

UUSIA URIA KIELITIETEEN IKUISIIN KYSYMYKSIIN

Tietokoneiden suorituskyvyn hulma kasvu on kehittänyt ratkaisevasti kielitekniologiaa.

Nousussa olevan tieteenhaaran ajatusmallit voivat joskus levitä pitkälle muiden tieteiden alueelle. 1900-luvun jälkipuoliskolla niin sanottu tietokonemetafora on vaikuttanut vahvasti ihmistieteisiin. Henkisiä toimintoja, kuten ajattelua tai kielen käyttöä, kuvataan tietokoneiden rakenteen, toimintatapojen ja niiden ohjelmointiteorian avulla.

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

Kielitekniologia ei ole itsenäinen ala, vaan sateenvarjonimitys erälle 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 samojen 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 ikuisuuskyymyksiä. Uutta on ennen kaikkea niiden tarkastelu käyttäen apuna formaalisten kielten, tietojenkäsittelytieteen sekä tekoälyn tutkimuksen teorioita, hahmotuksia ja työkaluja.

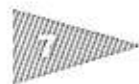
Silti kielitekniologia liittyy läheisesti yliopistollisiin oppiaineisiin yleinen kielitiede ja tietokoneolingvistiikka.

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

Tietokoneiden suorituskyvyn hulma kasvu on vaikuttanut ratkaisevasti kielitekniologian kehittämiseen, 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 kieliin arabiaa, japania ja samojedia myöten. Siitä on käytännössä tullut tietokonemorfologian kansainvälinen standardi. Koskenniemi nimettiin 1992 Helsingin yliopiston tietokoneolingvistiikan professoriksi.

Edistysaskeleeksi on myös osoittautunut rajoituskieliopin teoria, jonka esitin 1990. Sen avulla kielille voidaan kirjoittaa kielioppeja, joilla



Tavallinen kielen käyttäjä ei yleensä ole juurikaan selvillä yksittäisten sananmuotojen mahdollisista useammista tulkinnoista. Esimerkiksi lukiessaan seuraavan lauseen, sen jokainen sana saattaa tuntua itsestään selvältä ja ongelmattomalta:

Persianlahdelta pääsi kuusi alusta ennen taistelujen alkua.

Kuitenkin varsin monet sananmuodoista osoittautuvat moniselitteisiksi, kun niitä tarkastellaan erillisinä. Seuraavassa sananmuotoja tunnistavan Kimmo Koskenniemen kehittämän FINTWOL-nimisen ohjelman analyysi virkkeen kustakin sanasta:

persian-		alusta	("alku" N ELA SG)
lahdelta	("persian_lahti" PROP N ABL SG)		("alusta" N NOM SG)
pääsi	("pää" N NOM PL 2SG)		("alustaa" V IMPV ACT NEG SG)
	("pää" N NOM SG 2SG)		("alustaa" V IMPV ACT SG2)
	("pää" N GEN SG 2SG)		("alustaa" V PRES ACT NEG)
	("päästä" V PAST ACT SG3)		("alus" N PTV SG)
kuusi	("kuusi" NUM NOM SG)		("alunen" N PTV SG)
	("kuu" N NOM PL 2SG)	ennen	("ennen" PP)
	("kuu" N NOM SG 2SG)		("ennen" ADV)
	("kuu" N GEN SG 2SG)	taistelujen	("taistelu" DV-U N GEN PL)
	("kuusi" N NOM SG)	alkua	("alku" N PTV SG)

Havaitsemme, että *pääsi* voisi olla myös *pää*-sanana possessiivisuffiksillinen muoto, *kuusi* voisi myös olla *puu* ja peräti *kuu*-sanana possessiivisuffiksillinen muoto. Monitulkintaisiin on *alusta*, joka voisi olla myös *alku*-, *alusta*-, tai *alunen*-substantiivin muoto sekä *alustaa*-verbin erinäisiä muotoja. Sopivassa asiayhteydessä kukin näistä ylimääräisistä tulkinnoista tuntuu puolestaan itsestään selvältä.

Eräs lauseenjäsenyyksen ensimmäisistä tehtävistä on siis sananmuotojen yllä kuvattujen ylimääräisten tulkintojen eliminointi. Tehtävän automaattinen ratkaisu on osoittautunut varsin vaikeaksi, sillä tietokoneelle ei voida nykyisellään ohjelmoida kielen merkitystä ja tietoja ympäröivästä maailmasta läheskään riittävässä määrin. Pelkän lauserakenteen perusteella ongelma ratkeaa kuitenkin pääosin, kuten esim. englannin kielen osalta Atro Voutilainen tutkimusyksiköstämme on osoittanut.

voidaan suorittaa automaattinen lauseenjäsennys. Kimmo Koskenniemi, Atro Voutilainen ja Pasi Tapanainen ovat merkittäväällä tavalla vieneet eteenpäin rajoituskieliopin teoriaa.

Tutkijamme 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 cräänlainen tekstianalyysin maailmanennätys.

Tieteellisten teorioiden tärkeä piirre 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 tär-

keää.

Monen tutkijaryhmän arkea on kilpailu EU:n rahoituksesta. Yhdessä laajassa informaatioonhakua koskevassa Esprit-hankkeessa olemme jo olleet mukana. Nyt neuvottelemme mukanaolosta kahdessa tutkimuskonsortiossa.

Toinen kehittäisi menetelmiä ohivirtaavan tekstin lajitteluun sisältöluokkiin. Siinä olisi mukana myös useita tietotoimistoja. Toisen hankkeen tavoitteena on luoda laajoja, standardoituja tekstimassoja ja perustavantarkeitä koneismuotoisia sanakirjoja kaikille Euroopan keskeisille kielille. Humanisteilla voi siis hyvinkin olla kyntää EU:n teknologiahankkeissakin.

Kieliteknologia kansallisessa tietostrategiassa

Kieliteknologia on näkyvästi mukana Suomen uudessa kansallisessa tietotekniikkastrategiassa. Aikamme muoti-ilmauksia on "verkottuva tietoyhteiskunta". Jos tästä on tuleva totta, 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 lauseenjäsennyksen teorian rikastaminen, tekstissä esiintyvien termien tunnistaminen, informaatiohakumenetelmien kehittäminen luonnollisia kieliä käyttäen, monimerkityksisten sanojen tulkinta sekä tekstin sisällön päättely ja luokitus.

FRED KARLSSON

Yleisen kielitieteen professori

Helsingin yliopisto