

Kieliteknologia käyttää tietojenkäsittelyn ja tekoälyn tutkimuksen teorioita ja työkaluja

Uusia uria kielitieteen ikuisiin kysymyksiin

Nousussa olevan tieteenhaaran ajatusmallit voivat joskus levitä pitkälle muiden tieteiden alueelle. 1900-luvun jälkipuoliskolla

niin sanottu tietokone-metafora on vaikuttanut vahvasti ihmistieteisiin. Henkisiä toimintoja, kuten ajattelua tai kielen käyttöä, kuvataan tietokoneiden rakenteen, toimintatapojen ja niiden ohjelmoinnin teorian avulla.

Esimerkiksi vuosisadan alkupuolella kielitieteessä syntynyt struktuurin käsite levisi muihin tieteisiin kansantaloustiedettä ja biologiaa myöten.

Kieliteknologia ei ole itsenäinen ala, vaan sateenvarjonimitys eräille mainitun kaltaisille kehityskaarille. Tekniseltä kalskahtavasta nimestä huolimatta kysymys on humanistisesta perustutkimuksesta. Tarkoituksena on ymmärtää luonnollisten kielten ominaisuuksia entistä syvällisemmin.

Sen ongelmia ovat muun muassa: Voidaanko kielten sanojen rakenne, esimerkiksi taipuminen ja yhdyssanojen muodostuminen, kuvata sanojen yleisten periaatteiden mukaan?

Kuinka lauseiden rakenne voidaan päätellä niistä tiedoista lähtien, joita sisältyy lauseen sanoihin ja niiden päätteisiin?

Miten yksittäisten sanojen merkitys on riippuvainen tekstiyhteydestä, jossa ne esiintyvät? Miten tekstin sisältö ja tyyppi on pääteltävissä siinä enemmän vai vähemmän suorasti ilmaistujen merkitysten avulla?

Kieliopilla automaattinen lauseenjäsennys

Tämänkaltaiset ongelmat ovat oikeastaan kielitieteen ikuisuus-kysymyksiä. Uutta on ennen kaikkea niiden tarkastelu käyttäen apuna formaalisten kielten, tietojenkäsittelytieteen sekä tekoälyn tutkimuksen teorioita, hahmotuksia ja työkaluja.



■ Tietokoneiden suorituskyvyn huima kasvu on kehittänyt ratkaisevasti kieliteknologiaa, kirjoittaa Fred Karlsson.

Silti kieliteknologia liittyy läheisesti yliopistollisiin oppiaineisiin yleinen kielitiede ja tietokone-lingvistiikka.

Toinen uusi ulottuvuus on kielen rakenteiden ja merkitysten tutkiminen tulkinallista päättelyä vaativina ongelmina. Esimerkiksi sana "kuusi" voi vaihtoehtoisesti tarkoittaa joko lukusanaa tai tiettyä puulajia (tai sen yksittäistä puuta).

Tietokoneiden suorituskyvyn huima kasvu on vaikuttanut ratkaisevasti kieliteknologian kehittymiseen, samoin entistä huomattavasti suurempien tekstimassojen helppo saatavuus koneismuodossa. Yhdeksi perusmenetelmäksi onkin vakiintunut jostakin kielestä laadittujen kuvausmallien testaaminen suurissa tekstimassoissa.

Tutkimuksemme ensimmäinen läpimurto oli Kimmo Koskenniemen väitöskirja vuodelta 1983. Siinä hän esitti sanojen rakenteen automaattista käsittelyä varten yleisen teorian, joka kulkee nimellä kaksitasomorfologia. Mallia on sittemmin sovellettu kymmeneen kieleen arabiaa, japania ja samoja myöten. Siitä on käytännössä tullut tietokone-morfologian kansainvälinen standardi. Koskenniemi nimettiin 1992 Helsingin yliopis-

ton tietokone-lingvistiikan professoriksi.

Edistysaskeleeksi on myös osoittautunut rajoituskieliopin teoria, jonka esitin 1990. Sen avulla kielille voidaan kirjoittaa kielioppeja, joilla voidaan suorittaa automaattinen lauseenjäsennys. Kimmo Koskenniemi, Atro Voutilainen ja Pasi Tapanainen ovat merkittävällä tavalla vieneet eteenpäin rajoituskieliopin teorian.

Tutkimamme Atro Voutilainen, Juha Heikkilä ja Arto Anttila ovat näitä teorioita hyödyntäen ja niitä toteuttavia tietokoneohjelmia käyttäen laatineet englannin kielen analyysiohjelmiston. Siihen kuuluu koneismuotoinen sanakirja, morfologia eli sanarakenteen analysointori ja syntaksi eli lauseenjäsennin.

Kokonaisuus on herättänyt runsaasti kansainvälistä huomiota. Kun suuri sanakirjakustantaja HarperCollins Publishers pari vuotta sitten päätti perustaa peräti 200 miljoonan sanan tekstipankin "Bank of English", helsinkiläinen tutkimusyksikkömme valittiin suorittamaan koko tekstimassan automaattinen analyysi.

Tämä on laajin minkään kielen tekstianalyysi. Työn suoritti opiskelija Timo Järvinen, jolla siis lienee eräänlainen tekstianalyysin maailmanennätys.

Tieteellisten teorioiden tärkeä piire on yleisyys. Meidän tapauksessamme se tarkoittaa muun muassa pyrkimystä teorioihin, joilla (toivon mukaan) voitaisiin kuvata mitä tahansa kieltä.

Olemme itse soveltaneet teorioitamme englannin lisäksi suomeen, ruotsiin, venäjään (Liisa Vilkki) ja saksaan (Mariikka Haapalainen, Ari Majorin). Professori Arvi Hurskainen on tehnyt swahilinkielisen sanojen analyysiohjelman.

Tällä hetkellä olemme yhteistyössä muun muassa San Sebastianin yliopiston tutkijoiden kanssa ohjelmiamme soveltamisesta baskin kielen automaattiseen analyysiin. Tukholman ja Uumajan yliopiston kanssa syntyy miljoonasanainen ruotsin kielen perusaineisto. Parhaiden kansainvälisten tutkijakoulujen hyödyntäminen on tärkeää.

Monen tutkijaryhmän arkea on kil-

pailu EU:n rahoituksesta. Yhdessä laajassa informaationhakua koskevassa Esprit-hankkeessa olemme jo olleet mukana. Nyt neuvottelemme mukanaolosta kahdessa tutkimuskonsortiossa.

Toinen kehittäisi menetelmiä ohjelmavirtaavan tekstin lajitteluun sisältöluokkiin. Siinä olisi mukana myös useita tietotoimistoja. Toisen hankkeen tavoitteena on luoda laajoja standardoituja tekstimassoja ja perustavanlaatuisia koneismuotoisia sanakirjoja kaikille Euroopan keskeisille kielille. Humanisteilla voi siis hyvin olla kysyntää EU:n teknologia-hankkeissakin.

Kieliteknologia kansallisessa tietostrategiassa

Kieliteknologia on näkyvästi mukana Suomen uudessa kansallisessa tietotekniikkastrategiassa. Aikamme muoti-ilmauksia on "verkottuva tietoyhteiskunta". Jos tästä on tulevaisuutta, on välttämätöntä parantaa mahdollisuuksia luonnollisten kielten monipuoliseen käyttöön, kun kommunikoidaan automaattisten tietolähteiden kanssa.

Tällaisia tavoitteita ei voi edistää ilman pitkäjännitteistä teoreettista perustutkimusta. Ryhmän tutkijoiden ajankohtaisia ongelmia ovat lauseenjäsennys-teorian rikastaminen tekstissä esiintyvien termien tunnistaminen, informaatiohakumenetelmien kehittäminen luonnollisia kieliä käyttäen, monimerkityksisten sanojen tulkin sekä tekstin sisällön päättely ja luokitus.

FRED KARLSSON

■ Kirjoittaja on Helsingin yliopiston yleisen kielitieteen professori ja monimerkityksisten sanojen kieliteknologian tutkimusryhmän johtaja.

■ Kirjoitussarjassa esitellyt Suomen Akatemian valitsemat Suomen tieteen huippuyksiköt.