2

UHKA KANSALLISKIELILLE ON HAASTE KIELITEKNOLOGIALLE

Olemme todistamassa digitaalista vallankumousta, jonka vaikutukset viestinnän toimivuuteen ja sitä kautta koko yhteiskuntaan tulevat olemaan merkittäviä. Tieto- ja viestintätekniikan viimeaikaista kehitystä on toisinaan verrattu Gutenbergin keksimään kirjapainotekniikkaan. Millaisia oletuksia Euroopan tietoyhteiskunnan ja erityisesti kieltemme tulevaisuudesta voimme vertauksen pohjalta tehdä?

> Digitaalisen vallankumouksen vaikutukset yhteiskuntaan tulevat olemaan merkittäviä.

Gutenbergin keksinnöstä seurasi todellisia läpimurtoja viestinnässä ja tiedon siirrossa, kuten Lutherin Raamatun käännös kansankielelle. Gutenbergin ajan jälkeen kuluneina vuosisatoina on kehitetty eri kulttuurien tarpeisiin monenlaisia teknikoita parantamaan kielenkäsittelyä ja tietämyksen siirtoa:

- suurten kielten ortografinen ja kieliopillinen standardisointi mahdollisti
- uusien tieteellisten ja henkisten saavutusten nopean levittämisen;
- virallisten kielten kehittyminen mahdollisti kansalaisten kommunikoinnin tiettyjen (usein poliittisten) rajojen sisällä;
- kielten opetus ja kääntäminen mahdollisti kieltenvälisen viestinnän;
- tekstin toimittamisen ja bibliografian laatimisen suositusten luominen takasi painotuotteiden laadun;
- erilaiset viestintäkanavat, kuten sanomalehti, radio, televisio ja kirja, tyydyttivät erilaisia viestinnällisiä tarpeita.

Informaatioteknologia on kuluneiden kahdenkymmenen vuoden aikana auttanut automatisoimaan asioita ja helpottanut monia toimintojamme arjessa:

- tietokoneavusteinen julkaisuohjelma on korvannut kirjoituskoneen ja ladonnan;
- piirtoheitinkalvot tehdään nykyisin esitysmateriaalien tuottamista varten tehdyillä ohjelmilla, kuten OpenOfficen esitysgrafiikat tai Microsoft PowerPoint;
- sähköposti lähettää ja vastaanottaa tiedostoja nopeammin kuin faksi;
- voimme puhua edullisia tai jopa ilmaisia Internet-puheluja ja kokoontua virtuaalisesti verkkokeskusteluohjelmien avulla;
- äänen ja kuvan tallennusformaatit tekevät multimediasisällön jakamisen helpoksi;
- hakukoneet tarjoavat asiasanaperusteista verkkosivujen hakumahdollisuutta;
- verkossa olevat palvelut kuten Googlen Kääntäjä tuottavat nopeita, summittaisia käännöksiä;

- sosiaalisen median alustat kuten Facebook, Twitter ja Google+ mahdollistavat kommunikaation, yhteistyön ja tiedonjaon.

Vaikka mainitut työkalut ja sovellukset ovat hyödyllisiä, ne eivät vielä kykene tukemaan kaikkien kansalaisten tavoittamaa monikielistä Euroopan yhteisöä, jossa tieto ja tavarat voivat liikkua vapaasti.

2.1 KIELTEN VÄLISET RAJAT ESTEENÄ EUROOPAN TIETOYHTEISKUNNAN KEHITYKSELLE

Emme kykene ennustamaan tarkasti, millaiselta tulevaisuuden informaatioyhteiskunta näyttää, mutta on hyvin todennäköistä, että tietotekniikan vallankumous tuo eri kieliä puhuvia ihmisiä yhteen uusilla tavoin. Kansalaisille syntyy tarpeita oppia uusia kieliä ja sovellusten kehittäjille tilaus luoda uusia teknologisia sovelluksia, joiden avulla voidaan varmistaa, että ymmärrämme toisiamme ja saavutamme kaiken tarvitsemamme tiedon.

> Yhä enemmän kieliä, puhujia ja sisältöä on jatkuvassa vuorovaikutuksessa keskenään.

Maailmanlaajuisten talousmarkkinoiden alueella ja tiedonkulun kentällä yhä enemmän kieliä, puhujia ja sisältöä on jatkuvassa vuorovaikutuksessa keskenään uusien viestintävälineiden avulla entistä nopeammin. Sosiaalisen median (Wikipedia, Facebook, Twitter, YouTube) suuri suosio on vain jäävuoren huippu.

Voimme nykyisin siirtää gigatavujen kokoisia tekstejä ympäri maailmaa muutamassa sekunnissa huomaamatta, että toimimme kielellä, jota emme edes ymmärrä. Euroopan komission tuoreen raportin mukaan 57% Internetin käyttäjistä Euroopassa ostaa tavaroita ja palveluja käyttäen muuta kuin äidinkieltään kaupanteossa. Englanti on kaikkein tavallisin vieras kieli, ja seuraavina tulevat ranska, saksa ja espanja. 55% käyttäjistä lukee sisältöä vieraalla kielellä, kun taas vain 35% käyttää vierasta kieltä kirjoittaessaan sähköposteja tai lisätessään kommentteja verkkoon [4]. Vielä muutama vuosi sitten englannin asema verkon lingua franca -kielenä oli kiistaton – suurin osa verkossa olevasta sisällöstä oli englanniksi – mutta tilanne on nyt ratkaisevasti muuttunut. Muilla eurooppalaisilla kielillä samoin kuin Aasian ja Lähi-idän kielillä tuotetun sisällön määrä on kasvanut räjähdysmäisesti.

Kielellisten raja-aitojen aiheuttama kuilu sähköisessä kanssakäymisessä on saanut hämmästyttävän vähän julkista huomiota. Sen tiedostaminen nostaa kuitenkin esiin oleellisen kysymyksen: Mitkä Euroopan kielistä tulevat kukoistamaan verkottuneessa tieto- ja osaamisyhteiskunnassa, ja mitkä katoamaan?

2.2 KIELET KOHTAAVAT UUSIA UHKIA

Samalla kun painotekniikka edisti tiedonvälitystä Euroopan sisällä, se myös johti monien Euroopan kielten katoamiseen. Paikallisilla kielillä ja vähemmistökielillä julkaistiin harvemmin. Joitakin kieliä, kuten kornin kieli ja dalmatian kieli, käytettiin vain suullisessa viestinnässä, mikä puolestaan rajoitti niiden käytön alaa. Tuleeko Internetillä olemaan sama vaikutus kieleemme?

> Euroopan kielten moninaisuus on sen tärkeimpiä voimavaroja.

Euroopan noin 80 kieltä muodostavat yhden sen rikkaimmista ja tärkeimmistä kulttuurien varaan rakentuvista kilpailuvalteista [5]. Vaikka isot kielet, kuten englanti ja espanja, tulevat todennäköisesti selviytymään kasvavilla digitaalisilla markkinoilla, voivat monet eurooppalaisista kielistä joutua verkostoituneessa yhteis-

kunnassa yhdentekevän kielen asemaan. Tällainen kehitys heikentäisi Euroopan asemaa maailmassa ja haittaisi Euroopan strategiaan sisältyvää tavoitetta taata kaikille Euroopan kansalaisille yhtäläinen oikeus osallistumiseen kielestä riippumatta. Unescon raportti monikielisyydestä osoittaa, että kielet ovat elintärkeitä perusoikeuksien turvaamisessa, joita ovat esimerkiksi oikeus koulutukseen, oikeus ilmaista poliittinen mielipiteensä ja oikeus osallistua yhteiskunnalliseen toimintaan [6].

2.3 KIELITEKNOLOGIA TUKEE KIELTEN SÄILYMISTÄ

Tähän asti toimenpiteet kielen säilymisen puolesta ovat kohdistuneet lähinnä kielen opetukseen ja kääntämiseen. Eurooppalaiset käännöstoiminnan, tulkkauksen ja lokalisoinnin markkinat vuonna 2008 olivat 8,4 miljardin euron arvoiset ja niiden odotetaan yhä kasvavan 10 prosentin vuosivauhdilla [7]. Luku kattaa kuitenkin vain pienen osan kieltenvälisen viestinnän nykyisistä ja tulevaisuuden tarpeista. Tavoitteena on varmistaa, että tulevaisuuden Euroopassa kansallisia kieliä voidaan käyttää laaja-alaisesti kaikkiin tarkoituksiin. Tarkoituksenmukainen teknologia on avuksi tavoitteen saavuttamisessa samalla tavoin kuin teknologia ratkaisee mm. kuljetuksen ja energiatalouden kysymyksiä ja vastaa erityisryhmien tarpeisiin.

> Kieliteknologiat auttavat meitä ottamaan osaa monikieliseen sosiaaliseen ja poliittiseen keskusteluun.

Kieliteknologian tutkimuskohteita ovat kaikki kirjoitetun ja puhutun kielen muodot. Sovellukset auttavat meitä tekemään yhteistyötä, hoitamaan liikeasioita, jakamaan tietoa ja ottamaan osaa sosiaaliseen ja poliittiseen keskusteluun kielellisistä rajoitteista ja tietotekniikan taidoista riippumatta. Usein ne toimivat apunamme näkymättömällä tavalla monimutkaisten tietokonejärjestelmien syvyyksissä ja auttavat:

- löytämään tietoa Internetin hakukoneen avulla;
- tarkistamaan tekstinkäsittelyohjelman sisällä oikeinkirjoituksen ja kieliopin;
- saamaan tuotetta koskevia suosituksia näkyviin verkkokaupassa;
- kuuntelemaan puhuttua ohjeistusta auton navigaattorista;
- kääntämään verkkosivuja verkossa olevan palvelun avulla.

Kieliteknologiat koostuvat erilaisista keskeisistä ydinteknologioista, joita käytetään laajemmissa tehtäväkokonaisuuksissa monenlaisten tehtävien suorittamiseen. Tavoitteena META-NET valkoisten kirjojen julkaisusarjassa on selvittää, missä vaiheessa eurooppalaisten kielten ydinteknologiat tänään ovat.

> Eurooppa tarvitsee vakaata, kohtuuhintaista ja tärkeimpiin ohjelmistoympäristöihin integroitua kieliteknologiaa.

Jotta voisimme säilyttää asemamme kehityksen etujoukoissa maailmassa, tarvitsemme kaikille Euroopan kielille sovitettua kieliteknologiaa, joka on vakaata, kohtuuhintaista ja tärkeimpiin ohjelmistoympäristöihin tiiviisti integroitua. Ilman kieliteknologiaa emme pääse käyttäjinä nauttimaan todella tehokkaista, interaktiivisista ja multimediaa tehokkaasti hyödyntävistä monikielisistä sovelluksista lähitulevaisuudessa.

2.4 KIELITEKNOLOGIAN MAHDOLLISUUKSIA

Painotuotteiden maailmassa todellinen teknologinen läpimurto oli paperilla olevan kuvan (tekstin) nopea

monistaminen käyettävissä olevalla tekniikalla toimivan kirjapainokoneen avulla. Ihmisten piti noina aikoina tehdä tiedon etsimisen, omaksumisen, kääntämisen ja tiivistämisen edellyttämä työ käsityönä. Puheen nauhoittamiseksi piti odottaa Edisonia – ja silloinkin tuloksena oli vain analogisia kopioita.

Nykyisin kieliteknologia tarjoaa mahdollisuuden automatisoida kääntämisen, sisällöntuotannon ja tietämyksen hallinnan prosesseja kaikilla Euroopan kielillä. Sitä tarvitaan myös mahdollistamaan helppokäyttöisiä kieleen tai puheeseen pohjautuvia käyttöliittymiä kotitalouksille suunnattuihin elektronisiin tuotteisiin, ajoneuvoihin, tietokoneisiin ja robotteihin. Vaikka kaupalliset ja teolliset sovellukset ovat todellisuudessa vielä kehityksen esiasteita, tutkimuksen ja tuotekehityksen saavutukset luovat aitoja mahdollisuuksia tulevaisuuden ratkaisuihin. Erikoisalojen konekäännös toimii esimerkiksi jo suhteellisen tarkasti, ja kokeelliset sovellukset sisältävät monikielisiä informaation ja tietämyksen hallintatyökaluja samoin kuin sisällöntuotantoa tukevia ohjelmia useilla eurooppalaisilla kielillä.

Kieliteknologia auttaa vastaamaan monikielisyyden haasteisiin.

Useimpien teknologioiden tavoin ensimmäiset kieliteknologiset sovellukset, kuten äänipohjaiset käyttöliittymät ja dialogijärjestelmät, kehitettiin hyvin erikoistuneille aloille ja niiden suorituskyky on usein rajallinen. Toisaalta opettamisen puolella ja viihdeteollisuudessa löytyy huikeita kaupallisia mahdollisuuksia integroida kieliteknologioita peleihin, kulttuuriperintösivustoihin, opetusviihdepaketteihin, kirjastojen palveluihin, erilaisiin simulaatioympäristöihin ja harjoitteluohjelmiin. Mobiilit tietopalvelut, tietokoneavusteinen kielen oppiminen, verkko-opetusympäristöt, itsearvioinnin työkalut ja plagioinnin tunnistusohjelmat ovat vain joitakin esimerkkejä sovellusaloista, joissa kielitekнologialla voi olla tärkeä rooli. Sosiaalisen median sovellusten kuten Twitterin tai Facebookin suosio osoittaa, että jatkossakin tarvitaan kehittyneitä kieliteknologioita, joiden avulla voidaan tarkkailla viestiliikennettä, tehdä yhteenvetoja keskusteluista, havaita trendejä erilaisten kyselyjen perusteella, dokumentoida tunnepohjaisia reaktioita tai tunnistaa tekijänoikeusloukkauksia.

Kieliteknologia tarjoaa Euroopan unionille monenlaisia ratkaisuja. Se auttaa meitä vastaamaan Euroopan moninaisiin monikielisyyden haasteisiin – siihen arkipäivään, jossa eri kielet elävät luonnostaan sovussa eurooppalaisessa liike-elämässä, organisaatioissa ja kouluissa. Mutta kansalaisten tulee voida kommunikoida ristiin rastiin Euroopan yhteismarkkina-alueella kielten rajojen yli – ja tätä kieliteknologia voi edesauttaa tarjoamalla ratkaisuja, jotka ovat kaikkien kansalaisten saavutettavissa ja joiden avulla kommunikointi onnistuu kaikilla kielillä. Kieliteknologia voidaan nähdä avustavana teknologiana, kun ratkaistaan kielellisen monimuotoisuuden kysymyksiä ja helpotetaan kieliyhteisöjen välistä viestintää. Eräs aktiivisista tutkimuskohteista on kieliteknologian hyödyntäminen pelastusoperaatioissa katastrofialueilla, kun toimintakyvyn ripeys on elämän ja kuoleman kysymys: tulevaisuuden useita kieliä taitavat älykkäät koneet voivat pelastaa ihmishenkiä.

Panostamalla tulevaisuudessa innovatiiviseen eurooppalaiseen monikieliseen kieliteknologiaan Eurooppa voi näyttää suuntaa muulle maailmalle.

2.5 KIELITEKNOLOGIAN HAASTEITA

Vaikka kieliteknologia on tutkimus- ja sovellusalueena jo ottanut isoja edistysaskeleita, on teknologinen edistys ja tuotekehitys nykyisellään liian hidasta. Laajalti käytössä olevat teknologiat, kuten oikeinkirjoituksen ja kieliopin tarkistusohjelmat, ovat tyypillisesti yksikielisiä ja niitä on saatavissa vain kouralliselle kieliä. Verkon tar-

joamat käännöspalvelut, vaikka ovatkin hyvä apu tiedoston sisällön likimääräisen vastineen tuottamisessa, ovat hankaluuksissa heti, kun tarvitaan oikein tarkkoja ja yhdenmukaisia käännöksiä. Ihmiskielen monimutkaisuudesta johtuen kielten mallintaminen ohjelmallisesti ja niiden testaaminen todellisessa elämässä on pitkä ja kallis liiketoiminnan muoto, joka edellyttää pitkän aikavälin rahoitussitoumuksia.

Teknologinen edistys ja tuotekehitys tapahtuvat liian hitaasti.

Euroopan tulee siksi pitää kiinni edelläkävijän roolistaan monikielisen yhteisön teknologisten haasteiden kohtaamisessa ja kehittää uusia menetelmiä kehityksen nopeuttamiseksi koko Euroopassa. Nämä voivat tarkoittaa sekä tietoteknisiä edistysaskeleita että uusia tekniikoita, kuten yleisön osallistamisen menetelmä kansalaisten tietämyksen hyödyntämisessä.

2.6 KIELEN OMAKSUMISESTA

Ennen kuin lähdemme pohtimaan tarkemmin sitä, miten tietokoneet käsittelevät kieliainesta ja miksi niitä on vaikeaa ohjelmoida hyödyntämään kieltä, tarkastelemme lyhyesti ihmisten ensimmäisen ja toisen kielen omaksumista ja sen jälkeen tutustumme tarkemmin kieliteknologisten järjestelmien toimintaan. Ihmiset oppivat kieltä kahdella tavalla, oppimalla esimerkeistä ja tekemällä niistä yleistyksiä. Vauvat omaksuvat kielen kuuntelemalla ja osallistumalla itse aitoihin vuorovaikutustilanteisiin vanhempiensa, sisarustensa ja muiden perheenjäsenten kanssa. Noin kaksivuotiaista eteenpäin lapset alkavat tuottaa sanoja ja lyhyitä fraaseja itse. Tämä on mahdollista ainoastaan siksi, että ihmisillä on geneettinen taipumus matkimiseen ja kuulemansa puheen analysointiin.

Ihmiset oppivat kieltä kahdella tavalla, oppimalla esimerkeistä ja tekemällä niistä yleistyksiä.

Vanhempana lapsen vieraan kielen oppiminen vaatii enemmän vaivannäköä, pääosin siksi, että oppija ei enää ole osa kieltä äidinkielenään puhuvien kieliyhteisöä. Koulussa vieraat kielet usein omaksutaan opettelemalla kielen kieliopillista rakennetta, sanastoa ja oikeinkirjoitusta harjoitusten avulla, jotka kuvaavat käsitystämme kyseisestä kielestä abstraktien sääntöjen, taulukoiden ja esimerkkien kautta. Vieraan kielen oppiminen vaikeutuu iän myötä. Kieliteknologisten menetelmien kaksi päätyyppiä oppivat tietoa kielestä samalla tavoin. Tilastolliset (tai 'aineistolähtöiset') lähestymistavat eristävät kielitietoa valtavista aitojen esimerkkitekstien kokoelmista. Vaikka esimerkiksi oikeinkirjoituksen tarkistimelle riittää harjoitusaineisoksi yksikielinen teksti, konekäännösjärjestelmien treenaamiseen tarvitaan rinnakkaistekstejä kahdesta tai useammasta kielestä. Konekäännösalgoritmi oppii niiden rakenteita ja päättelee, miten sanat, lyhyet fraasit ja kokonaiset virkkeet on niissä käännetty.

Kieliteknologisten menetelmien päätyypit oppivat tietoa kielestä samalla tavoin.

Tilastollinen lähestymistapa saattaa edellyttää miljoonien virkkeiden aineistoa, ja menetelmien laatu paranee analysoidun tekstin määrän kasvaessa. Tämä on yksi syy siihen, että hakukoneiden kehittäjät keräävät niin suuria määriä kirjoitettua kieliainesta kuin mahdollista. Google-haku ja Googlen Kääntäjä perustuvat kaikki tilastollisiin menetelmiin. Tilastoista saatava suuri hyöty syntyy koneen kyvystä oppia nopeasti sille jaksoittaisena tarjotusta harjoitusaineksesta, vaikkakin oppimistulosten laatu voi vaihdella.

Toinen kieliteknologian ja erityisesti konekääntämisen lähestymistapa on sääntöpohjaisten järjestelmien rakentaminen. Kielitieteen, tietokonelingvistiikan ja tietojenkäsittelytieteen asiantuntijat koodaavat aluksi kieliopillisia analyysejä (kääntämisen sääntöjä) ja kokoavat sanastoja (leksikkoja). Jotkin johtavista sääntöpohjaisista konekäännösjärjestelmistä ovat olleet tekeillä jo yli kaksikymmentä vuotta. Sääntöpohjaisten järjestelmien suuri etu piilee siinä, että asiantuntijat voivat kontrolloida kielen prosessointia tarkemmin. Näin heidän on mahdollista korjata ohjelman virheitä systemaattisesti ja antaa yksityiskohtaista palautetta käyttäjälle, erityisesti tilanteessa jossa sääntöpohjaisia järjestelmiä käytetään kielen oppimisessa. Mutta työn kalleudesta johtuen on sääntöpohjaisia kieliteknologisia menetelmiä tähän asti kehitetty vain isoille kielille.

Koska tilastollisten ja sääntöpohjaisten järjestelmien vahvuudet ja heikkoudet tapaavat olla toisiaan täydentäviä, tutkimushankkeissa keskitytään molemmat menetelmät yhdistäviin hybridimalleihin. Näiden osalta menestystä on toistaiseksi koettu enemmän tutkimuslaboratoriossa kuin teollisten sovellusten maailmassa.

Kuten olemme tässä osiossa nähneet, monet nykyisessä informaatioyhteiskunnassa hyödynnettävät sovellukset perustuvat kieliteknologisiin menetelmiin. Tämä on erityisen tyypillistä Euroopan monikieliselle talousmarkkinoiden ja tiedonjaon alueelle. Vaikka kieliteknologian parissa on viime vuosina saavutettu merkittäviä edistysaskeleita, on kieliteknologisten järjestelmien laadullisessa parantamisessa vielä valtavasti työtä ja mahdollisuuksia. Seuraavissa osioissa tarkastellaan suomen kielen roolia eurooppalaisessa tietoyhteiskunnassa ja arvioidaan kieliteknologian tämänhetkistä tilaa suomen kielen näkökulmasta.

http://www.springer.com/978-3-642-27247-9

The Finnish Language in the Digital Age
Rehm, G.; Uszkoreit, H. (Eds.)
2012, VI, 81 p. 24 illus. in color., Softcover
ISBN: 978-3-642-27247-9