ELORE (ISSN 1456-3010), vol. 16 – 2/2009.
Julkaisija: Suomen Kansantietouden Tutkijain Seura ry.
[http://www.elore.fi/arkisto/2_09/katsart_palonen_et_al_2_09.pdf]

KATSAUSARTIKKELI

SEMANTTINEN KALEVALA – KULTTUURISAMMON TAONTAA

Tuomas Palonen, Jouni Hyvönen, Joeli Takala ja Eero Hyvönen

KALEVALAN ENSIMMÄINEN ”KÄÄNNÖS” TIETOKONEKIELELLE

Kalevala on ollut suomalaisen kulttuurin, taiteen ja kansallisen identiteetin keskiössä 1800-luvulta lähtien. Tällä Elias Lönnrotin koostamalla ja toimittamalla kansalliseepoksella (1) on ollut merkittävä rooli suomalaisuuden rakentamisessa ja yhä edelleen sen innoittamana luodaan taidetta ja kulttuuria.

Kalevalan merkityksellisyyttä on haluttu hyödyntää ja ilmentää *Semanttisessa Kalevalassa*, jota tässä artikkelissa esitellään. *Semanttinen Kalevala* on kansalliseepoksemme verkkoversio ja osa kansallista Kulttuurisampo-portaalia (Hyvönen 2008a; Hyvönen ym. 2009). *Semanttinen Kalevala* on toteutettu *Uuden Kalevalan* (1849), suomalaisille kaikkein tutuimman ja kansalliseepokseksi kanonisoituneen version, pohjalta. Hankkeessa *Kalevala* on siirretty digitaaliseen lukuympäristöön, jossa sekä uudet lukijakunnat että toisenlaiset luennat tuottavat uudenlaisia variaatioita kansalliseepoksen reseptioon samalla vahvistaen *Kalevalan* merkitystä suomalaisuuden keskeissymbolina. Voidaan sanoa, että kansalliseepoksestamme on *Vanhan Kalevalan* 175-vuotisjuhlan kunniaksi valmistunut ensimmäistä kertaa käännös tietokoneiden ymmärtämälle formaalille kielelle.

Semanttinen merkitsee *Semanttisen Kalevalan* yhteydessä sitä, että tietokone kykenee alkeellisessa mielessä ”ymmärtämään” *Kalevalan* tarinaa ja näin ollen liittämään automaattisesti esimerkiksi eepoksen eri kohdat niihin liittyviin selityksiin, toisiin *Kalevalan* kohtiin tai muihin kansallisiin kulttuurisisältöihin. Esimerkiksi Kullervoa käsittelevät runot voidaan yhdistää Akseli Gallen-Kallelan maalaamaan teokseen Ateneumin taidemuseossa ja Kansallisbiografiassa oleviin Lönnrotin ja Gallen-Kallelan elämänkertoihin.

Semanttinen Kalevala -hankkeen tavoitteena on ollut johdattaa nykylukijaa ymmärtämään eepoksen lähderunojen ja itse teoksen kertomusmaailmaa, joka saattaa jäädä nykylukijalle hämäräksi arkaaisen runokielen vuoksi. *Semanttinen Kalevala* toimii lukuoppaana myös Lönnrotin työskentelyyn: siihen, millä tavalla hän teoksen kokoajana sommitteli ja yhdisti kansanrunolähteiden narratiiviset rakenne-elementit yhtenäiseksi juonelliseksi kokonaisuudeksi. *Semanttinen Kalevala* johdattaa nykylukijat uuden tietotekniikan avulla *Kalevalan* juonellisten rakenneratkaisujen, runojen kielen ja kulttuuristen merkitysten äärelle. Tässä katsausartikkelissa esittelemme *Semanttisen Kalevalan* ja *Kulttuurisammon* toimintaperiaatteita sekä niiden toteuttamiseen käytettyjä teknisiä puitteita. Aluksi valotamme vielä semanttisen webin taustaa sekä avaamme hypertekstin ja narratiivin käsitteitä semanttisen webin kontekstissa. Lopuksi esittelemme muita vastaavia projekteja ja tarkastelemme *Semanttiseen Kalevalaan* liittyviä haasteita ja mahdollisuuksia.

Semanttinen web – merkitysten verkko

Perinteisessä mielessä semantiikka on kielitieteen haara, joka tutkii sanojen ja lauseiden merkityksiä. Kielen puhujalla on oltava semanttista tietoa sanoista, jotta hän voi käyttää niitä ymmärrettävällä tavalla. (Saeed 1997, 3.) Tekoälyssä, tietojenkäsittely- ja informaatiotieteessä termi ”semanttinen” viittaa merkkien ja viestien formaaliin, täsmällisesti määriteltyyn merkityssisältöön ja sisältöä koskevaan tietoon, jota voidaan tulkita automaattisesti (Russell & Norvig 2005). Lähtökohtaisesti tietokone ei osaa tulkita web-dokumenttien sisältöä. Semanttisen webin tarkoituksena onkin saattaa verkkoaineistot tietokoneelle ymmärrettävälle kielelle, jolloin tietokone voi käsitellä annettuja tietosisältöjä. Näin voidaan rakentaa tehokkaampia ja tarkempia selailu- ja hakujärjestelmiä, jotka auttavat käyttäjää tiedon etsimisessä, seulomisessa ja yhdistelyssä. (Antoniou & van Harmelen 2008, 1–3.) Semanttisen webin keskeisin tavoite onkin tehdä verkossa olevasta datasta ”semanttisesti rikastettua”, jolloin hakukoneet ”osaavat” seuloa laajoista korpuksista pelkästään loppukäyttäjää kiinnostavia tietosisältöjä.

Semanttisen webin kehittämiseen kohdistetaan maailmanlaajuisesti tällä hetkellä runsaasti tutkimuksellisia voimavaroja. Suomessa alan laajin hanke on kansallinen *Suomalaiset semanttisen webin ontologiat* -projekti *FinnONTO* (2003–2007 ja 2008–2010), jossa on valmistunut prototyyppi kotimaisiin sanastoihin ja käsitteistöihin eli ontologioihin perustuvasta infrastruktuurista (Hyvönen 2008b). Esittelemme seuraavaksi lyhyesti ontologia-käsitteen, jolla on keskeinen merkitys semanttisen webin toiminnassa.

Ihmiskielen käsitteiden muuttaminen tietokoneelle luettavaan muotoon tapahtuu ontologioiden avulla. Filosofiassa ontologialla tarkoitetaan sen tutkimusta, mitä on olemassa, millaisia käsityksiä tuosta olevaisesta on, millaisista ominaisuuksista se rakentuu sekä miten se on suhteessa toisiin objekteihin. Tietojenkäsittelytieteessä ontologia-käsite mielletään tietokoneelle kirjoitetuksi konkreettiseksi kuvaukseksi maailmasta. Ontologialla on tietojenkäsittelytieteessä lähinnä deskriptiivistä merkitystä, kun pyritään kuvaamaan mahdollisimman tarkkaan olioiden, ilmiöiden, esineiden tai luontokappaleiden olemusta ja näiden keskinäisiä suhteita. Ontologia on monipuoli-

nen, sähköinen versio perinteisestä sanastosta, tietynlainen käsitteiden verkko, jossa yksittäisten käsitteiden keskinäiset suhteet on määritelty. Ontologiat ovat semanttisen webin perusta. Niiden avulla tietokoneohjelmat pystyvät tehokkaammin paikantamaan hakijaa kiinnostavia kohteita. Ontologiat sisältävät siis fragmentaarisia käsitteellisiä kuvauksia maailmasta. Niiden avulla verkkoaineistot tulevat ”ymmärrettäväksi” tietokoneille ja tietojenkäsittelyohjelmille. Ontologiat rakennetaan usein olemassa olevien sanastojen pohjalta ja ne on yleensä rajattu koskemaan jotakin tiettyä aihepiiriä. Tietojenkäsittelytieteessä ontologian perustavoitteena on määrittää tietyn aihepiirin käsitteistöä esimerkiksi jonkin yksilön tai luokan mukaan sekä avata eri käsitteiden välisiä suhteita. Aineistojen kuvailua tietokoneelle nimitetään *annotoinniksi*. Annotointi on ontologian käsitteistön avulla tapahtuvaa sisällön kuvaamista, luokittelua ja jäsentelyä. (Antoniou & van Harmelen 2008, 8–12; Hyvönen 2005.)

Kulttuurisampo – suomalaisen kulttuuriperinnön portaali

Annotointi tarkoittaa käytännössä *Kalevalan* sisällön ja eepoksen kertomusmaailmaan liittyvän toiminnan kuvailua tietokoneelle käyttämällä semanttisen webin menetelmiä. *Kalevalan* annotointihanke onkin osa laajempaa semanttisen webin tekniikoiden tutkimukseen ja sovellusten kehittämiseen keskittynyttä *FinnONTO* -tutkimusprojektia. Se on osa *Kulttuurisampo*-portaalia, johon on koottu aineistoja kansallisesti keskeisten museoiden, arkistojen ja muiden organisaatioiden kokoelmista ja verkkopalveluista. Suurin osa *Kulttuurisammon* sisällön annotoinnista on toteutettu kokoelmien luetteloinnin yhteydessä. Kulttuurisammossa on maalauksia, veistoksia, etnografisia museoesineitä, tietoja historiallisista paikoista ja rakennuksista, digitoituja aineistoja kuten karttoja, valokuvia, äänitteitä sekä vanhojen kalevalamittaisten runojen tekstitiedostoja. Kulttuurisampo-portaalissa on tätä kirjoitettaessa yhteensä lähes 700 000 kokoelmakohdetta ja yli viisi miljoonaa muuta kohdetta, kuten viittauksia paikkoihin tai henkilöihin. Kulttuurisammon sisältö on annotoitu semanttisesti ontologioiden avulla. Tähän on käytetty *FinnONTO*-projektissa kehitettyä kansallista KOKO-ontologiaa (http://www.ysofi/onto/koko/). Sen perustana on *Yleinen suomalainen ontologia YSO*, joka pohjautuu Kansalliskirjaston kehittämään *Yleiseen suomalaiseen asiasanastoon* (YSA). (2)

Semanttisella Kalevalalla on Kulttuurisammossa kahtalainen rooli. Ensinnäkin se muodostaa itsessään oman kokoelmansa, joka sisältää *Uuden Kalevalan* kaikki 50 runoa nivoutuen portaalin muihin kokoelmiin. Toiseksi se toimii yhtenä Kulttuurisammon yhdeksästä temaattisesta tulokulmasta suomalaiseen kulttuuriperintöön. Ideana on, että *Semanttista Kalevalaa* lukemalla pääsee käsiksi portaalin muihin kokoelmiin ja kohteisiin (Hyvönen 2007). Kulttuurisampo suosittelee semanttisen samankaltaisuuden ja muiden sääntöjen perusteella kohteita, jotka liittyvät tavalla tai toisella siihen tekstikohtaan, jonka käyttäjä on valinnut luettavakseen. Kun käyttäjä selaa esimerkiksi tiettyä kalastusaiheista kohtaa *Kalevalasta*, Kulttuurisampo saattaa suositella

kalastukseen liittyviä museoesineitä tai maalauksia muista kokoelmista. Semanttiseen Kalevalaan on luotu erilliset paikka- ja toimijaontologiat, jotka sisältävät kuvaukset *Kalevalan* keskeisistä toimijoista ja paikoista. Kustakin näistä on kirjoitettu käyttäjää varten kuvailuteksti, joka esittelee kalevalaisen runoperinteen ja sen ajatusmaailman taustoja. Toimijoina esiintyvät ihmishahmojen, kuten Väinämöisen ja Louhen, lisäksi myös myyttiset toimijat, yliluonnolliset oliot sekä muita hahmoja, kuten myyttinen Hiiden hirvi. Tässä vaiheessa hanketta *Kalevalan* paikka- ja toimijaontologiat eivät kata vielä kaikkia eepoksessa esiintyviä paikkoja ja toimijoita.

Käytetty tekniikka ja työkalut

Kaikkien Kulttuurisammossa olevien hakukohteiden ”semanttinen rikastaminen” perustuu kohdeyksiköiden tietosisältöjen kuvailuun ja annotointiin. Kun Kulttuurisammossa tietoa etsivä tekee portaalissa hakuja, tiedon seulonta toteutuu sekä ”semanttisen rikastamisen” että tietokoneelle avauttujen tietosisältöjen välisten yhteyksien loogisten mallien avulla. Sisältökohteiden kuvaus rakentuu RDF-kuvauskieleen, joka käsittelee tietoa kolmikkoina (ks. esim. Antoniou & van Harmelen 2008, 61–84). Portaalin yhtenä keskeisenä haasteena on hyvin heterogeenisten sisältöjen kuvailu keskenään yhteismitallisella tavalla, joka mahdollistaa sisältöjen yhdistämisen. Valittu sisällönkuvaustapa on esittää kohteiden tiedot tapahtumien (engl. event) mukaan jäsennettynä (Junnila ym. 2008). Tämän mallin esikuvana ovat tekoälyn esittämisessä käytetyt menetelmät ja luonnollisen kielen semanttiseen kuvailuun tietokoneelle tarkoitettu, John Sowan kehittämä syväsijamalli (Sowa 2000), mutta *Semanttisessa Kalevalassa* tarkkuustaso on kyseistä mallia yksinkertaisempi. Sowan syväsijamallin mukaisesti ”semanttinen rikastaminen” perustuu *Kalevalan* sisältöjen toiminnallisuuden avaamiseen tietokoneelle ymmärrettävään muotoon ja itse toiminnallisuuden sisältämän logiikan tekemiseen tietoteknisesti jäsenneltäväksi ja käsiteltäväksi.

Semanttisen Kalevalan tapahtumarakenteiden ja sisällönkuvailu on toteutettu käsityönä käyttämällä Saha-annotointityökalua, joka on yhdistetty ONKI-ontologiapalveluun (Viljanen ym. 2009). *Kalevalan* annotaatio koostuu kuudesta eri osiosta:

1) varsinaisesta semanttisesta kerroksesta, jossa on kuvattu tarinan alkeistapahtumat,
2) narratiivisesta kerroksesta, joka kuvaa tarinan etenemistä suurempana kokonaisuutena,
3) käsitemalleista, jotka kuvaavat toimijoita, henkilöitä ja paikkoja tarkoilla selitysteksteillä ja joita käytetään asiasanoina muualla Kulttuurisammossa,
4) rivinumeroista, jotka kiinnittävät annotaatiomallin kuvailut annotoitavaan dokumenttiin,
5) ontologioista, joilla geneeriset asiasanoitukset tehdään ja
6) suosittelusäännöistä, joilla liitetään samankaltaisia kohteita muualta Kulttuurisammosta Kalevala-sivujen yhteyteen.

Annotoidut osiot yhdessä muodostavat semanttisen sisällönkuvailun osalta oleellisimman informaation sisältävän semanttisen tietosisältöjen tason. Niihin perustuvat myös toiminnallisuutta jäsentävät suosittelutoiminnot. Tietokone pystyy annotoinnin avulla seulomaan *Kalevalaa* ja muita Kulttuurisammon kohteita huomioiden loppukäyttäjän näkökulman hakien merkityksellisiä ja relevantteja yhteyksiä sekä *Semanttisen Kalevalan* eri annotointitasojen välillä että muiden Kulttuurisammon kohteiden joukosta. Kaikki tapahtumarakenteiden ja sisällönkuvailuun osallistuvat erilliset kerrokset ovat itsenäisiä osia *Kalevalaan* liittyvää tietosisältöjen korpusta. Erillisinä ja osittain itsenäisesti toimivina kokonaisuuksina ne voidaan tarvittaessa vaihtaa tai poistaa ilman, että kokonaisuus hajoaa. Näin ollen portaalissa voidaan helposti kehittää ja kokeilla erilaisia semanttisia sisällönkuvailu- ja suosittelumalleja.

Hypertekstit ja lineaarisuuden murtumia

Semanttista Kalevalaa voi luonnehtia kansalliseepoksen hypertekstikäännökseksi. George P. Landow:n mukaan hyperteksti ”rakentuu tekstilohkoista – – ja sähköisistä linkeistä, jotka liittävät näitä lohkoja yhteen” (Landow 2006, 3). Hypermedian Landow puolestaan määrittelee hypertekstin laajentumaksi, johon kuuluu ”visuaalista informaatiota, ääntä, animaatiota ja muuta dataa”. Hän käyttää tekstilohkoista nimitystä leksia (engl. lexia). Olennaista on, että hyperteksti ja -media eivät kumpikaan koske painettua tekstiä vaan liittyvät nimenomaan tietokoneiden mahdollistamaan tekstuaalisuuteen.

Hyperteksti liittyy laajempaan sähköisen kirjallisuuden käsitteeseen, jonka muita ilmentymiä ovat muun muassa interaktiivinen fiktio (eli tekstipelit), tietokoneen generoimat tekstit sekä tietokoneella luodut taideinstallaatiot, jotka sisältävät tekstiä (Hayles 2008, 1–42). Landow määrittää kaksi linkkejä sisältävän tekstin rakennetyyppiä, jotka ovat aksiaalinen rakenne ja verkkorakenne. Ensin mainittu sisältää runkotekstin, johon on ripustettu leksioita eli tekstejä. Tällaisia ovat esimerkiksi monet kanonisoitujen kirjallisten tekstien sähköiset tai painetut kriittiset editiot. Verkkorakenne taas muodostaa edellistä monisyisemmän leksioiden verkon, joka tuottaa verkosta itsestään sen yksittäistä tekstinkappaletta merkittävämmän kokonaisuuden. Myös näiden välimuotoja esiintyy. (Landow 2006, 69–76, 99–109.) Myös Kulttuurisampo voidaan käsittää tällaiseksi niin kutsutuksi hypermediaaliseksi verkoksi. Portaalissa on kuitenkin systeemin sisäinen, ontologioista ja metatiedoista muodostuva, tietokoneen tulkittavaksi tarkoitettu semanttinen käsiteverkko, joka automaattisesti luo ”hypertekstuaalisuutta” portaalin käyttäjille. Yksi *Semanttisen Kalevalan* tavoitteista onkin ollut laatia teoksesta narratiivinen representaatio, jossa tekstinsisäiset viittaukset ja juonelliset jatkumot visualisoidaan.

Hypertekstin rakenne vaikuttaa myös lukutapaan. Yleisesti tekstiä voidaan lukea monella eri tavalla, esimerkiksi lineaarisesti, visuaalisesti tai hypähdellen. Lineaarinen lukutapa liittyy yleensä painetun tekstin lukemiseen edeten sanasta sanaan ja sivulta toiselle. Visuaalisessa lukemisessa keskitytään sisällön ohella myös tekstin ulkoasuun. Hypähtelevästä lukutavasta on aiheellista puhua erityisesti hypertekstien yhteydessä, sillä

tällöin lukeminen ei etene lineaarisesti vaan lukuprosessissa tapahtuu siirtymiä, hyppyjä ja takautumia tekstin sisällä tai sen ulkopuolelle. On selvää, että hypähtelevää lukutapaa voi käyttää myös painetun tekstin yhteydessä, mutta hypertekstin rakenne korostaa tätä lukemisen muotoa. (Liu & Smith 2008; Tanselle 2008; Mchoul & Roe 1996.)

Narratiivi on käsite, josta on käyty runsaasti keskustelua hypertekstien tutkimuksen yhteydessä. Monet narratologit määrittävät narratiivin ihmisajattelun peruspilariksi, jonka avulla välitetään yhteistä todellisuutta koskevia yleismaailmallisia viestejä (esim. White 1980, 1–2). Postmodernin ajattelun innoittama hypertekstuaalisuuden tutkimus on kyseenalaistanut tämän näkemyksen: hyperteksti ei vastaa aristoteelista käsitystä tarinasta, jolla on alku, keskikohta ja loppu (Landow 2006, 218–226; Aarseth 1997, 41–51). Hypertekstien tutkimus on korostanut lineaarisuuden rinnalla narratiivien epälineaarisuutta. (3) *Semanttinen Kalevala* sijoittuu haastavasti tämän keskustelun ytimeen yhtäältä vanhana, auktorisoituna teoksena, toisaalta hypertekstikontekstissa julkaistavana narratiivina. Epälineaarisuuden käsitteessä on kuitenkin nähty myös ongelmia: lukuprosessi on joka tapauksessa lineaarinen, vaikka teksti tarjoaisikin lukijalle useita vaihtoehtoisia luentoja. Niinpä epälineaarisuuden rinnalla käytetään myös termejä multilineaarinen ja interaktiivinen. Multilineaarinen viittaa hypertekstin tarjoamiin vaihtoehtoisiin tai vaihtuviin luentoihin. Interaktiivinen painottaa lukijan aktiivista roolia: lukija vastaa itse osaltaan siitä, minkälainen lopullinen luettu teksti on. Hyperteksti ei poista inhimillistä taipumusta tekstin lineaariseen prosessointiin ja siihen tyydytykseen, jonka kerronnallisesti koherentin tekstin lukeminen tuottaa. (Aarseth 1997, 2–3, 41–47; Landow 2006, 41–43, 221–229; Douglas 2000, 122.)

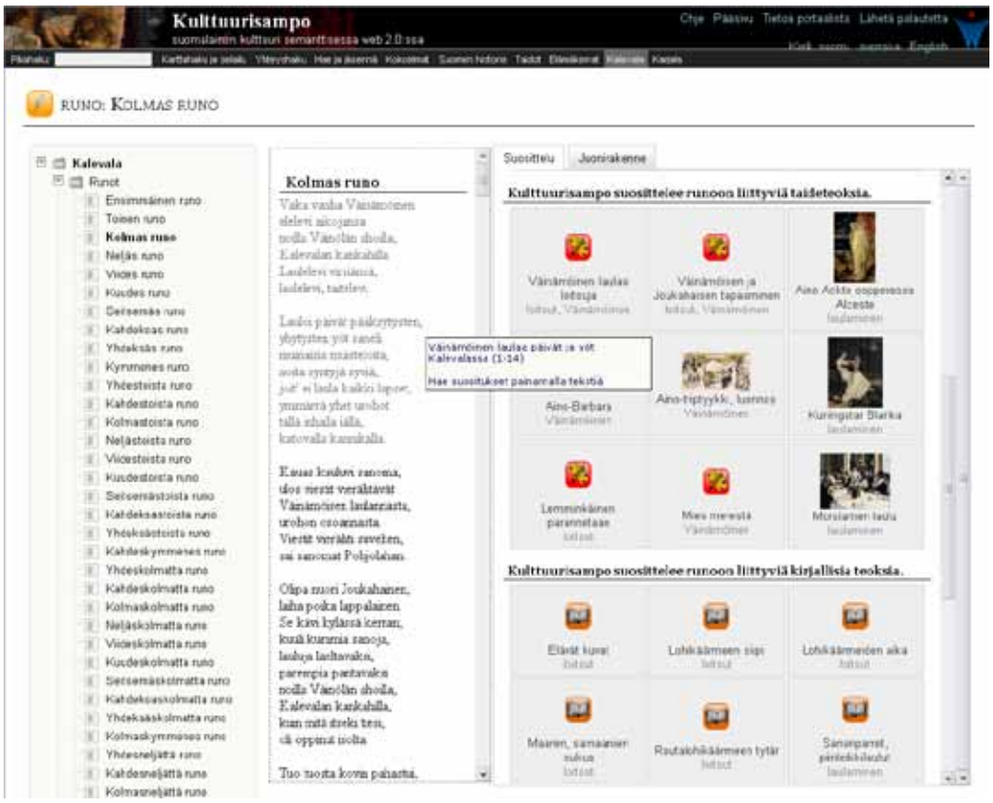
Juuri tämän epälineaarisuuden osalta hypertekstin ja suullisen perinteen välillä voidaan nähdä kiintoisaa samankaltaisuutta. John Miles Foleyn (2009) mukaan molemmissa on kyse yksiköiden linkittämisestä toisiinsa, ei staattisesta tekstistä. Lisäksi molemmat tuottavat yksittäisiä, ainutlaatuisia kokemuksia. Walter J. Ong on toisaalta jo 1980-luvulla, ennen internetin yleistymistä, puhunut tietokoneen ja muun modernin teknologian aikaansaamasta sekundaarisesta suullisuudesta (Ong 1982, 11), jonka piiriin myös hypertekstit voidaan luontevasti hahmottaa.

Kun muistamme, että *Kalevalan* pohjana on vanha suullinen perinne, *Semanttinen Kalevala* voidaan edellä esitetyn näkemyksen pohjalta nähdä kalevalaisen kansanrunon paluuna staattisesta eepoksesta epälineaariseen ja verkostomaiseen luentaan. Hypertekstinä sen voi ajatella kumoavan lineaarisuuden ja epälineaarisuuden vastakkainasettelun: molemmat selitys- ja lukutavat sopivat *Semanttiseen Kalevalaan* yhtä hyvin. Samaan tapaan se peilautuu myös Landowin dikotomiaan tekstin rakenteesta: *Semanttinen Kalevala* voidaan nähdä sekä aksiaalisena että verkkorakenteisena tekstinä tai ehkä pikemminkin näiden kahden rakenteen hybridinä. Se on osa verkkoa, mutta samalla myös akseli, jonka ympärille verkko kutoutuu.

Kalevalan annotointi voidaan jakaa semanttiseen ja rakenteelliseen annotointiin. Semanttinen annotointi pureutuu teoksen kertomusmaailman ja runojen kulttuurisen taustan sisältöihin ja merkityksiin. Rakenteellisessa annotoinnissa kuvaillaan teoksen juonellista rakennetta. Siinä läpäistään narratiivin semanttinen pintataso keskittymällä rakenteeseen sekä sen sisäisiin hienojakoisempiin sidoksiin.

SEMANTTINEN ANNOTOINTI

Kalevalan annotointiskeema on tapahtumapohjainen. Tapahtuman annotointi koostuu tapahtuman otsikoinnista, sen paikallistamisesta tiettyyn tekstikorpuksen kohtaan (tiettyihin säkeisiin) sekä asiasanoittamisesta KOKO-ontologian käsitteiden avulla. Asiasanakäsitteiden kautta voidaan kertoa tilanteen luonne, siihen osallistuvat tahot, paikka, aika ja niin edelleen. Jokainen annotoitu tapahtuma on ankkuroitu tiettyihin *Kalevalan* säkeisiin jakamalla eepoksen 50 runoa pienempiin osiin. Tässä tapahtumajaossa on pyritty noudattamaan Lönnrotin tekemää kappalejakoa. Oletuksena on ollut, että Lönnrot itse on käyttänyt jakoa ilmaisemaan tiettyjen narratiivisten kokonaisuuksien rajoja. Lopputuloksena teksti on jaettu pääsääntöisesti yhden tai kahden kappaleen pituisiin tapahtumiin, poikkeuksena loitsut ja pitkät monologit. Annotoinnissa on toteutettu kahta toisiaan täydentävää päämäärää. Ensiksi *Kalevala* on haluttu kuvata niin tarkasti ja lähdeuskollisesti kuin se ontologian avulla on mahdollista ja sovelluksen kannalta tarpeen. Toiseksi on pyritty siihen, että tietokone voi automaattisesti tuottaa kiinnostavia ja mielekkäitä suosituslinkityksiä Kulttuurisammossa (ks. kuva 1).



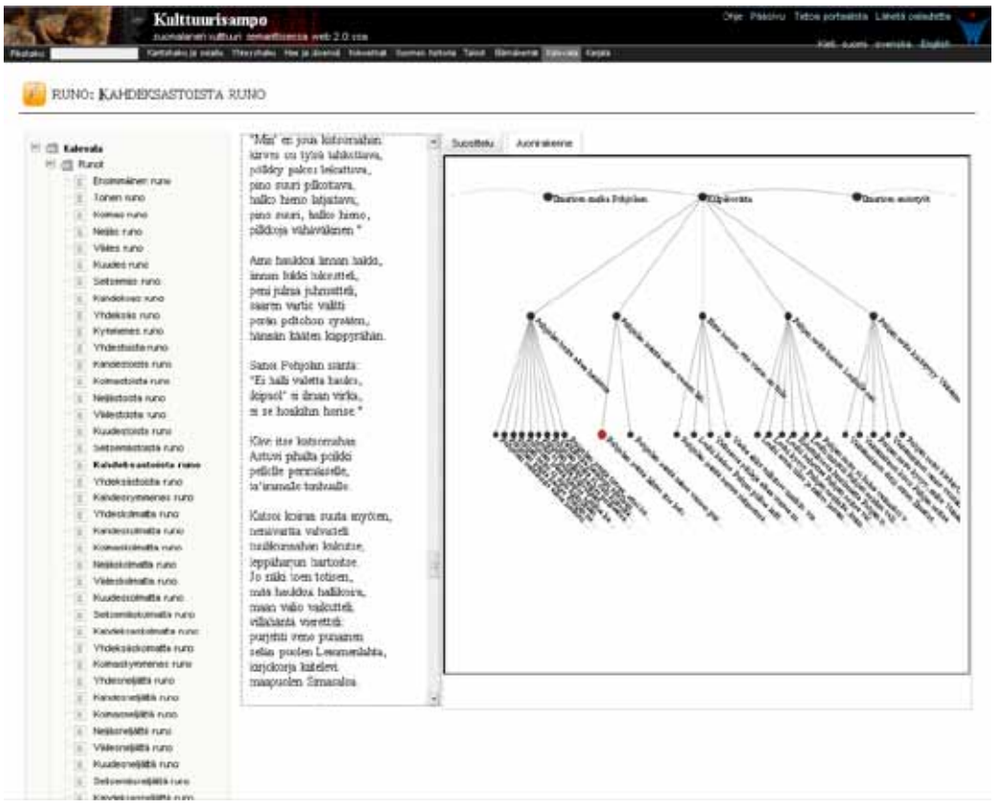
Kuva1: Semanttisen Kalevalan suosituksia. Oikealla runoluettelo, keskellä runoteksti ja vasemmalla Kulttuurisammon suosittelemia kohteita kuvakkeineen.

Annotoinnissa käytetyt asiasanakäsitteet voidaan jakaa karkeasti neljään kategoriaan: toimijat, paikat, toiminnot ja muut asiasanat. *Kalevalan* toimijoita ja paikkoja varten luotiin omat ontologiat, jotka eivät sisälly KOKO-ontologiaan. KOKO ja etenkin sen perustana oleva *Yleinen suomalainen asiasanasto YSA* on alun perin tarkoitettu kirjasto- ja arkistokäyttöön, joissa kuvailun tarkkuustaso on suurelta osin abstrakti ja geneerinen. *Semanttisessa Kalevalassa* samaa käsitteistöä käytetään kuvaamaan kertomuksen sisäistä toimintaa, joka tapahtuu konkreettisemmalla, arkisemmalla tasolla. KOKO-ontologiaa on tästä syystä täydennetty toiminnan osalta. Toiminnallisen kuvailun ulkopuolella ovat vielä muut asiasanat, joihin kuuluvat esimerkiksi esineet, mentaaliset ilmiöt sekä luontoon ja kulttuuriin liittyvät käsitteet. Tyypillisesti juuri tämä konkretian, materian ja arjen taso yhdistää *Kalevalan* luontevimmin Kulttuurisammon muuhun sisältöön.

Seuraavassa tarkastelemme esimerkkinä annotoinnista ja sen haasteista 40. runon säkeitä 159–172. Annotoinnissa tapahtuman otsikoksi valitaan lyhyt tiivistelmä sen sisällöstä, tässä ”Väinämöinen pistää haukea miekalla ja saa sen katkeamaan kahtia”. Tapahtumaa kuvaamaan valitaan joukko asiasanakäsitteitä, jotka esiintyvät säkeiden tekstissä tai kuvaavat niiden tapahtumaa, tässä esimerkiksi: ”Väinämöinen”, ”tappo”, ”hauki”, ”vesi”, ”poisto” ja ”pää”. Jotta kone voisi liittää nämä käsitteet ja tekstin laajempaan kontekstiin, niiden pitää kuulua järjestelmän perustana olevaan ontologiaverkkoon. Yhtenä haasteena annotoinnissa on, mitä tehdä käsitteille, joita ontologiasta ei löydy. Esimerkiksi Väinämöinen-käsitettä ei löytynyt alkuperäisestä KOKO-ontologiasta. Ratkaisuna on laajentaa ontologiaa tarpeen mukaan ja tarjota uudet käsitteet eri annotoijille yhteiseen käyttöön, jotta samaa käsitettä ei määritellä useaan otteeseen ja kenties monella eri tavalla. Lievempi muoto käsitteen puuttumisen ongelmasta on tilanne, jossa ontologiassa oleva käsite ei vastaa täysin haluttua merkitystä. Esimerkiksi käsite ”tappo” löytyy KOKO:sta, mutta sen määrittely periytyy *Yleisestä suomalaisesta asiasanastosta*, jossa termin merkityksenä on henkirikostyyppi, ei hengenriisto eläimeltä, mihin sitä tässä yhteydessä on käytetty. *Kalevalan* kaltaisen tekstin sisällönkuvailu on haasteellista, kun välineenä on yleisemmän tason käsitteistö, joka sisältää eri alojen terminologiaa ja jota ei ole alun perin kehitetty narratiivin kuvailuun. (4) Käsitteistö on monesti sekä puutteellista että epätäsmällistä ja saattaa liittyä ontologiatasolla toiseen viitekehykseen, kuten edellä juridiikkaan. Ongelmia tulee varsinkin silloin, kun säkeiden merkitys on epäselvä, monimutkainen tai triviaali. On vaikea päättää, mitä kuvailla ja mitä jättää kuvailematta. Ongelma on tuttu muun muassa kirjastoissa tehtävässä julkaisujen asiasanoituksessa, vaikka siinä kuvailun tarkkuus ja luonne ovatkin erilaisia.

Rakenteellinen annotointi

Semanttisen Kalevalan kautta on mahdollista hahmottaa narratiiviyksiköiden välisiä suhteita, kohtausketjuja ja eepoksen juonirakenteen kokonaisuutta. Rakenteellinen annotointi voidaan jakaa hierarkkiseen ja intratekstuaaliseen annotointiin. Hierarkkinen annotointi viittaa siihen, että *Kalevala* on jaettu kolmeen tasoon, joita kutsutaan



Kuva 2. Kalevalan narratiivisen hierarkian visualisointi. Vasemmalla narratiivin kolmen tason yksiköt otsikoineen puumallina. Punainen väri osoittaa tarkasteltavan tapahtuman sijainnin hierarkiassa.

narratiiviyksiköiksi. Tasot ovat episodi, kohtaus ja tapahtuma. Episodien ja kohtausten annotointi sisältää tiedon siitä, mistä alemman tason yksiköistä kukin narratiiviyksikkö koostuu, mutta episodeja ja kohtauksia ei ole annotoitu semanttisesti. Narratiiviyksiköt ovat myös Kulttuurisammon kohteita. Hierarkiajaon pohjalta *Semanttiseen Kalevalaan* on toteutettu eepoksen tarinahierarkian visualisointi (ks. kuva 2).

Hierarkiajaon lähtökohtana on tapahtumataso. Tässä narratiiviyksikössä esiintyy jokin melko yksinkertainen toiminto. Kohtaus on sarja toisiinsa liittyviä tapahtumia ja episodi sarja toisiinsa liittyviä kohtauksia. Esimerkiksi *Kalevalan* 26. runoon on annotoitu episodi ”Lemminkäisen valmistautuminen Pohjolaan”, jonka yhdeksi kohtaukseksi on annotoitu ”Lemminkäinen varustautuu” ja edelleen tämän kohtauksen yhdeksi tapahtumaksi ”Lemminkäinen kehuu miekkaansa”.

Intratekstuaalinen annotointi auttaa eepoksen sisäisten kytkösten ja juonikomposition kokonaisrakenteen ymmärtämisessä. Siihen kuuluu kaksi osa-aluetta, joita nimitämme kohtausketjuiksi ja kohtausreferensseiksi. Kyseisten piirteiden visualisointi *Semanttisen Kalevalan* näyttökerrokseen on tällä hetkellä vielä toteuttamatta, mutta varsinainen annotointi on suoritettu.

Juonirakenteen skematisoinnin taustatyöhön on käytetty Vladimir Proppin, Claude Bremondin ja Roland Barthesin teorioita narratiivin rakenteesta (5). Teoriat eivät kuitenkaan ole skematisoinnin käytännön toteutuksen kanssa täysin yhteismitallisia.

Semanttisen Kalevalan juonirakenteen annotointi on toteutettu kohtaustasolla yhdistämällä kohtauksia toisiinsa. Tämä johtuu siitä, että kuvaustarkkuus on liian yksityiskohtainen tapahtumatasolla ja toisaalta liian harva episoditasolla. Annotointi yhdistää kohtauksia kahdella tapaa: kohtausketjuina sekä kohtausreferensseinä. Kohtausketju on kolmi- tai nelivaiheinen sarja kohtauksia. Kohtausketjulla osoitetaan pitkäkestoisia ja monivaiheisia juonellisia kehitelmiä. Kohtauksen vaiheet ovat 1) alku, 2) kehittely, 3) kohokohta ja 4) ratkaisu. (6) Kolmivaiheiseen ketjuun ei kuulu kehittelyvaihetta. Temaattisesti tärkeimmät vaiheet ovat kohokohta ja ratkaisu: kohtausketjujen seulomisen lähtökohtana on käytetty kerronnallisia huipentumia eli rakennettu sellaisia ketjuja, joiden kohokohdaksi tai ratkaisuksi muodostuu jokin kerronnan kannalta tärkeä kohtaus. Kohtausketjuja on annotoitu toimintaperusteisesti: niissä toteutuu kahden tai useamman toimijan välinen tai yhden toimijan suorittama toiminta. Kohtausketjut voivat myös olla limittäisiä siten, että ratkaisu on samalla uuden ketjun alku. Annotoinnissa on suosittu ketjuja, jotka eivät muodostu ainoastaan peräkkäisistä kohtauksista vaan ovat harvempia ja harppovat joidenkin triviaalien kohtausten ylitse. Hyviä esimerkkejä tästä ovat ketjut, jotka liittyvät useampaan kuin yhteen runoon. Esimerkiksi Lemminkäisen surmaan liittyvät tapahtumat saavat alkunsa runossa 12 ja päättyvät varsinaiseen surmaan runossa 14. Lisäksi ketjuannotoinnissa on suosittu kohtauksia, jotka vievät kertomusta eteenpäin.

Hyvä esimerkki kohtausketjusta on Kullervo-runostoon kuuluva ”Ilmarisen emännän petos”, jossa saa alkunsa dramaattisesti päättyvä Ilmarisen emännän ja Kullervon välinen konflikti. Ketjuun liittyy neljä kohtausta: ”Ilmarisen emäntä leipoo Kullervolle kivisen leivän”, ”Kullervo paimentaa”, ”Kullervo katkoo veitsensä kiviseen leipään” sekä ”Kullervo vannoo kostoa Ilmarisen emännälle”. Ketjun kohokohdaksi on valittu Kullervon veitsen katkeaminen, joka onkin kerronnallisesti merkittävä kohtaus ja yksi Kullervo-runoston avainkohdista.

Juonellisten kuvioiden lisäksi narratiivit sisältävät tyypillisesti myös sisäisiä referenssejä: tekstissä viitataan aiemmin tapahtuneeseen tai ennakoidaan tulevia. Kalevala sisältää tällaisia runsaasti ja nimitämme niitä tässä yhteydessä kohtausreferensseiksi. Lähinnä kyse on kohtauksista, jotka sisältävät vuorosanoja, mutta viittaus voi tapahtua myös epäsuorasti. Viittaus voi liittyä myös useampaan kuin yhteen kohtaukseen. Referenssejä on kahta tyyppiä, viittaukset menneisiin ja tuleviin kohtauksiin. Molemmista löytyy runsaasti esimerkkitapauksia. Esimerkiksi 12. runosta löytyy selkeä referenssi tulevaan: Lemminkäinen jättää äidilleen ja Kyllikille ennusmerkiksi suan. Jos suka alkaa vuotaa verta, Lemminkäinen on pulassa. Myöhemmin 15. runossa suka alkaa vuotaa verta merkiksi siitä, että Lemminkäinen on juuri kuollut Tuonelan joella. Annotoinnissa suan ennusmerkiksi asettaminen on linkitetty veren vuotamiseen.

Annotoinnissa on myös pyritty tekemään nykylukijoille näkyväksi sellaisia rakenteellisia piirteitä, jotka kertovat kerrontastrategisten käytäntöjen lisäksi Lönnrotin ideologisista tavoitteista. Lönnrot käytti kertomusmaailman toimijoita ja esitti esimerkiksi

Väinämöisen vuorosanoilla senaikaisen sivistyneistön arvoja. Väinämöisellä kansansa johtajana oli auktoriteetti ohjata, valistaa ja kieltää kansaa harhautumasta kunnialliselta ja oikealta polulta. Runoissa, joissa Lönnrot on halunnut esittää neuvoja ”yhteiselle kansalle”, kertoja päästetään ääneen vasta runojakson lopussa. (7) Eeposkirjailijana Lönnrotin keskeiseksi päämääräksi nousi teoksen kokonaisrakenteen hahmottaminen. Suullisen perinteen runoaiheiden hajanaisuuden ja variaation synnyttämien lukuisten toisintojen edessä Lönnrotin oli löydettävä jokin luokitteleva ja edustavuutta arvottava kriteeri kokonaisuuden jäsentelylle. Toimitusprosessin myötä Lönnrot loi itselleen kuvan perinteen taustalla olevasta maailmankuvasta. Tämä helpotti lähderunoston hajanaisuuden ja intertekstuaalisten yhteyksien jäsentämistä.

Lönnrotin työskentelymenetelmien ja ideologisten tavoitteiden todentaminen on mahdollista nostamalla esiin kattavasti sekä itse teoksessa että sen lähteissä olevaa intertekstuaalista verkostoa. Kohtaustason ja episoditason annotoinnissa pyrkimyksenä on ollut tehdä näkyväksi ne strategiat, joilla Lönnrot on koostanut juonellisen kokonaisuuden suhteellisen hajanaisista lähdeainesten kerronnallisista elementeistä. Juonirakenteen semanttisen annotoinnin myötä tulee esille Lönnrotin mieltymys symmetriaan ja toistoon. Myös kansanrunolähteet noudattavat kerronnallisessa parallelismissaan tiettyjä kertauskaavoja. Lönnrot on saanut mallin lähderunoista, mutta toteuttaessaan sitä staattisessa tekstiympäristössä ja myös laajempien kokonaisuuksien strukturoinnissa hän on turvautunut omiin mieltymyksiinsä palautuviin esteettisiin ratkaisuihin. Edellä kuvatut seikat, kuten tekstuurin tasolla ilmenevä symmetria tai eepostekstiin sisään kirjoitetut ideologiset ulottuvuudet, ovat haasteita *Kalevalan* semanttisessa rikastamisessa. *Kalevalan* suhde lähderunostoon on sangen mielenkiintoinen.

Semanttista Kalevalaa ollaan mahdollisesti tarkentamassa ja laajentamassa uusiin suuntiin. Yksi houkutteleva suunta on yhdistää Kalevala-korpus *Suomen Kansan Vanhoihin Runoihin* (SKVR), joka sisältää kansalta kerätyt runot alkuperäisasussaan. Kulttuurisampoon on liitetty osia SKVR:n runokokonaisuudesta. Näin *Kalevalan* säkeet voitaisiin johtaa alkulähteilleen. Työssä voitaisiin käyttää pohjana Väinö Kaukosen aiheeseen liittyvää tutkimustyötä. Samoin olisi mahdollista tutkia, miten samankaltaisia kohtia voitaisiin tunnistaa ja liittää toisiinsa automaattisesti eri korpusten välillä. Myös eeposkomposition kompleksisuuden näkyväksi tekeminen on merkittävä haaste *Semanttisen Kalevalan* tulevaisuudessa. Tässä kehitysvaiheessa *Semanttinen Kalevala* on vasta pintaraapaisu niistä merkitysverkostoista, jotka sekä teokseen itseensä että sen lähdepohjaiseen luentaan olisi mahdollista vielä liittää mukaan. (8)

Muita hypertekstiin tai semanttiseen webiin perustuvia hankkeita

Eeposten ja kirjallisten teosten hypertekstikäännöksiä on julkaistu paljon CD-ROM -formaatissa. Esimerkiksi *Raamatusta* on julkaistu useita tällaisia versioita (Landow 2006, 69–71). Myös *Kalevalasta* julkaistiin CD-ROM-versio vuonna 1996 nimellä *HyperKalevala*. Eepostekstin lisäksi *HyperKalevala* sisältää tarinaan ja runonlauluperinteeseen liittyviä kuvia, ääninäytteitä ja videoita, mutta se ei hyödynnä semanttisen webin teknologiaa, joka kehitettiin vasta myöhemmin. Nyttemmin useita eepoksia on julkaistu hypertekstimuodossa verkossa, ja niistä osa on vapaasti luettavissa. Esimerkiksi *Beowulf in hypertext* (http://www.humanities.mcmaster.ca/~beowulf/) ja *An eEdition of The Wedding of Mustajbey's Son Bećirbey* (http://www.oraltradition.org/zbm) tarjoavat alkutekstin rinnalle englanninkielisen käännöksen sekä eeposta taustoittavia artikkeleita ja tarkan kommentaarin. Näihinkään ei ole vielä sovellettu semanttista webiä, mikä mahdollistaisi automaattisen linkittämisen sisällönkuvailun perusteella. Semanttisen webin keinoin on sen sijaan toteutettu useita MuseoSuomen ja Kulttuurisammon tapaisia kulttuuriportaaleja, kuten *MultimediaN E-Culture Demonstrator* (http://e-culture.multimedian.nl/). Lisäksi tekoälyn ja semanttisen webin teknologioita on käytetty tekstien ja narratiivien kuvaamiseen. Tällaisesta tutkimuksesta osa on keskittynyt käytännön sovelluksiin, osa teoreettisiin ja teknisiin ulottuvuuksiin. Gian Piero Zarrin (2008) kehittämä Narrative Knowledge Representation Language -kieltä (NKRL) edustaa semanttisen tutkimuksen teoreettistä puolta. Lähempänä käytäntöä on *Historical Event Markup and Linking Project* (Helm), jossa historiallisia tapahtumia on järjestetty kronologisesti ja yhdistelty henkilöitä ja paikkoja tietokoneen ymmärtämään muotoon (Robertson 2009). Myös Text Encoding Initiative -skeemaa (TEI) on kokeiltu yhdistää semanttisen webin välineisiin. TEI on yleisesti käytetty annotointiskeema tekstin digitaaliseen säilytykseen ja representaatioon. Sitä on käytetty erityisesti kirjallisten teosten sähköisiin editioihin. (9)

Semantiikan ja tekstin yhteensovitus on edennyt tähän mennessä lähinnä teknisten ambitioiden ajamana eikä ohjaksia ole vielä annettu kulttuuriorganisaatioiden tai sisällön tuottajien käsiin. Tämä käänne edellyttäisi annotointivälineiden kehittymistä helppokäyttöisemmiksi ja kenties jonkinlaisen semanttisen annotoinnin standardin muodostumista. Tässä prosessissa *Semanttinen Kalevala* ja edellä mainitut muut hankkeet ovat askeleita oikeaan suuntaan.

Lopuksi

Semanttisen Kalevalan selailu verkossa herättää ajatuksia jatkokehittelymahdollisuuksista ja *Kulttuurisampo*-portaalin toiminnallisuuteen liittyvistä haasteista. Esimerkiksi tiettyihin *Kalevalan* kohtauksiin liittyvät maalaukset eivät välttämättä näy portaalin suosituksissa niiden kohtausten yhteydessä, joita maalaukset kuvaavat. *Semanttisen Kalevalan* hanke on osoittanut sen, että eepoksen tai kaunokirjallisen narratiivin asiasanoittamiseen, eri aineistojen asiasanoituksen yhdenmukaistamiseen, ontologian

johdonmukaiseen käyttöön ja sitä kautta koko semanttiseen webiin liittyy ongelmia ja haasteita. Yhden ontologian soveltaminen useaan eri aineistoon, vieläpä eri aineistotyyppeihin, johtaa aina kompromisseihin. Yhtenä ongelmana voidaan nähdä se, että annotoija tekee aina yksilöllisiä tulkintoja eli kaksi eri henkilöä tuskin kuvailisi samaa aineistoa identtisellä luennalla. Muodoltaan ja sisällöltään läheisiä kohteita on vaikea yhdistää, jos ne on asiasanoitettu eri käsitteillä. (10) Semanttisen webin sovellukset voidaan karkeasti hahmottaa kokonaisuuksiksi, jotka rakentuvat neljästä toisiinsa hierarkkisessa suhteessa olevasta osasta: 1) järjestelmästä, 2) ontologiasta, 3) aineistosta ja 4) annotoinnista. Sisällönkuvailun yhdisteltävyyden ongelmat voivat liittyä mihin tahansa hierarkiatasoista. Ontologia ja annotointi ovat tietyssä historiallisessa ja tieteellisessä kontekstissa rakennettuja kulttuurisia konstruktioita. Jotakin jää väistämättä semanttisen annotoinnin ulkopuolelle. On kuitenkin muistettava, että *Kalevala* on varsin poikkeuksellinen asiasanoituksen kohde ja lisäksi ensimmäinen narratiivinen kokonaisuus, johon semanttisen webin teknologiaa ylipäätään sovelletaan tässä mitassa.

Haasteista huolimatta semanttinen web tarjoaa uusia tapoja ja mahdollisuuksia *Kalevalan* lukemiseen, ymmärtämiseen ja hahmottamiseen laajemmassa kontekstissa. *Semanttinen Kalevala* tarjoaa nykylukijalle mahdollisuuden sukeltaa teoksen kertomusmaailmaan – sen kielelliseen ja runolliseen rikkauteen. Sen kautta syntyy myös ennennäkemättömiä ja yllättäviäkin siltoja suomalaiseen kulttuuriperintöön. Tapahtumajaon ja -otsikoinnin kautta vanha, monelle nykylukijalle kenties vaikeaselkoinen runomitta aukeaa tiiviiksi ja helppolukuiseksi proosaksi. *Kalevalan* paikka- ja toimijaontologia kuvailuineen perehdyttää lukijaa kalevalaiseen maailmaan. Juonirakenteiden visualisointi auttaa lukijaa hahmottamaan *Kalevalan* keskeisen juonen asettaen eeposkokonaisuuden samalla laajempaan perspektiiviin. Tällaisten ominaisuuksien vuoksi *Semanttista Kalevalaa* voi suositella esiteltäväksi ja käytettäväksi esimerkiksi kouluissa, kirjastoissa, erilaisissa nuorisotiloissa sekä muissa vastaavissa sivistykseen ja vapaa-ajan toimintaan liittyvissä laitoksissa. *Semanttinen Kalevala* viihdyttää, mutta myös sivistää.

Kalevala on toiminut suomalaisessa kulttuurissa yhtenä keskeisenä kansallisena identiteetin rakennuksen välineenä ja siihen on projisoitu symbolisesti lähes ikonisessa mielessä keskeisiä suomalaisuuden representaatioita ja kansallista peruskuvastoa. Vaikuttaessaan kulttuurissamme yleisellä tasolla ja yksilöiden mielissä se on tyypillinen kansallista identiteettiä rakentava, voimistava ja uusintava representaatio. *Semanttinen Kalevala* on osa tätä tulkintojen ja kansalliseepoksen luennan jatkumoa. Lauri Hongon esittämän näkemyksen mukaan Kalevala-prosessilla – teoksen tieteellisillä, populaareilla, taiteellisilla ja kansanomaisilla tulkinnoilla – ei ole näkyvissä päätepistettä (ks. Honko 1987, 126–127). *Kalevala* ei ole vain yksi tekstuaalinen artefakti, jolla olisi olemassa yksi kiinteä muoto, vaan teos, joka taipuu luennassa moneksi. Lönnrotin toteamus, että *Kalevalaa* olisi yhtä hyvin saanut toimitettua seitsemän toisistaan poikkeavaa versiota, on paljastava kansaneepoksen autenttisuuden, staattisuuden ja tulkintojen kannalta. Tämä lausunto kertoo omalla tavallaan Lönnrotin suhtautumisesta teoksen toimitustyön päättämättömyyteen. *Semanttinen Kalevala* on osoittanut vanhan eepoksen yhdistämisen nykyteknologiaan hedelmälliseksi raivaten tietä nykylukijoille eepoksen tulkintaan. Niin kauan kuin kansalliseeposta luetaan ja siihen viitataan, tämä tulkintojen jatkumo elää. *Semanttinen Kalevala* liittyy tähän Kalevala-prosessiin

edustaen yhtä uudenlaista luentaa. Semanttisen webin tekniikoiden mahdollistamassa toimintaympäristössä *Kalevalan* moniäänisyys tulee esille tuoreella tavalla.

On selvää, että Kulttuurisammossa nyt oleva semanttisesti annotoitu *Kalevalan* juonirakenne on vain pintaraapaisu sen osalta, miten monitasoisia ratkaisuja strategisesti ja toimitusteknisesti Lönnrot teki koostaessaan eeposta. Annotoinnin haasteena olikin pyrkimys tehdä näkyväksi Lönnrotin hyödyntämiä toimitusstrategioita ja eepoksen kollaasimaista juonellista rakennetta. Suhteessa lähteenä toimineisiin kansanrunoihin Kalevala on sangen monimutkainen tekstikokoonpano. Tämän johdosta tavoitteeksi ei edes asetettu sitä, että kaikki Lönnrotin hyödyntämät toimitukselliset ratkaisut olisi ollut mahdollista tehdä näkyväksi. Annotointiskeema kuitenkin mahdollistaa uusien semanttisten tulokulmien liittämisen myöhemmin osaksi *Semanttista Kalevalaa*, joten kaikkea ei ollut tarkoituksellista liittää mukaan hankkeen tässä kehitysvaiheessa.

Hankkeen eteneminen

Semanttisen Kalevalan ja Kulttuurisammon idean ja mallin kehitti vuonna 2003 Eero Hyvönen, joka on myös johtanut koko hanketta tämän jälkeen. Järjestelmän suunnitteluun, sisällöntuotantoon ja tekniseen toteuttamiseen on eri vaiheissa osallistunut suuri joukko Teknillisen korkeakoulun ja Helsingin yliopiston Semanttisen laskennan tutkimusryhmän (SeCo) tutkijoita. Ensi askeleet Kalevalan semanttisessa annotoinnissa otettiin vuosina 2004–2005, kun MuseoSuomi-järjestelmästä (http://www.museosuomi.fi/) kehitettiin ensimmäinen Kulttuurisammon versio (Hyvönen 2008a) kahden pro gradu -työn puitteissa (Junnila 2006; Salminen 2006). Mukana oli kaksi jaksoa *Kalevalasta*. Tämän jälkeen tapahtumien annotointiskeemaa yksinkertaistettiin, ja Kulttuurisammon seuraavaan versioon Jouni Hyvönen annotoi yksityiskohtaisesti neljä eepoksen 50 runosta ja laati *Kalevalan* toimija- ja paikkaontologioiden sisällöt. Tarkoitus oli testata sen hetkistä skeemaa ja välineistöä sekä narratiivin semanttista annotointia ylipäätään. *FinnONTO*-hankkeen työvälineet eivät kuitenkaan olleet vielä kovin kehittyneitä eikä käytettävissä ollut nykyistä kymmenistä tuhansista toisiinsa yhdistetyistä käsitteistä koostuvaa KOKO-ontologiaa.

Uusi vaihe annotoinnissa alkoi vuonna 2008, kun Kulttuurirahaston rahoituksella voitiin aloittaa koko *Kalevalan* semanttinen sisällönkuvailu. Suvantoaikana annotointieditori Saha oli saatu valmiiksi (Valkeapää ym. 2007) ja KOKO-ontologian ensimmäinen versio kehitettyä. Aiempia syväsijamalleja hyödyntävää annotointiskeemaa pelkistettiin aiemmista. Muutokset mahdollistivat helpomman ja käsitteistöltään rikkaamman annotoinnin. Koko eepoksen ensimmäinen annotaatio valmistui lokakuuhun 2008 mennessä Tuomas Palosen laatimana, Jouni Hyvösen ja muun tutkimusryhmän tukemana. Vuonna 2009 työtä jatkettiin vielä narratiivisten tapahtumarakenteiden osalta, joiden kautta jäsenneltiin eepoksen juonirakennetta pienempiin yksiköihin. Järjestelmän tietoteknisessä suunnittelussa ja toteuttamisessa keskeisissä rooleissa ovat olleet muiden muassa Joeli Takala (Kalevala-sovellus), Jussi Kurki (suosittelujärjestelmä ja grafiikka) ja Eetu Mäkelä (MuseoSuomen ja Kulttuurisammon julkaisualustat) sekä KOKO-ontologian osalta Katri Seppälä. Lauri Harvilahti ja Heli Kautonen ovat

osallistuneet *Semanttisen Kalevalan* kehittämiseen Suomalaisen Kirjallisuuden Seuran (SKS) edustajina ja asiantuntijoina.

Kulttuurisampo julkistettiin syyskuussa 2008 ja on käytettävissä osoitteessa http://www.kulttuurisampo.fi/. Tutkimushankkeen kotisivu on http://www.seco.tkk.fi/applications/kulttuurisampo/.

Semanttinen Kalevala ja Kulttuurisampo ovat osa kansallista *Suomalaiset semanttisen webin ontologiat* -projektia *FinnONTO* (2003-2010), jota rahoittaa Tekes ja 38 yrityksestä ja muusta organisaatiosta koostuva konsortio. Pääosan *Semanttisen Kalevalan* annotointityöstä on rahoittanut Suomen Kulttuurirahasto.

Viitteet

1. Kalevala-tutkijat laskevat, että Elias Lönnrot (1802–1884) teki yhteensä viisi *Kalevalaa*. *Kalevalan* esityöt valmistumisjärjestyksessä ovat seuraavat: 1) käsikirjoituksiksi jääneet pienoisrunoelmat *Lemminkäinen*, *Väinämöinen* ja *Naimakansan virsiä* (vuonna 1833), 2) alkuharjoitelmana / varhaisversiona tulevan eepoksen juonellisen perusrakenteen idean sisältävän *Runokokous Väinämöisestä* -käsikirjoituksen (1834), 3) *Vanhan Kalevalan* (1835), 4) *Uuden Kalevalan* (1849) sekä 5) kouluopetusta varten laaditun *lyhennetyn Kalevalan* (1862).
2. KOKO:on kuuluu myös useita erikoisalojen ontologioita toisiinsa yhdistettynä, kuten Museoviraston ylläpitämä museoalan ontologia MAO, Viikin tiedekirjaston maa- ja metsätalousalan ontologia AFO, taideteollisuusalan ontologia TAO sekä valokuvausalan ontologia VALO.
3. Espen Aarseth näkee hypertekstinlukijan aktiivisena pelaajana, kun lineaarisen tekstin lukija määrittyy passiiviseksi vastaanottajaksi. Hänen mukaansa hypertekstissä on kyse pelistä, ei narratiivista, vaikka kategorioiden välinen ero onkin häilyvä. (Aarseth 1997, 4–5.) Aarseth tosin käyttää hypertekstin sijaan termiä kyberteksti, joka kattaa sähköisen tekstin lisäksi myös painetun tekstin ja tarkoittaa ”mahdollisten tekstuaalisuuksien laajaa kirjoa (tai näkökulmaa), joka ymmärretään koneiden typologiaksi, erilaisiksi kirjallisiksi kommunikaatiosysteemeiksi, joissa toiminnalliset erot mekaanisten osien välillä määräävät esteettistä prosessia” (Aarseth 1997, 22; käännös: Markku Eskelinen, Anna-Kaarina Kippola & Raine Koskimaa). Toisaalla Aarseth (1994, 51) määrittelee kybertekstin itsestään muuttuvaksi tekstiksi, jonka sisällön jäsentymistä kontrolloi mekaaninen tai inhimillinen toimija. Epälineaarisen tekstin Aarseth (1994, 51) määrittelee verbaalisen viestinnän objektiksi, jolla ”ei ole yhtä kiinteää kirjainten, sanojen ja lauseiden sekvenssiä, vaan sekvenssi voi vaihdella lukemiskerrasta toiseen tekstin muodon, konventioiden tai mekanismien vuoksi”.
4. Tosin valitsemalla sisällön kuvailuun käsitteitä, joiden merkitys viime kädessä yksilöidään webin URI-tunnisteilla eikä niiden ilmiasulla kuten perinteisessä asiasanoituksessa, voidaan kuvailu tehdä ontologialla asiasanastoa tarkemmin muiden muassa synonyymisten ja polyseemisten ilmausten osalta. Näin esimerkiksi annotaatio ”pää” saadaan merkitsemään ruumiinosaa eikä esimerkiksi perheen, nuolen

tai valtionpäätä. Näin sisällöt eivät mene sekaisin tiedon haussa ja suosittelussa haettaessa Kulttuurisammosta vaikkapa taidesisältöjä, joissa esiintyy jokin ruumiinosa. Annetussa ”tappo”-esimerkissä Kulttuurisampo linkittää säkeitä erilaisiin rikoksiin liittyviin sisältöihin, mikä ei tässä tapauksessa ole toivottavaa, joten käsitteistöä olisi syytä laajentaa yleisellä tappamisen käsitteellä, jonka voisi sitten liittää nykyiseen ”tappoon” semanttisesti. Vaihtoehtona on joko sivuuttaa kyseinen kohta annotaatiossa tai käyttää jotakin muuta käsitettä (esim. ”väkivalta”), mutta usein tällaisessa tapauksessa annotoinnin tarkkuus heikkenee.

5. Vladimir Propp käsittää tarinan joukkona funktioita, saussurelaisia merkitystason muuttumattomia yksiköitä, joiden ilmentymiä yksittäiset aktit ovat. Tämä käsitys periytyy myös Claude Bremondille ja Roland Barthesille. Proppin mukaan tarinan metarakenne on 31 funktion sekvenssi: yksittäisissä tarinoissa toteutuu vaihteleva määrä funktioita, mutta funktioiden keskinäinen järjestys on muuttumaton (Propp 1958 [1928], 13–32). Bremondin mukaan Proppin malli ei kuitenkaan kuvaa tarinoihin sisältyvää variaatiota. Bremondin (1970, 250) mallissa tarina rakentuu ”päällekkäisistä perussekvensseistä, jotka on kiedottu yhteen. Jokainen tapahtuma voi täyttää yhtä aikaa useita funktioita tarinassa eli edistää useita rinnakkaisia perussekvenssejä.” Bremond jakaa sekvenssin nelivaiheiseksi sykliksi, jonka vaiheet ovat puutos, kehitys, tyydytys ja taantuminen. Funktiot määräytyvät näkökulman mukaan: samalla tapahtumalla voi olla eri funktio eri toimijoille. (Bremond 1970, 247–252.) Barthes puolestaan jakaa funktiot neljään luokkaan sen mukaan, miten tärkeitä ne ovat tarinan etenemisen kannalta. Tärkeimpiä ovat kardinaalifunktiot, jotka avaavat, jatkavat tai sulkevat vaihtoehdon, joka oleellisella tavalla vaikuttaa tarinan kehitykseen. Bremondin tavoin Barthes puhuu funktioiden sekvensseistä, jotka limittyvät ja ketjuuntuvat. Barthes käsittää tarinan hierarkiana, jossa limittäisten sekvenssien ketjun päättyminen on merkki rajaviivasta tarinan korkeammalla tasolla, episoditasolla (Barthes 1996 [1966], 45–54.)
6. Ketjun vaiheita voidaan määritellä sisällöllisesti esimerkiksi seuraavalla tavalla. Alkuvaihe esittää alkuasetelman, lähtökohdan, puutteen tai pyrkimyksen. Kehittely sisältää toimintaa, joka tähtää alussa nimettyyn päämäärään, tai alkuasetelman jälkeisen muutoksen tai ongelman. Kohokohdassa kuvataan jonkinlainen onnistuminen, epäonnistuminen, yllätys, kohtaaminen, konflikti tai muu draamallinen huippu. Kohokohtaa seuraa ratkaisu, lopputulos, tilanteen asettuminen tai alkuasetelman muuttuminen.
7. Tätä kerronnallista strategiaa edustaa toimittajan opetusfunktiossa esitetyt mielipiteet ja arviot esimerkiksi Kullervo-runostossa (UK 36), Kultaneidon taonnassa (UK 37) ja Tuonella käynti -runossa (UK 16). Esimerkiksi Kullervo-runon loppuepisodissa Väinämöinen varoittaa, kieltää, ohjaa tulevia sukupolvia toimimaan tietyllä tavalla, kasvattamaan jälkikasvunsa rakkaudella ja oikeudentunnolla, tyytymään osaansa ja välttämään turhamaisuutta ja luonnottomuutta. Esimerkiksi Ilmarisen taottua itselleen Kultaneidon ja tämän osoittautuessa ”kylmäksi kumppaniksi” (UK 37: 181–196), seuraa tätä episodi, jossa Väinämöinen kehottaa olemaan ”tekemättä kullasta puolisoa” (UK 37: 221–250). Kyseisen kaltainen vuorosanojen käyttäminen oli toimitusstrategisesti tyypillistä Lönnrotin eepostyöskentelylle.

8. Kansanrunoaineksista oli rakennettava juonellinen kokonaisuus homeeristen eeposten ja eurooppalaisen eeppisen runotradition esikuvien mukaisesti. *Semanttinen Kalevala* tarjoaa mahdollisuuden uppoutua Lönnrotin työn haasteisiin ja eepostekstin komposition perusteisiin. Jos hankkeen tulevat suunnitelmat toteutuvat, tätä lähdekriittistä kuvausta syventää entisestään *Kalevalan* yksityiskohtainen linkitys lähteinä toimineisiin kansanrunoihin. Tässä SKVR-linkityksessä hyödynnetään Väinö Kaukosen ja A. R. Niemen tekemää perustutkimusta, mutta semanttiset tekniikat mahdollistavat uusien yhteyksien jäsentämisen laajemmin SKVR-korpukseen. Tämä mahdollisesti vaikuttaisi perustutkimuksen sangen säekeskeiseen tarkastelutapaan ja tarjoaisi tilaisuuden valottaa Lönnrotin työskentelytapoja suhteessa laajempaan lähdepohjaan.
9. Henry III Fine Rolls -projektissa tekstimateriaali, englantilainen kokoelma kuninkaallisia kirjanpidollisia asiakirjoja 1200-luvulta, on koodattu TEI:llä ja lisäksi tekstissä esiintyvistä toimijoista ja paikoista on luotu instanssit eli tiettyä yksilöä ilmentävät yksiköt ontologiaan (Ciula, Spence & Vieira 2008). Oslon yliopistossa on niin ikään luotu kulttuuriperinnön tapahtumakohtainen tietokanta, jossa TEI-enkoodaus ja ontologia yhdistyvät. (Ore & Eide 2009.) Tapahtumapohjaista semanttista sisällönkuvailua on sovellettu myös videomateriaaliin, mikä on hyvin lähellä *Semanttisen Kalevalan* kuvailutapaa (ks. esim. Salminen 2006; Jung ym. 2004).
10. Kalevala-tekstin ja Kalevala-taulujen asiasanoitusta vertailemalla tämän problematiikan voi huomata. Yhtäältä taulussa on visualisoitu asioita, joita teksti ei välttämättä kuvaile (esim. ympäristö, vaatetus), ja toisaalta taulussa ei voi kuvata kaikkea, mitä tekstissä kuvataan (esim. keskustelun aihe, tunnetilat). Esimerkiksi Gallen-Kallelan *Kullervon kirous* -maalauksessa päähenkilön ympärille on asetettu esineitä ja muita elementtejä, joita ei esiinny Kalevalan eepostekstissä. Semantiikka ei siis ole näkökulmasta riippumatonta, vaan esimerkiksi maalaustaiteella ja runonlauluperinteellä on oma semantiikkansa.

Kirjallisuus

AARSETH, ESPEN 1997: *Cybertext: Perspectives on Ergodic Literature*. Baltimore: The Johns Hopkins University Press.

AARSETH, ESPEN 1994: Nonlinearity and Literary Theory. – Landow, George P. (toim.), *Hyper/Text/Theory*. Baltimore: Johns Hopkins University Press.

ANTONIOU, GRIGORIS & VAN HARMELEN, FRANK 2008: *A Semantic Web Primer*. Lontoo: The MIT Press.

BARTHES, ROLAND 1996: Introduction to the Structural Analysis of Narratives. – Onega, Susana & Landa, José Angel García (toim.), *Narratology: An Introduction*. Lontoo: Longman. [1966]

BEOWULF IN HYPERTEXT [online]. <http://www.humanities.mcmaster.ca/~beowulf/> [10.2.2009.]

BREMOND, CLAUDE 1970: Morphology of the French Folktale. – *Semiotica* 2(3).

CIULA, ARIANNA & SPENCE, PAUL & VIEIRA, JOSÉ MIGUEL 2008: Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project. – *Literary and Linguistic Computing* 23(3).

DOUGLAS, J. YELLOWLEES 2000: *The End of Books - or Books Without End? Reading Interactive Narratives*. Ann Arbor: University of Michigan Press.

AN eEDITION OF THE WEDDING OF MUSTAJBEY'S SON BEĆIRBEY [online]. <http://oraltradition.org/zbm> [10.2.2009.]

FOLEY, JOHN MILES 2009: *Getting started: How to Surf the Pathways Project* [online]. <http://www.pathwaysproject.org/pathways/show/GettingStarted> [26.8.2009.]

HAYLES, N. KATHERINE 2008: *Electronic Literature. New Horizons for the Literary*. Notre Dame: University of Notre Dame Press.

HONKO, LAURI 1987: Kalevala: aitouden, tulkinnan ja identiteetin ongelmia. – Honko, Lauri (toim.), *Kalevala ja maailman eepokset*. Helsinki: SKS.

HYVÖNEN, EERO 2005: *Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita* [online] <http://www.seco.tkk.fi/publications/2005/hyvonen-miksi-asiasanastot-eivat-riita-2005.pdf> [26.8.2009.]

— 2007: Älykäs semanttinen web tietämyksenhallinnan rajoja siirtämässä – esimerkkinä suomalainen kulttuuri semanttisessa webissä. – Raivio, Kari & Rydman, Jan & Sinnemäki, Anssi (toim.), *Rajalla – tiede rajojaan etsimässä*. Helsinki: Gaudeamus.

— 2008a: *Kulttuurisampo – suomalainen kulttuuri semanttisessa webissä. Muistiorganisaatioiden ja kansalaisten yhteisöllinen kansallinen julkaisujärjestelmä* [online]. Espoo: Teknillinen korkeakoulu, mediatekniikan laitos. <http://www.seco.tkk.fi/publications/2008/hyvonen-Kulttuurisampo-2008.pdf> [26.8.2009.]

— 2008b: *FinnONTO-malli kansallisen semanttisen webin sisältöinfrastruktuurin perustaksi - visio ja sen toteutus* [online]. Espoo: Teknillinen korkeakoulu, mediatekniikan laitos. <http://www.seco.tkk.fi/publications/2008/hyvonen-ONKI-yleisesitys-2008.pdf> [26.8.2009.]

HYVÖNEN, EERO & MÄKELÄ, EETU & KAUPPINEN, TOMI & ALM, OLLI & KURKI, JUSSI & RUOTSALO, TUUKKA & SEPPÄLÄ, KATRI & TAKALA, JOELI & PUPUTTI, KIMMO & KUITTINEN, HEINI

& VILJANEN, KIM & TUOMINEN, JOUNI & PALONEN, TUOMAS & FROSTERUS, MATIAS & SINKKILÄ, REETTA & PAAKKARINEN, PANU & LAITIO, JOONAS & NYBERG, KATARIINA 2009: CultureSampo – Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user. – *Proceedings, Museums and the Web 2009, Indianapolis, USA, April 15-18, 2009* [online]. Archives & Museum Informatics, Toronto, Canada. <http://www.seco.tkk.fi/publications/2009/hyvonen-et-al-culsa-mw-2009.pdf> [26.8.2009.]

JUNG, BYUNHGEE & KWAK, TAETEONG & SONG, JUNEHWA & LEE, YOONJOON 2004: Narrative Abstraction Model for Story-oriented Video. – *Proceedings of the 12th Annual ACM International Conference on Multimedia*. New York.

JUNNILA, MIIKKA 2006: *Tietosisältöjen semanttinen yhdistäminen toimintakuvausten avulla*. Pro Gradu -tutkielma [online]. Helsinki: Helsingin yliopisto, tietojenkäsittelytieteen laitos. <http://www.seco.hut.fi/publications/2006/junnila-tietosisaltojen-semanttinen-yhdistaminen-2006.pdf> [26.8.2009.]

JUNNILA, MIIKKA & HYVÖNEN, EERO & SALMINEN, MIRVA 2008: Describing and Linking Cultural Semantic Content by Using Situations and Actions. – Robering, Klaus (toim.), *Information Technology for the Virtual Museum*. Berliini: LIT Verlag.

LANDOW, GEORGE P. 2006: *Hypertext 3.0. Critical Theory and New Media in an Era of Globalization*. Baltimore: The Johns Hopkins University Press.

LIU, YIN & SMITH, JEFF 2008: *A Relational Database Model for Text Encoding* [online]. <http://www.chass.utoronto.ca/epc/chwp/CHC2007/Liu_Smith/Liu_Smith.htm> [26.8.2009.]

MCHOUL, ALEC & ROE, PHIL 1996: *Hypertext and reading cognition* [online]. <http://wwwmcc.murdoch.edu.au/ReadingRoom/VID/cognition.html> [26.8.2009.]

ONG, WALTER J. 1982: *Orality & Literacy. The Technologizing of the Word*. Lontoo: Routledge.

ORE, CHRISTIAN-EMIL & EIDE, ØYVIND 2009: TEI and cultural heritage ontologies: Exchange of Information? – *Literary and Linguistic Computing* 24(2).

PROPP, VLADIMIR 1958: *Morphology of the Folktale*. Indiana: Indiana University. [1928]

ROBERTSON, BRUCE 2009: Exploring Historical RDF with Heml. – *Digital Humanities Quarterly* 3(1) [online]. <http://www.digitalhumanities.org/dhq/vol/003/1/000026.html> [26.8.2009.]

RUSSEL, STUART & NORVIG, PETER 2005: *Artificial Intelligence: A Modern Approach*. New York: Prentice Hall.

SAEED, JOHN I. 1997: *Semantics*. Oxford: Blackwell Publishers.

SALMINEN, MIRVA 2006: *Kuvien ja videoiden semanttinen sisällönkuvailu*. Pro gradu -tutkielma. Helsinki: Helsingin yliopisto, tietojenkäsittelytieteen laitos.

SOWA, JOHN 2000: Knowledge *Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove: Brooks Cole Publishing Co.

VALKEAPÄÄ, ONNI & ALM, OLLI & HYVÖNEN, EERO 2007: Efficient content creation on the semantic web using metadata schemas with domain ontology

services. – *Proceedings of the 4th European Semantic Web Conference (ESWC 2007).* Innsbruck: Springer-Verlag.
VANHOUTTE, EDWARD 2006: *Electronic Textual Editing: Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text-Encoding for Electronic Editions* [online]. <http://www.tei-c.org/About/Archive_new/ETE/Preview/vanhoutte.xml> [26.8.2009.]
WHITE, HAYDEN 1980: The Value of Narrativity in the Representation of Reality. – Mitchell, W. J. T. (toim.), *On Narrative.* Chicago: University of Chicago Press.
ZARRI, GIAN PIERRO 2008: *Representation and Management of Narrative Information: Theoretical Principles and Implementation.* Berliini: Springer.

Filosofian maisteri Tuomas Palonen toimii tutkijana kansallisessa FinnONTO-projektissa.

Filosofian lisensiaatti Jouni Hyvönen valmistelee väitöskirjaa Elias Lönnrotin tekstualisointistrategioista ja Kalevala-eepoksen synnystä.

Tekniikan ylioppilas Joeli Takala valmistelee diplomityötä narratiivien kuvailusta semanttisessa webissä.

Professori Eero Hyvönen toimii Teknillisen korkeakoulun Semanttisen laskennan tutkimusryhmän tutkimusjohtajana.