Korpusten siirrosta (11.7.2007):

Laitoin ensimmäiset esimerkkikorpukset CSC:lle viime perjantaina (6.7.2007). Kysymyksessä on inkeroisen ja hantin korpukset, joissa kummassakin on morfologisesti koodattua aineistoa ja jotka on kokonaan (inkeroinen) tai osaksi (hanti) käännetty jollekin toiselle kielelle (englanti, saksa, venäjä). Tavoitteena on, että käytännössä kaikki korpukset ovat siinä kunnossa, että ne voitaisiin siirtää CSC:lle ensi viikon jälkeen. Ongelmana on vielä sopimukset (ks. alla). Työvaiheista:

Metadata-tiedostot:
Olen käynyt läpi korpusten metadata-tiedostoja ja korpusten kuvauksia. Jonkin verran on löytynyt virheitä: vastuu on minun, koska en ehtinyt tarkastaa kuvauksia silloin, kun tutkimusavustaja laati ne. Kuitenkin työ on nyt pian valmis. Kopioin korpusten metadata-tiedostot kuhunkin korpushakemistoon, mistä niitä voidaan katsella, jos joskus vielä päästään siihen, että korpuksia voi käyttää jonkin käyttöliittymän kautta. Toimitan mukaan myös tiedoston, johon olen optimistina koonnut korpusten polut korpusten luo UHLCS:ssa. Metadatan yksi idea on, että metadata-tiedostot ovat julkisesti avoimia ja siis jollakin verkkoselaimella luettavissa. Alkuperäiset tiedostot jäävät UHLCS:n verkkosivun yhteyteen.

Korpukset:
Korpuksissa itsessään ei ole paljon tekemistä tässä vaiheessa. Jos aikaa jää, työstän korpusten UNICODE-konvertointeja ja täydennän rakennekuvauksia.

Sopimukset:
Olen jatkanut yhteydenpitoa korpusten omistajiin. Kaikkiin en ole vielä saanut yhteyttä: tämä heinäkuu on lomakuukausi eivätkä kaikki ole olleet viekä tavoitettavissa. Jatkan työtä. Kun olen saanut metadatakuvaukset valmiiksi, teen luettelon niistä korpuksista, joiden sopimukset eivät ole kunnossa. Tämän ehdin tekemään ensi viikon puolella.

Ryhmät:
Ryhmien selvitystä koskeva työ jää kuun loppuun ja ensi kuun alkuun.

Terveisin,
Pirkko