1861

Pirkko Tuulihovi

MONIKIELISET VERKKOJULKAISUT

Tietojärjestelmätieteen
pro gradu -tutkielma
9.7.1999

Jyväskylän yliopisto
Tietojenkäsittelytieteiden laitos
Informaatioteknologian maisteriohjelmat
Digitaalinen media

ABSTRACT

Tuulihovi, Pirkko Eliisa
Multilingual publications on the Web/ Pirkko Tuulihovi
Jyväskylä: University of Jyväskylä, 1999
74 s.
Master's thesis

The number of multilingual documents on the Web has increased together with the increase of the use of Internet. Many companies and public institutions use more and more intranets and extranets as a document publishing and delivering channel in their operation. Many documents are published in several languages as the operation becomes international, and the processing and management of them is a considerable problem. One solution to the management of multilingual web publications could be the structuring of their content.

First this thesis studies the international standards Unicode and Standard Generalized Markup Language (SGML). Unicode standard is fully compatible with the international standard ISO 10646-1:1993 and provides the capacity to encode all of the characters used for the written languages of the world. SGML and the specification languages XML, HTML and TEI based on it are briefly studied. Especially it has been examined how they could be applied to the management of multilingual documents published on the Web. Then the TIPSTER architecture and applications based on it are introduced. The result of this thesis is a mapping of the issues related to the management of multilingual web publications. Finally the applications based on TIPSTER have been evaluated shortly on the basis of this mapping.

KEYWORDS: Multilinguality, web publications, document management, SGML, Unicode, TIPSTER architecture

TIIVISTELMÄ

Tuulihovi, Pirkko Eliisa
Monikieliset verkkojulkaisut/ Pirkko Tuulihovi
Jyväskylä: Jyväskylän yliopisto, 1999
74 s.
Tutkielma

Internetin käytön laajentuessa on myös monikielisten dokumenttien määrä verkossa lisääntynyt. Yritykset ja laitokset käyttävät toiminnassaan dokumenttien julkaisu- ja jakelukanavana yhä enemmän intranet- ja extranetverkkoja. Toiminnan kansainvälistyessä monet dokumentit julkaistaan useilla kielillä ja niiden käsittely ja hallinta muodostavat merkittävän ongelman. Yksi ratkaisu monikielisten verkkojulkaisujen hallintaan saattaa olla niiden sisällön rakenteistaminen.

Tässä tutkielmassa tarkastellaan monikielisten tekstien koodaamiseen soveltuvaa Unicode-standardia, joka on täysin yhtenevä kansainvälisen standardin ISO 10646-1:1993 kanssa. Unicoden kapasiteetti riittää kaikkien kirjoitetussa muodossa olevien kielten merkistöjen koodaamiseen yksikäsitteisesti. Lisäksi tarkastellaan lyhyesti dokumenttien rakenteistamisessa käytettävää Standard Generalized Markup Language-metakieltä (SGML) ja siihen pohjautuvia määrittely- ja merkkauskieliä XML, HTML ja TEI. Erityisesti tutkielmassa selvitetään niiden soveltuvuutta monikielisten verkkojulkaisujen hallintaan. Sen jälkeen esitellään TIPSTER–arkkitehtuuri ja sen pohjalta kehitettyjä sovelluksia. Tutkielman tuloksena on monikielisten verkkojulkaisujen hallinnan yleiskuvaus, jonka pohjalta on arvioitu TIPSTER:iin pohjautuvia sovelluksia.

AVAINSANAT: Monikielisyys, verkkojulkaisut, dokumenttien hallinta, SGML, Unicode, TIPSTER-arkkitehtuuri

SISÄLLYS

1. JOHDANTO

Internetin käytön lisääntyessä on myös verkon kautta tapahtuva tiedonvälitys kasvanut voimakkaasti ja kehityksen uskotaan yhä jatkuvan. Tämän johdosta verkossa tulee olemaan paljon erikielisiä dokumentteja. Oudetin (1997) mukaan on tosin myös ennustettu, että englanti tulee hallitsemaan ja muiden kielten käyttö verkossa katoaa. Sellaisella suuntauksella on kuitenkin omat haittapuolensa: valtaosalle ihmisistä englanti on toinen kieli, jonka käyttö riittää perustietojen ymmärtämiseen. Sen sijaan syvemmälle menevässä keskustelussa useimmat käyttävät mieluiten omaa äidinkieltään, kuten Oudet (1997) artikkelissaan toteaa. Jotta Internet palvelisi tehokkaasti kansainvälisen tiedonvälityksen kanavana, täytyy jokaisella verkon käyttäjällä olla mahdollisuus hakea ja käyttää siellä olevaa materiaalia omalla äidinkielellään. (Oudet, 1997; ks. myös Goldfarb & Prescod, 1998, s. 143).

1.1. Tutkimuksen tausta

Monikielisellä dokumentilla tarkoitetaan tässä tutkielmassa tekstidokumenttia, joka on saatavissa vähintään kahtena erikielisenä versiona tai dokumenttia, joka sisältää erikielisiä osia. *Monikielisellä verkkojulkaisulla* tarkoitetaan puolestaan Internetin, extranetin tai intranetin välityksellä julkaistavaa tai jaettavaa monikielistä dokumenttia.

Monikielisiä, digitaalisessa muodossa olevia dokumentteja tarvitaan useilla eri sovellusalueilla. Esimerkkejä sellaisista ovat muun muassa:

- koneiden käyttö- ja huolto-ohjeet, joista useinkin tarvitaan monta kieliversiota
- verkossa julkaistava etäopetusmateriaali
- julkiset asiakirjat, jotka on tarkoitettu kaikkien kansalaisten saataville heidän omalla äidinkielellään (esimerkiksi EU-asiakirjat)
- yritysten tai tuotteiden esittely ja mainostaminen verkossa
- tietojärjestelmien ja tietokoneohjelmistojen lokalisointi.

Markkinoiden kansainvälistyminen merkitsee sitä, että tuotteiden jakelu tapahtuu paikallisesti. Silloin tuotteiden täytyy vastata kohdemarkkinoiden kielellisiä, kulttuurisia

ja lakisääteisiä vaatimuksia. Lokalisaatio voi koskea tuotetta itseään, kuten esimerkiksi ohjelmistotuotteita, joiden käyttöliittymiin liittyy kielellistä informaatiota. Erityisesti lokalisaatio liittyy tuotteiden mukana tulevaan dokumentaatioon, joka on tarkoitettu helpottamaan niiden käyttöä ja ylläpitoa. (Hartley & Paris, 1997).

Monikielisten tekstien käsittely aiheuttaa monenlaisia ja monentasoisia ongelmia näppäimistöltä syötössä, koodauksessa, muokkauksessa, tulostuksessa sekä tiedonsiirrossa. Boualemin ja Harién (1997) mukaan siinä esiintyy esimerkiksi seuraavanlaisia ongelmia:

- Monet näppäimistöt esittävät vain ISO 14962:1997 –merkistön, joka on ASCII-koodattua englantia (ISO, 1999a) tai ISO 646 -merkistön, joka on informaation siirtoon tarkoitettu 7-bittinen merkkijoukko (ISO, 1999b). Tällöin jotkut kirjaimet vaativat kahden näppäimen painallusta (kuten ï ja ê); joidenkin kirjainten esittäminen ei tietyillä näppäimistöillä ole ollenkaan mahdollista (muun muassa ñ-kirjainta ei voida esittää ranskalaisella näppäimistöllä).
- Esimerkiksi kiinan kielessä on yli 6000 ideogrammia ja arabiassa arviolta neljä 28 kirjaimen joukkoa ja 10 vokaalia, joten niiden mukaan ottaminen vaatii erityisten syöttöohjelmien määrittelemistä.
- Tietokoneiden valmistajat ja ohjelmistojen kehittäjät käyttävät useita erityisiä ja yhteen sopimattomia merkkikoodeja.
- Kielten tunnistaminen useita erikielisiä osia sisältävästä tekstistä on monesti vaikeaa.
- Erilaiset kielten kirjoitussuunnat, kuten oikealta vasemmalle (arabia ja heprea) tai ylhäältä alas (kiina ja japani), aiheuttavat ongelmia varsinkin tekstin editoinnissa.
- Tulostimien fonttien vähyys aiheuttaa usein tulostusongelmia.

Monikielisiin teksteihin liittyy myös monia muita ongelmia, kuten päivämäärien ilmaiseminen, kiinteät sanonnat (esimerkiksi latinalaiset de facto, ad hoc jne.), vieraskieliset sanat lauseiden keskellä, maantieteelliset nimet, lyhenteet ja kirjainsanat eli akronyymit sekä koodit ja kaavat (Jaeger, Devillers, Cruickshank & Cencioni, 1991). Yleensäkin nimien erottaminen muista substantiiveista, esimerkiksi historiallisissa ja vanhoissa kaunokirjallisissa teksteissä, on usein ongelmallista. Niiden löytäminen on kuitenkin tärkeää, koska nimet ovat tutkijoiden kiinnostuksen kohde ja monesti heidän tutkimuksensa lähtökohta (Flanders, Bauman, Caton & Cournane, 1998).

Tiedonhaku monikielisessä ympäristössä asettaa myös dokumenttien hallinnalle erityisvaatimuksia. Kyselylausekkeet on voitava muotoilla niin, että haun tulos sisältää mahdollisimman kattavasti juuri sitä tietoa, mitä käyttäjä haki. Tällöin on kiinnitettävä erityistä huomiota käytettyyn terminologiaan ja tarjottava apuvälineitä käyttäjälle oikeiden hakutermien valitsemisessa. Tällaisia voivat olla esimerkiksi monikieliset termipankit.

1.2. Tutkimusongelma ja tutkimusalueen rajaus

Tässä työssä tutkimusongelmana on: Miten hallitaan monikielisiä verkkojulkaisuja? Osaongelmia ovat muun muassa:

- Miten monikielisten verkkojulkaisut ja niiden metatiedot organisoidaan?
- Miten monikielisiä hakuja suoritetaan?
- Miten monikielisiä verkkojulkaisuja ylläpidetään?
- Mikä on monikielisten verkkojulkaisujen käytettävyys?

Tässä tutkielmassa käsitellään Internetin, intranetin tai extranetin kautta julkaistavia tai jaeltavia monikielisiä verkkojulkaisuja, jollaisia esimerkiksi yritykset tai julkishallinnon laitokset julkaisevat omaan sisäiseen tai sidosryhmiensä käyttöön. Sellaisia dokumentteja ovat muun muassa tuotteiden mainokset ja esitteet sekä käyttö- ja huolto-ohjeet, joiden saaminen asiakasyritysten työntekijöiden äidinkielellä on välttämätöntä ja joita yhä enemmän julkaistaan digitaalisesti ja verkkojen kautta.

Vaikka tutkielmassa käsitelläänkin kaikkien kielien koodaamiseen tarkoitettua Unicode-merkistöä tekstien koodausjärjestelmänä, rajaudutaan tässä tutkielmassa lähinnä kieliin, joita käytetään yleisesti Euroopassa (indoeurooppalaiset ja suomalais-ugrilaiset kielet) ja Amerikassa. Muun muassa Aasiassa ja Afrikassa puhutut kielet samoin kuin venäjä jätetään tarkastelun ulkopuolelle lyhyitä esimerkinomaisia mainintoja lukuun ottamatta. Tutkielmassa keskitytään dokumentteihin, joiden varsinainen sisältö on tekstimuotoista. Tekstidokumenteissa saattaa olla kuvia ja ääntä mukana, mutta niissä esiintyvään kieleen ei kiinnitetä huomiota. Dokumenttien tallennustapojen tarkastelu rajataan myös tämän tutkielman ulkopuolelle, koska tutkielmassa on näkökulmana monikielisten

verkkojulkaisujen esittäminen verkon käyttäjille. Tutkielman tavoitteena on kartoittaa monikielisten dokumenttien hallintaan liittyviä asioita eli kuvata ongelma-alue yleisesti.

1.3. Tutkielman sisältö

Luvussa kaksi esitellään aluksi Unicode-standardi, joka mahdollistaa kaikkien kirjoitettujen kielten koodaamisen digitaaliseen muotoon. Sitten käsitellään SGML-metakieltä ja siihen pohjautuvia määrittelykieliä: XML:ää, HTML:ää ja TEI:tä. Niistä esitellään ensin peruskäsitteet lyhyesti. Sen jälkeen tarkemmalla tasolla tutkitaan, mitkä niiden piirteet tukevat monikielisyyden toteuttamista verkkojulkaisuissa.

Luvussa kolme esitellään ja arvioidaan TIPSTER-arkkitehtuuria ja sen pohjalta rakennettuja sovelluksia sekä niiden soveltuvuutta monikielisten verkkojulkaisujen hallintaan. Luvussa neljä kuvataan monikielisten verkkojulkaisujen hallintaan liittyviä osa-alueita ja ongelmakohtia. Luvussa on myös arvioitu lyhyesti TIPSTER-arkkitehtuuriin pohjautuvia sovelluksia. Luvussa viisi on tutkimuksen yhteenveto.

2. UNICODE- JA SGML-STANDARDI

Tässä luvussa tarkastellaan ensin Unicode-standardia, joka on ISO/IEC 10646-1 – standardin kaupallinen vastine. Se tarjoaa mahdollisuuden käytännössä kaikkien kielten koodaamiseen. Toinen luvussa tarkasteltava standardi on ISO 8879 eli Standard Generalized Markup Language (SGML), joka on tarkoitettu tekstien loogisen rakenteen merkkaamiseen. Lisäksi luvussa käsitellään SGML:n sovelluksia HTML, XML ja TEI. Näkökulmana tarkastelussa on dokumenttien monikielisyyden toteuttaminen.

2.1. Unicode-standardi

Unicode tarjoaa yhdenmukaisen tavan tekstin koodaamiseen ja helpottaa tekstitiedostojen kansainvälistä vaihtamista. Unicode helpottaa myös monikielisten tekstien käsittelyä, koska sen kapasiteetti riittää kaikkien tunnettujen merkistöjen koodaamiseen. Unicode-standardista on huomattavasti apua myös matemaattisten ja teknisten dokumenttien koodaamisessa, koska ne sisältävät usein erikoissymboleja. (The Unicode® Standard, 1998).

ISO/IEC 10646 -standardi määrittelee *UCS:n (Universal Multiple-Octet Coded Character Set)*, joka soveltuu tekstin koodaamiseen tietokoneella tapahtuvaa käsittelyä varten. *Unicode-standardi* on UCS:n kaupallinen vastine ja täysin yhteensopiva kansainvälisen standardin ISO/IEC 10646-1 kanssa, jossa on määritelty *BMP (Basic Multilingual Plane)* (ISO, 1999c). BMP on nimitys ISO 10646 -standardissa määritellyille ensimmäiselle 65 535 merkille. (Johnson, 1997; Noerr, 1998; The Unicode® Standard, 1998). SGML:n pohjalta kehitetyssä XML-määrityksessä on sitouduttu UCS-merkistön käyttöön (Bray, Paoli & Sperberg-McQueen, 1998), ja Unicode-standardi on myös HTML 4.0 –version kehityksen peruslähtökohtana (Raggett, Le Hors & Jacobs, 1998).

Unicoden suunnittelu perustuu *ASCII*:n *(American Standard Code for Information Interchange)* yksinkertaisuudelle ja johdonmukaisuudelle, mutta se ulottuu paljon laajemmalle kuin ASCII, joka rajoittuu ainoastaan latinalaisen aakkoston koodaamiseen. ASCII-standardi perustui 7-bittiseen koodaukseen eli sillä voitiin koodata 128 merkkiä.

Se oli riittävä englanninkieliselle merkistölle ja siihen sisältyi myös välimerkkejä, joitakin ohjauskoodeja sekä numerot. (Davis, 1994; Noerr, 1998).

Unicode-standardin tavoitteena on, että sillä voidaan koodata kaikkien kirjoitettujen kielten merkistöt. Jokainen tekstin pieninkin elementti on voitava koodata ja jokaiselle elementille annetaan yksikäsitteinen koodi ja myös nimi. Lisäksi Unicode tarjoaa perussäännöt tekstin koodaamiselle ja tulkinnalle niin, että ohjelmat kykenevät lukemaan ja käsittelemään tekstiä. (The Unicode® Standard, 1998).

Unicode käyttää 16-bittistä koodausta, joka mahdollistaa yli 65 000 merkin esittämisen. Se ei käytä koodauksessa monimutkaisia tapoja eikä myöskään kooditaulun vaihtokomentoja (*escape codes*), vaan osoittaa jokaiselle merkille yksikäsitteisen 16 bitin arvon, jolloin merkkien koodaaminen pysyy yksinkertaisena ja tehokkaana. Merkkien moninkertainen koodaaminen on vältetty yhtenäistämällä ne eri kielissä eli muodoltaan samanlaisille merkeille on annettu yksi koodi. Esimerkki sellaisesta on *CJK (Chinese/Japanese/Korean) -konsolidaatio* eli *-yhdistäminen*, joka on saatu aikaan osoittamalla yksittäinen koodi jokaiselle ideograafille, joka on tavallinen useammassa kuin yhdessä näistä kielistä. (The Unicode® Standard, 1998).

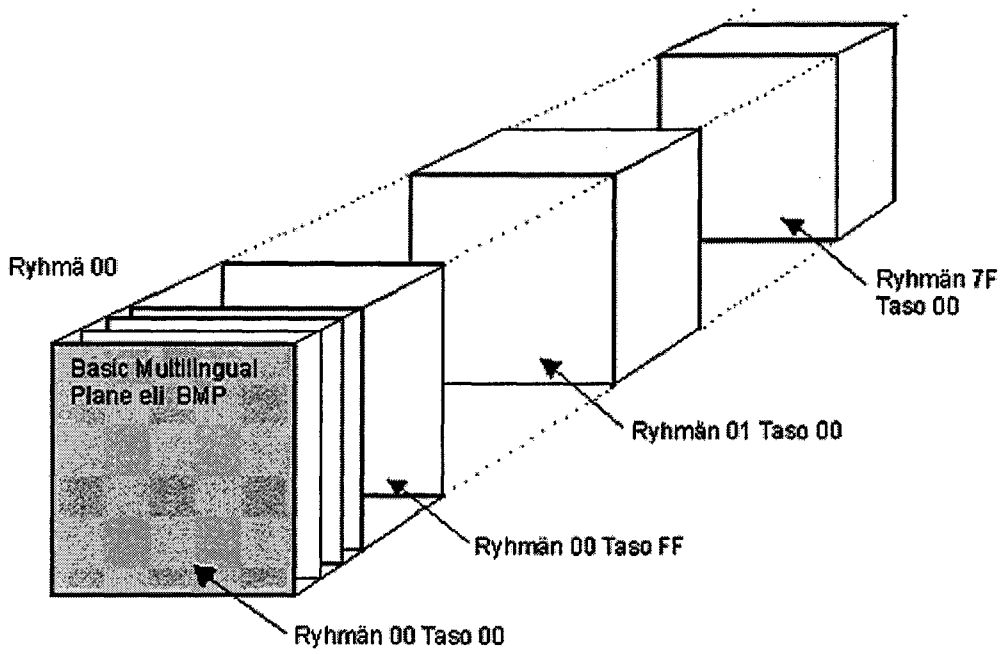
Unicoden perusperiaatteet voidaan tiivistää seuraavaan kolmeen kohtaan. Se on

1. **yleismaailmallinen** eli koodi kattaa kaikki nykyiset kirjoitetut kielet
2. **yksiselitteinen** eli jokaiselle merkille on täsmällisesti vain yksi koodaustapa
3. **yhtenäinen** eli jokainen merkki esitetään kiinteällä määrällä bittejä.

(Bettels & Bishop, 1993).

2.1.1. UCS-koodausjärjestelmä

Unicoden perustana oleva UCS on rakenteeltaan neliulotteinen *koodausavaruus (coding space)*, joka on 128 kolmiulotteisesta ryhmästä koostuva kokonaisuus. Jokainen ryhmä koostuu 256 kaksiulotteisesta tasosta, joissa on 256 riviä ja jokaisella rivillä 256 solua. Koodattu merkki sijaitsee koodausavaruudessa olevassa solussa tai sitten solun kohdalla ilmoitetaan, ettei sitä ole käytetty. (Ogg, 1996). Kuviossa 1 on esitetty UCS:n yleinen rakenne.



KUVIO 1. UCS:n yleinen rakenne (Ogg, 1996).

Jokainen merkki esitetään neljänä kahdeksan bitin joukkona eli oktettina (octets), jotka määrittävät ryhmän, tason, rivin ja solun. Tätä muotoa kutsutaan UCS-4:ksi. BMP (Basic Multilingual Plane) on kahdella oktetilla koodattu merkkijoukko, jota kutsutaan UCS-2:ksi ja joka on Unicoden kanssa ekvivalentti. (Ogg, 1996).

2.1.2. Muuntamiskaavat UTF-8 ja UTF-16

Merkkien koodausstandardit määrittelevät jokaisen merkin identiteetin ja numeerisen arvon sekä sen, kuinka tämä arvo esitetään bitteinä. Unicode-standardi hyväksyy kaksi koodausmuotoa, jotka vastaavat ISO 10646 -standardin muuntamiskaavoja *UTF-8 (UCS Transformation Format)* ja *UTF-16*. (The Unicode® Standard, 1998).

UTF-8:lla voidaan muuntaa kaikki Unicode-merkit vaihtelevanmittaisiksi tavukoodeiksi. UTF-8:n etuna on se, että tuttuja ASCII-merkkejä vastaavilla Unicode-merkeillä on samat arvot. Näiden merkkien koodit sijaitsevat BMP:n ensimmäisen rivin

ensimmäisellä puoliskolla (Järnefors, 1996). UTF-8:aan muunnettuja Unicode-merkkejä voidaan siis käyttää monissa olemassa olevissa ohjelmistoissa ilman, että ohjelmistoja tarvitsee paljonkaan päivittää. Unicode Consortium hyväksyy UTF-8-muuntamiskaavan käytön Unicode-standardin toteuttamisessa. Mikä tahansa Unicode-merkki, joka on ilmaistu 16-bittisellä UTF-16-muodolla, voidaan konvertoida UTF-8-muotoon ja takaisin ilman informaation hukkaamista. (The Unicode® Standard, 1998).

UTF-16:ssa oletuksena ovat 16-bittiset merkit ja siinä käytetään tiettyä valikoimaa merkkejä laajennusmekanismina, jolla voidaan koodata lisää miljoona merkkiä käyttämällä 16-bittisiä merkkipareja. Tämä on riittävä määrä kaikille koodausvaatimuksille, mukaan lukien kaikki historialliset käsikirjoitukset maailmassa. Unicode-standardilla voidaan myös esittää välimerkkejä, matemaattisia ja muita teknisiä merkkejä sekä sellaisia perusmerkkien muunteluun tarkoitettuja merkkejä kuten ~ (tilde), jonka avulla esitetään esimerkiksi kirjain ñ. Unicode-standardin versioon 2.0 on otettu UTF-16-muuntamiskaava sellaisena kuin se on määritelty ISO/IEC 10646-standardissa. (The Unicode® Standard, 1998).

2.2. SGML

SGML (Standard Generalized Markup Language) on tekstidokumenttien merkkaukseen kehitetty metakieli. SGML on ISO:n (International Organization for Standardization) vuonna 1986 hyväksymä standardi 8879. Keskeistä SGML-merkkauksessa on dokumentin esitysmuodon ja sisällön erottaminen toisistaan sekä pyrkimys laite- ja sovellusriippumattomuuteen (Goldfarb, 1990). Seuraava SGML:n peruskäsitteiden esittely perustuu, ellei toisin ole mainittu, Goldfarbin teokseen The SGML Handbook (1990).

Peruskäsitteet

Dokumentilla tarkoitetaan SGML:ssä loogista rakennelmaa, joka sisältää *dokumenttielementin*. Dokumenttielementti on ylin solmu dokumentin *sisällön* muodostavien elementtien puurakenteessa.

SGML-dokumentti koostuu seuraavista osista:

- *SGML-esittelyosa (SGML Declaration)*, jossa ilmoitetaan muun muassa käytettävä merkistö ja merkkauksessa käytettävät merkit. Esittelyosa voi puuttua ja silloin käytetään SGML-standardin määrittelemiä oletusarvoja.
- *Dokumenttityyppimäärittely* eli *dokumentin rakennemäärittely (Document Type Definition, DTD).*
- Dokumentin sisältö merkatussa muodossa. Sisältää varsinaisen tekstin sekä dokumentin merkkauksen ja voi esiintyä vain DTD:n jälkeen.

Dokumenttityyppimäärittely (Document Type Definition, DTD) määrittää

- kyseessä olevassa dokumenttityypissä sallittujen elementtityyppien *tunnisteet (generic identifiers, GI)*
- jokaiselle elementtityypille mahdolliset *attribuutit*, niiden arvojen vaihteluvälit sekä oletusarvot
- jokaiselle elementtityypille *elementtien* sisällön rakenteen.

SGML:n peruskomponentteja ovat elementit ja *entiteetit.* Elementit merkataan *aloitus- (start-tag)* ja *lopetustunnisteilla (end-tag),* jotka tavallisesti ilmaistaan < >- ja </>-merkinnöillä, mutta niiden tilalla voidaan käyttää muitakin merkkejä. Entiteetit ovat mitä tahansa tekstin nimettyjä osia. Niitä on kahta tyyppiä: *yleiset entiteettiviitteet (general entity references)* ja *parametrientiteettiviitteet (parameter entity references).* Yleinen entiteettiviite alkaa &-merkillä ja parametrientiteettiviite puolestaan %-merkillä. Entiteettiviitteet päättyvät ;-merkkiin. Entiteettien avulla voidaan SGML-dokumentissa ilmaista muitakin merkkejä kuin DTD:ssä määritellyssä merkistössä on, mikä on tärkeä ominaisuus monikielisyyden kannalta. Jos DTD:n mukaan käytetään esimerkiksi vain englantilaisen aakkoston kirjaimia, niin entiteettien avulla voidaan ilmaista myös vaikkapa å-, ä- ja ö-kirjaimet å-, ä ja ö-entiteeteillä.

Yksittäisen dokumentin merkkaustunnisteet kuvaavat sen elementtien muodostaman rakenteen. Toisin sanoen merkkaustunnisteet osoittavat, mitkä elementit esiintyvät dokumentin sisällössä ja mikä on niiden järjestys. Esimerkissä 1 on merkattu SGML:llä Keski-Suomen Viro-seura Ry:n eestinkielinen kotisivu (Keski-Suomen Viro-seura ry, 1997).

Esimerkki 1. Hierarkkisen rakenteen merkkaaminen SGML:llä.

```
<KOTISIVU kieli="eesti">
  <YHDISTYS>Kesk-Soome Eesti Selts</YHDISTYS>

  <PERUSTETTU>
  Kesk-Soome Eesti Selts asutati
  <PERPVM>29. m&#228;rtsil 1995</PERPVM>
  </PERUSTETTU>

  <TOIMINTA>
    <TOIM-AJATUS>Seltsi eesm&#228;rgiks on v&#245;imalda eesti
    keele &#245;ppimist ja edistada organisatsioonide ja inimeste vahelisi
    kontakte  Eesti ja Kesk-Soome vahel.
    </TOIM-AJATUS>

    <TOIM-TAVAT>
    Selleks otstarbeks korraldab Selts kultuuri-, koolitus- ja
    informatsioonikoosolekuid, seminare ja ekskursioone, ja
    v&#245;ib ka v&#228;lja  anda b&#252;llet&#228;&#228;ne
    v&#245;i muid  tr&#252;kiseid.
    </TOIM-TAVAT>
  </TOIMINTA>

  <HALLITUS>
    <HALL-OTS>Kesk-Soome Eesti Seltsi Juhatus 1997</HALL-OTS>
    <HALL-AS-JASENET>
      <ASEMA>Esimees</ASEMA>:
      <NIMI>Vesa Olkkonen</NIMI>.
      <ASEMA>Abiesimees</ASEMA>:
      <NIMI>Pekka Lilja</NIMI>.
      <ASEMA>Sekret&#228;r</ASEMA>:
      <NIMI>Olli J. Ojanen</NIMI>.
      <ASEMA>Laekur</ASEMA>:
      <NIMI>Sirkka Varpula</NIMI>.
    </HALL-AS-JASENET>

    <MUUT-OTS>Muud juhatuse liikmed:</MUUT-OTS>
    <HALL-MUUT-JASENET>
      <NIMI>Heikki Pentik&#228;inen</NIMI>,
      <NIMI>Andres Perendi</NIMI>,
      <NIMI>Kauko Saarinen</NIMI>,
      <NIMI>Rauno V&#228;lim&#228;ki</NIMI>.
    </HALL-MUUT-JASENET>

  </HALLITUS>
...
  <PAIVITYS>
    <PAIV-OTS>Viimati ajakohastatud</PAIV-OTS>
    <PAIV-PVM>17.1.1997</PAIV-PVM>
```

```
    <PAIV-EROTIN>/</PAIV-EROTIN>
    <PAIVITTAJA>ks.</PAIVITTAJA>
  </PAIVITYS>

</KOTISIVU>
```

Elementteihin voidaan liittää attribuutteja, joissa on lisäinformaatiota elementtien ominaisuuksista. Attribuuttien on oltava yksikäsitteisiä ja ne määritellään erityisessä *attribuuttien määrittelylistassa (attribute definition list).* Jokaisessa attribuutin määrittelyssä on attribuutin nimi, sen sallitut arvot ja oletusarvo. Attribuuttien avulla voidaan esimerkiksi määritellä elementeissä käytetyt kielet ja mitä kieltä käytetään oletusarvona, ks. esimerkki 2.

Esimerkki 2. Attribuutin ja sen oletusarvojen määrittely.

```
<!ATTLIST kappale    kieli (su | ru | sa | ra | eng)   su >
```

Esimerkissä on elementin kappale attribuutin kieli määrittely. Sallitut arvot ovat su, ru, sa, ra ja eng; su on oletusarvo. Elementti kappale voi siis esiintyä viidellä eri kielellä, oletusarvoisesti se on suomenkielinen. Attribuutti kieli voitaisiin määritellä myös parametrientiteetin % KIELI avulla (ks. esimerkki 3). Tästä on se hyöty, että attribuuttien arvoja ei tarvitse listata erikseen jokaiselle elementille. Riittää, kun ne määritellään kerran parametrientiteetille, jota sitten käytetään kaikkien niiden elementtien attribuuttien määrittelyssä, jotka käyttävät samaa kielivalikoimaa.

Esimerkki 3. Parametrientiteetin käyttö attribuutin määrittelyssä.

```
<!ENTITY  % KIELI ("SU","RU","SA","RA","ENG") "SU">
                    <!ATTLIST kappale  kieli  (%KIELI)  "SU" >
```

SGML-standardi on suunniteltu laite- ja sovellusriippumattomaksi ja se on myös täysin riippumaton mistään tietystä kielestä. Merkkauksessa käytettävät elementtityyppien tunnisteet ja attribuuttien nimet sekä muut nimet voivat sisältää erityisiä kansallisia merkkejä. Käyttäjä voi nimetä ne haluamallaan tavalla elementtien ja entiteettien määrittelyssä. Standardin määrittelyn mukaan monikielisissä dokumenteissa käytettäviä monia erilaisia merkistöjä tuetaan. Käytettävän kielen ja sen merkistön on luonnollisesti sovittava yhteen käytössä olevien tekstinkäsittelylaitteiden ja -järjestelmien kanssa. (Goldfarb, 1990, s. 242-243).

SGML on tarkoitettu sopivaksi rajapinnaksi tiedon syöttöön ja siirtämiseen ilman esikäsittelyä. Se on mahdollista laajalti mukauttaa vastaamaan erilaisten näppäimistöjen ja näyttöjen asettamia vaatimuksia. (Goldfarb 1990, s. 243). SGML-dokumentin esittelyosassa voidaan määritellä käytettävä merkistö, muutoin käytetään standardin oletusarvoja. Haluttaessa tukea mahdollisimman monia kieliä määritellään merkistöksi ISO 10646 eli Unicode (Travis & Waldt, 1995).

Monikielisyyden toteuttaminen SGML-dokumenteissa

Monikielisyys voidaan toteuttaa SGML:llä seuraavilla tavoilla:

1. Eri kielet ovat eri SGML-dokumenteissa, jolloin kussakin kieliversiossa käytetty kieli ilmaistaan DTD:ssä dokumenttielementin attribuutin avulla.
2. Dokumentin erikieliset osat voidaan merkata SGML:llä omiksi elementeikseen (ks. esimerkki 4).
3. Dokumentissa voi olla toistuvia elementtejä, joissa kieli ilmaistaan attribuutilla (ks. esimerkki 5).
4. Dokumentti voi sisältää rinnakkain ja sisäkkäin olevia erikielisiä elementtejä, joiden kielivalikoimat ilmaistaan attribuuttien avulla.

Kohdan 1 mukainen tapa sopii hyvin silloin, kun koko dokumentti on julkaistu useilla eri kielillä. Dokumentin sisältö eri kieliversioissa on siis pääosin yksikielinen, mutta siinä voi kuitenkin olla pienempiä erikielisiä osia, lähinnä sanoja ja sanontoja, joita käytetään vakiintuneesti alkuperäisellä kielellä kääntämättä niitä. Tällaisia ovat esimerkiksi sellaiset latinankieliset sanonnat kuten de facto ja force majeure. Sellaiset

osat voidaan merkata tekstiin omina elementteinään, joille on määritelty oma kieliattribuutti.

Jos dokumentti sisältää joitakin osia useammalla kuin yhdellä kielellä, esimerkiksi jotkut tietyt kappaleet tekstissä, kuten tiivistelmä tai johdanto, kannattaa niiden erikieliset osat merkata omiksi elementeikseen (ks. esimerkki 4), joille DTD:ssä on määritelty attribuuttina kieli. Tällöin ne löytyvät merkatusta tekstistä helposti. Hyvin lyhyet erikieliset tekstin osat kannattaa merkata tällä tavalla omiksi elementeikseen. Erikieliset osat voidaan myös merkata toistuviksi elementeiksi, jolloin niissä käytetyt kielet ilmaistaan attribuuttien avulla (esimerkki 5).

Esimerkki 4. Erikielisten tekstin osien merkkaaminen omiksi elementeikseen.

```
...
<ABSTRACT>
...
The number of multilingual documents in the web has increased together
with the the increase of the use of Internet. Many companies and public
institutions use more and more intranets and extranets as a document
publishing and delivering channel in their operation. ...

</ABSTRACT>

<TIIVISTELMÄ>
...
Internetin käytön laajentuessa on myös monikielisten dokumenttien
määrä verkossa lisääntynyt. Yritykset ja laitokset käyttävät toiminnassaan
dokumenttien julkaisu- ja jakelukanavana ...
</TIIVISTELMÄ>
...
```

Seuraavat esimerkit ovat (hieman mukailtuina) peräisin RASKE-projektin raportista (Lyytikäinen, Päivärinta, Salminen & Tiitinen, 1997), missä on kuvattu valtion talousarvioon liittyvien asiakirjojen rakenteistamista. Näissä asiakirjoissa tarvitaan maamme molempia virallisia kieliä, niin kuin esimerkistä 5 käy ilmi.

Esimerkki 5. Kieli-attribuutin käyttö SGML-merkatussa tekstissä, jossa on toistuvia elementtejä.

```
<yleisper>
        <otsikko kieli="suomi">YLEISPERUSTELUT</otsikko>
        <otsikko kieli="ruotsi">ALLMÄN MOTIVERING</otsikko>
</yleisper>
...
<nimi kieli="suomi">EDUSKUNTA</nimi>
...
```

Esimerkissä näkyy kieli-attribuutin käyttötapa RASKE-projektin raportin SGML-merkatussa esimerkkitekstissä. RASKE-projektissa valittiin tämä tapa, koska tällöin ei tarvinnut kirjoittaa uudestaan sellaisia rakenneosia, jotka sisälsivät pelkästään kuvia tai taulukoita (Lyytikäinen ym., 1997, s. 65). Lyytikäisen ym. (1997) mukaan ratkaisu on myös joustava, koska asiakirjaan voidaan myöhemmin lisätä uusi kieli tarvitsematta muuttaa rakennemäärittelyä. Sama voitaisiin toteuttaa myös määrittämällä omat elementit erikielisille otsikoille, ks. esimerkki 6:

Esimerkki 6. Omien elementtien merkkaaminen eri kielille.

```
<yleisper>
        <s-otsikko>YLEISPERUSTELUT</s-otsikko>
        <r-otsikko>ALLMÄN MOTIVERING</r-otsikko>
</yleisper>
...
```

2.3. HTML versio 4.0

HTML 4.0 -määritys vuodelta 1998 laajentaa monin tavoin HTML:n aikaisempien versioiden ominaisuuksia. Se on kehitetty tukemaan monikielisten www-sivujen toteutusta. Yksi hyvin tärkeä askel on ollut ISO/IEC 10646 -standardin eli Unicoden

ottaminen HTML-dokumenttien merkkijoukoksi, jolloin HTML soveltuu entistä paremmin erikielisten verkkojulkaisujen toteuttamiseen.

Muun muassa seuraavat seikat, jotka ovat tärkeitä monikielisyyden ja hakujen toteuttamisen kannalta, ovat olleet HTML 4.0:n kehittämisen pohjana:

- Dokumenttien rakenteen ja esittämisen erottaminen toisistaan suosimalla tyylitiedostojen (style sheets) käyttöä sen sijaan, että käytettäisiin HTML:n elementtejä ja attribuutteja.
- Mahdollisuus merkata tekstin kuvaus dokumenttiin liitetylle objektille (käyttäen OBJECT-elementtiä).
- Vaihtoehtoisen tekstin liittäminen kuviin (images) ja kuvakarttoihin (image maps) sekä tuki pistekirjoituksen (braille) käyttämiselle.
- Lang- ja title-attribuuttien käytön tuki kaikille elementeille.
- Dir-attribuutin käyttö länsimaisesta lukusuunnasta poikkeavaan suuntaan luettaville kielille (esimerkiksi heprea ja arabia).
- ABBR- ja ACRONYM-elementtien tuki.
- Taulukoiden, kuvien, kehysten jne. pitkät kuvaukset.

(Raggett ym., 1998).

Seuraavassa esitellään tarkemmin edellä mainittuja ominaisuuksia sekä SPAN- ja BDO-elementit ja annetaan esimerkkejä niiden käytöstä. Esimerkit on pääosin peräisin HTML 4.0-määrityksestä (Raggett ym., 1998) jonkin verran mukailtuina. Ne ovat vain kuvitteellisia, koska selainten nykyiset versiot eivät tue HTML 4.0:aa.

OBJECT-elementillä voidaan liittää www-sivuun useita vaihtoehtoisia esityksiä samaa asiaa sisältävistä sivuista. Esimerkiksi sivulla voi olla upotettuna samaan asiaan liittyen vaihtoehtoisesti linkki videopätkän tai kuvan sisältämään tiedostoon tai tekstimuotoisen kuvauksen sisältävään sivuun. Samalla tavalla OBJECT-elementillä voitaisiin myös linkittää verkkojulkaisun kieliversioita. Esimerkissä 7 on sisäkkäisiä samansisältöisiä kieliversioita, joihin viitataan lähtien englanninkielisestä sivusta. Jos sille sivulle ei päästä, viitataan seuraavaksi suomenkieliseen sivuun jne.

Esimerkki 7. OBJECT-elementillä toteutetut sisäkkäiset kieliversiot.

```
<P> <!-- Ensin yritetään englanninkielistä sivua -->
    <OBJECT title="Englanninkielinen sivu">
            classid="http://www.jyu.fi/~tuulihov/engl.html"
      <!-- Toiseksi yritetään suomenkielistä sivua -->
      <OBJECT title="Suomenkielinen sivu">
              classid="http://www.jyu.fi/~tuulihov/fin.html"
        <!-- Kolmanneksiksi yritetään vironkielistä sivua -->
        <OBJECT title="Vironkielinen sivu">
                classid="http://www.jyu.fi/~tuulihov/est.html"
        </OBJECT>
      </OBJECT>
    </OBJECT>
...
```

Esimerkissä 8 HTML-dokumentin kieleksi on määritelty ranska, joten oletusarvoisesti kaikki teksti tulkitaan ranskaksi. Sekaan voidaan määritellä eri kielellä olevia kappaleita <P>-merkkaustunnisteella samoin kuin erikielisiä tekstin osia esimerkiksi EM-(emphasis) elementillä. EM-elementtiä käytetään jonkin tekstissä olevan osan korostamiseen eli sitä voidaan käyttää esimerkiksi silloin, kun halutaan merkata jollakin tietyllä kielellä olevat sanat tai sanonnat tekstistä.

Esimerkki 8. Dokumentin tekstisisällön kielen määrittely lang-attribuutilla.

```
<HTML lang="fr">
<HEAD>
<TITLE>Un document multilingue</TITLE>
</HEAD>
<BODY>
... où la romance à l'eau de rose déploie sa magnificence au son du
lyrisme échevelé d'un Michel Legrand. ...
<P lang="en"> Interpreted as English....<BR>
<P> ... de la romance-bonbon et de la belle musique...
<P> ... En 1964, le "cinéma-pompier" triomphe sur les écrans grâce à
Jacques Demy et ses Parapluies de Cherbourg.<EM lang="en">This is
written in English</EM> Pour tous ceux et celles qui acceptent les
conventions de la comédie musicale,...<BR>
...
```

Dir-attribuutilla määritellään tekstin ja taulukoiden lukusuunta eli luetaanko vasemmalta oikealle vai oikealta vasemmalle. Attribuutin arvo voi olla joko LTR (Left-to-right) tai RTL (Right-to-left). Esimerkissä 9 on HTML-dokumentti, jossa on eri suuntiin luettavaa tekstiä.

Esimerkki 9. HTML-dokumentti, jossa on eri suuntiin luettavaa tekstiä.

```
<HTML dir="LTR">
<HEAD>
<TITLE> ... vasemmalta oikealle luettava otsikko...</TITLE>
</HEAD>
<BODY>
<P>
... vasemmalta oikealle luettavaa tekstiä...
<P>Seuraava kappale luetaan eri suuntaan:</P>
<P dir="rlt">... naatnuus ire naateul ämäT...</P>
<P>Tästä jatketaan taas tavalliseen tapaan...</P>
...
</BODY>
</HTML>
```

ABBR-elementillä voidaan ilmaista tekstissä olevia lyhenteitä ja ACRONYM-elementillä puolestaan kirjainsanoja. Niihin voidaan liittää title-attribuutti, jossa on ilmaisu kokonaisuudessaan, sekä lang-attribuutti, joka ilmaisee, minkä kielinen lyhenne tai kirjainsana on kyseessä, ks. esimerkki 10.

Esimerkki 10. ABBR- ja ACRONYM-elementtien käyttäminen.

```
<P>
   Monikielisten
   <ABBR title="World Wide Web">www</ABBR>-dokumenttien
   kieliversioita hallitaan...
...
<P>
   <ABBR lang="fi"
```

```
        title="Jyv&auml;skyl&auml;n yliopisto">
        JY
    </ABBR> on kasvanut voimakkaasti viime vuosina...
...
<P>
    Latinankielisen lyhenteen<ABBR lang="la"> et al. </ABBR>
    tilalla Hirsjärvi suosittelee käytettävän suomenkielistä...
...
<P>
    <ACRONYM title="Federal Bureau
     of Investigation">FBI</ACRONYM> on
     USA:n liittovaltion poliisi.
...
```

2.4. XML-standardiperhe

XML-metakieli (Extended Markup Language) on kehitetty SGML:n pohjalta ja on sen osajoukko. Se on tarkoitettu rakenteisten dokumenttien esittämiseen erityisesti verkossa. World Wide Web Consortiumilla (W3C) oli XML:n kehitystyössä periaatteena kehittää sellainen SGML:n alijoukko, joka säilyttäisi SGML:n hyvät puolet, mutta samalla sopisi verkossa käytettäväksi. XML:ään liittyen kehitettiin myös standardit hyperlinkeille (*XLink*) ja tyylitiedostoille (*XSL*). (Goldfarb & Prescod, 1998).

XML

Yksi XML:n päätavoitteista oli, että merkkauskielen säännöt olisivat niin yksinkertaisia, että niistä voitaisiin pitää tiukasti kiinni. Dokumenttia, jonka merkkaus on metakielen sääntöjen mukainen, kutsutaan *hyvin muodostetuksi (well-formed)*. Dokumentti on myös *validi*, jos sen esittelyosassa (document type declaration) ilmoitetaan yhteensopivuus jonkun tietyn DTD:n kanssa ja se myös todella noudattaa kyseistä DTD:tä. (Bray ym., 1998; Goldfarb & Prescod, 1998; ks. myös Leinonen, 1998). XML eroaa SGML:stä muun muassa siinä, että XML-dokumentissa ei tarvitse olla mukana DTD:tä, jonka mukaisesti se on merkattu. Erityisesti verkkoympäristössä tämä on hyvä piirre, koska tällöin XML-merkatut dokumentit eivät tule liian raskaiksi käsitellä. Goldfarbin ja

Prescodin (1998) mukaan saattaa riittää pelkästään se, että XML-dokumentti on hyvin muodostettu, jos se on kooltaan pieni tai se on ainoa laatuaan. Jos dokumentti on suurikokoinen tai se on osa jotakin tietojärjestelmää, on sille syytä kirjoittaa DTD ja sen validiteetti on myös tarkistettava. XML-standardi sopii hyvin monikielisten verkkojulkaisujen merkkaamiseen, koska siinä on sitouduttu ISO 10646-standardin mukaisen Unicode-merkistön (ks. kohta 2.1) käyttöön, jonka koodauskapasiteetti riittää kaikille kirjoitetussa muodossa oleville kielille.

XSL

Extensible Stylesheet Language (XSL) määrittelee tyylitiedostot XML-dokumenttien ulkoasulle. Monikielisten verkkojulkaisujen kannalta tärkeä ominaisuus XSL-määrityksessä on ns. *kirjoitustapa (writing-mode)*, jonka avulla voidaan ilmaista eri kielten kirjoitussuunnat sivulla tai näytöllä. XSL sisältää monipuoliset määrittelyt erilaisille kirjoitussuunnille ja niiden yhdistelmille. Kirjoitustavat on määritelty XSL:ssä seuraavasti:

- left-to-right – top-to-bottom (”lr-tb”)
- right-to-left – top-to-bottom (”rl-tb”)
- top-to-bottom – right-to-left (”tb-rl”)
- jne.

Näistä ensiksi mainittu eli kirjoitussuunta vasemmalta oikealle ja ylhäältä alas on sivuilla oletusarvona. (Deach, 1999). Nykyiset HTML:llä toteutetut sivut, joilla on esimerkiksi oikealta vasemmalle luettavaa tekstiä, perustuvat useinkin frame-rakenteen käyttöön ja vaikuttavat XSL:llä toteutettuja kömpelömmiltä. Tosin asiaa on vaikea käytännössä todentaa, koska tämänhetkiset www-selaimet eivät tue XML:ää eivätkä myöskään HTML-kielen 4.0.-versiota.

XLink

XLink on tällä hetkellä keskeneräinen standardiluonnos, jonka kehitystyössä ei ole tapahtunut edistystä version 1.0 julkaisemisen jälkeen. Luonnoksessa on esitelty mahdollisuus liittää linkkeihin tietoa linkitettävästä resurssista, esimerkiksi sivun otsikko. Maler ja DeRose (1998) toteavat kuitenkin luonnoksessa, että kuvattu mekanismi on riittämätön suoriutuakseen kansainvälistämisestä, ja viittaavat tulevaan versioon, joka tarjoaa mahdollisuuden myös verkon kansainvälistämiseen, mikä pitää sisällään muun muassa monikielisten sivujen toteutuksen. Monikielisyyden toteuttamisen kannalta XLink-luonnos ei siis vielä tarjoa riittäviä ominaisuuksia eri kielillä olevien sivujen linkitykseen.

2.5. TEI

Kansainvälinen yhteistyöprojekti Text Encoding Initiative (TEI) aloitettiin v. 1987. Sitä sponsoroivat Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL) sekä Association for Literary and Linguistic Computing (ALLC). Runsaan viiden vuoden työn tuloksena projektissa kehitettiin SGML-standardiin pohjautuen *TEI-määritykset (TEI Guidelines)*, jotka on tarkoitettu erityisesti humanististen tekstien koodaamiseen elektroniseen muotoon sekä niiden välittämiseen elektronisessa muodossa (Sperberg-McQueen & Burnard, 1994). TEI-määritykset on suunniteltu siten, että niitä voidaan käyttää laajalti eri sovelluksissa. Burnardin ja Sperberg-McQueenin (1993) mukaan ne ovat joustavia, helppoja käyttää ja niitä voidaan laajentaa käyttäjän tarpeiden mukaan.

TEI-standardin perusteet

TEI-standardin juuret ovat humanistisessa tutkimusyhteisössä ja se on alunperin suunniteltu palvelemaan tutkimustyötä. Siksi TEI-standardin tavoitteena on olla mahdollisimman ymmärrettävä, joustava ja laajennettava. Täsmällisemmät TEI-standardin kehityksen tavoitteet ovat olleet:

- tarjota standardi formaatti tiedon vaihtoon
- tarjota suuntaviivat tekstien koodaamiseen tässä formaatissa
- tukea kaikenlaisten tutkimuksen kohteena olevien tekstien kaikkien piirteiden koodaamista
- olla sovellusriippumaton

Näiden tavoitteiden seurauksena on suunnittelun pohjaksi otettu SGML-metakieli ja 7-bittiseen koodaukseen perustuva ISO 646 -standardin merkistö (ISO, 1999b). TEI-standardi tarjoaa laajan ennalta määritellyn merkkaustunnisteiden joukon sekä mahdollisuuden koodata erilaisia näkymiä teksteihin ja käyttää vaihtoehtoisia tapoja koodata samoja tekstien ominaisuuksia. Käyttäjät voivat lisäksi laajentaa koodausta omilla määrityksillään. (Sperberg-McQueen & Burnard, 1994).

Kaikissa TEI-standardin mukaisissa teksteissä on kaksi osaa: *TEI-otsake (TEI header)* ja *varsinainen tekstielementti*. TEI-otsakkeessa on samankaltaisia tietoja kuin painetun tekstin otsikkosivulla. Se merkataan <teiHeader>-tunnisteella ja se sisältää seuraavat neljä osaa (Sperberg-McQueen & Burnard, 1994, s. 89-91; ks. myös Burnard & Sperberg-McQueen, 1993):

- *<fileDesc>*: elektronisen tekstin bibliografiset tiedot; osio on pakollinen (Sperberg-McQueen & Burnard, 1994, s. 93)
- *<encodingDesc>*: kuvaus tekstin koodauksessa käytetystä menetelmästä; ei pakollinen, mutta sen käyttöä suositellaan standardissa (Sperberg-McQueen & Burnard, 1994, s. 109)
- *<profileDesc>*: kuvaus tekstin ei-bibliografisista piirteistä kuten kieli, genre ja käyttötarkoitus; osio on valinnainen (Sperberg-McQueen & Burnard, 1994, s. 127)

- *<revisionDesc>*: tekstin yksityiskohtaiset päivitystiedot; osio on valinnainen, mutta sen käyttöä suositellaan vahvasti (Sperberg-McQueen & Burnard, 1994, s. 132)

TEI-otsake voidaan joko luoda osaksi dokumentin alkuperäistä koodausta tai sillä voidaan kuvata jo olemassa olevia dokumentteja. TEI-otsakkeet voidaan tallentaa erillään niistä dokumenteista, joita ne koskevat, ja niihin voidaan linkittää myös *MARC-tietueita (Machine Readable Cataloging Record)*, joiden avulla voidaan tunnistaa bibliografisen tiedon elementtejä. (Coleman & Willis, 1997; Attig, Klimczyk & Mangin, 1999).

TEI-tekstit ovat joko *unitaarisia (unitary)*, jolloin teksti muodostaa yhden kokonaisuuden, tai *yhdistettyjä (composite)*, jolloin teksti koostuu useista komponenteista, jotka ovat jollakin merkittävällä tavalla riippuvaisia toisistaan. TEI-dokumenttityyppimäärittelyssä tekstiosa merkataan <text>-tunnisteella ja se on jaettu <front>-, <body>- tai <group>- sekä <back>-elementteihin, joista ensimmäinen ja viimeinen ovat valinnaisia. Kun tekstiosa muodostuu unitaarisesta tekstistä, käytetään <body>-elementtiä. Jos taas tekstiosa on useiden alisteisten tekstien tai ryhmien muodostama eli yhdistetty, merkataan se <group>-elementillä. Ennen tekstiä tulevat alkutiedot, kuten esimerkiksi nimiösivu ja omistuskirjoitukset, esitetään <front>-elementissä. Tekstin jälkeen tulevat liitteet ovat <back>-elementissä. (Sperberg-McQueen & Burnard, 1994, s. 217-218). Liitteessä 1 on osa TEI:llä merkatusta artikkelista, jossa pääkielenä on englanti ja joka sisältää joitakin muunkielisiä sanoja.

TEI-merkkauskieli pohjautuu SGML-metakieleen ja sen periaatteelliseen oletukseen, että kaikille dokumenteille voidaan kuvata hierarkkinen rakenne. Todellisuudessa monet tekstit kuitenkin sisältävät useita erilaisia rakenteita, joiden elementit menevät limittäin. Varsinkin dokumentti, joka sisältää rinnakkaisia erikielisiä osia, voi olla rakenteeltaan hyvin erilainen eri kielten kohdalla, johtuen kielten erilaisista rakenteista ja ilmauksista samojen asioiden kohdalla. Tekstin looginen rakenne on käyttökelpoinen tutkimuksessa ja navigoinnissa, mutta myös fyysinen rakenne saattaa olla tutkijoiden kiinnostuksen kohde. Tällaisia tekstejä koodattaessa on hyvin tärkeää tarkasti pohtia, mikä on tärkein käyttötapa ja suunnitella merkkaus sen pohjalta. Koodattaessa yksi rakenne usein saakin

etuoikeutetun aseman ja muut hierarkkiset rakenteet koodataan vähemmän eksplisiittisesti. (Mah, Flanders & Lavagnino, 1997).

Monikielisyyden toteuttaminen TEI-standardissa

TEI-otsakkeessa voidaan jokainen tekstissä käytetty kieli määritellä <language>-elementillä. Monikielisessä tekstissä jokaisen kielen osuudesta voi olla viittaus asiaankuuluvaan <language>-elementtiin lang-attribuutin avulla. Lang-attribuutin arvot ovat kaksikirjaimisia kielikoodeja, jotka on määritelty ISO 639-standardissa (ISO, 1999d; ks. myös Seaman, 1992). Lang-attribuutti on sovellettavissa mihin tahansa elementtiin TEI-järjestelmässä, joten kielen vaihtuminen toiseksi voidaan ilmaista rakenteen kaikilla tasoilla aina yksittäisiin sanoihin saakka, ks. esimerkki 11. (UKOLN Metadata Group, 1998).

Esimerkki 11. Tekstissä käytettävän kielen merkkaaminen TEI-standardissa.

```
...
<langUsage>
        <language id="fr" iso639="fr">French</language>
        <language id="en" iso639="en">English</language>
        <language id="et" iso639="et">Estonian</language>
        <language id="la" iso639="la">Latin</language>
</langUsage>
...
Their estonian friend Katrin ate <foreign lang=fr>croissants</foreign>.
<foreign lang=et>"See oli väga maitseva, aitäh!"</foreign>, she said
smiling.
...
In the afternoon Katrin and and her cousin Joanna were sitting in the
garden with David and Jane. Joanna was reading Jerome's Latin Vulgate
Bible. <foreign lang=la>"... pone me ut signaculum super cor tuum ut
signaculum super brachium tuum quia fortis est ut mors dilectio dura sicut
inferus aemulatio lampades eius lampades ignis atque
flammarum..."</foreign>, she read loud and clear so that even aunt Mary
could hear it...
```

Vaikka TEI-määritykset ovat saaneet osakseen kritiikkiä liiallisen joustavuutensa vuoksi, tarjoavat ne käyttäjille mahdollisuuden soveltaa niitä hyvinkin erilaisiin tarkoituksiin. Niillä voidaan muun muassa kätevästi koodata tekstidokumentteja, joihin halutaan tavanomaisesta poikkeavia ominaisuuksia mukaan. Esimerkkinä tästä voidaan mainita koreankielisen sanakirjan merkkaus TEI-määritysten avulla, jota Kang (1997) tarkastelee artikkelissaan. Koreankielisissä sanakirjoissa muun muassa ilmoitetaan kasveille ja eläimille myös niiden tieteelliset nimet. TEI tarjoaa sanakirjojen merkkausta varten perusmallin, jota voidaan muokata esimerkiksi lisäämällä DTD:hen elementti <sciName> ilmaisemaan tieteellistä nimeä. Kang pitääkin TEI:n määrittelyä sanakirjaelementeille riittävän joustavana, jotta sillä voidaan merkata länsimaisesta käytännöstä poikkeava sanakirjamuotoinen julkaisu. Kangin mukaan tiukempi rakennemäärittely aiheuttaisi ongelmia koreankielisen sanakirjan merkkaamisessa. Tällainen joustavuus on hyvä ominaisuus, mutta toisaalta se aiheuttaa sen, että TEI-määrityksen soveltaminen vaatii sen käyttäjältä syvällisempää perehtymistä sen rakenteeseen ja tarjoamiin mahdollisuuksiin, ennen kuin niitä voi kunnolla hyödyntää.

Olsen (1996) on kritisoinut voimakkaasti TEI-määrityksiä muun muassa siitä, että ne ovat liian vaihtelevia ja joustavia ollakseen käyttökelpoisia tiedonsiirtoformaattina. Osaltaan tämä johtuu Olsenin mukaan siitä, että niiden määrittelyssä on pyritty ottamaan monien eri osapuolten tarpeet ja toiveet huomioon. Lisäksi Olsen arvostelee TEI-projektia siitä, että sen tuloksia ei ole riittävässä määrin kriittisesti arvioitu kansainvälisessä tiedeyhteisössä, vaan niitä on lähinnä vain esitelty Computers and the Humanities -aikakauslehdessä. Myös Coleman ja Willis (1997) viittaavat raportissaan TEI-määrityksiin kohdistuneeseen kritiikkiin: Niitä pidetään liian monimutkaisina ja kalliina, niiden arvellaan olevan käytettäviä vain ylimmillä tasoilla johtuen niiden joustavuudesta eikä niitä myöskään pidetä riittävän yksityiskohtaisina täyttämään erityisalojen vaatimuksia. Kuitenkin Coleman ja Willis pitävät TEI-määrityksiä niin merkittävinä, että niiden soveltuvuutta on tarkasti arvioitava ryhdyttäessä toteuttamaan humanististen tekstien koodausta.

3. TIPSTER-ARKKITEHTUURI JA SEN SOVELLUKSIA

Tässä luvussa käsitellään TIPSTER-arkkitehtuuria ja sen pohjalta kehitettyjä sovelluksia. Ne ovat monikielisyyden kannalta tärkeitä ja mielenkiintoisia, koska niiden suunnittelussa yhtenä lähtökohtana on ollut nimenomaan monikielisten dokumenttien käytön ja hallinnan kehittäminen.

Aluksi esitellään TIPSTER Text Program –projektia ja sitten TIPSTER-arkkitehtuurin perusperiaatteita. Sen jälkeen esitellään TEMPLE- ja GATE-järjestelmät, jotka molemmat ovat TIPSTER-arkkitehtuurimallin sovelluksia. Lopuksi kuvataan lyhyesti TEMPLE-projektin jatkona käynnistettyä URSA (UNICODE Retrieval System Architecture) –projektia, jossa suunniteltu monikielisyyttä tukeva arkkitehtuuri pohjautuu myös TIPSTER:iin. URSA:ssa kehitettyjen MUNDIAL- ja ARCTOS-hakujärjestelmien toimintoja esitellään ja arvioidaan lyhyesti verkossa toimivien demonstraatioiden pohjalta.

3.1. TIPSTER

TIPSTER Text Program on DARPA:n (Defense Advanced Research Projects Agency) johtama USA:n hallituksen projekti, jonka tarkoituksena on ollut edistää tekstinkäsittelyteknologiaa. Projekti alkoi 1991 ja päättyi virallisesti syksyllä 1998. Projektissa toimi yhteistyössä Yhdysvaltojen hallituksen, teollisuuden ja yliopistojen tutkijoita. Sen pyrkimyksenä oli parantaa dokumenttien käsittelyn tehoa sekä kustannustehokkuutta ja se keskittyi seuraavaan kolmeen teknologiaan, jotka muodostavat perustan lähes kaikille muille informaation käsittelytehtäville:

- *dokumenttien jäljittäminen:* kyky paikallistaa dokumentit, jotka sisältävät käyttäjän haluamaa tietoa, joko tekstivirrasta tai dokumenttivarastosta
- *informaation poimiminen:* kyky paikallistaa määrätty informaatio tekstistä
- *yhteenvetojen tekeminen:* kyky tiivistää dokumentti tai kokoelma säilyttäen samalla sen sisältämät avainajatukset

(TIPSTER Text Program Overview, 1998)

TIPSTER Text Program –projektin tärkein tulos on niin sanottu TIPSTER-arkkitehtuurimalli, jonka pääpiirteitä kuvataan projektin yleisesittelyn jälkeen.

3.1.1. TIPSTER Text Program -projekti

TIPSTER Text Program -projektin ensimmäisessä vaiheessa (v. 1991-1994) luotiin algoritmeja, joilla voidaan jäljittää dokumentteja ja poimia informaatiota tekstistä, sekä kehitettiin näiden tekniikoiden arviointia muun muassa sellaisissa konferensseissa kuin Message Understanding Conferences (MUC) ja Text Retrieval Conferences (TREC). Projektin ensimmäisen vaiheen tuloksena oli aikaisempiin tekniikoihin verrattuna huomattavasti parempi kyky tunnistaa yksityiskohtaista tietoa tekstidokumenteista. Sellaista tietoa ovat esimerkiksi nimet (sekä henkilöiden että organisaatioiden), päivämäärät, sijaintipaikat, ajankohdat ja puhelinnumerot. (TIPSTER Text Program Overview, 1998). Jaegerin ym. (1991) mukaan tällaisten yksittäisten tietojen löytäminen ja tunnistaminen teksteistä on usein vaikeaa, varsinkin kun käsitellään monikielisiä tekstidokumentteja. Monesti on vaikeaa erottaa tekstistä, mitkä elementit ovat henkilöiden tai yritysten nimiä, mitkä taas esimerkiksi paikannimiä, joilla joku yritys sijaitsee. Ongelma on hyvin konkreettinen muun muassa saksankielisessä tekstissä, jossa kaikki substantiivit kirjoitetaan isolla alkukirjaimella riippumatta siitä, ovatko ne erisnimiä vai eivät. Päivämäärien ja ajankohtien ilmaiseminen eri kielissä voi olla myös hyvin erilaista ja siksi niiden löytäminen saattaa olla vaikeaa.

TIPSTER-projektin toisessa vaiheessa (huhtikuu 1994 – syyskuu 1996) keskityttiin ohjelmistoarkkitehtuurin luomiseen. Vaiheen tavoitteina oli standardoida teknologiakomponentit, tehdä mahdolliseksi kehitettyjen välineiden yhteensopivuus keskenään sekä sallia ohjelmiston jakaminen eri osapuolille. Monikielisten dokumenttien käsittelyä kehitettiin Multilingual Entity Task (MET) -konferenssin puitteissa. (TIPSTER Text Program Overview, 1998).

Informaation tiivistäminen monikielisistä dokumenteista on usein hankalaa muun muassa sen vuoksi, että sanoja on vaikea erottaa toisistaan (sanojen rajakohtia ei kyetä tunnistamaan). Tämän vuoksi projektissa on kehitetty väline tekstien segmentointiin, jolloin tekstistä voidaan tunnistaa esimerkiksi erisnimiä ja teknisiä termejä, joita ei

löydy valmiista sanastoista. SGML-merkkausta on käytetty apuna sanojen tunnistamisen toteutuksessa. (Multilingual Entity Task Conference, 1998).

Projektin kolmas vaihe käynnistyi lokakuussa 1996. Monikielisyyden kannalta tärkeimpiä osa-alueita, joilla tutkimustyötä jatkettiin, olivat:

- yhdessä ympäristössä toimivien tekniikoiden ja välineiden räätälöinti (mukauttaminen), niin että ne toimivat toisilla kielillä ja/tai toisissa käyttöympäristöissä
- monikielisyyden tukeminen sekä informaation jäljittämisessä että tekstin tiivistämisessä
- käyttöliittymien suunnittelu ja käytettävyyden testaus.

(Phase III Overview, 1998)

3.1.2. TIPSTER-arkkitehtuuri

TIPSTER-arkkitehtuuri on suunniteltu yleiseksi malliksi, jota voidaan käyttää perustana rakennettaessa eri tarkoituksiin tulevia dokumenttien hallintajärjestelmiä. Arkkitehtuurimalli on kuvattu joukkona objektiluokkia ja niihin liittyviä toimintoja.

TIPSTER-objektiluokka koostuu luokan nimestä sekä joukosta nimettyjä ominaisuuksia ja operaatioita (Grishman, 1998, s. 3). Luokan ominaisuuksien arvoina voi olla objekteja ja objektijoukkoja, merkkijonoja, numeerisia arvoja, totuusarvoja (Boolean) tai jäseniä luetteloista. Ominaisuuden arvo voi olla myös tyhjä. Operaatiot voivat olla joko proseduureja, jotka eivät palauta mitään arvoa, tai arvon palauttavia funktioita. Osa objekteista ja toiminnoista on pakollisia eli ne täytyy toteuttaa missä tahansa arkkitehtuurin kanssa yhdenmukaisessa järjestelmässä. Osa on valinnaisia eli niitä ei ole pakko toteuttaa. Jos ne toteutetaan, on niiden silloin noudatettava arkkitehtuurin määrityksiä.

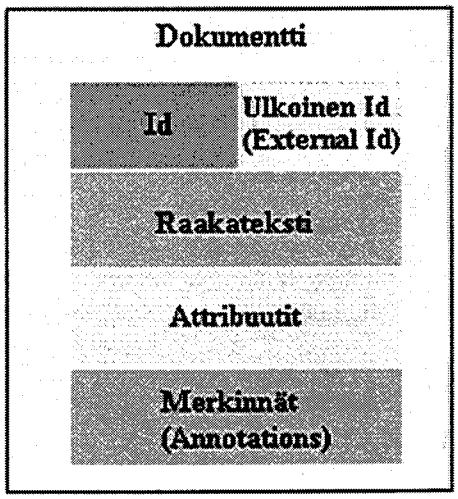
Dokumenttien hallinta on TIPSTER-arkkitehtuurin ydin, ja se sisältää kolme tärkeää objektiluokkaa: dokumentit, dokumenttikokoelmat ja merkinnät (annotations) (Grishman, 1995). Seuraavassa esitellään niitä hieman tarkemmin.

TIPSTER:in dokumenttien hallintajärjestelmän keskeisin objektiluokka on dokumentti. Dokumentti voidaan käsittää informaatioyksikkönä, joka palvelee useita arkkitehtuurin perustoimintoja:

- se muodostaa tekstiin liittyvän tiedon varaston attribuuttien ja merkintöjen (annotation) avulla
- se on perusyksikkö muodostettaessa kokoelmia
- se on perusyksikkö tiedon jäljittämisessä

(Grishman, 1998, s. 11).

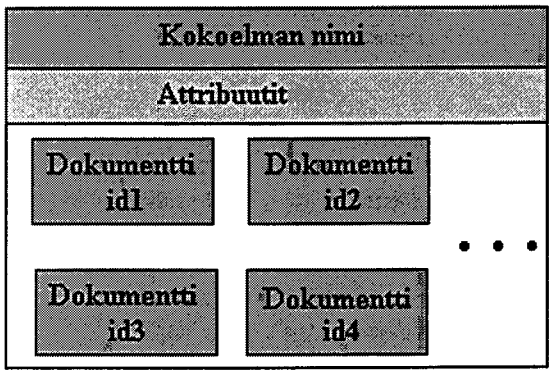
Dokumentin yleinen rakenne on esitetty kuviossa 2. Jokainen dokumentti on osa yhtä tai useampaa kokoelmaa ja sille on annettu yksikäsitteinen identiteetti Id-ominaisuuden avulla. Id on tunniste, joka annetaan dokumentille, kun se luodaan. Ulkoisen Id:n antaa dokumentille sovellus, jolla sitä käsitellään. Raakateksti on se sisältö, joka liitetään dokumenttiin sitä luotaessa. Sisältöä ei voida myöhemmin muuttaa. Attribuutit ovat ominaisuus-arvo –pareja, joissa ominaisuuden nimi on merkkijono (string) ja arvo voi olla esimerkiksi jokin merkkijono tai viittaus kokoelmaan, dokumenttiin, attribuuttiin tai merkintään (Grishman, 1998, s. 7). Merkinnät sisältävät suurimman osan siitä tiedosta, jonka TIPSTER-järjestelmä lisää dokumenttiin.



KUVIO 2. TIPSTER-dokumentin rakenne (Grishman, 1995).

Dokumentteihin voidaan kohdistaa monia operaatioita, kuten dokumentin luominen, kopiointi, merkintä (annotation) sekä WriteSGML ja ReadSGML. WriteSGML konvertoi dokumentin ja siihen kuuluvat merkinnät (annotations) SGML-muotoon. Tulosdokumentti on normalisoidussa SGML-muodossa, jossa on kaikki attribuutit ja lopputunnisteet (end tags) eksplisiittisesti. ReadSGML lukee merkkijonon, joka on merkattu normalisoidulla SGML:llä ja muodostaa dokumentin, jolla on määritelty ulkoinen Id (ExternalId) ja merkintäjoukko (AnnotationSet). Merkintäjoukko sisältää merkinnän jokaiselle SGML-tekstielementille, joka on merkattu syötetekstissä. (Grishman, 1998, s. 12).

Dokumentit on järjestetty kokoelmiin (collections), ks. kuvio 3. TIPSTER-järjestelmä ylläpitää luetteloa pysyvistä objekteista. Hajautetussa järjestelmässä eli siis silloin, kun toimitaan verkkoympäristössä, luettelo sisältää sekä isäntien (host) nimet että objektien yksikäsitteiset nimet isäntien sisällä (Grishman, 1998, s. 8). Pysyvät objektit ovat kokoelmia ja hauissa käytettyjä indeksejä. Jokainen dokumentti on osa jotakin kokoelmaa ja se voidaan hakea käyttämällä sen kokoelman nimeä ja sisäistä tai ulkoista tunnistetta (Id).



KUVIO 3. Kokoelmaobjekti TIPSTER-järjestelmässä (Grishman, 1995).

TIPSTER-arkkitehtuurissa käytetään ensisijaisesti merkintöjä (annotations) välineenä, jolla varastoidaan tietoa dokumenteista eli ne muodostavat pääosan dokumenttien metatiedoista. Tällaista tietoa on esimerkiksi dokumenttianalyysin tuloksena syntynyt informaatio, jota tarvitaan tekstin käsittelyssä. Se voi olla tietoa dokumentin rakenteesta, kuten esimerkiksi

- header-body –jaottelu
- osa- ja kappalejaottelu
- lauseiden jaottelu
- nimien tunnistaminen
- esityksen rakenne eli aiheen mukainen jaottelu.

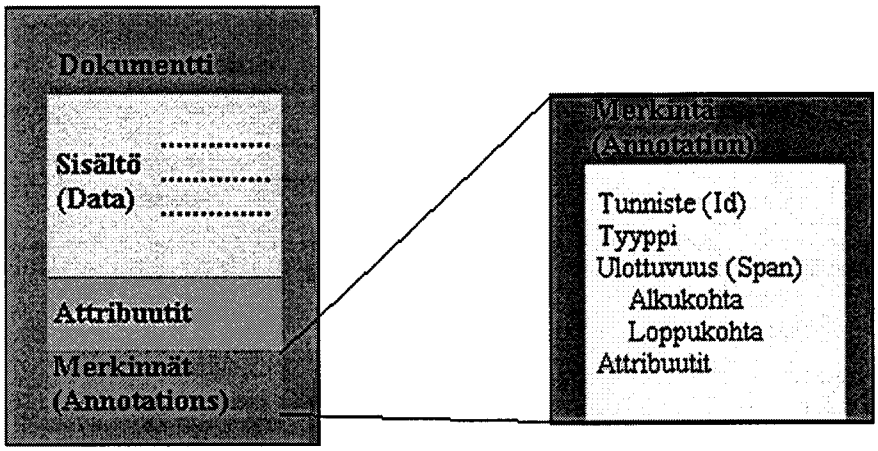
(Grishman, 1995).

Merkintä koostuu seuraavista komponenteista (ks. myös kuvio 4 ja esimerkki 12):

- tyyppi; esimerkiksi merkki, lause, kappale tai nimi
- merkinnän ulottuvuus (span), joka voi sisältää yhden tai useampia aloitus- ja lopetuskohtia dokumentissa
- joukko attribuutteja esimerkiksi nimille, merkeille, toisista merkinnöistä johdetuille merkinnöille ja oikolukutoiminnon käyttämille korvaaville arvoille.

(Grishman, 1995; Grishman, 1998).

Merkinnät voivat myös muodostaa hierarkkisia rakenteita, koska attribuutin arvo voi olla osoitin toiseen merkintään tai dokumenttiin, ks. esimerkki 13.



KUVIO 4. Merkinnän rakenne (Grishman, 1995).

Ulottuvuus eli jänneväli (span) ilmaisee, mitä kohtaa dokumentista merkintä koskee. Dokumentissa voi olla useita merkattuja kohtia. Kuviossa 4 on yksinkertaisuuden vuoksi esitetty vain yksi merkintä.

Esimerkki 12. Lauseen merkintä TIPSTER-järjestelmässä.

```
W 3 C   k e h i t t ä ä   X N L : ä ä .
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 0
```

Lauseen tekstialkiot on merkattu seuraavasti:

Tyyppi	Alku	Loppu	Attribuutit
Tekstialkio (token)	0	3	
Tekstialkio (token)	4	12	
Tekstialkio (token)	13	20	

Nimen tunnistaja (name recognizer) lisäisi merkinnän

Nimi	0	3	Nimi_tyyppi="organisaatio"

ja oikolukutoiminto lisäisi merkinnän

Oikeinkirj	13	20	Korvaus="XML:ää."

Esimerkki 13. Osoittimien käyttö merkinnöissä.

Tiina Tomera, sihteeri, Virtuaalitodellisuus Oy.

```
T i i n a   T o m e r a , s i h t e e r i ,
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2

V i r t u a a l i t o d e l l i s u u s   O y .
3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
```

Tekstin merkinnät:

Id	Tyyppi	Ulottuvuus	Attribuutti	Arvo
A1	Nimi	0-11	Nimi_tyyppi	Henkilö
A2	Nimi	23-45	Nimi_tyyppi	Organisaatio
A3	Henkilöstö	0-46	Nimi	Merkintä: A1
			Org	Merkintä: A2
			Asema	"sihteeri"

Grishmanin (1995) mukaan merkinnät (eli metatiedot) ovat keskeisiä koko TIPSTER-arkkitehtuurissa, koska ne muodostavat perustan informaation välittämiselle arkkitehtuurin eri moduulien välillä. Ne myös palvelevat niitä kolmea tärkeää perustehtävää, joihin TIPSTER Text Program -projekti keskittyi: dokumenttien jäljittäminen, informaation poimiminen tekstistä sekä yhteenvetojen tekeminen.

3.1.3. TIPSTER:in soveltuminen monikielisten verkkojulkaisujen toteutukseen

Arkkitehtuurimallissa on useita ominaisuuksia, joiden perusteella sitä voidaan soveltaa, kun suunnitellaan monikielisten verkkojulkaisujen käsittelyjärjestelmän arkkitehtuuria:

- Siinä on mahdollista käyttää erilaisia ohjelmistokomponentteja (Grishman, 1995).
- Sitä voidaan soveltaa monenlaisiin ohjelmisto- ja laiteympäristöihin (Grishman, 1995).

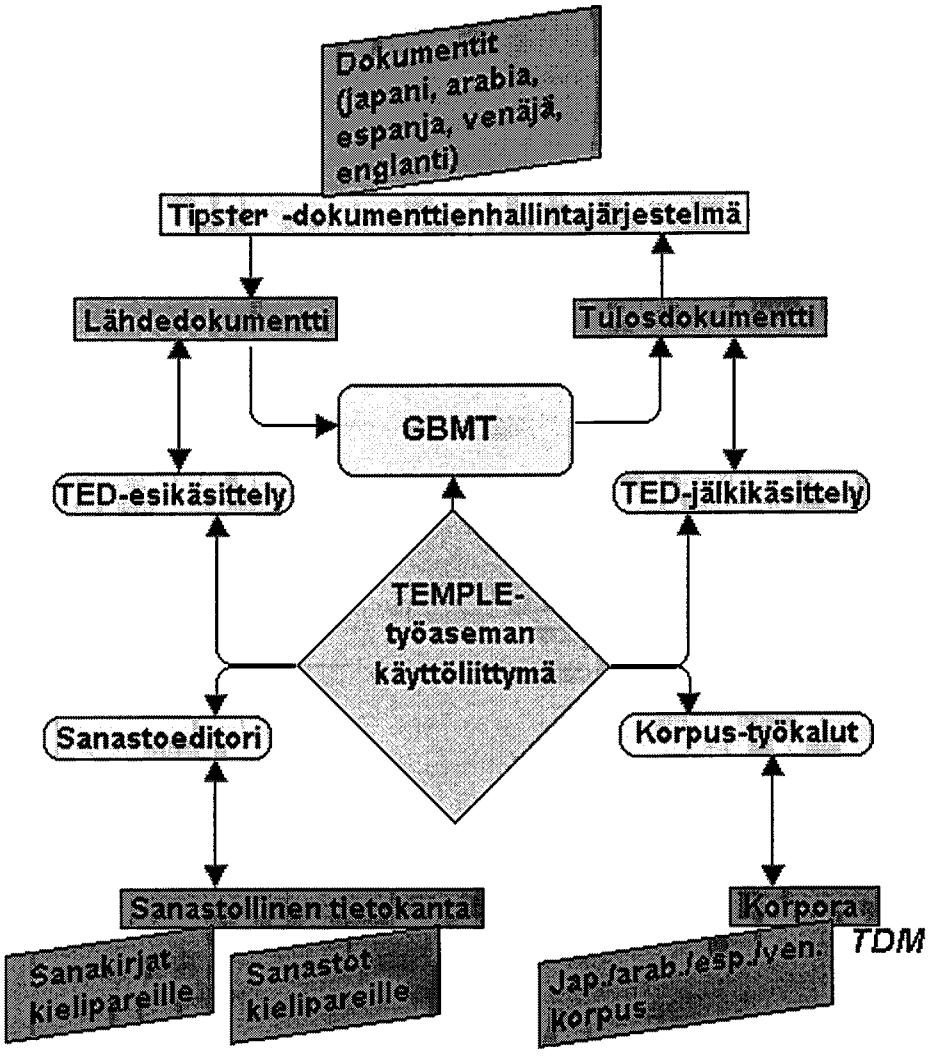
- Informaatiota voidaan jäljittää, tiivistää ja saada käyttöön tekstiin määriteltyjen merkintöjen (annotations) avulla (Grishman, 1995), jotka muodostavat dokumenttien metatietovaraston.
- TIPSTER Text Program -projektissa yksi keskeinen osa-alue on ollut monikielisten dokumenttien hallinnan kehittäminen (Architecture Committee for the TIPSTER Text Phase II Program, 1996a, s. 5 ja 1996b, s. 5; Multilingual Entity Task Conference, 1998).
- TIPSTER-arkkitehtuurissa voidaan esimerkiksi attribuuttien avulla ilmaista dokumenttikokoelmissa olevien erikielisten dokumenttien kieli ja käytetty merkistö (Architecture Committee for the TIPSTER Text Phase II Program, 1996b, s. 13). Merkintöjen (annotations) avulla puolestaan voidaan merkata dokumenttien sisällä olevia erikielisiä osia yksittäisiin sanoihin saakka (Grishman, 1995).
- Arkkitehtuurimalli tarjoaa mahdollisuudet muun muassa suorittaa hakuja eri kielillä oleviin dokumentteihin käyttäen englanninkielistä kyselylauseketta. Mallin pohjalta kehitetty järjestelmä tarjoaa käyttöliittymän, jonka avulla englanninkielinen kyselylauseke voidaan kääntää yhdelle tai useammalle vieraalle kielelle.(Architecture Committee for the TIPSTER Text Phase II Program, 1996b, s.19).
- Merkinnät on mahdollista konvertoida SGML-merkkaukseksi WriteSGML-operaatiolla ja vastaavasti voidaan SGML-merkattu teksti konvertoida ReadSGML-operaatiolla TIPSTER-mallin mukaiseksi tekstiksi, joka sisältää merkintöjä eli annotaatioita. (Grishman, 1998).

3.2. TEMPLE

TEMPLE-järjestelmä on kehitetty Yhdysvalloissa New Mexico State University:n (NMSU) Computing Research Laboratory:ssä (CRL). Projektissa on rakennettu avoin monikielinen arkkitehtuuri ja ohjelmisto TIPSTER-arkkitehtuurimallia soveltaen. Zajacin ja Casperin (1997) mukaan TEMPLE-järjestelmää ei alun perin suunniteltu verkossa tapahtuvaa selailua varten, mutta se sopii ominaisuuksiltaan hyvin myös verkkosovelluksen kehityksen pohjaksi, koska se on suunniteltu suhteellisen pienten, heterogeenisten, eri kielillä olevien dokumenttien selailuun, joita käytetään erilaisissa ympäristöissä ja jotka on toteutettu eri tyyleillä. Aluksi esitellään TEMPLE-prototyyppi ja sitten TEMPLE –verkkokäännösarkkitehtuuri.

TEMPLE-prototyyppi

Tavoitteena Temple-projektissa on ollut tukea konekääntämisen (Machine Translation, MT) nopeaa kehitystä sekä erityisesti harvinaisten tai vaikeasti saavutettavien kielten käyttöä. Prototyypissä ovat mukana seuraavat kielet: arabia, espanja, japani, venäjä ja englanti. Eri kielillä olevat lähdedokumentit käännetään prototyypissä aina englanniksi. Prototyyppi on esitetty kuviossa 5:



KUVIO 5. TEMPLE-järjestelmän perustoiminnot (Zajac & Vanni, 1997).

TEMPLE-protyypissä on:

- GBMT-kone (Glossary Based Machine Translation), jolla voidaan automaattisesti kääntää jokainen kielipari englanti & arabia/espanja/japani/venäjä
- kaksikieliset sanakirjat ja sanastot espanjan, arabian, japanin ja venäjän kielille sekä kielitieteellisiä apuohjelmia (morfologisia analysoijia ja englanninkielinen morfologinen generaattori)
- monikielisten dokumenttien editori (Tipster Editor for Documents, TED), jota käytetään dokumenttien ja niiden käännösten selaamiseen ja editointiin
- monikielisen sanakirjan ja sanaston editori sekä mahdollisuus jäsentää ja ladata sanakirja- ja sanastotiedostot systeemin sanastolliseen (leksikaaliseen) tietokantaan

(Overview of the Temple project, 1996).

TEMPLE:n prototyypissä saadaan automaattinen raakakäännös englanniksi lähdedokumenteista, joita on useilla kielillä (arabia, espanja, japani ja venäjä). Raakakäännökset tehdään käyttäen sanastoperustaista konekääntämistä (Glossary-Based Machine Translation, GBMT). Kieltä analysoivat henkilöt ja kielenkääntäjät voivat myös editoida sekä lähdedokumenttia että raakakäännöstä käyttäen projektissa toteutettua monikielistä editoria (multilingual editor) TED (Tipster Editor for Documents). Lähdedokumentteja ja niiden käännöksiä hallitaan projektissa kehitetyllä Tipster-dokumenttien hallintajärjestelmällä (Tipster Document Manager, TDM).

TIPSTER-arkkitehtuurimallia käytetään myös systeemin komponenttien integroinnin arkkitehtuuriperustana. Projektin yksi tärkeä tulos on järjestelmä, joka tukee luonnollisen kielen käsittelytyökalujen (ohjelmien) sekä tekstidokumenttien ja dokumenttikokoelmien uudelleenkäyttöä. Tällöin ulkopuolisista lähteistä hankitut työkaluohjelmat, joita käytetään esimerkiksi tekstien morfologiseen analysointiin tai merkkaukseen, voidaan integroida järjestelmään pienellä ohjelmointiponnistuksella. Järjestelmässä voidaan jäsentää heterogeeniset kielitieteelliset resurssit ja muuntaa ne yleiseksi monikieliseksi esitykseksi.

Temple-verkkokäännösjärjestelmä

Temple-prototyypin pohjalta on kehitetty järjestelmä www-sivujen kääntämiseen. Taustana tälle kehitystyölle on Internetin lisääntynyt käyttö ja se, että verkossa on yhä enemmän erikielisiä dokumentteja (Zajac & Casper, 1997). Erikielisten www-sivujen lisääntyminen on hyvä asia siinä mielessä, että silloin jokainen voi saada verkosta informaatiota omalla kielellään (Oudet, 1997). Siihen liittyy kuitenkin myös monenlaisia ongelmia, kuten esimerkiksi vaara, että muodostuu pieniä erillisiä informaatiosaarekkeita, joiden rajoja ei kyetä ylittämään, ts. niiden sisältämää tietoa ei saada laajemmin käyttöön johtuen dokumenteissa käytetystä kielestä.

Zajacin ja Casperin (1997) mukaan verkossa jo olevat lukuisat www-sivujen kääntämiseen tarkoitetut ohjelmat ovat pienikokoisia, riittävän nopeita (viive käännöksen saamiseen ei ole kovin suuri), halpoja ja ne kykenevät käsittelemään useimmat niille syötetyt HTML-dokumentit. Tällaisia ohjelmia ovat esimerkiksi joidenkin verkossa toimivien hakukoneiden (muun muassa AltaVista) käännösohjelmat, joilla käyttäjä voi käännättää haetun www-sivun jollekin toiselle kielelle. Niiden tuottaman käännöksen laatu ei ole kovin hyvä, mutta useimmille verkon käyttäjille se kuitenkin ilmeisesti riittää antamaan käsityksen dokumentin sisällöstä, ainakin karkealla tasolla.

Jos halutaan laadukkaita käännöksiä, on käytettävä pitkälle kehitettyjä konekääntämiseen (Machine Translation, MT) tarkoitettuja ohjelmia. Niiden puutteena on suuri koko, mikä hidastaa ja vaikeuttaa niiden käyttöä. Koska ne vaativat resursseja huomattavasti enemmän kuin halvat verkkokäännösohjelmat, on niiden käytössä rajoituttava muutamaan kieleen kerrallaan. Lisäksi ne on useinkin suunniteltu jonkin erityisalueen tarpeisiin, joten ne eivät kovin hyvin sovi yleisiksi käännösohjelmiksi www-sivuille. (Zajac & Casper, 1997).

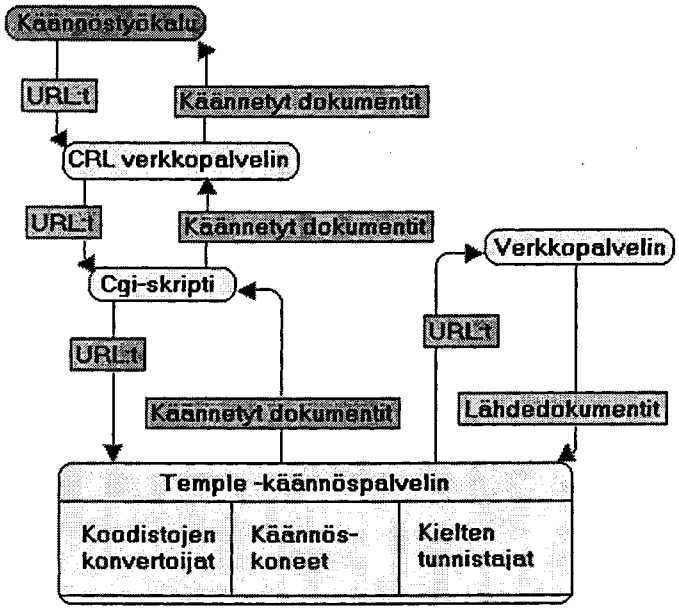
Sivujen sisältö on verkkosivujen selailijoille pääasiallinen kiinnostuksen kohde (Nielsen, 1999; Zajac & Casper, 1997) ja he haluavat sen käyttöönsä mahdollisimman nopeasti. Verkkokäännösohjelman on sen vuoksi oltava niin vaatimaton kuin mahdollista, jotta viive käännetyn sivun saamiseen ei kasva liian suureksi.

Nykyinen HTML-standardi ei tue monikielisten dokumenttien käsittelyä verkossa, ja muun muassa Carrasco Benitez (1996) sekä Yergeau, Nicol, Adams ja Duerst (1997) ovat esittäneet ajatuksia verkon kansainvälistämisestä ja monikielisyyden tukemisesta. W3C onkin käynnistänyt HTML:n version 4.0 kehitystyön, jossa on tarkoitus olla täysi tuki monikielisten dokumenttien käsittelylle verkossa (ks. kohta 2.3., jossa asiaa on tarkasteltu lähemmin).

Odoteltaessa monikielisyyttä tukevan HTML 4.0 –version valmistumista sekä erityisesti sellaisten selainten kehittämistä, jotka tukevat monikielisyyden kannalta tärkeitä HTML 4.0-version uusia ominaisuuksia (ks. kohta 2.3.), on Temple-projektissa pyritty toteutettamaan käyttöliittymä, joka tarjoaisi hyvälaatuisen käännöksen, mutta ei kuitenkaan olisi suurikokoinen eikä raskas ja hidas käyttää. Valitettavasti prototyypin kokeilu ja arviointi käytännössä ei ollut mahdollista, koska verkossa oleva demonstraatio ei ollut vapaasti kokeiltavissa.

Temple-verkkokäännösjärjestelmässä käyttäjä voi lähettää käännöspyynnön käännöspalvelimelle. Käyttäjän tarvitsee vain syöttää järjestelmälle käännettävän dokumentin URL. Käännöspalvelin palauttaa englanninkielisen käännöksen vieraskielisestä www-sivusta uuteen selainikkunaan.

Verkkokäännösjärjestelmän arkkitehtuuri on esitetty kuviossa 6. Käyttäjä voi syöttää käännettävän dokumentin URL:n JavaScriptillä toteutetulla käännöstyökalulla, joka tekee käännöspyynnön osoitetusta dokumentista. Käännöspyyntö välitetään Temple-verkkokäännöspalvelimelle CRL-verkkopalvelimen ja cgi-skriptin kautta. Verkko-käännöspalvelin hakee käännettävän dokumentin verkosta, jäsentää dokumentin rakenteen, kääntää sen tekstisisällön ja palauttaa käännetyn dokumentin alkuperäisessä HTML-muodossa käyttäjälle. Dokumentin käännös näytetään selaimessa erillisessä ikkunassa.



KUVIO 6. Temple-verkkokäännösjärjestelmän arkkitehtuuri (Zajac & Casper, 1997).

3.3. GATE

GATE (General Architecture for Text Engineering) on New Mexico State University:n Computing Research Laboratory:n ja Sheffieldin yliopiston yhteisprojekti (CRL Current Research, 1999). Sen tavoitteena on rakentaa laaja-alainen luonnollisen kielen käsittelyjärjestelmä, jossa voidaan hyödyntää uudella tavalla jo olemassa olevia ohjelmisto- ja tekstiresursseja (CRL Current Research, 1999; Cunningham, Wilks & Gaizauskas, 1996). Tekstiresurssit sisältävät muun muassa laajoja kielitieteellisiä tekstikokoelmia, joissa on monikielistä aineistoa. Niiden tallentamisessa ja uudelleenkäytössä on huomioitava sellaisten standardien käyttö, jotka tukevat mahdollisimman monien eri kielten merkistöjä. Standardeilta vaaditaan helppokäyttöisyyttä sekä joustavuutta, jotta niitä voidaan soveltaa eri käyttökohteissa halutulla tavalla. Standardien on oltava huomaamattomia (läpinäkyviä eli transparentteja) käyttäjille, koska muutoin niitä ei kokemusten mukaan haluta soveltaa

tekstiresurssien tallennuksessa tai käytössä, kuten Cunningham, Peters, McCauley, Bontcheva ja Wilks (1998) toteavat.

Dataresurssien uudelleen käytöstä on useita onnistuneita esimerkkejä kuten

- WordNet, joka on laaja sanastotietokanta
- Penn Tree Bank, joka on suuri englanninkielinen merkattu tekstikorpus, sekä
- Longmanin nyky-englannin sanakirja.

(Cunningham ym., 1996).

Ohjelmistoresurssien uudelleen käyttö on rajoittuneempaa. Syynä tähän on muun muassa kulttuurinen vastarinta (ei luoteta ulkomaalaiseen ohjelmakoodiin) ja integrointivaikeudet eri komponenttien välillä. Näitä ongelmia voidaan ratkaista seuraavilla tavoilla: Lisätään uudelleen käytettävien yksiköiden granulariteettia eli tarjotaan joukko pieniä rakennuspalikoita suurten kokonaisuuksien sijaan, parannetaan tutkijoiden luottamusta ohjelmistoresursseihin lisäämällä niiden käyttöä, testausta ja kehitystyötä, sekä tarjoamalla ohjelmistoarkkitehtuurimalli luonnollisen kielen käsittelysysteemeille. (Cunningham ym., 1996).

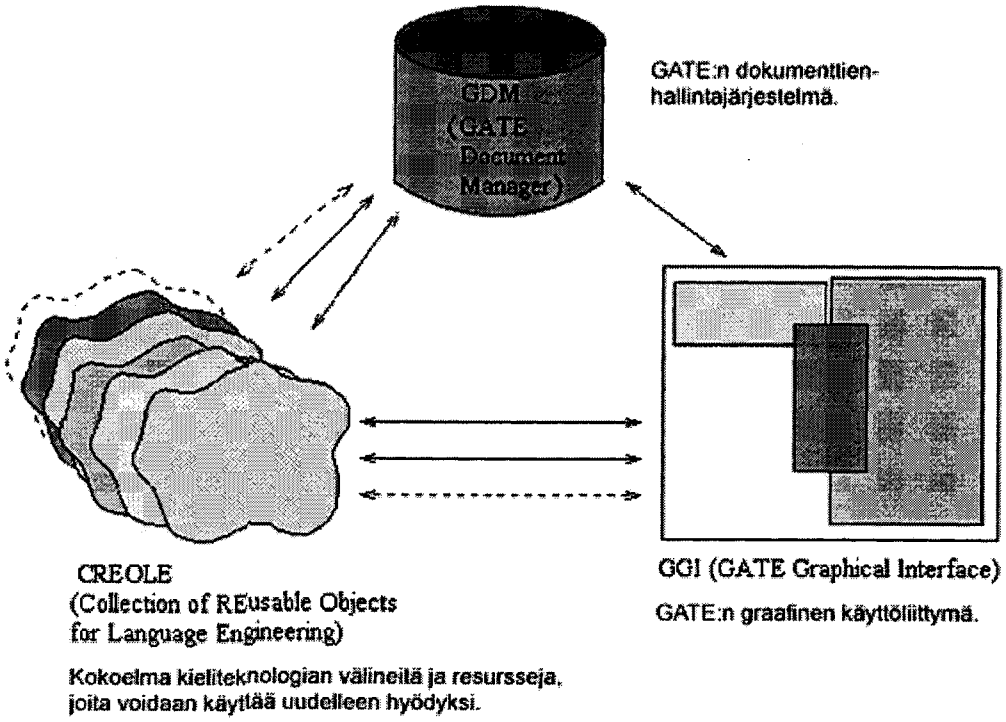
Cunninghamin ym. (1996) mukaan ohjelmistojen menestyksellinen uudelleen käyttö vaatii sellaisen yleisen arkkitehtuurimallin luonnollisen kielen käsittelyyn, mikä on erityisesti suunniteltu tekstinkäsittelyjärjestelmiä varten. GATE-järjestelmän kehitys pohjautuu tähän lähestymistapaan.

GATE-arkkitehtuurin pääpiirteitä ovat

- Olio-suuntautunut malli dokumenttien sisältämälle informaatiolle.
- Hajautettu dokumenttien hallinnan tietokanta, joka perustuu Tipster-arkkitehtuurille.
- Mahdollisuus lukea tai tulkita ja kirjoittaa SGML-merkattua tekstiä.

(GATE Architecture Overview, 1998).

GATE-arkkitehtuurimalli muodostuu kolmesta pääkomponentista, jotka on esitetty kuviossa 7.



KUVIO 7. GATE-arkkitehtuurin pääkomponentit (Gaizauskas, Rodgers, Cunningham, Humphreys & Robertson, 1998).

GATE:n dokumenttien hallintajärjestelmä (GATE Document Management, GDM) perustuu TEMPLE-projektissa kehitettyyn järjestelmään, joka on rakennettu TIPSTER-arkkitehtuurimallin pohjalta. (Cunningham ym., 1996).

Ide ja Veronis (1994) pitävät TEI-määrityksiä liian vähän testattuina varsinkin monikielisissä ympäristöissä ja katsovat, että TEI:n määrityksiä (ks. myös kohta 2.5.) on laajennettava, jotta ne soveltuisivat laajojen monikielisten tekstikorpusten merkkaukseen. Tämän vuoksi GDM-malliin on lisätty myös SGML-tuki.

GDM on tietokanta, joka toimii varastointipaikkana kaikelle sille tiedolle, jota järjestelmä tuottaa käsittelemistään tekstidokumenteista (Cunningham, Gaizauskas & Wilks, 1995). Kaikki järjestelmän komponentit kommunikoivat GDM:n kautta eli ne on eristetty toisistaan, ja GDM toimii niiden yhteisenä rajapintana. Etuna tästä on muun

muassa se, että näin voidaan tehokkaasti hyödyntää tietokannan ominaisuuksia, valvoa järjestelmää hajautetusti ja vähentää komponenttien riippuvuutta toisistaan. (Cunningham ym., 1996).

GATE:n graafinen käyttöliittymä on kehitetty Sheffieldin yliopistossa kieliteknologiaohjelmistojen käyttöön. Sen avulla voidaan selata ja testata tuloksia ja käyttää eri ohjelmistokomponentteja sekä liittää objekteja erilaisiin systeemikokoonpanoihin. (Cunningham ym., 1996).

GATE:een perustuvassa kieliteknologiajärjestelmässä kaikki työ tehdään CREOLE:n (Collection of REusable Objects for Language Engineering) sisältämillä moduuleilla. CREOLE:ssa on ohjelmistokomponentteja, joilla voidaan esimerkiksi analysoida tekstejä, tuottaa tiivistelmiä niistä, kääntää niitä tai suorittaa hakuja SQL-lauseilla. CREOLE:n sisältämät moduulit voivat olla koostumukseltaan paitsi algoritmeja, myös dataresursseja tai ne voivat sisältää myös molempia. (Cunningham ym., 1996).

Cunninghamin ym. (1998) mukaan GATE-projektin jatkotavoitteena on kehittää yleinen oliosuuntaunut malli, jossa erilaiset resurssit on integroitu yhteen ja joita voidaan käyttää www-protokollia hyödyntämällä. Mallissa eri resurssit voivat säilyttää oman alkuperäisen rakenteensa ja käyttää erilaisia standardeja tallennuksessa. Resursseja voidaan käyttää hajautetusti ja joustavasti verkon kautta, joten niitä ei tarvitse asentaa paikallisesti.

3.4. URSA

URSA eli UNICODE Retrieval System Architecture on myös New Mexico State University:n projekti kuten TEMPLE:kin, ja siinä jatketaan TIPSTER-projektin kolmannen vaiheen tutkimusta ja kehitystyötä. Tavoitteena URSA-projektissa on kehittää tekstinkäsittely- ja informaation hakutoiminnot dokumenteissa käytetyistä kielistä riippumattomaksi. (URSA Unicode Retrieval System Architecture, 1998).

URSA-projektin kehitystyö keskittyy Unicode-standardin integroimiseen TIPSTER-arkkitehtuurimalliin. Tämä merkitsee sitä, että URSA:ssa on mahdollista indeksoida ja hakea dokumentteja kaikilla kielillä, joita voidaan koodata Unicodella (Davis & Ogden,

1997). Projektissa kehitettävä malli hyödyntää dokumenteissa olevia merkintöjä (annotations), joilla kuvaillaan tekstiä indeksointia varten. URSA-kone (URSA engine) tulkitsee tekstistä merkityt lohkot tai sanat ja indeksoi ne hakuja varten. Indeksoinnin perusteella voidaan ilmaista monimutkaisia kyselylausekkeita ja suorittaa tekstihakuja. (URSA Unicode Retrieval System Architecture, 1998). Projektissa on kehitetty kaksi verkosta löytyvää demoversiota: MUNDIAL- ja ARCTOS-järjestelmät, joita käytetään monikieliseen tiedonhakuun verkosta. Seuraavassa on lyhyt esittely näistä molemmista.

MUNDIAL

MUNDIAL (MUNDIAL NET SEARCH III, 1998) on internet-hakujärjestelmän demonstraatio, jonka avulla voidaan hakea monikielisiä www-dokumentteja. Englanninkielinen kyselylauseke käännetään ensin listasta valitulle kielelle (huhtikuussa 1999 valittavissa 12 eri kieltä), minkä jälkeen MUNDIAL välittää etsintäpyynnön käyttäjän valitsemalle hakukoneelle. Hakukoneina ovat Alta Vista, Infoseek, Yahoo, Excite, Lycos ja venäläinen Rambler. Esittelyn mukaan haku suoritetaan nimenomaan käännetyillä hakusanoilla, mutta todellisuudessa haku tapahtuu sekä alkuperäisillä englanninkielisillä termeillä että niiden käännöksillä. MUNDIAL ei siis toimi www-sivulla olevan kuvauksen mukaan eikä siis käyttäjän odottamalla tavalla, mikä on käyttäjää harhaanjohtavaa.

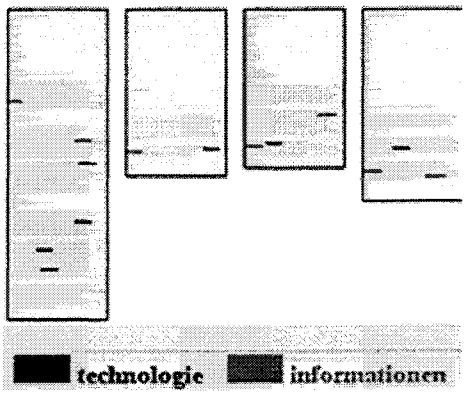
Huono piirre MUNDIAL:ssa on myöskin se, että tuloksena saattaa tulla sivuja, joissa ei ole näkyvissä haettua sanaa, vaan se sisältyy esimerkiksi sivulla olevaan valintalistaan. Verkosta tietoja hakevaa käyttäjää kiinnostaa sivulla näkyvä teksti, joten hän yleensä odottaa hakusanan esiintyvän siinä eikä alasvetovalikossa tai muualla näkymättömissä.

ARCTOS

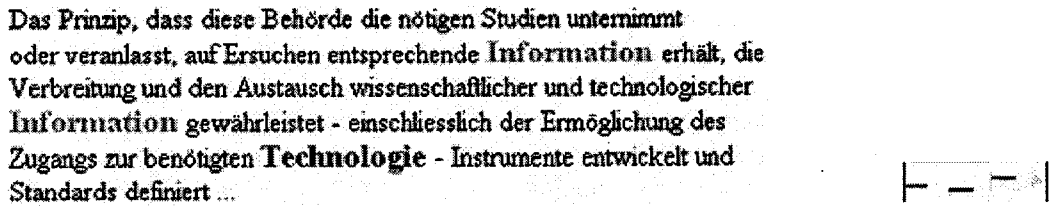
ARCTOS (Arctos: Interactive Multilingual Search with Ursa, 1998) on järjestelmä, joka on suunniteltu dokumenttien visualisointiin interaktiivisessa tiedonhaussa. ARCTOS:in demonstraatio on rakennettu www-sivujen monikielistä tiedonhakua varten. ARCTOS:issa kysely syötetään englanninkielisenä ja käännetään halutulle kielelle

(demonstraatioversion vaihtoehtoina ovat saksa, ranska ja italia). Käännöksen jälkeen kyselyä voidaan vielä korjata tai muuttaa, minkä jälkeen se lähetetään www-hakukoneelle suoritettavaksi.

Kuviossa 8 näkyy ARCTOS-järjestelmän näkymä hakusanojen esiintymisestä haun tuloksena saaduissa dokumenteissa. Hakulauseke on annettu englanninkielisenä ("information technology") ja käännetty ARCTOS:illa, jolloin kääntäjä antoi hakusanoiksi "informationen technologie". Tämän jälkeen on suoritettu haku demonstraatiota varten kootusta dokumenttivarastosta. Kukin hakusana näkyy dokumenteissa omalla värillään. Näytössä voidaan myös selata dokumenttien tekstisisältöjä, joissa hakusanat näkyvät eri väreillä merkattuina, ks. kuvio 9.



KUVIO 8. Avainsanojen esittäminen kuvakkeina (thumbnails) ARCTOS-hakujärjestelmässä (Arctos: Interactive Multilingual Search with Ursa, 1998).



KUVIO 9. Merkatut avainsanat dokumentin tekstisisällössä. (Arctos: Interactive Multilingual Search with Ursa, 1998).

4. MONIKIELISTEN RAKENTEISTEN VERKKOJULKAISUJEN HALLINTA

Tässä luvussa käsitellään monikielisten rakenteisten verkkojulkaisujen hallintaan liittyviä seikkoja. Perustan tarkastelulle muodostavat toisaalta monikielisyyden toteutuksen mahdollistavat ja dokumenttien rakenteen kuvaamiseen tarkoitetut standardit (luku 2), toisaalta TIPSTER -dokumenttien hallinnan arkkitehtuuri ja sen sovellukset (luku 3).

Monikielisten rakenteisten verkkojulkaisujen hallintaa tarkastellaan jaoteltuna seuraaviin osa-alueisiin:

- esittäminen ja organisointi,
- tuottaminen ja ylläpito,
- tiedon hakeminen monikielisessä ympäristössä.

4.1. Esittäminen ja organisointi

Esittämisessä on kiinnitettävä huomiota sekä julkaisun ulkoasuun että sisältöön niin, että se palvelee käyttäjän tarpeita mahdollisimman hyvin. Monikielisten verkkojulkaisujen esittämisessä näytöllä on huomioitava sellaisia eri kieliin ja kulttuureihin liittyviä seikkoja, kuten erilaiset käsitykset värien, symbolien ja sanontojen merkityksestä. Elementtien sijoittelu ruudulla ohjaa ja auttaa käyttäjää, ja se riippuu myös julkaisun sisällössä käytetyistä kielistä.

Julkaisun sisältö on se pääasia, jota käyttäjät sivuilta hakevat, kuten Nielsen (1999) toteaa, joten sen esitystapaan, muun muassa rakenteeseen ja laajuuteen, on kiinnitettävä erityistä huomiota. Monikielisissä verkkojulkaisuissa se tarkoittaa sitä, että käyttäjä saa tiedon julkaisun kieliversioista ja sen sisältämistä erikielisistä osista, sanonnoista ja yksittäisistä sanoista. Kaikki kieliversiot ja julkaisun erikieliset osat täytyy saada käyttöön tasavertaisesti. Verkkojulkaisu on pyrittävä esittämään näytöllä tiiviissä ja ytimekkäässä muodossa, koska käyttäjät eivät mielellään lue näytöltä pitkiä

kokonaisuuksia. Samoin metatietojen esitystapaan, muun muassa niiden sijoitteluun, laajuuteen ja käytettyyn kieleen, on kiinnitettävä huomiota.

Organisoinnilla tarkoitetaan verkkojulkaisujen ja niiden metatietojen tallentamistapaa, verkkojulkaisujen metatietojen, kieliversioiden ja erikielisten osien linkitystä toisiinsa sekä ryhmittelyä suuremmiksi kokonaisuuksiksi kielen tai aihealueen mukaan. Yksittäinen verkkojulkaisu voidaan tallentaa myös pienempinä, toisiinsa linkitettyinä osina, jos se sisältää paljon erikielisiä osia. Metatiedot ovat monikielisyyden kannalta keskeisessä asemassa, koska niihin tallennetaan kaikki tieto verkkojulkaisun erikielisistä osista ja siinä käytetyistä kielistä.

Esittäminen

Organisaation ja sen sidosryhmien henkilökunnalla saattaa olla hyvinkin suuria eroja muun muassa tietokoneen käyttötaidoissa, joten käyttöliittymän suunnitteluun on kiinnitettävä erityistä huomioita. Jos yrityksen intranetin tai extranetin käyttäjät ovat useista eri kansallisuuksista, on myös kulttuurierot ja erilaiset käytännöt, tavat ja mieltymykset otettava huomioon. Paras käyttöliittymä onkin ulkoasultaan mahdollisimman neutraali ja yksinkertainen sekä toiminnoiltaan helppotajuinen, odotettu ja hallittavissa oleva (Shneiderman, Byrd & Croft, 1998).

Käyttöliittymässä olevat symbolit ja värit on pyrittävä suunnittelemaan niin, että kaikki ymmärtävät ne samalla tavalla. Pelkästään symboleihin ja väreihin perustuvaa elementtien (esimerkiksi toimintopainikkeet) tunnistamista on syytä välttää niiden merkityksen kulttuurisidonnaisuuden takia. Samassa kulttuurissakin saattaa ihmisillä olla hyvinkin erilaisia merkityksiä joillekin symboleille, samoin värit koetaan eri tavalla miellyttävinä.

On myös ihmisryhmiä, jotka tarvitsevat sivuilla olevat ohjeet, toimintopainikkeet ja linkit mieluiten tekstimuodossa eivätkä väreihin perustuvina tai kuvakkeina. Värisokeat käyttäjät eivät saa selvää toimintojen ja ohjeiden merkityksestä tai ymmärtävät ne kokonaan väärin, jos ne perustuvat pelkästään eri värien käyttöön. Teksteinä olevat ohjeet ja kuvien korvaavat selitykset sekä tekstipohjaisen selaimen käyttö tulisi olla

mahdollista, jotta näkövammaiset käyttäjät voisivat käyttää pistekirjoitusta ja siten hyödyntää myös monikielisiä verkkojulkaisuja täysimääräisesti.

URSA-projektin ARCTOS-hakujärjestelmä (ks. kohta 3.4.) on suunniteltu nimenomaan visuaaliseksi käyttöliittymäksi. Toteutustavassa on juuri sellaisia heikkouksia, joista edellä mainittiin:

- Hakusanat näytetään tulosdokumenteissa eri väreillä, joista esimerkiksi värisokea käyttäjä ei saa selvää.
- Käytettyjen värien valinta kullekin hakusanalle saattaa olla käyttäjää harhaanjohtavaa, muun muassa punainen väri vetää huomion puoleensa voimakkaammin kuin vihreä. Punainen/vihreä -väriparin käyttäminen saattaa muutenkin aiheuttaa sekaannusta ja vääriä tulkintoja, koska punainen voi merkitä kieltoa ja vihreä taas lupaa (vrt. esimerkiksi liikennevalot).
- Jos on useita hakusanoja, tulee tulosdokumentista liian kirjava, eikä värien viesti hahmotu kunnolla.

Symbolien käyttäminen linkkeinä sekä eri vaihtoehtoja ja toimintoja osoittamaan on verkkojulkaisuissa hyvin yleistä. Tämä voi olla kuitenkin hyvin harhaanjohtavaa ja aiheuttaa monille käyttäjille sekaannusta tai jopa ärtymystä, koska he ymmärtävät käytetyt symbolit eri tavalla kuin verkkojulkaisun tekijä (kirjoittaja) on ajatellut. Esimerkkinä symbolien käytön harhaanjohtavuudesta ja siihen liittyvistä ongelmista tarkastellaan seuraavassa maiden lippujen kuvien käyttöä ilmaisemaan verkkosivuilla käytettyjä kieliä.

Monikielisillä verkkosivustoilla eri kieliversiot on usein ilmaistu maiden lippujen kuvilla. Lippujen tarkoitus on siis osoittaa, mitä kieliä on käytetty ja samalla ne toimivat linkkeinä näihin kieliversioihin. Tällä käytännöllä on kuitenkin monia huonoja puolia, kuten Pemberton (1998) toteaa:

- Kieliä lasketaan olevan 4000-5000, mutta maita on alle 200, joten lippujen kuvien käyttäminen kieliä edustamaan on huono käytäntö.
- Osalla maista saattaa olla sama nimi kuin enemmistön siellä puhumalla kielellä, esimerkiksi Kreikka, Ranska ja Saksa, kun taas asia ei ole näin esimerkiksi Australiassa tai Yhdysvalloissa.

- Useimmissa maissa puhutaan useampia kuin yhtä kieltä, joten millä tavalla ilmaistaan kunkin maan kielivähemmistöjen käyttämä kieli?
- Englannin kielen kohdalla käytetään usein Yhdysvaltojen tai Ison-Britannian lippua eli Union Jackia. Kuitenkin brittiläinen lippu edustaa useita eri kieliä puhuvia ihmisryhmiä, kuten esimerkiksi walesilaiset ja skotit. Jos taas käytettäisiin Englannin lippua, monetkaan käyttäjistä eivät todennäköisesti tunnistaisi sitä.
- Jos käytetään lippuja edustamaan kieliä, saatetaan usein loukata ihmisten kulttuurisia tai poliittisia tunteita. Lipun myötä he saattavat tulla esimerkiksi yhdistetyksi maahan, johon eivät halua itseään yhdistettävän. Esimerkiksi kansallistuntoinen itävaltalainen tai sveitsiläinen tuskin haluaa käyttää Saksan lippua kuvaamaan äidinkieltään saksaa.

Pembertonin (1998) mukaan onkin turhaa käyttää lippujen kuvakkeita. Kielen nimi on paras tapa ilmaista se. Samoin www-sivujen kieliversioiden osoitteissa ei pidä hänen mukaansa käyttää sellaisia ilmauksia kuin esimerkiksi ".../gb/" tai ".../us/" osoittamaan englannin kieltä, kun on olemassa standardi kaksikirjaimisille kielikoodeille, nimittäin ISO 639 –standardi vuodelta 1988 (ISO, 1999d).

Monikielisen verkkojulkaisun rakenne on suunniteltava niin, että sivuilla esitettävät kieliversiot tarjoavat samat toiminnot. Sivujen rakenteen on oltava mahdollisimman samanlainen, jotta käyttö olisi helppoa ja kätevää. Joskus on huomioitava kielen erityispiirteet, kuten erilainen lukusuunta kuin muissa kielissä, ja rakennettava sivu sen mukaisesti. Oikealta vasemmalle luettavaa tekstiä sisältävän sivun rakenne painottuu sivun oikeaan reunaan, esimerkiksi linkkilistat sijaitsevat oikealla. Sivut ovat siis ikään kuin peilikuvia länsimaisten sivujen tavallisesta rakenteesta. Tällä tavalla rakennettuja ovat esimerkiksi monet arabian- ja hepreankieliset www-sivut. Liitteessä 2 on esimerkkikuvat tällaisista sivuista.

Jos sivustossa on yksi laajempi, julkaisijan pääkielenään käyttämällä kielellä oleva versio ja joukko sitä suppeampia kieliversioita (niin kuin usein www-sivuilla nykyisin on), on tämä seikka ilmoitettava esimerkiksi metatiedoissa käyttäjille. Sivujen rakenne on silti pyrittävä saamaan samankaltaiseksi. Verkkojulkaisu saattaa sisältää vain esimerkiksi abstraktin usealla kielellä muun julkaisun ollessa vain yksikielinen. Kaikkien erikielisten osien täytyy löytyä helposti ja niiden on oltava saatavilla samalla tavalla.

Esittämisessä on myös ratkaistava se, näytetäänkö julkaisun kieliversioita rinnakkain, mikä voi olla käyttäjille tärkeää, esimerkiksi vertailtaessa erikielisiä tekstejä ja käännettäessä niitä tai haettaessa toisilla kielillä olevia erikoisalojen termejä ja sanontoja. Monikielisessä verkkojulkaisussa voi olla yksi kieli niin sanottu pääkieli, joka on esimerkiksi sama kuin yrityksen julkaisutoiminnassaan käyttämä. Muilla kielillä olevat versiot on sitten linkitetty tälle sivulle. Verkkojulkaisulle voidaan myös tehdä aloitussivu, jossa ovat julkaisun metatiedot ja kullekin kieliversiolle oma linkkinsä.

Metatietojen osalta on ratkaistava, näytetäänkö metatiedot käyttäjälle vai ovatko ne vain sivun lähdekoodissa, jolloin niitä lähinnä hyödyntävät vain verkossa toimivat hakukoneet. Jos metatiedot ovat näytöllä käyttäjän katseltavissa, voivat ne olla omalla sivullaan, josta on linkki varsinaiseen julkaisuun ja takaisin. Metatiedot voivat olla myös sijoitettu samalle sivulle kuin julkaisu, joko sen alkuun tai loppuun. Metatietojen julkaisukieli on sama kuin julkaisun pääkieli. Jossakin tapauksessa, jos metatiedot ovat kovin laajat, saatetaan tarvita myös metatiedot useammalla kielellä.

Organisointi

Verkkojulkaisu ja sen metatiedot yhdessä muodostavat dokumentin. Nämä osat voivat sijaita joko erillään, toisin sanoen ne on tallennettu eri paikkaan, tai metatiedot on tallennettu varsinaisen julkaisun yhteyteen eli osaksi sitä. Dokumentit voidaan ryhmitellä suuremmiksi kokonaisuuksiksi aiheiden tai kielten mukaan. Dokumentit muodostavat dokumenttikokoelman, jonka osat voivat fyysisesti sijaita useissa eri paikoissa eli esimerkiksi eri www-palvelimilla (vrt. luvussa 3 esitelty TIPSTER-arkkitehtuuri).

Metatiedot ovat keskeinen asia verkkojulkaisujen organisoinnissa. Dublin Core (DCMI, 1999) on erityisesti verkkojulkaisemiseen suunniteltu metatietomääritys, jossa on 15 elementtiä. Yksi niistä on Language-elementti, jolla ilmaistaan julkaisussa käytetyt kielet kaksikirjaimisena koodina, joka on ISO 639 –standardin (ISO, 1999d) mukainen. Language-elementin tulisi olla yhteensopiva RFC 1766 –määrityksen kanssa, missä on määritelty kielitunnisteen (language tag) esitystapa, silloin kun sen tarkoitus on ilmaista informaatio-objektissa käytetty kieli (ks. Alvestrand, 1995). Esimerkissä 14 on esitetty

Dublin Core –metatietoelementin syntaksi ja kielen ilmaiseminen Language-elementillä, kun se on koodattu HTML:ään (Hansen, 1999).

Esimerkki 14. Kielen ilmaiseminen Dublin Core –metatietoelementillä.

Syntaksi:
<META NAME = "DC.ElementName" CONTENT = "Value">
Käytetyn kielen ilmaiseminen:
<META NAME = "DC.Language" CONTENT = "fi">

Metatietojen avulla käyttäjä saa tiedot julkaisussa käytetyistä kielistä sekä löytää haluamansa kieliversiot ja erikieliset osat samoin kuin pienimmätkin erikieliset elementit, joita julkaisussa saattaa olla. Tällaisia käyttäjää kiinnostavia pieniä osia voivat olla sanonnat ja yksittäiset termit ja sanat, jotka esitetään aina vain yhdellä tietyllä kielellä eikä koskaan käännetä. Esimerkkejä sellaisista ovat muun muassa monet latinankieliset sanonnat, kuten force majeure ja de facto, joita käytetään teksteissä sellaisinaan.

Monikieliset verkkojulkaisut voidaan organisoida seuraavilla tavoilla riippuen siitä, mitkä osat julkaisusta ovat monikielisiä:

1. Verkkojulkaisusta on kieliversioita eli koko julkaisu on tallennettu useammalla kuin yhdellä kielellä. Tällöin kieliversiot muodostavat kokonaisuuden, jolla voi olla yhdet metatiedot, jotka koskevat yhteisesti kaikkia kieliversioita, ja kaikki julkaisussa käytetyt kielet ilmaistaan yhdessä metatietoelementissä. Toinen mahdollisuus on se, että jokainen kieliversio sisältää omat metatietonsa. Metatiedot voidaan tallentaa osaksi verkkojulkaisua tai sitten ne voivat sijaita erillään itse julkaisusta. Tämä on tarpeen ehkä silloin, kun metatietoja on paljon eikä niitä kätevästi voi ilmaista vain yhdellä kielellä.
2. Jos verkkojulkaisusta vain jokin osa on useammalla kuin yhdellä kielellä tai se sisältää erikielisiä pieniä osia aivan yksittäisiin sanoihin saakka, tulee metatietojen sisältää tarkat tiedot niistä. Julkaisun erikieliset osat (laajemmat kuin yksittäiset sanat, sanonnat tai lauseet) voivat olla tallennettu myös erillisinä ja olla linkitetty toisiinsa. Metatiedoissa tulee olla tiedot linkityksestä.

Metatietojen avulla voidaan siis hallita monikielisten verkkojulkaisujen saatavuus, löydettävyys ja uudelleen käytettävyys. Saatavuudella tarkoitetaan tässä sitä, että ne verkkojulkaisut ja metatiedot, joita käyttäjä kulloinkin tarvitsee ja joita hänellä on oikeus käyttää, ovat vaivattomasti ja aina tarvittaessa hänen saatavissaan. Löydettävyys tarkoittaa sitä, että haun tulos on mahdollisimman kattava, toisin sanoen tärkeitä, haun kriteerit täyttäviä verkkojulkaisuja ei jää tulosjoukon ulkopuolelle. Lisäksi verkkojulkaisun sisältöön täytyy voida tehdä hakuja, joista saadaan oikeanlaisia tuloksia kaikilla tuetuilla kielillä. Uudelleen käytettävyys tukee monikielisten verkkojulkaisujen tuottamista. Lähdetekstipankeissa voi olla tallennettuna valmiita, eri kielillä tuotettuja tekstejä ja tekstin osia (Hartley & Paris, 1997). Samoin erikielisten termien ja sanontojen tallentaminen termipankkeihin (ks. Danish, 1998) helpottaa tuottamista ja lisää uudelleen käytettävyyttä.

Dublin Core –määrityksessä on kielten ilmaisemiseen vain yksi elementti, jossa ovat käytettyjen kielten koodit. Lisäksi monikielisten verkkojulkaisujen metatiedoissa tarvitaan usein kielten osalta useampia elementtejä, jotka sisältävät esimerkiksi seuraavat tiedot:

- Onko koko dokumentista kieliversioita ja millä kielillä niitä on?
- Onko joitakin osia (esimerkiksi tiivistelmä, johdanto, avainsanat) useammalla kielellä, mitä osia ja millä kielillä?
- Tiedot tuottamisessa käytetyistä lähdeteksti- ja termipankeista.

4.2. Tuottaminen ja ylläpito

Hartleyn ja Parisin (1997) mukaan monikielisiä dokumentteja on perinteisesti tuotettu sarjaluonteisessa prosessissa, jolloin valmis ja vahvistettu lähdedokumentti on käännetty muille halutuille kielille. Tuottamisprosessissa keskeisellä sijalla on ollut kääntäminen ja keskeisenä henkilönä kielenkääntäjä. Koska tuotteet halutaan nopeasti markkinoille, on niihin liittyvien dokumenttien (esimerkiksi käyttö- ja huolto-ohjeet) käännöstyössä otettu käyttöön monenlaisia apuvälineitä. Hartley ja Paris (1997) jakavat nämä työkalut laajasti ottaen kolmeen ryhmään:

- Ensimmäiseen ryhmään kuuluvat online-sanastot sekä termien korvaaminen automaattisesti toisen kielen termeillä, mitkä toimivat kääntäjien apuvälineinä käännöstyön ollessa pääosin heidän vastuullaan.
- Toiseen ryhmään kuuluvat sellaiset työkalut, jotka vapauttavat kääntäjät suuresta osasta käännöstyötä. Kääntäjien tehtävä on tällöin lähinnä varmistaa, että lopputulos on yhtenevä käännöstyön lähdeaineiston kanssa. Tällaisia työkaluja ovat esimerkiksi sääntöihin perustuvat konekäännösjärjestelmät ja esimerkkeihin perustuvat järjestelmät eli käännösmuistit.
- Kolmannen ryhmän työkalujen avulla pyritään muodostamaan tekstejä yhdellä kielellä siten, että vastaavat tekstit muilla kielillä voidaan generoida ilman kääntäjien puuttumista prosessiin. Nämä työkalut on tarkoitettu siis tukemaan yksikielisiä henkilöitä eikä kääntäjiä. Kun tekstit generoidaan tietokannoista, jotka sisältävät samaa merkitseviä katkelmia erikielisistä teksteistä, merkitsee se sitä, että järjestelmä tukee kääntämisen sijasta monikielistä tekstin tuottamista. Erikieliset versiot tuotetaan siis rinnakkain eli samanaikaisesti eikä peräkkäin, niin kuin perinteisillä menetelmillä.

Monikielisten dokumenttien tuottamisessa pyritään yhä enemmän peräkkäisestä tuottamisesta rinnakkaiseen tuottamiseen. Hartleyn ja Parisin (1997) mukaan rinnakkaisessa tuottamisessa on tunnustettu se, että samaan tarkoitukseen eri kielillä kirjoitetuissa teksteissä saattaa olla erilainen sisältö ja rakenne. Rinnakkainen tuottaminen tukee lokalisaatiota ja lähtee lukijan tarpeista, mikä on tärkeää esimerkiksi käyttöohjeiden ollessa kyseessä. Tuotettaessa dokumentteja rinnakkain eli samanaikaisesti saadaan tuote ja sen käyttöohjeet markkinoille huomattavasti nopeammin kuin aikaisemmin, jolloin käännettiin alkuperäinen ohje kaikille halutuille kielille peräkkäisessä prosessissa.

Luvussa 3 esitellyt TIPSTER:iin pohjautuvat sovellukset painottuvat automaattisen kääntämisen puolelle, mikä antaa melko huonoja ja karkeita tulosdokumentteja. Konekääntämisen sijaan tulisikin keskittyä tuottamistapoihin, joissa kääntäjät ovat vastuussa monikielisten dokumenttien tuottamisesta ja heillä on apunaan tehokkaita monikieliseen tuottamiseen kehitettyjä ohjelmia, lähdetekstikantoja ja termipankkeja.

Verkkojulkaisujen sisällön laatuun, esimerkiksi käytettyyn sanastoon ja terminologiaan, on kiinnitettävä erityistä huomiota. Erikoissanastot ovat monille käyttäjille outoja ja siksi heidän verkossa tekemänsä kyselyt epäonnistuvat (ei saada mitään tuloksia tai sitten tulokset ovat aivan muuta, mitä käyttäjä haki), kuten Kambil ja Ginsburg (1998) toteavat. Kambilin ja Ginsburgin (1998) mukaan tiedon laatua voidaan parantaa soveltamalla standardeja tiedon esittämisessä. He pitävät XML-standardiehdotelmaa lupaavana vaihtoehtona, koska sillä voidaan merkata tekstistä osia, joihin voidaan kohdistaa hakuja.

Monikielisten verkkojulkaisujen tuottamisessa on otettava huomioon seuraavat seikat:

- Koska monikieliset verkkojulkaisut saattavat sisältää monia eri kieliä, jotka käyttävät erilaisia merkistöjä, tarvitaan sellainen koodausjärjestelmä, jonka kapasiteetti riittää periaatteessa kaikkien kielten merkistöjen koodaamiseen yksikäsitteisesti. Luvussa 2 esitelty ISO 10646 –standardin kanssa yhteensopiva Unicode on tällainen koodausjärjestelmä.
- Näytöllä käytettävät fontit olisi pyrittävä saamaan mahdollisimman yleispäteviksi eli laite- ja ohjelmistoriippumattomiksi.
- Verkkoa voidaan käyttää myös materiaalin jakelukanavana, jolloin dokumentit voivat olla pitkiäkin. Tällöin ne on muotoiltava tulostukseen sopiviksi ja niissä käytettyjen fonttien olisi oltava mahdollisimman yleisiä eli tavallisten tulostimien tunnistettavissa.
- Dokumentin käyttötarkoitus (esimerkiksi suoritetaanko hakuja) vaikuttaa myös tuottamiseen. Verkkojulkaisu voidaan tuottaa rakenteisessa muodossa eli pohjautuen XML-standardiin, jolloin siitä voidaan hakea kätevästi yksittäisiäkin tietoja, toisin kuin HTML-muotoisesta verkkojulkaisusta. Erikieliset osat voidaan merkata yksityiskohtaisesti XML:n avulla (ks. Luku 2), jolloin ne ovat hauissa löydettävissä.
- Verkkojulkaisussa käytettyjen kaikkien kielien kirjaaminen metatietoihin, esimerkiksi Dublin Core –määrityksen mukaisesti.
- Mahdollinen dokumenttien käännöstyö silloin, kun jostakin dokumentista tarvitaan uudella kielellä oleva versio.
- Valmiiden dokumenttien editointi esimerkiksi rakenteisen tekstin editoreilla (muuntaminen XML-muotoon) tai HTML-editoreilla (riippuen käyttötarkoituksesta).
- Tuottamisessa voidaan käyttää valmiita erikielisiä lähdetekstejä (Hartley & Paris, 1997) ja termipankkeja, joille on oltava myös omat metatietonsa.

Tuottamisessa dokumentti voidaan koostaa joko kokonaan tai osittain valmiista lähdetekstipankista saatavista tekstikappaleista tai sitten kirjoitetaan annetusta aineistosta. Termipankista saadaan erityisalojen termejä niillä kielillä, joita järjestelmä tukee. Tällaisia termipankkiin tallennettuja erikielisiä sanoja ja sanontoja voivat olla esimerkiksi tuotteiden osat tai käyttöohjeissa tarvittavat usein toistuvat (lyhyet, muutaman sanan pituiset) ilmaisut (ks. Danish, 1998). Metatietomäärityksen pohjalta kirjoittaja laatii uuden dokumentin metatiedot. Kirjoitettu verkkojulkaisu ja sen metatiedot konvertoidaan XML-editorilla XML- tai HTML-muotoon riippuen lähinnä dokumentin arkistointitarpeesta ja säilytysajasta.

Monikielisten verkkojulkaisujen ylläpidon osalta käsitellään tässä lyhyesti niiden päivittämistä ja poistamista. Päivitettäessä julkaisua on huolehdittava siitä, että kaikki kieliversiot tai julkaisun monikieliset osat päivitetään samalla tavalla. Julkaisun päivittäminen voi tarkoittaa seuraavia asioita:

- Kieliversioiden lisääminen tai poistaminen.
- Julkaisun monikielisen osan lisääminen tai poistaminen.
- Tekstissä olevien vieraskielisten termien ja sanontojen lisääminen tai poistaminen.
- Julkaisun metatietojen päivitys tehtäessä joku edellä mainituista muutoksista.
- Julkaisun ja sen metatietojen, kieliversioiden välisen tai julkaisun eri osien keskinäisen linkityksen pitäminen ajan tasalla muutoksia tehtäessä.

Monikielisen julkaisun, sen kieliversioiden tai osien poistaminen voi tarkoittaa ensinnäkin niiden siirtämistä ikään kuin arkistoon, jolloin ne eivät ole käytössä, mutta ovat kuitenkin tallessa. Tarpeen vaatiessa ne voidaan taas ottaa aktiiviseen käyttöön. Metatietoihin voidaan liittää tieto julkaisun ja sen osien arkistoinnista tai ottamisesta uudelleen käyttöön. Toiseksi julkaisun poistaminen voi tarkoittaa sen fyysistä poistamista, jolloin sitä ei voida enää palauttaa käyttöön. Koko julkaisun poistaminen tarkoittaa sitä, että sen kaikki kieliversiot poistetaan. Metatiedot on hyvä säilyttää julkaisun poistamisen jälkeenkin. Niihin voidaan kirjata poistamista koskevat tiedot, esimerkiksi kuka poisti julkaisun tai miksi ja milloin se poistettiin.

4.3. Tiedon hakeminen monikielisessä ympäristössä

Metatiedot muodostavat perustan tietojen löytymiselle suoritettaessa hakuja monikielisistä verkkojulkaisuista. Tiedon hakeminen voi tapahtua selailemalla verkossa olevia julkaisuja tai muodostamalla kyselylausekkeita, joilla pyritään määrittelemään mahdollisimman tarkasti, millaista tietoa sisältäviä ja millä kielellä olevia julkaisuja haetaan.

Selailuhaku

Hirashiman, Hachiyan, Kashiharan ja Toyodan (1997) mukaan aloittavat selailun monesti hyvin erilaisista, laajasti vaihtelevista syistä, jolloin mielenkiinnon kohteet myös usein muuttuvat. Useimmiten selailu on tehoton tapa löytää mitään, koska siinä helposti eksyy asiasta ja myöskin kadottaa jossakin vaiheessa kulkemansa polun. Selailun tehottomuutta lisää usein linkkivaihtojen liian suuri määrä sekä se, että sisällön kirjoittaja on määritellyt hypertekstin linkit omasta näkökulmastaan. Tämä poikkeaa yleensä käyttäjän näkökulmasta ja tarpeista, niin kuin Hirashima ym. (1997) toteavat.

(Hirashima ym., 1997) esittävät ratkaisua, jossa hypertekstin solmuihin liitetään joukko avainsanoja eli indeksitermejä. Samoja indeksitermejä sisältävät solmut on linkitetty toisiinsa, jolloin käyttäjä voi selaillen hakea tietoa kiinnostuksen kohteistaan eksymättä pois kohteenaan olevalta aihealueelta. Hypertekstin solmut voitaisiin myös linkittää käytetyn kielen perusteella, jolloin selailuhaun tuloksena saataisiin tietynkielisiä, samaa aihetta käsitteleviä verkkojulkaisuja.

Kyselyhaku

Tiedonhakua verkossa voidaan kuvata nelivaiheisella etsinnän viitekehyksellä (Shneiderman ym., 1998, vrt. Marchionini, 1995, s. 50). Luvussa 3 esitellyn Ursa-projektin Arctos-hakujärjestelmän suunnittelussa on huomioitu tämän viitekehyksen esittämät vaiheet:

- kyselyn muotoilu ennen kuin käyttäjä aloittaa etsinnän
- toiminta eli kyselyn toteuttaminen

- tulosten tarkastelu eli mitä käyttäjä näkee kyselyn tuloksena
- jalostus, parannus (refinement) eli mitä tapahtuu tulosten tarkastelun jälkeen ennen palaamista takaisin kyselyn muotoiluun

Tämän tutkielman kannalta keskeinen asia monikielisten verkkojulkaisujen käytössä on *monikielinen tiedonhaku* (*Multilingual Information Retrieval, MLIR*), jolle Hull ja Grefenstette (1996) ovat esittäneet muun muassa seuraavat määritelmät:

- tiedonhaku *rinnakkaisten dokumenttien kokoelmasta* (kokoelma, jossa erikieliset dokumentit ovat kokoelman erillisiä ja itsenäisiä osia) tai monikielisestä dokumenttikokoelmasta, missä haetaan vain kyselykielellä olevat dokumentit
- tiedonhaku yksikielisestä dokumenttikokoelmasta, johon voidaan tehdä kyselyjä monella eri kielellä
- tiedonhaku monikielisestä dokumenttikokoelmasta, johon voidaan tehdä kyselyjä monella eri kielellä
- tiedonhaku monikielisistä dokumenteista ts. dokumenteista, joissa voi esiintyä erikielisiä osia

Niin kuin edellä esitetyistä määritelmistä voidaan päätellä, monikielisessä tiedonhaussa on siis hyvin usein pystyttävä kääntämään kyselylauseke toisille kielille. Kun kysely käännetään, on tarpeen tarkastella sitä ja kenties uudelleen muotoilla niin, että se vastaa käyttäjän alun perin tarkoittamaa hakulauseketta. Kyselyn kääntämisessä voitaisiin hyödyntää samaa termipankkia, jota käytetään dokumenttien tuottamisvaiheessa erikoisalojen termien ja sanontojen erikielisten versioiden saantipaikkana.

Kyselylauseketta käännettäessä on myös otettava huomioon lauseen rakenne ja se, että eri kielillä sama asia ilmaistaan aivan erilaisella rakenteella. Tällöin on pyrittävä ryhmittelemään lauserakenteet toisiaan vastaaviksi, kuten Ballim, Coray, Linden ja Vanoirbeek (1998) artikkelissaan osoittavat. Monimutkainen kyselylauseke on ehkä tarpeen myös jakaa osalausekkeisiin (ks. Chidlovskii & Borghoff, 1998) ennen kääntämistä, jotta haun tuloksena todella saadaan sitä, mitä käyttäjä haluaa. Kyselyn kääntäminen ja mahdollinen uudelleen muotoilu käännöksen jälkeen on siis lisättävä Shneidermanin ym. (1998) esittämään viitekehykseen ennen vaihetta 2 eli kyselyn aloittaminen.

Haun tulosten esittäminen

Haun tulokset voidaan esimerkiksi esittää linkkilistana verkkojulkaisuihin, jotka sisältävät hakutermejä tai jotka sopivat hakulausekkeen määrittämään joukkoon. Näyttö voidaan jakaa kahteen osaan, jolloin toisessa reunassa on linkkilista ja toisessa reunassa on isompi ikkuna, jossa näkyy listasta valitun tulosdokumentin sisältö. URSA-projektin ARCTOS-demonstraatiossa (ks. kohta 3.4) on toteutettu tällainen haun tulosten esittäminen, mikä tuntuu kätevältä. Jos käyttäjä haluaa tarkastella erikielisiä julkaisuja rinnakkain, olisi niitä voitava esittää omissa ikkunoissaan vähintään pareittain.

Jos alkuperäinen hakulauseke oli eri kielellä kuin tulosdokumentit, on tarjottava myös tulosten käännösmahdollisuus käyttäjän haluamalle kielelle. Useimmat nykyiset verkkokäännösohjelmat antavat tuloksena melko karkean käännöksen, mutta Zajacin ja Casperin (1997) mukaan se useimmiten riittää antamaan käyttäjälle yleiskuvan julkaisun sisällöstä. Toisaalta esimerkiksi Watters ja Patel (1999) varoittavat käyttäjiä luottamasta paljonkaan näihin käännöksiin, koska muun muassa monitulkintaisten eli polysemisten (ks. esimerkiksi Harakka, 1999) sanojen käännökset voivat aiheuttaa sekaannusta ja hämmennystä. Tosin myös Watters ja Patel (1999) pitävät nykyisten käännösohjelmien tuotoksia rajoitetussa ympäristössä hyväksyttävinä ja riittävinä.

5. YHTEENVETO

Monikielisten dokumenttien julkaiseminen verkoissa on tullut yhä tärkeämmäksi Internetin käytön yleistyessä. Lisäksi monet yritykset, julkishallinnon laitokset ja yhteisöt julkaisevat ja jakelevat dokumentteja intranet- ja extranetverkkojen kautta. Koska yritysten toimipisteitä ja yhteistyökumppaneita voi olla monissa eri maissa, täytyy pystyä julkaisemaan myös monikielistä materiaalia verkoissa. Tässä tutkielmassa on kartoitettu monikielisten, verkossa julkaistavien dokumenttien hallintaan liittyviä ongelmia. Koska tarkasteltava alue on laaja ja sisältää monenlaisia ongelmia, on tämä esitys hyvin yleisellä tasolla eikä siinä ole voitu yksityiskohtaisesti tarkastella mitään osaongelmaa.

Aluksi on tarkasteltu Unicode-standardia, joka mahdollistaa periaatteessa kaikkien kirjoitetussa muodossa olevien kielten koodaamisen digitaaliseen muotoon. Dokumenttien sisällön rakenteistamiseen liittyen on esitelty SGML-metakieli sekä sen sovellukset HTML, XML ja TEI.

HTML on SGML-sovellus, joka on suunniteltu www-julkaisemiseen. XML on SGML:n osajoukko, joka on erityisesti kehitetty verkossa tapahtuvaa julkaisemista varten, ja sillä voidaan ilmaista dokumenttien looginen rakenne, mihin HTML ei anna mahdollisuuksia. XML:n ja myös uusimman HTML-version (versio 4.0) merkistöksi on otettu Unicode, joten niillä voidaan esittää monikielisiä dokumentteja.

TEI-standardi on kehitetty monivuotisessa projektissa SGML:n pohjalta. Se on tarkoitettu humanististen tekstien koodaamiseen ja siinä on jätetty käyttäjille melko paljon joustovaraa, joten sitä voidaan soveltaa erilaisten tekstien rakenteen kuvaamiseen. Kritiikkiä onkin kohdistettu sen liialliseen joustavuuteen ja laajuuteen, koska se asettaa käyttäjille melko suuria vaatimuksia.

Tutkielmassa on tarkasteltu jo olemassa olevia tai kehitteillä olevia monikielisten dokumenttien hallintajärjestelmiä. Tarkastelun lähtökohtana on ollut TIPSTER-arkkitehtuuri, joka on kehitetty monivuotisessa yhdysvaltalaisessa projektissa. TIPSTER:in pohjalta kehitettyjä sovelluksia TEMPLE, GATE ja URSA on esitelty lyhyesti sekä tarkasteltu niiden soveltuvuutta monikielisten dokumenttien hallintaan.

Näiden tarkastelujen sekä esiteltyjen standardien pohjalta on kuvattu monikielisten rakenteisten dokumenttien hallintaan liittyviä osa-alueita ja ongelmakohtia. Kuvauksen yhteydessä on samalla arvioitu lyhyesti joitakin TIPSTER:in pohjalta kehitettyjen sovellusten piirteitä.

Tämä tutkielman tuloksena on yleiskuvaus monikielisten verkkojulkaisujen hallintaan liittyvistä seikoista. Alue on hyvin laaja ja siinä on monia osa-alueita, jotka soveltuvat jatkotutkimusaiheiksi. Yksi hyvin tärkeä osa-alue on monikielisten kyselyiden toteutuksessa huomioon otettavien seikkojen tarkempi selvittäminen. Verkkojulkaisujen ja niiden metatietojen organisointi ja rakenteen suunnittelu on toinen tärkeä ja tarkempaa tutkimusta vaativa alue. Myös monikielisten julkaisujen esittämiseen näytöllä liittyy useita lisäselvitystä vaativia aihealueita, kuten esimerkiksi kulttuurista ja kielestä riippuvat seikat ja yleensä sivujen rakenteen suunnittelu nimenomaan monikielisyyden näkökulmasta.

LÄHDELUETTELO

Alvestrand, H. T. 1995. Tags for the Identification of Languages. [online]. [Viitattu 8.6.1999]. Saatavilla www-muodossa <URL: http://www.rfc-editor.org/rfc/rfc1766.txt>.

Architecture Committee for the TIPSTER Text Phase II Program. 1996a. TIPSTER Text Phase II Architecture Concept. [online]. [Viitattu 12.4.1999]. Saatavilla www-muodossa <URL: http://www.nist.gov/itl/div894/894.02/related_projects/tipster/docs/con112.doc>.

Architecture Committee for the TIPSTER Text Phase II Program. 1996b. TIPSTER Text Phase II Architecture Requirements. [online]. [Viitattu 12.4.1999]. Saatavilla www-muodossa <URL: http://www.nist.gov/itl/div894/894.02/related_projects/tipster/docs/req201.doc>.

Arctos: Interactive Multilingual Search with Ursa. 1998. [online]. [Viitattu 17.3.1999]. Saatavilla www-muodossa <URL: http://messene.nmsu.edu/ursa/arctos/>.

Attig, J., Klimczyk, L. & Mangin J. Understanding MARC Bibliographic. [online]. Library of Congress, 1999. [Viitattu 4.2.1999]. Saatavilla www-muodossa <URL: http://lcweb.loc.gov/marc/umb/umbhome.html>.

Ballim A., Coray G., Linden A., Vanoirbeek C., The Use of Automatic Alignment on Structured Multilingual Documents. Teoksessa: Hersch R. D., André J., Brown H. (toim.), Electronic Publishing, Artistic Imaging, and Digital Typography. 7th International Conference on Electronic Publishing, EP'98. Held Jointly with the 4th International Conference on Raster Imaging and Digital Typography, RIDT'98, St. Malo, France, March/April 1998. Proceedings.

Bettels, J. & Bishop, F. A. Unicode: A Universal Character Code. [online]. Digital Equipment Corporation, 1993. [Viitattu 8.1.1999]. Saatavilla www-muodossa <URL: http://www.digital.com/info/DTJB02/DTJB02SC.TXT>.

Boualem M., Harié S., MtScript: A Multilingual Text Editor. Computers and the Humanities, Vol. 31, No. 2, 1997, 135-151.

Bray T., Paoli J. & Sperberg-McQueen C. M.(toim.): Extensible Markup Language (XML) 1.0. W3C Recommendation 10-February-1998. [online]. W3C World Wide Web Consortium, 1998. [Viitattu 21.9.1998]. Saatavilla www-muodossa <URL: http://www.w3.org/TR/1998/REC-xml-19980210>.

Burnard, L. & Sperberg-McQueen, C. M. 1993. Living with the Guidelines: An Introduction to TEI Tagging. [online]. [Viitattu 25.09.1998]. Saatavilla www-muodossa <URL: http://www.sgmlu.com/documents/tei-guidelines.html>.

Carrasco Benitez, M. T. 1996. WInter (Web Internationalization & Multilinguism) Document. [online]. [Viitattu 9.4.1999]. Saatavilla www-muodossa <URL: http://www.w3.org/International/tomas.carrasco.benitez.htm>.

Chidlovskii, B., Borghoff, U. M., Query Translation for Distributed Information Gathering on the Web. Proceedings of the International Database Engineering and Applications Symposium. July, 1998, UK.

Coleman J., Willis D., SGML as a Framework for Digital Preservation and Access. The Commission on Preservation and Access, A Program of the Council on Library and Information Resources, Washington, DC, July 1997.

CRL Current Research. 1999. [online]. [Viitattu 8.3.1999]. Saatavilla www-muodossa <URL: http://crl.nmsu.edu/New%20Research/research.htm>.

Cunningham, H., Gaizauskas R. & Wilks, Y. A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D. [online]. Institute for Language, Speech and Hearing (ILASH), and the Department of Computer Science, University of Sheffield, UK, 1995. [Viitattu 23.3.1999]. Saatavilla www-muodossa <URL: ftp://ftp.dcs.shef.ac.uk/home/hamish/gate_rpt.ps>.

Cunningham, H., Peters W., McCauley C., Bontcheva K. & Wilks, Y. Uniform Language Resource Access and Distribution. [online]. Department of Computer Science, University of Sheffield, UK, 1998. [Viitattu 6.4.1999]. Saatavilla www-muodossa <URL: ftp://ftp.dcs.shef.ac.uk/home/hamish/auto_papers/Cun98a.ps>.

Cunningham, H., Wilks, Y. & Gaizauskas R. 1996. GATE – a General Architecture for Text Engineering. [online]. [Viitattu 6.4.1999]. Saatavilla www-muodossa <URL: ftp://ftp.dcs.shef.ac.uk/home/hamish/gate_coling96.ps>.

Danish, S., Building Database-driven Electronic Catalogs. SIGMOD Record, Vol. 27, No. 4, 1998, 15-20.

Davis, M. 1994. UCS Transformation Format 16 (UTF-16). [online]. WG2 Project, 1994. [Viitattu 28.9.1998]. Saatavilla www-muodossa <URL: http://www.stonehand.com/unicode/standard/wg2n1035.html>.

Davis, M. W. & Ogden, W. C. Design, Implementation and User's Guide to URSA, the UNICODE Retrieval System Architecture. [online]. New Mexico State University, Computing Research Laboratory, 1997. [Viitattu 17.12.1998]. Saatavilla www-muodossa <URL: http://crl.NMSU.Edu/Research/Projects/tipster/ursa/Papers/ursa2.ps>.

DCMI, Dublin Core Metadata Initiative. The Dublin Core: A Simple Content Description Model for Electronic Resources. 1999. [online]. [Viitattu 8.6.1999]. Saatavilla www-muodossa <URL: http://purl.org/dc/>.

Deach, S. (toim.) Extensible Stylesheet Language (XSL) Specification. W3C Working Draft 21 Apr 1999. [online]. [Viitattu 15.6.1999]. Saatavilla www-muodossa <URL: http://www.w3.org/TR/WD-xsl/>.

Flanders J., Bauman S., Caton P., Cournane, M., The Proper Names and the Improper. Applying the TEI to the Classification of Proper Nouns. Computers and the Humanities, Vol. 31, No. 4, 1997, 285-300.

Gaizauskas, R., Rodgers, P., Cunningham, H., Humphreys, K. & Robertson, S. GATE User Guide. [online]. Institute for Language, Speech and Hearing (ILASH), and the Department of Computer Science, University of Sheffield, UK, 1998. [Viitattu 23.3.1999]. Saatavilla www-muodossa <URL: http://www.dcs.shef.ac.uk/research/groups/nlp/gate/system_docs/user_guide/main/main.html >.

GATE Architecture Overview. 1998. [online]. [Viitattu 10.11.1998]. Saatavilla www-muodossa <URL: http://www.dcs.shef.ac.uk/research/groups/nlp/gate/architecture.html>.

Goldfarb C. F., The SGML Handbook. Clarendon, Oxford, 1990.

Goldfarb C. F., Prescod P., The XML Handbook. Prentice-Hall, Upper Saddle River, NJ, USA, 1998.

Grishman, R. TIPSTER Phase II Architecture, The Tinman Architecture. [online]. New York University, 1995. [Viitattu 15.3.1999]. Saatavilla www-muodossa <URL: http://cs.nyu.edu/pub/nlp/tipster/12monthsSlides.ps>.

Grishman, R. (toim.). TIPSTER Text Architecture Design, Version 3.1. [online]. New York University, 1998. [Viitattu 23.3.1999]. Saatavilla www-muodossa <URL: http://www.nist.gov/itl/div894/894.02/related_projects/tipster/docs/arch31.doc>.

Hansen, P. User Guidelines for Dublin Core Creation. [online]. Nordic Metadata Project, 1999. [Viitattu 9.6.1999]. Saatavilla www-muodossa <URL: http://www.sics.se/~preben/DC/DC_guide.html>.

Harakka, T. 1999. Terminologisen tutkimuksen perusteet 1999. [online]. [Viitattu 10.6.1999]. Saatavilla www-muodossa <URL: http://www.uwasa.fi/~tepa/Termino1.htm>.

Hartley A., Paris C., Multilingual Document Production From Support for Translating to Support for Authoring. Machine Translation, Vol. 12, No. 1 / 2, 1997, 109-129.

Hirashima T., Hachiya K., Kashihara A., Toyoda J., Information Filtering Using User's Context on Browsing in Hypertext. User Modeling and User-Adapted Interaction, Vol. 7, No. 4, 1997, 239-256.

How to browse the arabic language based web pages? [online]. 1999. [Viitattu 23.6.1999]. Saatavilla www-muodossa <URL: http://www.salafi.net/arabic.html>.

Hull, D. A., Grefenstette, G. 1996. Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. Teoksessa SIGIR '96 : proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in InformationRetrieval, August 18-22, 1996, Zurich, Switzerland, 49-57. New York (NY): ACM.

Ide, N. & Véronis, J. MULTEXT (Multilingual Tools and Corpora). [online]. Laboratoire Parole et Langage, CNRS & Université de Provence, France, 1994. [Viitattu 6.4.1999]. Saatavilla www-muodossa <URL: http://www.up.univ-mrs.fr/~veronis/pdf/1994multext.pdf>.

ISO (International Organization for Standardization). ISO 14962:1997. [online]. 1999a. [Viitattu 19.5.1999]. Saatavilla www-muodossa <URL: http://www.iso.ch/cate/d26096.html>.

ISO (International Organization for Standardization). ISO/IEC 646:1991. [online]. 1999b. [Viitattu 19.5.1999]. Saatavilla www-muodossa <URL: http://www.iso.ch/cate/d4777.html>.

ISO (International Organization for Standardization). ISO/IEC 10646-1:1993. [online]. 1999c. [Viitattu 27.5.1999]. Saatavilla www-muodossa <URL: http://www.iso.ch/cate/d18741.html>.

ISO (International Organization for Standardization). ISO 639:1988. [online]. 1999d. [Viitattu 7.6.1999]. Saatavilla www-muodossa <URL: http://www.iso.ch/cate/d4766.html>.

Jaeger N., Devillers C., Cruickshank G., Cencioni R., The EUROTRA front-end: Current state and perspectives. Teoksessa: Lichnerowicz, A. (toim.) INTELLIGENT TEXT AND IMAGE HANDLING. Proceedings of a Conference on Intelligent Text and Image Handling "RIAO 91", Barcelona, Spain 2-5 April 1991. Elsevier Amsterdam, 1991, 455-474.

Johnson, C. D. 1997. ISO-10646 Concept Dictionary. [online]. [Viitattu 8.1.1999]. Saatavilla www-muodossa <URL: http://cns-web.bu.edu/pub/djohnson/web_files/i18n/ISO-10646.html>.

Järnefors, O. A short overview of ISO/IEC 10646 and Unicode. [online]. 1996. [Viitattu 1.6.1999]. Saatavilla www-muodossa <URL: http://www.nada.kth.se/i18n/ucs/unicode-iso10646-oview.html>.

Kambil A., Ginsburg M., Public Access Web Information Systems: Lessons from the Internet EDGAR Project. Communications of the ACM, Vol. 41, No. 7, 1998, 91-97.

Kang, B., Modifying the TEI DTD: The Case of Korean Dictionaries. Computers and the Humanities, Vol. 31, No. 5, 1997, 433-449.

Keski-Suomen Viro-seura ry. 1997. [online]. [Viitattu 24.5.1999]. Saatavilla www-muodossa <URL: http://www.jyu.fi/yhd/Viro-seura/eesti-selts.html>.

Leinonen, A. 1998. XML:n käyttö verkkosovelluksissa. [online]. EVA-projektin raportti. [Viitattu 28.1.1999]. Saatavilla www-muodossa <URL: http://linnea.helsinki.fi/eva/saanle.html>.

Lyytikäinen V., Päivärinta T., Salminen A., Tiitinen P., Valtion talousarvioon liittyvien asiakirjojen rakenteistaminen. RASKE-projektin raportti, 1997.

Mah C., Flanders J., Lavagnino J., Some Problems of TEI Markup and Early Printed Books. Computers and the Humanities, Vol. 31, No. 1, 1997, 31-46.

Maler, E. & DeRose, S. (toim.). XML Linking Language (XLink). World Wide Web Consortium Working Draft 3-March-1998.). [online]. [Viitattu 15.6.1999]. Saatavilla www-muodossa <URL: http://www.w3.org/TR/WD-xlink>.

Marchionini, G. 1995. Information Seeking in Electronic Environments. Cambridge Series on Human Computer Interaction 9. Cambridge: Cambridge University Press.

Multilingual Entity Task (MET) Conference. [online]. [Viitattu 16.12.1998]. Saatavilla www-muodossa <URL: http://www.tipster.org/met/met.htm>.

MUNDIAL NET SEARCH III. 1998. [online]. [Viitattu 17.3.1999]. Saatavilla www-muodossa <URL: http://crl.nmsu.edu/Research/Projects/tipster/ursa/Mundial/mundial.html>.

Nielsen J., User Interface Directions for the Web. Communications of the ACM, Vol. 42, No. 1, 1999, 65-72.

Noerr, P. L. 1995. Character sets and UNICODE. [online]. Information Management and Engineering Ltd, London. [Viitattu 26.11.1998]. Saatavilla www-muodossa <URL: http://www.ua.ac.be/KB/pn/pnoerr0.html>.

Ogg, C. Solution Given by the Universal Character Set. Extract from the OII Technology Handbook. [online]. Technology Appraisals Ltd, 1996. [Viitattu 27.5.1999]. Saatavilla www-muodossa <URL: http://www.techapps.co.uk/ucs.html>.

Olsen, M. 1996. Text Theory and Coding Practice: Assessing the TEI. [online]. ARTFL Project, University of Chicago. [Viitattu 22.10.1998]. Saatavilla www-muodossa <URL: http://tuna.uchicago.edu/homes/mark/talks/TEI.talk.html>.

Oudet, B. Multilingualism on the Internet. [online]. Scientific American, March 1997. [Viitattu 28.1.1999]. Saatavilla www-muodossa <URL: http://www.sciam.com/0397issue/0397oudet.html >.

Overview of the Temple project. [online]. New Mexico State University, Computing Research Laboratory, 1996. [Viitattu 12.11.1998]. Saatavilla www-muodossa <URL: http://crl.nmsu.edu/Research/Projects/tide/presentation/Overview.html>.

Pemberton, S. Views and Feelings: Flags are not Languages. [online]. SIGCHI Bulletin, Vol. 30, No. 1, January 1998. [Viitattu 7.6.1999]. Saatavilla www-muodossa <URL: http://www.acm.org/sigchi/bulletin/1998.1/views.html>.

Phase III Overview. [online]. [Viitattu 16.12.1998]. Saatavilla www-muodossa <URL: http://www.tipster.org/phaseiii.htm>.

Raggett, D., Le Hors A. & Jacobs I. (toim.). HTML 4.0 Specification. W3C Recommendation, revised on 24-Apr-1998. [online]. W3C World Wide Web Consortium, 1998. [Viitattu 19.1.1999]. Saatavilla www-muodossa <URL: http://www.w3.org/TR/REC-html40/>.

Seaman, D. (toim.). Guidelines for SGML Text Mark-up at the Electronic Text Center, Appendix II: ISO 639 Language Codes. [online]. ElectronicText Center, University of Virginia, 1992. [Viitattu 8.2.1999]. Saatavilla www-muodossa <URL: http://etext.lib.virginia.edu/tei/iso639.html>.

Seaman, D. Guidelines for SGML Text Mark-up at the Electronic Text Center, The TEI header. [online]. Electronic Text Center, University of Virginia, 1999. [Viitattu 8.6.1999]. Saatavilla www-muodossa <URL: http://etext.lib.virginia.edu/tei/uvatei4.html>.

Shneiderman B., Byrd D., Croft W. B., Sorting Out Searching. A User-Interface Framework for Text Searches. Communications of the ACM, Vol. 41, No. 4, 1998, 95-98.

Sperberg-McQueen, C. M., Burnard, L. (toim.), Guidelines for Electronic Text Encoding and Interchange (TEI P3). Vol. 1. Oxford, Chicago, 1994.

The State of Israel, Ministry of Justice. [online]. 1999. Saatavilla www-muodossa <URL: http://www.justice.gov.il/>.

The Unicode® Standard. A Technical Introduction. [online]. The Unicode Consortium, 1998. [Viitattu 17.11.1998]. Saatavilla www-muodossa <URL: http://www.unicode.org/unicode/standard/principles.html>.

TIPSTER Text Program Overview. [online]. [Viitattu 16.12.1998]. Saatavilla www-muodossa <URL: http://www.tipster.org/overv.htm>.

Travis, B., Waldt, D., The SGML Implementation Guide. A Blueprint for SGML Migration. Springer-Verlag, Berlin Heidelberg, 1995.

UKOLN Metadata Group. 1998. Biblink - LB 4034 - Three SGML metadata formats: TEI, EAD, and CIMI. [online]. [Viitattu 5.10.1998]. Saatavilla www-muodossa <URL: http://hosted.ukoln.ac.uk/biblink/wp1/sgml/overview.html>.

University of Virginia Library. Electronic Text Center. 1992. [online]. [Viitattu 8.6.1999]. Saatavilla www-muodossa <URL: http://etext.lib.virginia.edu/>.

URSA Unicode Retrieval System Architecture. 1998. [online]. [Viitattu 9.11.1998]. Saatavilla www-muodossa <URL: http://crl.nmsu.edu/Research/Projects/tipster/ursa/index.html>.

Watters P. A. & Patel M., Semantic processing performance of Internet machine translation systems. Internet Research: Electronic Networking Applications and Policy, Vol. 9, No. 2, 1999, 153-160.

Yergeau, F., Nicol, G., Adams, G. & Duerst, M. Internatonalization of the Hypertext Markup Language. 1997. [online]. [Viitattu 13.4.1999]. Saatavilla www-muodossa <URL: http://www.cis.ohio-state.edu/htbin/rfc/rfc2070.html>.

Zajac R. & Casper M. The Temple Web Translator. [online]. Computing Research Laboratory, New Mexico State University, 1997. [Viitattu 1.9.1998]. Saatavilla www-muodossa <URL: http://crl.nmsu.edu/Research/Projects/tide/papers/twt.aaai97.html>.

Zajac R., Vanni M., Glossary-Based MT Engines in a Multilingual Analyst's Workstation Architecture. Machine Translation, Vol. 12, No. 1 / 2, 1997, 131-151.

LIITTEET

LIITE 1: Tieteellisen artikkelin merkkaus TEI:llä.

Liitteessä on mallina käytetty Virginian yliopiston elektronisten tekstien keskuksen (ks. University of Virginia Library, 1992) esimerkkiä lehtiartikkelin merkkauksesta (Seaman, 1999). Merkattu teksti on osa Wattersin ja Patelin (1999) artikkelista, joka on ilmestynyt Internet Research –aikakauslehdessä. Artikkeli on esimerkki julkaisusta, jossa pääosa tekstistä on samalla kielellä (tässä tapauksessa englanniksi) ja jonka tekstissä lisäksi esiintyy useammalla eri kielellä olevia sanoja.

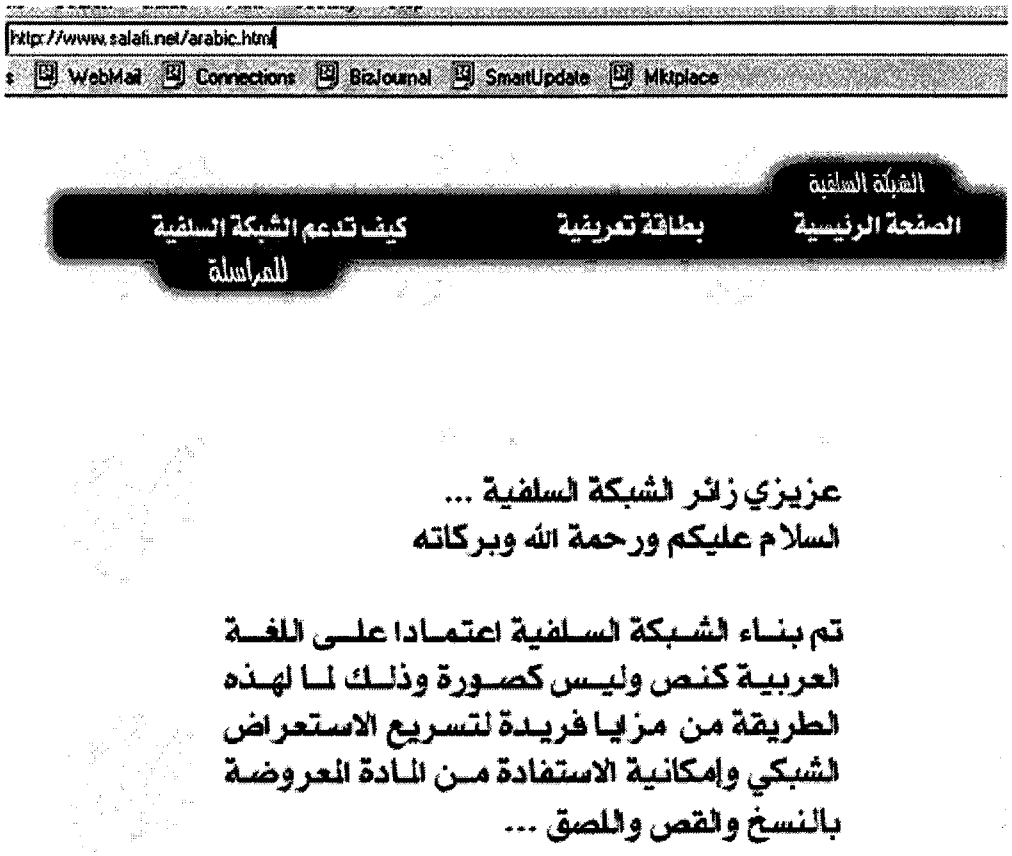
```
<teiHeader>
  <fileDesc>
      <titleStmt>
            <title>Semantic processing performance of Internet machine translation
            systems [a machine- readable transcription] </title>
            <author>Watters, Paul A.</author>
            <author>Patel, Malti</author>
            <respStmt>
               <resp>Creation of machine-readable version:</resp>
               <name>Pirkko Tuulihovi </name>
            </respStmt>
      </titleStmt>

      <publicationStmt>
            <publisher>MCB University Press</publisher>
            <availability>
               <p>Publicly accessible </p>
               <p n=public>URL: http://www.emerald-
            library.com/pdfs/17209bh2.pdf</p>
            </availability>
            <date>1999</date>
      </publicationStmt>

      <sourceDesc>
            <bibl>
               <titleStmt>
                        <title> Semantic processing performance of Internet
                        machine translation systems </title>
```

```
                    <author>Watters, Paul A.</author>
                    <author>Patel, Malti</author>
                </titleStmt>

                <notesStmt>
                    <note>This article appeared in Internet Research: Electronic
                    Networking Applications and Policy, v. 9, no. 2, p. 153-
                    160.</note>
                </notesStmt>
            </bibl>
    </sourceDesc>

</fileDesc>

<encodingDesc>

    <projectDesc>
        <p> This is an example of TEI markup prepared for my master
        thesis.</p>
    </projectDesc>

     <editorialDecl>
        <p>Spell-check and verification made against printed text using MS Word
        spell checker.</p>
        <p>All unambiguous end-of-line hyphens have been removed, and the
        trailing part of a word has been joined to the preceding line.</p>
    </editorialDecl>

    <refsDecl>
        <p>ID elements are given for each page element and are composed of
        the text's unique cryptogram and the given page number.</p>
    </refsDecl>

</encodingDesc>

<profileDesc>
    <creation>
        <date>1999</date>
    </creation>

    <langUsage>
        <language id=en>English </language>
        <language id=fr>French</language>
        <language id=es>Spanish</language>
        <language id=de>German</language>
        <language id=it>Italian</language>
    </langUsage>
</profileDesc>

<revisionDesc>
    <change>
        <date>June 1999 </date>
        <respStmt>
```

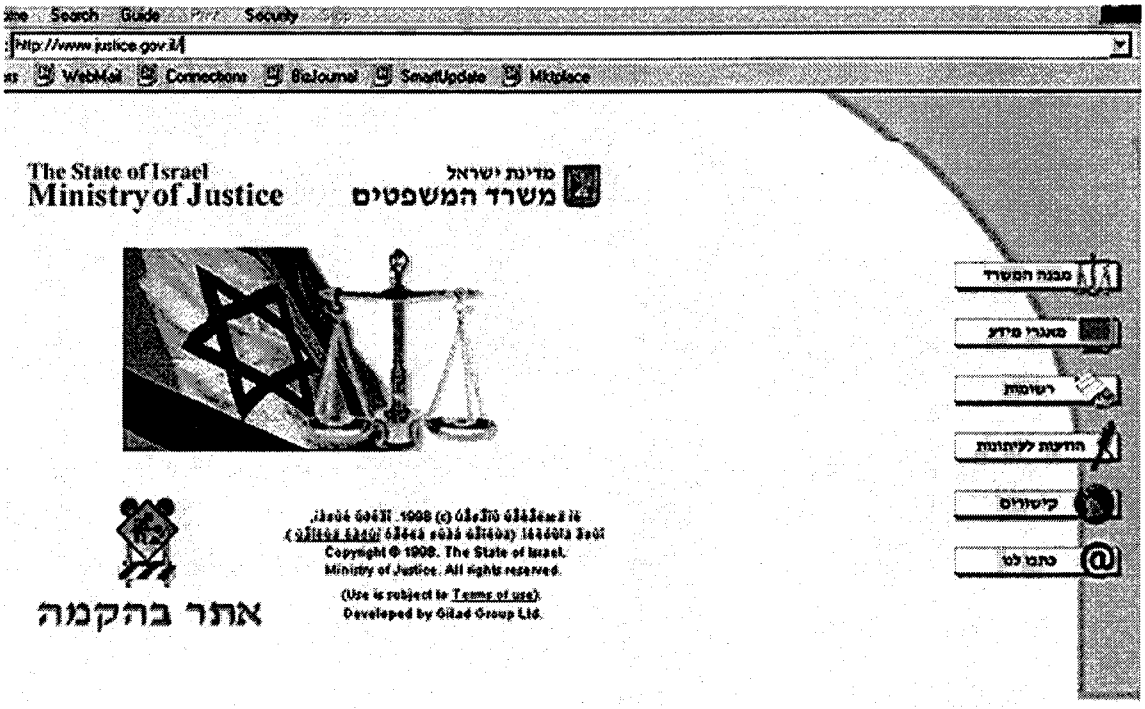
```
                <resp>corrector</resp>
                <name>Pirkko Tuulihovi, University of Jyväskyl&auml;</name>
            </respStmt>
            <item>Added TEI header and tags.</item>
        </change>
    </revisionDesc>

</teiHeader>

<text id='WattPat'>
    <front>
...
        <div type='abstract'>
            <p>The Internet has the potential to facilitate understanding across
            cultures and languages by removing the physical barriers to intercultural
            communication. One possible contributor...</p>
        </div>
...
    </front>
    <body>
        <div type ='chapter' n='1'>
            <head>Introduction</head>
            <p>The Internet, as a non-coercive technology, which facilitates the
            development locally-relevant and "culturally-appropriate" content...</p>
            ...
        </div>
        ...
        <div type ='chapter' n='3'>
            <head>Results</head>
            ...
            <p>...Alternatively, the error appears to occur in the French –> English
            inversion, as the verb in French which appears is correct ("<foreign
            lang="fr">sauvegarde</foreign>".</p>
            ...
            <p>...This is actually quite a silly and inexplicable error, since the
            Spanish word for moss ("<foreign lang="es">musgo</foreign>") is actually
            contained in the target Spanish sentence as translated from English.</p>
            ...
        </div>
    </body>
    <back>
        <div type=bibliograph>
            <head>References</head>
        ...
        </div>
    </back>
</text>
```

LIITE 2: Kuvat oikealta vasemmalle luettavista www-sivuista.

Arabiankielinen www-sivu on kuviossa 10 ja hepreankielinen kuviossa 11.



KUVIO 10. Arabiankielinen www-sivu (How to browse the arabic language based web pages?, 1999).



KUVIO 11. Hepreankielinen www-sivu (The State of Israel, Ministry of Justice, 1999).