FINLANCE

The Finnish Journal of Language Learning and Language Teaching

Volume VIII 1990

EDITED BY ARI HUHTA

Language Centre for Finnish Universities
University of Jyväskylä · Finland

FINLANCE

The Finnish Journal of
Language Learning and Language Teaching

FINLANCE Vol. VIII 1990

Volume VIII 1990
EDITED BY ARI HUHTA

Language Centre for Finnish Universities
University of Jyväskylä • Finland

ISSN 0359-0933

Kopi-Jyvä Oy
1991

PREFACE

The main theme of the present volume is language testing. The articles were written in 1989 or in 1990 and cover a wide range of testing systems and methods, mostly dealing with the language proficiency of university students and its measurement.

Anna Mauranen discusses language testing carried out at the language centres of Finnish universities. She deals with the central problems facing LSP language testers and suggests some solutions and approaches which could make testing more meaningful for all concerned. Ari Huhta's report on an ongoing study also deals with language centre testing but concentrates on a narrower problem. He has studied new testing techniques used to assess reading comprehension in English. The study also explored students' attitutes towards the tests.

Margaretha Corell and Gunilla Gentzel from the Institute for English Speaking Students at the University of Stockholm report on a fairly large-scale testing enterprise, the Rikstest (the National Test), a compulsory Swedish language test for foreign students wishing to attend a Swedish university. The test assesses the students' abilities in all the traditional macroskills - reading, writing, listening, and speaking - and has been designed to be easily administered and at the time as objective as possible.

Lynne Dotzenroth and Anu Virkkunen have explored the nature of language proficiency and FL learning by comparing the oral and written recall of negatively prefixed words by first and second language learners. Their approach is linked to testing by the use of a particular, though relatively rare, type of test to investigate a linguistic phenomenon. Using tests as a research tool has a long tradition. Recently, however, language testers such as Lyle Bachman have increasingly urged that information obtained through these studies be used to validate tests, i.e., to ascertain that they measure what they are supposed to measure. This approach would demand closer co-operation between language testers and researchers in L1 and L2 learning, which would definitely be beneficial for both groups.

Jyväskylä, December 1990

Ari Huhta

SISÄLTÖ

Anna Mauranen

KIELITAIDON MITTAAMISESTA KIELIKESKUKSISSA

1. YLEISTÄ

Kielikeskustestauksen tavoitteenmäärittelyllä on kaksi lähtökohtaa: korkeakoulusivistyksen saaneilta kansalaisilta odotettu kielitaidon laatu ja taso sekä korkeakouluopintojen menestyksekkään suorittamisen edellyttämä vieraan kielen taito. Nämä lähtökohdat painottuvat eri tavoin eri tutkinnoissa ja kielissä.

Koska tavoitelähtökohtia on kaksi, on myös testien pyrittävä ennustamaan kielikäyttäytymistä kahdenlaisissa tilanteissa: opiskelussa ja työelämässä. Näiden ennustamisessa ja ennusteen seuraamisessa on selvä ero - opiskeluun tarvittava kielitaito on suhteellisen suppea-alaista, kohtalaisen helposti spesifioitavissa ja sen riittävyyttä voidaan ainakin periaatteessa seurata. Työelämän vaatima kielitaito on kaikissa näissä suhteissa erilaista, koska työelämän vaatimukset kielitaidolle vaihtelevat paljon ja ovat toisinaan hyvin vaikeasti ennakoitavissa.

Koska kielitaidon mittaamisen tavoitteet kielikeskuksissa määräytyvät tulevien kielenkäyttötarpeiden kannalta pragmaattisesti, on tarkoituksenmukaista, että myös mittarien sisältö ja arvioinnin kriteerit perustuvat ensisijaisesti juuri kielen käyttötarpeisiin. Näin pyritään lähestymään tavoitteena olevaa käyttäytymistä mahdollisimman suoraan ja kattamaan kaikki normaalin kielenkäyttötilanteen vaatimat osataidot sekä niiden luonnollinen yhdistyminen kommunikaatiossa. Myös arvosteluperusteissa tavoitellaan luonnollisen kommunikaation näkökulmaa, joka painottaa kokonaisvaltaista kommunikaatiotilanteesta selviytymistä pikemminkin kuin kielitaidon yksityiskohtia. Tällainen kielitaidon mittaus edellyttää kielen käytön tarkastelua holistisesti. Kielitaitomittauksen keskeisiksi piirteiksi muodostuvat siten sellaisten kielenkäyttötilanteiden simulointi, joita pidetään kielitaitotavoitteiden kannalta tärkeinä ja vieraskielisen kommunikaatiotaidon kokonaisvaltainen arviointi.

Mittausmenetelmissä pyritään monipuolisuuteen ja joustavuuteen. Arviointitilanteet ja -olosuhteet vaihtelevat paljon, ja joka tarkoitukseen on voitava käyttää siihen

parhaiten soveltuvaa mittausmuotoa. Tämä asettaa arviointikriteereille vaatimuksen, että niiden on pysyttävä yhteismitallisina testausmuodon vaihdellessa. Mittauksen yleisperiaatteiden on oltava samat kaikissa kielikeskuksissa, jotta yhteismitallisia arviointikriteerejä voidaan kehittää eri tilanteisiin.

Mittaamismenetelmien monipuolisuutta edellyttää myös takaistusvaikutus opetukseen. Jos käytetään vain muutamaa, samana pysyvää testimuotoa, opetus saattaa kangistua näiden harjoittelemiseksi. Testaus voi antaa tukensa opetukselle, joka pyrkii autenttiseen kielenkäyttöön kehittämällä testimuotoja, jotka muistuttavat mahdollisimman paljon todellisia kielenkäyttötilanteita.

Takaistusvaikutuksen vuoksi testauksen täytyy myös perustua ajanmukaisiin teoreettisiin ja menetelmällisiin suuntauksiin kieltenopetuksessa. Kielikeskuksissa testauksen innovatiivisen roolin kehittäminen on erityisen tärkeää, koska yliopistoissa on suuri opetuksen vapaus, mutta kielikeskusten tehtävä vaatii vertailukelpoista testausta kaikkialla. Testauksella on tällaisessa tilanteessa mahdollisuus toimia kielenopetuksen uusien ajatusten ja menetelmien tunnetuksitekemisen kanavana.

Kielikeskusten käyttämien kielitaidon mittareiden on annettava luotettavaa tietoa opiskelijoiden kielitaidosta useille käyttäjäryhmille: työnantajille, yliopistojen ainelaitoksille, yliopistojen hallintoelimille, kielikurssien opettajille ja opiskelijoille itselleen. Jotta mittauksen tulosten mielekäs tulkinta olisi mahdollisimman helppoa myös kielikeskusten ulkopuolisille, on kehitettävä menetelmiä niiden kuvaamiseksi kaikille ymmärrettävällä ja käyttäjien kannalta relevantilla tavalla. Myös yliopistojen sisällä käytävässä keskustelussa kieltenopetuksen asemasta ja kehittämisestä on hyödyksi, jos kieltenopettajat pystyvät osoittamaan selkeitä ja yhteisesti hyväksyttäviä kriteereitä opiskelijoiden kielitaidolle sekä testituloksia, joista käy ilmi, miten hyvin opiskelijat näiden kriteerien valossa suoriutuvat. Seurannan ja kyselyjen avulla on hyvä varmistaa, että käyttäjäryhmät ovat tyytyväisiä testaukseen. Avainasemassa tässä ovat opiskelijat, joiden elämään testien tulokset eniten vaikuttavat.

Kielikeskusten tehtäviin ei asetuksessa eikä tutkintosäännöissä ole sisällytetty kirjoittamistaitojen opetusta tai testausta muissa kuin ruotsin kielessä. Vieraalla kielellä kirjoittamisen opettamista puoltaa kuitenkin kaksi seikkaa: taidon tarpeellisuus työelämässä (vrt. esim. Mehtäläinen, 1987; Bullivant et al., 1987) ja sen muita kielitaidon osa-alueita vahvistava vaikutus. Koulussa opetettava vieraskielinen kirjallinen ilmaisu on lähes yksinomaan ainekirjoitusta, mikä ei vastaa työelämässä tai tieteellisessä viestinnässä tarvittavia taitoja. Tällaisten taitojen sisällyttäminen

kielikeskusopetukseen olisi perusteltua ja toteutuisi laajimmin mikäli kirjoitustaitoja myös testattaisiin.

2. TAVOITTEIDEN KAKSINAISUUS

Useimmat suomalaiset korkeakouluopiskelijat tarvitsevat vieraan kielen taitoa sekä opiskeluaikanaan että valmistumisen jälkeen. Eri tutkinnoissa nämä tarpeet kuitenkin painottuvat erilailla sen mukaan, millainen ja miten hyvin ennakoitavissa on tuleva ammattikuva ja miten paljon toisaalta opintoihin sisältyy vieraskielistä kirjallisuutta. Opiskeluajan ja työelämän vaatimukset painottuvat eri tavoin myös eri kielien opetuksessa, sillä tutkintovaatimuksissa on yleensä teoksia vain muutamalla kielellä, jotka eivät välttämättä käy edes yksiin tulevan ammatin tarpeiden kanssa. Eroa on vielä siinäkin, miten ns. kielitaidon osa-alueet ovat tarpeen: opiskelussa useimmiten tarvitaan lukutaitoja, valmistumisen jälkeen niiden lisäksi puhumis-, kuuntelu- ja kirjoitustaitoja.

Tästä seuraa kielitaidon mittaukselle ristiriitaisia vaatimuksia, sillä työelämässä vaadittava monipuolinen ja joiltakin osin vaikeasti ennustettava kielitaito edellyttää mittaukselta laaja-alaisempaa ja monipuolisempaa otetta kuin pelkästään opintoihin liittyvän kielitaidon testaaminen. On ajateltavissa, että tällaiseen laajaan, kaikki kielitaidon osa-alueet ja kommunikaatiotaidot kattavaan, integratiiviseen testiin voitaisiin periaatteessa sisällyttää myös opintojen vaatima suppeampi kielitaito, ja näin vähennettäisiin yksi mittaamiskerta ylioppilaskokeen ja ammattisuuntautuneen kielikokeen välistä.

On kuitenkin käytännössä välttämätöntä sijoittaa kielikoe opintojen alkuun, jotta voidaan ohjata opetuksen pariin ne opiskelijat, joiden kielitaidon puutteet vaikeuttaisivat opintoja. Alkuvaiheessa ei toisaalta voida järjestää kaikenkattavaa, ammatissa tarvittavaan kielitaitoon perustuvaa koetta, koska tällainen koe, mikäli se pyrkisi olemaan ollenkaan realistinen, edellyttäisi kielitaitojen soveltamista sellaisiin tehtäviin, joissa testattava ottaa jonkinlaisen asiantuntijaroolin omaan alaansa nähden. Ensimmäisen tai toisen vuoden opiskelijoilla ei ole vielä tiedollisia valmiuksia tai tarvittavaa kypsyyttä selviytyä tällaisesta kokeesta. Kokeesta tulisi liian vaativa, tai jos se sovitettaisiin ensimmäisen opintovuoden tasolle, kyseenalainen ennustavuudeltaan, koska testattavissa on odotettavissa suuria muutoksia ennen opinnoista työelämään siirtymistä.

Kielikeskusten testaustavoitteiden kaksinaisuus mutkistaa myös ratkaisuja, jotka koskevat testien eriytymisastetta opinalojen mukaisesti. Eri kielten ja eri taitoalueiden tarpeet täytyy ottaa huomioon. Kaikkia yhteisesti koskevat periaatteelliset päätökset keskittyvät kahteen kysymykseen: mikä on yleissivistyksellisen ja tiukan ammatillisen kielitaidon suhde, sekä miten saavuttaa jokaisen opiskelijan kannalta optimaalinen eriytymisaste, jotta kukin voi tuoda taitonsa parhaiten esiin.

Edellinen kysymys liittyy Widdowsonin (1983) pohdiskelemaan dikotomiaan, joka erityisalojen kielenopetuksessa aina kohdataan: tehokkaalta tuntuva, tarkoin erityisalaan kohdistuva kielenopetus on toisaalta rajoittavaa ja sivistysihanteen vastaista. Toisaalta suomalaisten opiskelijoiden yliopistoa edeltävä koulutus on vahvasti yleissivistykseen suuntautunutta myös vieraissa kielissä. Yliopistossa alkavan kielen opiskelussa yleiskielitaito painottuu toki vahvemmin kuin niissä kielissä, joissa koulupohja on olemassa.

Toinen kysymys, millaisissa testeissä opiskelija voi parhaiten osoittaa taitonsa, on paljolti empiirisen selvittämisen asia. Ymmärtämistesteissä vaikuttaa siltä, että aihepiirin tuttuus ja valmiiden sisältöskeemojen olemassaolo helpottaa kuullun tai luetun diskurssin ymmärtämistä; samoin oman alan tekstien tyypillisten muotoskeemojen eli tekstin tyypillisen organisaation tuntemus (Carrell, 1984). Tämä herättää kysymyksen testien eriyttämistarpeesta opinalan mukaan. Kuinka pitkälle menevä eriyttäminen on mielekästä? Muualla on katsottu, että parasta olisi eriyttää akateemisten alojen kielitaitomittausta niin pitkälle kuin resurssit sallivat (Bingham Wesche, 1987). Vähimmäisvaatimuksena on humanistis-yhteiskunnallisten alojen mittaaminen erikseen teknis-luonnontieteellisistä aloista (vrt. Alderson ja Urquhart, 1983, 1985).

Myös opiskelijoiden näkemys siitä, miten pitkälle menevän eriyttämisen he katsovat olevan edullista, on tärkeää ottaa huomioon testejä kehitettäessä.

3. KIELEN KÄYTTÖTAITOON PERUSTUVA TESTAUS

Koska kielikeskustestien tavoitteena on ennustaa opiskelijoiden kykyä käyttää kieltä opintojensa ja ammattinsa vaatimissa yhteyksissä, on mittauksen keskeisenä periaatteena arvioida kommunikaation onnistumista tilanteissa ja tehtävissä, jotka edellyttävät kielitaidon ja -tiedon soveltamista opiskelun ja työelämän vaatimuksiin.

Kielitaidon testaus on perinteisesti pyrkinyt eristämään "puhtaan" kielitaidon mittauksen kohteeksi. Tällaiset testit ennustavat kuitenkin usein epätyydyttävästi testattavien kykyä selviytyä testin ulkopuolisista kommunikaatiovaatimuksista. Tällöin ilmeisesti mittaus on kohdistunut sellaisiin kielitaidon alueisiin, jotka ovat käytännön kommunikointitaidon kannalta toisarvoisia tai jotka eivät muodosta riittäviä edellytyksiä onnistuneelle kommunikoinnille.

Yksi ongelmien aiheuttaja on epäilemättä ollut muodollisen oikeakielisyyden korostunut asema. Muotovirheiden yksityiskohtaiseen tarkkailuun joudutaan helposti juuri silloin, kun kielikoodin hallinta eristetään sen käytöstä tilanneyhteydessä. Muotovirheiden summan voi tuskin kuitenkaan odottaakaan antavan kovin osuvaa arviota henkilön kommunikaatiotaidoista.

On myös ajateltavissa, että hyvä käytännön kommunikaatiokyky voi rakentua eri komponenteista eri yksilöillä ja jotkut komponentit voivat kompensoida toisia, ts. taitoprofiili voi vaihdella, vaikka kokonaistulos olisi hyvää kommunikaatiota. Joku pystyy käyttämään tehokkaasti hyväkseen vähäistäkin kielitaitoa siinä missä toinen vaatii itseltään laajoja tietoja ja varmaa taitoa ennen kuin uskaltautuu soveltamaan niitä käytäntöön.

Yleisesti voidaan todeta, että kommunikaatiotaito edellyttää enemmän kuin pelkkiä lingvistisiä taitoja, mutta joissakin rajoitetuissa olosuhteissa, kuten oman erikoisalan tekstien ymmärtämisessä on päinvastoin niin, että hyvät ei-kielelliset tiedot ja taidot saattavat korvata kielitaidon puutteita. Kummassakin tapauksessa ei puhtaasti kielitaitoon kohdistuva mittaus anna riittävää käsitystä kommunikaation onnistumisen edellytyksistä.

Testejä, jotka pyrkivät ennustamaan tulevaa tavoitteenmukaista kielikäyttäytymistä arvioimalla testattavan suoriutumista mahdollisimman autenttisissa tilanteissa, nimitetään usein performanssitesteiksi. Ajatus voidaan myös ilmaista sanomalla, että mittauksella pyritään selvittämään sitä, mitä testattava osaa kielellä tehdä, eikä sitä, mitä hän kielestä tietää. Performanssitestauksen käsite on alunperin lainattu ammatillisesta testauksesta (Jones 1985), missä se on paljon vanhempi kuin kielitestauksessa. Kielitaitoon sovellettuna performanssitestaus on varsin uutta, vaikka tätä termiä on käytetty vaihtelevissa merkityksissä jo notionaalis-funktionaalisen ja kommunikatiivisen opetussuuntauksen yhteydessä 1970-luvulta alkaen, ja sen sisältöä on tulkittu eri tavoin.

Yhteisiä piirteitä performanssitestauksen tärkeimmistä ominaisuuksista ovat ennen kaikkea realistisuuden tai autenttisuuden kaikinpuolinen tavoittelu, puhtaan lingivistisen taitoalueen rajojen ylittäminen, testitehtävien interaktiivisuuden korostaminen, testauksen kriteeriviitteisyys ja arvioijien järjestelmällinen opastus. Kielenopetuksen ja -testauksen suuntauksista nojaudutaan paljolti kommunikatiiviseen ajatusperinteeseen. Kommunikatiivisen ja performanssitestauksen tavoitteissa on paljon yhteistä, mutta realistisuuden tai autenttisuuden vaatimus esitetään paljon radikaalimmin performanssitestauksen yhteydessä.

Tulkintaerot keskittyvät pääasiassa performanssin käsitteeseen. Lingvistiikassa performanssi nähdään yleensä Chomskyn (1965) kompetenssi - performanssi - käsiteparin valossa, jolloin performanssilla tarkoitetaan ideaalisen puhuja - kuulijan kielellisen kompetenssin eli sisäisen kieliopin ilmenemistä ulkoisesti, kielen tuottamisena ja vastaanottamisena. Kompetenssi on ideaalinen, mentaalinen kielitieto ja performanssi sen reaalistumista, soveltamista käytäntöön. Tämän näkemyksen mukaan kompetenssia ei voi lainkaan suoraan mitata, koska se ilmenee vain performanssin välityksellä.

Chomskyn kompetenssikäsitys rajoittuu lingvistiseen kompetenssiin erotukseksi kommunikatiivisesta kompetenssista, jolla käsitteellä Hymes (1970) laajentaa kompetenssin koskemaan myös kykyä käyttää kieltä vaihtelevien kielenkäyttötilanteiden edellyttämällä tavalla. Myöhemmin kommunikatiivisen kompetenssin käsitettä on vielä laajennettu ja täsmennetty (ennen kaikkea Canale ja Swain 1980). Testauksen keskeisenä tavoitteena on yleensä ollut ilmikäyttäytymisen taustalla olevan kompetenssin arviointi, oli sitten kysymys lingvistisestä tai kommunikatiivisesta kompetenssista. Kompetenssin oletetaan olevan niin stabiili, että sen avulla voidaan ennustaa suoriutumista uusista kielenkäyttötilanteista myös tulevaisuudessa. Käytännössä ennusteissa on ollut vaikeuksia lyhyelläkin aikavälillä kuten yllä todettiin

Useimmat performanssitestauksen kehittäjät säilyttävät performanssi/kompetenssidikotomian (esim. Emmett, 1985, Bailey, 1985, Weir, ms.), mutta ovat ottaneet lähtökohdakseen kommunikatiivisen kompetenssin ja performanssin eivätkä lingvististä. Ajatuksena tällöin on, että realistinen performanssitestaus antaa luotettavampaa tietoa testattavan kommunikatiivisesta kompetenssista kuin analyyttinen lingvistinen testaus. Mitä enemmän testi muistuttaa autenttista tilannetta, sitä paremmin se kattaa tällaisessa tilanteessa tarvittavat taidot.

Jotkut taas (esim. Jones, 1985) jättävät huomiotta koko kompetenssin käsitteen (tai suhtautuvat siihen kielteisesti kuten Economou ms.) ja pyrkivät hyvän performanssiotoksen avulla ennustamaan performanssia laajemmin, postuloimatta väliin kompetenssia. Ajatus on siis varsin samantapainen kuin esimerkiksi ajokokeen tai koeluennon yhteydessä: testattava asetetaan tilanteeseen, jossa hänen on toimittava siten kuin todellista työtehtävää suorittaessaan. Tätä toimintaa sitten havainnoidaan ja arvostellaan. Kytkentä ammatilliseen testaukseen on ilmeinen.

Tämäntapaista testausta nimitetään joskus myös "suoraksi". Tässä testauksen problematiikka keskittyy voimakkaasti mittaustilanteiden ja -tehtävien realistisuuteen ja edustavuuteen sekä sosiaalisessa (testin "ekologinen validiteetti") että psykolingvistisessä mielessä. Testitehtäviä ei voi motivoida pelkästään sillä, että ne mittaavat hyvin jotain kielikyvyn komponenttia; niiden on edellytettävä testattavalta kaikkea sitä, mitä todellinenkin kommunikaatiotilanne edellyttäisi.

Käytännössä performanssitestaukseen pyrkivät testaajat tavoittelevat samoja asioita, vaikka lähtökohdat poikkeaisivatkin hiukan toisistaan, eivätkä he juuri ole pyrkineet teoreettisten sävyjen erittelyyn. Koko suuntaus on peräisin käytännön uudistamispyrkimyksistä, jotka eivät kaikilta osin ole vielä löytäneet teoreettista ilmaisua. Niinpä ratkaistavia ongelmiakin on monta.

Ensimmäinen performanssitestaukseen liittyvä ongelma on empiirisen näytön puute sitä, että "suora" performanssitestaus ennustaa tavoitteena olevaa kielikäyttäytymistä paremmin tai edes yhtä hyvin kuin "epäsuora", kenties helpommin toteutettava testaus. Systemaattista seurantaa ei ole vielä tehty. Toinen ongelma syntyy siitä, että täydellinen simulaatio on mahdotonta, koska emme tunne todellisuutta kokonaan ja vaikka tuntisimme, testitilanne asettaa omat rajoituksena. Simulaatiossa abstrahoidaan todellisuutta ja simuloitavien piirteiden valintaa on joka tapauksessa tehtävä tietoisesti, eli on tingittävä siitä autenttisuuden maksimoinnin periaatteesta, joka on performanssitestauksen keskeisiä ajatuksia. Mielekkään simuloinnin aikaansaaminen edellyttää adekvaattia teoreettista mallia kielenkäyttötapahtuman relevanteista piirteistä. Esimerkiksi Hallidayn (mm. 1978) malli kielenkäyttöä määrittävistä tilannepiirteistä tarjoaa mahdollisen lähtökohdan. Kielenkäyttötilanteiden piirteistä on olemassa moniakin luetteloita, mutta yleensä ne ovat eklektisiä taksonomioita, joista on kyllä apua uusien testien laadinnassa, mutta jotka teoreettisen viitekehyksen puuttuessa ovat hiukan satunnaisia ja epäsystemaattisia.

Kolmas ongelma on se, että vaikka observoisimme autenttista kommunikaatiotilannetta (tai simulaatio olisi täydellinen), on sen edustavuus kaikkiin tavoiteltuihin kommunikaatiotilanteisiin nähden voitava todeta jollain menetelmällä.

Neljänneksi on varsinaisen arvioinnin ongelma: mitä ilmiöitä pyritään kommunikaatiotilanteessa havainnoimaan ja miten niitä arvostellaan? Performanssitestin arviointi saattaa olla hankalampaa kuin tilanteesta irrotettua kielitaitoa mittaavan testin, ainakin tavanomaisen koulutuksen saaneille kieltenopettajille. Arviointikriteerien onnistuneisuus ja taitava soveltaminen ovat kuitenkin testeille yhtä tärkeitä kuin sopivat tehtävät. Myös kriteerien on ilmennettävä todellisen, testin ulkopuolisen maailman vaatimuksia. (vrt. Bingham Wesche 1987.)

Lopuksi on vielä kysymys siitä, soveltuuko performanssitestaus kielikeskukseen ylipäänsä, kun opiskelijoiden tulevat tarpeet ovat osittain tuntemattomia ja joka tapauksessa melko etäisiä, ja opiskelijoilta puuttuu monia tiedolliseen ja persoonalliseen kehitykseen kuuluvia valmiuksia, joita ammatti-ihmiseltä edellytetään. Jos lähennämme testejä opiskelijoiden nykytilanteeseen, joudumme ristiriitaan performanssi-idean kanssa, jossa testien realismi johdetaan tavoitteista käsin. Tämä koskee ennen kaikkea suullista kielitaitoa ja mahdollisesti mukaantulevaa kirjoitustaitoa. Tekstinymmärtämistä mitataan varsin samoihin aikoihin ja verrattain samanlaisissa olosuhteissa kuin sitä tarvitaankin.

Useimmat yllä esitetyistä ongelmista ovat luultavasti ratkaistavissa tai niihin voidaan kehittää mielekäs kompromissi kielikeskusten tarpeita vastaamaan. Teoreettisen ja periaatteellisen työn tueksi tarvitaan käytännön kokeilutyötä, jonka nojalla ratkaisujen kelvollisuutta voidaan arvioida ja parempia ratkaisuja kehitellä.

4. KRITEERIVIITTEINEN TESTAUS

Koska kielikeskusten testauksen tarkoituksena on arvioida opiskelijoiden kielitaidon riittävyyttä tavoitteisiin nähden pääsääntöisesti dikotomisesti, ei normiviitteinen mittaaminen ole yleensä tarkoituksenmukaista. Tärkeämpää kuin verrata opiskelijoiden suoritustasoja toisiinsa, mikä on normiviitteisen testauksen perusta, on selvittää kunkin opiskelijan osalta se, vastaako hänen taitonsa sitä tasoa, jolla odotettavissa olevista kielitaitovaatimuksista voidaan selviytyä. Tämä edellyttää, että tavoiteltu kielitaito ja vaadittu taso on määritelty selkein kriteerein. Kriteerien

spesifikaation tulee nojata kielenkäyttötarpeiden kartoituksiin ja kuvauksiin tavoitteena olevan kielenkäytön piirteistä.

Suomalaisten kielitaitotarpeita ollaan melko perusteellisesti kartoittamassa KTL:n tutkimusprojektissa (ks. Mehtäläinen 1987), mikä tarjoaa hyvän pohjan myös kielikeskustestien sisällönmäärittelylle eri aloilla. Tarvekartoitukset perustuvat kyselyihin ja haastatteluihin, joten niistä yleensä ilmenee, minkä tasoista kielitaitoa tarvitaan ja missä yhteyksissä.

Se, mikä tarvekartoituksesta ei käy ilmi, on mitä kielenkäyttötilanteissa todella tapahtuu ja mitkä puolet kommunikoinnissa korostuvat tai aiheuttavat ongelmia. Hyvien, realististen testien pohjaksi tarvitaan myös suomalaisille akateemisen koulutuksen saaneille kansalaisille keskeisten, autenttisten kommunikaatiotilanteiden havainnointiin ja analyysiin perustuvaa tietoa.

Näiden ohella tarvitaan myös päätöksenvaraisia, teoreettisiin kuvauksiin ja malleihin perustuvia spesifikaatioita. Keskeinen yleisperiaate on kielen käytön kokonaisvaltainen tarkastelutapa: kommunikaatiotapahtuma on nähtävä kokonaisuutena. Parhaiten tämä saavutetaan lähestymällä sitä yhtaikaa sekä sosiaalisen kielenkäyttötilanteen näkökulmasta että kielenpuhujan kapasiteetista käsin.

Kriteerien määrittelyn yhteydessä on hyvä konsultoida testien kaikkia käyttäjäryhmiä. Tärkein palaute käyttäjäryhmiltä saadaan kuitenkin testien kokeilu- ja käyttövaiheen seurannassa.

5. MYÖNTEINEN TESTAUS

Testeihin tavallisesti liittyvät kielteiset tunteet ja stressi, joita kokevat sekä testaajat että testattavat, on pyrittävä minimoimaan kielikeskuksissa ja lähestymään testausta pikemminkin myönteisen palautteen antajana. Tämä edellyttää suurilta tasokokeilta mahdollisesti jonkinlaista diagnostista tulosta ja kursseilta formatiivisen evaluaation kehittämistä.

Opiskelijoiden kannalta myönteinen testaus merkitsee lisääntyvää opiskelijakeskeisyyttä: opiskelijoiden omat näkemykset kielitaidostaan ja esteistä otetaan huomioon. On myös mahdollista, että opiskelijoiden arvio testimuotojen ja tehtävien onnistuneisuudesta vaikuttaa testisuoritukseen ja sitä kautta itse testin validiteettiin (vrt.

Low, 1985). Joka tapauksessa voidaan olettaa, että opiskelijoiden motivaatioon ja testauksen ja opetuksen ilmapiiriin on myönteinen vaikutus sillä, että opiskelijoiden arviot testeistä, omasta kielitaidostaan sekä näiden välisestä suhteesta otetaan huomioon testauksessa ja sen kehittämistyössä.

On tärkeätä, että testien laadinnassa pyritään sellaisiin tehtäviin, joissa opiskelija voi näyttää taitonsa eikä sellaisiin, joissa hänen puutteensa koetetaan mahdollisimman tehokkaasti paljastaa. Samoin arvostelussa on tavoiteltava suoritusten myönteistä, kokonaisvaltaista arviointia eikä virheiden etsimistä.

Testauksen tekeminen opettajien kannalta myönteisemmäksi edellyttää opettajien kolmen eri roolin huomioon ottamista testauksessa: ensinnäkin he joutuvat usein laatimaan testejä, jolloin työtä helpottavat yksi tulkintaiset laadintaohjeet; toiseksi he joutuvat toteuttamaan ja arvostelemaan testit, mikä merkitsee, että hyvä testi on vaivaton toteuttaa ja arvosteluperusteiltaan mahdollisimman selkeä eikä kohtuuttoman aikaavievä. Opettajien kolmas rooli on niiden kurssien opettajana, joille testi valikoi opiskelijat. Testin on siis annettava myös mielekästä tietoa opetukselle.

6. MITÄ PYRITÄÄN MITTAAMAAN

Testattavilta edellytettävää kielikäyttäytymistä (performanssia) voi kuvata sen sosiaalisen kontekstin näkökulmasta, jossa kieltä tulee osata käyttää tai niiden tieto- ja taitoelementtien valossa, jotka testattavan tulee hallita. Näiden lisäksi testitehtävien osalta voidaan kuvata niitä ominaisuuksia, joita "kommunikatiivisen realismin" nimissä pidetään tärkeinä ('dynamic communicative characteristics', Weir, ms.) sekä tehtävien muita dimensioita.

Kielikeskustestien johtava periaate on holistinen lähestymistapa kielenkäyttötaidon arviointiin. Tämä merkitsee sitä, että kaikki yllämainitut näkökulmat otetaan jollain tavoin huomioon, mutta keskeisinä on kommunikaatiotilanteen ja opiskelijan kielikäyttäytymisen tarkastelu kokonaisuutena.

Kielenkäytön sosiaaliseen kontekstiin perustuva kuvaus tuottaa tulokseksi testauksessa tavoiteltavan kielellisen aineksen spesifikaation, jonka nojalla määräytyvät testin pohjana olevien diskurssiotteiden laji (testistimulus) samoin kuin testattavan reaktion diskurssilajit. Tässä kuvauksessa voi hyvin hyödyntää Hallidayn (1978) teoreettista mallia, jossa kielenkäytön kannalta relevantti sosiaalisen tilanteen

kuvaus voidaan kaikkein abstrakteimmalla tasolla suorittaa käsitteiden 'field' (institutionaaliset puitteet, se, mistä toiminnasta on kyse ja mistä puhutaan), 'tenor' (osanottajien väliset suhteet, status, muodollisuusaste jne.) ja 'mode' (viestintäkanava, ts. kirjallinen, suullinen ja näiden täsmennyksiä) avulla. Nämä tekijät yhdessä määrittävät kielenkäytön 'rekisterin'. Tämä malli antaa mahdollisuuden määritellä tilannetyypin, mikä on tarpeen silloin kuin halutaan tuottaa uusia, keskenään verrattavissa olevia testitehtäviä eikä pitäytyä jatkuvasti samoissa. Realistisuuden kannalta testitehtävien vaihtelu on eduksi, koska se vähentää stereotyyppisten tilanteiden ennaltaharjoittelua.

Hallidayn malli ei käsittele kielenkäyttötaitoja, jotka ovat pikemminkin kompetenssin käsitteeseen kuuluvia. Kielitaidon testauksen kannalta yksi käyttökelpoisimmista kommunikatiivisen kompetenssin malleista lienee Canalen ja Swainin (1980, ja Canale 1984), joka koostuu neljän keskeisen tieto- ja taitoalueen yhdistelmästä: kieliopillinen kompetenssi (kielikoodin hallinta), sosiolingvistinen kompetenssi (taito sovittaa kielenkäyttö sosiaalisen tilanteen vaatimuksiin), diskurssikompetenssi (taito tuottaa ja ymmärtää yhtenäistä, koherenttia ja kohessiivista kieltä) ja strateginen kompetenssi (kyky kompensoida kommunikaation puutteita ja tehostaa viestintää).

Kommunikatiivisen kompetenssin osa-alueet muodostavat testitehtävien laadinnalle psykolingvistisen pohjan, ja samalla ne jäsentävät diagnostista arviointiskaalaa. Toisin sanoen opiskelijan suoritusta havainnoidaan ja arvioidaan kullakin näistä osa-alueista. Tavoitteena ei ole arvioida opiskelijan kompetenssia sinänsä vaan hänen suoriutumistaan kommunikaatiotehtävistä näistä neljästä näkökulmasta. Tämä lähestymistapa poikkeaa ratkaisevasti sellaisista, joissa kenties kommunikatiivistenkin tehtävien avulla tähdätään ensisijaisesti kielitaidon arviointiin. Kun tavoitteena on kommunikatiivisen performanssin arviointi, ovat esim. strategista kompetenssia ilmentävät kiertoilmaukset positiivisia osoituksia taidosta, vaikka puhdasta kielitaitoa mitattaessa ne pikemminkin osoittavat taidon puutteita.

Yksittäisten testien laatimisperiaatteita varten on toivottavia taitoja edelleen täsmennettävä, koska opetuksen tarvitsemaa diagnostista informaatiota varten relevantteja osataitoja täytyy esimerkiksi lukemisessa testata eri tehtävin. Samoin on määriteltävä paitsi tehtävien ulottuvuudet (laajuus, kompleksisuus jne.), myös ainakin ymmärtämistehtävien osalta ymmärtämisen syvyys tai taso. Sekä kuullunettä tekstinymmärtämistehtävien sovittamisessa toivotulle ymmärtämisen tasolle voidaan syväsuuntautuneen lukemisen käsitettä (Marton et al., 1980) soveltaa myös vieraskielisen diskurssin sisällön ymmärtämiseen. Niinikään Biggsin ja Collisin

(1982) SOLO-taksonomiaa voi hyödyntää silloin kun sisällönymmärtämiskysymysten laatua arvioidaan muulta kuin lingvistiseltä kannalta.

Tällaista tehtävien arviointia on syytä tehdä nykyistä enemmän, sillä eri tehtävätyypit edellyttävät erilaisia ei-kielellisiä taitoja ja saattavat siten suosia eri persoonallisuustyyppejä tai eri kognitiivisia tyylejä. Samaan testikokonaisuuteen ei tietenkään haluta useita tehtäviä, jotka suosivat samoja persoonallisuudenpiirteitä.

7. TESTAAMISEN MENETELMÄT

Testaamisen muotoja ja menetelmiä hallitsee kaksi yleisperiaatetta: pyrkimys autenttisuuteen ja pyrkimys monipuolisuuteen.

Autenttisuuden tavoittelu testimuodoissa perustuu sekä itse testien validisuuden ja niiden antaman ennusteen luotettavuuden parantamiseen että testauksen ja opetuksen suhteisiin, jos testit jäljittelevät todellisen elämän kielenkäyttövaatimuksia, myös opetuksella on mahdollisuus suuntautua samoin, ilman että erikseen harjoitellaan testimuodoista selviytymistä.

Opetuksen vapautta edistää myös pyrkimys monipuolisuuteen ja useamman erilaisen osatestin käyttöön. Tämä on tarpeen myös siksi, että mittaamismenetelmien puutteet kompensoisivat toisiaan. Samaan testiin on syytä sisällyttää osia, jotka suosivat erilaisia tukitaitoja (enabling skills) ja erilaisia persoonallisuuden ja kognitiivisia tyylejä.

Siltä osin kuin pyritään diagnostiseen testaukseen, täytyy diagnosoitavien osa-alueiden määrittämisen perustua samanlaisiin kommunikaatiotilanteen kannalta relevantteihin kategorioihin kuin testien laatimisen ja arvostelunkin. Diagnostisen testauksen on palveltava opetuksen tarpeita niin, että se antaa käyttökelpoista tietoa. Yhdellä moniosaisella testikokonaisuudella voidaan esimerkiksi mitata kielellisen pintatason hallintaan perustuvaa tekstinymmärtämistä yhdellä osatestillä (esim. semanttinen cloze, kysymykset tekstistä) ja itsenäiseen, tekstisisällön pääkohtien löytämiseen perustuvaa ymmärtämistä toisella (esim. tiivistelmä): tulosten mukaan opetusta voidaan suunnata joko peruskielitaidon kohentamiseen, lukustrategioiden tehostamiseen tai molempiin. Tämän yksityiskohtaisempaan diagnostiseen analyysiin on tuskin aihetta pyrkiä, koska eri osatietojen lukumäärä ja

merkitys kokonaissuoritukselle ovat melko vähän tunnettuja. Kielellisten detaljien hallintaa ei ole tarkoituskaan selvittää.

Suunnittelussa on otettava huomioon myös opetuksen mahdollisuudet käyttää saatua tietoa esimerkiksi opetusryhmien muodostamisessa tai tukiopetuksen järjestämisessä - on turha hankkia sellaista tietoa, jota ei käytännössä pystytä hyödyntämään. Esimerkiksi tekstinymmärtämistaidon testaaminen erikseen silloin, kun sitä ei voida opettaa erikseen, on tarpeetonta, samoin vaikkapa lauserakenteiden tuntemuksen. Myös kurssien päättökokeissa saattaa olla tarpeetonta hankkia opiskelijoiden taitoprofiilista kovin yksityiskohtaisia tietoja, kun korjaavaa opetusta ei voida antaa.

Mittausmuodoissa tarvitaan myös kokeilevaa elementtiä: moniosaisiin testeihin voi melko helposti sijoittaa yhden kokeilevan osan, jos muut osat ovat tunnettuja ja koeteltuja. Testauksessa voi näin kaiken aikaa hyödyntää ja koetella kieltenopetuksen uusia teoreettisia ja menetelmällisiä virtauksia.

Realistinen testaus kielikeskuksissa edellyttää sellaisten testimuotojen suosimista, jotka vaativat testattavilta melko itsenäistä kommunikaatiotehtävien suorittamista: kirjallisesta tai suullisesta diskurssista on pystyttävä omatoimisesti löytämään tilanteen kannalta relevantti aines, ja tuottavassa kielenkäytössä on kieltä osattava käyttää kommunikaatiotavoitteen saavuttamiseksi myös silloin kun se edellyttää oma-aloitteisuutta.

Testausolosuhteet eivät kuitenkaan aina anna tilaisuutta vapaaseen ja itsenäisyyttä edellyttävään mittaukseen, joka tyypillisesti on varsin yksilöllistä ja jonka arvostelu on melko työlästä ja aikaaviepää. Sidotumpia testimuotoja tarvitaan näiden ohella jatkuvasti, erityisesti kun suuria joukkoja on testattava nopeasti. Myös näitä menetelmiä täytyy kehittää realistisuuden ja kokonaisvaltaisuuden hengessä. Sidottujen testien, kuten monivalintatestien, kehittäminen painottuu testin laadintavaiheeseen ja sen yhteydessä tapahtuvaan kokeiluun, kun taas vapaammissa testimuodoissa korostuu arvioitsijoiden osuus.

Suurten opiskelijajoukkojen testauksessa on yhtenä ratkaisuna monivaiheinen testaus: ensimmäisessä vaiheessa käytetään joukkotestiä suorittamaan alkukarsinta, jonka läpäisseet opiskelijat ohjataan yksilötestaukseen. Tavallista on ollut käyttää tekstin tai kuullun ymmärtämistestiä joukkotestinä ja siirtyä sitten yksilöllisempään tuottamistestiin. Jollei tällaisen testin tarkoituksena ole antaa opetusta ohjaavaa

diagnostista tietoa opiskelijoiden tasosta nimenomaan eri taitoalueilla, saattaisi parempi ratkaisu olla integratiivisen joukkotestin (esim. osittaissanelu, jokin cloze-variantti) käyttö ennen yksilötestiä. On kenties järkevämpää alkajaisiksi selvittää opiskelijan kielellinen yleistaso ja sitten siirtyä yksilöllisempiin soveltamistehtäviin tarpeen mukaan, kuin suorittaa ensimmäinen erottelu jonkin yksittäisen taitoalueen mukaan. Tällä alueella kaivataan selvitys- ja kokeilutyötä.

6. ARVIOINNIN PERUSTEISTA

Opiskelijoiden suoritusten arvioinnissa pääperiaatteena on kommunikaation kokonaisvaltainen huomioonottaminen. Toinen keskeinen tavoite on kriteerien sovittaminen autenttisen kommunikaatiotilanteen vaatimusten mukaisiksi.

Nämä periaatteet merkitsevät, että arvostelussa korostuu se, miten adekvaattia kielenkäyttö on käsillä olevan tilanteen sujumisen kannalta samoin kuin se, mitä kyseisessä tilanteessa pidettäisiin tärkeänä todellisessa elämässä. Mm. oikeakielisyyden vaatimukset voivat vaihdella tilanteen tarpeiden mukaan, samoin ymmärtämisen tarkkuus tai syvyys jne. Viestin sisällön ja viestijän tarkoituksen selkeä esiintuominen voi olla tärkeämpää kuin varsinaiset kielelliset muotoseikat.

On välttämätöntä, että arviointi tapahtuu samansuuntaisilla kriteereillä kuin itse testimateriaalin valinta ja tehtävien laadinta, jotta testauksen periaatteet pääsevät toteutumaan. Autenttisen testauksen tavoitteeseen ei voi päästä realistisimpienkaan tehtävien avulla, jos arvioinnissa keskitytään pelkästään esimerkiksi kielivirheisiin tai sanavaraston laajuuteen. Arvostelun osuus koko testissä on erityisen suuri sellaisten tehtävien yhteydessä, jotka edellyttävät opiskelijalta omaa tuotosta.

Yksittäisten testien arviointikriteerit on sovittava kunkin testaustilanteen vaatimuksiin. Hyvän lähtökohdan tälle tarjoavat ne yhteiset testien laatimisperiaatteet, joita käsiteltiin edellä kohdassa "mitä pyritään mittaamaan". Kommunikaation yleistä tasoa voi arvioida Hallidayn rekisteri-käsitteen avulla ja spesifimmät arvostelukriteerit voivat nojautua kommunikaation osa-alueisiin Canalen ja Swainin kompetenssimallin pohjalta.

Arvioinnin luotettavuuden takaamiseksi täytyy varsinkin kokemattomia ja uusia arvioitsijoita perehdyttää tehtäväänsä mahdollisimman hyvin. Joidenkin tutkimustulosten mukaan opettajaryhmien keskenään laatimat yhteiset arviointikriteerit

eivät paranna arvioinnin luotettavuutta, ellei niiden pohjana ole valmiita ja selkeitä testaamisperiaatteisiin nojaavia normeja. Arviointiin perehdyttävät ohjeet tarpeellisine harjoitusesimerkkeineen esimerkiksi videonauhalla olisivat hyödyllisiä arvioitsijakoulutuksen tueksi. Yhdenmukainen arvioitsijakoulutus on tarpeen myös jatkuvan arvioinnin perustana.

Arviointiskaalojen täytyy olla yksinkertaisia, mahdollisimman yksiselitteisiä ja helppokäyttöisiä sekä pyrkiä välttämään haloeffektiä vapaissa tehtävissä esimerkiksi vaihtelemalla asteikkoa tehtävästä toiseen (vrt. Tung, 1985).

Yksi alue, jolla tarvitaan luotettavaa arviointia on kielitaitotodistukset. Niissä sovellettavat kuvailevat kategoriat täytyy sovittaa yhteen sekä testauksen arvostelutavan että todistusten ajatellun käytön kanssa. On tärkeää, että kielitaidon kuvaus on ymmärrettävä sekä opiskelijalle itselleen että tuleville työnantajille tai muille kysymykseen tuleville kielenopetuksen terminologiaan vihkiytymättömille henkilöille. Kielikeskusten itsensä kannalta on merkitystä myös sillä, että ne tiedekunnat ja korkeakoulut, joita kielikeskukset palvelevat, ymmärtävät ja hyväksyvät kielikeskuksissa sovelletut mittausperiaatteet siltä osin kun ne ovat niiden kannalta relevantteja eli juuri vaadittavan kielitaidon kuvauksen osalta. Tällainen tavoitteiden selkeys saattaa edesauttaa yhteisymmärrystä opetuksen järjestämisessä.

Todettakoon lopuksi periaatteellisena yhteenvetona, että kielikeskustestejä on kehitettävä yhä elämänläheisemmiksi eli autenttisuuteen pyrkivän performanssitestauksen suuntaan, joka perustuu kokonaisvaltaiseen näkemykseen kommunikaatiotapahtumasta. Tärkeätä on myös, että testaus on myönteistä ja käyttäjäystävällistä: opiskelijoiden täytyy kokea saavansa oikeudenmukaista ja hyödyllistä palautetta kielitaidostaan sekä päästä osallistumaan opetukseen silloin kun se on tarpeen.

LÄHTEET

Alderson, J. Charles (1986), Innovations in language testing? teoksessa Portal, M., (toim.), Innovations in Language Testing. Windsor: NFER-NELSON

Alderson; J. C. ja Urquhart, A. H. (1983). The Effect of Student Background Discipline on Com prehension: a Pilot Study. Teoksessa Hughes, A. ja Porter, D. (toim.), Current Developments in Language Testing. Academic Press

Biggs, J. B. ja Collins, K. F. (1982). Evaluating the Quality of Learning. New York: Academic Press

Bingham Wesche, M. (1987). Second language performance testing: the Ontario Test of ESL as an example. Language Testing 4/1

Bullivant, D., Lönnfors, P., Nordlund J., Satchell, R., (1987). Needs Analysis for the Lay Person. Kielikeskusuutisia 9/1987

Canale, M. ja Swain, M., (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1/1, 1-47

Canale, M. (1984). A communicative approach to language proficiency assessment in a minority setting. Teoksessa Rivera, C. (toim.), Language Proficiency and Academic Achievement. Clevedon: Multilingual Matters

Carrell, P. (1984). Evidence of a Formal Schema in Second Language Reading Comprehension. Language Learning 34, 2, 87-112

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, Mass.,: M.I.T.Press

Economou, D. (käsikirjoitus), Language Testing in Academic Contexts

Halliday, M. (1978). Language as Social Semiotic. London: Edward Arnold

Hauptman, P. C., LeBlanc R., Bingham Wesche, M. (toim.) (1985). Second Language Performance Testing. Ottawa: University of Ottawa Press

Hymes, D. H. (1970). On Communicative Competence. Teoksessa Gumperz, J. J. ja

Hymes, D. H. (toim.), Directions in Sociolinguistics. Holt, Rinehart and Winston

Jones, R. L. (1985). Second Language Performance Testing: an Overview. Teoksessa Hauptman et al. (toim.)

Low, G. (1985). Validity and the problem of direct language proficiency tests. Teoksessa Alderson, C. (toim.), Evaluation. Lancaster Practical Papers in English Language Education, Vol. 6. Oxford: Pergamon

Marton, F., Dahlgren, L. O., Svensson, L., Säljö, R., (1980). Oppimisen ohjaaminen. Espoo: Weilin & Göös

Mehtäläinen, J. (1987). Teollisuuden toimihenkilöiden kielitaidon tarvetutkimus. Teollisuuden koulutusvaliokunta, monistesarja 8

Tung, P. (1985). Designing Oral Proficiency Tests in EFL for Hong Kong Secondary Schools. Teoksessa Hauptman et al. (toim.)

Widdowson, H. (1983). Learning Purpose and Language Use. Oxford University Press

Ari Huhta
Korkeakoulujen kielikeskus

UUDET ENGLANNIN TEKSTIN YMMÄRTÄMISEN KOKEET

Osana Korkeakoulujen kielikeskuksen kielitaidon mittaamisen kehittämisprojektia perustettiin syksyllä 1986 englannin tekstin ymmärtämisen opettajista ja tutkijoista työryhmä laatimaan uusia luetun ymmärtämisen kokeita englannin kieleen. Ryhmään kuuluu edustajia Korkeakoulujen kielikeskuksesta ja viidestä paikallisesta kielikeskuksesta. Tähän mennessä on saatu valmiiksi n. 80 mallikoetta, joista puolet on perinteisiä monivalintatestejä tai monivalintacloze-testejä ja puolet avoimia kysymyksiä, tiivistelmärungon ja tekstin rakennekaavioiden täydennystehtäviä. Cloze-testejä ovat laatineet paljon myös sellaiset Helsingin yliopiston kielikeskuksen opettajat, jotka eivät ole kuuluneet työryhmän ydinjoukkoon.

Alunperin tavoitteena oli vain luoda erityyppisiä mallitestejä, joiden avulla kielikeskusopettajat voivat itse laatia tarvitsemiaan testejä. Nyttemmin on päädytty siihen, että mallikokeiden lisäksi kootaan testipankkia niitä opettajia varten, joiden ei ole mahdollista itse laatia omia kokeitaan mallien perusteella. Tarkoitus on siis kerätä mahdollisimman monien kielikeskusopettajien laatimia testejä, kokeilla niitä käytännössä ja levittää niitä muiden opettajien käyttöön sitten, kun ne on analysoitu ja niiden on todettu toimivan hyvin. Mallikokeisiin laaditaan kokeiluista saatavan palautteen perusteella mahdollisimman selkeät laatimis- ja suoritusohjeet. Nämä sisältävät tietoa testityypin käyttötarkoituksesta, pisteytyksestä sekä tehtävien ja vastausohjeiden laatimisesta, jotta kokeiden teko ja käyttö olisi mahdollisimman vaivatonta.

Pääsyy uusien testien laatimiseen on välttää liian yksipuolisesta lukemistaidon mittaamisesta. Tähän asti yleisin testaustapa on ollut monivalintatekniikka, vaikka toki muitakin menetelmiä on eri kielikeskuksissa käytetty. Testaustekniikalla on kuitenkin oma vaikutuksensa koetuloksiin (ks. esim. Bachman 1990), ja jos käytetään vain yhtä tekniikkaa, osa testattavista voi kärsiä siitä suhteettomasti, koska eivät joko tunne kyseistä kielitaidon mittaamismenetelmää tai pidä siitä. Lisäksi monivalintaa on kritisoitu mm. kommunikatiivisuuden puutteesta (hyvä katsaus monivalintaa vaivaavista ongelmista löytyy Weiriltä, 1988). Useiden testityyppien käyttö vähentäisi menetelmästä johtuvia virheitä kielitaidon mittaamisessa. Toinen perustelu monien erilaisten testaustapojen käyttämiseen tulee kieliteorioista, tai

oikeammin niiden puuttumisesta: luetun ymmärtämisen tutkijat eivät näytä vielä päässeen yksimielisyyteen esimerkiksi siitä, kuinka monesta osataidosta lukutaito koostuu. Varman tiedon puutteessa turvallisinta on käyttää useita, lukutaidon eri puolia mittaavia testejä, jos halutaan saada kattava kuva testattavan lukutaidosta.

Käyttämällä uudentyyppisiä testejä kielikeskusten luetun ymmärtämiskokeisiin saadaan lisää kattavuutta ja samalla yksipuolisesta monivalintatekniikasta johtuvat haitat vähenevät. Testaus saadaan myös paremmin vastaamaan opetusta: nykyään monilla kursseilla opetetaan esimerkiksi tekstin tiivistämistä ja sen rakenteen hahmottamista, mutta tiivistelmien käyttö kokeissa ei vielä ole kovin yleistä.

TUTKIMUSMENETELMÄT

Testin ominaisuuksia voidaan tutkia sekä ennen testin käyttöä että sen jälkeen. Testin ominaisuuksista tärkeimmät ovat validiteetti (mittaako testi luetun ymmärtämistä vai jotain muuta taitoa), reliabiliteetti (miten johdonmukaisesti testi mittaa haluttua ominaisuutta) ja käyttökelpoisuus. Etukäteen pyritään jo ennen kokeen järjestämistä varmistamaan, että se on mahdollisimman hyvä ja käyttökelpoinen. Testaajilla on mahdollisimman selkeä käsitys siitä, mitä ollaan mittaamassa ja kuinka hyvä testi laaditaan. He yrittävät laatia kokeen, joka mittaa juuri niitä ominaisuuksia, joita halutaan testata; he käyvät läpi toistensa tuotoksia ja yhdessä niitä parannellen valmistavat lopullisen version kokeesta. Heillä voi myös olla mahdollisuus kokeilla testiä ennen varsinaista koepäivää. Pääpaino etukäteisvalmisteluissa on kuitenkin asiantuntemuksella ja teoreettisella tiedolla, empiirinen tutkimus tulee vasta, kun koetulokset ovat selvillä. Eniten etukäteen tapahtuva tutkimus vaikuttanee kokeen validiteettiin.

Jälkikäteen testien hyvyyttä on yleensä tutkittu tilastollisilla menetelmillä; tällainen tutkimus keskittyy yleensä kokeen reliabiliteettiin, joskus myös validiteettiin.

Kokeen käytännöllisyys selviää sekä ennen että jälkeen kokeen järjestämisen; testin valistamisen helppous käy ilmi tietysti jo etukäteen, mutta esimerkiksi pisteytyksen ja tulosten tulkinnan helppous selviää usein vasta sitten, kun koetta on jo käytetty.

Testin validiteetin, reliabiliteetin ja käytännöllisyyden lisäksi voi olla tärkeää saada selville, miten uskottavana, miten hyvänä kielitaidon mittarina testattavat tai testitulosten käyttäjät (hallinto, työnantajat, jne.) koetta pitävät.

Tässä tarkasteltavina olevia kokeita on tutkittu etukäteen. Laatijoilla on ollut tietoa siitä, miten laatia toimivia kysymyksiä ja vastausvaihtoehtoja; heillä on myös ollut käsitys siitä, mitä taitoja pitäisi mitata. Tärkeänä pidetty taito on esimerkiksi tekstin pääasioiden ymmärtäminen, mitä osoittaa tiivistelmätestien yleistyvä käyttö sekä se, että monet muidenkin testien kysymyksistä kohdistuvat tekstin pääsisältöön. Lisäksi cloze-testien laadinnassa suositaan aukkoja, joiden täydentämiseksi olisi käytettävä hyväksi mahdollisimman laajaa kontekstia. Toisaalta käytännön syistä on haluttu käyttää mahdollisimman paljon helposti korjattavia monivalintakokeita huolimatta niihin liittyvistä ongelmista (vrt. Weir 1988).

Pääpaino tässä raportoitavassa tutkimuksessa on kuitenkin testien jälkikäteen tapahtuvalla analyysillä. Helpointa oli tutkia reliabiliteettia, mikä onkin laskettu hyvin monille testeille. Sekään ei kuitenkaan ole aivan ongelmatonta: ensinnäkin monivalintakokeista lasketaan yleensä niiden sisäinen johdonmukaisuus eli missä määrin eri kysymykset mittaavat samaa taitoa, mutta esimerkiksi tiivistelmäkokeen reliabiliteetissä on kyse eri asiasta, nimittäin eri arvioijien pisteytyksen yhdenmukaisuudesta. Eri kokeiden reliabiliteetti-arvoissa voi siis olla kyse eri asioista. Tämän tutkimuksen reliabiliteetti-arvot ilmaisevat kaikki kokeen sisäistä johdonmukaisuutta; muille paitsi monivalintakokeille (mukaan lukien monivalintaclozet) ei reliabiliteetteja ole vielä laskettu.

Tärkeämpää, mutta myös vaikeampaa on selvittää kokeiden validiteetti eli niiden kyky mitata juuri sitä taitoa, mitä ne on suunniteltu mittaamaan. Validiteetin tutkimus jakautuu itse asiassa useampaan osaan. Perinteinen tutkimusasetelma on se, että uutta koetta verrataan jo olemassaolevaan kokeeseen: mitä paremmin uuden kokeen tulokset korreloivat vanhan kokeen tuloksiin sitä todennäköisemmin uusi koe mittaa samoja taitoja kuin vanha koe. Tässä tutkimuksessa vertailukokeina eli ns. ulkopuolisina kriteereinä olivat Korkeakoulujen kielikeskuksessa 1970-luvun lopulla valmistetut ja standardoidut monivalintakokeet: niissä kunkin tekstin jälkeen on kymmenen ymmärtämistä mittaavaa kysymystä, joista jokaisessa on neljä vastausehdotusta. Ongelmana tutkimusasetelmassa on se, mitä vanhat kokeet mittaavat: koko lukemistaitoa, joitakin sen osia vai muitakin taitoja kuin lukemista? Korrelaatioiden tulkinnan luotettavuus riippuu siis täysin siitä, miten valideja lukemistestejä vertailussa käytetyt monivalintatestit ovat. Yleisin tutkimusasetelma tässä raportoitavissa kokeiluissa oli sellainen, jossa testistö koostui yhdestä kokeilukokeesta ja 2 - 4 vanhasta monivalintakokeesta, joissa kussakin oli kymmenen kysymystä.

Validoinnissa käytettiin muitakin vertailukohtia kuin monivalintatestit. Joissakin kokeiluissa opiskelijat vastasivat testin hyvyyttä ja omaa lukutaitoaan kartoittavaan kyselyyn. (Liite 1) Opiskelijoiden arviota omasta englannin kielisen tekstin lukutaidosta ja heidän ylioppilastodistuksensa englannin arvosanaansa verrattiin menestykseen kielikeskuksen kokeessa. Hyvä korrelaatio oman lukutaitoarvion ja testimenestyksen välillä antaa aihetta uskoa, että testi mittaa lukutaitoa. Ylioppilaskoe voi olla hiukan huonompi vertailukohta, koska sen arvosana määräytyy muidenkin taitojen perusteella kuin lukemisen.

Testissä käytetyn tekstin vaikutusta koemenestykseen haluttiin myös tutkia. Tämä tapahtui kysymällä opiskelijoilta arviota kunkin tekstin kiinnostavuudesta ja vaikeudesta. Yksi keskeisiä ongelmia erityisalojen kielitaidon mittaamisessa on, miten lähellä opiskelijan omaa alaa tekstin on oltava, jotta testi olisi oikeudenmukainen. Tekstin aihepiirin on nimittäin havaittu vaikuttavan menestymiseen sen pohjalta tehdyssä testissä (vrt. Alderson ja Urquhart 1985).

Toisinaan testin uskottavuutta (ilmeisvaliditeetti, face validity) pidetään myös validiteetin osana. Opiskelijoilta kysyttiin näin myös heidän arviotaan testistön eri osista nimenomaan lukemistaidon mittareina. Epäuskottava testi saattaa laskea testattavan motivaatiota ja vaikeuttaa menestymistä, joten ei ole aivan merkityksetöntä, miltä testi "näyttää".

Testejä järjestäneiltä kielikeskusopettajilta kyseltiin, millaista oli järjestää ja pisteyttää erityyppisiä kokeita. Palautetta tästä käytetään hyväksi esimerkiksi laadittaessa pisteytysohjeita ja mallivastauksia kokeisiin ja parannettaessa kokeiden käytännöllisyyttä.

Kokeista selvitettiin myös niiden vaikeustaso, koska usein on tärkeää tietää onko koe vaikea, tavanomainen vai helppo. Lisäksi hankittiin tietoa opiskelijoiden muusta kieli- ja opiskelutaustasta: opintojen vaiheesta ja oleskelusta englanninkielisissä maissa, joilla voi olla vaikutusta testitulokseen.

TULOKSIA

Tässä esitetyt tulokset perustuvat useisiin kymmeniin kokeiluihin eri kielikeskuksissa; useimmiten kyseessä on ollut kurssin alkukoe, joissa opiskelijat ovat valtaosin juuri opintonsa aloittaneita. Tarkemmin on analysoitu Jyväskylässä syksyllä 1988

järjestetyt humanistien ja kasvatustieteilijöiden suuret alkukokeet. Myös Tampereella syksyllä 1987 järjestettyjä yhteiskuntatieteen ja humanistien alkukoetta analysoitiin tarkemmin. Jyväskylän ja Tampereen kokeisiin liittyi opiskelijakysely (liite 1).

Taulukko 1: TARKEMMIN ANALYSOITUJEN KIELIKOKEIDEN RAKENNE JA OSALLISTUJAT
(mv = vanha standardoitu monivalintatesti)

Tampere 1987
1) 37 yhteiskuntatieteilijää
- cloze-testi ja 2 mv
2) 60 humanistia
- cloze-testi ja 2 mv

Jyväskylä 1988
1) 210 kasvatustieteilijää
- cloze-testi ja 4 mv
2) 137 humanistia
- täydennettävä tiivistelmä ja 3 mv

TESTITYYPPIEN VAIKEUSEROT

Yhtenä tutkimuksen tavoitteena oli saada selville ovatko jotkin testitekniikat luonnostaan muita helpompia tai vaikeampia. Yksittäisiä testejä vertailtaessa on tietenkin aina vaarana se, että toinen niistä on muutenkin helpompi esimerkiksi tekstin rakenteen ja kieliasun tai kysymysten perusteella. Kun verrataan kymmeniä ja taas kymmeniä eri tekstien pohjalta tehtyjä testejä, voitaneen kuitenkin tehdä joitain päätelmiä testitekniikan vaikutuksesta kokeen vaikeuteen. Eri testityyppien vertailua voi vaikeuttaa hieman niiden pisteytyksen erilaisuus. Monivalintatyyppiset kokeet arvioidaan aina samalla tavalla, mutta tiivistelmien ja tekstin rakenteen hahmottamista mittaavien kokeiden arviointi on jonkin verran subjektiivista.

Liitteen 2 ja taulukko 2:n perusteella näyttäisi siltä, että monivalintatestit ovat helpompia kuin clozet tai tiivistelmät: monivalintojen ratkaisuprosentit ovat usein 70:n tietämillä, muiden kokeiden taas n. 60%. Vaikka tutkimuksessa mukana olleet cloze-testitkin ovat monivalintatyyppisiä, ne eivät olleet yhtä helppoja kuin useimmat monivalintatestit. Jyväskylän kokeilussa mitattiin T-testillä ovatko vaikeuserot monivalintojen ja muiden kokeiden välillä tilastollisesti merkitseviä. Taulukossa 2 näkyvät erot osoittautuivatkin merkitseviksi: cloze oli selvästi vaikeampi kuin

monivalinnat (t-arvo 8.05, P=.000), samoin täydennettävä tiivistelmä (t-arvo 7.35, P=.000). Kaikenlaisten monivalintakokeiden suhteellinen helppous selittynee suurelta osin sillä, että ne ovat jo teknistä syistä helpompia kuin monet muut kokeet. Pelkästään arvaamalla on 25%:n mahdollisuus saada kukin osio oikein. Mahdollisesti myös monivalinnan tuttuus koemenetelmänä auttaa selviytymään siitä suhteellisen helposti. Mielenkiintoista tässä vertailussa ei niinkään ole monivalinnan helppous, mikä oli odotettua, vaan monivalintaclozen vaikeus verrattuna useimpiin tavallisiin monivalintakokeisiin.

Taulukko 2: JYVÄSKYLÄN KOKEILUN TULOKSIA

Ala	Koetyyppi	Ratkaisu %	Keskihaj.	Min.	Max.
Hum.	tiivistelmä	60.3	23.6	10	100
Kasv.	cloze (moniv.)	61.6	13.9	17	89
Hum.	monivalinnat	73.0	12.2	30	93
Kasv.	monivalinnat	68.9	14.1	20	95

Humanisteja oli kokeessa 136, kasvatustieteilijöitä 210.

KOKEIDEN RELIABILITEETTI

Reliabiliteettia eli luotettavuutta tutkittiin vain testeissä, joissa se voidaan laskea suoraan koepisteistä. Tällaisia ovat perinteiset monivalintatestit ja monivalintaclozet; tiivistelmien ja muiden muodoltaan avoimempien kokeiden reliabiliteettia ei ole vielä tutkittu, vaikka se periaatteessa on mahdollista esimerkiksi vertaamalla kahden eri arvioijan pisteytyksiä toisiinsa.

Reliabiliteetin laskemiseen käytettiin ns. Cronbachin alfaa, toiselta nimeltään Kuder-Richardson 20, joka on yksi kokeen sisäisen yhdenmukaisuuden testeistä (vrt. Hatch ja Farhady 1982, 247). Menetelmää käytetään, kun halutaan saada selville mittaavatko testin eri osiot kaikki samaa ominaisuutta tai taitoa. Analyysi ei kuitenkaan selvitä, mikä taito on kyseessä; sitä yritetään selvittää kokeen validiteetin tutkimuksessa. Cronbachin alfaa käytetään sellaisten kokeiden analysointiin, jotka koostuvat useasta erillisestä **osiosta**. Osio tarkoittaa tavallisen monivalintakokeen yhteydessä yhtä kysymystä vastausvaihtoehtoineen, cloze-testissä taas yhden osion muodostaa yksi aukko (mahdollisine vastausvaihtoehtoineen).

Suurin yksittäinen kokeen reliabiliteettiin vaikuttava tekijä on osioiden lukumäärä: paljon osioita sisältävä testi on todennäköisemmin luotettavampi kuin lyhyempi testi. Myös kokeeseen osallistujien lukumäärä vaikuttaa saatuun reliabiliteettiarvoon; kovin pienestä testattavien joukosta saatuihin tuloksiin täytyy suhtautua varovasti. Mikäli koe on erittäin helppo, niin että melkein kaikki saavat siitä huippupisteet, kokeen reliabiliteetti on väistämättä huono.

Vanhoja monivalintatestejä käytettiin kokeiluissa yleensä kolmen kappaleen ryhminä, jolloin niiden yhteinen osiomäärä kussakin kokeessa oli 30, mitä yleensä pidetään suositeltavana määränä, jotta koe olisi kohtuullisen luotettava. Liitteestä 2 käy ilmi, että kolme monivalintaa sisältäneissä kokeissa monivalintaosan reliabiliteetti vaihteli .60 ja .90 välillä. Shohamyn (1985, 70) mukaan riittävä reliabiliteetti on yleensä vähintään .70, mutta raja-arvo riippuu testistä. Mitä tärkeämpi testitulos on sitä korkeampi reliabiliteetin tulee olla, sillä matala reliabiliteetti on yleensä osoitus kokeen epätarkkuudesta ja virheellisistä koetuloksista. Näyttää siis siltä, että aivan kaikkien monivalintatestien reliabiliteetti ei ole tarpeeksi korkea, kun niitä käytetään kolmen ryhminä; alunperinhän kokeita on käytetty ja analysoitu kuuden testin paketteina. Ilmeisesti jotkin testeistä ovat liian helppoja. Korkeakoulujen kielikeskus onkin alkanut kerätä tietopankkia yksittäisistä vanhoista monivalintakokeista, jotta liian helppojen tai muuten vanhentuneiden testien käyttöä voitaisiin välttää.

Reliabiliteetti kokeen sisäisen yhtenäisyyden ilmaisijana ei kuitenkaan ole ongelmaton käsite. Erityisen selvästi tämä tulee ilmi analysoitaessa cloze-testejä. Perusoletushan on, että kokeen eri osioiden tulisi mitata samaa taitoa, jotta koe olisi tilastollisesti mitattuna reliaabeli. Vanhoissa monivalintakokeissa tähän on ilmeisesti pyritty ja usein päästykin, sillä ne ovat yleensä hyvin reliaabeleja. Voidaan kuitenkin kysyä johtuuko joidenkin uusien kokeiden huono reliabiliteetti siitä, että eri kysymyksiin vastaamiseen vaaditaan erilaisia lukemistaitoja. Vaaditaanko esimerkiksi tekstin pääajatuksen ymmärtämisessä samaa taitoa kuin yhden tai kahden lauseen mittaisen kokonaisuuden tajuamisessa? Semanttiset monivalintaclozet on laadittu sillä periaatteella, että aukkojen täydentämiseksi olisi luettava vähintään se lause, jossa aukko sijaitsee, usein paljon enemmänkin (ks. Mauranen 1987). Testeissä ei siis pitäisi olla aukkoja, jotka vaativat vain parin sanan ymmärtämistä aukon ympäriltä. Testeissä on kuitenkin sekä sellaisia poistoja, joita varten täytyy lukea koko kappale tai enemmän että sellaisia, joista selviää lauseella parilla. Mahdoton ajatus ei siis ole, että näissäkin clozeissa eri aukkojen ratkaiseminen vaatisi erilaisia taitoja.

Liitteessä 2 olevista tuloksista käy ilmi, että 35-osioiset semanttiset monivalinta-clozet ovat yllättävän reliaabeleja jo ensimmäisellä kokeilukerralla: reliabiliteetit vaihtelevat noin .60 ja .85 välillä. Melko korkea reliabiliteetti viittaa siihen, että valtaosa cloze-aukoista mittaakin yhtä ja samaa taitoa, vaikka jotkin niistä edellyttävät laajemman kontekstin ymmärtämistä kuin toiset. Vaihtoehtoinen selitys on, että erilaiset osiot ehkä vaativat eri taitoja, mutta että taidot kehittyvät rinta rinnan, jolloin hyvin clozessa menestyvät testattavat ovat hyviä kaikissa clozeen liittyvissä taidoissa. Selitys semanttisen clozen yhtenäisyyteen voi liittyä myös testaustekniikkaan: monivalintamuoto, jossa on kyse ymmärtämisestä, "tasapäistää" osiot niin, että niissä vaaditaan kaikissa samaa taitoa. Kielen tuottamista vaativa avoin cloze saattaa saada aikaan sen, että eri tasoiset osiot vaativat eri taitoja (vrt. Huhta 1989).

VERTAILU VANHOIHIN MONIVALINTAKOKEISIIN

Osa tutkimusta oli yrittää saada selville, missä määrin uudet kokeet mittaavat samoja taitoja kuin vanhat monivalintakokeet. Korrelaatioita on toistaiseksi laskettu lähinnä vain cloze-testien ja monivalintojen välille, pääasiassa sen vuoksi, että cloze-testien tuloksia on ollut eniten saatavilla helposti tilastolliseen käsittelyyn sopivassa muodossa. Liitteessä 2 näkyvät kaikkien suurehkojen koeanalyysien tulokset, joita tähän mennessä on saatu. Liitteestä käy ilmi, että kaikissa tapauksissa clozen ja monivalinnan tulokset korreloivat toisiinsa erittäin merkitsevästi. Vain parissa kokeessa korrelaation merkitsevyys jää "vain" merkitsevälle tasolle (P=.001 tai .002). Listan oikeassa laidassa ovat kokeilutestien korrelaatikertoimet vanhoihin monivalintoihin. Suluissa oleva luku on korrelaatio, jossa kokeiden reliabiliteetin vaikutus on mukana. Matala reliabiliteetti vähentää automaattisesti korrelaation määrää, koska se kertoo kokeen epätarkkuudesta mitata aiottua taitoa. Jos koe ei mittaa tarkasti haluttua ominaisuutta, sen tulokset *eivät voi* korreloida sellaisen kokeen kanssa, joka ko. ominaisuutta mittaa. Kokeissa oleva epätarkkuus voidaan kuitenkin ottaa huomioon ja saada selville kokeiden välinen todellinen korrelaatio (vrt. Hatch ja Farhady 1982, 259). Todellisia korrelaatioita kuvaavat luvut ovat liitteessä 2 korjaamattomien korrelaatioiden alapuolella (sulkumerkkien sisällä).

Korjatut clozen ja vanhojen monivalintakokeiden korrelaatiot ovat melko korkeita vaihdellen .56 ja .90 välillä; yleisiä näyttäisivät olevan .75:n vaiheilla olevat korrelaatiot. Tulokset voidaan tulkita niin, että jos kokeiden välinen korrelaatio on esimerkiksi .80, ne mittaavat 64 % osuudelta samoja taitoja (0.80 x 0.80 = 0.64).

Korjaamattomat korrelaatiokertoimet jäävät yleensä kymmenystä tai kahta pienemmiksi.

Mukana on vain yksi tiivistelmäkoe, jonka reliabiliteettia ei ole laskettu, joten siitä on käytössä vain korjaamaton korrelaatio vertailutesteihin. Korrelaation suuruus on .52, mikä on samaa suuruusluokkaa kuin cloze-testien vastaavissa korrelaatioissa keskimäärin.

Useimmat monivalintaclozet näyttäisivät mittaavan ainakin puoleksi samoja taitoja kuin perinteiset monivalinnat. Havainto on rohkaiseva siinä mielessä, että clozetestit on ilmeisesti pystytty laatimaan sellaisiksi, että niissä menestyminen edellyttää myös korkeamman tason ymmärtämistaitojen käyttöä. Monet vanhojen monivalintatestien kysymykset kohdistuvat nimittäin tekstien keskeiseen sisältöön. Triviaalejakin kysymyksiä on vanhoissa kokeissa pakosta mukana, koska kukin teksteistä on noin 500 sanan mittainen ja jokaisesta on tehty kymmenen kysymystä. Myös täydennettävä tiivistelmä näyttää mittaavan ainakin osaksi samoja taitoja kuin vanhat testit.

Tuloksen tulkintaan on kuitenkin suhtauduttava tietyllä varovaisuudella, sillä on mahdollista, että testimetodi vaikuttaa tuloksiin. Sekä vanhat testit että uudet clozet perustuvat monivalintatekniikkaan, joten on mahdollista, että korkea korrelaatio näiden testien välillä johtuisikin monivalintatekniikan hallinnasta eikä siitä, että ne mittaisivat molemmat samoja kielitaidon piirteitä. On vaikea päätellä, miten paljon korrelaatioissa on metodin vaikutusta; todennäköistä on kuitenkin, että korrelaatiot olisivat jonkin verran matalampia, jos metodin vaikutus voitaisiin poistaa. Näyttää toisaalta siltä, että ainakin vanhat monivalintatestit on pystytty laatimaan siten, ettei pelkällä logiikalla, epäuskottavien vaihtoehtojen eliminoinnilla ja muilla monivalintatekniikan "hallintakeinoilla" pysty useimpia kysymyskohtia ratkaisemaan.

Koska uusien ja vanhojen testien korrelaatiot eivät kuitenkaan ole läheskään täydellisiä, on ilmeistä, että uudet testit mittaavat myös paljon sellaista, mitä entiset kokeet eivät tee. Onko tämä hyvä vai huono asia, sitä on vaikea yksiselitteisesti sanoa. Jos uudet taidot, joita uudet testit mittaavat katsotaan tärkeiksi, niin on vain hyvä, etteivät testit korreloi liian hyvin toisiinsa: koko kokeen mittaama lukemistaidon ala vain kasvaa uusien testien myötä. Jos taas testien ajatellaan mittaavan taitoja, joita ei ole tarpeen mitata, uudet kokeet ovat pahimmillaan vain ajanhukkaa. Tarkempien vastausten saaminen empiiristä tietä vaatii laajempia kokeiluja, joissa yhtä testityyppiä verrataan useampaan muuhun testiin, ei vain yhteen, niin kuin asianlaita on ollut tässä tutkimuksessa.

Mielenkiintoinen ja jossain määrin huolestuttava piirre on vanhojen monivalinta-kokeiden suhteellisen matalat ja vaihtelevat keskinäiset korrelaatiot. Alustavien havaintojen perusteella, joissa kuitenkin on mukana muutama kymmenen eri koetta, vanhojen kokeiden keskinäiset (korjaamattomat) korrelaatiot ovat .2 - .5 luokkaa. Kyse on toki vain 10-osioisista kokeista, joista jotkut ovat niin helppoja, ettei koepisteissä ole tarpeeksi vaihtelua - vastausten varianssihan on korrelaatioiden ilmenemisen ehdoton edellytys. Kuitenkin näyttää siltä, etteivät kaikki vanhat monivalinnat mittaa aivan samoja taitoja.

TESTAUSTEKNIIKAN TUTTUUS JA TESTIN USKOTTAVUUS

Tavallinen monivalintatesti on selvästi kaikkein tutuin kaikista kokeiluissa mukana olleista testityypeistä, mikä ei ole mikään ihme, sillä niin paljon sitä on käytetty koulussa ja ylioppilaskokeissa. Taulukosta 4 käy ilmi, että useimmissa kokeiluissa opiskelijat arvioivat monivalinnan tuttuudeksi 4 tai enemmän asteikolla 1 - 5 arvioitaessa. Kaikkien muiden testityyppien tuttuus jää yleensä alle kahden. Jyväskylän kokeilussa laskettiin merkitsevyystestit arvioissa havaituille eroille (monivalinta v. tiivistelmä ja monivalinta v. cloze). Erot osoittautuivat tilastollisesti erittäin merkitseviksi, vaikka ko. tapauksessa cloze-testi olikin normaalia tutumpia opiskelijajoukolle (taulukko 3). Lisäksi monivalintojen tuttuusarvioiden keskihajonta on pienempi kuin muiden testien, mikä osoittaa, että ne ovat erittäin tuttuja melkein kaikille opiskelijoille, kun taas muissa testeissä arviot hajoavat selvästi: monille opiskelijoille testityypit ovat täysin outoja, mutta on myös paljon niitä, jotka ovat sellaisia ennenkin kohdanneet.

Taulukko 3: JYVÄSKYLÄN KOKEILUKOKEIDEN TUTTUUS JA HYVYYS LUETUN YMMÄRTÄMISEN TESTEINÄ OPISKELIJOIDEN ARVIOIMANA: EROT KOKEIDEN VÄLILLÄ

Koe:	Moni-valinta	Tiivis-telmä	Cloze	Ero mv-testeihin		N
Humanistit:						
Hyvyys	3.5	3.0	-	t=4.7	P=.000	121
Tuttuus:	4.7	1.9	-	t=19.6	P=.000	112
Kasvatust.:						
Hyvyys:	3.5	-	3.2	t=4.2	P=.000	183
Tuttuus:	4.7	-	3.9	t=8.8	P=.000	166

t = t-arvo
P = havainnon merkitsevyys

Testityypin tuttuudella/outoudella ei yleensä näyttänyt olevan yhteyttä siinä menestymiseen. Vain pienemmässä Tampereen kokeilussa cloze-testillä ja toisella monivalinnalla oli yhteys niissä menestymiseen (cloze korr. .47, P=.008; monival. korr. .58, P=.001). Opiskelijoita vain oli melko vähän (26), mikä vähentää tuloksen luotettavuutta. Havainto, että testimuodon tuttuus ei näyttäisi auttavan siinä menestymistä on mielenkiintoinen, sillä voisi olettaa päinvastaista.

Testin uskottavuus eli hyvyys (face validity) ja testissä menestyminen eivät näyttäneet olevan yhteydessä toisiinsa. Jälleen Tampereen pienempi kokeilu muodostaa poikkeuksen: yhteiskuntatieteilijöillä clozen (.42, P=.015) ja yhden monivalinnan (.51, P=.002) hyvyys korreloi koepistemäärään. Kun tarkasteltiin heidän suhtautumistaan koko testiin (kaikkien osatestien hyvyysarvioiden summa) ja kokonaismenestystä, havaittiin, että ne, jotka pitivät koetta hyvänä luetun ymmärtämisen mittarina myös menestyivät siinä paremmin (.42, P=.008). Näyttää siis siltä, että useimmiten kokeen tuttuus ja uskottavuus eivät vaikuta menestymiseen, mutta joillakin ryhmillä niin voi käydä.

Testien välillä oli kuitenkin merkittäviä eroja siinä, miten hyvinä kielitaidon mittareina niitä pidettiin. Tavallinen monivalinta katsottiin paremmaksi testaustavaksi kuin cloze-testi tai täydennettävä tiivistelmä. Kummassakin Jyväskylän kokeilussa monivalinnan "hyvyydeksi" asteikolla 1 - 5 arvioituna tuli 3.5, tiivistelmän jäädessä 3.0 ja clozen 3.2:een. Erot arvioissa olivat tilastollisesti erittäin merkitseviä, kuten taulukosta 3 näkyy. Koska cloze ja tiivistelmä olivat eri testeissä, ei niiden arvostuksessa havaittua pientä eroa voitu todentaa tilastollisesti.

Muiden kuin monivalintatestien uskottavuus näyttää olevan "keskimääräistä" luokkaa eli arviot sijoittuvat usein arviointiasteikon puoliväliin (vrt. taulukko 4). Tämä saattaa osoittaa, etteivät opiskelijat ole oikein varmoja siitä, mitä testit mittaavat ja siksi sijoittavat arvionsa asteikon keskelle "en tiedä"-alueelle. Luetun ymmärtämisessä on ilmeisesti vaikeaa ilman tarkempaa perehtymistä kielitaidon mittaukseen arvioida testien hyvyyttä, mikä ehkä selittää sen, etteivät arviot tässä tutkimuksessa näyttäneet korreloivan testimenestykseen. "Luonnonmukaista", autenttista luetun ymmärtämisen testiä ei ehkä voidakaan laatia samalla tavalla kuin esimerkiksi suullisen kielitaidon arvioinnissa voidaan käyttää performanssitestejä, joissa testattavan on suoriuduttava tehtävistä, joihin hän todellisuudessakin joutuu. Esimerkiksi Shohamy (1982) havaitsi, että haastattelu oli uskottava suullisen kielitaidon testi kaikkien testattaviensa mielestä, mutta clozeen puhumistaidon testinä uskoivat vain ne, jotka menestyivät siinä hyvin. Tässä tutkimuksessa tehty

vertailu testin uskottavuuden ja testimenestyksen välillä sai osaksi kimmokkeen Shohamyn edellä mainitusta havainnosta, että testiin suhtautuminen vaikuttaa siinä menestymiseen. Aiemmin mainitusta syystä johtuen tuntuu epätodennäköiseltä, että samanlaista systemaattista yhteyttä voisi esiintyä luetun ymmärtämisen puolella.

Taulukko 4: YHTEENVETOA VUOTEEN 1989 MENNESSÄ TUTKITUISTA KOKEISTA

TESTITYYPPI:	Tuttuus vaiht.väli	Hyvyys vaiht.väli	Ratkaisuprosentti vaiht.väli
avoimet kysym.	**2.6** 2.2-3.1	**3.5** 2.9-3.9	**80%** 71-85%
cloze	**2.4** 2.3-2.5	**2.7** 2.6-2.9	**63%** 45-80%
monivalinta	**4.4** 3.6-4.8	**3.4** 3.0-3.7	**79%** 63-98%
tekstin rakenne-kaavion täydenn.	**2.7** 2.3-2.9	**3.6** 3.5-3.7	**73%** 70-75%
tiivistelmä	**2.8** 2.4-3.6	**3.0** 2.7-3.3	**77%** 75-79%
täydennettävä tiivistelmä	**2.6** 2.0-3.3	**3.0** 2.5-4.0	**73%** 60-87%

Taulukossa 4 "tuttuus" ja "hyvyys" arvioitiin asteikolla 1-5 (vrt. Liite 1)
Lihavoitu luku on tyypillisin arvo, "vaiht.väli" (vaihteluväli) lihavoidun luvun alla osoittaa, miten paljon eri koekerroilla saadut tulokset vaihtelivat ko. testityypin osalta. Vain clozessa ja monivalinnoissa luvut ovat jokseenkin luotettavia, koska ne perustuvat kymmeniin eri testeihin; muita koetyyppejä on kokeiltu kutakin vain 3-5 eri kielikokeessa.

TEKSTIN KIINNOSTAVUUS JA VAIKEUSTASO

Seuraavaksi tarkastellaan testin pohjatekstin vaikutusta testissä onnistumiseen. Opiskelijoilta kysyttiin arviota tekstin kiinnostavuudesta ja vaikeudesta (ks. liite 1) käyttäen arviointiasteikkoa 1 - 5. Jyväskylän kokeilun tulosten valossa näyttää siltä, ettei niissä käytettyjen clozen ja tiivistelmän pohjatekstien vaikeus tai kiinnostavuus olleet yhteydessä testeissä menestymiseen. Yksi poikkeus kuitenkin oli: clozessa

tekstin vaikeus oli yhteydessä menestykseen: ne, jotka arvioivat tekstin helpoksi, menestyivät siinä yleensä keskimääräistä paremmin. Yhteys oli kuitenkin melko heikko (r=-15, P=.018). Cloze-testissä tulosten tulkintaa vaikeuttaa se, että aukotus tekee väistämättä tekstistä vaikeamman, joten opiskelijoiden arvio cloze-*testin* vaikeudesta vaikuttanee heidän arvioonsa *tekstin* vaikeudesta.

Useissa monivalinnoissa sen sijaan löytyi yhteys koetekstin vaikeuden ja kokeessa menestymisen välillä. Kaikilla kasvatustieteen monivalinnoilla oli samanlainen yhteys tekstin vaikeuteen kuin clozessa: mitä paremmin kokeessa menestyttiin, sitä helpompana tekstejä pidettiin tai kääntäen: mitä helpompi teksti, sitä parempi koemenestys. Korrelaation suuruus vaihteli .20 ja .25 välillä (P=.006-.000). Yleensä nämä tekstit arvioitiin hiukan keskimääräistä vaikeammiksi (3.4 - 3.7), ja testien ratkaisuprosentit vaihtelivat 65:stä 71:een.

Vain yhdessä humanistien monivalintatestissä tekstin vaikeudella ja testissä menestymisellä oli yhteyttä toisiinsa. Tässäkin tapauksessa tekstiä helppona pitäneet saivat parempia pistemääriä kuin muut.

Pohjatekstin kiinnostavuus näyttää kuitenkin olevan yhteydessä siihen, miten hyvänä luetun ymmärtämisen mittarina testiä pidetään. Näin oli laita ainakin Tampereen humanistien alkukokeen useimmissa testeissä, samoin lähes kaikissa Jyväskylän kokeilujen testeissä. Korrelaatiot olivat melko lieviä, .15 - .30 -luokkaa.

Löydetyt korrelaatiot eivät kuitenkaan kerro, kumpi on syy ja kumpi seuraus; luontevinta on kuitenkin ehkä ajatella, että tekstin kiinnostavuus parantaa testinkin uskottavuutta. Kyse on tässä ilmeisesti LSP-testauksen peruskysymyksestä: oikeudenmukaisinta on käyttää kielitaidon mittauksessa aihepiiriltään tuttuja tekstejä. Testattavat tietävät yleensä, että kielikokeen on määrä mitata heidän oman alansa tekstien ymmärtämistä. Tällöin he pitävät vain sellaisia testejä uskottavina, jotka perustuvat heidän oman alansa teksteihin. Oman alan tekstit taas lienevät useimmiten kiinnostavia.

Toisaalta on varmasti niitäkin, joille on vakiintunut tietty käsitys eri testimuotojen hyvyydestä jo kouluaikana ja jota käsitystä ei tekstin kiinnostavuus muuta. Tässä tutkimuksessa kuitenkin tekstin kiinnostavuus ja arvio testin hyvyydestä olivat yhteydessä toisiinsa.

Jos tekstin mielenkiintoisuus lisää luottamusta testiin - ja tämä toisinaan parantaa siinä menestymistä - tekstien valintaan on kiinnitettävä paljon huomiota. Tässä törmätään tietenkin vakaviin käytännön ongelmiin: samaan kokeeseen osallistuu monesti usean eri aineen opiskelijoita, joten kaikki tekstit eivät voi olla juuri kunkin omalta alalta. Lisäksi on varmasti vaikea valita aiheeltaan kaikkia kiinnostavia tekstejä, vaikka kaikki kokeeseen osallistujat olisivat saman aineen opiskelijoita.

Tekstin vaikeuden ja kiinnostavuuden välillä havaittiin noin puolessa testeissä sellainen riippuvuus, että helppoa tekstiä pidettiin kiinnostavana ja vaikeaa taas vähemmän kiinnostavana. Korrelaatiot olivat tässäkin melko pieniä, korkeintaan .3 tasoa. Havainto näyttäisi tukevan ajatusta, että tekstiä pidetään mielenkiintoisena, koska se käsittelee tuttua aihepiiriä. Alan tuntemus taas voi auttaa vastaamaan paremmin kysymyksiin. Lukija myös todennäköisesti hallitsee paremmin aiheeseen liittyvän termistön kuin alaa tuntematon. Aihepiirin tuntemuksen on todettu merkitsevästi vaikuttavan menestymiseen nimenomaan cloze-testeissä (ks. Alderson ja Urquhart 1985).

OMA ARVIO KIELITAIDOSTA

Opiskelijoilla näyttää olevan jonkinlainen käsitys omasta luetun ymmärtämisen taidostaan, ainakin kun vertailukohtana on menestys tutkituissa alkukokeissa: oma arvio (1 - 5 -asteikolla) korreloi menestykseen lähes jokaisessa testissä sekä Tampereen että Jyväskylän kokeiluissa. Suuruusluokaltaan yhteys oli .4 - .5.

Ylioppilaskokeen englannin arvosana oli myös yhteydessä koemenestykseen: korrelaatio (.5 - .6) ilmeni lähes kaikissa kokeissa ja oli useimmiten hiukan suurempi kuin oman arvion korrelaatio koemenestykseen. Yo-kokeen arvosana korreloi myös omaan arvioon luetun ymmärtämistaidosta: Tampereella korrelaatio koko alkukokeeseen oli .44 molemmissa kokeissa, Jyväskylässä .42 ja .62. Järkevää lienee olettaa, että opiskelijoiden arvio englannin lukemistaidosta perustuu pääosin menestykseen ylioppilaskokeessa ja muissa koulun kielikokeissa.

Se, että opiskelijoiden käsitys lukutaidostaan ja heidän ylioppilastodistuksen englannin arvosanansa olivat yhteydessä kielikeskusten lukemistesteissä menestymiseen antaa aihetta uskoa, että nämä testit mittaavat ainakin osittain lukemistaitoa vieraalla kielellä. Puutteet vertailukriteereissä (oman arvion epäluotettavuus;

yo-arvosana ei ole pelkän luetun ymmärtämisen indikaattori) estävät kuitenkin arvioimasta, miten hyvin tai tarkasti testit mittaavat lukemistaitoa.

MUUT TAUSTATEKIJÄT

Opiskelijoiden ilmoittamalla opintojensa määrällä (pääaineen opintoviikkoina ja opiskeluvuotena) ei havaittu olevan yhteyttä kielikokeissa menestymiseen. Hypoteesina oli, että mitä enemmän omasta alastaan tietää sitä paremmin ymmärtää oman alansa tekstejä. Opiskelijat olivat kuitenkin valtaosin vasta ensimmäisen opiskeluvuotensa aloittaneita, mikä selittää yhteyden puuttumisen.

Ulkomailla oleskelun pituus ei yleensä ennustanut menestymistä luetun ymmärtämiskokeessa. Ainut poikkeus oli kasvatustieteen alkukoe Jyväskylässä, missä osassa kokeita löytyi noin .2 suuruinen korrelaatio.

OPISKELTAVAN ALAN VAIKUTUS KOEMENESTYKSEEN

Alderson ja Urquhart (1985) havaitsivat, että yliopisto-opiskelijoiden ala vaikutti heidän menestykseensä luetun ymmärtämistesteissä. Oli mielenkiintoista katsoa olisivatko tulokset samansuuntaisia myös suomalaisilla opiskelijoilla. On huomattava tosin, että Aldersonin ja Urquhartin opiskelijat edustivat hyvinkin erilaisia aloja kuten humanisteja ja matemaatikkoja. Tässä tutkimuksessa tarkasteltiin kahta alkukoetta Jyväskylässä syksyllä 1988, joissa opiskelijat edustivat kasvatustieteitä ja humanistisia aloja. Kumpikin ryhmä jakautuu eri koulutusohjelmiin, mutta on todennäköistä, etteivät opiskeltavat alat eroa toisistaan niin paljon kuin englantilaisessa tutkimuksessa. Lisäksi suurin osa opiskelijoista oli vasta-alkajia, joille ei vielä ole voinut kertyä kovin paljon oman alan erikoistietoa.

Kasvatustieteen opiskelijat jakautuivat seuraaviin koulutusohjelmiin: 1) OKL (102 opiskelijaa), 2) lastentarhan opettajat (34), 3) psykologit (28), 4) erityispedagogit (18) ja 5) muut (28). Humanistien koulutusalat olivat puolestaan seuraavat: 1) taideaineet (50 opiskelijaa), 2) vieraat kielet (34), 3) suomen kieli (25), 4) historia (15), 5) viestintä (9) ja 6) muut (4).

Kokeiluissa käytetyt testit olivat:

Kasvatustieteilijät
1. cloze (Learning to Talk)
2. mv 1 (The Uses of Enchantment)
3. mv 2 (Developmental Dyslexia)
4. mv 3 (Learners of a Second Language in a Formal Setting)
5. mv 4 (Compensatory Education)

Humanistit
1. täydennettävä tiivistelmä (Popular Literature)
2. mv 1 (Hitler's War)
3. mv 2 (Cultural Habits)
4. mv 3 (Stroking those Wild Beasts)

Taulukko 5: KASVATUSTIETEEN OPISKELIJOIDEN TESTITULOKSET KOULUTUSALOITTAIN

Koe: Ryhmä	Cloze	Mv 1	Mv 2	Mv 3	Mv 4	MvYht	Kaikki testit yhteensä
OKL	62	71	69	66	74	70	67
Lastent.	51	61	55	54	57	57	55
Psykol.	68	82	71	72	81	77	74
Er.pedag.	65	69	67	66	66	67	66
muut	66	72	74	69	76	73	70

Tulokset ovat prosentteja maksimituloksesta. Yhteistuloksessa clozen osuus on yksi kolmasosa koko pistemäärästä. 'Mv' on lyhennys monivalintatestistä; 'MvYht' on kaikkien neljän monivalintatestin yhteistulos.

Silmämääräisesti arvioiden lastentarhanopettajiksi opiskelevien ryhmä on kielitaidoltaan heikoin, psykologeiksi aikovat taas parhain. Tuloksille laskettiin varianssianalyysi ryhmien välisten erojen todentamiseksi. Sen mukaan lastentarhanopettajat todellakin menestyivät tilastollisesti merkitsevästi (0.050-tasolla) heikommin kuin kaikki muut ryhmät cloze-testissä ja koko kokeessa. Myös monivalintakokeissa he olivat erityispedagogeja lukuunottamatta muita ryhmiä selvästi heikompia. Ensimmäisessä monivalintakokeessa psykologit olivat merkitsevästi parempia kuin opettajankoulutuslaitoksella opiskelevat ja lastentarhanopettajat. Muut erot ryhmien välillä eivät olleet merkitseviä. Aldersonin ja Urquhartin tutkimuksessa ilmeni, että joissakin tapauksissa syynä eroon tiettyjen ryhmien testituloksissa oli todennäköisemmin ero kielitaidossa kuin tekstin sisällössä. Todennäköisesti tämä on syynä

myös lastentarhanopettajiksi opiskelevien huonompaan menestymiseen, koska tekstien aiheet eivät näyttäisi olleen heille sen vieraampia kuin muille ryhmille.

Vaikka ryhmien keskihajontoja ei tässä sen tarkemmin käsitellä, mainittakoon, että kaikissa alkukokeen testeissä lastentarhanopettajiksi opiskelevilla tulosten hajonta oli suurinta. Heidän ryhmänsä oli siis kielitaidoltaan jonkin verran heterogeenisempaa kuin muut.

Myös humanistisen tiedekunnan opiskelijoiden eri koulutusaloja verrattiin toisiinsa Jyväskylän kokeilun yhteydessä, mutta ryhmät eivät eronneet merkitsevästi toisistaan, joten ryhmäkohtaisia tuloksia ei tässä esitetä.

YHTEENVETO

Uudet englannin luetun ymmärtämistä mittaavat cloze-testit näyttävät olevan melko reliaabeleja eli niiden osiot mittaavat johdonmukaisesti samaa taitoa. Jäljellä olevat puutteet lienee helppo korjata osioanalyysien jälkeen. Muiden uusien koetyyppien kuten tiivistelmän arvioinnin luotettavuutta ei vielä ole systemaattisesti tutkittu. Vanhojen standardoitujen monivalintatestien luotettavuus ei yleensä ole ongelma, jos niitä käytetään alkuperäisinä kuuden osakokeen paketteina. Mikäli paketteja aletaan pilkkoa pienemmiksi osiksi, tarvitaan tietoa yksittäisten testien luotettavuudesta tai paremminkin vaikeustasosta; eräät testit ovat osoittautuneet liian helpoiksi, jolloin ne eivät enää voi luotettavasti erotella hyviä ja huonoja lukijoita. Korkeakoulujen kielikeskus onkin alkanut kerätä tietopankkia vanhojen monivalintakokeiden ominaisuuksista, jotta liian helppoja testejä voidaan välttää.

Testien validiteetti eli tarkkuus on hankalammin määriteltävä ominaisuus, koska ei ole täysin varmaa, mitä vertailussa mukana olevat muut kokeet mittaavat. Puhtaasti korrelaatioiden valossa näyttää siltä, että semanttiset clozet mittaavat 1/3 - 2/3 osalta samoja taitoja kuin vanhat monivalintakokeet (korrelaatiot .56 - .90). Koska molemmat kokeet perustuvat monivalintatekniikkaan, tämän testimuodon hallinta voi jonkin verran vaikuttaa tulokseen. Jos testimetodin vaikutus pystyttäisiin eliminoimaan, korrelaatiot eivät todennäköisesti olisi aivan näin korkeita. Yhden tutkitun tiivistelmäkokeen korrelaatio monivalintatesteihin oli samaa luokkaa kuin clozeilla keskimäärin. Opiskelijoiden oma arvio lukemistaidosta korreloi yleensä jonkin verran testimenestykseen, mikä antaa lievää tukea testien validiteetille.

Samoin on laita opiskelijoiden ylioppilastodistuksen englannin kielen arvosanan ja koemenestyksen suhteen.

Testien uskottavuudessa (face validity) oli suuria eroja: yleensä perinteistä monivalintatestiä pidetään parhaana kokeena, clozea taas huonoimpana. Muista koetyypeistä ei ole kovin paljon tietoa, mutta näyttää siltä, että tiivistelmät, avoimet kysymykset ja erilaiset tiivistelmätestit ovat joko yhtä suosittuja kuin monivalinta tai vain hieman vähemmän suosittuja. Testityypin uskottavuus ei kuitenkaan näyttänyt yleensä auttavan siinä menestymisessä, mikä ehkä johtuu siitä, ettei opiskelijan ole kovin helppoa arvioida, mikä on hyvä luetun ymmärtämisen testi.

Koetekniikan vaikutusta kokeen vaikeustasoon ei aina ole helppo saada selville, koska pisteytystavat saattavat erota kovasti toisistaan ja tekstin sisältö ja kielellinen vaikeus vaikuttavat todennäköisesti enemmän kuin käytetty tekniikka. Tarkasteltavana olevissa kokeissa on tarpeeksi tietoa vain perinteisistä monivalinnoista ja monivalintaclozeista. Niiden osalta näyttää siltä, että cloze on selvästi vaikeampi testimuoto kuin monivalintakysymykset.

Testaustavoista monivalinta oli ylivoimaisesti tutuin. Muut tekniikat olivat toisiinsa verrattuina yhtä tuttuja. Testausmenetelmän tuttuus ei näyttänyt kuitenkaan yleensä auttavan testissä menestymistä.

Yleensä koetekstin kiinnostavuus tai vaikeus eivät olleet yhteydessä koemenestykseen. Tekstin kiinnostavuus sen sijaan oli usein yhteydessä siihen, miten hyvänä lukemistaidon mittarina testiä pidettiin. Havaittiin myös, että tekstin kiinnostavuus ja helppous liittyivät toisiinsa.

Opiskelijoiden koulutusohjelma ei useimmiten ennustanut menestymistä testeissä eli testien/tekstien aihepiiri ei näyttänyt erityisesti suosivan mitään tiettyä ryhmää. Kasvatustieteilijöistä lastentarhanopettajat olivat kuitenkin ilmeisesti kielitaidoltaan muita heikompia, mikä näkyi koetuloksissa. Humanisteilla eroja ei havaittu. Tarkastellut koulutusohjelmat ovat niin lähellä toisiaan mitä tieteenaloihin tulee, ettei muualla (esim. Alderson ja Urquhart 1985) havaittua koulutusalan vaikutusta ilmennyt näissä testeissä.

Tässä raportoidut tulokset ovat osa koko ajan jatkuvaa kielikokeiden seurantaa, jonka tarkoituksena on parantaa käytettyjen kokeiden luotettavuutta, tarkkuutta ja käyttökelpoisuutta. Tavoitteena on myös mahdollisimman monipuolinen lukemis-

taidon testaus, mikä edellyttää entistä parempaa uusien testausmenetelmien tuntemusta. Tämä tutkimus on vain raapaissut koko ongelmakentän pintaa; toivottavasti se kuitenkin antaa uusia ideoita jatkotutkimuksille.

Lopuksi vielä kiitokset kokeiluja järjestäneille opettajille; ilman heidän aktiivista osallistumistaan tällaisen laajahkon, vertailevan tutkimuksen tekeminen olisi erittäin vaikeaa.

VIITTEET

ALDERSON, J. Charles 1979. The Effect on the Cloze Test of Changes in Deletion Frequency. Journal of Research in Reading, Vol.2, No.2, 108 - 119.

ALDERSON, J. Charles ja A. H. Urquhart 1985. The Effect of Students' Academic Discipline on their Performance on ESP Reading Tests. Language Testing Vol.2, No.2, 192-204.

BACHMAN, Lyle F. 1985. Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. TESOL Quarterly, Vol.19, No.3, 535 - 556.

BACHMAN, Lyle F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

HATCH, Evelyn ja Hossien Farhady 1982. Research Design and Statistics for Applied Linguistics. Rowley, Mass.: Newbury House.

HUHTA, Ari 1989. Testing Language Proficiency at University Level - Can the Cloze and C-Tests Help? Pro Gradu -tutkielma, Jyväskylän yliopisto, soveltavan kielitieteen laitos.

MAURANEN, Anna 1987. Monivalintacloze englannin tekstinymmärtämisen mittarina. Korkeakoulujen kielikeskuksen julkaisuja N:o 27.

SHOHAMY, Elana 1982. Predicting Speaking Proficiency from Cloze Tests: Theoretical and Practical Considerations for Tests Substitution. Applied Linguistics, Vol.3, No.2, 161 - 169.

WEIR, Cyril 1988. Communicative Language Testing. University of Exeter.

LIITE 1 Opiskelijakyselylomake

_____ YLIOPISTO **KIELIKESKUS**

Nimi: _____ Henkilötunnus: _____
Tiedekunta: _____
Koulutusohjelma/pääaine: _____

VASTATTUASI KOEKYSYMYKSIIN PYYDÄMME SINUA ARVIOIMAAN KOKEEN OSAT:
Suorita arviosi merkitsemällä sopivaksi katsomasi numero asteikolta.

__ **TESTI 1:** (testin 1 pohjana olevan tekstin nimi)

1.Miten hyvin uskot sen mittaavan lukemistaitoasi:	eritt.huonosti	1	2	3	4	5	eritt.hyvin
2.Tekstin vaikeus:	eritt.helppo	1	2	3	4	5	eritt.vaikea
3.Tekstin kiinnostavuus:	ei kiinnostava	1	2	3	4	5	eritt.kiinnost.

__ **TESTI 2:** (testin 2 pohjana olevan tekstin nimi)

1.Miten hyvin uskot sen mittaavan lukemistaitoasi:	eritt.huonosti	1	2	3	4	5	eritt.hyvin
2.Tekstin vaikeus:	eritt.helppo	1	2	3	4	5	eritt.vaikea
3.Tekstin kiinnostavuus:	ei kiinnostava	1	2	3	4	5	eritt.kiinnost.

__ **TESTI 3:** (testin 3 pohjana olevan tekstin nimi)

1.Miten hyvin uskot sen mittaavan lukemistaitoasi:	eritt.huonosti	1	2	3	4	5	eritt.hyvin
2.Tekstin vaikeus:	eritt.helppo	1	2	3	4	5	eritt.vaikea
3.Tekstin kiinnostavuus:	ei kiinnostava	1	2	3	4	5	eritt.kiinnost.

(... ja niin edelleen; em. kohtien lukumäärä riippuu kokeen testien lukumäärästä)

Onko testityyppi (esim. monivalinta) sinulle ennestään tuttu ?

	täysin tuntematon				usein koulussa tms käytetty
1. monivalinta	1	2	3	4	5
2. aukkotesti (cloze)	1	2	3	4	5

TOIVOISIMME SINUN EHTIVÄN VIELÄ VASTATA SEURAAVIINKIN KYSYMYKSIIN:

Minä vuonna aloitit opintosi ? v _____
Paljonko arvioisit suorittaneesi pääaineesi opintoja ? noin _____ ov

Miten hyväksi arvioisit oman englannin luetun ymmärtämistaitosi ?
erittäin huonoksi 1 2 3 4 5 erittäin hyväksi

Minkä arvosanan sait ylioppilaskokeiden englannin kielessä ?
joko a) pitkässä englannissa (A-kielenä): _____
tai b) lyhyessä englannissa: _____

Oletko oleskellut englanninkielisessä maassa ? ( ) en ole
( ) alle 1/2 vuotta ( ) 1/2 - 1 vuotta ( ) yli vuoden

Monennenko kerran osallistut englannin tekstinymmärtämiskokeeseen? __

KIITOS VAIVANNÄÖSTÄSI JA AVUSTASI TUTKIMUKSELLEMME !

Lyhenteet:
n = kokeeseen osallistujien lukumäärä
mv = vanha standardoitu monivalintatesti
P = korrelaation merkitsevyys
Hki = Helsinki
Jkl = Jyväskylä

Taulukon tulkinta: ensinnä taulukossa on kokeen/opiskelualan nimi (esim. Biologia), paikka (Hki) ja osallistujien lukumäärä (n=49). Seuraavalla rivillä on kokeillun uuden testin nimi (Bacteria Come...) ja tyyppi (cloze); samalla rivillä on ko. testin vaikeustaso (ratkaisu %), reliabiliteetti ja sen korrelaatio samassa kokeessa olleisiin vanhoihin monivalintatesteihin. Seuraavalla rivillä on vastaavat tiedot kokeessa käytetyistä vanhoista monivalintatesteistä (3mv = 3 monivalintatestiä à 10 kysymystä). Korrelaatiosarakkeessa ylempi arvo on korjaamaton korrelaatio; alempi, suluissa oleva arvo on korjattu niin, että reliabiliteetin vaikutus on otettu huomioon.

OPISK.ALA KOKEILUTESTIN NIMI MV-TESTIEN LUKUM.	TYYPPI	RATKAISU %	RELIAB.	KORRELAATIO MONIVALINTOIHIN yllä korjaamaton (alla korjattu)
Syksy 1989				
Biologia, Hki (n=49)				
Bacteria Come to the Aid of Wounded Plants	cloze	72%	.76	.40 (P=.002)
3 mv	mv	86/69/68%	.67	(.56)
Elintarvikeala, Hki (n=97)				
Recent Developments in Flavour Research	cloze30	52%	.69	.43 (P=.000)
3 mv	mv	78/72/60%	.80	(.58)
Farmasia, Hki (n=91)				
Pharmacotherapy of Depression	cloze	40%	.58	.32 (P=.001)
1 mv	mv	76%	.44	(.63)
Historia+taid., Hki (n=95)				
Language, Society and Culture	cloze	65%	.85	.59 (P=.000)
3 mv	mv	65/74/74	.85	(.69)
Kasvatustiede, Hki (n=158)				
Writing as Social Action	cloze	51%	.65	.60 (P=.000)
3 mv	mv	69/46/72%	.73	(.87)
Kemia, Hki (n=51)				
How Big can Aromatic Compounds Grow	cloze	72%	.72	.46 (P=.000)
3 mv	mv	66/75/75%	.90	(.57)
Kieliaineet, Hki (n=119)				
Language, Society and Culture	cloze	67%	.82	.55 (P=.000)
3 mv	mv	80/72/66%	.82	(.67)
Luonnontieteet, Jkl (n=200)				
6 mv	mv	68%	.89	-

Maatalous, Hki (n=103)

Bacteria Come to the Aid of Wounded Plants	cloze	65%	.80	.65 (P=.000)
3 mv	mv	84/67/57%	.83	(.80)

Metsäala, Hki (n=93)

Battle of the Bog	cloze	68%	.84	.52 (P=.000)
3 mv	mv	87/76/64%	.78	(.64)

Oikeustiede, Hki (n=180)

Sources of Knowledge about Deviance	cloze	58%	.82	.73 (P=.000)
3 mv	mv	80/83/74%	.80	(.90)

Taideteoll., Hki (n=87)

Henry Moore	cloze	74%	.61	.57 (P=.000)
3 mv	mv	70/82/69%	.70	(.87)

Tietojenkäs., Hki (n=67)

Computer Software for Intelligent Systems	cloze	68%	.84	.53 (P=.000)
3 mv	mv	87/77/87%	.60	(.75)

Teologia, Hki (n=109)

The Prophet	cloze(50os)	65%	.85	.52 (P=.000)
1 mv	mv	67%	.56	(.75)

Valtiotieteet, Hki (n=184)

Knowledge as a Commodity	cloze	66%	.83	.62 (P=.000)
3 mv	mv	90/72/75%	.81	(.76)

Yhteiskuntatieteet, Jkl (n=149)

6 mv	mv	75%	.83	

Syksy 1988

Biologia (n=106)

Acid Rain	cloze	75%	.78	.61 (P=.000)
3 mv	mv	73/83/70%	.74	(.80)

Hist.kielitiede, Hki (n=127)

Communication	cloze50	75%	.85	.50 (P=.000)
Ahead of Us	mv	65%	.71	.61 (P=.000)
2 mv	mv	80/74%	.74	(.63 & .84)

Humanistit, Jkl (n=137)

Popular Literature	täyd.tiiv.	60%		.52 (P=.000)
3 mv	mv	81/71/66%		-

Kasvatustiede, Jkl (n=210)

Learning to Talk	cloze	62%	.73	.57 (P=.000)
4 mv	mv	71/68/65/72%	.78	(.76)

Oikeustiede, Hki (n=176)

Courts' Resistance to Change	cloze	67%	.84	.65 (P=.000) (.77)
3 mv	mv	79/75/74%	.85	

Taideteoll., Hki (n=112)

Ahead of Us	mv	56%	.62	.53 (P=.000)
3 mv (Hki omia?)	mv	66/64/50%		-

SUMMARY IN ENGLISH

The study explored new testing techniques used to assess students' reading comprehension at language centres of Finnish universities. New tests were compared with old standardized multiple-choice (mc) tests. The new multiple-choice cloze tests turned out to be as reliable as ordinary mc tests: reliability coefficients (Cronbach's alpha) for the 35-item cloze tests varied from .60 to .85 before any improvement of the items. A validity study showed that the correlations between the clozes and the old mc tests were fairly high (.56 - .90) indicating a substantial (1/3 - 2/3) overlap in skills measured by the two kinds of tests. The fact that both test types employ the multiple-choice technique may, however, explain part of the correlation. The students' estimates of their reading ability and their English grades on the Finnish matriculation examination correlated only moderately with test scores (.4 - .6), giving only some support to the hypothesis that the new clozes measure reading comprehension. The above statistics were computed from dozens of tests taken by a total of about 2500 students (for details, see Appendix 2).

The face validity of the techniques varied: the traditional - and the most familiar - mc tests were considered the best by the students, whereas the cloze was the poorest. Tasks based on summarization or open questions were considered as good as the multiple-choice tests. The face validity estimates did not have any association with success in the tests, probably because it is not easy to assess the validity of a reading comprehension test.

The cloze technique seemed to make the tests clearly more difficult than the old mc tests, although both use the multiple-choice technique.

The effect on the test scores of the familiarity of the test technique, and the difficulty of, and amount of interest in, the text (students's estimates) was also explored. The familiarity of the technique seemed to have no effect on the score, neither did the two other variables. It was found out, however, that if the text was considered interesting, the test as a whole was often considered a good test of reading. Also, if the text was considered interesting, it was likely to be regarded as a relatively easy text, too. This probably implies that the face validity of a test depends on how interesting the text is.

There seemed to be no systematic difference between students of various subjects within the humanistic (137 students) and educational (210) faculties. The only exception were the students who are to become nursery school teachers; they showed consistently lower reading proficiency than other students in the educational faculty. The reason for the general lack of differencies may be that the student's fields were not so far apart from each other that field specific knowledge could have affected the test results. In addition, most students were beginners without much of the specific knowledge typical of more advanced students.

Margaretha Corell
Gunilla Gentzel

THE SWEDISH "RIKSTEST"

BACKGROUND

In 1982 a new type of final examination in Swedish for Academic Purposes was introduced. It is called the RIKSTEST ("the NATIONAL TEST") as it is used throughout Sweden at six universities from Umeå to Lund. Appendix 1

The effect of this test has been far greater than we anticipated and so has the work involved. Foreign students passing the rikstest receive the certificate in the Swedish language which certifies their fulfilment of general requirements for university level studies.

IES

The Institute for English Speaking Students at the University of Stockholm was founded in 1947 with the help of Marshall aid money for the benefit of US war veterans. The IES consists of two sections: Swedish Language Courses (SLC) and The International Graduate School (IGS) for postgraduate studies in English in a number of subjects specifically to do with Sweden such as: The Welfare State, The Social Security System etc.

SAP - PREPARATORY COURSES

(Swedish for Academic Purposes) Aim and syllabus, see Appendix 2

Since 1968, when the Swedish Government, through the Ministry of Education, granted money to teach Swedish for Academic Purposes at the largest universities in Sweden, approx. 50% of the total number of classes have been arranged by the IES at the University of Stockholm. The remaining 50% of the courses have been held at the Universities of Lund, Gothenburg, Uppsala, Umeå and Linköping, in total six places.

Towards the end of the 70s feelings among the university preparatory course organizers were running high as we found that foreign students who had failed the exams in one place, travelled all over Sweden to take the final Swedish examination where it was easiest to pass. In those days, the number of subtests in the finals varied from 3 in one place to as many as 12 in another.

The situation became intolerable. There was a huge increase in numbers of students coming from countries with different school-systems, cultures and languages. More and more of our students spoke little or no English (or German or French). Moreover, the majority of the students now wanted to study technology, natural sciences, computer science, medicine, dentistry and economy.

THE AUTHORS OF THE RIKSTEST

In order to do something about this unfortunate situation a meeting was arranged where each university had at least one representative and thus the **rikstest-group** was formed. Anyone interested in the particulars can read a report (in Swedish) from 1983, where the background and growth of the rikstest is shown. (Corell, Nissen: Gemensamma prov i svenska för utländska studerande. Rapport om projektet Rikstest 81 - 83.) The actual tests have been produced within the group. The tendency today is that some members of the group have become more and more specialized at writing the test, while the others act as consultants, giving their opinions, checking for ambiguities and errors.

AIM OF THE RIKSTEST GROUP

During the initial discussions we agreed on five main points

- the students should be examined as fairly and objectively as possible
- we wanted a test in general proficiency not an achievement test i.e. related to goal not to a particular course
- the test should cover "the four skills" i.e. listening, speaking, reading and writing
- all the subtests should be based on vocabulary and coherence (**not** grammatical form)
- the test should be as easy as possible to mark and assess. There should be a balance between active (difficult to mark) proficiency and passive (easy to mark) proficiency.

VALIDITY - RELIABILITY - PRACTICABILITY

LANGUAGE NORM

One of the most difficult problems to solve was, and is, the question of **what kind of Swedish** is acceptable to other Swedes (i.e. other than language teachers). To what extent is the oral/written message transmitted/misunderstood/lost because of variations of pronunciation, form and syntax errors and lack of exact vocabulary? Are there great differences between the kind of academic Swedish we use as language teachers at an average age of 50 and the written and spoken Swedish of 20 year old students?

TEST CONTENT

In 1980 the prevalent fashion in language testing was in transition from a more direct to a more indirect method. Our students need

passively

- to be able to listen to lectures, seminars, discussions
- to be able to read **quickly** and summarize
- to have an extensive knowledge of vocabulary in context

actively

- to be able to write reports, speeches and answers to essay questions
- to be able to talk, present papers, take part in discussions at seminars and in social life.

VALIDITY - BEFORE

In our attempts to construct as valid a test as possible we learned among other things from **student advisers** and **university teachers** that the greatest problems lay in the foreign students' ability to express themselves in speaking and writing. They emphasized that vocabulary difficulties were NOT subject specific but rather of a general character.

Our **former foreign students** helped us by pointing to the fact that there was a gap between our language courses and the actual demands awaiting them in their subject areas. We studied a number of **comparable foreign tests** on the basis of

which we constructed a prototype. However, we rejected one part after another of this prototype, eventually finding that we had to design each subtest exclusively to meet our specific needs.

Consequently we know of no other second language test today that looks quite like the rikstest.

VALIDITY - DURING

In order to check whether the level of difficulty of language, content and format was acceptable to Swedes we tried out the rikstest on 200 secondary school students. As it turned out, however, all the students did not take it under the same conditions and there was number of students with Swedish names who turned out to be immigrants and vice versa.

In another attempt to validate the test 64 foreign students of social sciences who had acquired a minimum of 10 points at the University of Stockholm, were invited to take the test. Five turned up.

We also asked the SAP teachers to give preliminary grades before their students took the rikstest and correlated them with the test results.

At the end of the two year trial period, questionnaires were sent to the teachers of the preparatory courses and to 338 students who had recently taken the rikstest. 22 teachers and 103 students returned the questionnaires. Their answers are accounted for in the 1983 report.

RELIABILITY

Much time and effort goes into the actual construction of the three subtests with MCQs (Multiple Choice Questions). There are many steps:

1. construction (two teachers work together)
2. test sent out to advisory consultants (= other teachers)
3. editing
4. pre-testing in the class-room (min. 50 preferably > 100 students)
5. editing ("dead meat" cut out, difficulties erased or smoothed out)

6. trial-rikstest
7. item-analysis (computer) (200 students)
8. editing

To increase the reliability of the written and spoken subtest, score sheets are used to facilitate discussion between examiners/teachers. (There is a gradual change away from these after seven years of usage.)

WHERE AND WHEN CAN YOU TAKE THE RIKSTEST?

One can take the rikstest either by enrolling on a preparatory course, where the rikstest is the final examination, or if one's knowledge of Swedish is sufficient, just by taking the test at a cost of 550:- (1989).

The third possibility came in on 1st July 1988 when the Swedish Institute started training a couple of their Swedish lecturers abroad to administer the test. However all material, including the taped oral, is sent to the IES for grading. The IES also issues the certificate.

The rikstest takes place simultaneously at the six universities mentioned earlier, eight times a year. Step 1 consists of the three passive parts with MCQ i.e. Reading, Aural, Vocabulary/link words. A minimum of 65 of 90 possible points must be achieved in order to proceed to step 2 which consists of writing and oral proficiency. Following failure in one or both of these, one can return the next time and do that part/those parts again.

There are eight parallel sets of Reading/Aural/Vocabulary/link words i.e. they are individually interchangeable.

For the written/oral parts there are number of subjects which have been collected and added to over the years. Two or three are used at a time covering the main areas of future studies. (Technology/Economy/Computer Science/Medicine/Arts)

CONSTRUCTION OF NEW SUBTESTS

At present there is a renewal of the different passive parts at the rate of 1 Aural/ 1 Reading/1 Voc/Link per year.

In terms of effective time we have found that it takes **two** teachers working full time for **two weeks** to produce **one** test. In reality, of course, one has to work one's way through pretesting - analysis - editing etc a couple of times and that takes approximately one year.

RESULT

The test is graded Pass or Fail. If the students receives a Pass a certificate is issued, if he fails, he receives a letter giving the specific results of each subtest and information about the time and place of the next test.

CLASSIFIED AS SECRET

The rikstest has been scrutinized by the legal expert at the National Board of Universities and Colleges. (Universitets- och högskoleämbetet). As the test is regarded as an independent, scientifically developed product we have been allowed to classify it as secret as long as it is used for examination.

THE NUMBER OF FOREIGN STUDENTS TAKING THE RIKSTEST

During the introductory year in 1982, around 1 000 students took the test. Since then there has been a gradual increase each year and during 1988 more than 3 000 students sat the examination.

TESTS IN ENGLISH FOR ACADEMIC PURPOSES

In March 1987 we had the opportunity to visit Britain to establish contact with test constructors and have a close look at the most recent tests in English for Academic Purposes (EAP).

To sum up we found that on the whole the British EAP-tests are more exacting than the rikstest in so much as:

- students are only allowed to listen **once** in the aural part
- students are **pressed for time** throughout the tests

- the British tests are divided into many more short subtests
- no dictionaries are allowed at any stage
- the tests are divided into two parts. One General and one Subject Specific
- the result is given in the form of **bands** (from non user to expert user).

THE DIFFERENT SUBTESTS OF THE RIKSTEST

We are now going to describe the different parts of the test. We also want to comment on certain aspects of the test we feel could be improved, as well as describing some of the changes we are undertaking at present, since we think that tests should be developed and revised continuously.

Step 1
LISTENING COMPREHENSION

This is how we compile a listening comprehension test:
We invite a person to our recording studio at the Stockholm University and talk to him to her for about an hour. Among our guests have been an official from the National Red Cross, a member of the Swedish Parliament, a naval officer and so on. When we interview someone we do not try to adapt the level of the conversation to suit the students who will later listen to the tape. Our interviewee speaks freely about his or her special field in a way which is normal in spontaneous speech, that is with repetitions, unfinished sentences and the syntactical constructions typical of spoken language. After this we edit the interview and cut it down to approximately 15 minutes, and thus have a manuscript on the basis of which we can start constructing 20 multiple choice questions.

Here lies a difficulty, however. The recorded material must be edited in such a way that the content of the interview is absolutely clear and unambiguous, otherwise it is almost impossible to design accurate questions with plausible distractors.

We have found it easy, but not very valid to ask for details, such as figures and posers (kuggfrågor). It is much more relevant however to investigate the students' general understanding and therefore we place emphasis on testing their comprehension of the gist of a whole paragraph. If anyone is especially interested

in the background of the test design, we refer them to a report written after the item analysis of the listening comprehension tests in 1987. (See Corell, Gentzel: Itemanalys av rikstest, 1985-86)

The student's task is to show that they have correctly understood and to have marked one of three alternatives. One can always question whether the multiple choice format is a valid way of testing. When, in real life, are we forced to choose between one correct and two incorrect alternatives? Furthermore the students might have understood the meaning of the interview but failed to mark the correct answer because they didn't completely comprehend the vocabulary in the alternatives. To do well in our listening comprehension test one must have certain reading skills and be able to read sufficiently fast so as to finish within the time allowed.

The results of the item analysis show, however, that the accuracy of the listening comprehension test is satisfactory and this part of the test therefore remains unchanged for the time being because of the wash back effect on the class-room teaching. We find it necessary to keep this test although we are very much aware of the fact, that it is not possible to isolate and test one faculty at a time.

READING COMPREHENSION

It is easier to construct questions which test the student's general grasp of something written, because from the start one can choose a text which is lucid and logical. In the reading test we measure the reading speed and vocabulary of the students. If they take too much time simply reading through the text, they won't have time to answer the questions. However, this subtest is less of a strain for the students since the written material is in front of them all the time and does not have to be caught in mid air.

As with the listening comprehension test we regret that we have to use the multiple choice format. One would like to test reading skill more realistically. But the eternal problem then arises; if one moves away from the multiple choice format, and lets the students answer open ended questions, how can one then gain a fair and equal assessment? What is won in validity is lost in reliability. And besides a test needs to be easy and quick to mark.

VOCABULARY/LINK WORDS

This subtest is a complement to the listening and reading test. It tests vocabulary and understanding of context, not in general but in detail. It consists of a coherent text with 50 gaps. These gaps are chosen very carefully. Half of them are link words such as adverbs, conjunctions and pronouns and the other half are abstract verbs and abstract nouns. The item analysis showed that **this** subtest had the highest reliability of the three parts, due to the fact that the number of tasks are more than twice as many, 50 instead of 20 and that there are three distractors instead of two. So why only 20 questions and two distractors in the other parts? It proved to be unnecessarily difficult to construct three distractors to a paragraph in Listening and Reading. Two was feasible but the third often turned out to be silly or weird and revealed itself as an incorrect answer by being so odd. Furthermore we didn't want to prolong these tests and we have found that 20 survey-questions to a 15 minute interview is a maximum.

The item analysis showed us which items did not help in ranking or were otherwise inappropriate. In the previously mentioned report, those items kept and discarded are discussed.

Discussion at present centres on whether one should withdraw the listening and reading parts and only keep the vocabulary/link words test to use as an instrument for sorting out which students should be allowed to pass to step two, namely oral and written proficiency. The vocabulary test has high reliability and ranks very well. It is easier to construct than, for example, listening tests which demand recording, editing, question writing and so on. The problem is that the contents of the subtests become known in one way or the other and test security calls for continuous renewal of the test battery. It would be easier to store a number of vocabulary tests instead of going through the tedious work of constructing the listening and reading tests.

Step 2
WRITING PROFICIENCY

The essay or composition is probably that part of the rikstest which has been most discussed through the years for two obvious reasons, validity and assessment. Essay writing is not required in our students' university courses. The assessment must be

fair. The students must receive the same mark regardless of the teacher who marks their test, or the university at which they take it and the essays must be assessed according to the same standard regardless of the students' mother tongue (or handwriting!).

How do we know what is pass and what is fail in writing? We have found no better way than to trust the intuitive feelings of experienced teachers.

From 1982 until Spring Term 1989 we used a mark sheet where four skills were assessed separately, namely 1. overall impression and content, 2. vocabulary and idiom, 3. syntax, 4. grammatical form and spelling. The marks were then converted into sums and the total of this added up to a grade.

Two teachers read an essay and without consulting each other they filled in a mark sheet and compared it afterwards. Surprisingly often they had decided on the same grade. If they did not agree, however, a third and even a fourth assessor was called in.

Once a year teachers from the six universities in Sweden, where Swedish for Academic Purposes is taught, gather for an assessment meeting. Some controversial essays are then discussed, as well as some of those which are very typical for their grades. Just like musicians in an orchestra who have to tune their instruments to the same "a" from the oboeist, we have to listen in to each other in order to avoid gradual changes in our grading. The aim of these meetings is to increase the reliability of the assessment of the essay writing.

During the first years of the rikstest, the students had to write their essay on themes such as "An important historical person of my country" or "Living in a new country" etc. The students were allowed to use a Swedish dictionary and were given a few questions and comments to inspire them, otherwise they were left to their own imagination and verbal creativity. There was a lot of subjective thinking and very little substance in those essays and as they were so difficult to compare, the assessment turned out to be unreliable. Besides this test form where students expressed their own views, was obviously unfair to those students who were not familiar with this form of essay writing.

Furthermore we didn't find this a valid way of testing writing proficiency. In academic life they are not asked to give their opinion on a lot of general matters,

but they will have to know how to write surveys, summaries and lab reports. Nowadays the essay at the IES is based on a fact sheet which gives short articles from a newspaper, tables, charts and other information. This form is still not ideal. Assessing vocabulary for instance is harder, as so much is already given in the fact sheet. Gradually, however, we have found that the quality of the written part has improved as we have grown more aware of general academic writing skills in our teaching.

As from the autumn of 1989 we have been trying out another system of assessment at the IES which seems promising so far.

All the essays are distributed among the teachers who read and correct them quickly and give a preliminary grade. Then the essays are turned over to **the Assessment Group** consisting of four teachers each of whom reads through all the essays, writes down short remarks and gives a grade with four parameters in mind: 1. Organisation, 2. Content, 3. Vocabulary, 4. Grammar.

In the following discussion the four grades are compared after which they are checked against the preliminary grade and the team decide on the final grade. This procedure has so far shown three definite advantages:

- Greater emphasis is placed on organization and content
- Each teacher's assessment is checked against **four** others, which improves both **inter-** and **intra**reliability
- The difficult task of determining what is and what is not acceptable is simplified since all the "borderline cases " are assessed together.

Moreover, the procedure provides us with a good basis for discussion of essay writing, of what is considered important/less important in this type of writing.

ORAL PROFICIENCY

The oral is an interaction between the student and the teacher. On this occasion there is a second teacher present. He or she does not take part in the conversation but functions as an active listener taking notes of that which is linguistically good or bad in the student's speech. This assessor uses a mark sheet which looks like this:

1. pronunciation / phonemes and prosody
2. grammar / form and syntax
3. vocabulary / variety and idioms

4. communication / interaction
5. content / reading material and general topics

Immediately after the oral the two teachers make their assessment and agree on a grade. Why **two** teachers?

When testing oral proficiency it is very exacting to conduct the conversation, raise the level to a maximum, simultaneously assess vocabulary and grammar, and also note if the student avoids using complex language. We feel it is a great advantage that another teacher is present, a person who can fully concentrate on this without being involved in the interaction. Using two assessors is also a means of increasing the reliability of the oral test.

In order to achieve better comparability between different orals the students are now given some reading material one hour before the exam. During this hour they study the material themselves in order to give a presentation of the content at the oral. They can use any kind of dictionary. During the oral the student is asked to make a summary, after which the conversation continues, perhaps along the lines of the text, perhaps about something entirely different. The oral takes 20 minutes.

THE PREDICTIVE VALUE OF THE RIKSTEST

Our students sometimes ask us: Will I succeed in my academic studies if I pass the rikstest? What they really want to know is the validity of the test and its predictive value. We have, so far in vain, applied for money to make some research into academic success and academic failure. We can only present some ideas that we have collected empirically.

Generally speaking, we feel we can claim that a pass obtained by a student from Asia or from Africa does not guarantee the same academic success as a pass obtained by a student from Europe, Australia or from North America, even if the students have achieved exactly the same amount of points in the subtests. We feel the reason for this is that when studying at a university you need not only a good command of the language but also a kind of cultural competence. The greater the cultural distance is between the students' home-country and Sweden, the greater difficulties that they will meet in their academic career and their chance of obtaining a degree will consequently be less. For example: An Iranian wants to study Psychology. In the first place he has never studied psychology at school, or he

has had a course which is very different from what the students of a Swedish gymnasium take. He may have no idea who Freud or Jung were and knows nothing about their influence, he has grown up in a country where family ties and relations between men and women are very different from ours and where mental problems express themselves in another way than we are used to. He knows nothing about and does not read the family pages of our daily papers, where he would frequently find articles about problems of everyday neuroses. In short, he is in a very disadvantageous position, by comparison with other foreign students from, say, Great Britain, Poland or Finland, and this is not due to language difficulties. Suppose the Iranian fails his exam in psychology. It is meaningless to recommend him a more advanced course in Swedish or to demand a higher level of the rikstest. For his part, if he wants to continue studying Social Sciences it is essential for him to learn to interpret and understand Swedish values, to open himself up to his new country's traditions, cultural habits and ways of looking at life and the world. To succeed in doing this, our Iranian will need great personal maturity and a long time in the country, perhaps 10 - 15 years. This process of adaptation cannot be taught, only learned.

You might find this example a little extreme but we have found that the majority of Asians and Africans face greater difficulties than other students, and therefore we try to teach something that might be called cultural understanding/awareness. We are at this moment planning to add this to the syllabus as a separate subject. Our view is that a student need not change his (the majority of our African and Asian students are male) cultural values and become "a Swede", but must be able to recognise what his own culture stands for and what Swedish culture implies so that he can move freely between the two and function in both.

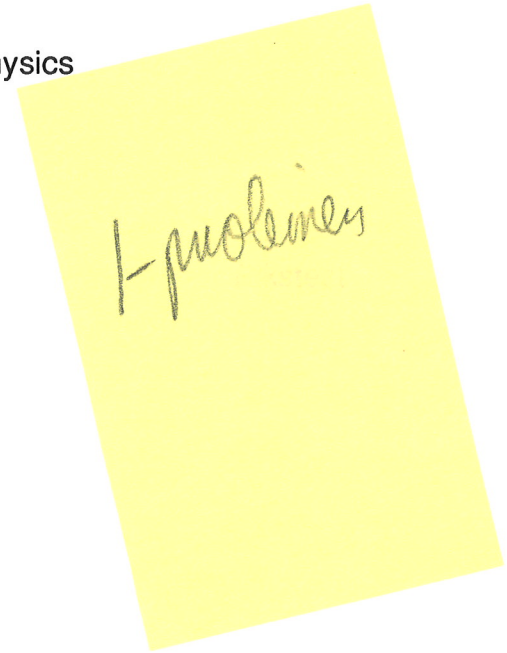
APPENDIX 1

Preparatory Year
Swedish for Academic Purposes

For whom	All foreign students (exception: the Nordic countries)
Aim	To give students compentence in the Swedish language which fulfills the general requirement for university level studies in Sweden
Extent	One academic year, full time Term 1, Term 2 (2 x 16 weeks)
Categories	Guest students (Term 1 + 2) Immigrant and refugee students (Term 2) Visiting researchers Special Programmes
Financing	Guaranteed from home country / state study loan

APPENDIX 2
SYLLABUS
Total 700 hours

Term 1:	introductory week	40 hrs
	Swedish language and Social Science	≈ 250 hrs 4 x 45 min/day 16 hrs/week
	* "Tutorials"	
Term 2:	Swedish language and Social Science	≈ 250 hrs 4 x 45 min/day 16 hrs/week
	History Study Skills Study Counselling	
	Maths, Biology, Physics	
	"Tutorials"	
Final examination:		



* *Each teacher gives remedial tuition to individuals/small groups in his/her class one hour a week*

Lynne Dotzenroth
Anu Virkkunen

NEGATIVE PREFIXATION IN TEXT REPRODUCTION

ABSTRACT

A study was conducted to compare the recall of negatively prefixed words by first and second language learners in oral and written communication. The performance of a group of Finnish EFL students was compared to the performance of a group of native English speakers in a recall task where they were asked to read a short text containing fifteen negatively prefixed words and later recall what they had read either orally or in written form. The results showed that the native English speakers reproduced more than twice the number of negatively prefixed words in written communication than they reproduced in oral communication. The second language learners of English, in contrast, produced slightly more of the negatively prefixed words in their spoken communication than they did in their written communication. The native speakers produced more than twice as many of the negatively prefixed words than the EFL students. It is hypothesized that these differences may be attributed to the differences in the active and passive vocabularies of the two groups, the active vocabularies of the native speakers being much larger than the active vocabularies of the EFL students. The fact that the native English speakers reproduced more than twice as many of the negatively prefixed words in written communication may be due to the **more formal** nature of the written mode as opposed to oral communication.

We wish to thank Prof. Helen Jorstad and Dr. Connie Walker as well as our colleagues Yukie Horiba, Daxing Chen, Carol Ann Pesola, Susan Ranney and Lucy Wu for constructive criticism and support.

PURPOSE

The purpose of this study is to compare the recall of negatively prefixed words by first and second language learners in oral and written communication. The authors, teachers of EFL, became interested in this topic because prefixes are usually taught as a component of the EFL curriculum, yet no experimental information exists on how these prefixes are used by first and second language learners.

BACKGROUND

In this study we refer to prefixes such as un-, il-, im-, ir-, and non- as semantic morphemes. Unlike inflectional or derivational affixes which play a grammatical role in the sentence and tend to be found at the end of a word, semantic morphemes appear to modify only the semantics of the root word and are usually found as prefixes.

Because semantic morphemes do not play a grammatical role in the sentence, they appear to have a different psycholinguistic status than those that do. While grammatical morphemes are stable, unchanging elements in a language, semantic morphemes are sometimes added to a language over a short period of time. For example, the prefix 'super' has recently been added to the lexicon of Spanish via English, and is heard frequently in some dialects of the language. Semantic morphemes are also very productive in creating new terms for science, industry and business. The novel term "the uncola", is readily understood by consumers without explanation.

Traditional grammars of English which teach prefixes such as un-, il-, im-, ir-, and non-, as well as other semantic morphemes, have mentioned that the melding of certain prefixes and root words does not always carry an equal semantic weight. Some prefixes have stronger independent meanings than others, and the same prefix combined with different roots may carry different semantic values. Dwight Bolinger (1952) proposes the examples of "return" and "religion". Both words were originally formed with the prefix **re**; however, the **re** in "religion" carries its meaning only in its historical sense.

The negative prefixes which we look at in this study also seem to be subject to variation in the meaning, that the prefix carries. For example, the **un** of

"unfortunately" does not seem to be as independent from the root word as the **un** in the less frequent word "uncaring".

From classic studies of the 1960s and 1970s on the "tip of the tongue phenomenon", it has been found that the first syllable of a word is particularly important for the word's recognition and recall (Brown and McNeill, 1966; Rubin, 1975; and Fay and Cutler, 1976). As such, we would expect that the semantic prefixes would also be psycholinguistically important for recall and recognition.

Written and oral recall protocols were chosen as the methods of investigation in this study as they tend to reflect "real life" speech better than more controlled experimental methods. There is, however, a drawback to using free recall in that it does not ensure that targeted forms will appear in the recalls. Graf and Mandler (1984) found that subjects reproduced far fewer primed words in free recall than they did when other procedures involving word recognition or completion were used. Based on this study, we expected that the priming provided by our text would improve later recognition of those words but would not necessarily lead to their use in recall protocols.

The only studies to address the possible differences between L1 and L2 learners in free written recall tasks of written texts are those of Patricia Carrell (1984) and Ulla Connor (1984). Carrell found that there was little difference between first and second language learners in their recall of stories which either followed common story schema or violated story schema. Connor found that there was also little difference between L1 and L2 speakers in their recall of higher level ideas from a text they read. The L1 subjects did, however, recall more propositions than the ESL students. From this study we might expect that the L1 speakers would elaborate more than the L2 speakers.

STUDY

In the process of designing and conducting this study, four hypotheses emerged: 1) The recall of a prefixed word will depend on its frequency, i.e. the more frequent the word, the more certain its recall. 2) Written recall is more formal and therefore will include more of the prefixed words than oral recall. 3) If the prefixed words have familiar, less complicated equivalent forms they will be used by the L2 learners provided that the learners have comprehended the term in the first place.

4) The L2 learners will not reproduce as many details of the text as the L1 learners though both will recall the most important ideas in the text.

SUBJECTS

Two groups of students were tested. The first group consisted of eight Finnish agriculture students from the University of Helsinki, seven men and one woman, all in their early twenties. They had all come to the U.S. approximately a year before the testing took place, and they were all participants in an international exchange program. Most of that year they had been working on farms but at the time of testing they had all just completed one quarter of undergraduate studies at the University of Minnesota. Their English exposure was generally uniform in that they all had completed ten years of English in Finnish schools, from the third grade through high school. Seven had taken a two-quarter English reading comprehension course at the University of Helsinki, in which formal instruction of affixes is a component. The student who had not had that course had passed a proficiency test containing a part on affixes. Two students had also spent a month in England, participating in a language course.

The group of native English speakers comprised eight students, high school graduates, of whom three had studied some foreign language. There were three men and five women in the group, and the ages varied from 18 to 25. When questioned, the students indicated that they had not had much formal instruction in English besides English grammar in the 12th grade and freshman composition.

INSTRUMENTATION AND PROCEDURES

The task was based on a text devised by the researchers. It was written in the form of a letter to the editor and contained 170 words out of which fifteen had negative prefixes. The text was divided into two paragraphs the first of which was narrative and included four words with negative prefixes, and the second was expository and included ten words with negative prefixes. The 15th word with a negative prefix was part of the pseudonym for the writer, used to sign the letter (cf. Appendix 1).

Each group was tested separately. To keep the anxiety level down, they were told that this was not a proficiency test and mistakes would not mean anything as the

focus of the study was not the correctness of the language. They were told that the focus of the research was on language.

All the students were asked to read the text at least five times. Though no time limits were given, all subjects completed the task within seven minutes. The papers were collected immediately after the reading task was completed. Then the students were randomly assigned to two subgroups and told that they were to reproduce the text they had read. One group was to reproduce the text in writing and the second group to do so orally. No time limits were given for the recall; the students were simply asked to reproduce everything they could remember of the text.

For the writing task, the students received a blank page. The ones who were to retell the story were interviewed separately. Everybody was also asked to answer two questions: 1) How much exposure have you had to English? (Finns) / How much formal language teaching, in English or foreign languages, have you had? (Americans), 2) What linguistic characteristic of the text do you think is the focus of this study? (everybody).

Before the scores were tabulated, The American Heritage Word Frequency Book was consulted to find out the relative frequencies of the fifteen negatively prefixed words in the text.

RESULTS

With such a small sample size, it was felt that the raw score data would be most informative and that statistical analyses would not yield additional relevant information. Therefore no statistical measures were used.

When the frequencies of the negatively prefixed words were established from the frequency dictionary, it was found that three of the terms 'uncaring', 'atypical' and 'inhumane' were too infrequent even to appear in the dictionary. The following order lists the frequencies of the remaining terms from most frequent to least frequent. The numbers represent frequencies per million words.

List of Frequencies

unfortunately	10.73
unkind	2.82
illegal	1.5
immature	.62
anonymous	.51
irresponsible	.25
mistreatment	.21
irresponsibility	.16
malnourished	.12
disowned	.12
non-profit	.02
devalued*	.02

* The term in the text is **devalues**, which was not in the dictionary; **devalued** was taken as close enough for the purposes of this study.

In examining the written and oral recalls it became apparent that we were dealing with several different kinds of phenomena other than just the reproduction of the negatively prefixes words or their replacement by synonyms.

Since judging the synonymity of particular words was sometimes very difficult, a test of synonymity was required. For the researchers this test of synonymity became substitution of the term or phrase in question in the context of the original text. Other terms, though closely related in meaning to the negatively prefixed words, were actually elaborations of those terms (e.g. **ruthless** for **irresponsible**).

There were also a number of errors made in the reproduction of the terms by both L1 and L2 subjects. In several of these items the mistake was made in the choice of the prefix (e.g. **unprofit** for **non-profit**). In one of the examples the word **malnourished** was obviously intended by the L1 student but the word **malnutrient** was produced. It was decided to create special categories for these cases: M = mentioned using a form with some kind of an error or mistake in the prefix and T = targeted but not reproduced accurately, an error or mistake in the root. These terms will be used throughout the article.

In one instance, an erroneous form was tabulated as a correct term. One L1 subject used **inhumanely** instead of **inhumane** in the syntactic context of an adjective.

Table 1 shows the total scores of the reproduced terms in the order in which they appeared in the text. Taking into account both forms of reproduction, the most frequently produced term is **devalues**, used eight times. Then comes **uncaring**, mentioned seven times, and **inhumane**, mentioned six times. The most frequent word according to the frequency list, **unfortunately**, was mentioned only once. Thus, the results were just the inverse of what was expected in the first hypothesis.

Table 1 REPRODUCTION OF INDIVIDUAL WORDS IN THE ORDER OF APPEARANCE IN THE TEXT

term	X	S/E	M/T	Total
uncaring	1	3+3	.	7
disowned	1	.	.	1
malnourished	3	1	1	5
unfortunately	1	.	.	1
irresponsible	3	.	.	3
atypical	.	1	.	1
immature	.	.	.	0
unkind	1	1	.	2
devalues	2	4+2	.	8
illegal	5	.	.	5
inhumane*	4	.	2	6
irresponsibility	2	.	.	2
mistreatment	.	.	.	0
non-profit	1	.	3	4
anonymous	2	.	.	2
	26	10+5	5+1	47

X mentioned in the original form
S a synonym
E elaboration, using an equivalent term or phrase
M mentioned using a form with some kind of an error or mistake in the prefix
T targeted but not reproduced accurately, an error or mistake in the root

* **inhumanely** for **inhumane** tabulated as an X because it was syntactically used in the same way (cf. **kindly** for **kind** in 'a kindly gentleman')

In the second hypothesis it was proposed that written recall is more formal and therefore would include more of the prefixes words. In our study, this proved to be the case. Referring to Table 2, the scores show that more than twice as many negatively prefixes words were produced in writing (39), as in speech (18). The most marked difference between written and oral communication was in the subcategory X (the exact reproduction of the negatively prefixed word). Table 3

shows that 28 of the terms were reproduced exactly in the written recalls, whereas only 8 were reproduced exactly in the spoken recalls.

Table 2 THE NUMBER OF TERMS REPRODUCED BY MODE OF REPRODUCTION (percentages in parenthesis)

	Mode of Reproduction		
	Written	Spoken	
X	28 (62%)	8 (44%)	Total Number
S/E	7 (24%)	8 (44%)	of
M/T	4 (14%)	2 (12%)	Terms Reproduced
Total	39	18	

X the same term used
S synonym
e an elaborated term
M term mentioned with a mistake in the prefix
T term targeted but not correctly reproduced

Addressing the third hypothesis, the results of the L1 and the L2 groups were compared. It was found that the L2 learners did not replace the negatively prefixes words with more common synonyms (e.g. **hungry** instead of **malnourished**) more often than their L1 counterparts. The results can be seen in Table 3. Thus, the hypothesis was not confirmed.

Table 3 THE NUMBER OF TERMS REPRODUCED BY L1 AND L2 SPEAKERS (percentages in parenthesis)

	L1	L21	
X	19 (59.4%)	7 (47%)	Total Number
S/E	9 (28.1%)	6 (40%)	of
M/T	4 (12.5%)	2 (13%)	Terms Reproduced
Total	32	15	

X the same term used
S synonym
e an elaborated term
M term mentioned with a mistake in the prefix
T term targeted but not correctly reproduced

Evidence to test hypothesis 4 was found by looking at the length of the recalls produced by each subject. Table 4 indicates that L2 Finns produced about half of

the amount of verbiage as the Americans in the written mode, though the group means show that the length of the oral recalls were almost the same. Looking more closely at the individual scores, in the oral mode, however, we find that the score of 206 produced by one of the subjects dramatically increased the Finnish oral mean which for the three other subjects would have been 66.6. As a natural consequence of the length of the recalls the L1 group also reproduced more than twice the number of negatively prefixed terms than the L2 group. The Americans reproduced 32 of these terms whereas the Finns reproduced only 15. Once again, the most marked difference between the groups was within the subcategory X, where the Americans reproduced 19 terms and the Finns only seven (cf. Table 3).

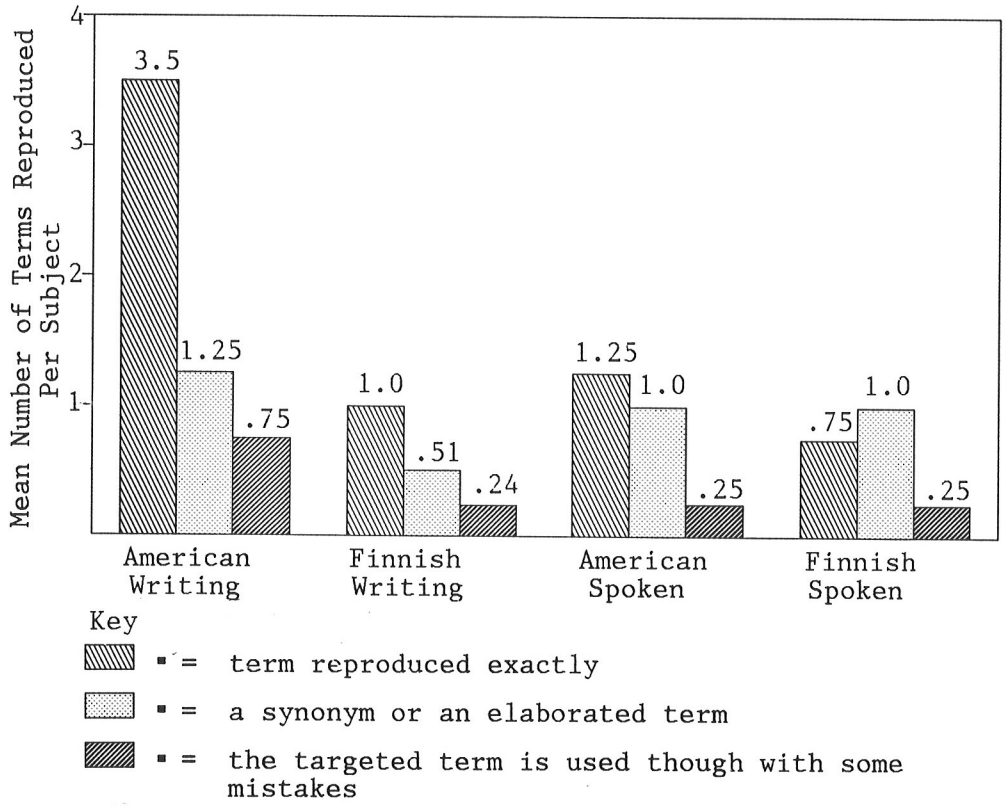
Table 4 INDIVIDUAL RECALL SCORES

subjects	words total	group means	neg. pref. words	group means
AW 1	127		4	
AW 2	124	116	8	5.5
AW 43	125		8	
AW 4	88		2	
FW 1	53		1	
FW 3	94	64.5	3	1.75
FW 5	53		2	
FW 7	58		1	
AS 5	63		2	
AS 6	107	97	3	2.5
AS 7	79		1	
AS 8	139		4	
FS 2	60		2	
FS 4	50	101.5	2	2
FS 6	206		4	
FS 8	90		0	

A American, L1 W written
F Finnish, L2 S spoken

The following graphs (1 and 2) and Table 5 summarize the results discussed so far, allowing the readers to distinguish between language group, mode of communication and subcategorization simultaneously. These sources show the dramatic contribution of subcategory X produced by the Americans in their written recalls while the other subcategories remain more or less similar in all categories.

Graph 1 THE STRATEGIES USED TO REPRODUCE THE VOCABULARY TARGETED BY NATIONALITY AND MODE OF COMMUNICATION



Graph 2 THE STRATEGIES USED TO REPRODUCE THE TARGETED VOCABULARY BY NATIONALITY AND MODE OF COMMUNICATION

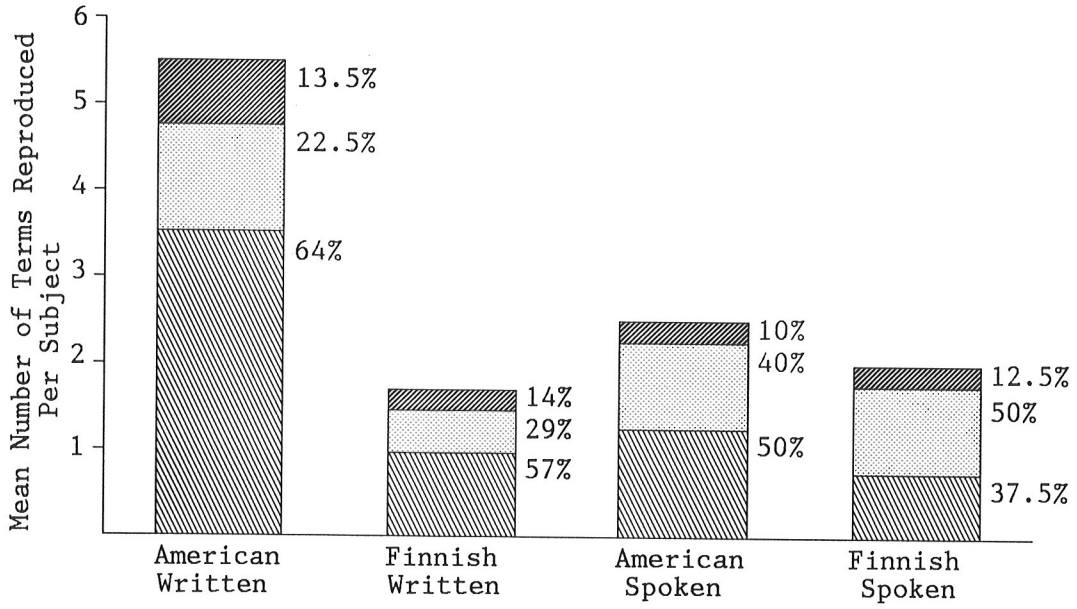


Table 5 THE TOTAL NUMBER OF TERMS REPRODUCED BY NATIONALITY AND MODE

	Nationality and Mode				
	AW	FW	AS	FS	
X	14	4	5	3	26
S/E	5	2	4	4	15
M/T	3	1	1	1	6
Total	22	7	10	8	

DISCUSSION

Perhaps the most salient finding in the data is the difference between the L1 and L2 written recalls. The relatively high number of terms exactly reproduced by the L1 Americans in the written mode was important. These differences may be due to differences in the passive and active vocabularies of first and second language speakers.

Because of the small sample size and the uneven numbers of male and female subjects sex was not regarded as a variable. As the Finnish subjects had had a considerable amount of exposure to both English and the American culture, differences in ethnic backgrounds were not regarded as an important factor either.

Considering the findings of Graf and Mandler (1984), we were not surprised that the number of the terms reproduced by the subjects was not high. In Graf and Mandler's experiment free recall did not yield as much material as did other experimental tasks. Despite the absence of high absolute numbers of recalled terms, the relative frequencies of these recalled terms provided interesting results for the study.

Several limitations should be noted. To begin with, a very small sample size was used. A larger sample size may yield different results. In the testing of the subjects, the different waiting times after reading the text and before providing the oral recalls may have influenced the results. As the subjects read the text together, and

only one researcher administered the test, the waiting time for some subjects was only a few minutes while others waited up to ten minutes before supplying their recall. This may have influenced the results, either giving subjects time to forget or rehearse; the unusually long oral recall was produced by the Finnish subject interviewed last. Another limitation appeared to be that the Finns were more cogniscent of the fact that they were being tested. When asked what the purpose of the study was several suggested that the study might be focusing on vocabulary. The Americans, when asked the same question, stated that the message of the article was the focus of the study, though they were asked to mention the linguistic characteristic.

Future studies may wish to look at differences in the way narrative and expository texts are recalled by L1 and L2 learners. In other words, does a story schema facilitate text recall in L1 and L2, and are there differences in the recalls of first and second language speakers?

Researchers and ESL teachers would find useful information about any universals that might emerge in the way that certain vocabulary items are reproduced/not reproduced in both kinds of text (narrative and expository). The answers to the following questions might also prove useful and interesting: What do the findings tell us about the intrinsic difficulty of certain classes of vocabulary for second language learners? Are certain classes of vocabulary more difficult to learn in a second language for speakers of language X than for speakers of language Y?

Further work could also be done using propositional analysis to see how the different vocabulary terms contribute to the macrostructure or microstructure of the text, that is, how much information in the proposition the specific terms carry and whether their information relates only to the parts of the phrase they themselves are a part of or whether the information bears meaning relating to more extensive parts of the text like a clause or a sentence or a paragraph etc. The fact that we found such curious results comparing the frequencies of the studied terms leads us to question the importance of the semantic load of the various terms as they relate to the larger structure of the text.

BIBLIOGRAPHY

Bailey, N., Madden, C., & Krashen, S. (1974). Is there a natural sequence in adult second language learning? Language Learning, 24, 235-244

Bolinger, D. W. (1952). Linear modification. PMLA, 67, 1117-1144.

Bolinger, D. W. (1965). The atomization of meaning. Language, 41, 55-573

Carrell, P. (1984). Evidence of a formal schema in second language comprehension. Language Learning, 34(2), 87-112

Chimombo, M. (1979). An analysis of the order of acquisition of English grammatical morphemes in a bilingual child. Working Papers on Bilingualism, 18, 202-230

Connor, U. (1984). Recall of text: Differences between first and second language recalls. TESOL Quarterly, 18(2)

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. Language Learning, 24, 37-53

Fay, D., & Cutler, A. (1977). Malaproxisms and the structure of the mental lexicon. Linguistic Inquiry, 3(8)

Graf, P., & Mandler, G. (1984). Activation makes words more accessible, but not necessarily more retrievable. Journal of Verbal Learning and Verbal Behavior, 23, 553-568

Hakuta, K. (1974). A preliminary report on the development of grammatical morphemes in a Japanese girl learning English as a second language. Working Papers on Bilingualism, 24(3), 18-44

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), Second language acquisition. Rowley, MA: Newbury House

Larsen-Freeman, D. (1976). An explanation for the morpheme acquisition order of second language learners. Language Learning, 26, 125-134

Lyons, J. (1968). Introduction to theoretical linguistics. Cambridge: Cambridge University Press

McNeill, D., & Brown, R. (1966). The tip of the tongue phenomenon. The Journal of Verbal Learning and Verbal Behavior

Moore, F. E. (1979). Assessing Secondary Students' Knowledge of Prefixes. Unpublished Doctoral Dissertation, University of Minnesota

Rosansky, E. J. (1976). Methods and morphemes in second language acquisition research. Language Learning, 26(2), 409-425

Rubin, D. (1975). Within word structure in the tip-of-the-tongue phenomenon. Journal of Verbal Learning and Verbal Behavior 14

APPENDIX 1

A LETTER TO THE EDITOR

A small abandoned dog was found last week near the St. Paul Animal Shelter. Some uncaring individual had apparently disowned the animal, and dumped him in a large hefty garbage bag on the side of the road. A passing paper boy noticed the bag moving and opened it up. Inside there was a small 5 month old malnourished springer spaniel puppy. The boy brought the dog to the Animal Shelter where it was adopted shortly thereafter.

Unfortunately, this kind of an irresponsible act is not atypical nowadays but occurs more frequently than most people are aware of. Such an immature and unkind act, which devalues the worth of life itself, is both illegal and inhumane. The irresponsibility of some pet owners sometimes requires intervention from authorities. However, the treatment of animals is not regulated in every state, and animal mistreatment occurs and will probably continue to occur until non-profit organizations can have more effect on legislation and regulations.

Anonymous Pet-lover

Saint Anthony Park

Korkeakoulujen kielikeskus
Jyväskylän yliopisto
PL 35
40351 JYVÄSKYLÄ

Language Centre for
Finnish Universities
PL 35
SF-40351 JYVÄSKYLÄ
FINLAND