Salo, O.-P., T. Nikula & P. Kalaja (toim.) 2007. Kieli oppimisessa – Language in Learning. AFinLAn vuosikirja 2007. Suomen soveltavan kielitieteen yhdistyksen julkaisuja no. 65. Jyväskylä. s. 197–213.

▼▼▼▼

TESTIAINEISTOSTA KIELENOPPIJAKORPUKSEKSI

Mirja Tarnanen
Jyväskylän yliopisto

This article concerns language learner corpus based on data from the National Certificate of Language Proficiency (NC) that provides tests in nine different languages on three different examination levels for adult language learners. The quantitative data of the corpus consist of background information of test takers, and test scores of reading and listening comprehension, speaking, writing and structure and vocabulary tests in numeric form, and the qualitative data contain speaking and writing performances of test takers in nine languages. The data are multilingual, and parallel across languages in terms of language proficiency levels and functions of test tasks. The corpus design is looked at from the viewpoint purposes of the corpora, sampling, representativeness and usability of interface. Further, usability of the corpus is discussed in the light of practicality and future possibilities.

Keywords: corpus design, multilingual test data, corpus based research, L2-learner

1 KORPUS OPETUKSEN JA TUTKIMUKSEN TUKENA

Tietotekniset mahdollisuudet ovat moninkertaistaneet korpusten koostamis- ja erilaisten käyttöympäristöjen soveltamismahdollisuuksia, mikä on herättänyt keskustelua myös siitä, millainen kor-

pus on laadukas, edustava tai käyttökelpoinen. Vastaus riippuu siitä, mitä ymmärrämme korpuksella, mikä sen käyttötarkoitus on ja millainen sen aineisto on luonteeltaan. Korpusten suunnittelusta ja koostamisesta ei ole yhtä oikeaa näkemystä, vaan niiden käyttötarkoitukset, aineistot ja rakenteet voivat vaihdella monin tavoin eri tieteenalojen, kuten lingvistiikan, kielen oppimisen, käännöstieteen, kirjallisuuden tai poliittisen retoriikan tutkimuksen mukaan.

Korpuksen aineiston laajuudelle ei ole myöskään minimirajoituksia, korpukseksi kun voidaan periaatteessa nimittää mitä tahansa yhtä tekstiä isompaa tekstikokoelmaa. Toisaalta nykylingvistiikan mukaisesti korpukselta edellytetään, että sen aineiston otostamiseen, edustavuuteen, lopulliseen kokoon, käytettävyyteen ja suunnitteluun on tietoisesti kiinnitetty huomiota ja että niihin liittyvät valinnat ja päätökset ovat perusteltuja (esim. McEnery & Wilson 2000; Meyer 2002).

Koska korpuksia on erilaisia, tietynlaisen korpustyypin valintaa ohjaa tutkimuksen aihe ja tavoite. Korpus voi olla synkroninen tai diakroninen sen mukaan, millaiseen ajanjaksoon aineiston kokoaminen perustuu. Korpuksen aineisto voi olla puhuttua ja/tai kirjoitettua kieltä ja aineiston koko voi olla staattinen, esimerkiksi miljoonan sanaa käsittävä tekstikokoelma, tai kasvava niin, että korpukseen lisätään koko ajan lisää aineistoa. Korpuksen aineisto voi olla annotoitu tai annotoimaton, jolloin korpuksen tekstit ovat raakamuodossa ilman lingvististä informaatiota. (ks. esim. McEnery & Wilson 2000; Meyer 2002.) Lisäksi korpuksen aineisto voi koostua jostakin tietystä teksti- tai diskurssilajista, esimerkiksi kokouskeskusteluista tai aikakauslehtien teksteistä, tai useista eri tekstilajien yhdistelmistä (esim. Heikkinen, Hurme, Lounela & Virtanen 2005; Jovanovic, op den Akker & Nijholt 2006).

Viimeisen kymmenen vuoden aikana toisen kielen oppijakorpusten koostamiseen liittyvät projektit ovat lisääntyneet. Korpusten sisällöt vaihtelevat kirjoitetuista esseistä ja kirjallisista koesuorituksista haastatteluihin ja syntyperäisen ja ei-syntyperäisen välisiin keskusteluihin (ks. myös Tono 2003). Korpuksissa oppijoiden taustat voivat vaihdella useista lähtökielistä joihinkin tiet-

tyihin kieliryhmiin, kuten myös oppijoiden kohdekielen taitotaso alkeistasosta edistyneisiin. Oppijakorpusten materiaali on usein kerätty formaalisista oppimistilanteista. Kielitaitotestien aineistoista on myös koostettu korpuksia. Esimerkiksi Cambridgen kielitutkinnoista koostettu korpus sisältää sekä kirjallista että suullista materiaalia eritasoisista testeistä (ks. Boyle & Booth 2000; Ball 2001; Barker 2006). Koko ajan kasvava korpus on kooltaan valtava, sillä siinä on jo nyt mukana 85 000 opiskelijaa, jotka edustavat 100 lähtökieltä ja 180 maata.

Tarkastelen artikkelissani Yleisten kielitutkintojen testiaineistosta koostettavaa korpusta, joka sopii sekä tutkimus- että opetuskäyttöön ja joka mahdollistaa aineistonsa puolesta hyvinkin erilaisia lähestymistapoja. Yleisten kielitutkintojen korpuksen aineisto koostuu suorittajien tasoarvioista, taustatiedoista sekä puhumisen ja kirjoittamisen suorituksista. Esittelen ensin lyhyesti korpuksen aineiston osana Yleisten kielitutkintojen arviointijärjestelmää. Käyn sitten läpi korpuksen koostamisen vaiheita ja sen rakennetta sekä pohdin näihin suhteuttaen, miten edustava ja käyttökelpoinen Yleisten kielitutkintojen testiaineistokorpus on.

2 YLEISET KIELITUTKINNOT TESTIAINEISTONA

Yleiset kielitutkinnot (YKI) on aikuisille tarkoitettu kielitaidon näyttötutkinto, jonka voi suorittaa kuka tahansa aikuinen riippumatta siitä missä ja miten kielitaidon on hankkinut. Tutkinnossa on valittavana yhdeksän kieltä (englanti, espanja, italia, ranska, ruotsi, saame, saksa, suomi ja venäjä) kolmella eri tutkintotasolla (perus-, keski- ylin taso). Tutkinto pohjautuu toiminnalliseen kielitaitokäsitykseen ja taitotasoajatteluun, jonka mukaan kielitaidon edistyminen on jaettu kuuteen tasoon siten, että tasot 1–2 arvioivat perustason, tasot 3–4 keskitason ja tasot 5–6 ylimmän tason kielitaitoa. Jokaisen kielen ja tason tutkinnot sisältävät viisi osakoetta: tekstin ymmärtäminen, kirjoittaminen, puheen ymmärtäminen, puhuminen sekä rakenteet ja sanasto. Testin suoritettuaan osallistujat

saavat kielitaitotodistuksen, jossa on arvio jokaisesta osataidosta erikseen sekä niihin perustuva yleistasoarvio. (Yleisten kielitutkintojen perusteet 2002.)

Yleisten kielitutkintojen testi on nk. paperi ja kynä -testi eli testin suorittaja saa testimateriaalin eteensä osakoekohtaisina vihkoina, joihin hän kirjoittaa vastauksensa annetun ajan kuluessa. Luokkaosa (tekstin ymmärtäminen, kirjoittaminen sekä rakenteet ja sanasto) suoritetaan yleensä peräkkäin siten, että tekstin ymmärtämisen tekemistä suositellaan ennen kirjoittamista. Kirjoittamisen koe koostuu kaikissa kielissä ja kaikilla tutkintotasoilla kolmesta tehtävästä, joiden tekstilajit, aihepiirit, kielenkäyttötarkoitukset ja vaativuustaso eroavat toisistaan. Esimerkiksi keskitason tutkinnon kirjoittamisen kokeessa voi olla kielestä riippumatta seuraavanlaiset tehtävät: tuttavallinen kirje, muodollinen sähköpostiviesti ja mielipidekirjoitus.

Puhumisen koe suoritetaan kielistudiossa ja/tai haastattelijan kanssa siten, että perustason ja keskitason puhumisen koe suoritetaan kielistudiossa suomen kielen perustasoa lukuun ottamatta. Kaikkien kielten ylimmällä tasolla puhumisen kokeessa on sekä studio-osa että haastattelu. Perus- ja keskitason puhumisen suoritukset tallennetaan kasetille tai CD-levylle ja ylimmän tason haastattelut videoidaan. Puhumisen studiokoe koostuu 3−4 tehtävästä, jotka voivat olla esimerkiksi itsestä kertominen, tilanteissa reagointi, simuloitu keskustelu ja puheenvuoron esittäminen. Haastatteluissa keskustellaan haastattelijan kanssa ajankohtaisista ilmiöistä ja argumentoidaan omia näkemyksiä niihin liittyen. Aineistoina puhumisen kokeen suoritukset ovat siis vihkoissa oleviin tehtäviin perustuvia äänitallenteita.

Tutkinnon suorittajia pyydetään täyttämään testin suorittamisen yhteydessä taustatietolomake, jossa kysytään mm. sukupuoli, ikä, äidinkieli, koulutustausta, ammattiala, kohdekielen opiskelupaikkaa ja -aikaa, kohdekielen käytön taajuutta kotona, työssä ja vapaa-aikana. Lisäksi kysytään, mistä tutkinnon suorittaja on saanut tietoa tutkinnosta, mitä tarkoitusta varten hän on suorittamassa tutkintoa sekä millaisiin tarkoituksiin hän aikoo käyttää tutkinto-

todistusta. Taustatietolomakkeen kysymykset ovat suurimmaksi osaksi rastitettavia. Näin ollen taustatietolomake on optisesti luettava ja se edellyttää käsin koodaamista vain avokysymysten osalta.

Yleisten kielitutkintojen aineisto on siis monikielistä ja tuottamistaitoja koskevat testisuoritukset ovat näytteitä eritasoisten kielenkäyttäjien taidosta kirjoittaa ja puhua kohdekieltä. Tutkimusaineistona käytetään vain niiden testiin osallistuneiden suorituksia, jotka ovat myöntäneet luvan käyttää suorituksiaan nimettöminä Internetin välityksellä tutkimus- ja opetustarkoituksiin. Testin suorittajien henkilöllisyyttä suojaa se, että he saavat id-numeron jo testiin ilmoittautumisvaiheessa, kun heidän tietonsa syötetään suorittajatietokantaan. Saman id-numeron perusteella heidät identifioidaan arviointivaiheessa ja sen avulla eri aineistojen yhdistäminen toisiinsa, esimerkiksi taustatietojen todistusarvioihin, on mahdollista eri tarkoituksia varten.

Yleisten kielitutkintojen testiaineisto on kiinnostanut opinnäytetyön tekijöitä lähes tutkinnon alkuvaiheesta lähtien. Samoin näytteitä esimerkiksi eritasoisten kielen käyttäjien suorituksista on kysytty opetustarkoituksiin. Koska testiaineistosta on elektronisesti tallennettu todistusarviot tutkintokerta-, tutkintopaikka ja -aikakohtaisesti suorittajatietokantaan sekä taustatiedot SPSS-muotoisina tiedostoina, on niiden käsittely ja koostaminen tutkimustarkoituksia varten ollut helpompaa kuin laatikoihin arkistoitujen testivihkojen ja kasettien. Korpuksen koostamisen motiivina onkin yhtäältä aineiston säilyttämisen ja selaamisen helpottaminen mutta myös eri aineistojen yhdistäminen nykyistä käytännöllisemmällä tavalla.

Korpuksen koostamisen käynnistymistä edesauttoi Jyväskylän yliopiston saama Akatemia-rahoitus eri alojen, kuten historian ja psykologian, tutkimusinfran kehittämiseen. Rahoitus on tarkoitettu tutkimusaineistojen käsittelyyn ja hallintaan liittyvien ohjelmistojen ja laitteiden hankintaan sekä tietokantojen rakentamiseen. Vaikka rahoituksella otettiin vasta ensiaskel, se oli suuri askel käyntiin pääsemisen kannalta. Neuvottelut korpuksen si-

joittamisesta Yhteiskuntatieteelliseen tietoarkistoon, joka sisältää muun muassa aineistoja työelämää, hyvinvointia ja vaalikäyttäytymistä tutkiville, oli puolestaan osoitus Yleisten kielitutkintojen aineiston monitieteisestä luonteesta.

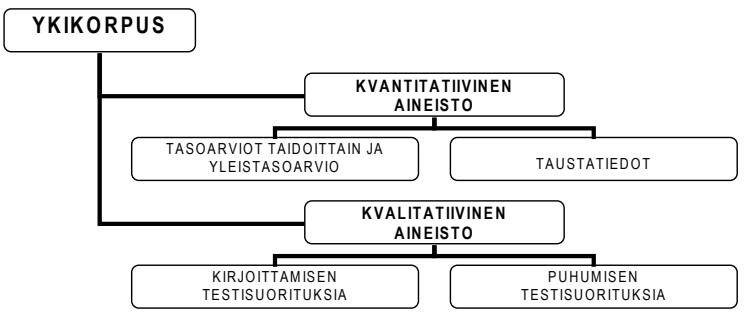
Korpuksen koostaminen on pitkä prosessi, jossa on monenlaisia vaiheita päätöksineen ja kokeiluineen (esim. Douglas 2003; Heikkinen ym. 2006). Osa päätöksistä on kompromisseja ihanteellisella tavalla käyttäjää palvelevan korpuksen ja käytännön sanelemien rajoitusten ja taloudellisten tai teknisten resurssien välillä. Yleisten kielitutkintojen korpuksen aineiston otostamista ohjasi pääasiassa käytettävissä oleva materiaali, jota kielitutkintojen järjestämisestä kerääntyy. Valintaperusteiden pohtiminen ja soveltaminen kytkeytyi seuraaviin aineistoon ja teknisiin ratkaisuihin liittyviin kysymyksiin: mitä kaikkea testimateriaalista olisi mahdollisista testisalaisuuden vuoksi tallentaa, minkä tallentaminen olisi käyttötarkoitukseen nähden tarkoituksenmukaista, millainen määrä tallennettua materiaalia olisi edustava otos toisaalta tutkinnon sisältöihin ja toisaalta keskimääräiseen suorittajajoukkoon nähden sekä millainen käyttöliittymän rakenne olisi toimiva hakujen kannalta. Tarkastelen seuraavassa näihin kysymyksiin liittyviä valintoja.

3 YKIKORPUKSEN AINEISTO JA RAKENNE

3.1 AINEISTON KOOSTAMINEN

Yleisten kielitutkintojen testiaineistokorpuksen koostaminen alkoi siihen tulevan aineiston valinnasta, jota ohjasi suurelta osin aineiston luonne, koska testitehtävien paljastaminen ei ollut mahdollista. Yleisten kielitutkintojen tehtävät eivät ole kertakäyttöisiä vaan ne tallennetaan analyysien jälkeen tehtäväpankkiin. Koska puheen ja tekstin ymmärtämisen taidon sekä rakenteiden ja sanaston hallinnan tarkasteleminen ilman tehtäviä on mahdotonta, niitä koskevien tehtävien tallentamisesta korpukseen luovuttiin. Edellä

mainituista syistä oli luontevaa, että korpus koostuisi tuottamistaitojen suorituksista ja numeerisesta tasoarvio- ja taustatietodatasta. Korpuksen aineisto koostuukin kvalitatiivisesta ja kvantitatiivisesta aineistoista kuvion 1 mukaisesti:



KUVIO 1. Yleisten kielitutkintojen korpuksen aineisto.

Korpukseen tulevasta aineistosta tasoarviot ovat jo valmiina korpukseen siirrettävässä muodossa. Taustatietolomakkeiden osalta korpukseen ei ole tarkoituksenmukaista siirtää kaikkia tietoja, koska osa siinä kysyttävistä tiedoista on tarkoitettu testin järjestäjien käyttöön. Taustatietolomakkeista tallennetaan korpusaineistoon sukupuoli, syntymäaika, äidinkieli, koulutus, sosioekonominen asema, kielen opiskeluaika ja -paika(t), testikielen käyttöä koskevat kysymykset sekä todistuksen käyttötarkoitus. Näiltä osin SPSS-muodossa oleva taustatietoaineisto täytyy muokata korpuksen valintakriteerien ja hakuehtojen mukaiseksi.

Laadullisen aineiston osalta korpuksen koostaminen aloitettiin päättämällä siitä, kenen suoritukset korpukseen siirrettäisiin. Tässä vaiheessa ne suorittajat, jotka olivat kieltäneet suorituksensa käyttämisen, jätettiin pois. Niissä kielissä, joissa suorittajamäärät ovat pieniä, päätettiin ottaa mukaan kaikki tutkimuslupaan myöntävästi vastanneet ja tutkintotason hyväksyttävästi suorittaneet. Englannissa ja suomessa, joissa testikertakohtaiset suorittajamäärät ovat suuria, valittiin yksi tutkintokerta, jonka suoritukset päätettiin siirtää ensimmäisenä korpukseen, jos niiden kirjoitta-

jat olivat vastanneet myöntävästi tutkimuslupaan ja jos he olivat suorittaneet hyväksyttävästi kirjoittamisen kokeen. Kaikilta korpukseen syötettäväksi valittavilta kirjoittajilta päätettiin tallentaa kaikki kolme kirjoitustehtävää. Tässä vaiheessa ei vielä mietitty sitä, miten monen suorittajan suoritukset kieli- ja tutkintotasokohtaisesti tallennettaisiin korpukseen.

Kirjoittamisen suoritukset kirjoitettiin aluksi XML-muotoon siten, että digitalisoinnissa pyrittiin säilyttämään testivihkoon käsinkirjoitetun tekstin muotoilu ja asettelu mahdollisimman alkuperäisenä esimerkiksi sisennyksineen, erilaisine merkkeineen, yli- ja alleviivauksineen. Eri kielten digitalisoijilta edellytettiin syötettävän testikielen hallintaa, jotta esimerkiksi romaanisten kielten aksentit tulivat mahdollisimman oikein merkittyä. Tekstin kohdille, joissa käsiala oli niin epäselvää, että sitä ei pystynyt lukemaan, sovittiin yhteinen merkitsemistapa. Kirjoittamisen aineiston digitalisoijat koulutettiin ennen työn alkamista ja heidät velvoitettiin validoimaan syöttämänsä tekstit jokaisen suorittajan jälkeen. Koska XML-koodin käyttöön liittyy virheriskejä, kirjoittamisen suoritukset syötetään tietokantaan nykyisin verkkopohjaisella lomakkeella, jossa käytetään tavallisia tekstieditorin komentoja ja jossa tekstin metatiedot valitaan pudotusvalikosta ja/tai annettua vaihtoehtoa hiirellä klikkaamalla.

Kirjoittamisen suoritusten digitalisoijien kanssa sovittiin, että he raportoivat ongelmatapauksista ennen kuin tekevät päätöksiä niiden suhteen. Näin ratkaisut voitiin koota yhteiseksi ohjeistoksi mahdollisimman aikaisessa vaiheessa ja parantaa digitalisoitujen tekstien yhdenmukaisuutta. Yleisimpiä tekstin digitalisoijien raportoimia ongelmia olivat:

- epäselvä käsiala
- kappalejaon puuttuminen
- hyvin samanlaiset isot ja pienet alkukirjaimet
- a:n ja o:n erottaminen toisistaan
- a:n ja e:n erottaminen toisistaan
- sekava lauserakenne
- pyyhitty teksti, joka näkyy yhä suorituksen alta
- himmeästi kirjoitetut välimerkit

- osoitetiedot, joita on sekä sivun vasemmassa että oikeassa reunassa (xml-koodin vuoksi on päätettävä, miten osoitetiedot tallentaa tietokantaan).

Koska tekstit on kirjoitettu käsin, niissä on eittämättä aina tulkinnanvaraisuutta, joka aiheuttaa jonkin verran epäyhdenmukaisuutta aineistoon. Toisaalta ongelma olisi sama, jos korpus koostuisi käsinkirjoitetuista versioista. Tosin tällöin kukin aineiston käyttäjä tekisi tulkintansa epäselvistä kohdista omaan näkemykseensä nojaten.

Kirjoittamisen suorituksista tallennetaan itse tekstien lisäksi seuraavat metatiedot: osallistujan ID, kieli, tutkintokausi, tutkintotaso, tekstilaji, tehtävätyyppi sekä tehtävän otsake, kuten taulukosta 1 käy ilmi. Testin tekstilajeja/-tyyppejä ovat esimerkiksi viesti, muodollinen tai epämuodollinen kirje, mielipidekirjoitus, muistio, kutsu ja hakemus. Tehtävätyyppi viittaa tässä yhteydessä siihen, onko tehtävä ohjattu vai ei. Ohjatuissa kirjoittamistehtävissä tekstin sisältöä ja pituutta ohjataan kysymyksillä, esimerkiksi seuraavasti: ”Kerro viestissä, mitä sinulle kuuluu, milloin lomasi alkaa, mitä aiot tehdä lomalla”. Tehtävän otsikko puolestaan viittaa testivihkossa olleeseen tehtävän otsikkoon. Taulukossa 1 on esimerkki englannin perustason testiin osallistuneen henkilön yhden tehtävän kirjoittamisen suorituksesta metatietoineen.

Puhumisen aineiston käsittely korpusta varten edellytti kasetilla olevien suoritusten digitalisoimista MP3-muotoisiksi sekä henkilötietojen poistamista suoritusten alusta, jossa testin suorittajat sanovat nimensä nauhoituksen ja arvioitavan henkilön vastaavuuden varmistamiseksi. Puhumisen suorituksista tallennetaan korpukseen suorituksen lisäksi seuraavat metatiedot: kieli, tutkintotaso, puhumisen koetyyppi sekä tehtävätyyppi. Koetyyppi viittaa siihen, onko kyseessä studiokoe vai haastattelu ja tehtävätyyppi puolestaan siihen, millaiseen tehtävään suoritus on vastaus. Mahdollisia studiopuhumisen tehtävätyyppejä ovat itsestä kertominen, simuloitu keskustelu, tilanteissa reagointi ja puheenvuoro. Perustason haastattelun tehtävätyyppejä voivat puolestaan olla esimerkiksi itsensä esittely ja kertominen ja ylimmän tason haastattelun mielipiteen perusteleminen.

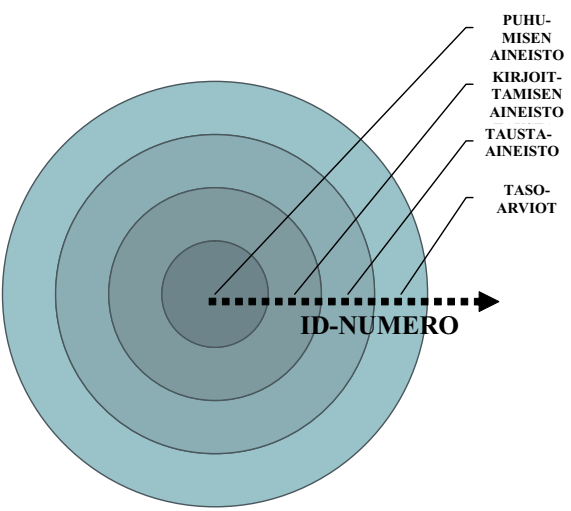
TAULUKKO 1. Esimerkki englannin perustason kirjoittamisen suorituksesta.

Suorituksen yhteiset tiedot	
Osallistujan ID	37593
Kieli	Englanti
Kausi	Kevät 2006
Testitaso	Perustaso
Tehtävä 1	
Tekstilaji	Epämuodollinen viesti
Tehtävätyyppi	Ohjattu kirjoittamistehtävä
Otsake	Kiitoskortti
Teksti	Dear Jill, Thank you for remembering my birthday. I like those shoes very much, how did you know my size? We have a great party at my grandmothers summercottage last weekend. All family and friends were there. I think we all will remember that day and I tell you why. My mother forget to make the cake...well, we have a lovely time without chocolate cake. I hope that you can come and see me and my family soon. with love, Maija

Koska Yleisten kielitutkintojen aineistoa tulee koko ajan lisää, korpuksen lopullista kokoa ei ole päätetty etukäteen. Tässä mielessä korpusta voisi kutsua dynaamiseksi ja kasvavaksi verrattuna synkroniseen korpukseen, jonka aineisto on eräänlainen leikkaus jostakin kielellisestä ilmiöstä, esimerkiksi tietystä tekstilajista tiettynä aika ja jonka koko on rajattu (Meyer 2002). Kvantitatiivisen aineiston osalta aineiston siirtäminen eri aikoina korpukseen ei ole yhtä haastavaa kuin laadullisen aineiston osalta. Laadullinen aineisto myös kyllääntyy todennäköisemmin ylittäessään jonkin numeerisen rajapyykin. Puhumisen aineiston osalta myös tiedostojen vaatima tila asettaa omat haasteensa. Tässä vaiheessa näyttää siltä,

että runsasaineistoisessa suomen ja englannin keskitason tutkinnon kirjoittamisen aineistossa tallennettavien suorittajien yläraja liikkuu reilun tuhannen suorittajan paikkeilla, mikä tarkoittaa yli 3000 tekstiä tutkintotasoa ja tutkintokieltä kohden. Kielissä ja/tai tutkintotasoilla, joissa suorittajamäärät ovat pieniä, kaikki tallennusluvan saaneet suoritukset tallennetaan tutkintokierroksen jälkeen.

Kvantitatiivista ja kvalitatiivista aineistoa yhdistää toisiinsa osallistujan id-numero, kuten kuviosta 2 näkyy. Kuvio 2 osoittaa myös aineistojen keskinäisen todennäköisyyden olla osa aineistoa ja samalla myös aineistojen keskinäistä kokoa. Kaikkein suurin aineisto suorittajamääriin nähden on tasoarvioaineisto. Seuraavaksi suurin on taustatietoaineisto. Taustatietoja ei ole kuitenkaan kaikilta niiltä testin suorittajilta, joiden tasoarviot ovat korpuksessa. Kvalitatiivisesta aineistosta kirjoittamisen aineisto on puhumista suurempi eli korpuksessa on useamman testin suorittaneen henkilön kirjoittamisen kuin puhumisen suoritukset.



KUVIO 2. Aineistojen keskinäinen koko.

Jos aineistoa lähestytään sen käytön tai hakujen kannalta ja haun lähtökohtana on kvalitatiivinen aineisto, joko kirjoittamisen tai puhumisen suoritukset, tarkasteltavalle henkilölle löytyy korpuksesta varmasti tasoarviot sekä hyvin todennäköisesti myös taustatiedot.

3.2 KÄYTTÖLIITTYMÄN TOTEUTUS

Yleisten kielitutkintojen korpus tulee sijoittumaan Tampereen yliopiston Yhteiskunnalliseen tietoarkistoon, joka myös koordinoi korpuksen käyttölupien hakemista ja myöntämistä. Korpus itsessään on verkkopohjainen ja aineistojen haku tapahtuu käyttöliittymän kautta. Käyttöliittymää voi käyttää suomen- tai englanninkielellä. Käyttöliittymän suunnittelun lähtökohtana on ollut, että aineistoja voi lähestyä mistä tahansa toisesta aineistosta käsin. Käyttöliittymän kautta aineistohakuja voi tehdä siten, että haetaan näytettävät tiedot pudotusvalikkojen kautta, määritellään vertailuehdot tai syötetään avokenttään haluttu tieto. Käytännössä tämä tarkoittaa esimerkiksi seuraavanalaisia sisääntuloja aineistoon:

1) Jos hakuintressinä on yli 40-vuotiaiden suomalaisten miesten englannin kielen taito, valitaan näytettävistä tiedoista 1) kieli eli englanti, 2) sukupuoli sekä 3) määritellään vertailuehdoista ikäkategoriaksi yli 40 vuotta. Näin saadaan kaikkien yli 40-vuotiaiden miesten englannin tutkinnon tasoarviot näkyviin.
2) Jos kiinnostuksen kohteena on suomi toisen kielenä -puhumisen taito erilaisine piirteineen henkilöillä, joiden taito on Yleisten kielitutkintojen taitotasolla 3, joka vastaa Yhteiseurooppalaisen viitekehyksen taitotasoa B1 ja on myös mm. perusopetuksen päättövaiheessa hyvää osaamista vastaava taitotaso, valitaan käyttöliittymän näytettävistä tiedoista 1) kieleksi suomi, 2) puhumisen arvioiden pudotusvalikosta 3 ja 3) kvalitatiivisen aineiston luokista puhumisen suoritukset. Jos puhujien taustatietoja ei tarvita vielä tässä vaiheessa, niihin voi palata myöhemmin syöttämällä avokenttään niiden henkilöiden ID-numerot, joiden taustatiedot kiinnostavat. Jos taas jo ensimmäisessä vaiheessa esimerkiksi äidinkieleltään venäjänkielisten puhumisen taito on kiinnostuksen kohteena, valitaan muiden hakukriteerien ohella äidinkielen ehdoksi venäjän kieli pudotusvalikosta. Tällöin

saadaan näkyviin niiden henkilöiden suomen kielen puhumisen suoritukset, joiden äidinkieli on venäjä ja joiden puhumisen tasoarvio on 3.

Käyttöliittymän ohjelmoinnissa ja käyttäjäystävällisyyden varmistamisessa haasteita ovat erilaiset datat ja niiden muutokset aikojen kuluessa. Esimerkiksi tutkinnon taitotasoasteikko on muuttunut 9-portaisesta 6-portaiseksi vuonna 2002, minkä vuoksi hakuehtona taitotaso 5 hakisi kaikki tason 5 suoritukset erittelemättä sitä, kummalla asteikolla suoritus on arvioitu. Sekaannusten välttämiseksi ja käyttäjän kannalta yksinkertaisemman ratkaisun hakemiseksi 9-portaisella asteikolla annetut tasoarviot päätettiin konvertoida 6-portaisiksi valmiiksi. Muutoksia on tapahtunut myös taustatietolomakkeissa moneen otteeseen tutkinnon kolmetoistavuotisen historian aikana. Käyttöliittymän hakukriteerien ehdoiksi päätettiin valita viimeisimmän taustatietolomakkeen kysymysmuotoilut muuttujineen. Käyttöliittymä tarjoaa kuitenkin mahdollisuuden tutustua taustatietolomakkeiden eri versioihin ja se tekee eron puuttuvien tietojen osalta siinä, onko kyseessä vastaamatta jättämisestä vai eriversioisesta taustatietolomakkeesta.

Käyttöliittymän haasteita ovat olleet myös monimuuttujainen taustatietoaineisto, joka koostuu sekä numeerisista muuttujista että avovastauksista. Korpusta varten osa avovastauksista on koodattu ja niistä on muodostettu uusia muuttujia. Äidinkieltä koskevissa avovastauksissa on puolestaan useita erilaisia kirjoitusmuotoja samalle kielelle tai samaan kieleen viitataan erilaisin termein, mikä vammauttaa avokenttähakua. Sen toimimisen ehtona on, että äidinkielien kirjoitusasut yhdenmukaistetaan ja/tai aineisto koodataan uudelleen. Tämäntyyppiset aineiston läpikäynnit ovat aikaa vieviä ja tulevat esille vasta sitä mukaa, kun eri hakutyyppien vaihtoehtoja käydään läpi konkreettisten hakuesimerkkien kautta. Tässä mielessä käyttöliittymä on aina vähän keskeneräinen ja käyttäjän kannalta kompromissi käytettävissä olevien resurssien ja ideaalisti toimivien hakuehtojen välillä.

4 KORPUKSEN MITTAILUA

Yleisten kielitutkintojen testiaineistosta koostettu korpus on kasvava korpus. Eri kielten kvalitatiiviset aineistot ovat keskenään erisuuruisia ja joissakin kielissä, kuten ranskassa ja italiassa, korpus ei saavuta vuosiin englannin ja suomen kielen aineistojen määriä. Korpuksen monikielisyydestä pidetään kuitenkin kiinni, koska se on eittämättä korpuksen vahvuus etenkin kieltenvälisen vertailtavuuden kannalta. Korpuksen vahvuuksia on myös se, että monikielisyydessään ja suoritusten taitotasojen vertailtavuuden perusteella se on tulkittavissa ja hyödynnettävissä myös monissa kansainvälisissä yhteyksissä. Yleisten kielitutkintojen taitotasoasteikko on kalibroitu Yhteiseurooppalaisen viitekehyksen asteikon kanssa, joten aineiston tasoarviot ja niihin perustuvat kirjoittamisen ja puhumisen suoritukset ovat rinnastettavissa soveltuvin osin viitekehyksen asteikkoon (Kaftandjieva & Takala 2003; Takala & Kaftandjieva 2004).

Vaikka aineisto on rinnastettavissa Yhteiseurooppalaiseen viitekehykseen ja testin suorittajat ovat oikeita kielenkäyttäjiä eritasoisine taitoineen, aineiston autenttisuutta voi syystäkin kritisoida. Testitilanne on oikean elämän tilanne mutta ei tietystikään saavuta todellisen elämän sävyjä ja vivahteita eikä esimerkiksi erilaisten puhekumppaneiden tuomaa panosta vuorovaikutustilanteisiin. Emme puhu todellisuudessa armottoman nauhan kanssa, joka ei toista pyydettäessä, tai emme valitse oikeaa vaihtoehtoa A, B, C lukiessamme aamun lehteä, vaikka jokapäiväisessä elämässä samantyyppisiä funktioita voisi liittyäkin puhekumppanin tai aamun lehden ymmärtämiseen. Tilanteiden keinotekoisuuden lisäksi testitilanteissa voi suoritukseen vaikuttaa todellista elämää enemmän jännitys, aikapaine ja toisten kokelaiden suoritukset (Tarnanen & Mäntylä 2006). Toisaalta korpuksen aineisto ei ole sen keinotekoisempaa kuin mikä tahansa oppija-aineisto, joka on kerätty formaalisissa yhteyksissä.

Aineiston edustavuutta voidaan tarkastella monesta eri näkökulmasta sen mukaan, viittaako edustavuus kielenoppijoihin vai

valitun testiaineiston laatuun ja sen määrään. Testin suorittajat edustavat monipuolisesti eri-ikäisiä ja eri kielten oppijoita, joiden sosioekonomiset ja kielenoppimistaustat vaihtelevat suorittajasta toiseen (ks. Härkönen, Kärkkäinen, Immonen, Kärkkäinen & Takala 2000). Tässä mielessä he ovat varioivampi joukko kuin jonkin yhden kurssin suorittajat tai tietyn työpaikan edustajat. Suorittajamäärien osalta aineisto ei ole edustava suhteessa koko aikuisväestöön muissa kielissä kuin englannissa ja suomessa ja niissäkin vain keskitason osalta. Testiaineiston sisällöllisen edustavuuden kannalta korpuksen laadullinen aineisto edustaa melko hyvin yleensä ottaen kaikkien kielten ja kielikohtaisten testipatteristojen sisältöjä, koska testien rakenne, tehtävien funktiot ja vaikeustasot ovat melko tarkasti määriteltyjä. Yksittäisten tehtävien aihepiirien ja kontekstien osalta korpuksen aineisto ei tietenkään tee oikeutta koko testivariaatiolle sisällöllisesti ja määrällisesti. Millaisten tutkimuskysymysten osalta tämä ero on sitten olennainen, on toinen kysymys.

Korpuksen käytettävyyttä on vielä ennenaikaista arvioida, koska se on vielä testivaiheessa. Suunnitteluvaiheessa asetettu tavoite mahdollistaa tietojen haku aineistojen poikki on kuitenkin toteutunut. Toisaalta korpuksen aineisto jättää paljon tehtävää tutkijalle itselleen, jos hän on kiinnostunut lingvistisistä piirteistä, koska korpus on annotoimaton. Korpuksen kvalitatiivisen aineiston annotointi olisikin hedelmällinen lisä korpukseen ja monipuolistaisi sen käyttömahdollisuuksia tuntuvasti. Kehittämishaasteita liittyy myös aineistojen metatietojen monipuolistamiseen: mitä enemmän korpuksen käyttäjällä on tietoa aineiston taustoista, sitä todennäköisemmin hän pystyy tekemään osuvia tulkintoja aineistoista erilaisia tarkoituksia varten.

Yleisten kielitutkintojen korpuksen hyödyntäminen tutkimustarkoituksia varten on jo alkanut. Suomen Akatemian rahoittama Jyväskylän yliopiston tutkimushanke CEFLING hyödyntää tutkimusaineistona ykikorpuksen englannin ja suomen kielen kirjoittamisen suorituksia. Hankkeessa tarkastellaan sitä, millaiset kielelliset piirteet erottavat kielitaidon eri tasoja toisistaan (ks. lisää

http://www.jyu.fi/hum/laitokset/solki/en/research/projects/cefling). Opinnäytetutkimusten lisäksi korpusta voidaan hyödyntää myös Yleisten kielitutkintojen kehittämis- ja tutkimustyössä, kun haetaan aineistoa esimerkiksi sitä varten, miten osuvasti ja monipuolisesti puhumisen ja kirjoittamisen tehtävien avulla voidaan arvioida eri tasojen kielitaitoa tai miten kieltenvälinen vertailtavuus ilmenee tuotoksissa. Korpuksen olemassa olo mahdollistaa testiaineiston hyödyntämisen helpommin lähestyttävällä tavalla ja toimii välittäjänä testiaineiston ja erilaisten tutkimusintressien välillä. Testiaineisto ei näin jää kertakäyttöiseksi, vaan se jatkaa elämäänsä korpuksen kautta.

KIRJALLISUUS

Ball, F. 2001. Using corpora in language testing. *Research Notes,* 6, 6−8.

Barker, F. 2006. Corpora and language assessment: trends and prospects. *Research Notes,* 26, 2−4.

Boyle, A & D. Booth 2000. The UCLES/CUP Learner Corpus. *Research Notes,* 1, 10.

Douglas, F. M. 2003. The Scottish corpus of texts and speech. Problems of corpus design. *Literacy and Linguistic Computing***,** 18 (1), 23−37.

Heikkinen, V. P., T. Hurme, M. Lounela & M. T. Virtanen 2006. Teksti, aihe ja laji. Diakronisen korpuksen koostaminen ja käyttäminen. Teoksessa A. Pajunen & H. Tommola (toim.) *XXXII Kielitieteen päivät Tampereella 19.−20.5.2005* Tampere studies in language. Translation and culture. Series B 2. Tampere: Tampere University Press, 218−238.

Härkönen, R., A. Kärkkäinen, H. Immonen, K. Kärkkäinen & S. Takala 2000. *Yleisten kielitutkintojen satoa - tietoa ja tilastoja suorituksista ja suorittajista 1994−2000*. Helsinki: Opetushallitus.

Jovanovic, N, R. Akker op den & A. Nijholt 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation,* 40, 5−23.

Kaftandjieva, F. & S. Takala 2003. Development and validation of scales of language proficiency. Teoksessa W. Vagle (toim.) V*urdering av språkferdighet*. Trondheim: Institutt for språk- og kommunikasjonsstudier, 31–38.

McEnery, T. & A. Wilson 2000. Corpus linguistics. ICT4LT Module 3.4. [online]. [luettu 19.1.2007]. Saatavissa: http://www.ict4lt.org/fi/index.htm.

Meyer, C. F. 2002. *English corpus linguistics. An introduction*. Cambridge: Cambridge University Press.
Takala, S. & F. Kaftandjieva 2004. Using the Common European Framework: Some Finnish experiences. Teoksessa K. Mäkinen, P. Kaikkonen & V. Kohonen (toim.) *Future perspectives in language education*. Oulun yliopiston kasvatustietieden tiedekunnan tutkimuksia 101/2004. Oulu: University of Oulu, 45−53.
Tarnanen, M. & K. Mäntylä 2006. Toisen ja vieraan kielenoppijat Yleisissä kielitutkinnoissa. Teoksessa P. Pietilä, P. Lintunen & H-M. Järvinen (toim.) *Kielenoppija tänään. Language learners of today.* AFinLAn vuosikirja n:o 64. Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA, 105–123.
Tono, Y. 2003. Learner corpora: design, development and applications. [online]. [luettu 3.5.2007]. Saatavissa: http://ucrel.lancs.ac.uk/publications/CL2003/papers/tono.pdf.
Yleisten kielitutkintojen perusteet 2002. Määräys 55/011/2001. Helsinki: Opetushallitus.