksista. Funktionaalinen riippuvuus kuvaa alkion suhteita muihin alkioihin. Morfosyntaksiset ominaisuudet kuvaavat alkion syntaktisia sekä pintasyntaktisia ja morfologisista ominaisuuksista.

Taulukko 8. Esimerkki Conexor Functional Dependency Grammar 3.7 for Finnishin tuottamasta jäsennyksestä lauseelle: ”Puolustuksen suomalaista tukipilaria kaivataan kommentoimaan tappioita, mutta voiton jälkeen media kirmaa maalintekijöiden perässä.”

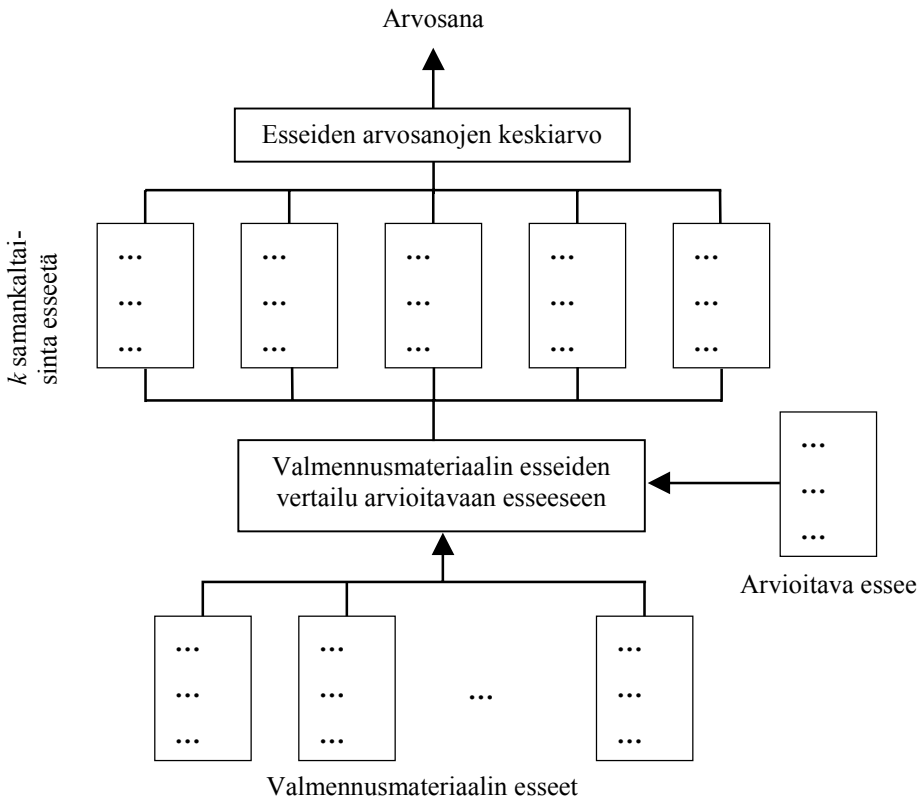
Paikka	*Sana*	*Perusmuoto*	*Funktionaalinen riippuvuus*	*Funktionaalinen, pintasyntaktinen ja morfologinen merkki*
1	Puolustuksen	puolustus	attr:>3	&A> N SG GEN
2	Suomalaista	suomalainen	attr:>3	&A> A SG PTV
3	Tukipilaria	tuki#pilari	obj:>5	&NH N SG PTV
4	Kaivataan	kaivata	main:>0	&+MV V PASS IND PRES
5	Kommentoimaan	kommentoida	obj:>4	&-MV V ACT INF3 SG ILL
6	Tappioita	tappio		&NH N PL PTV
7	,	,		
8	Mutta	mutta		&CC CC
9	Voiton	voitto		&NH N SG GEN &NH A SG NOM
10	Jälkeen	jälkeen	goa:>5	&ADV ADV
11	Media	media	subj:>12	&NH N SG NOM
12	Kirmaa	kirmata		&+MV V ACT IND PRES SG3
13	Maalintekijöiden	maalin#tekijä		&NH N PL GEN
14	Perässä	perässä	pm:>13	&PM PSP
15	.	.		
16	<p>	<p>		

Latentin semanttisen analyysin suorittamiseksi valmennusmateriaalista ja arvioitavista esseistä muodostetaan vektoriesitykset Conexor FDG:n tuottamia sanojen perusmuotoja käyttäen. Vektoriesitykseen ei sisällytetä sulkusanalistalla olevia sanoja. Järjestelmää kykenee käyttämään käyttäjän laatimia erilaisia sulkusanalistoja sekä suorittamaan analyysin ilman sulkusanojen poistoa. Sulkusanojen lisäksi tekstistä poistetaan välimerkit sekä numerot.

Järjestelmä käyttää termien painottamiseen kaavan 6 (s. 28) mukaista entropiaan perustuvaa painotusmallia. Analyysi voidaan suorittaa myös ilman termien painotusta. Dokumenttien samankaltaisuuden vertaaminen tapahtuu käyttäen mittana niiden sisältövektorien välistä kosinia kaavan 10 (s. 31) mukaisella tavalla.

Automaattinen arviointi voidaan suorittaa järjestelmän avulla kahdella eri menetelmällä: vertaamalla arvioitavia esseitä 1) valmennusmateriaalin sisältämiin esseisiin tai 2) esseetehtävän aihealuetta käsittelevään oppimateriaaliin.

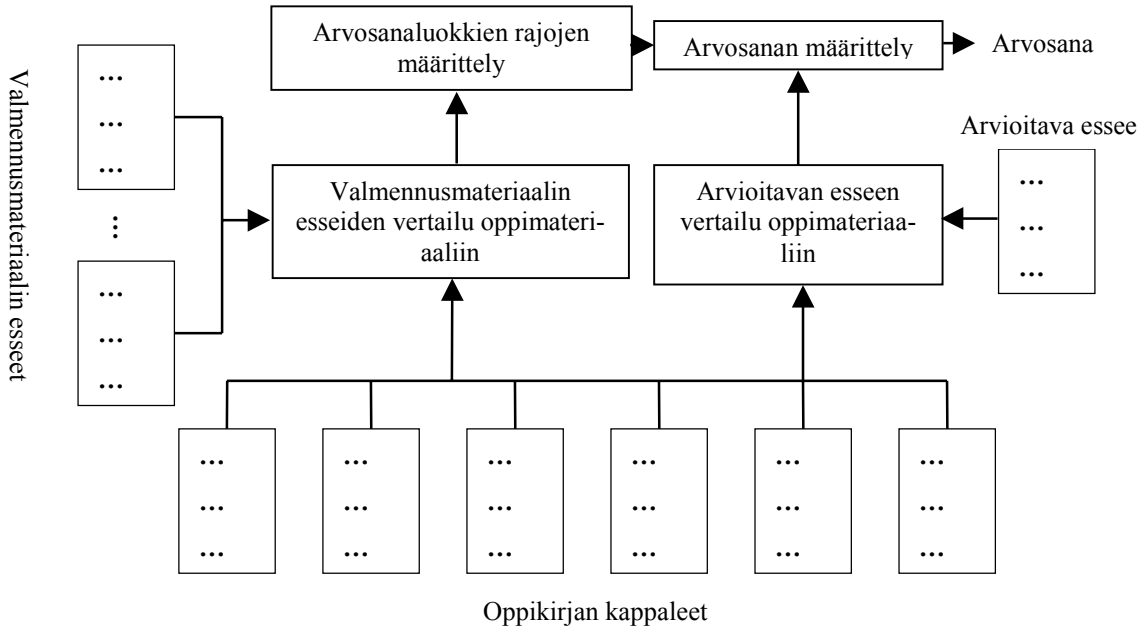
1) Valmennusmateriaalin esseisiin vertailuun perustuvassa menetelmässä esseen arvosana määräytyy *k*:n sitä lähimmin vastaavan valmennusmateriaalin esseen arvosanojen keskiarvon mukaisesti. Keskiarvo voidaan laskea joko aritmeettisena keskiarvona tai painotettuna keskiarvona, siten että kunkin *k*:n esseen arvosanaa painotetaan sen ja arvioitavan esseen yhdenmukaisuusarvon (sisältövektorien välisen kosinin) perusteella. Kuva 10 havainnollistaa menetelmän toimintaa.



Kuva 10. *K*:n lähimmän valmennusmateriaalin esseen arvosanojen keskiarvoon perustuva malli. Kuvan esimerkissä *k*:n arvo on viisi.

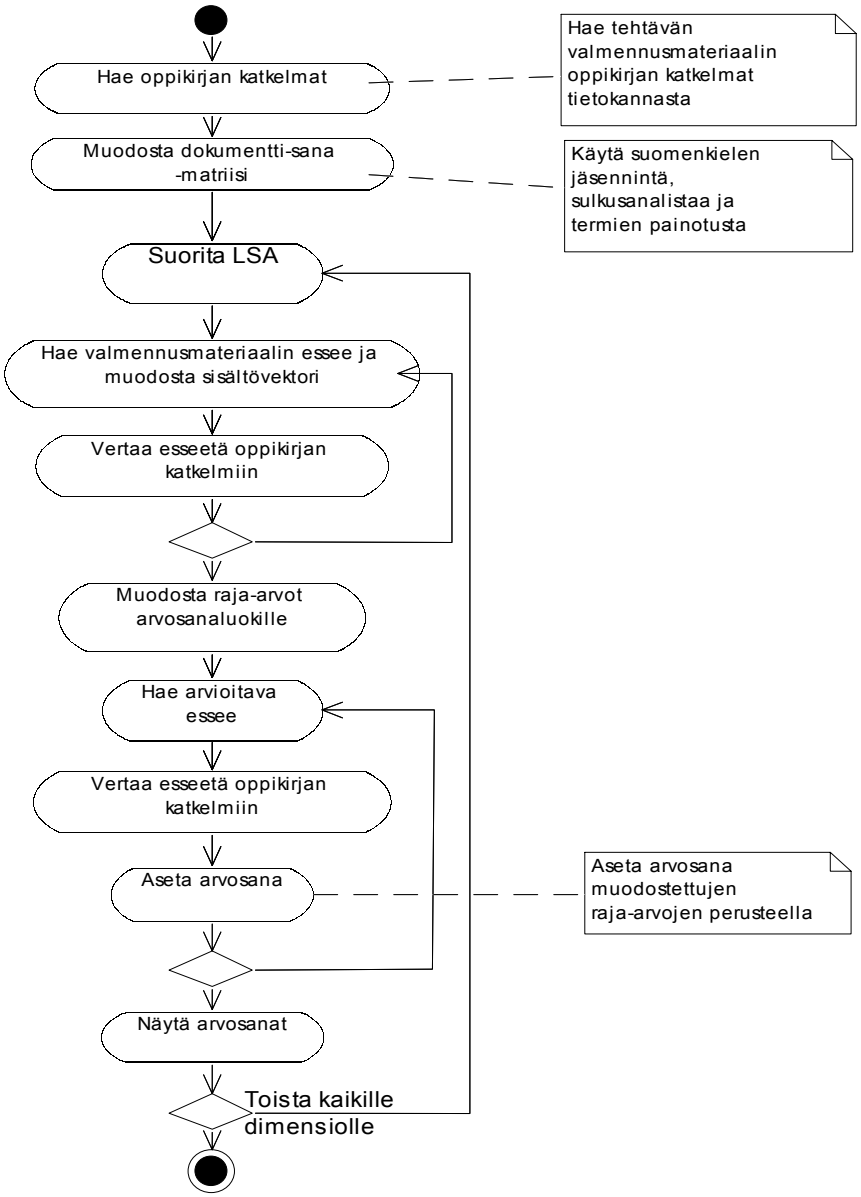
2) Oppimateriaaliin vertailuun perustuvassa menetelmässä arvioitavia esseitä verrataan esseetehtävän aihealuetta koskevaan oppikirjan osaan. Valmennusmateriaalin esseistä analysoidaan niiden yhdenmukaisuus oppimateriaaliin ja etsitään raja-arvot kutakin arvosanakategori-

aa vastaaville yhdenmukaisuusarvoille. Esseen arviointi suoritetaan vertaamalla sen samankaltaisuutta oppimateriaaliin ja asettamalla esseelle arvosana, johon se samankaltaisuusarvonsa mukaan kuuluu. Kuvassa 11 on esitetty esimerkki menetelmän toiminnasta.



Kuva 11. Oppimateriaalin ja arvioitavan esseen vertailuun perustuva malli. Kuvan esimerkissä vertailumateriaalina käytetään oppikirjan kuutta kappaletta.

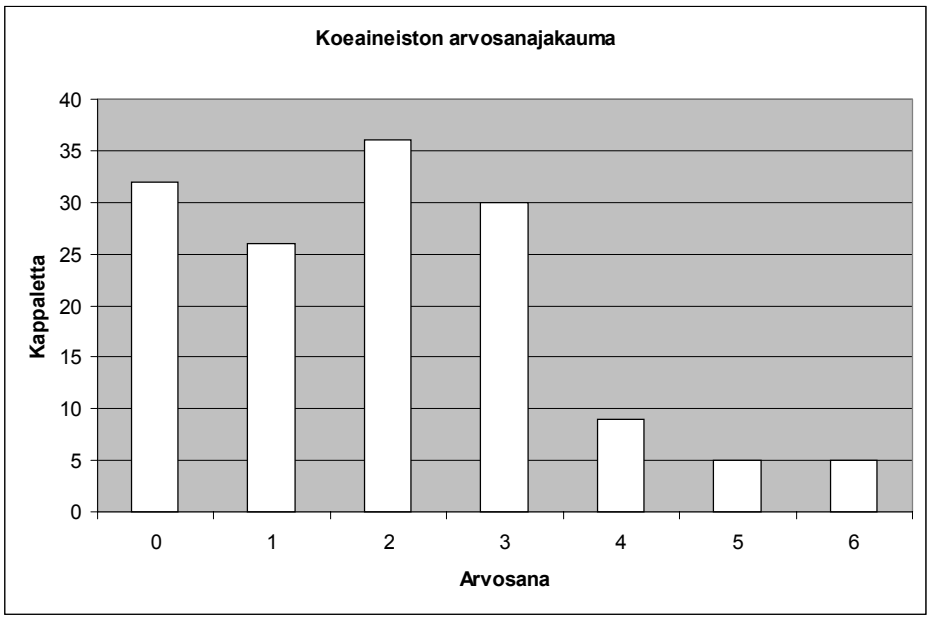
Latentin semanttisen analyysin toimivuuden kannalta optimaalisen dimension löytäminen semanttista avaruutta muodostettaessa on tärkeää. Dimension valinta tapahtuu toistamalla analysointi eri dimensiovaihtoehdoilla ja valitsemalla se dimensio, jolla saavutetaan paras yhdenmukaisuus arvioijien ja järjestelmän antamien arvosanojen välillä. Kuva 12 esittää järjestelmän toiminnan vaiheet UML-tilakaaviona oppikirjan katkelmiin vertaavaa mallia käytettäessä.



Kuva 12. Tilakaavio järjestelmän toiminnasta oppikirjan katkelmiin vertailevaa mallia käytettäessä.

5.3 Tulokset

Järjestelmää testattiin aineistolla, joka koostui 143 kasvatustiedettä käsittelevästä esseevastauksesta, joiden arvioinnin oli suorittanut yksi arvioija arvosana-asteikolla 0-6. Kuva 13 esittää aineiston arvosanajakauman.



Kuva 13. Koeaineiston arvosanajakauma.

Vastausten pituutta ei ollut tehtävänannossa rajattu ja esseiden pituus vaihteli aineistossa 18 ja 445 sanan välillä. Esseet jaettiin kahteen koeaineistoon satunnaisesti, joista toisessa oli 70 valmennusmateriaalin esseetä sekä 73 arvioitavaa esseetä ja toisessa 86 valmennusmateriaalin esseetä ja 57 arvioitavaa esseetä. Oppimateriaaliin perustuvassa mallissa valmennusmateriaalina käytettiin oppikirjan katkelmaa, joka käsittelee esseetehtävän aihealuetta. Oppikirjan katkelma oli 2397 sanan mittainen. Arviointiprosessia testattiin siten, että oppikirjan katkelma jaettiin erikseen sekä viiteen lukuun, 27 kappaleeseen että 147 lauseeseen. K:n lähimmän valmennusmateriaalin esseen vertailuun perustuvassa mallissa arviointi suoritettiin *k*:n arvoilla 2…7.

Järjestelmää testattiin käyttäen kolmea erilaista sulkusanalistaa sekä kahta eri painotusmallia. Painotusmalleina käytettiin kaavan 6 mukaista entropiaan perustuvaa mallia sekä pelkästään sanojen frekvensseihin perustuvaa mallia. Sulkulistana käytettiin 338-sanaista van Rijsbergenin (1979) englanninkielisen sulkusanalistan pohjalta laadittua listaa sekä 27 sanasta koostuvaa listaa. 27-sanainen sulkulista koostui suomen kielen pronomineista sekä eräistä yleisimmistä sanoista, kuten *ei*, *ja tai*, *jos*, *kuin*. Lisäksi järjestelmää testattiin ilman sulkusanojen poistoa.

Taulukossa 9 on esitetty osa koetuloksista, jotka on tuotettu käyttämällä erilaisia painotusmalleja, sulkusanalistoja sekä k:n lähimmän esseen ja oppimateriaalin vertailuun perustuvia malleja.

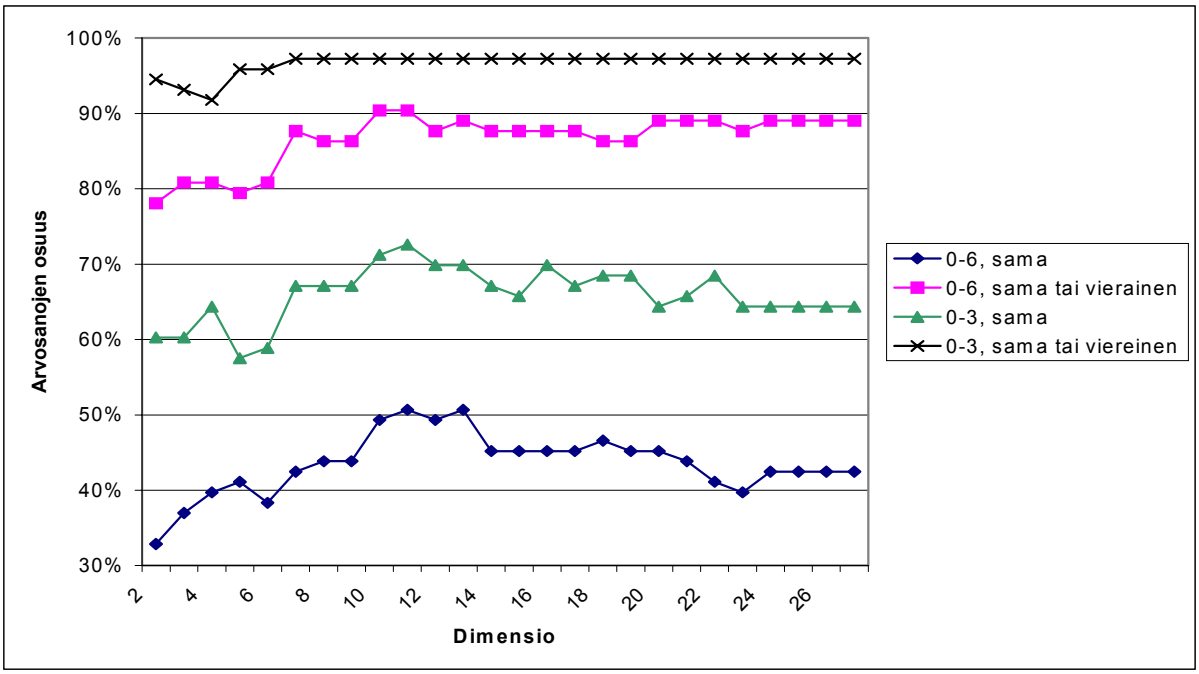
Taulukko 9. Tulokset.

Valmennusmateriaali	Sulkulistan koko	Painotusmalli	Arvioitavia esseitä	Dimensiot		Tulokset					
						0-3			0-6		
				Max	Optimi	Sama	Sama tai viereinen	Korrelaatio	Sama	Sama tai viereinen	Korrelaatio
Oppikirjan viisi lukua ja 70 esseetä.	27	Entropia	73	5	3	65,8	97,3	0,73	47,9	86,3	0,79
Oppikirjan 27 kappaletta ja 70 esseetä.	27	Entropia	73	28	11	72,6	97,3	0,78	50,7	90,4	0,80
Oppikirjan 147 lausetta ja 70 esseetä.	27	Entropia	73	147	102	63,0	97,3	0,72	46,6	90,4	0,80
Oppikirjan viisi lukua ja 70 esseetä.	0	Entropia	73	5	4	63,0	97,3	0,72	50,7	83,6	0,79
Oppikirjan viisi lukua ja 70 esseetä.	338	Entropia	73	5	4	60,3	95,9	0,68	38,4	80,8	0,73
Oppikirjan viisi lukua ja 70 esseetä.	27	Ei painotusta	73	5	4	46,6	86,3	0,53	32,9	65,8	0,59
Oppikirjan viisi lukua ja 86 esseetä.	0	Entropia	57	5	4	63,2	96,5	0,77	45,6	82,5	0,82
Oppikirjan viisi lukua ja 86 esseetä.	27	Entropia	57	5	4	63,2	96,5	0,77	47,4	84,2	0,82
Oppikirjan viisi lukua ja 86 esseetä.	338	Entropia	57	5	3	61,4	94,7	0,74	50,9	82,5	0,81
86 esseetä. K:n arvo 5	338	Entropia	57	86	25	47,4	93,0	0,27	33,3	78,9	0,35
86 esseetä. K:n arvo 5	27	Entropia	57	86	23	50,9	93,0	0,39	28,1	75,4	0,34
86 esseetä. K:n arvo 6	0	Entropia	57	86	24	47,4	93,0	0,31	29,8	77,2	0,49

Valmennusmateriaali-sarakkeessa on kuvattu valmennusmateriaalin rakenne ja sulkulistan koko -sarake kertoo käytetyn sulkulistan koon. Arvioitavien esseiden lukumäärä sekä käytetty painotusmalli on kuvattu omissa sarakkeissaan. Dimensiot-sarakkeessa max esittää suurimman mahdollisen dimension latentissa semanttisessa analyysissä käytetyllä valmennusmateriaalilla. Optimi-sarake ilmaisee dimension, joka tuotti parhaan arviointitarkkuuden. Tulokset

on jaettu kahteen osaan siten, että 0-6 merkityt sarakkeet kertovat järjestelmän tarkkuudesta käytettäessä kaikkia aineiston seitsemää arvosanaluokkaa ja 0-3 -sarakkeet tarkkuutta, kun arvosanaluokat 1-2, 3-4 ja 5-6 on yhdistetty. Yhdistettyjä arvosanaluokkia käytettiin, koska testiaineiston arvosanajakauma oli sellainen, että varsinkin arvosanakategorioissa 4-6 oli hyvin vähän esseitä. Sama tai viereinen -sarake kuvaa sen prosenttiosuuden esseistä, joille järjestelmä antoi saman tai viereisen arvosanan kuin arvioija. Sama -sarake kertoo vastaavasti prosenttiosuuden esseistä, joille järjestelmä antoi täysin saman arvosanan kuin arvioija. Korrelaatio-sarakkeessa on esitetty järjestelmän ja arvioijan antamien arvosanojen välinen Spearmanin korrelaatiokerroin.

Kuvassa 14 on esitetty esimerkki dimensioiden valinnan vaikutuksesta arviointitarkkuuteen aineistolla, joka sisältää 73 arvioitavaa esseetä. Analyysissa suoritettiin käyttäen 27-sanaista sulkusanalistaa ja entropiaan perustuvaa painotusmallia.



Kuva 14. Arvioinnin tarkkuus eri dimensioilla oppikirjan kappaleisiin verrattaessa käyttäen 27 sanan sulkulistaa. Arvioitavana 73 esseetä.

Kuvan 14 tapauksessa valmennusmateriaali koostui oppikirjan katkelmasta, joka oli jaettu 27 tekstikappaleeseen. Tällöin latentti semanttinen analyysi voitiin suorittaa dimensiolla 2…27. Parhaan arviointitarkkuuden tuotti dimensio 11.

5.4 Yhteenveto

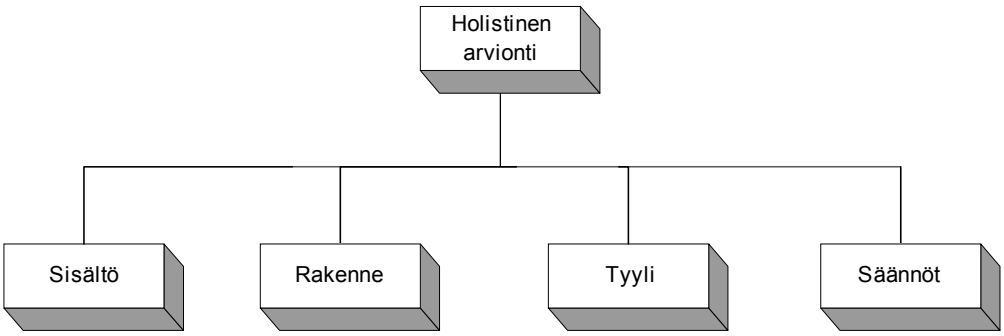
Tarkimpia tuloksia vertailluista malleista tuotti oppimateriaaliin vertaamiseen perustuva malli, jonka tuottamat arvosanat olivat huomattavasti k:n lähimmän esseen mallia tarkempia. Käytettäessä arvosana-asteikkoa 0-6 arvioijan ja järjestelmän antamien arvosanojen korrelaatio oli parhaimmillaan 0,78...0,82. Neljän arvosanan asteikolla käytettäessä 27-sanaista sulkulistaa korrelaatio oli 0,73…0,78. Tulokset ovat vertailukelpoisia tutkielmassa esitellyillä menetelmillä saavutettujen tulosten kanssa (vrt. taulukko 7, s. 46).

Mallit, joissa oppikirjan katkelma jaettiin vertailua varten lukuihin ja kappaleisiin olivat tarkempia kuin lauseisiin jakava malli. Lukuihin ja kappaleisiin jakavat mallit eivät eronneet suorituskyvyltään kovin merkittävästi. Painotusmalleista entropiaan perustuva painotus oli selkeästi tarkempi kuin painottamaton, sanojen frekvensseihin perustuva malli. 27 sanan sulkulistan käyttäminen tuotti tarkempia tuloksia kuin 338 sanan sulkusanalistan käyttö. Arvioinnin suorittaminen ilman sulkusanalistaa osoittautui tarkemmaksi kuin 338-sanaisen listan käyttö, mutta epätarkemmaksi kuin 27 sanan listan käyttö. K:n lähimmän esseen arvosanojen keskiarvoon perustuvassa mallissa *k*:n arvot 5 ja 6 tuottivat suurimman korrelaation arvioijan ja järjestelmän arvosanojen välillä. Sulkusanalistojen vaikutus ei ollut tässä mallissa yhtä ilmeinen kuin oppikirjan katkelmiin verrattaessa. 27-sanainen sulkusanalista tuotti suurimman korrelaation (0,39) neljän arvosanan asteikolla, mutta seitsemän arvosanan asteikkoa käytettäessä ilman sulkusanojen poistoa suoritettu arviointi tuotti tarkimman tuloksen (korrelaatio 0,49).

Syynä oppikirjaan vertailevan mallin parempaan toimivuuteen k:n lähimmän esseen malliin verrattuna koeaineistolla johtunee siitä, että tehtävä oli luonteeltaan sen kaltainen, että se soveltui oppikirjaan vertailuun hyvin. Tehtävänannossa oli viitattu tietyn oppikirjan lukuun, jolloin sopivan katkelman määrittely valmennusmateriaalia varten oli yksinkertaista. K:n lähimmän esseen arvosanojen keskiarvoon perustuvan mallin heikommalle toimivuudelle koeaineistossa voi syynä olla se, että aineiston esseiden pituus vaihteli varsin runsaasti. Tutkittavien dokumenttien pituus vaikuttaa latentin semanttisen analyysin kykyyn tunnistaa niiden samankaltaisuutta. Sovellettaessa LSA:ta esseiden arviointiin on aiemmissa tutkimuksissa usein käytetty esseitä, joiden tehtävänannossa vastauksen pituudeksi on määritelty tietty tavoitesanamäärä esim. 250 sanaa (mm. Landauer et al. 1997). Lisäksi käytettävissä ollut valmennusmateriaali oli suhteellisen pieni, toisessa aineistossa 86 ja toisessa 70 esseetä.

Varsinkin ylemmissä arvosanaluokissa oli vähän valmennusmateriaalin esseitä, joka aiheutti virhettä erityisesti niihin kuuluvia esseitä arvioitaessa.

Tutkielman osana toteutettu järjestelmä perustuu esseen sisällön arviointiin, eikä se huomio esimerkiksi rakenteeseen liittyviä seikkoja. Puhtaasti asiasisällön holistisesta arvioinnista järjestelmä suoriutuu testitulosten valossa varsin hyvin. Jotta järjestelmä kykenisi antamaan arvosanojen lisäksi yksityiskohtaisempaa palautetta, olisi sen kyettävä arvioimaan esseestä sisällön lisäksi muitakin osa-alueita. Kuva 15 esittää näkemykseni kokonaisesta, kaikki esseen tärkeimmät osa-alueet huomioivasta arviointijärjestelmästä.



Kuva 15. Kokonaisen arviointijärjestelmän rakenne.

Rakenne perustuu International Study of Written Compostionin -tutkimuksen tuloksiin (Blok ja de Glopper 1992). Siinä kirjoitettujen tehtävien arviointi jaetaan viiteen osaan: yleinen laatu, sisältö, rakenne, tyyli ja sääntöjen noudattaminen (conventions). Yleinen laatu perustuu arvioijan yleiskuvaan tekstin laadusta. Sisällön arvioinnissa kiinnitetään huomiota siihen, onko tekstissä esitetty vaadittu tiedollinen sisältö. Rakenteen tarkastelussa tutkitaan kirjoituksen koostumusta ja selkeyttä. Tyyli pitää sisällään lauseopillisen ja sanastollisen vaihtelun sekä sanojen ja ilmausten valinnan. Sääntöjen noudattamisen selvittämiseksi tekstistä tutkitaan sen kieliopillista oikeellisuutta sekä oikeinkirjoitusta.

Nykyiset menetelmät suoriutuvat parhaiten sisällön ja sääntöjen noudattamisen arvioinnista sekä näiden pohjalta luodun yleisarvion luomisesta. Myös rakenteen tarkasteluun on kehitetty varsin toimivia menetelmiä. Useimmat niistä on kehitetty kuitenkin vain englanninkielisin tekstin rakenteen tulkintaan, eikä niiden avulla ylletä täysin oikeelliseen tulokseen. Heikoimmin nykymenetelmät soveltuvat kirjoitustyylin arviointiin. Tyylin mittaaminen ja pisteytys on

käsitteellisestikin hieman hankalaa, eroavathan ihmistenkin käsitykset tyylin osalta varsin merkittävästi toisistaan. Eräs tyylin mittaamiseen soveltuva menetelmä on esseen koherenssin tutkiminen, jonka tutkimuksissa on saavutettu varsin lupaavia tuloksia (Miltsakaki ja Kukich 2000).

6. YHTEENVETO

Tässä tutkielmassa on perehdytty automaattisen esseen arvioinnin lähtökohtiin, menetelmiin ja saavutettuihin tuloksiin. 1960-luvulla alkaneissa tutkimuksissa on päästy yksinkertaisten ominaisuuksien, kuten esseen pituuden, mittaamisesta yhä lähemmäs ihmisten käyttämiä arviointikriteerejä. Tämä on tärkeä kehityssuunta, ei ainoastaan sen avulla saavutettavan paremman arvioinnin tarkkuuden, vaan myös automaattisen arvioinnin yleisen hyväksyttävyyden kannalta. Jatkuvasta menetelmien kehittymisestä ja rohkaisevista tuloksista huolimatta järjestelmien laajamittaiseen käyttöönottoon ja hyväksyntään ei ole vielä ylletty.

Kuten tutkielmassa on esitetty, on esseiden arvioinnin automatisointiin kohdistettu myös varsin voimakasta kritiikkiä. Kritiikkiin on pyritty vastaamaan kehittämällä yhä parempia ja toimivampia järjestelmiä. Tutkimustulokset ovat osoittaneet varsin selkeästi, että kehitetyillä menetelmillä saavutetaan tuloksia, jotka mahdollistavat niiden käytön arvioinnin apuvälineenä.

Tutkielmassa esitettyjen seikkojen pohjalta olen muodostanut käsitykseni johdannossa esittämiini tutkimuskysymyksiin. Tuntien nykyisten menetelmien rajoitukset on todettava, ettei nykyisiä automaattisen arvioinnin menetelmiä voida pitää puhtaasti vaihtoehtona arvioijien käytölle ainakaan korkean oikeellisuusvaatimuksen omaavissa esseiden arviointitehtävissä, kuten tenttien tai pääsykokeiden arvioinnissa. Arvioijan tukena tai osana useamman arvioijan raatia ne kuitenkin ovat varsin luotettava ja toimiva ratkaisu. Nykyisiä järjestelmiä voidaan käyttää, ja tällä hetkellä jo käytetäänkin, ainoana arvioijana erilaisissa harjoitustehtävissä ja vaikkapa verkko-opetuskursseilla.

Menetelmien käyttö vaikuttaa soveltuvan parhaiten kirjoitustaidon arviointiin eli arviointiin jossa yhdistyvät sisällön, rakenteen ja tyylin arviointi. Tietokone soveltuu erityisen hyvin kieliopin, oikeinkirjoituksen ja välimerkkien käytön kaltaisten ominaisuuksien mittaamiseen. Sisällön, tyylin ja rakenteen arvioinnin tarkkuutta rajaa se tosiseikka, ettei nykytietämyksen avulla ole mahdollista kehittää tietokoneistettua menetelmää, joka ”ymmärtää” tekstiä täysin. Tästä huolimatta myös näiden ominaisuuksien arviointi onnistuu nykyisillä menetelmillä yllättävänkin luotettavasti.

Tärkeimpiä kehityssuuntia jatkotutkimuksessa tulevat olemaan yhä suuremman arvioinnin tarkkuuden saavuttaminen sekä yksityiskohtaisemman ja sanallisen palautteen antamiseen pelkän pisteytyksen sijaan kykenevien järjestelmien kehittäminen. Lisäksi pyritään kehittämään menetelmiä, jotka mahdollistavat poikkeuksellisten, joko järjestelmän harhauttamiseksi tai tahattomasti kirjoitettujen, esseiden tunnistamisen ja luotettavamman arvioinnin.

Lähdeluettelo

Abney, S.: Part-of-Speech Tagging and Partial Parsing. Teoksessa: Young, S., Bloothooft, G.: *Corpus-Based Methods in Language and Speech Processing.* Kluwer, Dordrecht, Netherlands, 1996.

Alkula, R.: *Merkkijonoista suomen kielen sanoiksi*. Tampereen yliopisto, 2000.

Baeza-Yates, R., Ribeiro-Neto, B.: *Modern information retrieval.* Addison-Wesley, 1999.

Bayes, T.: Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*. London, England, 1764.

Benford, S., Burke, E., Foxley, E., Gutteridge, N., Mohd Zin, A.: Ceilidh: A Course Administration and Marking System. *Proceedings of the International Cenference on Computer Based Learning*. Vienna, Austria, 1993.

Berry, M., Do, T., O'Brien, G. Krishna, V., Varadhan, S.: *SVDPACKC (Version 1.0) User's Guide.* 1993. Internet WWW-sivu, URL: http://www.netlib.org/svdpack/ (26.9.2002).

Blok, H., de Glopper, K.: Large-scale writing assessment. Teoksessa: Verhoeven L., De Jong J. H. A. L.: *The Construct of Language Proficiency: Applications of Psychological Models to Language Assessment*. John Benjamins Publishing Company, Amsterdam, Netherlands, 1992.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M.: Enriching Automated Essay Scoring Using Discourse Marking. *Proceedings of the Workshop on Discourse Relations & Discourse Marking.* Montreal, Canada, 1998a.

Burstein, J., Kukich, K., Wolff, S., Lu C., Chodorow, M., Braden-Harder, L., Harris, M. D.: Automated Scoring Using A Hybrid Feature Identification Technique. *Proceedings of the Annunal Meeting of the Association of Computational Linguistics.* Montreal, Canada, 1998b.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M.: *Computer Analysis of Essays*. Paper presented at annual meeting of the National Council on Measurement in Education. San

Diego, USA, 1998c. Internet WWW-sivu, URL: http://www.ets.org/research/dload/ncmefinal.pdf (18.7.2002).

Burstein, J., Chodorow, M.: Automated Essay Scoring for Nonnative English Speakers. *Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies.* Maryland, USA, 1999.

Burstein, J., Marcu, D.: Towards Using Text Summarization for Essay-Based Feedback. *Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000,* Lausanne, Switzerland, 2000a.

Burstein, J., Marcu, D.: Benefits of Modularity in an Automated Essay Scoring System. *The Coling-2000 Workshop on Using Toolsets and Architectures to Build NLP Systems.* Luxembourg, 2000b.

Burstein, J., Chodorow, M.: Directions in Automated Essay Analysis. Teoksessa: *The Oxford Handbook of Applied Linguistics*. (ed. Kaplan, R.B). Oxford University Press, New York, USA, 2002.

CAA Centre: *Frequently Asked Questions*. Internet WWW-sivu, URL: http://www.caacentre.ac.uk/resources/faqs/index.shtml (1.10.2002). 2002.

Callan, J. P., Croft, W. B., Broglio, J: TREC and Tipster Experiments with Inquery. *Information Processing and Management,* **31**(3), 327-343, 1995.

Chodorow, M., Leacock, C.: An Unsupervised Method for Detecting Grammatical Errors. *Proceedings of First Meeting of the North American Chapter of the Association for Computational Linguistics.* Morgan Kaufmann, San Francisco, USA, 2000.

Chung, G. K. W. K., O'Neil, H. F.: *Methodological Approaches to Online Scoring of Essays*. *CSE Technical Report 461*. National Center for Research on Evaluation, Los Angeles, USA, 1997.

Conexor Oy: *Conexor Functional Dependency Grammar 3.7 – User's Manual*, 2002.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science,* **41**(6), 391-407, 1990.

ETS Technologies: *ETS Technologies, Inc. and EdVISION Announce Partnership.* Lehdistötiedote. Internet WWW-sivu, URL: http://www.etstechnologies.com/EdVISION.pdf (26.6.2002).

Foltz, P. W., Britt, M. A., Perfetti, C. A. : Reasoning from multiple texts: An automatic analysis of readers' situation models. *Proceedings of the 18th Annual Cognitive Science Conference*. Lawrence Erlbaum, USA, 1996.

Foltz, P. W., Kintsch, W., Landauer, T. K.: The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, **25**(2&3): 285-307, 1998.

Foltz, P. W., Gilliam, S. and Kendall, S.: Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments,* **8**(2): 111-129, 2000.

Foxley, E., Higgins, C., Hagazy, T., Symeonidis, P., Tsintsifas, A.: The CourseMaster CBA System: Improvements over Ceilidh. *Proceedings of the Fifth Annual Computer Assisted Assessment Conference*. Loughborough University, England, 2001.

Hearst, M. et al.: The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, *Trends & Controversies feature*, **15**(5): 22-37, 2000.

Herrington, A., Moran, C.: What Happens When Machines Read Our Student's Writing? *College English*, **63**(4):480-499, 2001.

Hopkins, K. D., Stanley, J. C., Hopkins, B. R.: *Educational and Psychological Measurement, Seventh Edition.* Prentice-Hall, Englewood Cliffs, USA, 1990.

Joy, M. S., Luck, M.: The BOSS System for On-line Submission and Assessment of Computing Assignments. Teoksessa: Charman, D. D., Elmes, A.: *Computer Based Assessment*

(Volume 2): Case studies in Science & Computing. SEED Publications, University of Plymouth, England, 1998.

Kaplan, R. M., Wolff, S., Burstein, J., Li, C., Rock, D., Kaplan B.: *Scoring Essays Automatically Using Surface Features (GRE Board Professional Report No. 94-21P)*. Educational Testing Service, New Jersey, USA, 1998.

Landauer, T. K., Laham, D., Rehder, B., Schreiner, M. E.: How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. *Proceedings of the 19th annual meeting of the Cognitive Science Society.* Mawhwah, New Jersey, USA, 1997.

Landauer, T. K., Foltz, P. W., Laham, D.: An Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**(2&3): 259-284, 1998a.

Landauer, T. K., Laham, D., Foltz, P. W.: Learning human-like knowledge by Singular Value Decomposition: A progress report. Teoksessa: Jordan, M. I., Kearns, M. J., Solla, S. A.: *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, USA, 1998b.

Landauer, T. K., Psotka, J.: Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments*, **8**(2), 2000.

Landauer, T. K.: On the computational basis of learning and cognition: Arguments from LSA. Teoksessa: Ross, N.: *The psychology of learning and motivation 41.* Academic Press, San Diego, USA, 2002.

Larkey, L.: Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of 21st Annual International Conference on Research and Development in Information Retrieval.* Melbourne, Australia, 1998.

Lewis, D. D.: *Representation and Learning in Information Retrieval*. PhD thesis, University of Massachusetts, USA, 1992.

Malmi, L.: Automaattinen tarkastaminen opetuksen apuvälineenä. *Tietojenkäsittelytiede,* 17, 2002.

Marcu, D: Discourse trees are good indicators of importance in text. Teoksessa: Mani, I., Maybury M.: *Advances in Automatic Text Summarization*. The MIT Press, USA, 1999.

McCallum, A., Nigam, K.: *A comparison of event models for Naive Bayes text classification*. In AAAI-98 Workshop on Learning for Text Categorization, 1998. Internet WWW-sivu, URL: http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf (26.6.2002).

Miltsakaki, E., Kukich, K.: Automated Evaluation of Coherence in Student Essays. *Proceedings of the Workshop on Language Resources and Tools in Educational Applications*, LREC 2000. Athens, Greece, 2000.

Myllymäki, P., Tirri, H.: *Bayes-verkkojen mahdollisuudet*. Teknologian kehittämiskeskus (TEKES), 1998.

Page, E. B.: The imminence of grading essays by computer. *Phi Delta Kappan*, **47**(1): 238-243, 1966.

Page, E. B.: The use of the computer in analyzing student essays. *International Review of Education*, **14**(1): 253-263, 1968.

Page, E. B.: New Computer Grading of Student Prose, Using Modern Concepts and Software. *Journal of Experimental Education*, **62**(2): 127-142, 1994.

Page, E. B., Petersen, N. S.: The computer moves into essay grading. *Phi Delta Kappan*, **76**(7): 561-565, 1995.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., Kukich, K.: *Comparing the Validity of Automated and Human Essay Scoring (GRE Board Professional Report No. 98-08aR)*. Educational Testing Service, New Jersey, USA, 2000.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., Kukich, K.: Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior,* **18**(2): 103-134, 2002.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., Kintsch, W.: Using Latent Semantic Analysis to Assess Knowledge: Some Technical Considerations. *Discourse Processes*, **25**(2&3): 337-354, 1998.

van Rijsbergen, C. J.: *Information Retrieval*. Butterworth, London, England, 1979.

Rudner, L. M.: Reducing Errors Due to the Use of Judges. *Practical Assessment, Research & Evaluation*, **3**(3), 1992.

Rudner, L. M., Liang, T.: Automated Essay Scoring Using Bayes' theorem. *Journal of Technology, Learning and Assessment*, **1**(2), 2002.

Rudner, L. M. : *Bayesian Essay Testing sYstem - BETSY*. 2002. Internet WWW-sivu, URL: http://ericae.net/betsy/ (26.06.2002).

Salton, G.: *Automatic Text Processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, Reading, USA, 1989.

Shermis, M. D., Mzumara, H. R., Olson, J., Harrington, S.: On-line Grading of Student Essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education,* **26**(3): 247-259, 2001.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., Harrington, S.: Trait Ratings for Automated Essay Grading. *Educational and Psychological Measurement*, **62**(1): 5-18, 2002.

Takala, S.: Arviointi - ongelma ja mahdollisuus. Teoksessa: Linnakylä, P., Pollari, P., Takala, S.: *Portfolio arvioinnin ja oppimisen tukena*. Jyväskylän yliopisto, 1994.

Tapanainen, P.: *Parsing in two frameworks: finite-state and functional dependency grammar*. PhD thesis, Department of General Linguistics, University of Helsinki, Finland, 1999.

Thorndike, R., L.: *Educational measurement (2nd ed.)*. American Council of Education, Washington, D. C., USA, 1971.

Wresch, W.: The Imminence of Grading Essays by Computer – 25 Years Later. *Computers and Composition*, **10**(2): 45-58, 1993.

Yang, Y., Pedersen, J. O.: A Comparative Study on Feature Selection in Text Categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*. Nashville, Tennessee, USA, 1997.