*Pertti Väyrynen**

Kohti digitaalista infrastruktuuria kieliteknologiassa

Pertti Väyrynen: Kohti digitaalista infrastruktuuria kieliteknologiassa [Towars a digital infrastructure in Language Tecnology] Informaatiotutkimus 21 (1), 3-8.

In this article the exploitation of potentially useful linguistic knowledge as part of a digital infrastructure is studied. Three points of view should be taken into consideration: 1) linguistic knowledge, 2) the technology and 3) the user.

Address: Pertti Väyrynen, MediaTeam Oulu Group, University of Oulu, Erkki Koiso-Kanttilan katu 3, FIN-90014 University of Oulu, Finland.
Email: pertti.vayrynen@ee.oulu.fi.

1. Johdanto

Informaatioyhteiskunnan tulo on tosiasia, olimmepa valmiita siihen tai emme. Informaatioyhteiskunnan oleellisin piirre on *suurten tietomassojen hallinta*, johon kuuluu esim. uuden tiedon tuottaminen, tiedon varastointi ja distribuutio. Suurten tietomassojen hallinta tapahtuu käytännössä tietokoneympäristössä, mikä edellyttää yleensä jonkinlaista kommunikaatiota tietokoneen kanssa mieluiten käyttäjän omalla äidinkielellä. Monilla elämän alueilla kuten esim. hallinnossa ja liike-elämässä oikean tiedon löytäminen erityisesti suurten tietomassojen joukosta on erittäin tärkeää informaatioyhteiskunnassa.

Erilaisiin tiedonhallinnan tehtäviin tarvitaan tietokoneympäristöön soveltuvia käytännön sovelluksia, joiden kehittämisessä voidaan hyödyntää *kieliteknologiaa*. Kieliteknologia tarkoittaa yleisesti määriteltynä kielen tietokonepohjaista käsittelyä. Käytännön sovellusten kehittämiseen informaationteknologian eri sovellusalueille tarvitaan Cucchiarini:n et al. (2001) mukaan *digitaalinen infrastruktuuri* kullekin kielelle erikseen, johon sisältyy esim. dataa kielestä ja puheesta eli lingvististä tietoa, korpuksia ja sanastoja.

Koska informaationhallintasovellusten prosessoima informaatio on pääosin kielellistä informaatiota, joko tekstiä tai puhetta, joka on ilmaistu jollakin luonnollisella kielellä, olettaisi helposti, että *lingvististä tietoa* eli tietoa kielestä ja sen rakenteesta hyödynnetään yleisesti näissä sovelluksissa. Tämä ei kuitenkaan pidä tarkkaan ottaen paikkaansa. Toisin kuin olettaisi, informaatioteknologian menestyneimmät käytännön sovellukset hyödyntävät nykyisin itse asiassa vielä toistaiseksi varsin vähän eksplisiittistä lingvististä tietoa. Näidenkin sovellusten taustalla saattaa itse asiassa piillä implisiittistä lingvististä tietoa kuten jäljempänä osoitetaan.

Jotkut sovellukset kuten esim. kaupalliset puheentunnistussovellukset perustuvat jopa suorastaan *vääriin olettamuksiin* puheen akustisesta invarianssista tai puheen lineaarisesta luonteesta (Suomi 1987). Toisaalta näidenkin menestyneimpien sovellusten *suorituskyky* on riittävän hyvä vain tarkoin rajatulla sovellusalueella; kun sovellusaluetta laajennetaan, niin sovellusten suorituskyky laskee usein dramaattisestikin. Sovellusten suorituskyky ei enää myöskään parane (suoraviivaisesti) teknisen kehityksen myötä: monet väittävät, että tilastolliset tekniikat, jotka

*Kirjoittaja FL Pertti Väyrynen työskentelee lingvistinä Oulun yliopiston MediaTeam-tutkimusryhmässä sen Language and Audiology Team:ssä. Lisätietoja osoitteesta http://www.mediateam.oulu.fi?/lang=fi

perustuvat tyypillisesti (perinteisiin) kätkettyihin Markovin malleihin, ovat jo saavuttaneet suorituskykynsä ylärajan, esim. King et al. (2000).

Tällöin tarvitaan uusia keinoja sovellusten suorituskyvyn parantamiseen. Lingvistinen tieto voi olla tärkeää tässä suhteessa, erityisesti kun sovelluksia kehitetään yhä uusille käyttöalueille. Lingvistisen tiedon hyödyntämisessä eri informaatioteknologian sovellusalueilla osana digitaalista infrastruktuuria voidaan itse asiassa erottaa kolme näkökulmaa: 1. lingvistisen tiedon itsensä näkökulma, 2. teknologian näkökulma ja 3. sovellusten käyttäjän näkökulma. Käytännössä nämä eri näkökulmat pitää sovittaa yhteen ohjelmistokehityksessä.

Tämän artikkelin tarkoitus on yrittää kartoittaa potentiaalisesti hyödyllistä lingvististä tietoa informaatioteknologian joillakin keskeisillä sovellusalueilla yllä mainitusta kolmesta näkökulmasta käsin lähinnä englannin kielen osalta. Sinällään kieliteknologiasovellukset eri kieliin ja kielityyppeihin ovat entistä tärkeämpiä varsinkin nykyisin: ellei käyttäjä voi käyttää omaa äidinkieltään kieliteknologiasovelluksissa, on vaarana, että pienten kielten asema heikkenee (Miettinen & Toivanen 2001). Systemaattiset kartoitukset potentiaalisesti hyödyllisestä lingvistisestä tiedosta edes keskeisimmillä informaatioteknologian sovellusalueilla näyttävät niin ikään myös puuttuvan kirjallisuudessa.

2. Lingvistisen tiedon näkökulma

Periaatteessa on paljon kielellistä tietoa eli tietoa kielestä ja sen rakenteesta, joka on potentiaalisesti hyödyllistä informaatioteknologian eri sovellusalueilla. Käytännössä kaikkea tätä potentiaalisesti hyödyllistä tietoa ei kuitenkaan voitane hyödyntää: perusedellytys kaikenlaiseen kielellisen tiedon hyödyntämiseen tietokoneympäristöissä on sen *soveltuvuus tietokoneympäristöön*. Soveltuakseen tietokoneympäristöön, kielellisen tiedon on oltava *algoritmista* tai se on tehtävä algoritmiseksi. Yritykset hyödyntää perinteistä deskriptiivistä lingvististä tietoa sellaisenaan tietokoneympäristössä sovellusalueilla kuten esim. automaattinen puheentunnistus tai puhesynteesi ovat pääsääntöisesti epäonnistuneet juuri lingvistisen tiedon deskriptiivisen, ei preskriptiivisen, luonteen takia (Huckvale 1996:1).

Vielä nykyisin lingvistisen tiedon hyödyntäminen monilla informaatioteknologian osa-alueilla näyttää keskittyvän etupäässä kielen *tilastollisen rakenteen* hyödyntämiseen. Kielen tilastollista rakennetta hyödynnetään etupäässä sen takia, että se on helposti implementoitavissa tietokoneympäristössä. Jopa niinkin yksinkertaisesti implementoitavaa tietoa kuten esim. tietoa sanan ja virkkeen pituudesta voidaan hyödyntää tekstin luokittelussa eri genreihin (Niemikorpi 1974) tai tekstin tilastollisesti mallinnettavien tyylillisten piirteiden hyödyntäminen tiedonhaussa (Karlgren 1999). Luokittelu ei kuitenkaan aina välttämättä ole täysin yksiselitteinen. Tekstin segmentointi, mikä sinällään edustaa erästä keskeistä tehtävää luonnollisen kielen prosessointia hyödyntävissä kieliteknologiasovelluksissa, erillisiin virkkeisiin koneellisesti ei ole välttämättä niin yksinkertaista kuin saattaisi olettaa: yksinkertaisilla tekniikoilla saatetaan päästä yllättävän pitkälle, mutta tulosten parantaminen perustasolta mukaanlukien vaikeiden tapausten segmentointi saattaa vaatia paljon lingvististä tietoa ja intuitiota kielen rakenteesta tutkimuskorpusten huolellisen analyysin lisäksi (Bayer et al. 1998, 243).

Sovellusalueilla kuten esim. tiedonhaussa on menestyksellä hyödynnetty hakukyselyjen *synonyymilaajennusta* thesaurustyyppisten sanastojen avulla, joka käyttäjien mielestä parantaa hakutuloksia jopa maagisesti (Croft 1995). Muita esimerkkejä lingvististen piirteiden hyödyntämisestä tiedonhaussa esitellään lähteessä Strzalkowski (1999) esim. lingvistisesti motivoitujen piirteiden hyödyntämisestä indeksoinnissa (Karen Sprack Jones) tai leksikaalisten resurssien käytöstä perinteellisten tilastollisten indeksointi- ja tiedonhakumenetelmien toiminnan tehostamisessa (Alan Smeaton).

Syntaktista tietoa voidaan hyödyntää esim. puheenymmärtämistä (speech understanding) vaativilla sovellusalueilla, jossa syntaktisen tiedon tehtävä on tyypillisesti disambiguoida potentiaalisesti monitulkintaisia ilmauksia kuten esim. *John saw the man with the telescope* englannin kielessä, joihin sisältyy rakenteellista monitulkintaisuutta. Varsinaisessa puheentunnistuksessa syntaktista tietoa ei vielä juurikaan hyödynnetä sen takia, että lauseenjäsennystekniikat eivät ole vielä täydellisiä (Voorhees 1999, 42) tai koska puheentunnistuksessa tarvitaan tyypillisesti vasemmalta oikealle etenevää lauseenjäsennystä, jota nykyiset lauseenjäsennystekniikat eivät vielä pääsääntöisesti mahdollista (Charniak 1997, 42).

Osittain sen takia, että lauseenjäsennys ei (vielä) toimi virheettömästä sovellusalueilla kuten esim. puheentunnistus ja -ymmärtäminen, kiinnostus

on virinnyt *semanttista tietoa* hyödyntävään lauseenjäsennykseen lähestymistavoissa, joissa yritetään tulkita lauseen merkitystä sekä sen syntaktisen jäsennyksen että semanttisen prosessoinnin avulla (katso esim. Ng & Zelle 1997). (Manning & Schutze 2000, 457-458).

3. Teknologian näkökulma

Teknologian näkökulmasta tärkein asia insinööreille, jotka tyypillisesti kehittävät erilaisia informaatioteknologian sovelluksia, on usein pelkästään *sovelluksen suorituskyky* (Huckvale 1996). Tällöin mikä tahansa tieto, joka parantaa sovelluksen suorituskykyä, kelpaa insinööreille oli se sitten millaista tietoa tahansa, esim. tietoa, jonka mielekkyys käyttäjän näkökulmasta saattaa olla kyseenalaista.

Monilla sovellusalueilla kuten esim. puheentunnistuksessa sovelluksen suorituskyky on itse asiassa useiden eri tekijöiden summa (Ainsworth 1999, 738); lingvistinen tieto, jonka tyypillisin funktio on usein juuri sovelluksen *suorituskyvyn parantaminen*, on tyypillisesti vain yksi osa sovelluksen suorituskykyä. Lingvistisen tiedon ainoa funktio kieliteknologiasovelluksessa ei välttämättä ole sovelluksen suorituskyvyn parantaminen kuten jäljempänä osoitetaan. Sovelluksen suorituskyvyn *mittaaminen* on myös ongelma monilla sovellusalueilla: tyypillisesti käytetyt globaalit suorituskyvyn mittarit kuten esim. virheellisesti tunnistettujen sanojen määrä (word error rate) puheentunnistuksessa, eivät itse asiassa kerro paljonkaan sovelluksen suorituskyvystä.

Tietokoneympäristössä ideaalitilanne olisi löytää sellaista kielellisesti merkityksellistä tietoa, joka on helposti implementoitavissa tietokoneympäristössä käytännön sovellusten tarpeisiin informaatioteknologian eri sovellusalueilla. Käytännön sovelluksilla tarkoitetaan tässä esityksessä sovelluksia, joiden suorituskyky on riittävän hyvä, mutta ei useinkaan täysin virheetön. Tyypillisesti nämä informaationprosessointisovellukset ratkaisevat joitakin käytännön ongelmia, jotka liittyvät ihmisten erilaisiin tiedontarpeisiin (Piotrowski 2000). Tällaiset digitaalisen infrastruktuurin *perusohjelmistokomponentit* (modules, semiproducts), joilla ei vielä sellaisenaan ole paljon *kaupallista merkitystä*, mutta jotka ovat siitä huolimatta tärkeitä komponentteja digitaalisessa infrastruktuurissa (Cucchiarini et al 2001) - varsinkin, jos käy ilmi, että näitä peruskomponentteja voidaan käyttää useilla eri informaatioteknologian sovellusalueilla yhdessä tai jopa useammassa (rinnasteisessa) kielessä. Esimerkiksi *morfologinen dekompositio* ja sanan vartalon tunnistus (stemming) ovat yleisesti ottaen hyödyllisiä kielissä, joilla on rikas morfologinen systeemi kuten esim. saksa, hollanti ja italia. Molempia voidaan käyttää hyödyksi sekä *tiedonhaussa että puheentunnistuksessa*. Puheentunnistuksessa morfologista dekompositiota, joka käytännössä on aika pitkälle sama asia kuin sanan vartalon tunnistus, voidaan hyödyntää tunnistusjärjestelmän käyttämän sanaston kattavuuden (lexical coverage) optimoinnissa sekä sanojen erilaisten ääntöasujen tuottamisessa käytettyä sanastoa varten (Piotrowski 2000, Adda-Decker & Adda 2000).

Muita lingvistisen tiedon perusohjelmistokomponentteja voisivat olla esim. analyysiyksikön distribuution muutos, jota voisi nimittää vaikkapa *suhdeparametriksi*. Esimerkkinä voitaisiin mainita vaikkapa ns. *hapax legomena* sanojen eli sanojen, jotka esiintyvät tekstissä vain kerran, distribuution muutos tekstissä, joka Bruce:n (1999) mukaan toimii yhtenä *merkityksellisen informaation* tekstuaalisena korrelaattina englannin kielessä.

Tätä voidaan hyödyntää esim. tiedonhaussa yhtenä *tekstitason metadatapiirteenä*, jonka avulla voidaan lokalisoida merkityksellistä informaatiota tekstissä kiinnittämättä vielä huomiota tämän merkityksellisen informaation tarkkaan sisältöön. Tarkoituksena on siis tällöin vain löytää nämä merkitykselliset informaationkohdat tekstissä esim. osana tekstiaineistojen metadataindeksointia.

Esimerkkinä lingvistisen tiedon hyödyntämisestä osana digitaalisen infrastruktuurin perusohjelmistokomponentteja voisi olla myös *analyysiyksikön toisto*, jota voisi nimittää vaikkapa *toistoparametriksi*. Toistoa esiintyy useilla eri kielen rakennetasoilla, ja sillä on erilaisia funktioita, joita voidaan hyödyntää *kieliteknologiasovelluksissa*: eräs hyväksi havaittu tekniikka tekstinennakoinnissa (word prediction) on sanojen taipumus esiintyä uudelleen tekstissä pian sen jälkeen kun sitä on käytetty ensimmäisen kerran (ns. recency of mention - tekniikka), jossa siis tunnistetaan analyysiyksikön eli saman sanan toistuminen tekstissä tyypillisesti lyhyellä aikaikkunalla. Yleensä saman sanan toistoa pidetään tekstin tilastollisena ominaisuutena, mutta psykolingvistiikassa siihen viitataan englanninkielisellä termillä *priming effect* (Carpenter 1999), joka tarkoittaa juuri kielellisen yksikön kuten esim. sanan tai jopa virkerakenteen taipumusta

toistua tai toistaa itseään. Sovellusten toiminnan kannalta analyysiyksikön toisto tekee niiden toiminnan jossain määrin *ennustettavammaksi*, ja ilmiö perustuu siis itseasiassa lingvistiseen tietoon, vaikka sitä ei ehkä tunnetakaan yleisesti. Diskurssin tasolla toiston funktio saattaa olla *lokaalisen topiikin identifiointi* (Berber Sardinha 1996, 5), jota voitaan hyödyntää topiikin tunnistuksessa tiedonhaussa. Analyysiyksikön toistolla eri kielen tasoilla on todennäköisesti myös muita funktioita kuin ylläkuvatut kaksi funktiota.

Ylläkuvatun kaltaisia kielellisen tiedon perusohjelmistokomponentteja löytyy todennäköisesti lisääkin, jos analysoidaan systemaattisesti lingvististä tietoa eri kieliteorioissa ja kielimalleissa ja kyetään näkemään niiden relevanssi informaatioteknologian eri sovellusalueilla.

Perusohjelmistokomponenttien lisäksi myös muunlainen lingvistinen tieto saattaa olla hyödyksi informaatioteknologian eri sovellusalueilla. Esim. *lingvistisiä säännönmukaisuuksia* (linguistics insights) kuten esim. sitä, että samaa sanaa käytetään yleensä samassa merkityksessä saman tekstin sisällä voidaan hyödyntää sovellusalueilla kuten esim. tiedonhaku.

Teknologian näkökulmasta katsottuna perusohjelmistokomponenttien lisäksi lingvistisen tiedon hyödyllisyyttä sovelluskehityksessä voidaan arvioida myös useista sovelluksen ominaisuuksiin liittyvistä lähtökohdista käsin kuten esim sovelluksen *tehokkuus, robusti toiminta* eli virhetilannesietoisuus ja ennen kaikkea virhetilanteista toipuminen tai käytetyn kielellisen tiedon *kielestä rippumattomuus*. Käytännössä joudutaan tekemään jokin optimaalinen kompromissi lingvistisen tiedon käytössä kieliteknologia-sovelluksissa suhteessa haluttuihin ohjelmaominaisuuksiin. Esim. tehokkaimmissa sovelluksissa joudutaan tyypillisesti käyttämään alemman tason kielellistä prosessointia (esim. vain sanatasolle saakka yltävää prosessointia) sen takia, että korkeamman tason kielellisten ilmiöiden (virketason yläpuolisten, tilastollisesti harvinaisten, kielellisten ilmiöiden prosessointi) vaatii jo enemmän prosessointitehoa (Liddy 1998).

Vaadittu prosessointiteho ei sinällään ole välttämättä ongelma tulevaisuudessa, varsinkin koska tietokoneiden teho lisääntyy jatkuvasti. Sovelluksen tehokkuus liittyy myös siihen, missä määrin käytetty kielellinen tieto riittää kattamaan erilaisia inputmuotoja, joita sovellus joutuu prosessoimaan käytännön sovelluksissa siten, että kattavuuden lisäys voi laskea sovelluksen tehokkuutta (Uszkoreit 1996).

4. Käyttäjän näkökulma

Lingvistisen tiedon hyödyntämisessä sovelluskehityksessä sovellusten käyttäjän näkökulma on erittäin tärkeä. Ohjelmistojen käytettävyys yhä tärkeämpää teknistyvässä maailmassamme (Hunt 2000, 1).

Käyttäjän näkökulmasta ohjelmiston *tehokkuus* ei välttämättä aina merkitse sitä, että teknologian näkökulmasta katsottuna tehokkain tai monipuolisin sovellus olisi aina myös paras tai hyväksyttävin käyttäjän kannalta. Esimerkiksi Walker et al. (1998) huomasivat, että käyttäjät pitivät parempana sovellusta, jossa sovellus itse kontrolloi dialogia kuin sovellusta, jossa myös käyttäjällä oli mahdollisuus joustavasti kontrolloida dialogia sähköpostisysteemissä, jossa oli puhekäyttöliittymä. Tähän oli luultavasti syynä se, että systeemiä, jossa sovellus kontrolloi dialogia, oli helpompi oppia käyttämään ja sen toiminta oli *ennustettavampaa* kuin systeemin, jossa käyttäjällä oli mahdollisuus kontrolloida joustavasti dialogia sovelluksen ja ihmisen välillä.

Sovelluksen helppokäyttöisyyden lisäksi sen tulisi käyttäjän näkökulmasta toimia myös *älykkäästi* erityisesti sovellusalueilla, joihin sisältyy jonkinlaista kielellistä prosessointia. Jos esim. tekstinennakointiohjelma (word prediction software) englannin kielessä esittää määräävän artikkelin "the" jälkeen tilastollisesti yleisimmäksi sanaksi verbin, jonka se ennustaa seuraavaksi kirjoitettavaksi sanaksi, niin se ei toimi käyttäjän näkökulmasta älykkäästi, koska se ei huomioi kielen lauserakenteen vaikutusta ennustettavien sanojen sanaluokkaan.

Ohjelmiston *intuitiivisuus* saattaa taas olla ristiriidassa sovelluksen tehokkuuden kanssa: on osoitettu, että lausetason informaation hyödyntäminen tekstinennakoinnissa parantaa sovellusten suorituskykyä eli kirjoittamatta jäävien merkkien näärää vain muutamalla prosentilla (Wood 1996, 123-129), mutta käyttäjän näkökulmasta lauserakenteen huomioonottaminen tekstinennakoinnissa vastaa paremmin käyttäjän kielellistä intuitiota. Ohjelman käytettävyyden/hyväksyttävyyden kannalta katsottuna sovelluskehityksen pitäisikin olla *iteratiivista*, jossa ohjelmaa testataan mahdollisesti useaan otteeseen sen käytettävyyden selvittämiseksi.

5. Päätäntä

Lingvistisen tiedon hyödyntämisessä informaatioteknologiassa osana kieliteknologiasovellusten tarvitsemaa digitaalista infrastruktuuria voidaan erottaa kolme eri näkökulmaa, jotka ovat 1. lingvistisen tiedon itsensä näkökulma, 2. teknologian näkökulma ja 3. käyttäjän näkökulma. Käytännön sovelluksia kehitettäessä, nämä erilaiset näkökulmat on sovitettava yhteen.

Lingvistisen tiedon hyödyllisyydestä puhuttaessa täytyy erottaa lingvistisen tiedon *potentiaalinen hyödyllisyys* vs. sen *aktuaalinen hyödyllisyys*, mitkä ovat kaksi eri asiaa. Yleisimmin tunnutaan puhuvan lingvistisen tiedon potentiaalisesta hyödyllisyydestä, kuten tässäkin esityksessä, kuin sen aktuaalisesta hyödyllisyydestä, mikä heijastelee lingvistisen tiedon vähäistä käyttöä informaatioteknologian eri sovellusaluilla vielä nykyisin.

Sinällään lingvistinen tieto eri informaatioteknologian sovellusalueilla on tärkeää varsinkin tulevaisuudessa kun uudentyyppisiä sovelluksia kehitetään yhä uusille sovellusalueille, joissa käytetään yhä laajempia sanastoja. Tällöin joudutaan hyödyntämään lingvististä tietoa esim. puheentunnistuksessa varmistamaan tunnistustulosten oikeellisuutta; tällöin lingvistisen tiedon rooli on sen perinteinen rooli eli tunnistustulosten verifiointi kielellisen tiedon avulla. (Ueberla 1994, 89-90.)

Uusienkaan mediatyyppien tulo ei poista lingvistisen tiedon tärkeyttä: jopa *multimediassa*, joka yhdistää tekstiä, grafiikkaa, ääntä ja videokuvaa, eri datamuodoissa olevan informaation rakentaminen, indeksointi ja informaation sisällä navigointi voidaan tehdä vain kielen avulla (Uszkoreit 2000, 2).

Luovuus ja kyky selittää ilmiöitä lingvistisesti ovat niin ikään tärkeitä lingvistisen tiedon hyödyntämisessä informaatioteknologian eri sovellusalueilla, esim. sovellusten optimaalisen käytön kannalta. Eräässä sanelukoetestissä testattiin sukupuolten välisiä eroja puheentunnistuksessa (PC Magazine Online, the October 20, 1998 issue). Yleensä naisääni on vaikeammin tunnistettavissa kuin miesääni. Testituloksissa saatiin sikäli yllättävä tulos, että naisilla sanelutulos oli muutaman prosentin parempi kuin miehillä. Tämä selittynee sillä, mitä tiedämme sukupuolten välisistä eroista kielenkäytössä yleensä: naisten väitetään olevan miehiä tietoisempia kielenkäyttönsä oikeellisuudesta, esim. huolitellumman ääntämyksen osalta, mikä luultavasti heijastui saaduissa sanelukoetuloksissa. Tällöin yritysten kannattaisi käyttää naisia tekstien sanelussa puheentunnistuksen avulla, koska tekstinkorjaukseen, joka on vielä usein varsin oleellinen osa tekstin tuottamista puheentunnistuksen avulla, kuluu tällöin vähemmän aikaa (Soltau & Waibel 1998, 1).

Sinällään ajatus digitaalisen infrastruktuurin kehittämisestä kieliteknologiasovellusten kehittämistä varten on mitä kannatettavin. Tämän infrastruktuurin edelleen kehittämiseen voidaan antaa mielestäni seuraavat suositukset:

1. lingvistien tiedon perusohjelmistokomponenttien identifiointi analysoimalla lingvistisen tiedon käyttöä joillakin informaatioteknologian keskeisillä sovellusalueilla sisällytettäväksi osaksi digitaalista infrastruktuuria;
2. muunlaisen potentiaalisesti hyödyllisen lingvistisen tiedon kartoitus perusohjelmistokomponenttien lisäksi;
3. karkeiden kriteerien määrittäminen, joilla potentiaalisesti hyödyllistä lingvististä tietoa voidaan valita suhteessa erilaisiin ohjelmistoominaisuuksiin, esim. lingvistisen tiedon kielikohtaisuus, skaalautuvuus eri sovellusalueille, etc.;
4. käyttäjän näkökulman huomioiminen ohjelmistokehityksessä entistä enemmän.

Hyväksytty julkaistavaksi 7.2.2002.

Lähteet:

Adda-Decker, M. & Adda, G. (2000). Morphological Decomposition for ASR in German. PHONUS 5:129-143.

Ainsworth, W. A. (1999). Some Approaches to Automatic Speech Recognition. The Handbook of Phonetic Sciences. Toimittaneet Hardcastle, W. J. & Laver, J. Blackwell Publishers.

Bayer, S., Aberdeen, J., Burger, J., Hirschman, L., Palmer, D., Vilain, M. (1998). Theoretical and Computational Linguistics: Toward a Mutual Understanding. Using Computers in Linguistics: A Practical Guide. Toimittaneet Lawler, J. M. & Dry, H. A. London and New York: Routledge.

Berber Sardinha, A. P. (1996). A Window on Lexical Density in Speech. 14th of January 2002. http://www.liv.ac.uk/~tony1/homepage.html.

Bruce, T. R. (1999). Peak in Discourse. 3rd of June 2001. http://www.ovc.edu/discourse/peak.htm.

Carpenter, B. (1999). Human versus Machine: Psycholinguistics Meets ASR. 17th of December 2001.asru99.research.att.com/abstracts/4_984_invited.pdf.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. AI Magazine. Vol. 18 (4): 33-43.

Croft, W. B. (1995). What Do People Want from Information Retrieval? D-Lib Magazine, November 1995. 10th of November 2001. http://sunsite.anu.edu.au/mirrors/dlib/november95/11croft.html.

Cucchiarini, C., Daeleman, W., Strik, H. (2001). Srengthening the Duch Human Language Technology Interface. The ELRA Newsletter: 3-7.

Huckvale, M. (1996). Learning from the Experience of Building Automatic Speech Recognition Systems. UCL Working Papers, Speech, Hearing and Language.

Hunt, C. (2000). Natural Language Processing and the Role of Linguistic Analysis. 4th of June 2001. http://www.slis.unalberta.ca/cap00/clhunt/paper.htm.

Karlgren, J. (1999). Stylistic Experiments in Information Retrieval. Natural Language Information Retrieval. Toimittanut Strzalkowski, T. Dordrecht/Boston/London: Kluver Academic Publishers.

King, S., Taylor, P., Frankel, J., Richmond, K. (2000). Speech Recognition via Phonetically-Featured Syllables. PHONUS 5:15-34.

Liddy, E. D. (1998). Enhanced Text Retrieval Using Natural Language Processing. 9th of February 2001. http://www.asis.org/Bulletin/Apr~98/liddy.html.

Manning, C. D. & Schutze, H. (toim.) (2000). Foundations of Statistical Natural Language Processing. Cambridge: The MIT Press.

Miettinen, M. & Toivanen, J. (toim.) (2001). Puheentutkimuksen resurssit Suomessa (The Resources of Speech Research in Finland). CSC - Scientific Computing Ltd.

Ng, H. T. & Zelle, J. (1997). Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing. AI Magazine 18:45-64.

Niemikorpi, A. (1974). Saneiden ja virkkeiden pituudet suomen 1960-luvun kirjakielen kumulatiivisina ominaisuuksina. Lisenssiaattitutkielma. Oulun yliopisto.

Piotrowski, M. (2000). NLP-Supported Full-Text Retrieval. Master's thesis. Friedrich-Alexander-Universität Erlangen-Nürnberg. Institut für Germanistisk Abteilung für Computerlinguistik.

Smeaton, A. F. (1999). Using NLP or NLP Resources for Information Retrieval Tasks. Natural Language Information Retrieval. Toimittanut Strzalkowski, T. Dordrecht/Boston/London: Kluver Academic Publishers.

Soltau, H. & Waibel, A. (1998). On the Influence of Hyperarticulated Speech on Recognition Performance. Proceedings of the ICSLP 98:1-4.

Sprack Jones, K. (1999). What is the Role of NLP in Text Retrieval? Natural Language Information Retrieval. Toimittanut Strzalkowski, T. Dordrecht/Boston/London: Kluver Academic Publishers.

Suomi, K. (1987). On Spectral Coarticulation in Stop-Vowel-Stop Syllables: Implications for Automatic Speech Recognition (abstrakti). Journal of Phonetics 15: 85-100.

Strzalkowski, T. (1999). Natural Language Information Retrieval. Dordrecht/Boston/London: Kluver Academic Publishers.

Ueberla, J. (1994). Analyzing and Improving Statistical Language Models for Speech Recognition. Ph.D. thesis. Simon Fraser University.

Uszkoreit, H. (1996). Mathematical Methods. The Survey of the State of the Art in Human Language Technology. 27th of March 2000. http://cslu.cse.ogi.edu/HLTsurvey. Toimittanut Cole, R.

Uszkoreit, H. (2000). What is Computational Linguistics? 12th of September 2001. http://www.coli.uni-sb.de/~hansu/what_is_cl.html.

Voorhees, E. M. (1999). Natural Language Processing and Information Retrieval. Information Extraction: Towards Scalable, Adaptable Systems: 32-48. Toimittanut Pazienza, M. T. Germany: Springer.

Walker, M. A., Fromer, J., Di Fabbrizio, G., Mestel, C., Hindle, D. (1998). What Can I Say?: Evaluating a Spoken Language Interface to Email. Conference on Human Factors in Computing Systems - Proceedings: 582-589.

Wood, M. E. J. (1996). Syntactic Pre-Processing in Single-Word Prediction for Dis-abled People. Ph.D. thesis. University of Bristol.