Historiallisen tiedon louhinnan tulevaisuus puhutti Transkribus-yhteisöä Innsbruckissa

Käsin kirjoitetun tekstin tekoälypohjainen tunnistaminen (HTR) on viime vuosina kasvattanut merkitystään historiantutkimuksen apuvälineenä. Tähän on vaikuttanut tunnistamisen teknologian kehitys, joka on ollut miltei yhtä rajua kuin ChatGPT:n kaltaisten tekoälysovellustenkin.

Yksi HTR-teknologian kehittäjistä on READ-COOP-osuuskunta, jonka Transkribus-alusta on räätälöity nimenomaan käsin kirjoitettujen historiallisten dokumenttien prosessointiin. Transkribuksella on jo 180 000 käyttäjää ympäri maailmaa, ja alustan kehittäjien mukaan sen avulla on prosessoitu yli 80 miljoonaa sivua.

READ-COOP pyrkii luomaan Transkribuksen käyttäjäkunnasta yhteisön, joka osallistuu sovelluksen kehittämiseen ja virittelee sen ympärille yhteisiä hankkeita. Yhteisön rakentamisen keskeisenä välineenä ovat tapahtumat, kuten Transkribus User Conference, joka järjestettiin tänä vuonna viidettä kertaa osuuskunnan kotipaikassa Innsbruckissa 15.–16.2.2024.

Konferenssin teemana oli tiedonlouhinnan tulevaisuus. READ-COOPin edustajien puheenvuorot konferenssissa toivat hyvin esiin, että osuuskunnalla on louhinnan tehostamiseksi suuria suunnitelmia. Niiden toteuttaminen vaatii kuitenkin kasvavia resursseja, joita on hankittava muun muassa käyttäjämaksuilla.

Riittävän tulovirran varmistamiseksi osuuskunta uudisti vastikään Transkribuksen tilausmallit kuukausitilauspohjaisiksi. Alustaa voi edelleen käyttää ilmaiseksi, mutta ilmaistilauksella voi prosessoida vain sata sivua kuukaudessa. Tutkijoille markkinoidaankin maksullista tilausvaihtoehtoa, joka tarjoaa myös erilaisia lisätoimintoja ja -palveluita.

Suurten tekstimassojen prosessointi nielee Transkribuksen käyttäjiltä rahaa. Halvimmat tekstintunnistushinnat osuuskunta tarjoaa jäsenasiakkailleen, joiden joukossa ovat Suomesta Kansallisarkisto ja Svenska litteratursällskapet. Jotkin organisaatiot pitävät Transkribusta kuitenkin liian kalliina miljoonien sivujen työstämiseen ja ovat kehitelleet sen rinnalle tai korvikkeeksi omia käsin kirjoitetun tekstin tunnistusjärjestelmiään.

Markkinoilla on myös käyttäjille täysin ilmaisia HTR-sovelluksia. Konferenssissa READ-COOPin väki kuitenkin pyrki erottamaan Transkribuksen kilpailijoistaan korostamalla sen kykenevän paljon muuhunkin kuin tekstintunnistukseen. Sovelluksella voi esimerkiksi kouluttaa tekstikenttä- ja taulukkomalleja, jotka helpottavat kaavaketyyppisten aineistojen prosessointia.

Lisäksi Transkribuksen avulla aineistoja voi rikastaa eri tavoin, ja tähän avautuu jo kuluvan vuoden uusia mahdollisuuksia. Sovellukseen aiotaan yhdistää esimerkiksi nimettyjen kokonaisuuksien tunnistamisen (NER) työkalu, jota voidaan opettaa tunnistamaan prosessoiduista teksteistä esimerkiksi nimiä, rahasummia tai lyhenteitä.

Myös laajojen kielimallien sisällyttämisestä Transkribukseen käytiin runsaasti keskustelua konferenssissa, ja sovelluksen kehittäjät etsivät tähän ratkaisuja. Laajojen kielimallien kuten OpenAI:n GPT:n käyttöön liittyy kuitenkin monia ongelmia, joista tietosuojaan liittyvät eettiset kysymykset ja energiankulutus eivät ole vähäisimpiä.

READ-COOP tekee yhteistyötä myös Wikimediasäätiön kanssa. Säätiön ohjelmistovastaava Satdeep Gill puhui konferenssissa pilottihankkeista, joissa Transkribuksen teknologiaa hyödynnettiin käsin kirjoitettujen lähdedokumenttien viemiseen Wikisource-alustalle avoimesti saataville.

Tekstintunnistuksen saralla Transkribus tuo käyttäjien ulottuville niin kutsuttuja supermalleja, jotka sopivat useilla kielillä kirjoitettujen tekstien tunnistamiseen. Konferenssissa esitelty Text Titan II -supermalli kattaa 17 kieltä, suomi niiden joukossa. Mallin opettamiseen on käytetty 85 miljoonan sanan tekstiaineistoa, ja se julkaistaan käyttäjille vielä tänä keväänä.

Suomalaisittain kiinnostava on myös 1600–1800-luvuilla kirjoitettujen ruotsinkielisten tekstien tunnistamiseen tarkoitettu malli The Swedish Lion I, jota esi-

teltiin konferenssissa. Mallin kehittämiseen on useiden ruotsalaisten organisaatioiden ohella osallistunut Suomen Kansallisarkisto, ja työssä on hyödynnetty joukkoistamista.

Konferenssin mielenkiintoisimpaan antiin lukeutui työpaja, joka käsitteli Transkribusta hyödyntävän projektin toteuttamista ideointivaiheesta rahoituksen hakemiseen, datanhallintasuunnitelman tekemiseen ja projektinhallintaan saakka. Työpaja toi esiin, että Transkribuksella työskentely ei merkitse pelkkää fyysisen aineiston muuntamista digitaaliseksi tekstiksi vaan sen tuloksena syntyy kokonaan uudenlainen rikastettu aineisto, jota voi käyttää tutkimuksessa hyödyksi aivan eri tavoin kuin fyysisisiä objekteja.

Aineistojensa kanssa yksin puurtavalle tutkijalle konferenssi osoitti myös selvästi, että informaation erottelu historiallisista aineistoista onnistuu tehokkaimmin ryhmässä. Transkribukseen ollaankin tuomassa lisää ominaisuuksia, jotka helpottavat ryhmätyöskentelyä ja tuovat yhteen samantyyppisillä tai -kielisillä aineistoilla operoivia käyttäjiä.

Yksi konferenssin pääpuhujista oli kielentutkija Christa Schneider Bernin yliopistosta. Hän esitteli käytännön esimerkkejä siitä, miten erilaisia koneoppimiseen perustuvia menetelmiä voi käyttää informaation erotteluun suurista historiallisista aineistoista. Schneider on tutkinut muun muassa Bernin noitaoikeudenkäyntien asiakirjoja 1500- ja 1600-luvuilta ja käyttänyt tutkimuksessaan esimerkiksi nimettyjen kokonaisuuksien tunnistamisen ja sentimenttianalyysin työkaluja. Esitelmässä nousivat esiin myös näiden menetelmien ongelmat, kuten se, että työkalujen algoritmit on usein esikoulutettu nykypäivän aineistoilla.

Transkribuksen käyttöä opetuksessa käsitteli muun muassa Jérémy Scraire Montrealin yliopistosta. Hän esitteli toisen asteen opiskelijoille suunnattua kurssiaan, jonka aikana opiskelijat prosessoivat ryhmissä Transkribuksella 200-sivuisen matkapäiväkirjan 1910-luvulta, analysoivat sitä ja perehtyivät samalla historialliseen kontekstiin. Scrairen mukaan opiskelijat selvisivät matkapäiväkirjan tekstintunnistuksesta hyvin, mutta tekstin annotointi osoittautui heille haastavaksi. Projekti vei opiskelijoilta myös oletettua enemmän aikaa ja kirvoitti valituksia työmäärästä. Itselläni on samantapaisia kokemuksia historianopiskelijoiden kanssa tehdyistä Transkribus-harjoituksista.

Konferenssin puhujista valtaosa prosessoi Transkribuksella latinalaisin aakkosin kirjoitettuja tekstejä, mutta esillä oli myös sovelluksen hyödyntäminen aasialaisilla kielillä tai jiddišillä kirjoitettujen tekstien tunnistamiseen. Lisäksi sovellusta on käytetty nuottien ja jopa kuvien tunnistamiseen.

Esiintyjiä konferenssissa oli yli 30 maasta, kaukaisimmat Argentiinasta ja Japanista, mutta valtaosa toki Euroopasta. Suomalaisedustus oli tällä kertaa harvalukuinen. Svenska litteratursällskapet esitteli julisteella projektiaan, jossa Transkribuksella on prosessoitu historiallisia ruokareseptejä 1700- ja 1800-luvuilta. Omassa esitelmässäni käsittelin projektiani, jossa olen hyödyntänyt Transkribusta kansalaisten vallankäyttäjille osoittamien kirjeiden puhtaaksikirjoittamiseen ja rikastamiseen.

Konferenssi osoitti, että READ-COOP kurottaa moneen suuntaan ja pyrkii palvelemaan hyvin erilaisia käyttäjäryhmiä ja -tarpeita. Tämä puolestaan edellyttää jatkuvaa kasvua, jota osuuskunta yrittää edistää ponnekkaalla markkinointi- ja brändityöllä. Hetkittäin konferenssin tunnelma muistuttikin herätysjuhlaa tai Slush-tyyppistä kasvuyritystapahtumaa. Konferenssin järjestäjistä ja osanottajista huokuneen innostuksen perusteella Transkribukselle voi kuitenkin ennustaa valoisaa tulevaisuutta.

Sami Suodenjoki