Indeksointimetatiedon eristäminen ja arviointi

Mika Wahlroos

Helsinki 17.2.2013

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Mika Wahlroos

Työn nimi — Arbetets titel — Title

Indeksointimetatiedon eristäminen ja arviointi

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	17.2.2013	59 sivua + 0 liitesivua

Tiivistelmä — Referat — Abstract

Tiedonhallinnassa käytetään usein metatietona tiedon sisältöä kuvaavia avainsanoja parantamaan tiedon hallittavuutta tai löydettävyyttä. Sisällön kuvailua luonnollisen kielen termein tai käsittein kutsutaan indeksoinniksi. Yhdenmukaisuuden vuoksi voidaan käyttää tarkoitusta varten laadittua asiasanastoa, joka kattaa toimialan kannalta keskeisen termistön. Semanttisessa webissä ja yhdistetyssä tiedossa käytettävät ontologiat vievät ajatuksen pitemmälle määrittelemällä termit käsitteinä ja niiden välisinä merkityssuhteina.

Metatiedon tuottamisen helpottamiseksi ja tehostamiseksi on kehitetty erilaisia menetelmiä, joilla sisältöä kuvailevia termejä voidaan tuottaa tekstiaineistosta automaattisesti. Tässä tutkielmassa keskitytään avaintermien automaattiseen eristämiseen tekstistä sekä metatiedon laatuun ja sen arvioinnin menetelmiin. Esimerkkitapauksena käsitellään ontologiaa hyödyntävän Maui-indeksointityökalun käyttöä asiakirjallisen tiedon automaattiseen asiasanoittamiseen.

Automaattisesti eristetyn metatiedon laatua verrataan alkuperäiseen ihmisten määrittämään asiasanoitukseen käyttäen tarkkuus- ja saantimittauksia. Lisäksi evaluointia täydennetään aihealueen asiantuntijoiden esittämillä subjektiivisilla laatuarvioilla. Tulosten perusteella selvitetään tekstin esikäsittelyn ja sanaston hierarkian merkitystä automaattisen asiasanoituksen laadun kannalta sekä pohditaan keinoja annotointimenetelmän jatkokehittämiseksi.

ACM Computing Classification System (CCS):
H.3.1 [Content Analysis and Indexing]: Indexing methods
I.2.4 [Knowledge Representation Formalisms and Methods]: Semantic networks
I.2.6 [Learning]: Induction

Avainsanat — Nyckelord — Keywords
indeksointi, avainsanojen eristäminen, arviointimenetelmät, semanttinen web

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Tiedonhallinnassa aineistoon, esimerkiksi asiakirjoihin tai kuviin, yhdistetään usein metatietona sisältöä kuvaavia avain- tai asiasanoja. Sisällön kuvailua luonnollisen kielen termein kutsutaan *indeksoinniksi* (engl. subject indexing). Eräänä indeksoinnin muotona voi pitää esimerkiksi webissä ja etenkin sosiaalisessa mediassa käytettäviä vapaamuotoisia avainsanoja (tag), joilla käyttäjät voivat mielensä mukaan merkitä ja kuvailla vaikkapa musiikkikappaleita tai uutisia. Kirjastojen tietojärjestelmissä teoksiin on liitetty esimerkiksi teemoja ja aiheita kuvaavia asiasanoja. Indeksointi sisällön kuvailuna poikkeaa esimerkiksi tiedonhaussa hyödynnettävistä kokotekstihakemistoista ja niiden muodostamiseen käytetyistä algoritmeista.

Avain- ja asiasanoja voidaan käyttää muun muassa tiedon luokitteluun, osana tiivistelmien automaattista laadintaa [BC00, DM05] sekä rajauskriteerinä moninäkymähaussa [Pol98, HSV04]. Sisältöä hyvin kuvaavat avaintermit voivat myös auttaa tietoa hakevaa käyttäjää erottamaan tiedontarpeeseensa vastaavat tulokset nopeammin epäolennaisista. Semanttisessa webissä [BHL01, SHB06] ja yhdistetyssä tiedossa (linked data) [BHB09] metatietoa käytetään liittämään tieto osaksi laajempaa yhdistetyn tiedon muodostamaa verkostoa. Sisällön kuvailua metatiedolla kutsutaan myös *annotoinniksi*.

Informaatiotutkimuksen alalla on kiinnitetty huomiota etenkin metatiedon yhdenmukaisuuden merkitykseen laadun kriteerinä [Rol81, Par09]. Yhdenmukaisuuden parantamiseksi voidaan käyttää tarkoitusta varten laadittua *kontrolloitua sanastoa* (controlled vocabulary), joka kattaa käsiteltävän aihealueen kannalta keskeisen termistön. Näin voidaan välttää esimerkiksi asiasanan esiintyminen eri taivutusmuodoissa tai usean samaa tarkoittavan termin epäjohdonmukainen käyttö [Gar04]. Aineiston indeksointia ennalta määritellyn asiasanaston termein kutsutaan *asiasanoitukseksi*.

Semanttisessa webissä ja yhdistetyssä tiedossa käytettävissä *ontologioissa* (ontology) termistö määritellään käsitteinä ja niiden keskinäisinä merkityssuhteina. Käsitteet ja muut tietoalkiot esitetään tekstiliteraalien sijaan *resursseina* (resource), jotka voivat viitata toisiinsa yksikäsitteisesti muodostaen näin yhdessä *semanttisen verkon* (semantic network). Semanttisella metatiedolla tarkoitetaan tässä yhteydessä varsinaiseen sisältöön liitettyjä annotaatioita, jotka viittaavat ontologisiin käsitteisiin tai muihin semanttisen verkon resursseihin. Tässä tutkielmassa keskitytään metatiedon eri muodoista erityisesti ontologioissa määriteltyjä luonnollisen kielen termejä

käyttävään asiasanoitukseen.

Huolellinen indeksointi ihmisvoimin on työlästä ja aikaavievää. Suljetuissa järjestelmissä, joissa laaditaan ja käsitellään esimerkiksi virallisia asiakirjoja, ei yleensä myöskään ole käytettävissä webin suuren yleisön kaltaista voimavaraa valjastettavaksi tiedon kuvailuun ja luokitteluun. Metatiedon tuottamiseksi koneellisesti tekstistä onkin kehitetty erilaisia automaattisia ja puoliautomaattisia menetelmiä [Tur99, FPW+99, Seb02].

Tässä tutkielmassa esitetään katsaus rakenteettomien tekstiaineistojen automaattisessa indeksoinnissa käytettyihin algoritmeihin sekä menetelmiin, joilla tuotetun metatiedon laatua on arvioitu. Yleiskatsauksen lisäksi esitellään tarkemmin Waikaton yliopistossa kehitetyn ontologiaa hyödyntävän Maui-indeksointityökalun [MFW09, Med09] käyttöä asiakirjallisen tiedon annotointiin ja verrataan sen tuottaman asiasanoituksen laatua alan aiempaan tutkimukseen. Esimerkkinä käsitellään tapaustutkimusta, jossa puolustusvoimien asianhallintajärjestelmästä saatu otos asiakirjoja ja niiden metatietoja muunnettiin semanttisessa webissä käytettävään RDF-muotoon[1] ja asiasanoitettiin Mauia käyttäen. Erityisesti tarkastellaan automaattisesti eristetyn metatiedon laatua ja sen arviointiin soveltuvia menetelmiä sekä selvitetään mahdollisia laatuun ja kattavuuteen vaikuttavia tekijöitä. Tehdyn analyysin pohjalta etsitään keinoja automaattisen asiasanoituksen laadun parantamiseksi.

Automaattisen asiasanoituksen voi nähdä luokitteluongelmana, ja siihen käytettyjen algoritmien suorituskykyä onkin yleisesti arvioitu koneoppimisen ja tiedonhaun evaluoinnissa käytettävin menetelmin [Seb02]. Informaatiotutkimuksen näkökulmasta metatiedon laatu on kuitenkin monitahoinen ongelma, ja laatua voidaan arvioida laajalla kirjolla eri kriteereitä ja menetelmiä [Par09]. Tässä tutkielmassa pyritäänkin selvittämään erilaisia menetelmiä ja mittareita, joita automaattisesti tuotetun asiasanoituksen laadun arviointiin on käytetty. Erityisesti etsitään vastauksia seuraaviin tutkimuskysymyksiin:

1. Miten indeksointimetatietoa voidaan tuottaa automaattisesti?
2. Millaisin menetelmin metatiedon laatua voidaan arvioida?
3. Kuinka laadukasta Mauilla [MFW09, Med09] tapaustutkimuksen yhteydessä eristetty metatieto on ja miten se vertautuu aiempaan tutkimukseen?

[1] http://www.w3.org/RDF/

4. Miten tapaustutkimuksessa käytettyä annotointimenetelmää voitaisiin kehittää?

Mauin suorituskyvyn osalta havainnot tukevat jossain määrin aiempien tutkimusten tuloksia, joiden mukaan Mauin eristämien annotaatioiden laadun on todettu lähestyvän ihmistyönä tehdyn indeksoinnin tasoa [Med09, SSH11]. Etenkin kattavuuden osalta automaattisesti tuotettu metatieto jää kuitenkin aiempiin tuloksiin verrattuna puuttelliseksi, mikä selittyy osin aineiston erityispiirteillä. Sekä annotoinnin tarkkuuteen että sen kattavuuteen voidaan kuitenkin jossain määrin vaikuttaa pienehköillä indeksointimenetelmään tehtävillä muutoksilla.

Tutkielman luku 2 esittelee lyhyesti semanttisen webin perustekniikoita sekä sanastojen ja ontologioiden käyttöä. Luvussa 3 tehdään katsaus automaattisen annotoinnin menetelmiin yleisesti sekä tapaustutkimuksessa sovellettavan Mauin lähestymistapaan ja sen käyttämiin algoritmeihin, ja luku 4 käsittelee asiasanoituksen laatua ja sen arviointiin käytettäviä menetelmiä. Luvussa 5 esitellään puolustusvoimien asiakirjoja ja niiden metatietoja koskeva tapaustutkimus, jonka tuloksia esitellään ja analysoidaan luvussa 6. Luvussa 7 ehdotetaan mahdollisia jatkotutkimuksen kohteita ja luvussa 8 esitetään työn yhteenveto ja päätelmät.

2 Semanttisen webin tekniikat

Tässä luvussa esitellään lyhyesti semanttisen webin perustekniikoita ja ontologioiden käyttöä. Näihin keskitytään kuitenkin vain siinä laajuudessa kuin on tarpeen ontologioita hyödyntävien asiasanoitusalgoritmien ja arviointimenetelmien sekä tapaustutkimuksen ja sen tulosten analyysin ymmärtämiseksi. Laajemman yleiskuvan semanttisen webin ja yhdistetyn tiedon tekniikoista antavat esimerkiksi Bizer ja kumppanit [BHB09].

2.1 RDF-tietomalli

Semanttisessa webissä ja yhdistetyssä tiedossa metadata esitetään RDF-standardin (Resource Description Framework) mukaisessa muodossa. RDF on tietomalli, jossa yksittäiset tietoalkiot esitetään resursseina, joista kullakin on yksilöivä URI-tunniste (Uniform Resource Identifier). Tunnisteen avulla kuhunkin resurssiin voidaan viitata yksikäsitteisesti.

RDF-mallissa metatieto esitetään joukkona varsinaista tietoa kuvailevia väittämiä, jotka koostuvat subjektista (subject), predikaatista (property) ja objektista (object). Kolmikossa subjektilla tarkoitetaan kuvailtavaa tietoalkiota, predikaatilla jotakin sen ominaisuutta ja objektilla ominaisuuden arvoa. Arvona voi olla toinen resurssi tai tekstiliteraali. Tällaisilla kolmikoilla toisiinsa viittaavat resurssit ja niiden väliset suhteet muodostavat suunnatun graafin, jota kutsutaan semanttiseksi verkoksi. Verkon kaaret vastaavat resurssien keskinäisiä yhteyksiä koskevia väittämiä.

2.2 Ontologiat

Semanttisessa webissä ontologialla tarkoitetaan käsiteverkostoa, jossa tietyn sovellusalan tai aihealueen kannalta keskeiset käsitteet ja niiden väliset merkityssuhteet on esitetty rakenteellisessa, koneella käsiteltävässä muodossa. Yleisemmin muotoiltuna ontologia on käsitteistön muodollinen ja eksplisiittinen määrittely [Gru93].

Informaatiotutkimuksessa tesaurukseksi kutsutaan kontrolloitua sanastoa, joka esittää jonkin alan termistön lisäksi joukon käsitteiden välisiä merkityssuhteita, esimerkiksi keskinäistä hierarkiaa. Hierarkkisten suhteiden avulla voidaan määritellä esimerkiksi jonkin termin olevan toista termiä tarkentava ja merkitykseltään suppeampi (narrower term). Vastaavan tiedon voi ilmaista myös käänteisesti määrittelemällä jälkimmäisen käsitteen olevan edellistä laajempi (broader term). Keskenään hierarkkisissa suhteissa olevat käsitteet muodostavat siis kokonaisuutena puun kaltaisia rakenteita. Hierarkian lisäksi tesaurukset voivat sisältää termien välisiä ekvivalenssieli vastaavuussuhteita sekä assosiatiivisia suhteita (related term), joilla määritellään kahden käsitteen liittyvän merkitykseltään toisiinsa jotenkin muutoin kuin hierarkkisesti [Gar04].

Ontologiat laajentavat tesauruksen kaltaista mallia ilmaisemalla yksinkertaisen hierarkian lisäksi myös muunlaisia tai tarkemmin määriteltyjä yhteyksiä. Tesaurusten muotoa koskevat standardit määrittelevät ilmaistavissa olevien suhteiden joukon kiinteästi, kun taas ontologioissa erilaisia merkityssuhteita esittäviä uusia predikaatteja voidaan esitellä periaattessa vapaasti [Gar04].

Yleinen laajennos yksinkertaiseen laajempien ja suppeampien termien taksonomiaan nähden on esimerkiksi erotella toisistaan hyponymiset alakäsitesuhteet (subclass-of) ja meronymiset osakäsitesuhteet (part-of). Jälkimmäisillä voidaan määritellä jonkin käsitteen olevan toisen, laajempaa kokonaisuutta edustavan käsitteen looginen tai rakenteellinen osa [GW02, HVTS08]. Esimerkiksi kaupunki hallinnollisena alueena

on erääntyyppinen kunta, ja ontologiassa kaupunkia vastaavan käsitteen voitaisiin määritellä olevan kunnan käsitteen aliluokka. Kaupunginosa puolestaan ei ole kaupungin alakäsite vaan kaupunkiin sisältyvä osa.

Tällaisten yhteyksien täsmällisempi määrittely ja erottaminen alakäsitesuhteista on eräs keskeinen ero perinteisten tesaurusten ja ontologioiden välillä [HVTS08]. Tämän tutkielman aihepiirin kannalta riittää kuitenkin käsitellä ontologioita sanastoina, joissa käsitteet muodostavat keskenään hierarkkisia rakenteita.

Semanttisen webin ontologioiden esitystapana käytetään RDF-pohjaisia standardeja [SHB06]. Aineistojen uudelleenkäytettävyyden ja helpon yhdisteltävyyden vuoksi suhteiden ja väittämien esittämiseen on mahdollisuuksien mukaan syytä käyttää yleisessä käytössä olevia RDF-pohjaisia määrittelyjä. Kun eri aineistoissa käytetään yhteensopivia määritelmiä käsitteiden välisten suhteiden ilmaisuun, eri aineistojen ja niiden osien yhdisteleminen on periaatteessa helppoa.

Ontologioiden määrittelyyn voidaan käyttää esimerkiksi esimerkiksi World Wide Web Consortiumin standardoimaa OWL:ia[2] (Web Ontology Language). Toinen, erityisesti tesaurusten kaltaisten rakenteeltaan kevyiden asiasanastojen määrittelyyn sopiva W3C:n suositus on SKOS[3] [MMWB05] (Simple Knowledge Organization System).

Aihealuekohtaisten erikoisontologioiden lisäksi on kehitetty myös yleisontologioita, joissa on määritelty suuri joukko luonnollisen kielen käsitteitä semanttisine suhteineen. Tällainen yleisontologia on muun muassa Yleinen suomalainen ontologia YSO[4] [HVTS08], joka on kehitetty Yleisen suomalaisen asiasanaston[5] (YSA) pohjalta.

RDF-perustaisten aineistojen yhdisteltävyyden ansiosta yleisontologioita voidaan käyttää erikoisontologioiden pohjana, jolloin erikoisalan termistö voidaan rakentaa yleisluontoisen käsitteistön rinnalle ja sitä tarkentamaan [HVTS08]. YSO:n yhteyteen kehitetty erikoisontologia on esimerkiksi Puolustushallinnon ontologia PUHO[6], jota käytetään luvussa 5 esiteltävässä tapaustutkimuksessa.

[2] http://www.w3.org/TR/owl-features/
[3] http://www.w3.org/2004/02/skos/
[4] http://www.seco.tkk.fi/ontologies/yso/
[5] http://www.kansalliskirjasto.fi/kirjastoala/asiasanastot/ysa.html
[6] http://onki.fi/fi/browser/overview/puho

3 Automaattisen indeksoinnin menetelmät

Aineiston kuvailemisella erilaisin avaintermein tai asiasanoin on pitkät perinteet. Myös erilaisia menetelmiä tekstiaineiston luokittelun automatisoimiseksi on esitetty jo vuosikymmenten ajan. 1980-luvulle asti suosittiin usein sovelluskohtaisesti kehitettyjä sääntöpohjaisia luokittelujärjestelmiä, mutta sittemmin koneoppimista hyödyntävät menetelmät ovat syrjäyttäneet ne [Seb02].

Aineistoa kuvaavien keskeisten termien määrittämiseen tarkoitettuja algoritmeja voidaan jaotella sekä erilaisten algoritmisten lähestymistapojen että itse tehtävän luonteen mukaan. Tässä luvussa esitetään katsaus erilaisiin menetelmiin, joita rakenteettomien tekstidokumenttien automaattiseen indeksointiin on kehitetty.

3.1 Tehtävän määrittely ja lähestymistapa

Avain- tai asiasanojen tuottamista voidaan tarkastella eri näkökulmista riippuen siitä, käytetäänkö ennalta määriteltyä sanastoa vai ei. Sisällön kuvailuun käytettyjä mielivaltaisia luonnollisen kielen ilmauksia kutsutaan tässä yhteydessä *avainsanoiksi* ja kontrolloidusta sanastosta valittuja termejä puolestaan *asiasanoiksi*. Vastaavasti asiasanoituksella tarkoitetaan tässä yhteydessä nimenomaan sanastoa hyödyntävää indeksointia.

Tehtävän luonteen ohella toinen tapa jaotella automaattisessa indeksoinnissa käytettäviä menetelmiä on algoritmien erittely niiden toisistaan perustavanlaatuisesti poikkevien strategioiden perusteella. Indeksoinnissa käytettävät menetelmät voi karkeasti jakaa *tekstin luokitteluun* (text classification) [Seb02] ja *avainlausekkeiden eristämiseen* (keyphrase extraction) [FPW+99, Tur00, MFW09] perustuviin algoritmeihin. Tehtävän luonne vaikuttaa osaltaan siihen, millaisia algoritmisia lähestymistapoja ongelman ratkaisemiseen on luontevaa käyttää.

Kontrolloitua asiasanastoa käytettäessä indeksoinnin voi luontevasti nähdä luokitteluongelmana, jossa kukin annotoitava teksti pyritään liittämään joukkoon kategorioita. Luokittelun kategorioina toimivat sanaston termit, ja indeksointiin voidaan käyttää ohjattuun koneoppimiseen perustuvia luokittelualgoritmeja [Seb02]. Tällöin valmiiksi annotoidun opetusdatan perusteella rakennetaan tilastollinen malli, jonka avulla algoritmi pyrkii ennustamaan sille syötettäville uusille tekstidokumenteille sopivimmat asiasanat. Luokittelussa käytettävät piirteet perustuvat annotoitavan tekstin ominaisuuksiin, esimerkiksi siinä esiintyviin sanoihin.

Avainlausekkeiden eristämisellä puolestaan tarkoitetaan tekstin sisältöä kuvailevien sanojen tai lausekkeiden poimimista suoraan tekstistä. Tähän lähestymistapaan perustuvat algoritmit pyrkivät siis eristämään tekstissä sellaisenaan esiintyvät termit ja valikoimaan niistä sisällön kuvailun kannalta keskeisimmät [FPW+99, Tur00, MT04, MW08]. Sekä ehdokkaiden poiminta tekstistä että lopullisten avainsanojen valinta voidaan toteuttaa ohjaamattomilla menetelmillä ilman opetusdatan käyttöä [BC00, MT04, LLZS09]. Jälkimmäistä vaihetta voidaan kuitenkin käsitellä myös luokitteluongelmana, jossa voidaan hyödyntää ohjattua koneoppimista [FPW+99, Tur00, MW08]. Perinteisestä tekstin kategorisoinnista poiketen tavoitteena on tällöin oppia luokittelemaan tekstissä esiintyvät sanat tai ilmaukset sisällön kuvailun kannalta relevantteihin ja ei-relevantteihin niiden omien piirteiden perusteella.

Avainsanojen sijaan alan kirjallisuudessa käsitellään usein avainlausekkeiden eristämistä, koska tekstin keskeisiä aiheita tai teemoja kuvaavat käsitteet voivat olla yksittäisten sanojen sijaan myös usean sanan yhdistelmiä [Tur00], esimerkiksi substantiivilausekkeita [BC00, Hul03] tai sanaliittoja. Esimerkiksi ilmaus ”puolueiden kannatus” koostuu kahdesta substantiivista ja ”synteettinen materiaali” puolestaan adjektiivista ja substantiivista, mutta kumpikin muodostaa silti kokonaisuutena itsenäisen käsitteen, jonka merkitys on huomattavasti yksittäistä sanaa rajatumpi. Useimmat tekstin läpikäyntiin perustuvat indeksointialgoritmit eristävätkin yksittäisten sanojen sijaan n peräkkäisen sanan jonoja eli n-grammeja. Tässä tutkielmassa termejä käytetään yhtäläisessä merkityksessä, ja sekä avainsanalla että -lausekkeella tarkoitetaan yhden tai useamman sanan muodostamaa tiedon sisältöä kuvaavaa ilmaisua.

Avaintermien eristäminen muistuttaa tehtävänä näennäisesti *tiedon eristämistä* (information extraction), jonka menetelmiä voidaan niin ikään käyttää metatiedon tuottamiseen tekstiaineistosta automaattisesti [EMSS00, HSC02, AKM+03]. Tiedon eristämisessä tekstistä pyritään kuitenkin löytämään sovelluskohtaisen tehtävän kannalta keskeisiä tietoja kuten esimerkiksi määrätynlaisia tapahtumia tai erisnimiä, kun taas indeksoinnin tavoitteena on tuottaa yleiskäyttöisiä, asiakirjan sisältöä käsitteiden tasolla kuvaavia indeksointitermejä [Tur00].

Tekstissä esiintyvien termien eristämiseen perustuvien menetelmien keskeisenä rajoitteena on se, etteivät ne pysty tuottamaan termejä, jotka eivät esiinny jossakin muodossa itse tekstissä. Tämä voi olla ongelma, jos metatiedon halutaan esimerkiksi kuvaavan sisältöä korkealla abstraktiotasolla käsittein, jotka eivät välttämättä esiinny tekstissä sellaisinaan [PSI03].

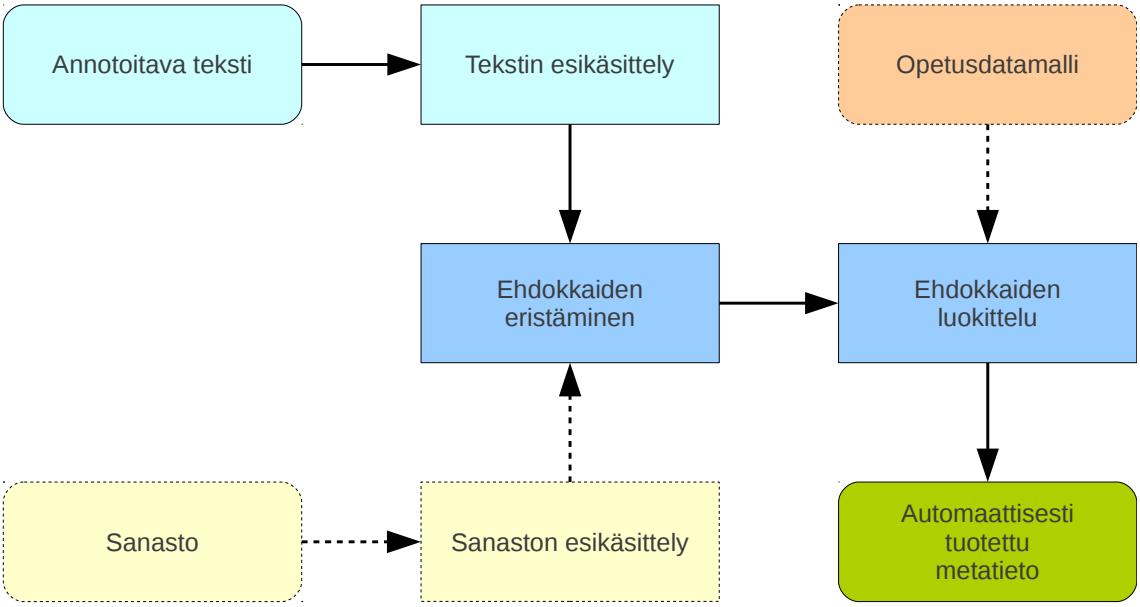
Toisaalta tekstistä on tarvittaessa mahdollista eristää myös mielivaltaisia, sanaston ulkopuolisia termejä. Eristettävien termien ei myöskään tarvitse välttämättä esiintyä sellaisinaan opetusdatassa voidaakseen tulla valituiksi, koska myös ohjattua koneoppimista hyödyntävät eristysalgoritmit pyrkivät oppimaan, millaiset tekstissä esiintyvän lausekkeen piirteet yleisesti ennustavat sen olevan aiheen kuvailun kannalta keskeinen, eivät yksittäisten sanaston termien ja tekstidokumenttien suhdetta. Perinteiset sanaston termejä kategorioina käsittelevät luokittelualgoritmit puolestaan voivat valita asiasanoiksi myös sellaisia käsitteitä, jotka eivät esiinny tekstissä, mutta rajoittuvat toisaalta sanastossa tai opetusdatassa eksplisiittisesti esiteltyihin kategorioihin.

Tekstin luokittelu vaikuttaa siten luontevalta lähestymistavalta kontrolloituun asiasanoitukseen ja tekstistä eristäminen puolestaan vapaiden avainsanojen automaattiseen tuottamiseen. Myös kontrolloituun sanastoon perustuvaan asiasanoitukseen on kuitenkin kehitetty algoritmeja, jotka eristävät sanaston termejä tekstistä [MW06b]. Keskeisenä etuna suoranaiseen tekstin luokitteluun nähden on vähäisempi opetusdatan tarve etenkin, jos käytettävä sanasto on hyvin laaja [MW08].

Termejä kategorioina käsittelevä luokittelu tarvitsee opetusdataksi riittävää otosta valmiiksi luokiteltua aineistoa jokaista sanaston termiä kohden, mikä tuhansista käsitteistä koostuvan sanaston tapauksessa edellyttää varsin suurta valmiiksi annotoitua aineistoa. Avainsanoja tai käsitteitä tekstistä eristävien algoritmien opetusdataksi voi riittää muutamista kymmenistä [FPW+99] tai sadoista [Tur00, MW08, Med09, s. 146] annotoiduista teksteistä koostuva aineisto.

Esimerkiksi luvussa 5 esiteltävässä tapaustutkimuksessa valmista indeksointimetadataa oli käytettävissä suhteellisen vähän. Sanasto koostui useista tuhansista käsitteistä, joiden käyttö olemassa olevassa aineistossa oli jakautunut varsin epätasaisesti eri termien kesken. Monet sanaston termit eivät esiintyneet otoksen metatiedoissa kertaakaan.

Opetusdatan riittävyys ja edustavuus voivat tällöin olla ongelma. Tässä tutkielmassa keskitytäänkin menetelmiin, joilla avaintermejä pyritään eristämään suoraan tekstistä. Yleiskuvan suoranaiseen tekstin kategorisointiin perustuvista menetelmistä esittää esimerkiksi Sebastianin katsaus [Seb02].



Kuva 1: Yleinen prosessi avaintermien eristämiseksi tekstistä. Yhtenäisellä viivalla esitetyt vaiheet ovat yhteisiä useimmille eristysmenetelmille riippumatta sanastojen tai ohjatun koneoppimisen käytöstä. Valinnaiset vaiheet on merkitty katkoviivoin.

3.2 Avaintermien eristäminen

Keskeisten termien tekstistä eristämiseen perustuvat algoritmit ovat yleensä kaksivaiheisia. Ensimmäisessä vaiheessa koko teksti käydään läpi ja siitä pyritään tunnistamaan avainsanaehdokkaiksi kelpaavat lausekkeet, ja toisessa vaiheessa kandidaattien joukosta pyritään valitsemaan tekstin sisällön kuvailun kannalta keskeisimmät painottamalla niitä heuristisiten sääntöjen mukaan tai ohjatulla koneoppimisella. Näitä edeltää usein aineiston esikäsittely. Prosessin yleinen kulku on esitetty kuvassa 1.

3.2.1 Tekstin esikäsittely

Yleisin tekstin automaattisessa annotoinnissa käytettävä esikäsittelyvaihe on eri taivutusmuodoissa esiintyvien sanojen normalisointi siten, että saman sanan esiintymät eri muodoissa voidaan samaistaa. Eri taivutusmuodoissa olevat sanat voidaan normalisoida yhtenäiseen muotoon joko muuntamalla sanat perusmuotoonsa tai eristämällä niistä sanan vartalo. Ensinmainittua kutsutaan perusmuotoistamiseksi eli lemmatisoinniksi (lemmatization) ja jälkimmäistä typistämiseksi eli stemmaukseksi (stemming). Normalisointi voidaan suorittaa joko esikäsittelynä tai sitä mukaa kuin tekstiä käydään läpi varsinaisessa eristysvaiheessa.

Kun indeksoinnissa käytetään kontrolloitua sanastoa, typistämisen tai perusmuotoistamisen ansiosta tekstissä mainitut sanat voidaan yhdistää vastaaviin sanaston käsitteisiin myös silloin, kun ne esiintyvät tekstissä eri taivutusmuodossa kuin sanastossa. Myös sanaston termit on syytä normalisoida vastaavaan tapaan, jos ne eivät jo ole tekstin normalisoitua asua vastaavassa muodossa. Taivutusmuodon normalisointi sanojen typistyksellä kasvattaa tekstin läpikäynnissä löydettävien potentiaalisten asiasanaehdokkaiden lukumäärää [MW08, Med09, s. 82]. Typistäminen voi parantaa myös vapaamuotoisten avainsanojen eristämiseen perustuvien indeksointimenetelmien suorituskykyä [Hul03].

Sanan luotettava perusmuotoistaminen edellyttää sen sanaluokan määrittämistä ja on yleensä monimutkaisempaa kuin typistäminen. Esimerkiksi tekstissä esiintyvä sana ”alusta” voi asiayhteydestä riippuen olla jokin taivutusmuoto useasta eri substantiivista (alku, alunen, alus, alusta) tai verbistä (alustaa). Perusmuotoa ei siis voida päätellä yksikäsitteisesti pelkästään tekstissä esiintyvän sanan itsensä perusteella. Edistyneen perusmuotoistuksen käytöstä yksinkertaisemman typistämisen sijaan voi kuitenkin olla hyötyä esimerkiksi suomenkielisen tekstin automaattisessa indeksoinnissa [SSH11].

Jotkin annotointialgoritmin osana käytettävistä kielikohtaisista komponenteista voivat olla vaihdettavissa tarpeen mukaan, ja esimerkiksi lauseenjäsennykseen tai sanojen perusmuotoistamiseen on olemassa yleisesti tunnettuja algoritmeja useille eri kielille. Indeksointialgoritmit, jotka hyödyntävät vain tällaisia luonnollisen kielen menetelmiä ilman muita oletuksia käytettävästä kielestä, voivat siten soveltua indeksointiin useilla eri kielillä [BRK05, SSH11, Med09, s. 141–142]. Tunnettuja esimerkkejä englanninkielisten sanojen typistämiseen kehitetyistä menetelmistä ovat Lovinsin [Lov68] ja Porterin [Por80] algoritmit. Suomenkielisen tekstin lemmatisointiin puolestaan soveltuvat esimerkiksi Connexor Oy:n kaupallinen FDG-jäsennin [TJ97] ja avoimena lähdekoodina julkaistu Omorfi [LSP09].

3.2.2 Ehdokkaiden eristäminen

Muodollisesti kelvollisina avainsanoina pidetään usein vain substantiiveja ja substantiivilausekkeita [Hul03, BRK05]. Ehdokkaiden eristämisvaiheessa voidaan käyttää esimerkiksi kielikohtaista lauseenjäsennysalgoritmia määrittämään kunkin sanan todennäköinen sanaluokka, jolloin ehdokkaiksi voidaan valita vaikkapa vain substantiivit [DM05]. Asiasanastoa käytettäessä ongelma voidaan ratkaista valitsemalla tekstistä ehdokkaiksi vain ne sanat tai lausekkeet, joiden perusmuoto tai sanavartalo

vastaa jotakin sanastossa määriteltyä termiä [MW06b].

Läpikäynnissä jätetään usein huomioimatta merkkijonot, joita ei katsota sopiviksi avainsanaehdokkaiksi. Tällaisia voivat olla esimerkiksi numerot, erisnimet ja hukkasanat [FPW+99]. Hukkasanoja (stop words) ovat tekstissä esiintyvät sanat, jotka saattavat olla esimerkiksi kieliopillisesti tarpeellisia mutta jotka eivät kuitenkaan sinällään ole semanttisesti merkittäviä. Ne tunnistetaan käsin määritellyn kielikohtaisen listan perusteella. Lisäksi tekstistä poimitut sanat usein normalisoidaan esimerkiksi typistämällä [FPW+99, Hul03] tai perusmuotoistamalla [SSH11], jolloin voidaan huomioida saman sanan esiintyminen useampaan kertaan esimerkiksi eri taivutusmuodoissa.

3.2.3 Ehdokkaiden luokittelu

Kun avaintermiehdokkaat on saatu poimittua tekstistä, niiden joukosta on valittava parhaat lopullisiksi annotaatioiksi. Tätä kutsutaan myös ehdokkaiden suodattamiseksi (candidate filtering) [MW06b]. Vaiheen voi nähdä luokitteluongelmana, jossa ehdokkaat on jaoteltava relevantteihin ja epärelevantteihin. Tulos voi olla myös jatkuva-arvoinen ehdokaskohtainen luottamusarvo, jonka perusteella ehdokkaille voidaan määrittää pelkän binäärisen jaottelun sijaan keskinäinen järjestys.

Luokitteluun voidaan käyttää käsin määriteltyjä heuristisia sääntöjä [BC00], ohjaamattomia tilastollisia menetelmiä [MT04, LLZS09] tai ohjattua koneoppimista [FPW+99, Tur00, Med09]. Luokittelussa käytettävät säännöt ja ehdokkaiden piirteet voivat liittyä esimerkiksi termin esiintymien lukumäärään tekstissä [FPW+99, Tur00, BC00], esiintymien sijaintiin [FPW+99, Tur00, NK07], sanan tai lausekkeen pituuteen [Tur00, BC00, MW06b] tai termin ominaisuuksiin osana sanaston hierarkiaa [HKJ+01, MW06b, Med09].

3.2.4 Vapaiden avainsanojen eristäminen

Suurimmassa osassa avaintermien eristämistä koskevaa tutkimusta on keskitytty pääosin vapaamuotoisten avainlausekkeiden eristämiseen ilman kontrolloitua sanastoa. Tässä aliluvussa tarkastellaan joitakin tunnettuja menetelmiä, joita avainsanojen eristämiseen on kirjallisuudessa esitetty.

GenExiä [Tur99, Tur00] on pidetty ensimmäisenä algoritmina, joka hyödynsi avaintermien eristämisessä ohjattua koneoppimista [PDB+10, Med09, s. 37]. Algoritmi etsii tekstistä kaikki siinä esiintyvät yksittäiset sanat, karsii pois hukkasanat ja nor-

malisoi sanat vartaloiksi typistämällä ne määrättyyn maksimipituuteen. Tämän jälkeen sanavartalot asetetaan paremmuusjärjestykseen niiden esiintymiskertojen ja painokertoimen mukaan. Painokertoimeen vaikuttaa esimerkiksi vartalon ensimmäisen esiintymän sijainti tekstissä [Tur99, Tur00].

Vartaloista valitaan parhaat, minkä jälkeen tekstistä etsitään kaikki korkeintaan kolmen sanan pituiset lausekkeet, joiden osana kukin vartalo esiintyy. Lopuksi näin valikoiduille avainlausekkeille lasketaan painoarvot esimerkiksi esiintymien ja lausekkeen pituuden mukaan. Algoritmin parametrit, kuten sanavartalon maksimipituuden ja painokertoimen laskentaan käytettävät piirteet, GenEx asettaa opetusaineistosta geneettisellä algoritmilla muodostetun mallin mukaan [Tur99, Tur00].

Viidellä eri aineistolla suoritetuissa testeissä 20–29 prosenttia GenExin eristämistä termeistä vastasi alkuperäisiä ihmisten määrittämiä avainsanoja [Tur99]. Subjektiiviseen evaluointiin osallistuneet katsoivat 62,0 % GenExin tuottamista termeistä hyviksi ja 18,1 % huonoiksi [Tur00]. Myös monet sellaiset GenExin eristämät termit, jotka eivät esiinny alkuperäisessä metadatassa, vaikuttavat siis subjektiivisten arvioiden mukaan kelvollisilta. Arvioijat ovat kuitenkin tutkimushankkeen verkkosivuilla vierailleita käyttäjiä, eikä vertailua muilla algoritmeilla tuotetun tai alkuperäisen metatiedon subjektiiviseen laatuun esitetä. Siten ei ole selvää, miten esimerkiksi osanottajien mahdollinen valikoituminen tai muut tekijät vaikuttavat arvioinnin tuloksiin.

Kea [FPW+99] jakaa tekstin ensin lauseisiin välimerkkien perusteella ja poimii sen jälkeen avainsanaehdokkaiksi kaikki korkeintaan kolmesta samaan lauseeseen kuuluvasta peräkkäisestä sanasta koostuvat lausekkeet. Sen jälkeen algoritmi normalisoi ehdokkaiden kirjainkoon ja typistää kunkin lausekkeessa esiintyvän sanan iteroidulla Lovinsin algoritmilla [Lov68]. Toisessa vaiheessa tekstistä poimitut ehdokkaat luokitellaan avaintermeihin ja ei-avaintermeihin naiivilla Bayes-luokittelijalla. Ehdokkaiden piirteinä luokittelussa algoritmi käyttää TF×IDF:ää ja termin ensimmäisen esiintymän suhteellista sijaintia tekstissä.

TF×IDF on muun muassa tiedonhaussa yleisesti käytetty laskennallinen arvo, joka koostuu termin esiintymistiheydestä käsiteltävässä dokumentissa (term frequency, TF) kerrottuna sen käänteisellä yleisyydellä koko tekstikorpuksessa (inverse document frequency, IDF). Sen arvo on siis suurin termeille, jotka esiintyvät useita kertoja käsiteltävässä dokumentissa mutta jotka kuitenkin ovat harvinaisia koko aineistossa. TF×IDF:n katsotaan korostavan termejä, jotka ovat juuri käsiteltävän dokumentin kannalta keskeisiä.

Termin esiintymän suhteellinen sijainti tekstissä puolestaan on esiintymää edeltävien sanojen määrä jaettuna tekstin kaikkien sanojen määrällä. Kolmas luokittelussa käytettävä attribuutti on se, kuinka moneen kertaan termiä on yhteensä käytetty avainsanana koko opetusaineistossa. Algoritmi diskretisoi piirteiden arvot ennen luokittelua [FPW+99].

Ohjattua koneoppimista hyödyntävänä menetelmänä Kea edellyttää valmiiksi annotoitua esimerkkiaineistoa. Sitä vaaditaan kuitenkin huomattavasti vähemmän kuin sanaston termejä kategorioina käyttävässä tekstin luokittelussa. Noin viidestäkymmenestä annotoidusta tekstiasiakirjasta koostuva opetusaineisto riittää kelvollisten tulosten saavuttamiseksi [FPW+99]. Lisäksi jopa eri aihealuetta käsittelevä aineisto voi soveltua opetusdataksi [FPW+99, JP02].

Kahden eri alan tekstidokumenteilla suoritetuissa kokeissa viidestä Kealla eristetystä termistä keskimäärin 1,35 tai 1,46 vastasi alkuperäisiä avainsanoja. GenExin vastaavat tulokset olivat 1,45 ja 1,43. Ero algoritmien välillä ei ole tilastollisesti merkittävä. Kean naiivi Bayes-luokittelija on kuitenkin opetusvaiheessa merkittävästi GenExin geneettistä algoritmia nopeampi [FPW+99]. Subjektiivisten arvioiden perusteella Kean eristämien avainsanojen laadun on havaittu olevan lähellä ihmisten tasoa [JP02].

D'Avanzo ja Magnini [DM05] esittelevät algoritmin, johon tässä tutkielmassa viitataan nimellä **D&M**. Myös D&M käyttää TF×IDF:ää ja ensimmäisen esiintymän sijaintia piirteinä ehdokkaiden keskinäiseen järjestämiseen naiivilla Bayes-luokittelijalla. Kandidaattien eristämiseen se kuitenkin soveltaa edistyneempiä luonnollisen kielen käsittelyn menetelmiä kuten lauseenjäsennystä ja erisnimien tunnistusta (named entity recognition).

D'Avanzo ja Magnini [DM05] käyttävät avaintermien eristystä osana automaattista tiivistelmien muodostamista, jolloin esimerkiksi erisnimillä on merkittävä informaatioarvo toisin kuin yleensä asiasanoituksessa. Järjestelmä pyrkii myös tunnistamaan tekstistä useamman yksittäisen sanan muodostamat sanaliitot etsimällä vastaavia termejä WordNet-sanastosta [Fel98]. Tutkimus keskittyy englanninkielisten uutistekstien tiivistämiseen eikä ota kantaa järjestelmän soveltuvuuteen muunkielisten aineistojen käsittelyyn. Järjestelmää myös evaluoidaan vain lopputuloksena syntyvien tiivistelmien kautta, joten arviointi ei tarjoa vertailukohtia muihin tässä esitettyihin menetelmiin.

Barkerin ja Cornacchian [BC00] **B&C**-algoritmi perustuu niin ikään substantiivilausekkeiden paikantamiseen sanaluokkien perusteella. Varsinaisen lauseenjäsennyk-

sen sijaan algoritmi kuitenkin hakee termejä sanaluokkatiedon tarjoavasta sanakirjasta substantiivien ja niitä mahdollisesti edeltävien adjektiivien löytämiseksi tekstistä. Ehdokkaiden suodattamisessa menetelmä käyttää lausekkeen pääsanan ja koko lausekkeen esiintymistiheyksiä tekstissä sekä lausekkeen pituutta. Piirteitä käytetään yksinkertaisina heuristisina sääntöinä, eikä luokittelussa käytetä koneoppimista. Tällöin valmiiksi annotoitua opetusdataakaan ei tarvita.

Barkerin ja Cornacchian [BC00] mukaan algoritmi vaikuttaa suorituskyvyltään vertailukelpoiselta GenExin kanssa. Ihmisten suorittamien arvioiden perusteella GenEx saavuttaa marginaalisesti paremmat tulokset yksittäisten avainsanojen osalta, kun taas B&C:n tulos on hieman parempi vertailtaessa kaikkien samasta dokumentista eristettyjen avainsanojen muodostamia metadatakokonaisuuksia keskenään. Barker ja Cornacchia pidättäytyvät kuitenkin vahvoista päätelmistä algoritmien suorituskyvyn suhteen.

Hulth [Hul03] esittelee algoritmin, joka hyödyntää lauseenjäsennystä sekä ehdokkaiden eristys- että suodatusvaiheissa. Potentiaalisten avainlausekkeiden löytämiseksi tekstistä Hulth vertailee kolmea eri eristyskriteeriä: n-grammeja, substantiivilausekkeita ja ennalta määritettyjä sanaluokkakaavoja. Ensin mainittu muistuttaa muun muassa Keassa ja GenExissä käytettyjä menetelmiä, kun taas keskimmäisessä tekstistä poimitaan lauseenjäsennyksen perusteella ainoastaan substantiivilausekkeet. Viimeisimmässä vaihtoehdossa ehdokkaiksi valitaan pelkkien substantiivilausekkeiden sijaan kaikki jotakin ennalta määrättyä sanaluokkien yhdistelmää vastaavat tekstissä esiintyvät jonot. Esimerkkeinä Hulth mainitsee muun muassa pelkästä substantiivista, adjektiivista ja substantiivista sekä kahdesta peräkkäisestä substantiivista koostuvat jonot. Kaikki esimerkkeinä mainitut sanaluokkakaavat voidaan tulkita myös substantiivilausekkeiksi, eikä Hulth mainitse esimerkkejä käyttämistään muunlaisista sanaluokkahahmoista. Käytettävät 56 hahmoa on määritetty englanninkielisessä alkuperäisaineistossa esiintyvien avainsanojen pohjalta.

Substantiivilausekkeisiin perustuva ehdokkaiden eristämisen strategia tuottaa kokeiden perusteella parhaan indeksointitarkkuuden. Sanaluokkahahmoihin ja n-grammeihin perustuvat menetelmät puolestaan tuottavat huomattavasti korkeamman saannin [Hul03]. Jälkimmäinen tulos lienee odotettava, koska ainoastaan substantiivilausekkeiden kelpuuttaminen on kolmesta vertaillusta kriteeristä rajaavin ja tuottanee siis alunperinkin suppeimman joukon ehdokkaita.

Ehdokkaiden suodatuksessa Hulth [Hul03] käyttää piirteinä termin esiintymistiheyttä dokumentissa ja koko korpuksessa, ensimmäisen esiintymän sijaintia sekä lausek-

keen sanojen sanaluokkia. Luokittelualgoritmina hän soveltaa ohjattua sääntöjen oppimista (rule induction). Hulth toteaa sanaluokkatiedon hyödylliseksi luokittelun piirteeksi kaikkien kokeilemiensa ehdokkaiden eristämisen strategioiden kanssa, mutta eniten siitä on etua yksinkertaiseen kaikkien n-grammien poimintaan perustuvan eristysmenetelmän yhteydessä.

Kaikista kokeilluista ehdokkaiden eristys- ja suodatusmenetelmien yhdistelmistä paras tarkkuus saatiin substantiivilausekkeiden eristämisellä yhdistettynä sanaluokkatietoa hyödyntävään suodatukseen. Tarkkuuden ja saannin yhdistävällä F-luvulla mitattuna korkeimman tuloksen saavuttaa n-grammeihin perustuva eristys sanaluokkatietoa hyödyntävään luokitteluun yhdistettynä [Hul03]. Suoraa vertailua muihin algoritmeihin ei esitetä.

3.2.5 Rakenteellisten sanastojen käyttö

KEA++ [MW06b] on Kean pohjalta kehitetty menetelmä, joka eristää tekstistä mielivaltaisten sanojen sijaan kontrolloidun sanaston termejä. Sanastona KEA++ voi käyttää mitä tahansa SKOS-standardin mukaista termistöä. Algoritmi normalisoi tekstissä esiintyvien n-grammien tapaan myös sanaston termien nimikkeet ja valitsee ehdokkaiksi tekstin lausekkeet, jotka normalisoinnin jälkeen vastaavat jotakin sanaston termiä.

SKOS-muotoisissa sanastoissa voidaan määritellä termeille muun muassa keskinäisiä ylä- ja alaluokkasuhteita sekä synonyymeja, ja algoritmi hyödyntääkin tätä samaa tarkoittavien sanojen samaistamiseen. KEA++:n yhdistää siten avainlausekkeiden eristämisen ja sanastoperustaisen luokittelun etuja parantamalla kontrolloidun sanaston avulla indeksoinnin yhdenmukaisuutta ja vaatimalla samanaikaisesti vähemmän opetusdataa kuin perinteiset luokittelumenetelmät [MW06b].

Ehdokkaiden normalisointiin KEA++ lisää hukkasanojen poiston ja typistämisen ohelle kolmannen vaiheen, jossa typistetyn n-grammin sanat järjestetään aakkosjärjestykseen [MW06b]. Normalisointistrategia vastaa bag of words -mallia, jossa sanojen alkuperäisellä järjestyksellä lausekkeen sisällä ei ole merkitystä. Tällöin esimerkiksi englanninkieliset lausekkeet ”keyphrase extraction” ja ”extraction of keyphrases” tuottavat saman normalisoidun muodon ”extract keyphras”, ja algoritmi samastaa toisistaan poikkeavat mutta mahdollisesti samaa tarkoittavat merkkijonot.

Ehdokkaiden eristämisvaiheessa käytettävän sanaston sekä muokatun normalisointistrategian lisäksi KEA++ lisää edeltäjäänsä nähden myös ehdokkaiden luokittelu-

vaiheeseen uusia piirteitä. Niistä ensimmäinen on lausekkeen sanojen lukumäärä ja toinen sanaston hierarkiassa termiä vastaavan solmun aste [MW06b].

Jos sanaston käsitehierarkiaa ajatellaan suunnattuna verkkona, solmun asteella tarkoitetaan tässä tapauksessa sellaisten solmusta lähtevien semanttisia suhteita kuvaavien kaarien lukumäärä, joiden toinen päätepiste on jokin toinen samassa tekstissä esiintyvä käsite. Algoritmi pyrkii siis hyödyntämään luokittelussa sanaston eksplisiittisesti kuvaamia käsitteiden välisiä semanttisia suhteita. Taustalla on hypoteesi, jonka mukaan tiettyä aihepiiriä käsittelevässä tekstissä esiintyvät todennäköisemmin myös muut semanttisesti läheiset termit, jolloin tällaiset ehdokkaat ovat todennäköisemmin asiakirjan aihepiirin kannalta keskeisiä. YK:n elintarvike- ja maatalousjärjestö FAO:n aineistolla suoritetuissa kokeissa KEA++ saavutti F-luvun 0,252, mikä on selvä parannus Kean vastaavassa kokeessa saavuttamaan 0,120:aan nähden [MW06b].

Myös Hulth ja kumppanit [HKJ+01] ehdottavat eksplisiittisiä semanttisia suhteita hyödynnettäviksi ehdokkaiden luokittelussa. He toteavat sanaston hierarkiaan perustuvien piirteiden parantavan algoritminsa suorituskykyä huomattavasti verrattuna asetelmaan, jossa piirteinä käytetään ainoastaan TF:ää ja IDF:ää. Toisaalta vertailua esimerkiksi muunlaisia epätriviaaleja piirteitä hyödyntäviin algoritmeihin ei esitetä. Myös aineisto ja koeasetelma poikkeavat muista tässä mainituista tutkimuksista, eivätkä tulokset siten ole suoraan vertailukelpoisia.

Medelyan ja Witten [MW08] kokeilevat KEA++:n kehittämiseksi jo Keassa käytetyn ensimmäisen esiintymän lisäksi termin kaikkien esiintymien jakaumaan perustuvia piirteitä, mutta ne eivät heidän mukaansa paranna luokittelutulosta. Lisäksi he testaavat naiivin Bayes-luokittelijan ohella muun muassa tukivektorikoneita, lineaarista regressiota, päätöspuita sekä bagging-menetelmällä [Bre96a] koostettuja yksitasoisia päätöspuita (decision stump). Näistä naiivi Bayes-luokittelija tuottaa kuitenkin parhaan tuloksen. Lisäksi KEA++ toimii myös ranskan- ja espanjankielisillä aineistoilla, joskin suorituskyky jää englanninkielisillä asiakirjoilla tehtyjä kokeita heikommaksi [MW08].

Medelyan ja kumppanit [MWM08] esittelevät KEA++:sta laajennetun version, joka käyttää sanastona Wikipedian[7] artikkeleiden otsikoita. Tällöin algoritmi voi käyttää ehdokkaiden suodatuksessa piirteinä myös artikkeleiden välisten linkkien ja artikkeleiden kategorisointitietojen perusteella määritettäviä semanttisia suhteita.

Maui [MFW09, Med09] on KEA++:aan perustuva algoritmi, joka lisää käytettäviin

[7] http://www.wikipedia.org

piirteisiin termin ensimmäisen ja viimeisen esiintymän välisen etäisyyden sekä termiä vastaavan Wikipedia-artikkelin sijainnin sanakirjan kategoriahierarkiassa. Ensinmainitulla voidaan painottaa termejä, jotka esiintyvät sekä käsiteltävän tekstin alussa että lopussa [MFW09]. Jälkimmäisen piirteen tarkoituksena on määrittää laskennallisesti, ovatko ehdokkaat merkitykseltään yleisluontoisia vai erityisiä [Med09, s. 100–101]. Luokittelijana Maui käyttää oletusarvoisesti bagging-algoritmilla [Bre96a] aggregoituja päätöspuita. Luokittelijan toteutukseen Maui käyttää WEKA-kirjastoa[8] [HFH+09], ja päätöspuut muodostetaan C4.5-algoritmiin [Qui93] perustuvalla menetelmällä.

Baggingissa (bootstrap aggregating) opetusdatasta poimitaan joukko satunnaisia otoksia takaisinpanolla. Kunkin otoksen perusteella muodostetaan luokittelija, ja näin saadut luokittelijat äänestävät lopullisesta tuloksesta [Bre96a]. Menetelmä parantaa luokittelun tarkkuutta epävakailla luokittelumenetelmillä kuten päätöspuilla [Bre96a, Qui96], joissa pieni muutos opetusdatassa ja puun rakenteessa voi aiheuttaa merkittävän eron luokittelun tulokseen [Bre96b].

Edeltäjistä poiketen Mauissa koostetut päätöspuut tuottavat naiivia Bayesia paremman tuloksen, kun uudet piirteet otetaan käyttöön [MFW09, Med09, s. 134–135]. Jotkin piirteistä ovat tilastollisesti riippuvaisia toisistaan, mikä heikentää naiivin Bayes-luokittelijan suorituskykyä [Med09, s. 135–136].

Maui annotoi tieteellisiä artikkeleita vapain avainsanoin edeltäjäänsä Keaa paremmin tuloksin [MFW09]. Myös maa- ja elintarviketalouden, lääketieteen ja hiukkasfysiikan alojen aineistoilla tehdyissä kokeissa Maui suoriutuu paremmin kuin Kea ja KEA++ [Med09, s. 137–139]. Maui soveltuu myös suomenkielisten aineistojen asiasanoitukseen, kunhan käytettävissä on sopiva sanasto, hukkasanalista ja tasokas lemmatisointialgoritmi [SSH11]. Lisäksi menetelmä toimii useita eri aihealueita käsittelevillä tekstiaineistoilla [SSH11, Med09, s. 138–141].

3.2.6 Aihealuekohtaiset tekniikat

Edellämainitut menetelmät ovat sikäli riippumattomia aineiston erityispiirteistä, etteivät ne käytä tiettyä toimialaa tai aineistolajia varten kohdennettuja tekniikoita. Erityiskäyttöisiä menetelmiä tai mukautuksia on kehitetty esimerkiksi tieteellisten artikkeleiden indeksointiin [KMKB10].

HaCohen-Kerner ja kumppanit [HGM05] osoittavat tieteellisen artikkelin tyypilli-

[8] http://www.cs.waikato.ac.nz/ml/weka/

seen rakenteeseen perustuvien piirteiden parantavan eristysalgoritminsa suorituskykyä. Myös Nguyenin ja Kanin [NK07] algoritmi hyödyntää luokittelun piirteenä muun muassa tavanomaisen akateemiseen jäsentelyn mukaisia lukuja ja saavuttaa Keaa paremman tuloksen tilastollisesti merkittävällä erolla. SemEval-työpajan tieteellisten artikkeleiden annotointia käsitelleessä tehtävässä [KMKB10] monet kilpailuun osallistuneet menetelmät suoriutuivat yleiskäyttöistä Mauia paremmin.

Aineistolajin tai aihealueen erityispiirteet huomioivat mukautukset voivat siis parantaa tuloksia yleiskäyttöisiin algoritmeihin verrattuna. Luvussa 5 esiteltävää tapaustutkimusta silmällä pitäen tämä tutkielma keskittyy kuitenkin yleiskäyttöisiin indeksointialgoritmeihin.

3.3 Puoliautomaattinen indeksointi

Järjestelmäsuunnittelun näkökulmasta metatiedon automaattista eristämistä voidaan käyttää eri osissa työnkulkua. Puoliautomaattisessa indeksoinnissa automatiikkaa käytetään ehdottamaan aineiston perusteella potentiaalisia avaintermejä käyttäjälle, joka kuitenkin määrittää lopulliset indeksointitermit [LC96]. Tällöin automatiikalla voidaan helpottaa varsinaisena indeksoijana toimivan asiantuntijan työtä [BCF02]. Ihmisen tekemän lopullisen tarkistuksen ansiosta puoliautomaattiset indeksointimenetelmät voivat olla täysin automaattista luokittelua sopivampi sovellustason ratkaisu silloin, kun pelkkä automaattinen luokittelu ei ole riittävän luotettavaa esimerkiksi opetusdatan heikon laadun tai edustavuuden vuoksi [Seb02].

Toisaalta indeksointiautomatiikan rooli järjestelmän osana voi puolestaan vaikuttaa siihen, millaisia tekijöitä algoritmin parametrien säädössä ja suorituskyvyn arvioinnissa on syytä painottaa. Esimerkiksi puoliautomaattisen annotoinnin tapauksessa voi olla perusteltua pyrkiä annotoinnin ehdottoman tarkkuuden sijaan tuottamaan riittävän suuri joukko kelvollisia ehdokkaita käyttäjän valikoitaviksi. Kaikkien automaattisesti eristettyjen kandidaattien ei välttämättä tarvitse olla hyviä, jos lopullisen asiasanoituksen tekevä ihminen voi helposti valita niistä sopivat. Jos taas automaattista eristystä käytetään tuottamaan suoraan valmista metatietoa ilman ihmisen tekemää systemaattista tarkistusta, algoritmin valitsemien termien oikeellisuutta voidaan haluta painottaa suhteellisesti enemmän.

Jos indeksointialgoritmi tuottaa pelkän binäärisen relevanssiluokittelun sijaan kullekin ehdottamalleen avain- tai asiasanalle jatkuva-arvoisen, termin relevanssin todennäköisyyttä kuvaavan luottamusarvon, ehdokkaat on luontevaa esittää käyttä-

jälle sen mukaisessa järjestyksessä. Edellä kuvatuista termien tekstistä eristämiseen perustuvista algoritmeista luottamusarvon tai keskinäisen järjestyksen tuottavat ainakin GenEx [Tur99] ja Kea [FPW+99] sekä sen seuraajat KEA++ [MW08] ja Maui [Med09]. Lisäksi tällöin on kiinnostavaa, kuinka hyvin luokittelijan määrittämä ehdokkaiden keskinäinen järjestys vastaa asiasanoituksen todellista laatua. Luvussa 6 käsitellään tapaustutkimuksen muiden tulosten ohella lyhyesti myös Mauin eristämilleen termeille asettamien luottamusarvojen ja asiantuntijoiden esittämien subjektiivisten laatuarvioiden keskinäistä vastaavuutta.

Muilta osin tässä tutkielmassa esitettävissä kokeissa ja niiden tulosten analyysissa ei kuitenkaan oteta kantaa asiasanoituksen käytön yksityiskohtiin sovellustasolla, vaan suorituskykyä pyritään arvioimaan sovellusriippumattomasta näkökulmasta.

4 Arviointimenetelmät

Automaattisia indeksointialgoritmeja ja niiden eristämän asiasanoituksen laatua arvioidaan usein vertaamalla koneellisesti tuotettua metatietoa tavalla tai toisella ihmisten tekemiin annotaatioihin. Tässä luvussa käsitellään joitakin evaluointimenetelmiä, joita asiasanoituksen laadun tutkimiseen ja vertailuun on käytetty.

Arviointimenetelmät voi jakaa karkeasti automaattisesti suoritettaviin laskennallisiin mittauksiin sekä ihmistyönä tehtäviin evaluointeihin. Huomionarvoista on myös se, mitataanko arvioinnilla annotaatioiden laatua, kattavuutta vai molempia.

Useimmat tässä esiteltävät laadun ja kattavuuden mittarit soveltuvat yhtä lailla vapaiden avainsanojen eristyksen kuin kontrolloidun asiasanoituksenkin arviointiin. Poikkeuksen muodostavat jotkin indeksointitermien semanttisen samankaltaisuuden huomioivat menetelmät, jotka edellyttävät sanastoa taustatiedoksi.

4.1 Tarkkuus ja saanti

Eräs suoraviivainen arviointitapa on mitata asiasanoituksen tarkkuutta (precision) ja saantia (recall), jotka ovat yleisiä evaluoinnin perustyökaluja etenkin tiedonhaun ja luokittelun aloilla. Tiedonhaussa tarkkuudella tarkoitetaan relevanttien tulosten osuutta kaikista hakujärjestelmän kyselyyn tuottamista tuloksista, ja saanti on vastaavasti palautettujen relevanttien tulosten osuus kaikista kyselyn kannalta olennaisista tuloksista, joita järjestelmässä on tarjolla.

Luokittelun ja asiasanoituksen tapauksessa järjestelmän palauttamia tuloksia vastaavat arvioitavan järjestelmän tuottamat annotaatiot ja relevantteja tuloksia puolestaan testiaineiston alkuperäiset asiasanat. Muodollisemmin tarkkuudeksi p ja saanniksi r saadaan

$$p = \frac{tp}{tp + fp}$$

$$r = \frac{tp}{tp + fn}$$

missä tp on aitojen positiivisten, fp väärien positiivisten ja fn väärien negatiivisten tulosten lukumäärä. Sekä tarkkuus että saanti saavat arvon väliltä $[0, 1]$, ja suurempi arvo tarkoittaa parempaa tulosta.

Yksinkertaisuuden lisäksi tarkkuus- ja saantimittausten etuna on, ettei niiden suorittamiseksi tarvita yhteen kertaan valmiiksi annotoidun testiaineiston lisäksi muuta taustamateriaalia tai -työtä. Ohjattua oppimista hyödyntävää algoritmia käytettäessä annotoitua aineistoa tarvitaan joka tapauksessa jo opetusdataksi, jolloin testidata saadaan jakamalla aineisto satunnaisesti opetus- ja testiosioihin.

4.1.1 Ristiinvalidointi

Jotta datan satunnainen jako opetus- ja testijoukkoihin ei vaikuttaisi tuloksiin merkittävästi, voidaan käyttää ristiinvalidointia. Siinä aineisto jaetaan satunnaisesti k yhtäsuureen osaan, minkä jälkeen koe suoritetaan k kertaa käyttäen kullakin suorituskerralla testiaineistona yhtä aineiston osaa ja opetusaineistona vastaavasti kaikkia muita $k - 1$ osaa. Tätä kutsutaan k-kertaiseksi ristiinvalidoinniksi. Kukin aineiston alkio, esimerkiksi tekstiasiakirja indeksointitermeineen, päätyy mukaan testiaineistoon täsmälleen kerran, ja koko kokeen tulokseksi lasketaan kaikkien suorituskertojen tulosten keskiarvo.

Eräs menetelmän erikoismuoto on niinsanottu yksi pois (leave-one-out) -ristiinvalidointi. Tämä tarkoittaa k-kertaista ristiinvalidointia, jossa k on yhtä suuri kuin aineiston koko, esimerkiksi annotoitavien asiakirjojen lukumäärä. Opetusdatan koko saadaan tällöin maksimoitua, koska kullakin kierroksella vain yksi alkio jätetään testidataksi. Etenkin suurilla aineistoilla menetelmä on kuitenkin laskennallisesti melko raskas, koska testi on suoritettava k kertaa koko aineiston kattamiseksi.

4.1.2 F-luku

Tarkkuus tai saanti eivät välttämättä ole toisistaan erillisinä mielekkäitä suorituskyvyn mittareita, koska kumman tahansa arvon voi yleensä helposti kasvattaa suureksi toisen kustannuksella. Tarkkuuden voi saada korkeaksi palauttamalla koko testissä vaikkapa vain yhden erittäin varman positiivisen tuloksen. Saannin taas saa triviaalisti maksimoitua tuottamalla kaikki mahdolliset tulokset, esimerkiksi palauttamalla jokaisen indeksoitavan dokumentin kohdalla kaikki käytettävän sanaston termit.

Yleinen tapa yhdistää tarkkuus ja saanti yhdeksi arvoksi on niinsanottu F-luku, joka on molempien painotettu harmoninen keskiarvo. Kumpaakin arvoa tasapuolisesti painottavaa erikoistapausta kutsutaan F_1-luvuksi, ja se voidaan laskea kaavalla

$$F_1 = 2\frac{pr}{p+r}$$

missä p on tarkkuus ja r saanti. Myös jompaakumpaa suhteellisesti enemmän painottavat mittarit ovat mahdollisia. Myöhemmin tässä tutkielmassa F-luvulla tarkoitetaan kuitenkin nimenomaan F_1-lukua.

4.1.3 Interpoloitu keskimääräinen tarkkuus

Eräs tarkkuutta ja saantia soveltava suorituskyvyn mittaustapa on muun muassa tiedonhaussa käytetty yhdentoista pisteen interpoloitu keskitarkkuus (eleven-point interpolated average precision). Siinä hakujärjestelmää testataan säätämällä luottamuksen raja-arvoa siten, että järjestelmä antaa tulokset saannin arvoilla {0,0; 0,1; 0,2; ...; 1,0}. Testin tulokseksi lasketaan näin saatujen yhdentoista suorituskerran tarkkuuksien keskiarvo. F-luvun tapaan keskitarkkuus siis huomioi samaan aikaan sekä tarkkuuden että saannin, ja menetelmää onkin ehdotettu käytettäväksi myös tekstin luokittelun arvioinnissa [Seb02].

Pohjimmiltaan termien tekstistä eristämiseen perustuvat indeksointimenetelmät eivät kuitenkaan voi taata mielivaltaisen korkeaa saantitulosta millään luottamustasolla. Kaikki alkuperäiset ihmisten määrittämät asiasanat eivät välttämättä esiinny itse tekstissä, eikä termejä tekstistä eristävä algoritmi tällöin tuota näitä termejä lainkaan. Korkein kulloisellakin algoritmilla saavutettava saantitulos riippuu lisäksi paitsi aineistosta myös ehdokkaiden eristämiseen käytettävästä strategiasta ja vaihtelee algoritmista toiseen. Arviointimenetelmät, jotka perustuvat algoritmin parametrien säätämiseen tiettyjen ennaltamääriteltyjen saantiarvojen saavuttamiseksi

tai kaikkien mahdollisten tulosten palauttamiseksi jossakin mielekkäässä keskinäisessä järjestyksessä, eivät tästä syystä sovellu käytettäviksi avainsanoja tekstistä eristävien indeksointialgoritmien evaluointiin sellaisinaan.

Edellämainitusta syystä interpoloitua keskimääräistä tarkkuutta ei käytetä automaattisen asiasanoituksen evaluointiin tässä tutkielmassa. Tapaustutkimuksen tulosten esittelyn yhteydessä luvussa 6 palataan kuitenkin lyhyesti myös tarkkuuden ja saannin suhteeseen ja sen kehitykseen luottamuksen raja-arvoa säädettäessä.

4.2 Yleiset testiaineistot

Tarkkuus- ja saantitulokset vaihtelevat luonnollisesti aineistosta toiseen. Vertailukelpoisten tulosten saamiseksi eri algoritmien arviointi on suoritettava samalla aineistolla ja standardoidulla koeasetelmalla [Seb02]. Esimerkiksi tiedonhaun arvioinnissa käytetään yleisesti standardiaineistoja, jotka tarjoavat yhtenäisen mittapuun eri algoritmien vertailuun [Voo02].

Avain- tai asiasanojen eristämiseen on julkisesti saatavilla useita eri aineistoja [ZG09, KMKB10]. Laajoja tekstin luokittelun tutkimukseen ja evaluointiin soveltuvia englanninkielisiä aineistoja ovat esimerkiksi Reutersin uutiskorpus[9] [LYRL04] sekä Yhdysvaltain National Library of Medicinen kokoelmiin perustuvat aineistot [ABC+00]. Hulth [Hul03] arvioi indeksointialgoritmiaan Inspec-tietokannasta[10] kerätyillä tieteellisten artikkeleiden tiivistelmillä ja niihin liitetyillä avainsanoilla. Samaa koeaineistoa on käytetty sittemmin melko laajasti muissakin tutkimuksissa [MT04, LLZS09, ZG09, HN10].

Koeaineistot, joita on käytetty yhtenäisesti useiden indeksointimenetelmien arviointiin tai vertailuun, edustavat usein tiettyjä aihealueita tai aineistolajeja. Monien yleiskäyttöisten indeksointialgoritmien evaluoinnissa eri aineistoja on käytetty melko hajanaisesti. Turney [Tur00] käyttää kokeissaan useita eri aineistoja, mukaan lukien tieteellisiä artikkeleita ja sähköpostiviestejä. Osaa samoista aineistoista hyödyntävät myös Kea-algoritmin esitelleet Frank ja kumppanit [FPW+99] sekä omaa menetelmäänsä evaluoineet Barker ja Cornacchia [BC00].

KEA++:n arvioinnissa Medelyan ja Witten [MW06b] käyttävät edellisistä poiketen YK:n elintarvike- ja maatalousjärjestö FAO:n aineistoa, jota he hyödyntävät myös myöhemmässä työssään [MW08, Med09]. Hulthin ja kumppaneiden [HKJ+01]

[9] http://about.reuters.com/researchandstandards/corpus/
[10] http://www.theiet.org/resources/inspec/

käyttämä Ruotsin valtiopäivien aineisto poikkeaa muista tutkimuksista.

Useimmat laajalti käytetyt koeaineistot ovat lisäksi yksinomaan englanninkielisiä. Poikkeuksena tästä FAO:n aineisto on saatavilla myös espanjan- ja ranskankielisenä [MW08].

Erään tavan vertailla eri menetelmien suorituskykyä tarjoavat kilpailut ja haasteet, joissa osallistujien toteuttamia algoritmeja testataan ennalta määrätyllä aineistolla yhdenmukaisessa koeasetelmassa. Vuoden 2008 ECML PKDD Discovery Challengessa[11] eräänä tehtävänä oli vapaamuotoisten avainsanojen ehdottaminen tieteellisille artikkeleille automaattisesti niiden tekstin perusteella. Automaattisen semanttisen analyysin menetelmiin sekä niiden arviointiin ja vertailuun keskittyvässä SemEval-työpajassa[12] käsiteltiin vuonna 2010 niin ikään muun muassa avaintermien eristämistä tieteellisistä artikkeleista [KMKB10].

Molemmat edellä mainitut haasteet keskittyivät kuitenkin tietyn sovellusalan aineistoihin ja niiden annotointiin. Laajoja vertailuja esimerkiksi tässä tutkielmassa mainittujen yleiskäyttöisten annotointialgoritmien kesken ei ole kirjoitushetkellä tiedossa. Vertailukelpoisten tulosten ja standardiaineistojen sijaan algoritmien suorituskyvylle on käytetty vertailukohtana muun muassa ihmisten keskinäistä yhdenmukaisuutta samassa tehtävässä [MW06b].

4.3 Indeksoinnin yhdenmukaisuus

Suoraviivaisessa tarkkuuden ja saannin mittauksessa alkuperäiset, mahdollisesti vain yhden ihmisen määrittämät indeksointitermit oletetaan dokumentin sisällönkuvailun kannalta relevanteiksi ja siten hyviksi asiasanoiksi. Aineistoa kuvaavien termien valinta on kuitenkin subjektiivista, eikä hyvien asiasanojen valinta ole yksikäsitteistä. Yhteisestä sanastosta huolimatta eri indeksoijat valitsevat usein asiasanoiksi eri termejä [ZD69, Rol81], jotka voivat kuitenkin olla annotoijan näkökulmasta riippuen yhtä hyviä tai kuvaavia [Saa02]. Tämä rajoittaa saannin ja tarkkuuden käyttökelpoisuutta automaattisen asiasanoituksen evaluoinnissa, jos kunkin vertailuaineistona käytettävän dokumentin on annotoinut vain yksi henkilö.

Asiasanoituksen subjektiivisuuden ja monikäsitteisyyden vuoksi tarvitaan evaluointikeinoja, joissa huomioidaan annotoijien väliset erot. Eräs tällainen on indeksoijien keskinäistä yhdenmukaisuutta kuvaava Rollingin mitta [Rol81]. Yksittäisten anno-

[11] http://www.kde.cs.uni-kassel.de/ws/rsdc08/
[12] http://www.senseval.org/

toijien keskinäisten yhdenmukaisuuksien perusteella voidaan laskea, kuinka hyvin kunkin indeksoijan asiasanoitus keskimäärin vastaa kaikkien muiden annotoijien valintoja.

Indeksoijien a ja b välinen Rollingin mitta yksittäisen objektin annotoinnissa on

$$r(a,b) = \frac{2|I_a \cap I_b|}{|I_a| + |I_b|}$$

missä I_a on a:n ja I_b puolestaan b:n objektille määrittämien indeksointitermien joukko. Näin laskettuna Rollingin mitta painottaa saantia ja tarkkuutta tasapuolisesti, mutta F-luvun tapaan myös muunlaiset painotukset ovat mahdollisia [Rol81].

Kahden indeksoijan välinen tasapainoinen Rollingin mitta vastaa F_1-lukua [MW08], joten olennainen ero tarkkuuden ja saannin mittaukseen nähden on samanaikainen vertailu useamman indeksoijan kesken. Tällöin myös kunkin termin suhteellista merkittävyyttä asiakirjan sisällön kuvailussa voidaan mitata implisiittisesti sen perusteella, kuinka moni indeksoijista on käyttänyt kyseistä termiä asiasanana [ZD69].

Indeksoijien keskinäistä yhdenmukaisuutta on käytetty pääasiassa informaatiotutkimuksen alalla ihmistyönä tuotetun metatiedon laadun tutkimuksessa, kun taas algoritmien suorituskyvyn evaluointiin on perinteisesti käytetty standardiverrokkina yhteen kertaan annotoitua aineistoa. Myös automaattisen asiasanoituksen laatua voidaan kuitenkin arvioida käsittelemällä algoritmin eristämää metatietoa yhtenä muiden joukossa ja vertaamalla sen ihmisten kanssa saavuttamaa yhdenmukaisuutta ihmisten keskinäisiin tuloksiin [MW08].

Vaatimus useaan kertaan annotoidusta aineistosta on myös Rollingin mitan kaltaisten arviointimenetelmien keskeinen ongelma. Medelyanin väitöstutkimuksessa [Med09] yhden testiaineiston asiasanoitti toisistaan riippumatta kuusi asiantuntijaa, ja toisen aineiston annotoi niin ikään toisistaan riippumatta 15 kahden hengen ryhmää. Myös Sinkkilä ja kumppanit [SSH11] käyttävät aineistoa, jonka kuusi asiantuntijaa oli indeksoinut tutkimusta varten. Usein tällaista aineistoa ei ole valmiiksi saatavilla, joten aineiston valmisteluun vaadittava työpanos on huomattava rajoite yksinkertaisiin tarkkuuden ja saannin mittauksiin verrattuna. Niistä poiketen asiantuntijoiden välisen keskinäisen konsistenssin laskeminen tarjoaa kuitenkin suoran vertailukohdan algoritmin suorituskyvylle.

4.4 Semanttinen samankaltaisuus

Automaattisesti tuotettujen annotaatioiden suora vertailu ihmisen tuottamaan metadataan jättää huomioimatta alkuperäisistä poikkeavat mutta semanttisesti samankaltaiset termit, jotka rinnastuvat tällöin arvioinnissa täysin epäonnistuneisiin annotointeihin [KBK10]. Esimerkiksi käsite ”jalkaväen taisteluajoneuvot” on merkitykseltään täsmällisempi ja rajatumpi kuin pelkkä ”taisteluajoneuvot”, mutta jälkimmäinenkin saattaa silti kuvata tekstin aihepiiriä riittävällä tarkkuudella. Tällöin alkuperäisistä termeistä eroavien mutta merkitykseltään samankaltaisten asiasanojen huomioiminen arvioinnissa voi olla perusteltua.

Merkkijonoina toisistaan eroavien sanojen tai ilmausten semanttisen samankaltaisuuden määritys on yleinen ongelma kieliteknologiassa ja luonnollisen kielen koneellisessa käsittelyssä. Samankaltaisuuden laskentaan voidaan hyödyntää tesauruksissa tai ontologioissa määriteltyjä merkityssuhteita tai laajojen tekstikorpusten tilastollista analyysia [RM09], ja myös molempia lähestymistapoja yhdisteleviä menetelmiä on ehdotettu [Res95, LBM03]. Erilaisten semanttisen etäisyyden laskentaan esitettyjen menetelmien kirjo on laaja, ja aiheesta on runsaasti kirjallisuutta. Tässä yhteydessä esitellään muutama esimerkki tavoista, joilla semanttista samankaltaisuutta on ehdotettu käytettäväksi osana automaattisen indeksoinnin arviointia.

Zesch ja Gurevych [ZG09] ehdottavat arvioinnin välineeksi approksimoitua merkkijonojen vertailua, jossa alkuperäisen termin kanssa täsmälleen samojen avainsanojen lisäksi myös osittain täsmäävät termit huomioidaan arvioinnissa. He erottelevat toisistaan tapaukset, joissa automaattisesti eristetty avainlauseke sisältää jonkin ihmisen määrittämän avaintermin kokonaisuudessaan ja päinvastoin. Menetelmä huomioi toisistaan eroavien lausekkeiden osittaiset päällekkäisyydet ainoastaan sanojen tasolla, eikä merkkitason eroavuuksia esimerkiksi sanojen sisällä sallita. Edellä mainitun esimerkin termit katsottaisiin tällöin automaattisesti osittain täsmääviksi, mutta esimerkiksi sanat ”ote” ja ”tiedote” laskettaisiin täysin erillisiksi.

Ehdottamansa approksimoinnin pätevyyttä Zesch ja Gurevych [ZG09] evaluoivat kyselymuotoisella tutkimuksella, jossa vastaajia pyydettiin kunkin osittain täsmäävän ehdokas-verrokkiparin kohdalla arvioimaan, olisiko automaattisesti eristetyn ehdokkaan käyttö alkuperäisen termin sijaan hyväksyttävää. Evaluoinnin mukaan alkuperäisen termin sisältävistä ehdokkaista 80 % olisi ollut hyväksyttäviä avaintermejä. Päinvastaisessa tapauksessa hyväksyttäviksi arvioitiin kuitenkin vain 44 % ehdokkaista. Jälkimmäisessä tapauksessa kelvottomiksi arvioidut osittaiset täsmäykset olisivat Zeschin ja Gurevychin mukaan useimmiten johtaneet ehdokkaisiin, jotka

olisivat olleet merkitykseltään liian yleisluontoisia alkuperäisiin avainsanoihin verrattuna.

Kim ja kumppanit [KBK10] vertaavat useita n-grammien osittaiseen täsmäykseen perustuvia laadun arvioinnin menetelmiä. Zeschin ja Gurevychin [ZG09] menetelmän lisäksi vertailussa on mukana useita konekääntämisen ja tekstin automaattisen tiivistämisen arviointiin ehdotettuja semanttisen samankaltaisuuden mittareita. Kirjoittajat arvioivat kunkin mittarin soveltuvuutta osittain täsmäävien avainsanaehdokkaiden evaluointiin selvittämällä, kuinka vahvasti ne korreloivat ihmisten arvioiman laadun kanssa. Zeschin ja Gurevychin menetelmä menestyy vertailussa parhaiten. Lisäksi Kim ja kumppanit ehdottavat menetelmään laajennosta, joka huomioisi, missä kohtaa lauseketta päällekkäisyys sijaitsee, ja painottaisi eri osissa ilmeneviä päällekkäisyyksiä eri tavoin. Tästä ei heidän tulostensa mukaan ole kuitenkaan yksiselitteistä hyötyä.

Toisaalta käsitehierarkian sisältävää sanastoa käytettäessä termien eksplisiittisiä ylä- ja alakäsitesuhteita on mahdollista hyödyntää vastaavaan tapaan. Mittarina voidaan käyttää esimerkiksi käsitteiden välisen ylä- ja alakäsitesuhteiden muodostaman polun pituutta tai termien lähimmän yhteisen yläkäsitteen sijaintia hierarkiassa [PPM04].

Jos käytettävissä on moninkertaisesti annotoitu aineisto, semanttisen samankaltaisuuden huomioivia mittareita voidaan myös yhdistää useiden indeksoijien konsistenssia mittaaviin malleihin. Medelyan ja Witten [MW06a] ehdottavat vektorimallia, jossa toisistaan eroavat mutta semanttisesti samankaltaiset termit huomioidaan sanastossa määriteltyjen merkitysuhteiden perusteella. Kahden annotoijan valintoja kuvataan vektoreilla, joiden alkiot vastaavat sanaston termejä. Kukin alkio saa nollasta poikkeavan arvon, jos annotoija on valinnut sitä vastaavan termin asiasanaksi käsiteltävälle dokumentille. Semanttiset suhteet huomioidaan laskemalla toisesta vektorista muunnettu versio, jossa nollasta poikkeavan arvon saavat annotoijan täsmällisten valintojen lisäksi myös valittujen asiasanojen ylä-, ala- ja vieruskäsitteet. Kahden indeksoijan yhdenmukaisuutta kuvataan tämän jälkeen vektoreiden kosinilla.

Täsmällisten osumien sekä ylä- ja vieruskäsitesuhteisiin perustuvien osittaisten täsmäysten suhteelliset painoarvot Medelyan ja Witten [MW06a] määrittelevät kokeellisesti siten, että valitut painot maksimoivat heidän koeaineistossaan usean ihmisen annotaatioille saatavan yhdenmukaisuusluvun. He siis pitävät ihmisten suorittaman indeksoinnin yhdenmukaisuutta eräänlaisena ohjenuorana. He eivät kuitenkaan eva-

luoi mallia itseään kokeellisesti eivätkä vertaa sitä muihin semanttisen samankaltaisuuden huomioiviin mittareihin.

4.5 Asiantuntija-arviot

Suora vertailu alkuperäiseen annotointidataan on laadun mittarina objektiivinen mutta ei huomioi indeksoinnin monikäsitteisyyttä. Rollingin mitalla laatua voidaan arvioida keskinäisen yhdenmukaisuuden kautta, mutta siinäkin ihmisten tekemät annotaatiot oletetaan lähtökohtaisesti relevanteiksi ja hyviksi. Asiantuntijoidenkin tuottama indeksointimetatieto voi kuitenkin olla heikkolaatuista [HZ07]. Aineiston käsittelemän aihealueen asiantuntijoiden tekemissä subjektiivisissa arvioinneissa voidaan huomioida myös alkuperäisen asiasanoituksen laadun vaihtelu ja verrata sitä suoraan automaattisesti tuotettujen annotaatioiden laatuun.

Asiantuntija-arvioissa voidaan soveltaa esimerkiksi asennetutkimuksissa usein käytettävää Likert-asteikkoa [Lik32]. Arviointeja on kuitenkin tehty myös käyttäen pelkkää binääristä jaottelua kelvollisiin ja kelvottomiin annotaatioihin.

Esimerkiksi Turney [Tur00] arvioi GenEx-algoritmin eristämien avainsanojen laatua tarkkuuden ja saannin lisäksi kokein, joissa ihmiset luokittelivat yksittäisiä avaintermejä hyviin ja huonoihin. Noin joka viidenteen termiin vastaajat eivät ottaneet kantaa. Arvioijat olivat algoritmin web-käyttöliittymän verkkosivuilla vierailleita käyttäjiä, jotka pystyivät syöttämään järjestelmän annotoitavaksi minkä tahansa websivun ja evaluoimaan algoritmin eristämät avainsanat. Tutkimus ei tarjoa vertailukohtaa esimerkiksi ihmisten valitsemien avainsanojen laatuun.

Pouliquen ja kumppanit [PSI03] tutkivat luokittelualgoritminsa tuottaman asiasanoituksen laatua sokkokokein, joissa oli mukana sekä automaattisesti tuotettuja että alkuperäisiä ihmisten valitsemia asiasanoja. Asiasanoituksen ammattilaiset arvioivat käsitteitä hyviksi, huonoiksi tai laadultaan tuntemattomiksi. Hyviksi luokiteltujen asiasanojen osalta vastaajilla oli lisäksi mahdollisuus tarkentaa, että merkitykseltään joko suppeamman tai laajemman termin valinta olisi ollut heidän mielestään vielä osuvampi. Vastaavasti yksikäsitteisesti huonon sijaan vastaajat saattoivat luokitella asiasanan huonoksi mutta semanttisesti aiheeseen liittyväksi.

Aronson ja kumppanit [AMG+04] tutkivat automaattisten asiasanaehdotusten onnistumista automatisoitujen kokeiden lisäksi vapaaehtoisvoimin toteutetulla monivaiheisella evaluoinnilla. Ensimmäisessä vaiheessa he selvittivät asiasanoituksen relevanssia ja kattavuutta kokeella, jossa vastaajat määrittivät asiasanoituksen vastaa-

van artikkelin aihepiiriä joko täysin, osittain tai ei lainkaan. Huonoiksi arvioitujen asiasanojen osalta vapaaehtoiset vastasivat lisäksi kyselyyn hylkäämisen syistä. Evaluoinnin viimeisessä osassa arvoinnin toteuttaneet indeksoijat vastasivat vielä yleisluontoiseen kyselyyn puoliautomaattisen indeksointijärjestelmän toimivuudesta ja hyödyllisyydestä.

Yksittäisten asiasanojen lisäksi voidaan evaluoida kaikkien kuhunkin asiakirjaan liitettyjen termien muodostamaa kokonaisuutta. Barker ja Cornacchia [BC00] tutkivat yksittäisten asiasanojen laatua kyselyllä, jossa kaksitoista vastaajaa arvioi laatua kolmeportaisella asteikolla. Lisäksi he vertailevat asiasanojen muodostamia kokonaisuuksia toisiinsa pareittain kokeella, jossa vastaajat valitsivat kokonaisuuksista paremman. Barker ja Cornacchia selvittävät muun muassa yksittäisten avaintermien ja kokonaisuuden arvioidun laadun välistä yhteyttä ja toteavat sen melko heikoksi. Heidän mukaansa kumpikaan ei yksinään riitä evaluoinnin kriteeriksi, ja he ehdottavatkin molempien mittaamista erikseen. Barkerin ja Cornacchian oman algoritmin verrokkina kokeessa on GenEx [Tur99].

Suoraan ihmistyönä tehtävien arviointien heikkoutena on etenkin niiden vaatima suuri työmäärä, joka voi vaikeuttaa riittävän otoskoon saamista ja tilastollisesti merkittävien tulosten syntyä [BC00]. Automaattisesti suoritettavat mittaukset soveltuvat myös ihmistyönä suoritettuja arviointeja paremmin tilanteisiin, joissa kokeet halutaan toistaa useilla erilaisilla annotointialgoritmeilla tai niiden muunnoksilla esimerkiksi algoritmin parametrien hienosäätämiseksi [ZG09].

4.6 Sovellustason arviointi

Viime kädessä metatiedon laadun kenties tärkein mittari on annotaatioiden vaikutus niitä hyödyntävän sovelluksen toimintaan ja käytettävyyteen. Jos asiasanoituksen ensisijainen käyttötarkoitus on esimerkiksi tukea tiedonhakujärjestelmää, metatiedon laatua voidaan arvioida mittaamalla sen vaikutusta tiedon löydettävyyteen tiedonhaun tutkimuksen menetelmin [BRK05]. Tässä tutkielmassa ei oteta kantaa metatiedon täsmällisiin sovelluskohtaisiin käyttötarkoituksiin, ja sovellustason arvoinnit sivuutetaan.

5 Tapaustutkimus: puolustusvoimien normit

Semanttisen webin ja yhdistetyn tiedon tutkimuksessa painopiste on usein avoimessa tiedossa ja sen julkistamiseen, jakamiseen ja yhdistelemiseen liittyvissä tekniikoissa. Vastaavia ontologioita, metatiedon esitystapoja ja ohjelmistokomponentteja voidaan kuitenkin käyttää myös ei-julkisten tietojen käsittelyssä esimerkiksi organisaatioiden sisäisissä tietojärjestelmissä.

FinnONTO 2.0 -tutkimushankkeen[13] eräänä soveltavana tapauksena selvitettiin semanttisen webin tekniikoiden käyttöä puolustusvoimien sisäisen asiakirja-aineiston ja sen metatietojen käsittelyssä ja käytössä [FHW11]. Osana tutkimustyötä kokeiltiin tekstiasiakirjojen automaattista asiasanoitusta ja arvioitiin tuotetun metatiedon laatua luvussa 4 kuvatuin menetelmin. Tässä luvussa esitellään tutkimuksen taustaa ja siinä käytettyä aineistoa, asiakirjojen automaattista annotointia Mauilla [Med09] sekä aineiston erityispiirteitä automaattisen asiasanoituksen näkökulmasta. Tapaustutkimuksen yhteydessä tehdyn asiasanoituksen laadun arvioinnin tulokset ja niiden analyysi esitetään luvussa 6.

5.1 Tutkimusaineisto

Tutkimusaineistona käytettiin puolustusvoimien asianhallintajärjestelmästä saatua otosta organisaation sisäisistä normeista. Asiakirjat käsittelevät esimerkiksi toimintatapoja, turvallisuusmääräyksiä ja hallinnollisia ohjeita. Otokseen kuului yhteensä 3 904 suomenkielistä normia metatietoineen, jotka koostuivat muun muassa asiakirjojen otsikoista, kutakin normia kuvaavista asiasanoista ja erilaisista asian käsittelyyn liittyvistä kentistä kuten päivämääristä ja toimijoista.

Koko vajaan neljäntuhannen normin kokoelmasta vain osa soveltui automaattisen annotoinnin testaamiseen. Noin puoleen normien metatiedoista ei ollut liitetty lainkaan asiasanoja. Lisäksi osasta otoksen asiakirjoja puuttui teknisistä tai salassapitoon liittyvistä syistä kokonaan varsinaiset asiakirjatekstit, vaikka metatietoja olikin saatavilla. Tällaisten asiakirjojen metatiedot käsiteltiin esimerkiksi muunnoksen osalta samoin kuin muukin aineisto, mutta automaattisen asiasanoituksen evaluointiin käytetystä opetus- ja testiaineistosta ne rajattiin pois.

Myös asiasanoja sisältäneissä normien metatiedoissa indeksointitermejä oli usein melko vähän, esimerkiksi vain yksi. Vastaavasti osassa tekstiasiakirjan sisältäneistä

[13] http://www.seco.tkk.fi/projects/finnonto/index.fi.php

normeista varsinainen teksti oli esimerkiksi saamamme otoksen ulkopuolelle jääneessä liitteessä, ja saatavilla ollut tekstiosuus saattoi koostua vaikkapa vain varsinaisen asiakirjan sijainnin kertovasta tiedosta.

Varsinaiseen evaluointiin käytetystä opetus- ja testiaineistosta rajattiin pois paitsi tyystin asiasanoittamattomat normit myös ne, joihin oli joko liitetty vähemmän kuin kaksi asiasanaa tai joiden asiakirjatekstin yhteenlaskettu koko oli puhtaaksi tekstitiedostoksi muunnettuna alle 3 000 tavua. Tekstin pituuden alaraja valittiin tarkastelemalla tekstimuodoltaan lyhimpien normien sisältöä ja määrittämällä karkeasti, kuinka pitkä tekstitiedoston vähintään oli oltava, jotta se sisältäisi tekstiin itseensä sisältyvien otsaketietojen lisäksi ainakin jonkin verran normin varsinaista tekstisisältöä.

Rajausten jälkeen alkuperäisistä vajaasta neljästätuhannesta normista opetus- ja testiaineistoksi soveltuvia oli yhteensä 485. Asiakirjojen metatiedoissa oli yhteensä 1420 asiasanaa, joten normia kohden asiasanoja oli keskimäärin noin 2,9.

5.2 Aineiston valmistelu

Olemassa olevat metatiedot muunnettiin XML-muodosta RDF-muotoon tarkoitusta varten kehitetyllä muunnostyökalulla. Varsinaiset asiakirjatekstit puolestaan muunnettiin automaattista annotointia varten alkuperäisestä Microsoft Word -muodosta muotoilemattomaksi tekstiksi avoimen lähdekoodin wvWare-ohjelmistoon[14] kuuluvalla komentorivityökalulla.

Yksittäisen normin varsinainen tekstisisältö koostuu yhdestä tai useammasta liitetiedostosta, mutta metatiedot asiasanoineen koskevat kuitenkin normia kokonaisuutena. Jos normin teksti jakautui useampaan kuin yhteen tiedostoon, yksittäiset tekstitiedostot katenoitiin muunnoksen jälkeen yhdeksi tiedostoksi.

Puolustusvoimien asianhallintajärjestelmässä käytetään toimialan keskeiset termit käsittävää asiasanastoa. Sanaston pohjalta on aiemmin FinnONTO 2.0 -hankkeessa kehitetty Puolustushallinnon ontologia PUHO, joka koostuu Yleisestä suomalaisesta ontologiasta YSO:sta täydennettynä puolustushallinnon sanaston toimialakohtaisilla termeillä. Valtaosa normien metatiedoissa esiintyvistä asiasanoista on asiakirjan laatimisen yhteydessä valittu puolustushallinnon asiasanaston termien joukosta. Ennalta määrättyjen termien lisäksi laatijan on kuitenkin mahdollista lisätä metatietoihin myös vapaavalintaisia avainsanoja, jos sanaston tarjoamat termit eivät hänen

[14] http://wvware.sourceforge.net/

mielestään kuvaa asiakirjan sisältöä riittävän tarkasti.

Alkuperäisessä XML-muotoisessa metadatassa sekä asiasanat että vapaasti valitut avainsanat on esitetty tekstiliteraaleina. Metatietojen muunnostyökalu korvaa literaalit viittauksilla ontologian käsitteisiin seuraavalla nelivaiheisella prosessilla:

1. Ontologian käsitteiden perusmuotoistaminen
2. Metadatan esikäsittely
3. Asiasanaliteraalien perusmuotoistaminen
4. Literaalien kuvaus ontologiaan

Ensimmäisessä vaiheessa jokaisen ontologiassa määritellyn käsitteen tekstinimike (label) lemmatisoitiin eli perusmuotoistettiin. Käytetty ontologia ei alunperin sisältänyt käsitteiden nimikkeitä perusmuodoissaan, vaan käsitteiden nimikkeet on pääsääntöisesti ilmaistu monikossa. Nimikkeiden lemmat eli perusmuodot lisättiin ontologiaan alkuperäisten ensisijaisten nimikkeiden rinnalle. Perusmuotoistamiseen käytettiin Connexor Oy:n kaupallista FDG-jäsennintä [TJ97].

Itse metadataa muunnettaessa asiasanoina esiintyvät tekstiliteraalit normalisoitiin samaan tapaan perusmuotoihinsa, minkä jälkeen kullekin asiasanalle etsittiin vastaavaa käsitettä ontologiasta vertaamalla literaalia käsitteiden perusmuotoistettuihin nimikkeisiin täsmällisellä merkkijonovertailulla. Jos käsite löytyi, literaali korvattiin RDF-mallissa asianmukaisella viittauksella ontologiseen käsitteeseen. Muussa tapauksessa termin oletettiin olevan vapaa, asiasanaston ulkopuolelta valittu avainsana. Tällaiset termit lisättiin RDF-malliin omina erillisinä resursseinaan. Saman vapaan avainsanan erilliset esiintymät saatiin siten samastettua korvaamalla ne viittauksilla yhteiseen resurssiin, vaikka sen semanttisia yhteyksiä ontologian käsitteisiin ei voitukaan automaattisesti määrittää. Automaattisen annotoinnin opetus- ja testiaineistoon sisällytettiin kuitenkin vain ontologiassa esitellyt asiasanat.

Luonnollisen kielen sanat ovat usein monimerkityksisiä. Esimerkiksi kuusi voi viitata joko puulajiin tai numeroon, ja häviäminen puolestaan voi tarkoittaa joko tappion kärsimistä tai katoamista. Kirjoitusasultaan samanlaiset mutta merkitykseltään eriävät termit voidaan määritellä ontologiassa erillisiksi käsitteiksi. Algoritmi, joka pyrkii löytämään sanastosta metatiedoissa tai tekstissä esiintyvää tekstiliteraalia vastaavat käsitteet, voi siis joutua valitsemaan useista kirjoitusasultaan samanlaisista vaihtoehdoista merkitykseltään sopivimman. Tätä kutsutaan sanan yksikäsitteistämiseksi (word sense disambiguation) [IV98].

Myös YSO sisältää runsaasti termejä, jotka vastaavat perusmuotoistetulta kirjoitusasultaan jotakin muuta, merkitykseltään eriävää käsitettä. Eri käsitteet on siis määritelty erillisiksi resursseiksi, jotka voivat sijaita eri osissa sanaston hierarkkista rakennetta. Perusmuodoltaan samankaltaisten termien nimikkeisiin on myös lisätty sulkeissa esitetty selvennys, joka tarkentaa termin merkitystä. Esimerkiksi joukkueella voidaan tarkoittaa urheilijaryhmää tai sotilasosastoa. Termin kahta eri merkitystä vastaavien YSO-käsitteiden ensisijaiset nimikkeet ovat ”joukkueet (urheilu)”[15] ja ”joukkueet (sotilasosastot)”[16]. Molemmille resursseille on lisäksi määritelty vastaavuussuhde koostekäsitteeseen, jonka nimike on ”joukkueet (kooste)”[17]. Vastaavankaltaisia koostekäsitteitä on YSO:ssa kirjoitushetkellä yhteensä useita satoja.

Normien metatiedoista ontologiassa käytetyt tarkenteet kuitenkin puuttuivat. Pelkällä täsmällisellä merkkijonovertailulla asiasanoja ei siis voitu kuvata ontologiaan yksikäsitteisesti, jos perusmuodoltaan asiasanaa vastaavia käsitteitä oli useampia. Kaikista metatiedoissa esiintyneistä 362 uniikista asiasanasta tällaisia oli kuitenkin vain kymmenen. Koska näiden merkitys automaattisen annotoinnin kokonaistulosten kannalta lienee tästä syystä vähäinen, oikean käsitteen automaattiseen valintaan liittyvät ongelmat sivuutettiin metatietojen muunnoksen yhteydessä. Kun normien asiasanoja yhdistettiin ontologian käsitteisiin, mahdolliset tarkenteet sulkumerkkeineen poistettiin nimikkeestä ennen merkkijonovertailua, ja kukin asiasana korvattiin RDF-mallissa viittauksella ensimmäiseen perusmuodoltaan asiasanaa vastaavaan käsitteeseen.

5.3 Automaattinen annotointi

RDF-muunnoksen jälkeen aineisto asiasanoitettiin automaattisesti käyttäen aiemmin FinnONTO 2.0 -projektissa kehitettyä ARPA-verkkopalvelua[18], jonka tarkoitus on tarjota eri annotointimoottoreille yhtenäinen web-palvelurajapinta. Varsinaisena annotointimoottorina käytettiin Mauin [Med09] versiota 1.2, jonka lähdekoodi[19] on julkisesti saatavilla verkossa.

[15] http://www.yso.fi/onto/yso/p2085
[16] http://www.yso.fi/onto/yso/p22577
[17] http://www.yso.fi/onto/yso/p22578
[18] http://www.seco.tkk.fi/services/arpa/
[19] http://code.google.com/p/maui-indexer/

5.3.1 Koeasetelma

Maui edellyttää käytettävien rakenteellisten sanastojen olevan SKOS-muotoisia, joten ennen annotointikokeita ontologia muunnettiin SKOS-muotoon Skosify-työkalulla [SH12]. Muunnettu sanasto perusmuotoistettiin FDG:tä [TJ97] käyttäen.

Mauin opetusvaiheessa teksti perusmuotoistettiin niin ikään FDG:llä. Varsinaiset annotointikokeet toistettiin käyttäen tekstin lemmatisointiin sekä Omorfia [LSP09] että FDG:tä.

Omorfin tapauksessa koko annotoitava tekstiasiakirja lemmatisoitiin esikäsittelynä ennen tekstin syöttämistä Mauille. FDG-jäsentäjää puolestaan käytettiin Mauiin lisättynä sanojen normalisointitoteutuksena, jolloin perusmuotoistus tehtiin Mauin yleisesti käyttämään tapaan tekstiä läpikäytäessä yksittäisten n-grammien tasolla. Kokeiden erilaiset toteutukset johtuivat FDG-verkkopalvelussa havaitusta ongelmasta, joka johti virheellisiin tuloksiin, jos lemmatisointi suoritettiin ARPA-järjestelmässä esikäsittelynä. ARPA:n kokeissa käytetty versio puolestaan tukee Mauin sisäisenä normalisointitoteutuksena ainoastaan FDG:tä. Erilaisten koeasetelmien mahdollisia vaikutuksia tuloksiin pohditaan tulosten analysoinnin yhteydessä luvussa 6.

Metatietojen muunnoksen yhteydessä mainittu monikäsitteisyyden ongelma koskee myös automaattista annotointia. Tekstissä esiintyvät sanat voivat olla monimerkityksisiä, ja ontologia voi sisältää useita eri käsitteitä, joiden nimikkeet ilman tarkenteita vastaavat tekstissä esiintyviä merkkijonoja. Maui ei pääsääntöisesti onnistunut määrittämään, mikä mahdollisista vaihtoehtoisista käsitteistä on kulloisessakin tapauksessa oikea, ja palautti tuloksissa kaikki lauseketta perusmuotonsa perusteella vastaavat käsitteet.

Useita vaihtoehtoisia käsitteitä sisältävät indeksointitulokset olivat melko harvinaisia, ja yksinkertaisuuden vuoksi monikäsitteisyyden ongelma tässä tapauksessa sivuutettiin. Maui edellyttää opetusdatan olevan muodossa, jossa alkuperäiset asiasanat esitetään tekstiliteraaleina käsitteet yksilöivien URI:en sijaan. Tästä syystä sekä opetus- ja testidatana käytettyjä alkuperäisiä asiasanoja että tekstistä eristettyjä termejä käsiteltiin annotointikokeissa tekstiliteraaleina. Tuloksia arvioitaessa Mauin eristämäksi termiksi katsottiin vastaavasti eristetyn käsitteen nimike ilman mahdollista sulkumerkeissä ollutta tarkennetta. Näin tehtiin myös silloin, kun vaihtoehtoisia eristettyjä käsitteitä oli useita, ja tulosten analysoinnissa vaihtoehtoiset käsitteet katsottiin yhdeksi annotoinniksi.

Maui tuottaa eristämiensä asiasanojen lisäksi kullekin ehdottamalleen termille luot-

tamusarvon. Korkeamman luottamusarvon saanut ehdokas on opetusdatasta rakennetun tilastollisen mallin mukaan todennäköisemmin asiakirjalle relevantti asiasana. Algoritmin tuottamat indeksointitermit voidaan rajata joko määrättyyn lukumäärään tai pelkästään niihin termeihin, joiden luottamusarvo ylittää annetun ylärajan.

Aiemmissa Mauin suorituskykyä arvioineissa kokeissa asiasanat on rajattu lukumäärän perusteella esimerkiksi viiteen tai kuuteen parhaan luottamusarvon saaneeseen termiin [Med09, SSH11]. Tuolloin verrokkina käytetyt annotaatiot on kuitenkin tehty nimenomaan tutkimuskäyttöön, ja siten myös alkuperäisiä annotaatioita on ollut vastaava määrä kutakin dokumenttia kohden. Tässä tapauksessa rajaus päädyttiin asettamaan luottamusarvon perusteella, koska alkuperäisten annotaatioiden lukumäärä vaihtelee huomattavasti asiakirjasta toiseen.

Tarkkuutta ja saantia mitannut ristiinvalidointikoe toistettiin kaikilla luottamuksen raja-arvoilla 2,5 prosenttiyksikön välein. Alustavissa kokeissa korkein F-luku saavutettiin asettamalla raja-arvoksi 0,15, ja subjektiiviseen evaluointiin käytettiin Mauin tässä koeasetelmassa tuottamia asiasanoja. Perusmuotoistukseen käytettiin Omorfia.

5.3.2 Indeksointialgoritmin muunnelmat

Kahden eri lemmatisointistrategian lisäksi tarkkuus- ja saantikokeet toistettiin myös kahdella muulla indeksointialgoritmiin tehdyllä muunnelmalla sekä niiden yhdistelmällä. Ensimmäinen muunnos oli lisätä ehdokkaiden luokittelussa käytettäviin piirteisiin termin sijainti ontologian käsitehierarkiassa. Toisena variaationa kokeiltiin luokittelijoina käytettyjen päätöspuiden käsittelyä virhettä pienentävällä karsimisella (reduced error pruning, REP) [Qui87]. Karsimisalgoritmin toteutus sisältyy Mauin käyttämään WEKA-kirjastoon.

Alustavia tuloksia tarkasteltaessa aineiston ja sen aihealueen asiantuntijat mainitsivat eräänä huolenaiheenaan joidenkin automaattisesti eristettyjen asiasanojen olevan merkitykseltään liian laaja-alaisia, jolloin termit eivät rajaa asiakirjan aihepiiriä riittävän tarkasti. Liiallisen yleisluontoisuuden on myös aiemmin arvioitu olevan merkittävä automaattisen indeksoinnin laatua heikentävä tekijä [AMG+04]. Tämän vuoksi päätettiin testata hypoteesia, jonka mukaan termin hierarkkinen asema voisi olla hyödyllinen piirre asiasanaehdokkaiden järjestyksen määrittämisessä.

Medelyan [Med09, s. 136] toteaa Mauin indeksointitulosten parantuvan, kun mallista jätetään pois joitakin piirteitä, vaikka piirteet yksittäisinä parantaisivatkin mallin

ennustusvoimaa. Syyksi hän mainitsee mallin ylisovittumisen kaikkia piirteitä käytettäessä. Tällöin malli kuvaa opetusdatan yksityiskohtaisia piirteitä liiankin tarkasti, eikä malli enää yleisty ennustamaan opetusdatan ulkopuolista aineistoa. Siten mallin tarkkuus varsinaisen testiaineiston luokittelussa kärsii. Päätöspuiden karsiminen voi vähentää ylisovittumista, ja tämän vuoksi kokeiltiin myös yksinkertaisen karsinta-algoritmin käyttöä.

Mauissa voidaan käyttää termin merkityksen yleisyyttä eli geneerisyyttä yhtenä piirteenä asiasanaehdokkaiden suodatuksessa, joskin vain silloin, kun sanastona käytetään Wikipedian artikkeleiden otsikoita [Med09, s. 100–101]. Tällöin termin geneerisyys voidaan laskea käyttämällä hyväksi käsitettä vastaavan Wikipedia-artikkelin sijaintia tietosanakirjan kategoriahierarkiassa.

Tässä esiteltyjen annotointikokeiden yhteydessä selvitettiin, voisiko ontologian sisältämää ylä- ja alakäsitehierarkiaa hyödyntää vastaavaan tapaan asiakirjojen indeksoinnin tarkkuuden parantamiseksi. Lähempänä puumaisen hierarkian juurta sijaitsevat käsitteet ovat merkitykseltään yleisempiä kuin ne, jotka sijaitsevat samassa puun haarassa lähempänä lehtisolmuja. Vastaavasti hierarkiassa syvemmällä olevat ovat termit ovat merkitykseltään rajatumpia tai erityisempiä. Koe toteutettiin muokkaamalla Mauia siten, että SKOS-sanastoa käytettäessä geneerisyyspiirteen arvot laskettiin Wikipedian kategorioiden sijaan sanaston hierarkian perusteella.

Mauin hyödyntämä Wikipedia Miner -kirjasto[20] määrittelee semanttisen erityisyyden eli spesifisyyden mitaksi kategorian etäisyyden hierarkian juuresta jaettuna kategoriapuun syvyydellä. Tällöin spesifisyysarvo on lähellä nollaa kategorioille, jotka sijaitsevat lähellä hierarkian juurta, ja kasvaa kohti yhtä lehtisolmuja lähestyttäessä. Myös ontologian ylä- ja alakäsitesuhteiden muodostaman verkon voi nähdä puumaisena rakenteena, jossa tavanomaisesta puusta poiketen yhdellä solmulla voi olla kuitenkin useampi kuin yksi vanhempi. Esimerkiksi Yleisessä suomalaisessa ontologiassa YSO:ssa urheiluhallit on tutkielman kirjoitushetkellä määritelty sekä liikuntatilojen että julkisten rakennusten alakäsitteeksi.

Puun tapaan YSO:n hierarkiassa on hyvin määritelty juurisolmu, joka on epäsuorasti kaikkien muiden käsitteiden yläkäsite. Siten kullekin termille voidaan määritellä rakenteessa syvyys ja korkeus lähestulkoon samoin kuin puussa. Käsitteen korkeus on siis pisin yksinkertainen polku käsitettä vastaavasta solmusta johonkin lehtisolmuun. Puusta poiketen polku solmusta juureen ei kuitenkaan ole yksikäsitteinen. Johdonmukaisuuden vuoksi myös käsitteen syvyys on tässä tapauksessa määritelty

[20] http://wikipedia-miner.cms.waikato.ac.nz/

pisimmäksi yksinkertaiseksi poluksi solmusta hierarkian juureen.

Pisimmän yksinkertaisen polun laskeminen verkossa on yleisessä tapauksessa NP-kova ongelma. Suunnatussa syklittömässä verkossa pisin polku on sen sijaan mahdollista löytää polynomisessa ajassa verkon solmujen lukumäärän suhteen esimerkiksi negatoimalla kaarien painot ja laskemalla muunnetussa verkossa kokonaispainoltaan lyhyimmät solmujen väliset polut.

Kirjoitushetkellä ajantasainen SKOS-määrittely[21] ei kiellä refleksiivisiä ylä- ja alakäsitesuhteita. Määrittelyn mukaan käsite voi siis olla laajentavassa tai tarkentavassa suhteessa itseensä, jolloin sanaston käsitehierarkia voi periaatteessa sisältää syklejä rikkomatta SKOS-määritystä. Sanaston esikäsittelyyn käytetty Skosify kuitenkin poistaa ylä- ja alakäsitehierarkioista mahdolliset syklit, koska ne voivat aiheuttaa ongelmia hierarkian automaattisessa käsittelyssä [SH12].

Hierarkkisten suhteiden lisäksi käsitteiden välisiä etäisyyksiä laskettaessa huomioitiin mahdolliset verkon solmujen väliset ekvivalenssisuhteet. Vastaavuussuhteiden muodostamat syklit poistettiin erikseen yhdistämällä keskenään ekvivalenteiksi määritellyt käsitteet yhdeksi solmuksi. Tällöin hierarkiaa voitiin käsitellä suunnattuna syklittömänä verkkona. Muunnettua ontologiaa käytettiin ainoastaan käsitteiden korkeuksien ja syvyyksien laskemiseen, joten esikäsittely ei aiheuta ongelmia ontologian muun käytön kannalta.

5.4 Arviointimenetelmien valinta

Annotoinnin laatua arvioitiin vertaamalla tuloksia alkuperäiseen asiasanoitukseen tarkkuus- ja saantimittauksin sekä sokkokokeilla.

Laadun perusmittareina päädyttiin käyttämään tarkkuutta ja saantia esimerkiksi Rollingin mitan sijaan, koska alkuperäinen tutkimusaineisto oli peräisin todellisesta tietojärjestelmästä, jossa metatiedot oli määritelty kullekin asiakirjalle vain yhteen kertaan. Indeksoijien keskinäistä konsistenssia hyödyntäviä menetelmiä ei siten voitu käyttää ilman alaan perehtyneiden asiantuntijoiden merkittävää ylimääräistä työpanosta aineiston moninkertaiseksi annotoimiseksi.

Myöskään käsitteiden semanttiseen samankaltaisuuteen perustuvia mittareita ei tässä yhteydessä käytetty, koska erilaisten mittareiden suhteesta indeksointitermien todelliseen relevanssiin ja siten metatiedon laatuun ei ole yksikäsitteistä selvyyttä.

[21] http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Mittareiden ja subjektiivisen laadun keskinäisen vastaavuuden selvittäminen puolestaan edellyttäisi laajaa erillistä tutkimusta, joka jää tämän tutkielman ulkopuolelle.

Tarkkuuden ja saannin mittaukset toteutettiin kymmenkertaisena ristiinvalidointina, johon päädyttiin yksi pois -ristiinvalidoinnin sijaan jälkimmäisen laskennallisen raskauden vuoksi. Luvussa 4.1 mainituista tarkkuuden ja saannin mittausten rajoitteista johtuen tuloksia täydennettiin subjektiivisella laatuarvioinnilla. Tähän päädyttiin, koska useiden indeksoijien yhdenmukaisuuteen perustuvia menetelmiä ei voitu käyttää. Sokkokokeessa saadun datan vähäisyyden vuoksi kokeen nojalla ei voida esittää vahvoja päätelmiä automaattisesti eristetyn metatiedon laadusta. Koetta voikin pitää pilottitutkimuksena, jonka tuloksia voidaan käyttää ohjaamaan tulevia tutkimusaiheita.

Asiantuntija-arvioissa koeaineistoksi valittiin satunnaisesti kolmekymmentä automaattisesti asiasanoitettua normia ja sekä alkuperäiset ihmisen määrittämät että Mauin eristämät asiasanat. Asiasanoja oli yhteensä 144, joista 76 oli koneellisesti tuotettuja ja 87 ihmisen valitsemia. Yhdeksääntoista termiin olivat päätyneet sekä alkuperäiset asiasanoittajat että Maui. Arvioijat jättivät vastaamatta yhteensä kymmenen asiasanan kohdalla, joten subjektiivisen evaluoinnin osalta vertailukelpoisia annotaatioita kertyi lopulta 134 kappaletta.

Sokkokoe toteutettiin taulukkomuotoisena kyselynä, jossa asiakirjojen keskinäinen järjestys oli sekoitettu. Samoin kunkin asiakirjan indeksointitermien keskinäinen järjestys oli satunnainen. Evaluoinnin teki toisistaan riippumatta kaksi asiantuntijaa, jotka arvioivat kunkin indeksointitermin kohdalla väittämän ”tämä termi on hyvä asiasana mainitulle normille” paikkansapitävyyttä neliportaisella Likert-tyyppisellä asteikolla. Kokeen vastausvaihtoehdot olivat ”täysin eri mieltä” (1), ”jokseenkin eri mieltä” (2), ”jokseenkin samaa mieltä” (3) ja ”täysin samaa mieltä” (4). Kokeen tulosten perusteella esitettäviä päätelmiä rajoittaa etenkin se, ettei evaluointiin ollut saatavilla useita henkilöitä, eikä arvioijien voida olettaa olevan keskenään samanmielisiä [BC00, AMG+04].

Vastausvaihtoehtoihin viitataan myöhemmin tässä tutkielmassa edellä mainituilla numeroilla tai lyhenteillä TEM, JEM, JSM ja TSM. Tuloksia analysoitaessa on kuitenkin huomioitava, etteivät vastauskategoriat ole varsinaisesti luonteeltaan numeerisia vaan ordinaalisia: eri vaihtoehtojen keskinäinen järjestys on hyvin määritelty, mutta niiden välinen etäisyys ei välttämättä ole mielekäs. Tämä rajoittaa periaatteessa datan analyysiin soveltuvia tilastollisia menetelmiä [Ste46], joskin aiheesta

on käyty pitkällistä ja vilkasta debattia [VW93, Han96]. Käsittelemme analyysissa vastausvaihtoehtoja pääosin ordinaalisena numeerisen datan sijaan.

Kyselyn vastausvaihtoehtojen määrän valinta on jossain määrin mielivaltainen. Eri alojen kirjallisuudessa on keskusteltu paljon niin vastausvaihtoehtojen optimaalisesta lukumäärästä [LG75, Cox80, TSV99] kuin siitäkin, onko syytä käyttää parillista vai paritonta määrää vaihtoehtoja [Gar91]. Keskeisin ero parillisten ja parittomien asteikkojen välillä on, että neutraalin vaihtoehdon puuttuessa parillinen asteikko pakottaa vastaajan tekemään valinnan positiivisen ja negatiivisen vaihtoehdon välillä.

6 Tulokset ja analyysi

Tässä luvussa esitellään ja analysoidaan tapaustutkimuksen annotointikokeissa saatuja tuloksia.

6.1 Tarkkuus- ja saantimittaukset

Tarkkuuden ja saannin mittaukset suoritettiin yhteensä kuudessa eri koeasetelmassa. Mauin alkuperäisen indeksointialgoritmin lisäksi kokeiltiin kahta eri muunnosta, joista ensimmäinen oli ontologian käsitehierarkian hyödyntäminen luokittelussa ja toinen oli päätöspuiden karsinta. Kaikki kokeet toistettiin erikseen käyttäen perusmuotoistukseen Omorfia ja FDG:tä. Vertailun vuoksi Mauin alkuperäistä muuntelematonta algoritmia kokeiltiin myös ilman perusmuotoistusta tai typistämistä. Kokeiden tulokset on esitetty taulukossa 1.

Tarkkuudella ja saannille tarkoitetaan tässä tapauksessa laskennallisia tuloksia, jotka saatiin vertaamalla eristettyjä termejä alkuperäisiin asiasanoihin. Intuitiivisesti tulkittuna tarkkuus on korkeampi asetelmassa, jossa suurempi osa automaattisesti eristetyistä termeistä vastaa jotakin alkuperäistä asiasanaa. Pelkkä tarkkuus ei siis kuvasta suoraan esimerkiksi pelkän ehdokkaiden luokitteluun käytetyn koneoppimismenetelmän kykyä erotella epäolennaiset ehdokkaat olennaisista, vaan laskennalliseen tarkkuuteen ja saantiin vaikuttavat luokittelijan lisäksi sekä koko annotointialgoritmi kaikkine vaiheineen että lopullisten tulosten valikointiin käytetty rajauskriteeri.

Jotkin annotointialgoritmiin tehdyt muunnokset saattavat esimerkiksi pienentää luokittelijan laskemia luottamusarvoja monien ehdokkaiden kohdalla, jolloin pienempi osa eristetyistä ehdokkaista ylittää raja-arvoksi valitun 0,15:n. Tällöin tark-

Algoritmi	Perusmuotoistus	Tarkkuus	Saanti	F-luku
Maui	–	0,274	0,146	0,190
Maui	Omorfi	0,254	0,144	0,183
Maui + karsinta	Omorfi	**0,331**	0,120	0,177
Maui + hierarkia	Omorfi	0,297	0,144	0,194
Maui + hierarkia + karsinta	Omorfi	0,327	0,119	0,174
Maui	FDG	0,175	**0,242**	0,203
Maui + karsinta	FDG	0,314	0,207	**0,249**
Maui + hierarkia	FDG	0,277	0,159	0,202
Maui + hierarkia + karsinta	FDG	0,308	0,209	0,249

Taulukko 1: Tarkkuus ja saanti annotointimenetelmän eri muunnelmilla

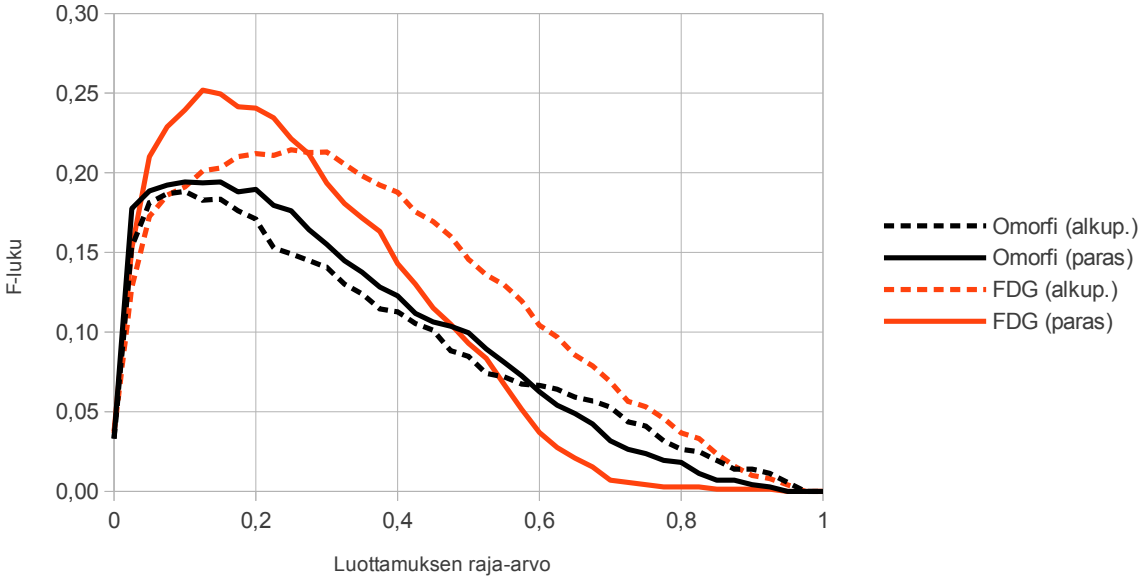
kuus voi parantua ja saanti vastaavasti huonontua, vaikka algoritmin tuottamat tulokset eivät välttämättä muuttuisi olennaisesti esimerkiksi ehdokkaiden keskinäisen järjestyksen suhteen. Tämä on syytä huomioida tuloksia tulkittaessa. Kuva 2 esittää tarkkuuden ja saannin yhdistävän F-luvun kehityksen joillakin indeksointimenetelmän muunnelmilla, kun luottamuksen raja-arvoa säädettiin välillä nollasta yhteen.

6.1.1 Perusmuotoistusmenetelmän vaikutus

Etenkin saanti jäi heikoksi kaikissa koeasetelmissa, kun perusmuotoistukseen käytettiin Omorfia. Mauin alkuperäisen algoritmin saanti jäi vain 14,4 prosenttiin. Saantitulos kuitenkin parani, kun lemmatisointiin käytettiin FDG-jäsentäjää, ja korkein 24,2 prosentin saantitulos saavutettiinkin FDG:n ja Mauin alkuperäisen indeksointialgoritmin yhdistelmällä. Toisaalta annotoinnin tarkkuus pääsääntöisesti huonontui vastaavaan Omorfilla saatuun tulokseen nähden. F-luku oli kaikissa tapauksissa hieman parempi FDG:tä käytettäessä.

Sekä FDG:tä että Omorfia käytettäessä Maui eristi huomattavasti suuremman joukon termiehdokkaita kuin täysin ilman perusmuotoistusta. Etenkin Omorfin tapauksessa suuri osa perusmuotoistuksen ansiosta löydetyistä lisäehdokkaista kuitenkin jäi luottamusarvoltaan alle raja-arvona käytetyn 0,15:n, jolloin tarkkuus- ja saantiluvut eivät rajauksen jälkeen poikenneet huomattavasti ilman perusmuotoistusta saaduista tuloksista.

FDG:n yhteydessä päätöspuiden virhettä pienentävä karsinta paransi tarkkuutta



Kuva 2: F-luku luottamuksen alarajan funktiona alkuperäisellä annotointialgoritmilla ja parhaat tulokset tuottaneilla muunnelmilla. Omorfia käytettäessä korkein F-luku saavutettiin hierarkiapiirteen kanssa ja FDG:n tapauksessa puolestaan päätöspuiden karsinnalla ilman hierarkiapiirrettä.

selvästi, ja saannin lievästä laskusta huolimatta myös F-luku nousi viisi prosenttiyksikköä alkuperäiseen luokittelijaan nähden. Myös Omorfin tapauksessa karsinta paransi annotoinnin tarkkuutta, mutta saanti ja F-luku jäivät alkuperäistä algoritmia huonommiksi.

Maui luo mahdollisten ehdokkaiden joukon kokonaisuudessaan jo eristysvaiheessa, joten luokitteluvaiheessa joukko voi ainoastaan pienentyä [Med09, s. 83]. Annotoinnin suurin mahdollinen saanti saavutetaan siis asettamalla luottamusarvon alaraja nollaan, jolloin algoritmi palauttaa lopullisena tuloksena kaikki eristysvaiheessa tuotetut ehdokkaat.

FDG:llä suurin mahdollinen saanti oli 46,3 %, ja Omorfin vastaava tulos oli 41,5 %. Ilman perusmuotoistusta korkein saanti jäi 28,8 prosenttiin. Koska FDG:tä käytettäessä ehdokkaiden joukosta saadaan kattavin, mahdollisuudet tulosten parantamiseen indeksointimenetelmän muita osia kuten luokittelijaa kehittämällä vaikuttavat hieman paremmilta kuin muita lemmatisointivaihtoehtoja käytettäessä.

Tuloksia tulkittaessa on syytä huomata, että Mauin opetusvaiheessa perusmuotoistus tehtiin FDG:llä myös silloin, kun annotointivaiheessa käytettiin Omorfia. Lisäksi Omorfia käytettäessä kukin annotoitava teksti perusmuotoistettiin kokonaisuudessaan esikäsittelynä ennen ehdokkaiden eristämistä, kun taas FDG:llä perusmuotois-

tus tehtiin n-grammi kerrallaan osana eristysvaihetta.

Ei ole täysin selvää, miten FDG:n käyttö opetusvaiheessa vaikuttaa Omorfilla saatuihin lopullisiin annotointituloksiin. Varsinaisten tulosten lisäksi myös korkein mahdollinen saanti jäi kuitenkin Omorfia käytettäessä FDG:tä pienemmäksi. Koska korkein saanti saavutetaan hyväksymällä luokitteluvaiheessa kaikki eristetyt ehdokkaat luottamusarvosta riippumatta, opetusdatan perusteella muodostetulla luokittelijalla ei ole tämän kannalta merkitystä.

Koska FDG suorittaa Omorfista poiketen morfologisen analyysin lisäksi lauseenjäsenanalyysin, on esitetty, että FDG saattaisi hyötyä tekstin analysoinnista koko tekstin tai lauseiden tasolla yksittäisten n-grammien tai sanojen sijaan [SSH11]. Tällöin n-grammien tasolla tehty lemmatisointi voi olla epäedullinen nimenomaan FDG:lle, joka tästä huolimatta vaikuttaa tuottaneen hieman Omorfia paremman tuloksen. Asian tarkempi analyysi jätetään kuitenkin jatkotutkimuksen aiheeksi.

6.1.2 Käsitehierarkian merkitys

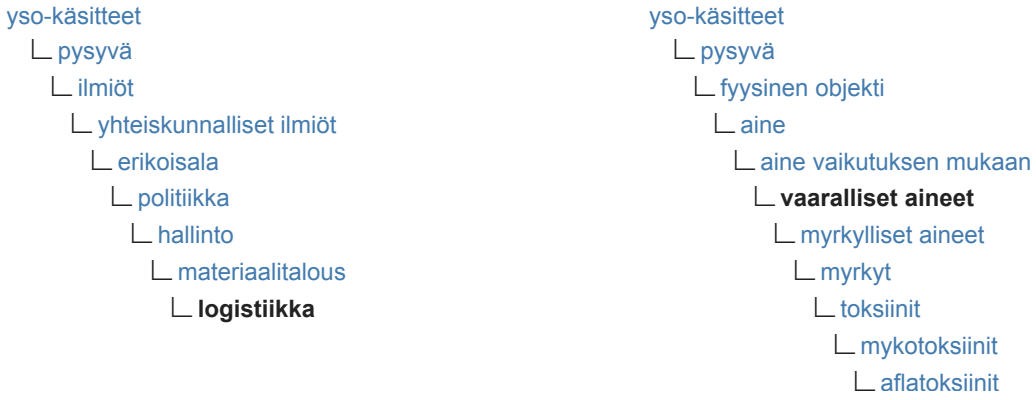
Omorfin yhteydessä käsitteen hierarkkisen aseman käytöllä ei ollut suurta vaikutusta tuloksiin. FDG:n tapauksessa hierarkiapiirteen käyttö muutti indeksoinnin tuloksia, mutta F-luku ei muuttunut mainittavasti. Käsitteen laskennalliseen spesifisyyteen perustuvasta piirteestä ei siten näyttäisi olevan hyötyä tapaustutkimuksen aineiston annotoinnissa Mauilla.

Mahdollisia selityksiä tulokselle on ainakin kaksi. Asiasanojen käsitteellinen rajaavuus ei välttämättä ole ollut keskeinen kriteeri alkuperäistä metatietoa laadittaessa. Toisaalta on myös mahdollista, että subjektiivinen rajaavuus on merkittävä tekijä asiasanojen valinnassa, mutta pelkkään taksonomiseen syvyyteen perustuva laskennallinen luku ei vastaa ihmisen tulkintaa käsitteen erityisyydestä.

Käsitteiden välinen painottamaton etäisyys taksonomiassa tai semanttisessa verkossa ei välttämättä ole paras mahdollinen tapa mitata niiden subjektiivista, käsitteellistä etäisyyttä [Res95, LBM03]. Sama voi koskea myös käsitteen yleisyyden tai erityisyyden laskentaa pelkästään taksonomisen hierarkian polkujen painottamattomia pituuksia laskemalla.

Lisäksi hierarkian kokonaissyvyys ontologian eri osissa vaihtelee. Esimerkiksi logistiikan käsitettä voi pitää merkitykseltään melko laaja-alaisena[22], mutta se on täs-

[22]Yleisen suomalaisen ontologian kuvaus logistiikan käsitteelle on kirjoitushetkellä ”materiaalitoimintojen kokonaisvaltainen ja keskitetty ohjaus”.



Kuva 3: Käsitteiden ”logistiikka” ja ”vaaralliset aineet” hierarkiat YSO:ssa

tä huolimatta sanastona käytetyssä yhdistelmäontologiassa taksonomisen haaransa lehtisolmu, ja sen laskennallinen erityisyys on siten suurin mahdollinen. Asiaa havainnollistaa kuva 3, joka esittää kahta metatiedoissa esiintynyttä käsitettä sekä pisimpiä polkuja niistä hierarkian juureen ja lehtiin. Kuvan hierarkiat ovat peräisin ONKI-verkkopalvelusta[23] [VTH09].

6.1.3 Vertailua aiempiin tuloksiin

Saavutettu tarkkuus vaikuttaa pintapuolisen tarkastelun valossa melko heikolta. Toisaalta myös aiemmissa tutkimuksissa raportoidut tulokset ovat vaihdelleet pääsääntöisesti noin 25 prosentista [Med09, s. 141–142] yli 50 prosenttiin [Med09, s. 139]. Automaattisen indeksoinnin tarkkuus ja saanti vaihtelevat luonnollisesti aineistosta toiseen, vaikka annotointiin käytettäisiin samaa algoritmia. Siten poikkeamat aiemmista tutkimustuloksista ovat odotettavia, eikä nyt saavutettu tarkkuus ole suoranaisesti poikkeuksellinen. Parhaisiin Mauilla saavutettuihin tuloksiin nähden normien asiasanoituksen tarkkuus jää joka tapauksessa selvästi huonommaksi.

Medelyan [Med09, s. 134–139] raportoi maatalousalaa, lääketiedettä ja hiukkasfysiikkaa käsittelevillä aineistoilla saaduiksi tarkkuuksiksi 32,9, 55,4 ja 38,4 prosenttia, kun luokittelijana käytettiin aggregoituja päätöspuita kaikilla Mauin tukemilla piirteillä. Kunkin alan teksteistä eristettiin 8–25 termiä asiakirjaa kohden. Sinkkilä ja kumppanit [SSH11] puolestaan eristivät kahden eri alan suomenkielisistä aineistoista neljä tai viisi asiasanaa tekstiä kohden ja raportoivat tarkkuuksiksi vastaavasti 45,4 ja 40,0 prosenttia. Tulosten perusteella Maui vaikuttaa heidän mukaansa toimivan sekä erikielisillä että eri aihealueita käsittelevillä aineistoilla, kunhan tekstin

[23] http://onki.fi/

ja sanaston perusmuotoistamiseen on käytettävissä tasokas kielikohtainen lemmatisointialgoritmi. Eri aineistoilla saavutetut tulokset ovat kuitenkin tästä huolimatta myös vaihdelleet huomattavasti, ja Medelyan [Med09, s. 141–142] raportoi espanjankielisillä maatalousalan dokumenteilla saaduksi tarkkuudeksi vain 24,7 %.

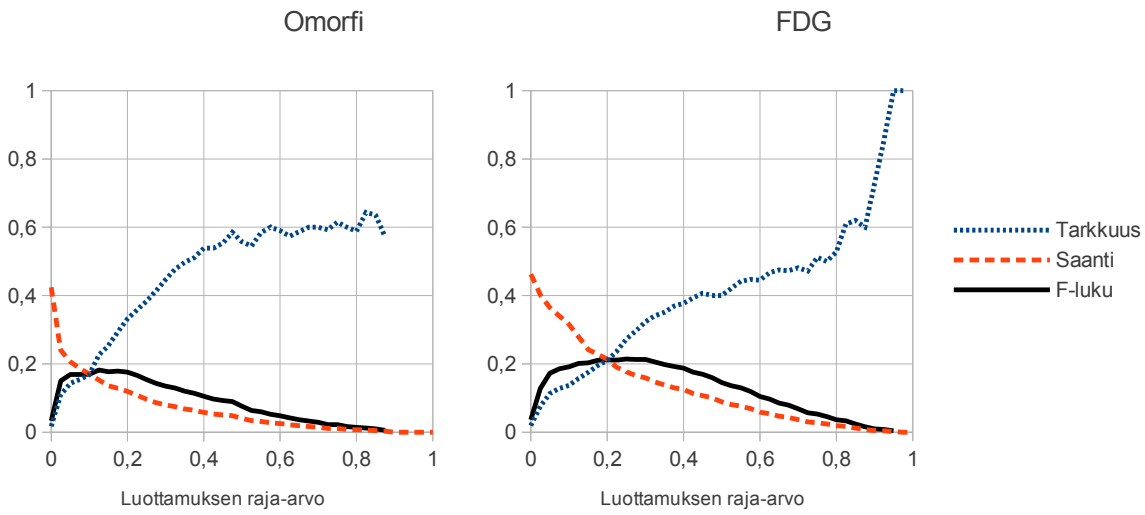
Toisaalta myös tiedonhallinnan ammattilaiset indeksoivat usein varsin epäyhdenmukaisesti [Rol81, Saa02]. Esimerkiksi Sinkkilän ja kumppaneiden [SSH11] tapauksessa aineiston indeksoineiden asiantuntijoiden kahdenväliset konsistenssit vaihtelivat 27,4 ja 36,6 prosentin välillä. Yksittäisen annotoijan ehdottamat asiasanat eivät siis ole kaikenkattavia, ja voidaan perustellusti olettaa myös monien alkuperäisistä poikkeavien termien olevan sisältöä kuvaavia. Tätä tukevat osaltaan myös subjektiiviset evaluointitulokset, joiden mukaan kelvollisten asiasanojen osuus on suurempi kuin alkuperäisiin metatietoihin vertailemalla laskettu tarkkuus.

Huomionarvoista normien annotoinnin tuloksissa on kuitenkin poikkeavan huonoksi jäänyt saanti. Esimerkiksi Sinkkilä ja kumppanit [SSH11] asiasanoittivat Mauilla suomenkielisiä sosiaalialan aineistoja ja raportoivat saantituloksiksi lemmatisointimenetelmästä riippuen 30–40 %. Myös Medelyan [Med09, s. 141–142] raportoi vastaavantasoisia tuloksia, joista heikoimmaksi jää espanjankielisen maatalousalan aineiston indeksoinnissa saavutettu 25,7 prosentin saanti.

Heikkoa saantia selittävät mahdollisesti ainakin aineiston ominaispiirteet. Sadasta satunnaisesti valitusta alkuperäisestä asiasanasta vain 49 esiintyi suoraan itse tekstissä jossakin taivutusmuodossa. Koska Maui perustuu termien eristämiseen tekstistä, tämä määrittää saavutettavissa olevalle saantiarvolle ylärajan, kun mittausmenetelmänä käytetään vain suoraa vertailua alkuperäiseen metatietoon.

Esimerkiksi Turneyn [Tur00] käyttämissä testiaineistoissa noin 75 % alkuperäisistä avainsanoista esiintyy suoraan itse tekstissä. Hulth [Hul03] mainitsee vastaavaksi osuudeksi 80 %. Kummassakin tapauksessa on kyse kontrolloidun sanaston termien sijaan vapaamuotoisista avainsanoista. Sanastoa hyödyntävien aineistojen osalta Medelyanin [Med09, s. 82] raportoimat saannin maksimiarvot vaihtelevat 45 ja 84 prosentin välillä riippuen korpuksesta, sanojen typistämiseen käytetystä menetelmästä ja ehdokkaiden eristämisen strategiasta.

Tarkkuuden ja saannin kehitystä luottamusarvon alarajaa muutettaessa havainnollistaa kuva 4, josta myös nähdään selvästi raja-arvon määrittämisen olevan kompromissi saannin ja tarkkuuden välillä. Kuvan kokeessa käytettiin Mauin alkuperäistä algoritmia ja parametreja.



Kuva 4: Tarkkuus ja saanti luottamuksen alarajan funktiona eri perusmuotoistusalgoritmeilla

Sinkkilä ja kumppanit [SSH11] eivät havainneet indeksointikokeiden tuloksissa merkittävää eroa FDG:n ja Omorfin välillä. Tutkimuksen alkuperäistä koeaineistoa ei kuitenkaan ole saatavilla vertailtavaksi, joten aineistojen eroavuuksia ja erilaisten tulosten mahdollisia syitä ei tässä yhteydessä analysoida.

Aiemmista Mauin suorituskyvyn evaluoinneista [Med09, SSH11] poiketen lopullisten ehdokkaiden joukko rajattiin tässä tapauksessa lukumäärän sijaan luottamusarvon perusteella. Automaattisesti eristettyjen termien keskimääräinen lukumäärä vaihteli tällöin perusmuotoistusmenetelmästä riippuen alle kahdesta noin neljään asiasanaan asiakirjaa kohden. Tulokset eivät siis ole suoraan vertailukelpoisia aiempien tutkimusten aineistojen kanssa.

6.2 Subjektiiviset arviot

Aihealueen asiantuntijoiden suorittamissa sokkokokeissa vastaajista ensimmäinen arvioi Mauin eristämistä asiasanoista 85 % vähintään melko hyviksi. Alkuperäisille ihmisen määrittämille annotaatioille luku oli 96 %. Toisen vastaajan tulos oli 61 % sekä Mauille että alkuperäisille termeille. Eri vastausvaihtoehtojen suhteelliset osuudet on esitetty taulukossa 2. Arviot tehneisiin henkilöihin viitataan myöhemmin tässä luvussa tunnisteilla A ja B.

Henkilön A näkemysten perusteella alkuperäisen ja automaattisesti tuotetun metatiedon laadussa näyttäisi siis olevan melko selvä ero. Alkuperäisistä annotaatioista

		Arvioija A	Arvioija B	Keskiarvo
Maui	Täysin eri mieltä (1)	0,01	0,17	0,09
	Jokseenkin eri mieltä (2)	0,14	0,23	0,18
	Jokseenkin samaa mieltä (3)	0,46	0,20	0,33
	Täysin samaa mieltä (4)	0,38	0,41	0,39
	$n = 71$	1,00	1,00	1,00
Ihminen	Täysin eri mieltä (1)	0,01	0,15	0,08
	Jokseenkin eri mieltä (2)	0,03	0,24	0,13
	Jokseenkin samaa mieltä (3)	0,48	0,16	0,32
	Täysin samaa mieltä (4)	0,49	0,45	0,47
	$n = 80$	1,00	1,00	1,00

Taulukko 2: Eri vastausvaihtoehtojen osuudet väittämään ”tämä termi on hyvä asiasana mainitulle normille”

lähes kaikki ovat A:n näkemyksen mukaan vähintään melko hyviä. Toisaalta joko täysin tai melko samanmielinen hän on myös valtaosasta Mauin tuottamia termejä. Henkilön B sen sijaan on esittänyt runsaammin myös negatiivisia arvioita, ja hänen mukaansa niin ihmisen kuin koneenkin määrittämistä asiasanoista vain noin kolme viidestä on vähintään melko hyviä. Parhaaseen luokkaan sijoittuu B:n mukaan 45 % alkuperäisistä ja 41 % Mauin tuottamista termeistä.

Kahden eri henkilön arviot poikkeavat huomattavasti toisistaan. Vastaavia tuloksia on saatu myös aiemmin. Esimerkiksi Barker ja Cornacchia [BC00] havaitsivat kahdentoista ihmisen arvioiden korreloivan keskenään vain keskinkertaisesti. Evaluoitaessa asiasanoituskokonaisuuksia yksittäisten termien sijaan yhdenmukaisuus on jo heikko. Barkerin ja Cornacchian mukaan koehenkilöille ei annettu tarkkoja ohjeita arvioinnin näkökulman suhteen, joten erot voivat selittyä esimerkiksi erilaisilla käsityksillä metatiedon tarkoituksesta. Myös Aronson ja kumppanit [AMG+04] toteavat arvioiden vaihtelevan.

Koska arviointiin osallistuneita henkilöitä ei ole useampia, koeasetelma kärsii tältä osin vastaavan tapaisesta ongelmasta kuin tarkkuuden ja saannin mittaus toteutettuna vain yhteen kertaan annotoituun aineistoon vertaamalla: yksinkertaisen indeksoinnin tapaan myös yksittäisen arvioijan vastaukset kuvaavat vain yhtä subjektiivista näkemystä relevanteista annotaatioista [Med09, s. 24]. Useampien arvioijien puuttuminen rajoittaa kaiken kaikkiaan tulosten analysointia paitsi asiasanoituksen laadun myös arvioijien keskinäisen yhdenmukaisuudenkin suhteen, eikä käytettä-

vissä olevan datan perusteella siten ole mielekästä esittää kummastakaan vahvoja päätelmiä. Edellä mainittuihin asiasanojen laatua koskeviin tuloksiin onkin syytä suhtautua automatisoitua arviointia täydentävänä ja jatkotutkimusta suuntaavana.

Tulosten tulkinnassa on lisäksi huomioitava, että koeasetelmassa käytetyillä parametreilla Maui tuotti normia kohden keskimäärin jonkin verran alkuperäisiä annotaatioita vähemmän asiasanoja. Jos luottamuksen alarajaa olisi laskettu niin, että lukumäärät olisivat vastanneet toisiaan, Mauin eristämien asiasanojen keskimääräinen laatu olisi ollut mahdollisesti heikompi.

6.3 Luottamusarvon yhteys laatuun

Luokittelijan ehdokkaille laskema luottamusarvo vastaa mallin perusteella laskettua todennäköisyyttä, jolla ehdokas on annotoitavan objektin kuvailuun soveltuva asiasana. Esimerkiksi puoliautomaattisen annotointijärjestelmän käyttöliittymässä asiasanaehdotukset on siten luontevaa esittää luottamusarvon mukaisessa järjestyksessä. Myös täysin automaattisessa asiasanoituksessa lopullisten termien joukko voidaan rajata joko kiinteään lukumäärään korkeimman luottamusarvon saaneita ehdokkaita tai vaadittavan luottamusarvon saavuttaviin ehdokkaisiin. Tällöin on olennaista, ovatko korkean luottamusarvon saavat ehdokkaat myös subjektiivisesti arvioiden hyviä.

Luottamusarvon ja subjektiivisen laadun suhde on esitetty taulukossa 3. Testissä käytetyillä parametreilla Mauin tuottamien asiasanojen luottamusarvojen mediaani oli noin 0,272. Mediaanin alapuolelle jääneistä 35 termistä 18 oli arvion B mukaan hyviä tai melko hyviä, kun taas mediaanin yläpuolelle sijoittuneista termeistä positiivisen arvion sai 25 termiä. Arvion A osalta hyvien tai melko hyvien termien osuus oli lähes identtinen mediaanin molemmin puolin, eikä luottamusarvon yhteyttä laatuun siten ole kaksijakoisen tarkastelun kautta havaittavissa. Myös henkilön A vastauksista kuitenkin ilmenee, että positiivisten arvioiden joukossa parhaiden osuus oli kasvanut ja melko hyvien pienentynyt. Luottamusarvon korrelaatiokerroin arvion A kanssa oli 0,33 ja arvion B kanssa 0,32.

Tulos antaa viitteitä siitä, että ainakin luottamusarvoltaan mediaanin yläpuolelle sijoittuvat asiasanat ovat pääsääntöisesti hyviä. Heikkolaatuiset ehdokkaat on siis mahdollista rajata pois luottamusarvon perusteella. Aiemmin mainituista koeasetelman rajoitteista, vastausdatan vähäisestä määrästä ja arvioijien erimielisyydestä johtuen myös tähän tulokseen on kuitenkin suhtauduttava varauksella. Lisäksi

		TEM	JEM	JSM	TSM	n
Arvioija A	luottamusarvo $< 0{,}272$	0,03	0,11	0,69	0,17	35
	luottamusarvo $> 0{,}272$	-	0,17	0,23	0,60	35
Arvioija B	luottamusarvo $< 0{,}272$	0,26	0,23	0,23	0,29	35
	luottamusarvo $> 0{,}272$	0,06	0,23	0,17	0,54	35

Taulukko 3: Subjektiiviset laatutasot suhteessa Mauin luottamusarvoon

tulokset ovat aineistokohtaisia, koska arviot riippuvat olennaisesti arvioijien näkökulmasta ja heidän asiasanojen sovelluskäyttöä koskevista oletuksistaan. Tuloksen vahvistamiseksi ja tarkentamiseksi tarvitaankin jatkotutkimusta.

7 Jatkotutkimus

Tässä luvussa pohditaan edellä esiteltyjen tulosten valossa mahdollisia jatkotutkimuksen kohteita sekä annotointimenetelmien suorituskyvyn kehittämiseksi että arviointimenetelmien parantamiseksi.

7.1 Indeksointialgoritmien vertailu ja arviointimenetelmät

Arviointikäytännöt avaintermien automaattisen eristyksen tutkimuksessa ovat jossain määrin hajanaisia. Useita kirjallisuudessa esiteltyjä annotointimenetelmiä on evaluoitu vain suppealla joukolla eri aineistoja, ja useiden annotointialgoritmien järjestelmälliset vertailut yhdenmukaisessa koeasetelmassa ovat olleet melko harvinaisia. Vertailukelpoisten tulosten saamiseksi olisi aiheellista selvittää eri algoritmien suorituskykyä usealla yhteisellä aineistolla. Järjestelmällistä vertailua voi tosin käytännössä rajoittaa myös algoritmien toteutusten saatavuus.

Kokeiden toistettavuus esimerkiksi algoritmin parametrien hienosäätämiseksi puoltaa automatisoitavien arviointimenetelmien suosimista suoran ihmistyön sijaan. Automaattinen evaluointi edellyttää kuitenkin laadukasta käsin annotoitua aineistoa, johon automaattisesti eristettyä metatietoa voidaan verrata. Toisaalta eri algoritmien järjestelmällisen vertailun kannalta myös subjektiiviset, useita tunnettuja algoritmeja yhtenäisessä koeasetelmassa vertailevat sokkokokeet voisivat olla aiheellisia.

Jos käytettävissä ei ole moninkertaisesti annotoitua koeaineistoa, indeksointitermien

semanttiseen etäisyyteen perustuvaa epätäsmällistä vertailua voidaan mahdollisesti käyttää indeksoinnin monikäsitteisyyden huomioimiseksi arvioinnissa. Asiasanojen semanttisen etäisyyden suhdetta indeksoinnin subjektiiviseen laatuun olisi kuitenkin syytä tutkia enemmän. Aiemmin esitetyt menetelmät semanttisen etäisyyden huomioimiseksi asiasanoituksen arvioinnissa ovat myös melko suppeita ja saattavat perustua esimerkiksi ainoastaan semanttisen verkon painottamattomien kaarien laskentaan.

Tilastollisten menetelmien käyttö yhdistettynä sanaston rakenteeseen on vaikuttanut lupaavalta tavalta mitata käsitteiden subjektiivista yhteenkuuluvuutta muissa yhteyksissä [Res95, LBM03, RM09]. Vastaavia menetelmiä voitaisiin mahdollisesti hyödyntää myös indeksointimetatiedon arvioinnissa, jos käytettävissä on tarkoitukseen soveltuva tesaurus tai ontologia.

7.2 Indeksointimenetelmien jatkokehitys

Tässä tutkielmassa esitettyjen kokeellisten tulosten valossa käsitteiden sijainnista ontologian käsitehierarkiassa ei voitu havaita olevan sellaisenaan hyötyä tekstistä eristettyjen termiehdokkaiden valikoinnissa. Hierarkkisen aseman määrittelyyn voitaisiin kuitenkin mahdollisesti käyttää myös muunlaisia semanttisen etäisyyden mittareita. Edistyneemmistä tavoista laskea käsitteiden välisiä semanttisia etäisyyksiä voisi olla hyötyä, jos asiasanojen käsitteellinen rajaavuus on metatiedon laadun kannalta merkityksellistä, mutta jos pelkkä semanttisen verkon kaarien laskenta ei riitä käsitteellisten etäisyyksien tai rajaavuuden mallintamiseen.

Toinen mahdollinen tutkimuskohde olisi selvittää, voidaanko ontologian rakennetta hyödyntää termiehdokkaiden luokittelussa jollain muulla tapaa. Jos rajattua aihealuetta käsittelevän aineiston kuvailuun käytetään esimerkiksi laajan yleisontologian termejä, asiakirjoja kuvaavat asiasanat saattavat esimerkiksi keskittyä muita useammin joihinkin tiettyihin käsitehierarkian haaroihin. Mahdollisia tapoja hyödyntää semanttisen verkon rakennetta annotointimenetelmien kehittämiseksi voitaisiin selvittää analysoimalla yhteyksiä verkon rakenteen ja olemassa olevan metatiedon välillä.

Sekä arviointimenetelmien että algoritmin jatkokehityksen osalta evaluointi on syytä tehdä usealla eri aineistolla laajemman kokonaiskuvan saamiseksi ja kokeiden toistettavuuden vuoksi.

8 Yhteenveto

Tässä tutkielmassa esitettiin katsaus tekstiaineistoa luonnollisen kielen termein kuvailevan indeksointimetatiedon automaattiseen tuottamiseen tekstistä eristämällä sekä erilaisiin mittareihin ja menetelmiin, joilla indeksointialgoritmien suorituskykyä ja metatiedon laatua on arvioitu. Lisäksi esiteltiin tapaustutkimus, jossa kokoelma asiakirjoja indeksoitiin Maui-algoritmia käyttäen ja selvitettiin näin tuotetun metatiedon laatua. Lopuksi etsittiin mahdollisia keinoja parantaa menetelmän suorituskykyä.

Esimerkkitapauksena esiteltiin puolustusvoimien normeja käsittelevien asiakirjojen automaattista annotointia Maui-indeksointialgoritmilla, joka perustuu keskeisten termien eristämiseen suoraan tekstistä. Metatiedon laatua selvitettiin vertaamalla algoritmin tuottamia asiasanoja alkuperäiseen asiakirjojen laatijoiden määrittämään metatietoon tarkkuus- ja saantimittauksin sekä yksittäisten termien asianmukaisuutta arvioinein sokkokokein. Lisäksi selvitettiin mahdollisuuksia parantaa annotoinnin tuloksia edistyneemmällä tekstin esikäsittelyllä ja avaintermien valikointiin käytettävään luokittelustrategiaan tehdyillä muunnoksilla.

Maui saavuttaa normien automaattisessa annotoinnissa aiemmin raportoitujen tulosten valossa kohtuullisen tarkkuuden. Tämä tukee aiempaa tutkimusta, jonka mukaan yleiskäyttöinen Maui soveltuu myös suomenkielisen tekstin asiasanoitukseen. Tarkkuus- ja saaantikokeita täydentävät aihealueen asiantuntijoiden sokkokokeissa esittämät arviot antavat viitteitä siitä, että sopivilla parametreilla Maui pystyy tuottamaan melko tasokkaita asiasanaehdotuksia. Sokkokokeiden osalta datan vähäinen määrä kuitenkin estää täsmällisten päätelmien tekemisen.

Annotointimenetelmän muunnelmia testattiin vertaamalla niillä saavutettua tarkkuutta ja saantia alkuperäisen koeasetelman tuloksiin. Esikäsittelyn osalta vertailtiin tekstin esikäsittelyssä kahta erilaista perusmuotoistusmenetelmää, joista kaupallinen FDG-jäsentäjä tuotti saannin ja siten automaattisesti eristetyn asiasanoituksen kattavuuden kannalta hieman Omorfia lupaavammat tulokset. Aiemmissa tutkimuksissa menetelmien välillä ei ole havaittu merkittävää eroa. Nyt havaittu ero saattaa selittyä koeasetelman ohella esimerkiksi edellisistä poikkeavan aineiston ominaispiirteillä, joskaan asiaa ei pystytty aiempien tutkimusaineistojen puutteen vuoksi selvittämään.

Asiasanaehdokkaiden luokittelun tarkkuuden parantamiseksi kokeiltiin rakenteellisen sanaston semantiikan hyödyntämistä sekä luokitteluun käytettyjen päätöspui-

den karsintaa. Käsitteiden hierarkkiseen asemaan perustuvasta luokittelupiirteestä ei havaittu olevan indeksoinnin tarkkuus- ja saantitulosten kannalta hyötyä. Sen sijaan päätöspuiden karsinta paransi indeksoinnin tarkkuutta ja F-lukua hieman, kun perusmuotoistukseen käytettiin FDG:tä.

Muunnelmien tuomista mahdollisista parannuksista huolimatta automaattisen indeksoinnin kattavuus jäi saantitulosten perusteella verrattain heikoksi. Aineiston alkuperäisistä asiasanoista vain noin puolet esiintyi suoraan tekstissä jossakin taivutusmuodossa, ja siten lähes puolet alkuperäisistä annotaatioista oli termejä tekstistä eristävän Mauin tavoittamattomissa. Sovellustason ratkaisuna kysymykseen tulisikin ensisijaisesti puoliautomaattinen indeksointijärjestelmä.

Kiitokset

Tutkielma on tehty Aalto-yliopiston perustieteiden korkeakoulun ja Helsingin yliopiston yhteisessä Semanttisen laskennan tutkimusryhmässä (SeCo) osana Suomalaiset semanttisen webin ontologiat (FinnONTO, 2003-2012) -projektia. Hanketta rahoittivat Teknologian ja innovaatioiden kehittämiskeskus Tekes sekä kymmenistä julkisen alan organisaatioista ja yrityksistä koostuva konsortio.

Haluan kiittää professori Eero Hyvöstä työn ohjaamisesta ja arvokkaista kommenteista sekä Matias Frosterusta yhteistyöstä tutkimuksessa, johon tutkielma perustuu. Lisäksi kiitän arviointiin osallistuneita puolustusvoimien työntekijöitä heidän työpanoksestaan ja koko Semanttisen laskennan tutkimusryhmää mukavasta työympäristöstä.

Tutkimuksessa käytetyn ARPA-rajapinnan ovat kehittäneet pääosin Joonas Laitio, Eetu Mäkelä ja Osma Suominen.

Lähteet

ABC+00 Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J., Rindflesch, T. C. ja Wilbur, W. J., The NLM indexing initiative. *Proceedings of the AMIA 2000 Annual Symposium.* American Medical Informatics Association, 2000, sivut 17–21.

AKM+03 Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H. ja

Shadbolt, N. R., Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18,1(2003), sivut 14–21.

AMG+04 Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M. ja Rogers, W. J., The NLM indexing initiative's medical text indexer. *Proceedings of the 11th World Congress on Medical Informatics (MEDINFO 2004)*, Fieschi, M., Coiera, E. ja Li, Y.-C., toimittajat, 2004, sivut 268–272.

BC00 Barker, K. ja Cornacchia, N., Using noun phrase heads to extract document keyphrases. Teoksessa *Advances in Artificial Intelligence*, Hamilton, H., toimittaja, osa 1822 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2000, sivut 40–52.

BCF02 Berrios, D. C., Cucina, R. J. ja Fagan, L. M., Methods for semi-automated indexing for high precision information retrieval. *Journal of the American Medical Informatics Association*, 9,6(2002), sivut 637–652.

BHB09 Bizer, C., Heath, T. ja Berners-Lee, T., Linked data – the story so far. *International Journal on Semantic Web and Information Systems*, 5,3(2009), sivut 1–22.

BHL01 Berners-Lee, T., Hendler, J. ja Lassila, O., The semantic web. *Scientific American*, 284,5(2001), sivut 34–43.

Bre96a Breiman, L., Bagging predictors. *Machine Learning*, 24,2(1996), sivut 123–140.

Bre96b Breiman, L., Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24,6(1996), sivut 2350–2383.

BRK05 Bracewell, D. B., Ren, F. ja Kuriowa, S., Multilingual single document keyword extraction for information retrieval. *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, loka–marraskuu 2005, sivut 517–522.

Cox80 Cox, E. P., The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17,4(1980), sivut 407–422.

DM05 D'Avanzo, E. ja Magnini, B., A keyphrase-based approach to summarization: the LAKE system at DUC-2005. *Proceedings of the DUC workshop*, 2005.

EMSS00 Erdmann, M., Maedche, A., Schnurr, H. ja Staab, S., From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. *Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*, Buitelaar, P. ja Hasida, K., toimittajat, 2000.

Fel98 Fellbaum, C., *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998.

FHW11 Frosterus, M., Hyvönen, E. ja Wahlroos, M., Extending ontologies with free keywords in a collaborative annotation environment. *Proceedings of the ISWC 2011 Workshop Ontologies Come of Age in the Semantic Web (OCAS)*. CEUR Workshop Proceedings, Vol 809, ISSN 1613-0073, lokakuu 2011.

FPW+99 Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C. ja Nevill-Manning, C. G., Domain-specific keyphrase extraction. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, San Francisco, CA, USA, 1999, Morgan Kaufmann Publishers Inc., sivut 668–673.

Gar91 Garland, R., The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, 2,2(1991), sivut 3–6.

Gar04 Garshol, L. M., Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. *Journal of Information Science*, 30,4(2004), sivut 378–391.

Gru93 Gruber, T. R., A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5,2(1993), sivut 199–220.

GW02 Guarino, N. ja Welty, C., Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45,2(2002), sivut 61–65.

Han96 Hand, D. J., Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159,3(1996), sivut 445–492.

HFH+09 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. ja Witten, I. H., The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11,1(2009), sivut 10–18.

HGM05 HaCohen-Kerner, Y., Gross, Z. ja Masa, A., Automatic extraction and learning of keyphrases from scientific articles. Teoksessa *Computational Linguistics and Intelligent Text Processing*, Gelbukh, A., toimittaja, osa 3406 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2005, sivut 657–669.

HKJ+01 Hulth, A., Karlgren, J., Jonsson, A., Boström, H. ja Asker, L., Automatic keyword extraction using domain knowledge. Teoksessa *Computational Linguistics and Intelligent Text Processing*, Gelbukh, A., toimittaja, osa 2004 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2001, sivut 472–482.

HN10 Hasan, K. S. ja Ng, V., Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, Stroudsburg, PA, USA, 2010, Association for Computational Linguistics, sivut 365–373.

HSC02 Handschuh, S., Staab, S. ja Ciravegna, F., S-CREAM – semi-automatic creation of metadata. Teoksessa *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, Gómez-Pérez, A. ja Benjamins, V., toimittajat, osa 2473 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2002, sivut 165–184.

HSV04 Hyvönen, E., Saarela, S. ja Viljanen, K., Application of ontology techniques to view-based semantic search and browsing. *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, 2004.

Hul03 Hulth, A., Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, Stroudsburg, PA, USA, 2003, Association for Computational Linguistics, sivut 216–223.

HVTS08 Hyvönen, E., Viljanen, K., Tuominen, J. ja Seppälä, K., Building a national semantic web ontology and ontology service infrastructure – the FinnONTO approach. *Proceedings of the European Semantic Web Conference ESWC 2008*. Springer, kesäkuu 2008.

HZ07 Hawking, D. ja Zobel, J., Does topic metadata help with web search? *Journal of the American Society for Information Science and Technology*, 58,5(2007), sivut 613–628.

IV98 Ide, N. ja Véronis, J., Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24,1(1998), sivut 2–40.

JP02 Jones, S. ja Paynter, G. W., Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. *Journal of the American Society for Information Science and Technology*, 53,8(2002), sivut 653–677.

KBK10 Kim, S. N., Baldwin, T. ja Kan, M.-Y., Evaluating N-gram based evaluation metrics for automatic keyphrase extraction. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10. Association for Computational Linguistics, 2010, sivut 572–580.

KMKB10 Kim, S. N., Medelyan, O., Kan, M.-Y. ja Baldwin, T., SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10. Association for Computational Linguistics, 2010, sivut 21–26.

LBM03 Li, Y., Bandar, Z. A. ja McLean, D., An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15,4(2003), sivut 871–882.

LC96 Larkey, L. S. ja Croft, W. B., Combining classifiers in text categorization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, New York, NY, USA, 1996, ACM, sivut 289–297.

LG75 Lissitz, R. W. ja Green, S. B., Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, 60,1(1975), sivut 10–13.

Lik32 Likert, R., A technique for the measurement of attitudes. *Archives of Psychology*, 140,140(1932), sivut 1–55.

LLZS09 Liu, Z., Li, P., Zheng, Y. ja Sun, M., Clustering to find exemplar terms for keyphrase extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, Stroudsburg, PA, USA, 2009, Association for Computational Linguistics, sivut 257–266.

Lov68 Lovins, J. B., Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11,1–2(1968), sivut 22–31.

LSP09 Lindén, K., Silfverberg, M. ja Pirinen, T., HFST tools for morphology – an efficient open-source package for construction of morphological analyzers. Teoksessa *State of the Art in Computational Morphology*, Mahlow, C. ja Piotrowski, M., toimittajat, osa 41 sarjasta *Communications in Computer and Information Science*, Springer Berlin Heidelberg, 2009, sivut 28–47.

LYRL04 Lewis, D. D., Yang, Y., Rose, T. G. ja Li, F., RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5,Apr(2004), sivut 361–397.

Med09 Medelyan, O., *Human-competitive automatic topic indexing*. Väitöskirja, University of Waikato, heinäkuu 2009.

MFW09 Medelyan, O., Frank, E. ja Witten, I. H., Human-competitive tagging using automatic keyphrase extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, Stroudsburg, PA, USA, 2009, Association for Computational Linguistics, sivut 1318–1327.

MMWB05 Miles, A., Matthews, B., Wilson, M. ja Brickley, D., SKOS Core: Simple knowledge organisation for the web. *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, 2005.

MT04 Mihalcea, R. ja Tarau, P., TextRank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, heinäkuu 2004.

MW06a Medelyan, O. ja Witten, I. H., Measuring inter-indexer consistency using a thesaurus. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, New York, NY, USA, 2006, ACM, sivut 274–275.

MW06b Medelyan, O. ja Witten, I. H., Thesaurus based automatic keyphrase indexing. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, New York, NY, USA, 2006, ACM, sivut 296–297.

MW08 Medelyan, O. ja Witten, I. H., Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology*, 59,7(2008), sivut 1026–1040.

MWM08 Medelyan, O., Witten, I. H. ja Milne, D., Topic indexing with Wikipedia. *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*. AAAI Press, 2008, sivut 19–24.

NK07 Nguyen, T. D. ja Kan, M.-Y., Keyphrase extraction in scientific publications. *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, ICADL'07, Berlin, Heidelberg, 2007, Springer-Verlag, sivut 317–326.

Par09 Park, J.-R., Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47,3–4(2009), sivut 213–228.

PDB+10 Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F. ja Tasso, C., Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems*, 25,12(2010), sivut 1158–1186.

Pol98 Pollitt, A. S., The key role of classification and indexing in view-based searching. Tekninen raportti, University of Huddersfield, 1998.

Por80 Porter, M. F., An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, 14,3(1980), sivut 130–137.

PPM04 Pedersen, T., Patwardhan, S. ja Michelizzi, J., WordNet::Similarity: measuring the relatedness of concepts. *Demonstration Papers at HLT-*

NAACL 2004, HLT-NAACL–Demonstrations '04, Stroudsburg, PA, USA, 2004, Association for Computational Linguistics, sivut 38–41.

PSI03 Pouliquen, B., Steinberger, R. ja Ignat, C., Automatic annotation of multilingual text collections with a conceptual thesaurus. *Proceedings of the Workshop Ontologies and Information Extraction at (EUROLAN'2003*, 2003, sivut 9–28.

Qui87 Quinlan, J. R., Simplifying decision trees. *International Journal of Man-Machine Studies*, 27,3(1987), sivut 221–234.

Qui93 Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Qui96 Quinlan, J. R., Bagging, boosting, and C4.5. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. AAAI Press, 1996, sivut 725–730.

Res95 Resnik, P., Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 1*, IJCAI'95, San Francisco, CA, USA, 1995, Morgan Kaufmann Publishers Inc., sivut 448–453.

RM09 Ruotsalo, T. ja Mäkelä, E., A comparison of corpus-based and structural methods on approximation of semantic relatedness in ontologies. *International Journal On Semantic Web and Information Systems*, 5,4(2009), sivut 39–56.

Rol81 Rolling, L. N., Indexing consistency, quality and efficiency. *Information Processing & Management*, 17,2(1981), sivut 69–76.

Saa02 Saarti, J., Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation*, 58,1(2002), sivut 49–65.

Seb02 Sebastiani, F., Machine learning in automated text categorization. *ACM Computing Surveys*, 34,1(2002), sivut 1–47.

SH12 Suominen, O. ja Hyvönen, E., Improving the quality of SKOS vocabularies with Skosify. Teoksessa *Knowledge Engineering and Knowledge Management*, Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H.,

d'Acquin, M., Nikolov, A., Aussenac-Gilles, N. ja Hernandez, N., toimittajat, osa 7603 sarjasta *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2012, sivut 383–397.

SHB06 Shadbolt, N., Hall, W. ja Berners-Lee, T., The semantic web revisited. *Intelligent Systems, IEEE*, 21,3(2006), sivut 96–101.

SSH11 Sinkkilä, R., Suominen, O. ja Hyvönen, E., Automatic semantic subject indexing of web documents in highly inflected languages. *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, kesäkuu 2011.

Ste46 Stevens, S. S., On the theory of scales of measurement. *Science*, 103,2684(1946), sivut 677–680.

TJ97 Tapanainen, P. ja Järvinen, T., A non-projective dependency parser. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.

TSV99 Tang, R., Shaw, W. M. ja Vevea, J. L., Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science*, 50,3(1999), sivut 254–264.

Tur99 Turney, P. D., Learning to extract keyphrases from text. Tekninen raportti, National Research Council Canada, Institute for Information Technology, 1999.

Tur00 Turney, P. D., Learning algorithms for keyphrase extraction. *Information Retrieval*, 2,4(2000), sivut 303–336.

Voo02 Voorhees, E., The philosophy of information retrieval evaluation. Teoksessa *Evaluation of Cross-Language Information Retrieval Systems*, Peters, C., Braschler, M., Gonzalo, J. ja Kluck, M., toimittajat, osa 2406 sarjasta *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2002, sivut 143–170.

VTH09 Viljanen, K., Tuominen, J. ja Hyvönen, E., Ontology libraries for production use: The Finnish ontology library service ONKI. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, toukokesäkuu 2009. Springer-Verlag.

VW93 Velleman, P. F. ja Wilkinson, L., Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47,1(1993), sivut 65–72.

ZD69 Zunde, P. ja Dexter, M. E., Indexing consistency and quality. *American Documentation*, 20,3(1969), sivut 259–267.

ZG09 Zesch, T. ja Gurevych, I., Approximate matching for evaluating keyphrase extraction. *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing (electronic proceedings)*, Borovets, Bulgaria, syyskuu 2009, sivut 484–489.