Informaatiotutkimuksen päivät 2014
6. - 7. marraskuuta, Oulun yliopisto, Oulu

ABSTRAKTI

Satu Niininen & Susanna Nykyri

Monikielistä ja yksikulttuurista Yleistä suomalaista ontologiaa rakentamassa

Yhteystiedot: Satu Niininen, Helsingin yliopisto, Kansalliskirjasto, satu.niinen@helsinki.fi
Susanna Nykyri, Helsingin yliopisto, Kansalliskirjasto, susanna.nykyri@helsinki.fi

Mennyt ja tuleva

Kansalliskirjasto kehittää paraikaa Yleistä suomalaista ontologiaa (jatkossa YSO/ALLFO), kielinään suomi, ruotsi ja englanti. Ontologia pohjautuu sanastoltaan Yleiseen suomalaiseen asiasanastoon (YSA) ja sen ruotsinkieliseen versioon (Allärs). Tesauruksen ja ontologian laadinnassa on paljon samoja perusperiaatteita, mutta myös eroavaisuuksia. Käytännön kehittämistyössä yksi suurista muutoksista on suhtautuminen monikielisyyteen.

Tesaurusten kehittämistyössä on perinteisesti jaoteltu tesaurukset yksi- ja monikielisiin, vaikka käytännön kehitystä palvelisi paremmin kulttuurinäkökulman kiinteä sisällyttäminen kehitystyöhön ja käännösstrategian tietoiseen valintaan (Nykyri 2010). Yleisen suomalaisen ontologian kehitystyössä luodaan yksikulttuurista (suomalaista) ontologiaa, joka on kolmikielinen (suomi, ruotsi, englanti). Käytännössä tämä tarkoittaa, että YSOn käsitepohja on suomalainen, ja se muodostetaan suomen ja suomenruotsin pohjalta. Englanti on käännöskieli, ja siinä käännösvastine haetaan vastaamaan suomalaista, ontologiaan jo muodostettua käsitettä. Englannin käännöstyön yhteydessä YSOa on myös ryhdytty siltaamaan kansainvälisiin sanastoresursseihin.

YSOn pohjana on Yleinen suomalainen asiasanasto, joka ontologisoitiin tutkimuksellisista lähtökohdista käsin Aalto-yliopistossa SeCo-ryhmässä (Seppälä & Hyvönen 2014). Sittemmin (2013) YSOn kehitys siirtyi Kansalliskirjastolle (projektista ks. Kansalliskirjasto 2014), ja sen kehittämisen ohjenuoraksi otettiin toimivuus sisällönkuvailun ja tiedonhaun tarpeisiin (ks. Lappalainen, Nykyri, Palonen 2013; Lappalainen, Frosterus, Nykyri 2014) ja käyttäjäryhmien huomioiminen (Nykyri & Palonen 2014).

Miksi monikielinen ja yksikulttuurinen

Nykyajan linkittyvässä ja globaalissa ympäristössä (*open linked data, semantic web*) tarvitaan eri ajassa ja paikassa laadittujen, sisällönkuvailtujen ja tarjolle saatettujen aineistojen yhteiskäytön mahdollistavia yhteisiä - tai yhdistäviä - käsitteistöjä ja sanastoja. Niillä halutaan mahdollistaa se, että yhdellä kielellä asiasanoitettu aineisto on haettavissa myös toisella kielellä, ja että samanaiheiset aineistot ovat löydettävissä yhteisten portaalien kautta riippumatta siitä, missä ja kuka ne on sisällönkuvaillut. Kun halutaan ylittää kieli- ja välillä myös diskurssirajat, on

mietittävä, että mihin käsitteistön muodostus perustuu, ja miten käsitteet rajataan ja kielellistetään varsin heterogeeniselle käyttäjäryhmälle.

Sanoilla on useita samanaikaisia merkityksiä - vallitsevia, jäänteenomaisia ja orastavia. Merkitysten maailmassa vallitsevat merkitykset ovat alati kyseenalaistuvia, ja merkityskarttoja on vallalla samaan aikaan useita. “Hallitsevassa asemassa olevien ryhmien merkityskartoilla on taipumusta tulla koko kulttuurin käyviksi tavoiksi luokitella ja järjestää todellisuutta.” (Lehtonen 2000. s. 25) Onkin siis niin, että tässäkin yhteydessä kun puhutaan yksikulttuurisesta YSOsta, niin puhutaan nimenomaan yleisestä ja eniten vallalla olevasta sanastosta, ei alakulttuurien ja diskurssien runsaista variaatioista.

Yhteisiä sanastollisia sisällönkuvailun ja tiedonhaun välineitä kehitettäessä törmätään varsin ikiaikaisiin haasteisiin. Kielen koetaan kuitenkin yleisesti heijastavan ympäröivää maailmaa ja kantavansa puhujiensa todellisuutta, arvoja ja asenteitakin (ks. esim. Eco 2005). Kontrolloidun asiasanaston, kuten tesauruksen, laadinnan yhteydessä kehittäjän tulisikin valita lähestymistapansa - ovatko käsitteet tarkkarajaisia vai sumeita (Lykke Nielsen 2002, 16). Yhteiskunnallisella tasolla monikulttuurisuus edellyttää paitsi eroavaisuuksien hyväksymistä, niin myös uusia yhteisiä monikulttuurisuudesta kumpuavia arvoja (ks. Fukuyama 2006).

YSOn nykykehityksessä on omaksuttu ajatus siitä, että YSO heijastaa suomalaisia sisällönkuvailutarpeita. Siinä ei pyritä mallintamaan maailmaa eikä lähestymään käsitteitä esimerkiksi japanilaisesta kulttuurista käsin. Kansainvälistäminen (ks. Nykyri 2010, 84) ja mahdollisimman globaali linkittyvyys toteutetaan englanninkielisen käännöksen siltaamisen kautta. Ihmiset tarvitsevat eri yhteyksissä erilaisia tapoja erotella asioita, mutta globaalissa linkittyvässä ympäristössä ne tulisi voida tarvittaessa yhdistää mahdollisimman saumattomasti.

Monikielisyyden haasteista

Kolmikielisessä työskentelyssä haasteita on monia. Standardeissa (ks. ISO 25964-2) tesaurusten käännösekvivalenssin ja muihin sanastoihin siltaamisen asteiden tunnustaminen ja erottelu pohjautuu sille ajatukselle, ettei täydellinen vastaavuus eri kielten ja kulttuurien välillä käytännössä ole aina tavoitettavissa. Tämä johtaa siihen, ettei käsitteelle välttämättä ole löydettävissä yksiselitteistä ja tyhjentävää vastinetta toisella kielellä vaan olemassa olevista vaihtoehdoista joudutaan tekemään valinta sen perusteella, mikä vaikuttaa käyttökontekstin kannalta parhaiten soveltuvalta. Erityisesti englannin kielen yhteensovittaminen on haastavaa, sillä koska englanti ei ole suomalaisen kulttuuripiirin kieli, joudutaan käännöstyössä ylittämään varsin suuri käsitteellinen kuilu.

YSO/ALLFOn sisältötyön lähtökohtana on ollut se, että suomi ja ruotsi ovat molemmat tasa-arvoisia käsitteenmuodostuskieliä kun taas englanti on asemaltaan käännöskieli, jonka funktiona on välittää suomalaisen kulttuuripiirin käsitteistö toiselle kielelle. Käytännössä tämä tarkoittaa, että silloin kun suomen ja ruotsin käsitteistöt eroavat toisistaan, toisessa kielessä joudutaan tekemään kompromisseja käsitteistön yhteensovittamiseksi. Tämä näkyy esimerkiksi käsitteessä *joet*, jossa ruotsin kielen kaikki kolme jokityyppiä *älvar, åar, floder* on kiinnitetty yhteen käsitteeseen (*älvar* suositeltavana terminä, *åar* ja *floder* korvattuina termeinä). Toiseen suuntaan on joustettu *keskustelu*-käsitteessä, joka on suomeksi jaettu merkityksiin *keskustelu (pohdinta), keskustelu (puhe)*, ja *keskustelu (väittely)*, koska ruotsissa on olemassa käsitteet kolmelle eri keskustelulle *debatt, diskussion* ja *samtal*. Kompromisseja pyritään tekemään puolin ja toisin, mutta käytännössä joustoja on jouduttu tekemään enemmän ruotsinkielisessä käsitteistössä.

Suomen- ja ruotsinkielisessä sisältötyössä jokaisen käsitteen ajatellaan olevan oma uniikki ajatusyksikkönsä, jolloin myös käsitteistä käytettävät termit ovat yksilöiviä. Englannin asema käännöskielenä sen sijaan mahdollistaa sen, että useammasta käsitteestä voidaan käyttää samaa

termiä, mikäli käsitteiden välillä ei englannissa tehdä eroa (esim. käsitteet *uudenaikaistaminen* ja *modernisaatio* kääntyvät molemmat *modernisation*). Tarvittaessa käännösvastineen epätarkkuutta voidaan täsmentää esimerkiksi sulkutarkenteiden tai käyttöhuomautusten avulla, jolloin sen merkityssisältö rajautuu haluttuun suuntaan. Tällöin esimerkiksi *tasa-arvo* kääntyy muotoon *equality (values)* ja yhdenvertaisuus muotoon *equality (fundamental rights)*.

Kielten erilaisuus näkyy myös siinä, että YSO/ALLFOn hierarkianäkymän selailu on osittain puutteellista muilla kielillä kuin suomeksi. Esimerkiksi *tarvikkeet*-käsitteen kaikki alakäsitteet ovat suomeksi *tarvikkeet*-päätteisiä yhdyssanoja, kuten *rakennustarvikkeet, sairaanhoitotarvikkeet* jne. Ruotsiksi ja englanniksi *tarvikkeet* kääntyy *tillbehör* ja *supplies*, mutta sen alakäsitteiden muoto vaihtelee (esim. *byggvaror / building materials* ja *sjukvårdsartiklar / nursing equipment)*, jolloin ne eivät aina asetu yläkäsitteensä alle yhtä luontevasti kuin suomen kielellä. Hierarkian toimimattomuus englanniksi ei johda sisältötyössä erityisiin toimenpiteisiin, mutta ruotsin kielen ongelmatapauksissa pyritään aina ensisijaisesti löytämään ratkaisu, joka toimisi molemmilla kielillä.

On kuitenkin huomattava, että käännöskielen asemasta huolimatta englanti osallistuu YSOn käsitteenmuodostukseen siinä mielessä, että käännösprosessi saattaa sivutuotteenaan paljastaa käsitteistä monimerkityksisyyttä, joka on ontologoinnissa jäänyt pimentoon ja edellyttää käsitteen tarkempaa ankkurointia haluttuun merkitykseen. Esimerkiksi käsite *heimosodat* kääntyy eri tavalla riippuen siitä, viitataanko sillä heimosotiin Suomen historian kontekstissa (*Finnish Kinship Wars*) vai heimosotiin yleensä (*tribal wars*).

Englannin käännöstyön yhteydessä YSOa on sillattu Library of Congress Subject Headingsiin (LCSH). Siltauksessa etsitään LCSH:sta YSO-käsitteitä vastaavat käsitteet, ja luodaan näiden välille linkitys.

Monikielisessä siltaamisessa haasteita on useita. Ongelmallisia ovat esimerkiksi käsitteet, jotka ovat merkityssisällöltään vain osittain päällekkäisiä tai joissa yhden käsitteen merkityssisältö ilmaistaan toisessa sanastoresurssissa useamman käsitteeen osittaisella yhdistelmällä (esim. YSO:ssa on käsite *uhmakkuushäiriö*, LCSH:ssa *Oppositional defiant disorder in children* ja *Oppositional defiant disorder in adolescence*). Vähintään yhtä ongelmallisia ovat tapaukset, joissa lähde- ja kohdesanastojen käsitteistö liikkuu eri spesifisyystasolla. Esimerkiksi YSO/ALLFOsta löytyy käsitteet *ampuma-aseet* ST *ilmakiväärit*, LCSH:sta vähemmän spesifi *Firearms* ST *Air guns*, joka kattaa ilmakiväärien lisäksi myös ilmapistoolit.

Ongelmatapausten yhteydessä on aina punnittava, onko siltaus niin vahva, että se täyttää ekvivalenssille määritetyt vaatimukset. Keskeisenä haasteena työn alkuvaiheessa onkin ollut ongelmakentän kartoitus, jotta voidaan tarkemmin määrittää, millaisella tarkkuudella vastaavuus määritellään ja millaisia linjauksia siltauksissa noudatetaan, jotta ne toteutuisivat läpi koko sanaston johdonmukaisina. Kaikille käsitteille ei löydy siltausvastinetta lainkaan, ja näissä tapauksissa käsite jää siltaamatta; toistaiseksi noin joka kolmas YSO/ALLFO-käsite on saatu sillattua LCSH:iin.

Ratkaisun haasteet

Dokumentaatiokieli ja sisällönkuvailu eivät ole kielellisistä ja kulttuurisista ilmiöistä, ihanteista eikä rajotteistakaan vapaita. Tehdyt periaatteelliset ratkaisut ovat usein kannatettavia makrotasolla - tavoite on edistää tiedon linkittyvyyttä ja saantia yli kieli- ja kansallisrajojen. Auktorisoidussa sanastossa ei haluta antaa painokertoimia tms. kielen puhujien määrän mukaan, vaan vähemmistökieli on yhtä tärkeässä asemassa kuin enemmistökieli. Suurimmat haasteet esiintyvätkin lähinnä ontologiatyön mikrotasolla, jolloin joudutaan välillä tinkimään tavoitteista - ns. realiteettien vuoksi. Jo yksin eri kielten rakenteiden eroavaisuus aiheuttaa sen, että täydellistä ekvivalenssia ei aina saavuteta edes saman kattokulttuurin - kuten suomalaisuus - alla.

Kansainvälistyminen kielellisen yksinkertaistamisen (vrt. globish) kautta ei kuitenkaan voi olla ratkaisu. YSO-ALLFO-työssä onkin pyritty säilyttämään eri kielten ilmaisuvoimaisuus niin pitkälle kuin se monikielisessä dokumentaatiokielen sanastossa on käytännössä mahdollista ja tarkoituksenmukaista.

Lähteet

Eco Umberto 2005 (1995): *The Search for Perfect Language*. 4.p. (Trans. James Fentress) Oxford: Blackwell.

Fukuyama 2006: "Identity, Immigration and Democracy". *Journal of Democracy*, 2, 5-19. URL: http://journalofdemocracy.org/sites/default/files/Fukuyama-17-2.pdf

ISO 25964-1:2011 = Information and documentation - Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval. 1st ed. Geneva: ISO, 2011 (ISO 25964-1-2011-08-05).

ISO 25964-2:2013 = Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 2: Interoperability with other vocabularies. 1st ed. Geneva: ISO, 2013 (ISO 25964-2- 2013-03-04).

Kansalliskirjasto 2014: *Finto : Suomalainen sanasto- ja ontologiapalvelu*. URL: https://wiki.helsinki.fi/display/ONKI/ONKI-projekti

Lappalainen, Mikko & Frosterus, Matias & Nykyri, Susanna 2014: "Reuse of library thesaurus data as ontologies for the public sector". Paper presented at: IFLA WLIC 2014 – Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. URL: http://library.ifla.org/819/1/086-lappalainen-en.pdf

Lappalainen Mikko & Nykyri, Susanna & Palonen, Tuomas 2013: "Semanttisesta ja funktionaalisesta tiedonhaun ja –tallennuksen ontologiasta : suomalaisen yleisontologian laadinnan haasteita". *Informaatiotutkimus* 32(3-4) – 2013. URL: http://ojs.tsv.fi/index.php/inf/article/view/9445/6731

Lehtonen, Mikko 2000 (1996): *Merkitysten maailma. Kulttuurisen tekstintutkimuksen lähtökohtia*. 4. p. Tampere: Vastapaino.

Lykke Nielsen, Marianne 2002: *The word association method : a gateway to work-task based retrieval*. Åbo : Åbo Akademi University Press.

Nykyri, Susanna 2010: *Equivalence and Translation Strategies in Multilingual Thesaurus Construction*. Åbo Akademis förlag, Åbo.

Nykyri, Susanna & Palonen, Tuomas 2014: *Ontologioiden käytön moninaisuudesta ja tulevaisuusnäkymistä: Finto-palvelun sidosryhmien ontologiapalvelulle kohdistamat kehitystoiveet*. Kansalliskirjasto. URL: http://urn.fi/URN:NBN:fi-fe201402111454

Seppälä, Katri & Hyvönen, Eero 2014: *Asiasanaston muuttaminen ontologiaksi : Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista.* Kansalliskirjasto. URL: http://urn.fi/URN:ISBN:978-952-10-9883-3