UHLCS:n korpusten siirto CSC:n koneelle: täydennys loppuraporttiin
(Pirkko Suihkonen)

1. Osa korpuksista (lueteltu tiedostossa ”puuttuvat-2-10-2007”) jätettiin siirtämättä siitä syystä, että halusin ensin selvittää, oliko aineistojen joukossa yksityishenkilöiden tiedostoja. Tiedustelin Jussi Piitulaiselta ja Sari Salmisuolta, oliko heidän henkilökohtaisia aineistojaan korpushakemistossa. Jussi poisti omat tiedostonsa. Sarilla ei ollut enää käyttölupaa laitoksen koneelle. Visa Rauste pyysi Hanna Westerlundia aktivoimaan tilapäisesti Sarin käyttäjätunnuksen 12.9.2007. Koska Hanna ei ollut ehtinyt aktivoimaan tunnusta 1.10.2007 mennessä, siirsimme aineistot Jukka Huhdan kanssa CSC:n koneelle 1.10.2007. Aineistot koostuivat laitoksella laaditusta materiaalista.

2. Edellisen siirron yhteydessä oli jäänyt siirtämättä osia ftc-hakemistosta. Ne siirrettiin Jukka Huhdan avustuksella CSC:n koneelle 2.10.2007.

3. Suuri osa korpuksista, jotka siirrettiin 1.10., siirrettiin hakemistoon /ADM/finnish, koska niissä oli aineistoja, jotka voidaan laittaa julkiseen käyttöön vasta sitten, kun on selvitetty niitä koskevat sopimukset. Korpukset, jotka siirrettiin 2.10, laitettiin kaikki hakemistoon /ADM/finnish/ftc-tmp. Ennen kuin ne voidaan siirtää julkiseen käyttöön, on selvitettävä niihin liittyvät sopimukset. Mukana on esim. laaja hakemisto /simple, jossa on SIMPLE-projektin aineistoa. 1.10. siirretystä parole-korpuksesta jätettiin hakemistoon /general-linguistics-kotus/uralic-lgs/finnish/parole vain alihakemisto /pankki, joka käsittää parole-korpuksen.

4. Sari Salmisuon nimellä olleet aineistot olivat tavallisia työstettyjä korpuksia, ja sama koski muita eri työntekijöiden nimellä nimetyissä hakemistoissa olleita aineistoja. Hakemistossa /katariina oli aineistojen joukossa joitakin sellaisia, joiden Jukka Huhta arveli olevan henkilökohtaisia. Ne aineistot jätettiin siirtämättä. Yritin selvittää laitoksen projektissa työskennelleen Katariinan yhteystietoja voidakseni tiedustella häneltä aineistoista, mutta en siinä onnistunut. Oletusarvona voidaan kuitenkin pitää sitä, että kaikki sellaiset aineistot, joiden voidaan päätellä olleen valmistetun laitoksen hankkiman rahoituksen turvin, on siirretty CSC:lle.

5. Eero Vitie on pyytänyt, että parole-korpus, joka on jo CSC:llä, laitetaan vain ryhmän li-adm käyttöön, ja jo CSC:llä oleva parole-korpus jää avoimeksi laajemmalle käyttäjäkunnalle. Jo laitoksella parole-korpus oli vain parole-ryhmän käytössä. Ryhmä ”parole” on jo poistettu hakemistosta /etc. Ryhmä oli suppea: siihen kuului laitokselta vain henkilöitä, jotka työskentelivät LE PAROLE -hankkeessa tai joilla oli jokin läheinen yhteys hankkeeseen. Ehdotan, että nytkin hakemisto laitetaan avoimeksi ryhmälle li-a1, johon kuuluu yleisen kielitieteen laitoksen henkilökunta, laitoksen tutkijat ja opettajat, ja jonka alue kattaa tutkimuksen. Tämä sama koskee myös englannin susanne-korpusta, josta siitäkin on kopio CSC:n koneella. Koska käyttäjäryhmä on pieni, ei ole oletettavissa, että niiden käyttö haittaa muiden korpusten käyttöä tai vaikuttaa niiden käyttöön.

Pirkko Suihkonen, 5.10.2007

5. Olen toimittanut 3.10. Fredille listan korpuksista, joiden sopimuksia ei ole ollut työn aikana saatavilla.

Pirkko Suihkonen, 5.10.2007