Sata ja yksi tapaa ilmaista topiikkia

Kirsi Sandberg & Juho Härme

Tutkimuksen ensimmäisessä vaiheessa aineistoon valikoidut tekstit analysoitiin kappaleittain[1] siten, että jokaiselle kappaleelle merkittiin topiikki eli se, mistä kappale kertoo (vrt. Lambrechtin määritelmä). Yhteensä neljäkymmentä tekstiä[2] kattaneen kokeiluaineiston perusteella päätettiin keskittyä *asumisesta* kertoviin kappaleisiin siitä syystä, että asumista käsiteltiin lähes jokaisessa analysoidussa näytetekstissä ja että asuminen aiheena havaittiin tavallisesti melko selvärajaiseksi. Koska tekstejä analysoi kaksi eri henkilöä (kumpikin 20 tekstiä), näyteaineiston avulla testattiin myös, kuinka samanlaisia tulkinnat kappaleiden aiheista olivat. (tähän joku maininta niistä tuloksista).

Kun kaikki tekstit kattava varsinainen kappalekohtainen analyysi saatiin päätökseen, lopulliseksi aineistoksi muodostui 415 asumisesta kertovan kappaleen sisältävää tekstiä. Tekstit sinänsä ovat keskimäärin 945 sanaa pitkiä ja sisältävät kaiken kaikkiaan keskimäärin 13 kappaletta, joista vähintään yhden topiikiksi on edellä kuvatussa tutkimusvaiheessa analysoitu asuminen.

Asumisesta kertovien kappaleiden syntaktisesta rakenteesta saatiin yleiskuva annotoimalla kappaleet koneellisesti dependenssijäsentimellä (ks. Haverinen ym. 2014). Koneellisesti tuotettu annotointi toimi lähtökohtana, kun kappaleita ryhdyttiin luokittelemaan eri ryhmiin sen mukaan, miten niissä indikoitiin asumistopiikkia.

Miten selvitettiin indikaattorisana=johdosasia.

(ps. liitä mukaan jonnekin linkki githubiin tilastoja ajatellen?)

(jotain tyyliin että mitä kaikkia automaattisesti tunnistettavissa olevia ominaisuuksia kullakin ryhmällä on. . . ?)

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski & Filip Ginter 2014. Building the essential resources for Finnish: The Turku Dependency Treebank. *Language Resources and Evaluation* 48:3, 1–39.

[1] Käyttäen kirjoittajien itsensä tekemeää kappaleiden ortografista erottelua.
[2] Tarkista määrä