Outi Meriläinen

Asiasanaekvivalenssi ja tesaurusten yhteensopivuus

Meriläinen, Outi. Asiasanaekvivalenssi ja tesaurusten yhteensopivuus [Descriptor equivalence and the compatibility of thesauri]. Informaatiotutkimus 15 (3): 83–91, 1996.

The concept of equivalence is considered in the context of compatibility of thesauri. Intra- and interlanguage equivalence in natural languages as well as linguistic equivalence vs. descriptor equivalence are specified. The consideration leads to defining two new concepts: dictionary equivalence and indexing equivalence which both are instances of descriptor equivalence. Indexing equivalence is further divided in potential and concrete indexing equivalence. The potential indexing equivalence between the descriptors of Liikunnan ja urheilun asiasanasto and the Canadian Sport Thesaurus is then described in the light of two different descriptor collections.

Address: Stakes/Information Service, P.O. Box 220, FIN-00531 Helsinki, Finland. E-mail: Outi.Meriläinen@stakes.fi

Tesaurusten yhteensopivuus tutkimuskohteena

Sisällönkuvailujärjestelmien yhteensopivuutta (compatibility) on tutkittu 1960-luvun alkupuolelta lähtien. Kiinnostus aihetta kohtaan laimeni 1970-luvulla, mutta elpyi jälleen 1980-luvun alussa, kun tietokoneet tulivat halvemmiksi ja niiden lisääntynyt kapasiteetti ja ohjelmien kehittyminen mahdollistivat yhä uusien eri kielisten, kansainväliseen käyttöön tarkoitettujen tietokantojen perustamisen. (Ks. esim. Dahlberg 1981; Svenonius 1983; Lancaster 1986, 179–229.) Innostus hiipui taas kymmenluvun loppua kohti, mutta aivan viime vuosina on yhteensopivuutta käsittelevä kirjallisuus lisääntynyt. Kansainvälisesti huomattavin merkki uudelleen virinneestä innostuksesta on ollut ISKO:n (International Society for Knowledge Organization) Puolassa syyskuussa 1995 järjestämä tutkimusseminaari "Compatibility and integration of order systems", josta on tekeillä proceedings-julkaisu.

Yhteensopivuus tutkimuskohteena alkoi kiinnostaa minua pari vuotta sitten, kun jouduin LISETI-sanastoa[1] (Meriläinen 1993) laatiessani ratkaisemaan suomen- ja englanninkielisten liikunnan asiasanojen välisiä vastaavuusongelmia sanakirjojen ja kohteena olleiden tesaurusten käytöstä saadun kokemuksen avulla. LISETI-kokemuksen innoittamana aloitin v. 1994 tutkimuksen asiasanojen vastaavuusongelmista kaksikielisen tiedonhakusanaston laatimisessa. Tutkimuksessani vertaan toisiinsa suomalaista Liikunnan ja urheilun asiasanastoa eli LUASia[2] (Liikunnan 1989 ja Liikunnan 1993) ja kanadalaista, englanninkielistä Sport Thesaurusta

eli SPORTia[3] (Sport 1990, Sport 1992). Vertailun avulla tutkin näiden tesaurusten yhteensopivuutta asiasanatasolla.

Tässä artikkelissa tarkastelen sisällönkuvailujärjestelmien yhteensopivuuden ja ekvivalenssin käsitettä yleensä sekä erityisesti niiden ilmenemistä TERVAS- ja INFSER-aineistossa. TERVAS-aineisto koostuu LUAS-SPORT-asiasanapareista ja vastineettomista LUAS-asiasanoista. INFSER-aineistossa ovat ne LUAS- ja SPORT-asiasanat, joita on käytetty suomalaisten dokumenttien indeksoinnissa vuosina 1989–1992.

Yhteensopivuus sisällönkuvailujärjestelmän ominaisuutena

Saksalaisen yhteensopivuutta tutkineen professori Ingetraut Dahlbergin mukaan yhteensopivuus on se sisällönkuvailujärjestelmän (ordering system) ominaisuus, joka mahdollistaa järjestelmän osioiden käytön toisen järjestelmän osioiden rinnalla tai sijasta (Dahlberg 1983, 5). Yhteensopivuus ei ole jakamaton kokonaisuus, vaan se muodostuu järjestelmien eri osa-alueiden yhteensopivuudesta. Yhteensopivuuden osa-alueet voidaan jakaa seuraaviin neljään alueeseen: tietosisällön jäsennys (mm. aihekate, luokitukset, hierarkiat ja notaatiot), lingvistinen taso (asiasanat ja niiden väliset suhteet), muodollinen esitystaso (mm. typografia, koodit, symbolit, asiasanojen kirjoitusasu ja järjestys), tietojenkäsittelyn taso (mm. isäntäjärjestelmä, tietueen rakenne, tietokannan organisointi) (Sager, Somers & McNaught 1981, 134–137).

Sisällönkuvailujärjestelmät voivat olla eri asteisesti yhteensopivia eri osa-alueilla. Mitä pitemmälle järjestelmät on tarkoitus käytännössä yhdistää sitä tärkeämmäksi tulee yhteensopivuus tietosisällön jäsennyksen ja tietojenkäsittelyn osa-alueella. Lisäksi yhteensopivuus on aina direktionaalista: järjestelmän A yhteensopivuus järjestelmän B kanssa voi olla suurempi tai pienempi kuin B:n yhteensopivuus A:n kanssa.

Yhteensopivuuden etuja ovat Dahlbergin mukaan

1) mahdollisuus hakea millä termillä tahansa useasta eri tiedostosta,
2) mahdollisuus löytää tiettyä käsitettä vastaava tieto mistä tahansa tietovarannosta, jonne se on indeksoitu jollakin varannossa käytetyistä indeksointikielistä,
3) mahdollisuus jonkin indeksointikielen käyttäjälle, saada tietää, että hänen hakemaansa aihetta on löydettävissä myös toisella tai toisilla indeksointikielillä kuvailtuna,
4) mahdollisuus löytää tietyn indeksointikielen asiasanoille vastineita toisesta tai toisista indeksointikielistä (Dahlberg 1981, 87).

Järjestelmien yhteensovittamiseksi on olemassa kaksi menetelmää variaatioineen. Ensiksikin voidaan laatia sanastojen välinen ns. switching language, joka kääntää indeksikielen A sanat muille järjestelmän indeksointikielille. Toiseksi voidaan yhdistää kahden tai useamman kielen sanastot keskenään yhdeksi yhteiseksi sanastoksi, jolloin menetelmää kutsutaan sanastojen sulauttamiseksi (integration) tai yhdistämiseksi (connecting, intertwining, mapping, matching, merging, reconciling) tai konkordanssin laatimiseksi (establishing of concordance)[4] (Scibor & Tomasik-Beck 1994; Hood & Eberman 1990; Buchan 1989; Mili & Rada 1988; Rada 1987; Lancaster 1986, 181–189; Svenonius 1983, 2; Dahlberg 1981; Sager, Somers & McNaught 1981). Se kumpaa perusmenetelmää käytetään riippuu siitä, millä tavalla järjestelmät halutaan käytännössä yhdistää ja miten paljon aikaa ja rahaa yhdistämiseen on käytettävissä.

LUAS- ja SPORT-tesaurusta ei ole tutkimustulosten avulla tarkoitus yhdistää kaksikieliseksi tesaurukseksi, vaan tavoitteena on hakemiston laatiminen suomenkielisten asiasanojen englanninkielisille tiedonhakuvastineille. Tähän tarkoitukseen sopivaa tietoa etsitään vertaamalla LUASia ja SPORTia toisiinsa lingvistisellä tasolla. Tietosisällön jäsennyksen, muodon ja tietojenkäsittelyn yhteensopivuutta käsitellään vain silloin, kun se vaikuttaa lingvistisen tason yhteensopivuuteen.

Kielen sisäiset ja kielten väliset vastaavuusongelmat

Yksikielistä asiasanastoa laadittaessa ongelmia ilmenee mm. synonyymien, kvasi-

synonyymien, laajojen ja suppeiden asiasanojen, kotoperäisten ja lainasanojen valinnassa. On ensinnäkin päätettävä mikä samaa käsitettä merkitsevistä sanoista on paras asiasanana ja mistä samaa tai lähes samaa käsitettä tarkoittavista sanoista tehdään käytäviittaukset. Valinnoissa tulisi noudattaa yksikielisten tesaurusten laatimista koskevaa standardia – suomenkielistä sanastoa laadittaessa siis SFS 5471 -standardia. Standardi ei kuitenkaan kata kaikkia erityistapauksia ja silloin ratkaisut on tehtävä tapaus- ja tesauruskohtaisesti. Tesaurusten rinnakkais- ja yhteiskäyttöä silmälläpitäen olisi toki suotavaa, että näissäkin ratkaisuissa pyrittäisiin sekä johdonmukaisuuteen tesauruksen sisällä että yhdenmukaisuuteen samanaiheisten tesaurusten välillä.

Toinen tilanne, jossa luonnollisen kielen sisäisiä vastaavuusongelmia joudutaan ratkomaan, on kahden samankielisen sanaston yhdistäminen. Ratkaisuja rajaa silloin tesauruskonventioiden lisäksi myös se, että yhdistämisen täytyy ottaa huomioon jo indeksoitu aineisto. Esimerkkinä tällaisesta tilanteesta on esim. vuonna 1991 tapahtunut SIRLS:n ylläpitämän Sport & Leisure-tietokannan[5] yhdistäminen Sport-tietokantaan. Sport & Leisure -tietokanta oli kanadalainen ja englanninkielinen kuten Sport-tietokantakin. Sen viitteiden indeksoinnissa oli käytetty SIRLS:n omaa 567:n asiasanan, strukturoimatonta luetteloa ja se piti sisällään lähinnä liikunnan, urheilun ja vapaa-ajan sosiologaa käsitteleviä dokumenttiviitteitä. Sen kattamat aiheet sisältyvät myös Sport-tietokantaan, johon tallennetaan laaja-alaisesti liikunnan eri osa-alueita käsitteleviä kirjallisuusviitteitä.

Ennen yhdistämistä SIRLSin asiasanoja oli verrattava Sport-asiasanoihin ja selvitettävä vastaavuudet. SIRLS:n käyttämistä asiasanoista 87 prosentille löytyi vastine (sama sana, synonyymi tai usean Sport-asiasanan yhdistelmä). Sanastojen samankielisyydestä huolimatta SIRLS:n asiasanoista 13 prosentille ei löydetty vastinetta Sportista, joten ne lisättiin Sport Thesaurukseen. Kaksi prosenttia SIRLSin asiasanoista hylättiin kokonaan. (Stark 1993, 3–4)

Sportin ja SIRLSin asiasanojen yhdistämistä käsittelevässä raportissa (Stark 1993) ei millään tavalla määritellä tai problematisoida asiasanojen ekvivalenssisuhteita. Voi vain olettaa, että vastinparien määrittelyssä tukeuduttiin samoihin periaatteisiin kuin yksikielistä sanastoa laadittaessa. TERVAS-aineistossa vastineet määritettiin sanakirjojen, käsitesanakirjojen ja LUASin ja SPORTin käytöstä kertyneen kokemuksen nojalla. INFSER-aineistossa vastineiden määrityksessä on lisäksi nojauduttu saman dokumentin kuvailussa käytettyjen suomen- ja englanninkielisten asiasanojen käsitepiirteiden vertailuun. Kun tarkasteltavana on kaksi erikielistä sanastoa, lisäävät luonnollisten kielten toisistaan eroavat lingvistiset ominaispiirteet ja sanaston ylläpito- ja käyttöympäristöjen kulttuuriset erot yhdistämisongelmia. Tätä kuvannee se, että yleisellä tasolla TERVAS-aineistossa jokaista vastineellista asiasanaa kohden on 0.35 vastineetonta asiasanaa, kun vastaava luku SIRLS/SPORT-aineistossa on 0.16.

Asiasanaekvivalenssi ja lingvistinen ekvivalenssi

Kielitieteessä on usein siteerattu lausetta "Equivalence in difference is the cardinal problem of language and the pivotal concern of linguistics" (Jakobson 1966, 233). Ekvivalenssi mielletään yleisesti keskeiseksi ongelmaksi, mutta on erilaisia käsityksiä siitä, mitä ekvivalenssilla oikeastaan tarkoitetaan.

Miten siis määritellä ekvivalenssi tai vastaavuus? Tesauruksen yhteydessä ekvivalenssi käsitetään suhteena, jossa asiasanaksi valittu termi korvaa indeksoitaessa ja haettaessa jonkin toisen termin. Ekvivalenssisuhteessa olevien sanojen oletetaan tällöin nimeävän saman tai melkein saman käsitteen. Mitä tässä yhteydessä 'sama tai melkein sama' tarkoittaa määritellään sekä suomalaisessa yksikielisen tesauruksen laatimisstandardissa että ISOn monikielisen tesauruksen laatimisstandardissa vain operationaalisesti luettelemalla ekvivalenssisuhteiksi soveltuvien suhteiden luokkia (SFS 5471 1988, s. 5–6 ja ISO 5964 1985, 7–9).

ISOn standardissa mainitaan suhdeluokkien lisäksi ekvivalenssin tason vaihtelu (indeksointikäytännön kannalta) täydestä vastaavuudesta täyteen vastaamattomuuteen:

"Due to the nature of language itself, terms selected from more than one natural language vary in the extent to which they represent the same concepts. These variations can be regarded as forming a continuum, one end of which is represented by terms that can, for the practical purposes of indexing, be regarded as exact equivalents, further points being marked by various degrees of partial of inexact equivalence, and the final point being represented by those extreme situations in which a term in one language refers to a concept which cannot be expressed by a single, direct and equivalent term in another language." (ISO 5964 1985, 7–8).

Tesauruskirjallisuuden suhteellisen epäproblemaattisena näyttäytyvä ekvivalenssikäsite on kuitenkin lingvistisenä käsitteenä sangen ongelmallinen määriteltävä. Osa kielitieteilijöistä on jopa sitä mieltä, ettei ekvivalenssikäsitettä voi ollenkaan käyttää varsinkaan, jos sillä tarkoitetaan täydellistä käsitetason vastaavuutta. Mm. amerikkalainen ekvivalenssia käännöstieteen näkökulmasta tutkinut Quian Hu toteaa artikkelissaan: "Linguistic facts prove that no full equivalence can ever be established between two languages. Even synonymy in the same language...does not yield equivalence" (Hu 1992, 291). Zürichin yliopistossa toimiva kielitieteen tutkija Mary Snell-Hornby ehdottaa ekvivalenssikäsitteen hylkäämistä kokonaan, koska se on hänen mielestään harhaanjohtava ja epämääräinen. Yhtenä todisteena ekvivalenssikäsitteen olemattomuudesta Snell-Hornby pitää saksankielen 'Äquivalenz' ja englanninkielen 'equivalence' termejä: nekään eivät hänen analyysinsa perusteella tarkoita täysin samaa käsitettä. (Snell-Hornby 1988, 434).

Edellä esitetyistä kriittisistä näkemyksistä huolimatta ekvivalenssi-termi esiintyy kielitieteellisessä ja erityisesti käännöstieteellisessä kirjallisuudessa tuhkatiheään itsestään selvänä ja määrittelemättömänä. Ekvivalenssikäsitteen ongelmaa pyritään myös kiertämään käyttämällä jakamattoman ekvivalenssikäsitteen sijasta suppeampia, rajattuja osakäsitteitä. Rajaus tehdään usein vastaavuussuhteen funktion avulla. Tällaisia osakäsitteitä ovat mm. funktionaalinen ekvivalenssi (Nida 1986), pragmaattinen ekvivalenssi (Kalisz 1981) ja dynaaminen ekvivalenssi (Tymczko 1985). Myös ekvivalenssikäsitteen kriitikko Hu päätyy siihen, että kirjallisuudesta löytyvät ekvivalenssikäsitteet "might well be subsumed under the dichotomy: formal equivalence and dynamic equivalence." (Hu 1992, 295). Tällöin kaikki yllä mainitut kolme osakäsitettä kuuluisivat dynaamisen ekvivalenssin nimen alle.

Asiasanaekvivalenssin kaksi lajia

Määrittelemäni kaksi asiasanaekvivalenssin lajia – sanakirjaekvivalenssi ja indeksointiekvivalenssi – kuuluvat lingvistisessä käsitteistössä dynaamisen ekvivalenssikäsitteen piiriin.

Sanakirjaekvivalenssi on sellainen asiasanojen välinen ekvivalenssisuhde, joka mahdollistaa niiden käytön toistensa sijasta (substituutio) luonnollisesta kielestä toiselle luonnolliselle kielelle käännettäessä ja joka voidaan todentaa yleisesti käytössä olevien sanakirjojen avulla.

Indeksointiekvivalenssi asiasanojen välillä vallitsee silloin, kun asiasanat nimeävät saman tai lähes saman käsitteen ja niitä voidaan käyttää toistensa sijasta indeksikieleltä toiselle käännettäessä. Indeksointiekvivalenssi voi olla potentiaalinen tai konkreettinen. Potentiaalinen indeksointiekvivalenssi on silloin, kun ekvivalenssin määritys perustuu vain asiasanalähteenä oleviin sanastoihin ja erilaisiin lingvistisiin menetelmiin ilman että nojaudutaan jo tehtyihin dokumenttien kuvailuihin. Konkreettiseksi indeksointiekvivalenssi muuttuu heti, kun sen osapuolia on konkreettisesti käytetty kuvaamaan saman dokumentin samaa tai lähes samaa käsitettä. Jos asiasanojen välillä vallitsee sanakirjaekvivalenssi, se lisää asiasanaparin indeksointiekvivalenssin todennäköisyyttä, mutta sanakirjaekvivalenssi ei ole indeksointiekvivalenssin edellytys.

Ekvivalenssin asteet

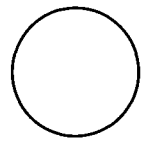
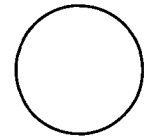
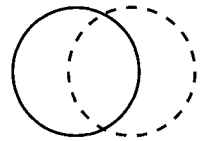
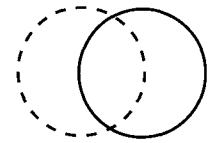
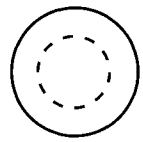
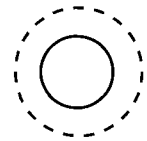
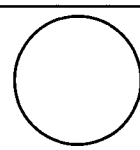
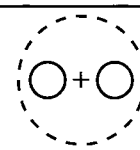
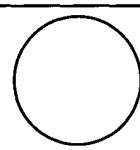
Sanakirjaekvivalenssi on periaatteessa kaksiluokkainen: jos suhde on ilmaistu käytetyssä leksikografisessa lähteessä, se on ole-

massa, muussa tapauksessa sanakirjaekvivalenssia ei ole. Jos sanakirjaekvivalenssia halutaan tarkentaa voidaan erottaa eri asteiksi yksi–yhteen ekvivalenssi (maastohiihto = cross-country skiing) ja yksi–moneen ekvivalenssi (suunnittelu = design; planning)

Sekä potentiaalinen että konkreettinen indeksointiekvivalenssi voivat vaihdella täydellisestä vastaavuudesta osittaisen vastaavuuden kautta täydelliseen vastaamattomuuteen. Indeksointiekvivalenssin asteet ovat samat kuin kuvassa 1 näkyvät monikielisten tesaurusten laatimista koskevan standardin esittelemät ekvivalenssin asteet.

Tarkka vastaavuus (exact equivalence) edellyttäisi oikeastaan, että asiasanat olisivat sekä muodollisesti että sisällöllisesti identtiset (täysi synonymia). Tällaisia synonyymeja ovat eri kielten välillä vain lainat, joissa lainatun käsitteen nimikin on pysynyt muuttumattomana (baseball = baseball). Tarkan vastaavuuden piiriin voidaan funktionaalisista syistä kuitenkin lukea myös ne tapaukset, joissa kielenkääntämisen kannalta on kyse

Table 2 – Degrees of equivalence

Case	Source language	Target language
1 – Exact equivalence		
2 – Inexact equivalence		
3 – Partial equivalence		
4 – Single-to-multiple equivalence		
5 – Non-equivalence		

 acceptable term exists  acceptable term does not exist

Kuva 1. Ekvivalenssin asteet (ISO 5964 1985, 9)

substituutiosta (uinti = swimming). Täydellisellä vastaamattomuudella (non-equivalence) puolestaan tarkoitetaan sellaista asiasanojen välistä suhdetta, jossa välittömiä muodollisia ja sisällöllisiä yhtymäkohtia ei ole lainkaan (puu – baseball)

Täydellisen vastaavuuden ja täydellisen vastaamattomuuden väliin jäävät osittaisen vastaavuuden luokat: epätäydellinen vastaavuus (inexact equivalence), osittainen vastaavuus (partial equivalence) ja yksi–moneen -vastaavuus (single–to–multiple -equivalence). Ensimmäinen vastaa assosiaatiosuhteita (diagnosointi – diagnosis) ja kaksi jälkimmäistä hierarkkisia suhteita asiasanojen välillä (synnytys – labour; veneily – boating/yachting).

Jatkotarkastelussani substituutiosuhde tarkoittaa yhteisesti synonyymi- ja substituutiosuhdetta, hierarkkinen suhde yhteisesti osittaista ekvivalenssisuhdetta ja yksi–moneen -ekvivalenssisuhdetta ja assosiaatiosuhde epätarkkaa ekvivalenssisuhdetta.

TERVAS- ja INFSER-aineiston ekvivalenssisuhteet

Ekvivalenssisuhteiden tarkastelussa käytän kahta eri aineistoa. TERVAS-aineiston muodostavat LUASin asiasanat ja niiden potentiaaliset indeksointivastineet SPORTissa. Vastinparin muodostavien asiasanojen väliset suhteet ovat joko substituutiosuhteita, hierarkkisia suhteita tai assosiaatiosuhteita. Jokainen asiasana muodostaa oman tietueensa, jossa on 40 kenttää. Kenttiin on koodattu mm. asiasanan sanaluokka, sanamuoto, viittaussuhteiden määrä, aiheluokka ja hierarkkinen taso. Jos asiasanalla on vastine, vastineesta on koodattu samat piirteet kuin LUAS-asiasanastakin ja ekvivalenssisuhteen laatu. Kaikkiaan asiasanatietueita on 2241.

Taulukosta 1 ilmenee Tervas-aineiston ekvivalenssisuhteiden suora jakauma. Neljännekselle suomalaisen tesauruksen asiasanoista ei löytynyt Sportista englanninkielistä vastinetta. Vastineellisista asiasanoista yli puolella (53,5 %) on sanakirjaekvivalenssisuhde (synonyymit ja substituutiot) ja lisäksi viidenneksellä (20,6 %) on hierarkkinen suhde tai assosiaatiosuhde vastineeseensa. Vastineellisia on siis kaikkiaan 84,1 prosenttia kaikista LUAS- asiasanoista. Puolet suomalaisten asiasanojen englanninkielisistä (luonnollisen kielen) substituutioista löytyy asiasanana tai viittausterminä SPORTista. Kun lisäksi neljännekselle LUASin asiasanoista löytyy samaan käsitteelliseen kokonaisuuteen kuuluva, LUAS-asiasanan merkityksen osittain kattava vastine, nousee vastaavuus asiasanatasolla 75 prosenttiin. Potentiaalinen indeksointiekvivalenssi on LUASin ja SPORTin välillä korkea.

INFSER-aineisto käsittää 79 viitettä, joita on kuvailtu molempien tutkittavien sanastojen asiasanoilla. Viitteet on indeksoitu LIKES-tietopalvelussa vuosina 1989–1992 suomalaiseen Finsport- ja kanadalaiseen Sport-tietokantaan. Otokseen sisältyy kaikkiaan 210 suomenkielistä ja 251 englanninkielistä asiasanaa, joista aineiston kriteerit täytti 426 asiasanaa. Asiasanoista ja niiden vastineista on INFSER- aineistoon tallennettu sekä sanojen ja niiden välisten suhteiden ominaispiirteet että lähdeviitteiden ominaisuudet. Tarkastelen tässä niitä INFSER-aineistossa esiintyviä

Taulukko 1: Ekvivalenssisuhteiden jakauma TERVAS-aineistossa (n=2241)

	f	%
Ei ekvivalenssisuhdetta	580	25,9
Synonyymisuhde	264	11,8
Substituutiosuhde	934	41,7
Hierarkkinen suhde	334	14,9
Assosiaatiosuhde	129	5,7
Yhteensä	2241	100,0

Taulukko 2: Niiden asiasanojen ekvivalenssisuhteiden jakauma jotka esiintyvät sekä TERVAS- että INFSER-aineistossa (n=426)

	f	%
Ei ekvivalenssisuhdetta	49	11,5
Synonyymisuhde	40	9,4
Substituutiosuhde	226	53,1
Hierarkkinen suhde	79	18,5
Assosiaatiosuhde	32	7,5
	426	100,0

asiasanoja, joilla TERVAS-aineiston perusteella on potentiaalinen vastine SPORTissa.

INFSER-aineistossa indeksointiin käytetyistä asiasanoista vain 11,5 prosenttia on sellaisia, joilla ei ole TERVAS-aineiston perusteella potentiaalista vastinetta. Potentiaalinen indeksointivastine on INFSER-aineistossa 88,5 prosentilla asiasanoista. Tulos siis vahvistaa TERVAS-aineiston pohjalta saatua käsitystä siitä, että tutkittavat asiasanastot ovat sanastoiltaan hyvin yhteensopivia. LUAS-indeksoinnin "kääntäminen" SPORT-indeksoinniksi on pääosin yksinkertaista, sillä vain noin 25 prosenttia kaikista LUAS-asiasanoista ja 14 prosenttia INFSER-aineiston konkreettisesti indeksointiin käytetyistä asiasanoista näyttää vaativan erityistoimenpiteitä, jotta niiden tarkoittamat käsitteet voitaisiin kääntää SPORTin asiasanoiksi.

Potentiaalisten indeksointiekvivalenssisuhteiden antama yleiskuva kertoo, että LUASin ja SPORTin yhteensopivuus asiasanatasolla on 75 prosenttinen. Vaikka tulos vielä tarkentuu, kun ekvivalenssisuhteiden analyysi syvenee, se kertoo jo, että mitään vakavia rakenteellisia tai käsitteellisiä esteitä näiden sanastojen rinnakkaiskäytölle ei asiasanatasolla ole. Mutta, vaikka kolme neljännestä LUAS-asiasanoista on käännettävissä SPORT-asiasanoiksi, joka neljännen asiasanan kääntäminen näyttäisi epäonnistuvan sopivan vastineen puuttuessa. Jatkotutkimukseni tavoitteena on selvittää kuinka todellista em. vastineettomuus on, minkälaisia käsitteitä ja/tai aiheita vastineettomat asiasanat kuvaavat ja miten niiden puuttuvat vastineet olisi mahdollista korvata SPORT-indeksointikielen keinoin.

Lopuksi

Yhteensopivuus on sisällönkuvailujärjestelmän ominaisuus, joka voidaan jakaa tiedonjäsennyksen, lingvistisen rakenteen, muodon ja tietojenkäsittelyn tason yhteensopivuuteen. Nämä tasot ovat erillisiä, mutta toisistaan riippuvaisia. Yhteensopivuutta voidaan empiirisen aineiston avulla tutkia kaikilla em. tasoilla sekä potentiaalisena yhteensopivuutena että konkreettisena, toteutuneena yhteensopivuutena. Tässä artikkelissa on kuvattu Liikunnan ja urheilun asiasanaston lingvistisen rakenteen yhteensopivuutta Sport Thesauruksen kanssa potentiaalisen indeksointiekvivalenssin avulla.

Hyväksytty julkaistavaksi 5.2.1996.

Lähteet

Buchan, R.L. (1989). Intertwining thesauri and dictionaries. Information Services & Use 9:171–175.

Dahlberg, Ingetraut (1983). Conceptual compatibility of ordering systems. International Classification 10(1):5–8.

Dahlberg, Ingetraut (1981). Towards establishment of compatibility between indexing languages. International Classification 8(2):86–91.

Hood, Martha W. & Ebermann, Christine (1990). Reconciling the CAB Thesaurus and AGRO-

VOC. Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists 35(4):181–185.

Hu, Qian (1992). On the implausibility of equivalent response. Part 1. Meta 37(2):289–301, 1992.

ISO 5964 (1985). Documentation – Guidelines for the establishment and development of multilingual thesauri.

Jakobson, Roman (1966). On linguistic aspects of translation. On translation (toim. Brower, R.A.), s. 232–238. New York.

Kalisz, Roman (1981). More on pragmatic equivalence. Linguistics across historical and geographical boundaries. Vol. 2: Descriptive, contrastive and applied linguistics. (toim. Kastowsky, D. & Szwedek, A.), s. 1247–1255. New York: Mouton de Gryuter.

Lancaster, F.W. (1986). Vocabulary control for information retrieval. 2nd ed. Arlington: Information Resources Press.

Liikunnan (1989). Liikunnan ja urheilun asiasanasto (toim. O. Meriläinen). Jyväskylä: Liikunnan ja kansanterveyden edistämissäätiö.

Liikunnan (1993). Liikunnan ja urheilun asiasanasto: uudet ja korjatut asiasanat: joulukuu 1990 – elokuu 1993. Jyväskylä: Liikunnan ja kansanterveyden edistämissäätiö.

Meriläinen, Outi (1993). Liikuntatieteellinen suomalais-englantilainen tiedonhakusanasto. Jyväskylä: Liikunnan ja kansanterveyden edistämissäätiö.

Mili Hafedh & Rada, Roy (1988). Merging thesauri: principles and evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(2):204–220.

Nida, E.A. & de Waard, J. (1986). From one language to another: Functional equivalence in Bible translating. Nashville.

Rada, Roy (1987). Connecting and evaluating thesauri: issues and cases. International Classification 14(2):63–69.

Sager, J.C., Somers, H.L. & McNaught, J. (1981). Thesaurus integration in the social science. Part I: Comparison of thesauri. International Classification 8(3):133–138.

Scibor, Eugeniusz & Tomasik-Beck, Joanna (1994). On the establishment of concordances between indexing languages having a universal or interdisciplinary scope (Polish experience). Knowledge Organization 21(4):213–223.

SFS 5471 (1988). Suomenkielisen tesauruksen laatimis- ja ylläpito-ohjeet = Guidelines for the establishment and maintenance of Finnish language thesauri. Helsinki: Suomen Standardoimisliitto SFS.

Snell-Hornby, Mary (1988). The role of text-linguistics in a theory of literary translation. Text-linguistik und Fachsprache: Akten des Internationalen Åbersetzungswissenchaftlichen AILA- Symposions Hildesheim 13.–16. April 1987 (toim. Arntz, R.), s. 433–448. Hildesheim: Georg Olms Verlag.

Sport (1990). Sport Thesaurus 1990 edition (toim. R. Stark & al.). Ottawa: Sport Information Resource Centre.

Sport (1992). Sport Thesaurus 1990 edition: Corrections and additions. November 1992 Update. Ottawa: Sport Information Resource Center.

Sport (1994). Sport Thesaurus 1994 edition (toim. R. Stark & al.). Ottawa: Sport Information Resource Centre.

Stark, Richard W. (1993). Database acquisition and integration in the field of sport and leisure: a case study. Paper presented in the 9th Scientific Congress "Sports information in the nineties" of the International Association for Sports Information Rome 7–10 June 1993. (Unpublished).

Svenonius, Elaine (1983). Compatibility of retrieval languages: Introduction to a forum. International Classification 10(1):2–4.

Tymoczko, Maria (1985). How distinct are formal and dynamic equivalence? The manipulation of Literature: Studies in Literary Translation (toim. Hermans, Theo), s. 63–86. New York: St. Martin's.

Viitteet

1. LISETI on Liikuntatieteellinen suomalais-englantilainen tiedonhakusanasto. Se on aakkosellinen luettelo niistä Liikunnan ja urheilun asiasanaston (LUAS ks. viite 2) asiasanoista, joille löytyy asiasanavastine Sport Thesauruksesta (SPORT ks. viite 3). LISETI valmistui vuonna 1993 ja sisältää 1394 LUAS-asiasanaa ja niiden SPORT-vastineet. Sanastossa on myös englanninkielisten asiasanojen mukaan aakkostettu hakemisto.

2. Liikunnan ja urheilun asiasanaston laatiminen aloitettiin 1984. Se saatiin koekäyttöön vuonna 1987 ja painettuna sanasto ilmestyi vuonna 1989. Syksyllä 1993 ilmestyi kolmas päivitystiedote, johon sisältyvät muutokset vuosilta 1990–1993 (Liikunnan 1993). Näiden muutosten jälkeen sanastossa on 1978 asiasanaa ja 184 synonyymia.
   Sanasto on tarkoitettu liikunta- ja urheiluaiheisten dokumenttien kuvailuun ja hakuun ja se on tesaurusmuotoinen. Se käsittää aakkosellisen pääosan ja aiheittain ryhmitellyn, 18-luokkaisen kategoriaosan sekä apuhakemistot henkilöryhmistä ja alkuaineista. Sen pääkäyttöalue on tällä hetkellä suomalaisen liikuntatietokannan Finsportin aineiston kuvailu ja haku ARTO- ja KATI-tietokantaan.
3. Sport Thesauruksen ensimmäinen versio ilmestyi vuonna 1981 ja viimeisin laitos on vuodelta 1994. Tutkimuksellisista syistä vertailussa käytetään vuoden 1990 laitosta ja sen lisäysosaa vuodelta 1992.
   Vuoden 1990 laitoksessa on lisäysosalla täydennettynä 6555 asiasanaa ja 1788 synonyymia järjestettynä aakkoselliseksi tesaurukseksi. Aakkosellisen tesaurusosan lisäksi sanastossa on maantieteellisten nimien ja yhteisönimien hakemisto sekä Sport Database Codes -luokitus, jonka koodeja voidaan käyttää asiasanojen lisänä sisällönkuvailussa ja tiedonhaussa. Sanastoa käytetään Sporttietokannan indeksointi- ja hakukielenä.
   SPORT-tietokanta on perustettu vuonna 1975 ja sitä ylläpitää kanadalainen Sport Information Research Center. Lähes 400000 viitettä sisältävänä SPORT on maailman laajin liikunta-aiheinen tietokanta. Sen vuosikartunta on noin 20000 viitettä, josta neljännes kerätään ja kuvaillaan hajautetusti eri maiden kansallisissa liikunnan ja urheilun tietopalvelukeskuksissa. Suomalaisten viitteiden käsittelystä vastaa LIKES-tietopalvelu Jyväskylässä.
4. Englanninkielisessä kirjallisuudessa käytetään useita eri termejä, jotka kaikki tarkoittavat sanastojen yhdistämistä. Eri termeillä haluttaneen korostaa yhdistämismenetelmien ominaispiirteitä.
5. SIRLS eli Specialized Information Retrieval and Library Service toimi University of Waterloossa 1970–1990. Se ylläpiti Sport & Leisure -tietokantaa, joka oli keskittynyt liikunnan sosiaalisiin ja psykososiaalisiin aspekteihin. Kun tietokannan viitteet päätettiin liittää SPORT-tietokantaan vuonna 1991, niitä oli yli 20000.