52-81 NOA norsk som andrespråk · Årgang 25 · 1/2009

Norsk andrespråkskorpus (ASK) – design og metodiske forutsetninger

Av Kari Tenfjord, Jon Erik Hagen og Hilde Johansen
Universitetet i Bergen

Sammendrag
Artikkelen gir en presentasjon av Norsk andrespråkskorpus (ASK) – et nytt elektronisk innlærerkorpus som er utviklet ved Universitetet i Bergen. Korpuset inneholder personopplysninger om og tekster skrevet av kandidater som har gått opp til to ulike tester i norsk: *Språkprøven i norsk for voksne innvandrere* og *Test i norsk – høyere nivå*. Artikkelens hovedhensikt er å diskutere bruk av feilkoder i korpuset, og hvorvidt bruk av feilkoder i seg selv betyr at en havner i "sammenlikningsfella", som Bley-Vroman (1983) advarer mot i artikkelen "The Comparative Fallacy", altså at vår behandling av språkdataene skjer på målspråkets premisser og dermed undergraver mellomspråk som egne, selvstendige språksystem. I artikkelen argumenteres det for at feilkoding slik det er gjort i ASK, må betraktes som en teorinøytral beskrivelse, en metodisk forutsetning for en teoretisk basert analyse av mellomspråk. 'Anklagen' om å havne i sammenlikningsfella, et angrep som gjerne blir rettet mot klassisk feilanalyse eller enhver annen befatning med *feil*, har sitt utspring i at en ikke skiller klart nok mellom metode og teori i andrespråksforskningen.

Nøkkelord: innlærerkorpus, sammenlikningsfella, feil, feilkoder.

Introduksjon

I denne artikkelen vil vi presentere Norsk andrespråkskorpus (ASK), et elektronisk korpus av innlærertekster som er utviklet ved Universitetet i Bergen.[1]

Hovedmålet med utviklingen av ASK har vært å skape et korpus av innlærertekster som kan styrke mulighetene for empiriske studier av norsk som andrespråk spesielt, og andrespråkslæring generelt. En viktig forutsetning for å starte dette utviklingsarbeidet var muligheten for tverrfaglig samarbeid. Dataene er hentet fra Norsk språktests arkiver, språkteknologikompetansen er hentet fra Aksis (Avdeling for kultur, språk og informasjonsteknologi) og andrespråksforskningskompetansen er hentet fra tidligere Nordisk institutt, nå en del av Institutt for lingvistiske, litterære og estetiske studier. Utviklingen av ASK inngikk i NFR-prosjektet *Parallelle korpus* som pågikk ved Aksis i perioden 2003-2006. Faglig prosjektleder for ASK var Kari Tenfjord.

Korpuset består ikke bare av innlærernes skriftlige tekstproduksjon, men inneholder også annotering av grammatiske kategorier, avvik fra målspråksnormen (feilkoder) og persondata knyttet til informantene.

Med denne presentasjonen av ASK vil vi gjøre rede for hvordan vi har gått frem ved utviklingen av korpuset, og ta opp noen grunnleggende spørsmål av mer epistemologisk art som er relevante for alle andrespråksstudier, nemlig forholdet mellom teori og metode innen vårt fagfelt, som til tider kan være problematisk. Vårt mål er å argumentere for den vitenskapsteoretiske holdbarheten i å bruke et korpus som dette i studiet av andrespråk, spesielt med tanke på at korpuset inneholder feilkoder.

Artikkelen er delvis ment som et motsvar til innsigelser vi har møtt i form av spørsmål om hvordan ASK håndterer kjente mellomspråksfenomener som variasjon, transfer, strategisk språkbruk, overgeneraliseringer osv., og antydinger om at vi ved å feilkode innlærerspråkene gjør oss skyldig i "The Comparative Fallacy" (Bley-Vroman 1983), altså at vår behandling av språkdataene skjer på målspråkets premisser og dermed undergraver mellomspråk som egne, selvstendig språksystem. Dette er spørsmål og innvendinger som i og for seg er velmotiverte og fornuftige, men med denne artikkelen ønsker vi å vise at det å kode en tekst for feil *i seg selv* ikke er å gå i en slik sammenlikningsfelle som Bley-Vroman med rette advarer mot.

I tillegg til spesiell vektlegging av feilkodingen i korpuset, ønsker vi også å gjøre rede for korpusets bruksmuligheter. Til tross for at feilkodingen har tatt mye oppmerksomhet og ressurser i utviklingsarbeidet, er søk etter ulike feiltyper bare en del av korpusets muligheter.

Andrespråkskorpus – hva og hvorfor

Et korpus er ikke bare en samling data, slike som vi ofte har brukt i norsk andrespråksforskning, for eksempel 20 eller 50 tekster hentet fra testbesvarelser eller fra egne elever/studenter evt. med noen personopplysninger om informantene. Den vanlige prosedyren har vært at forskeren har transkribert håndskrevne tekster til maskinskrevne, og behandlet dem videre manuelt, både søk etter spesielle strukturer eller ord, og kvantifisering. Eventuelt har man lagt tekstene inn i regneark for å effektivisere behandlingen. Dette er tidkrevende og i forhold til elektroniske korpora mindre effektivt, og mengden av tekster begrenser seg selv i forhold til den tiden forskeren har til disposisjon. Det at vi nå har muligheten til å behandle store tekstmengder elektronisk, sparer mye tid og gir et sikrere grunnlag for statistiske analyser. Men for at et korpus skal kunne brukes som forskningsverktøy, må det tilfredsstille visse kriterier. Grangers definisjon av et andrespråkskorpus som er basert på Sinclairs (1996) definisjon av (ordinære) korpora, er mye sitert:

> Computer learner corpora are electronic collections of authentic FL/SL [Foreign Language/Second Language] textual data assembled according to explicit design criteria for a particular SLA/FLT [Second Language Acquisition/Foreign Language Teaching] purpose. They are encoded in a standardised and homogeneous way and documented as to their original and provenance (Granger 2002:7).

Korpuslingvistiske undersøkelser er ikke et eget forskningsfelt, heller må det kalles en metode.

> Corpus linguistics can best be defined as a linguistic methodology which is founded on the use of electronic collections of naturally occuring texts, viz. corpora. It is neither a new branch of linguistics nor a new theory of language, but the very nature of the evidence it uses makes it a particularly powerful methodology, one which has the potential to change perspectives on language. (Granger 2002:4)

Granger påpeker videre at bruk av korpus heller er en komplementær datakilde enn en som erstatter andre, slik som for eksempel introspeksjon og elisitering. Men hun sier også at det er generell enighet om at korpus er ”[…] the only reliable source of evidence for such features as frequency.”

(McEnery & Wilson 1996:12). Når man bygger et *andrespråkskorpus (learner corpus),* må man også ta hensyn til at det er stor variasjon i andrespråksbruk (som delvis skiller seg fra variasjon man finner i vanlige monolingvale korpora) og at det er mange ulike faktorer som påvirker både språklæring og språkbruk.

Datamaterialet i ASK

ASK består foreløpig av noe over 1700 tekster på 200-400 ord som er skrevet av kandidater som har vært oppe til *Språkprøven i norsk for voksne innvandrere*[2] og *Test i norsk – høyere nivå*. I henhold til definisjonene i *The Common European Framework of Reference for Languages* (CEFR) måler disse to testene et nivå i nærheten av henholdsvis B1 (*Threshold level*) og B2 (*Vantage level*). Dessuten består ASK av et kontrollkorpus på 200 tekster av samme type skrevet inn av personer med norsk som morsmål. Kontrollkorpuset skiller seg uunngåelig fra hovedkorpuset fordi tekstene ikke er skrevet i en testsituasjon.

En opplagt fordel ved å bruke testbesvarelser som data, er at slike tekster er rimelig homogene når det gjelder både tekstsjanger og produksjonskontekst, i og med at alle kandidatene har tatt testen under de samme fysiske rammebetingelser, med den samme tidsbegrensningen, og at deres testprestasjoner er blitt evaluert av sensorer med den samme type instruksjon og opplæring. Dessuten tilfredsstiller de også krav til homogenitet i den forstand at det kun er tekster fra bestått besvarelser som er lagt inn i korpuset. Det vil si at tekstene i prinsippet er *på* eller *over* et visst funksjonelt nivå.[3] I tillegg gir persondataene som er lagt inn, opplysninger om flere variabler som kan ha påvirkning på språklæringen.

Utvalget av besvarelser er foretatt etter bestemte kriterier, motivert av følgende betraktninger:

1 Informantene skulle representere en rimelig typologisk variasjon av førstespråk *(morsmålskriteriet).*

2 Hver informant skulle ha en definert kommunikativ minimumskompetanse i norsk, noe som er sikret ved at de har bestått testen de har avlagt (*bestått-kriteriet*).

3 For hver informant skulle det foreligge personopplysninger som

kunne tenkes å være relevante for språkutviklingen, og som skulle kunne lenkes til kandidatens besvarelse (*info-kriteriet*)

Vi vil nå gå nærmere inn på de tre utvalgskriteriene.

Morsmålskriteriet
Det viktigste kriteriet for utvalg av tekster til korpuset har vært innlærernes morsmål. Dette er et teoretisk motivert kriterium siden virkningen av nettopp denne variabelen kanskje er den som har vært mest diskutert innenfor vårt fagfelt, samtidig som det ikke på noen måte er dristig å påstå at det i dag forventes av en hvilken som helst moderne teori om andrespråkslæring at den skal kunne si noe om morsmålets virkning på språklæringsprosessen, språkbruk og språkutvikling.

Det er imidlertid metodiske problemer forbundet med det å teste hypoteser hvor morsmålet inngår som en forklarende faktor. Vi mener at en potensielt fruktbar måte å imøtekomme disse problemene på, er å utføre statistiske analyser av språk produsert av innlærere med ulik morsmålsbakgrunn, samtidig som man holder andre faktorer konstante. Dette er også i tråd med moderne metodiske innsikter i transferforskningen (jf. f.eks. Jarvis 2007 og Jarvis & Pavlenko 2008). ASK-korpuset muliggjør transferforskning etter slike metodiske retningslinjer, både fordi mengden av data, bredde i morsmålsgrupper og fordi produksjonsforholdene har vært like for alle informantene.

Av statistiske grunner ønsket vi å ha så mange som 100 tekster fra hver morsmålsbakgrunn på hvert av testnivåene, men på grunn av et ønske om typologisk variasjon blant førstespråkene, mangler vi fremdeles tekster fra høyere nivå for språkene albansk, vietnamesisk og somali.

Til sist var det ønskelig at de ti språkene som ble valgt, skulle representere rimelig store minoritetsgrupper i Norge. I utvalget ble til slutt følgende morsmålsgrupper representert: tysk, nederlandsk, engelsk, spansk, russisk, polsk, bosnisk-serbisk-kroatisk, albansk, vietnamesisk og somali.

Bestått-kriteriet
Informantenes skriftlige prestasjoner er alle bedømt til å ligge *på* eller *over* visse definerte språkferdighetsnivåer. På begge de to testnivåene er *kommunikativ funksjonalitet* lagt til grunn for bedømmelsen. For å bestå disse testene måtte altså kandidatene være i stand til å kommunisere sitt intenderte innhold etter kravene på hvert nivå. Dette kravet er svært viktig, det er nem-

lig ikke bare et spørsmål om hva som er praktisk, men det er faktisk helt essensielt for den metodologiske validiteten til ASK. Dette poenget vil bli drøftet nærmere når vi gjør rede for feilkodingsprosessen, og i den bredere diskusjonen av sammenlikningsfella lenger ned.

Info-kriteriet
Selv om testomstendighetene for alle kandidatene prinsipielt er lik, gjelder ikke dette for innlæringskonteksten, og i så måte er det forskjeller på kandidatene. Derfor har vi i tillegg til selve testbesvarelsene opplysninger om hver av kandidatene som kan være relevant for forskning på andrespråkslæring eller tospråklighet, slik som alder, oppholdstid i Norge, antall timer med språkundervisning osv. Norsk språktest begynte tidlig med å innhente persondata fra testtakerne med tanke på muligheten for videre forskning på dataene. Det var frivillig for kandidatene å gi slike opplysninger til forskningsformål. Ved søknad til Datatilsynet om å legge inn tekster og personopplysninger i korpuset ble det stilt spesifikke krav til hvordan dataene skulle behandles.

Siden vårt korpus omfatter denne typen informasjon om alle kandidatene, er det mulig å kontrollere for en rekke ulikheter både i innlæringskontekst og i innlærerbakgrunn, noe som kan være av stor betydning. Vi er derfor rimelig trygge på at korpuset består det kriteriet som Granger (2002:9) foreslår: ”The usefulness of a learner corpus is directly proportional to the care that has been exerted in controlling and encoding the variables.”

Sammenligningfella i andrespråksforskning

Bley-Vroman (1983) påpeker innsiktsfullt at et *målspråksfiksert* språklæringssyn som går ut på at språktilegnelse handler om å tilegne seg flere og flere av målspråkets strukturelle og leksikalske entiteter helt til en er i mål, ikke åpner opp for mellomspråkets autonomi, og dermed hindrer en i å identifisere de mønstre og krefter som former dets utvikling og de faktorer som påvirker dette. Hvis man i forskningen begrenser seg til en systematisk analytisk sammenlikning mellom morsmålet og målspråket ved hjelp av *målspråkets* strukturelle kategorier, vil man ikke være i stand til å se andre, mer grunnleggende mønstre og strukturer i mellomspråket. Man vil simpelthen ikke kunne identifisere dets strukturelle og funksjonelle egenart.

Bley-Vroman understreker dermed hvor viktig det er å bevare mellomspråkets prinsipielle uavhengighet av *både* morsmålet og målspråket i struktur, virkemåte og utvikling.

Som påpekt av f.eks. Year (2004), er det en grunnleggende feil å tro at man mer eller mindre kan "lese av" informantens mellomspråkskompetanse med denne typen målspråksfiksert metodikk. Lakshmanan & Selinker (2001) utvikler også dette poenget, i tillegg til White (2003), Pienemann (1998), Larsen-Freeman & Long (1991). Som denne lista over henvisninger viser, er synet på *sammenlikningsfella* som en metodisk og teoretisk feil, noe det er bred enighet om innen andrespråksfeltet, til tross for høyst ulike teoretiske tilnærminger til feltet ellers. Som antydet ovenfor, deler også vi denne enigheten fullt ut.

Den form for feilkoding en kan finne i ASK, kan nok tilsynelatende virke som en øvelse i nettopp den type andrespråksforskning som det så sterkt advares mot. Når vi hevder at dette bare er tilsynelatende, er det fordi vi mener man må skille mellom den analytiske og *preteoretiske* informasjonen som korpuset gir, og de *teoretiske modeller* vi konstruerer for å *forklare* denne informasjonen. Det første er et spørsmål om *observasjon*, dvs. å identifisere fenomener som krever forklaring, mens selve forklaringen er et spørsmål om *teori*. Vi mener at disse to nivåene ofte er blitt forvekslet i andrespråksfeltets nyere historie, noe vi kommer nærmere inn på i slutten av artikkelen.

Vi håper at vår presentasjon av ASK og de observasjoner og refleksjoner vi gjør oss i den forbindelse, vil demonstrere at termen *språkfeil* i andrespråksforskning, til tross for det er en upopulær term blant mange av dagens forskere, er en fullstendig legitim metodisk begrepsanvendelse innen for rammen av en reflektert forskningsprosess. Det samme gjelder feilkoding som praksis. Det er *ikke* slik at det å registrere og kategorisere feil nødvendigvis forutsetter en teori som søker innsikt i andrespråkskompetanse ved hjelp av å systematisk beskrive andrespråksperformansen gjennom målspråkets strukturelle kategorier alene, eller som kategoriserer et andrespråktrekk som noe som *samsvarer* eller *avviker* fra målspråket. Vi vil komme tilbake til dette etter å ha beskrevet prinsippene som ligger til grunn for feilkodingen i ASK.

Feilkodingsprosessen

Helhetlig tolkning av teksten forut for rekonstruksjon og koding.
Hver enkelt tekst i korpuset har blitt lest og innholdsmessig tolket så nøy-

aktig som mulig. For å kunne bruke denne typen tekst som forskningsdata, må vi kunne forutsette at informantene uttrykker et rimelig klart, identifiserbart og koherent innhold, og at dette innholdet i tillegg er forståelig med innfødt kompetanse i normert norsk som eneste språklige ressurs. Siden kandidatenes språkbruk er vurdert av sensorer og sertifisert som kommunikativt adekvat, kan vi i utgangspunktet konkludere med rimelig sikkerhet at kandidatenes kommunikative intensjon er begripelig for en hvilken som helst innfødt norsktalende, selv om den logisk nok er formulert i hennes individuelle mer eller mindre målspråkslike mellomspråk. Et viktig poeng her har vært at systematisk mellomspråksanalyse *ikke* skulle være nødvendig for en vellykket tolkning, og en slik analyse har til og med vært eksplisitt forbudt i retningslinjene for feilkodingen.

Generelt vil en hvilken som helst tekst, enten den er produsert av en innfødt språkbruker eller ikke, ikke bli fortolket bare på grunnlag av sitt bokstavlige innhold, men i høy grad også på grunnlag av dens kontekst. Både den umiddelbare språklige konteksten, den litt mer omfattende situasjonskonteksten og den i siste instans brede kulturelle konteksten, er alle relevante nivåer for presis tolkning av forfatterens kommunikative intensjon. Vi mener derfor at denne typen betraktninger er relevante forut for rekonstruksjon og feilkoding, mens systematisk mellomspråksanalyse derimot *ikke* er relevant.

Grunnen til at bestått-kriteriet har vært så viktig, er at det å identifisere språkbruksfeil er meningsløst og logisk umulig hvis man ikke i utgangspunktet har en rimelig klar oppfatning av hva som er ytringens intenderte innhold. Denne fundamentale innsikt er en grunnpillar for Corders klassiske algoritme for feilanalyse (Corder 1973). Vår dokumentasjon av kandidatenes generelle kommunikative kompetanse kombinert med en rimelig kontekstuell tolkning imøtekommer etter vår mening, i alle fall til en viss grad, den generelle innvendingen mot feilanalyse som Lakshmanan & Selinker's (2001) har, nemlig at teksten kanskje ikke gir en adekvat gjengivelse av forfatterens kommunikative intensjon.

Prinsipper for valg av rekonstruksjon

Vi forutsetter altså at vi utelukkende på grunnlag av morsmålskompetanse i norsk i hovedsak har klart å identifisere en koherent intensjon fra forfatterens side. Når denne delen av jobben er gjort, har neste steg vært å identifisere de stedene hvor teksten skiller seg fra målspråkssnormen og velge en rekonstruksjon som kan danne grunnlaget for selve feilkodingen. Tekst-

utdrag med mer enn én mulig rekonstruksjon, har ofte forekommet, og ideelt sett burde man naturligvis registrere alle logisk tenkelige rekonstruksjoner, og kode hvert enkelt alternativ. Dette ville imidlertid ikke være en gangbar løsning fra et praktisk og økonomisk synspunkt. I stedet har vi ved utvalget fulgt to grunnleggende prinsipper som skal sikre en mer eller mindre homogen praksis blant ulike kodere:

Det pragmatiske probabilitetsprinsippet (PP-prinsippet)
Velg den tolkningen av innholdet som er *mest sannsynlig* fra et pragmatisk synspunkt når man tar i betraktning teksten som helhet, i tillegg til den situasjonelle og kulturelle konteksten.

Det minimale modifikasjonsprinsippet (MM-prinsippet)
Velg den rekonstruksjon av teksten som utgjør *minst mulig endring* av originalen. Med andre ord: Velg det alternativet som medfører minst korreksjon.

Disse to prinsippene er ikke alltid kompatible, noe som til tider har framstått som et dilemma, men dette er etter vårt syn uunngåelig uansett hvilken strategi som velges. Følgende eksempler med bruk av 'rød vin' kan illustrere hvordan PP-prinsippet trekker i en retning og MM-prinsippet trekker i en annen, men en skjønnsmessig vurdering tilsier at vi gir PP-prinsippet forrang:

*Det er vanlig å ha litt rød vin med måltiden.

Her vil to innfødte rekonstruksjoner være mulige:

Rekonstruksjon 1: Det er vanlig å ha litt *rød vin* til måltidet.
Rekonstruksjon 2: Det er vanlig å ha litt *rødvin* til måltidet.

Alternativ 1 forutsetter at det er snakk om *all rødfarget vin, uansett vinkategori,* mens alternativ 2 forutsetter at det er snakk om *rødvin, uansett farge*. Som innfødte språkbrukere vet vi at *rød vin* og *rødvin* betyr to forskjellige ting, og denne betydningsforskjellen er reflektert i skriftspråket ved særskrivning versus samskrivning. Velger vi alternativ 1, er jo frasen egentlig feilfri som den står, og derfor vil MM-prinsippet si at det er dette alternativet vi må velge - ingen korreksjon kan vel være mer minimal enn

ingen modifikasjon i det hele tatt! Men det er langt mindre sannsynlig at kandidaten her har ønsket å vektlegge farge alene, snarere enn vinkategori, så på dette pragmatiske grunnlag vil være legitimt å velge alternativ 2, i tråd med PP-prinsippet, til tross for at det er i strid med prinsippet om minimal modifikasjon.

Mens eksempelet ovenfor gjør det naturlig å gi PP-prinsippet forrang, er det MM-prinsippet som foretrekkes i kodingen av følgende eksempel:

*Denne retter bruker vi flest i vårt daglige livet.

Denne setningen kan rekonstrueres på minst tre forskjellige måter:

Alternativ 1: Disse rettene bruker de fleste av oss *i vårt dagligliv*.
Alternativ 2: Disse rettene bruker de fleste av oss *i vårt daglige liv*.
Alternativ 3: Disse rettene bruker de fleste av oss *i dagliglivet*.

De tre alternativene har ulike adverbialledd i slutten av setningen, og er altså strukturelt sett ganske forskjellige. Semantisk og pragmatisk er de derimot nokså like, og derfor kommer ikke PP-prinsippet til anvendelse her, men i stedet fremhever MM-prinsippet alternativ 2 som det naturlige valg, da det er den rekonstruksjonen som medfører minst endring fra originalen. Beskrivelsen av den feilen som er begått, kan da begrenses til galt valg av bøyningsform på substantivet (”livet” i stedet for det korrekte ”liv”), mens de to andre medfører mer omfattende korreksjoner som går over flere ord.

Her kan det innvendes at selv om alle tre alternativer i prinsippet er mulige, må man likevel kunne si at alternativ 3 er den mest idiomatiske varianten, men denne blir altså vraket av våre prinsipper. En kan likevel ikke se bort fra at enkelte kodere ville tillegge denne idiomfaktoren så stor vekt at den likevel ville bli foretrukket ut fra selve hovedprinsippet om at en innfødt rekonstruksjon skal være *innfødt*. Så her har det vært et visst rom for skjønn, og våre rekonstruksjonsprinsipper eliminerer derfor ikke arbitrær bedømmelse fullt ut. Dette rommet for skjønn er naturligvis en ulempe, men er ikke katastrofalt. For uansett kriterier vil forskeren ha mulighet til å vurdere og eventuelt forkaste vår valgte rekonstruksjon. Dette vil alltid være tilfellet uansett teknologisk hjelpemiddel - forskeren må selv inspisere sine data.

Klassifisering av avvik ved hjelp av feilkoder

Ved at alle tekstene i korpuset ble systematisk fortolket og rekonstruert i overensstemmelse med prinsippene skissert ovenfor, ble det mulig å identifisere feilene, dvs. de delene av teksten der mellomspråket ikke følger målspråksnormen, og kategorisere dem ved hjelp av *feilkoder*. Vi minner om at en klar tolkning av hver enkelt setning er en logisk forutsetning for å kunne rekonstruere og feilkode den. Tekstpassasjer som er konstruert slik at det intenderte innhold ikke lar seg tolke overhodet, kan derfor ikke gjøres til gjenstand for standard feilkoding. I slike tilfeller er hele passasjen kodet med "X" (*Uidentifiserbar feil*).

Der innlærertekstene skiller seg fra målspråket, er det blitt tildelt en feilkode og et forslag til korreksjon (se vedlegg 1). Feilkoden kategoriserer feilen ved grammatiske og leksikalske begreper som er relevante for beskrivelsen av målspråket, dvs. normert bokmål, men målspråkets kategorier er *ikke* brukt fordi vi tror at dette reflekterer den "rette" strukturen i mellomspråket, men av metodiske årsaker: For i det hele tatt å kunne snakke om språk trenger man et *metaspråk*, og som vi skal komme tilbake til, har vi valgt målspråket til denne funksjonen fordi vi ønsket et *teorinøytralt* metaspråk.

På samme metodiske grunnlag hadde vi heller ingen som helst motforestillinger mot å kategorisere ordforrådet i tekstene etter målspråkets ordklassekategorier. Ved hjelp av Oslo-Bergen-taggeren[4] har hver ordforekomst blitt automatisk tagget for ordklasse, morfologiske og syntaktiske trekk. Dette har gjort tekstene søkbare på samme måte som monolingvale norske korpora. Det er imidlertid problemer knyttet til å bruke en automatisk tagger som er konstruert for målspråket, på innlærerspråk, ettersom ord med ortografiske feil og andre former for avvik fra målspråksnormen, naturligvis er vanskelig å kjenne igjen for taggeren. I ASK er dette problemet løst ved at taggeren arbeider på *korreksjonen* av ortografiske feil (feiltypen ORT), og ikke på innlærerens ortografi. Men både syntaktiske og morfologiske forskjeller mellom norsk og innlærerspråkene er naturligvis med på å 'forvirre' taggeren, noe som kan føre til at taggeren gjør feil også når det gjelder ordklassetagging. Det har derfor blitt utviklet muligheter for å redigere den automatiske taggingen manuelt, og dette har løst problemet, iallfall til en viss grad. Redigeringen har først og fremst vært av ordklassetaggingen, mens vi har unnlatt å redigere syntaktiske og morfologiske kategorier. Sistenevnte er verken mulig eller ønskelig, siden det faktisk forutsetter analyse av mellomspråkene. Likevel har både den morfologiske taggingen en viss

verdi, særlig i mer eksplorative søk. Men her er det viktig å ikke falle i sammenligningsfella.

For illustrasjon, gjengir vi en liste av våre feilkategorier her (se for øvrig Kodeboken på www.ask.uib.no)

LEKSEMFEIL W galt ord ORT ortografisk feil PART sammensetningsfeil SPL særskrivingsfeil DER avledningsfeil CAP galt valg av stor/liten bokstav FL ord fra andre språk enn norsk	**SYNTAKSFEIL** M ord eller frase mangler R ord eller frase er overflødig O ordstillingsfeil Feilkategorien O har følgende underkategorier: INV inversjon mangler OINV overinversjon SCA setningsadverbial i leddsetninger plassert etter det finitte verbet MCA setningsadverbial i hovedsetninger plassert før det finitt verb
MORFEMFEIL F gal morfosyntaktisk kategori. INFL gal form, men riktig morfosyntaktisk kategori.	**TEGNSETTINGSFEIL** PUNC gal tegnsetting PUNC M mangler tegnsetting PUNC R overflødig tegnsetting
X uidentifiserbare feil, dvs. vanskelig/umulig å tolke tekstutdraget	

Illustrasjon 1: Feilkategorier i ASK

I tillegg til feilkodene som klassifiserer feilen, har koderne angitt et forslag til korreksjon. På grunnlag av denne korreksjonen, genereres det en rekonstruert utgave av hver enkelt tekst, som i teorien skal være grammatisk korrekt ifølge bokmålsnormen. På denne måten er ASK også et *parallellkorpus,* bestående av et korpus av originale innlærertekster og et korpus

med rekonstruksjoner av disse. De to parallelle korpusene er representert med hver sitt brukergrensesnitt, henholdsvis ASK og ASK-korrekt. Begge versjonene kan, sammen eller hver for seg, gjøres til gjenstand for empiriske studier gjennom de søkemulighetene som brukergrensesnittene gir anledning til. I visningen av søk har man også muligheten til å se originalsetningen parallellstilt med den rekonstruerte utgaven (se vedlegg 7).

Feilkoding som teorinøytral beskrivelse

Feilkoding har lenge vært etablert som en standardprosedyre i innlærerkorpora[5], en konsekvens av at korpora med denne typen tekster trenger egne teknikker i forhold til andre typer korpora, slik Granger påpeker:

> [...] computer learner corpora quite naturally call for their own techniques of analysis [...] such as error tagging, which are specially designed to cater for the anomalous nature of learner language. (2002:18)

En slik mer eller mindre mekanisk analyseprosedyre gir ganske mye informasjon om språket i teksten, i den forstand at den skaffer til veie systematisert data om kandidatenes *språkbruk.* Men det er veldig viktig å være klar over at slike teknikker må være i overensstemmelse med den relativt moderne innsikt at innlærerspråk har sin autonomi som egne språksystem, og at deres strukturelle egenart ikke uten videre kan tolkes som et derivat av morsmål eller målspråk. Det er derfor vi så sterkt vi bare kan, understreker at våre termer *feil* og *feilkoding* ikke benevner strukturelle trekk ved mellomspråket som sådant, men kun er tekniske termer for rent analytiske begreper som er empirisk og preteoretisk definert uten at noen spesiell teori om språklæring, språkbruk eller språkutvikling forutsettes. Vårt mål er å kunne bruke tekstkildene til å skaffe til veie og tilrettelegge data som det så er opp til andrespråksforskere, ikke oss som korpusutviklere, å foreslå gode teorier til å redegjøre for.

Det er også viktig å merke seg at ved feilkodingen er det *feilene*, ikke strukturtrekk ved mellomspråket, som har blitt kategorisert og definert, og derfor må ASK ikke forstås som en slags maskin som gir anledning til å avlese mellomspråksstrukturen hos informantene. I stedet kan en si at kodingen, både den automatiske ordklassetaggingen og feilkodingen, påfører tekstene en strukturell beskrivelse som er rimelig veldefinert og systematisk,

men i prinsippet *teorinøytral*. Det er en ekstern analyse som er påført tekstene og som gir en systematisk, sammenliknbar beskrivelse av hvordan mellomspråk kan avvike fra måten innfødte språkbrukere bruker målspråket på. Feilkodingen i korpuset er i så måte kun et redskap for identifikasjon og klassifikasjon av avvik fra den skriftlige målspråksnormen, en måte å påvise forskjeller mellom innlærerspråk og målspråket.

Vi kan altså si at feilidentifikasjon er en praktisk snarere enn en teoretisk oppgave, og at feilanalyse er et *metodisk hjelpemiddel*, ikke en *teori* om andrespråkslæring. Skillet mellom teori og metode har dessverre ofte blitt sammenblandet i vårt fagfelts historie, og når en leser faglitteraturen, kan en av og til få inntrykk av at det å befatte seg med feil overhodet, er å gå rett i Bley-Vromans sammenlikningsfelle. Men vi mener dette bare er en holdbar anklage så lenge man ikke skiller mellom feilanalyse som teori og feilanalyse som metode. Det er av stor viktighet at man ikke blander disse to nivåene sammen, noe Corder også poengterte så tidlig som i 1973, jf. følgende sitat, hvor temaet er utelatelse av målspråkets obligatoriske ubestemte artikkel i den engelske mellomspråkssetningen **there is bus stop*:

> The omission of the article (in this case) is only the surface evidence for an erroneous or idiosyncratic linguistic system. A full description of the error involves 'explaining' it in terms of the linguistic processes or rules which are being followed by the speaker [...] Of course, superficial description is a necessary condition for linguistic explanation but it is not a sufficient one [...]. (Corder 1973:277)

Vårt korpus er, med sin feilkoding, nettopp denne typen overflatebeskrivelse som Corder påpeker som en *nødvendig*, men ikke *tilstrekkelig* betingelse for forståelse av teksten og språket i den. Vår koding er en strukturell beskrivelse av mellomspråkstekstene ut fra målspråksgrammatikkens kategorier, *ikke* av mellomspråkets indre grammatiske struktur. Feilkodingen er derfor ikke et *resultat* av en mellomspråksanalyse, men et metodisk ledd som logisk nok må gå *forut* for en slik analyse i forskningsprosessen.

Håndtering av kjente mellomspråksfenomener

Vi har ved noen anledninger fått spørsmål om hvordan korpuset behandler kjente mellomspråksfenomener som transfer, feilkategoriseringer, over-

generaliseringer, kommunikasjonsstrategier osv. Vårt enkle svar er at vi *ikke* håndterer dem. Etter vår mening er termer som disse ikke empiriske teorinøytrale termer, altså *explanandum-termer*, men i stedet termer som forutsetter en spesifikk teori om hvordan spesifikt identifiserbare språktrekk kan forklares, altså *explanans-termer*. Oppgaven med å gi en teoretisk forklaring på de data ASK skaffer til veie, er forskerens, og ikke vår. Vi presenterer kun data som en hvilken som helst gangbar teori om andrespråkslæring har som oppgave å forklare.

Dette springende punkt kan illustreres med følgende konkrete eksempel: Innlærere med engelskspråklig morsmålsbakgrunn produserer gjerne substantivfraser av typen: **det hus, *den dag* og **den kirke.* Dette er strukturer som korresponderer med engelskspråklige uttrykk som *the house*, *the day* og *the church*, og det ville være nærliggende for noen å klassifisere disse innlærerformene som opplagte tilfeller av *transfer,* siden de så tydelig avspeiler de tilsvarende engelske formene. Andre igjen vil alternativt forklare denne mellomspråksforekomsten som et eksempel på *overgeneralisering*, siden man i norsk faktisk bruker 'den/det/de' som foranstilt bestemt artikkel på stort sett samme måte som den engelske 'the', men med et mer innskrenket bruksområde da de ikke kan brukes umiddelbart foran kjernessubstantivet. Noen kunne til og med påpeke at fenomenet kan være et eksempel på det såkalte *multiple effects principle* (jf. Laksmanan & Selinker 2001, Gass & Selinker 1992), siden to angivelige andrespråkslæringsprosesser, både *transfer* og *overgeneralisering* trekker i samme retning og gir samme resultat.

Disse formene blir i korpuset likevel ikke kodet som transferformer, overgeneraliseringer eller *multippel effect*-former, siden dette er termer som ikke er analytiske, men teoretiske. De er begreper som hører hjemme i teorier som er nyttige som redskap for å *forklare* våre data, men ikke for å *beskrive* dem. Vår jobb som korpusutviklere er å beskrive *hvordan*, men ikke *hvorfor*, de skiller seg fra målspråksgrammatikken, og det gjør vi ved hjelp av det målspråksgrammatiske metaspråket. I ASK ville eksempelvis ordet *det* i **det hus* bli kodet som "R" (redundant), mens ordet *hus* ville bli merket "F" (feil morfosyntaktisk form), og korrigert til *huset*. Videre kan det påpekes at både *det* og *hus* er gangbare ord i målspråksnorsk, hvor de har en ordklassetilhørighet som de blir tagget for i overensstemmelse med målspråkets leksikon. Følgelig vil en forsker som bruker korpuset og ønsker å teste en hypotese om hvordan engelskspråklige innlærere uttrykker bestemthet, kunne søke etter relevante avvik, enten det er strykninger,

tilføyelser eller omrokkeringer, og kan på den måten få bekreftet eller avkreftet hypotesen sin.

Vårt mål er som nevnt å være strengt teorinøytral i den rene beskrivelsen, og derfor tar vi overhodet ikke stilling til om den bruken av bestemt artikkel som er beskrevet ovenfor, skyldes transfer, overgeneralisering, den kombinerte effekt av begge eller ingen av delene. Vår måte å kode selve avviket på, er og blir den samme uansett hva feilen kan forklares med.

Ett annet eksempel kan illustrere dette: Setningen **I går jeg dro til byen,* inneholder et klassisk tilfelle av inversjonsfeil av typen underinversjon. Den ukontroversielt rette rekonstruksjon må her være: *I går dro jeg til byen.* Som vi vet, finnes det høyst ulike, til dels inkompatible, tilnærminger til dette fenomenet, og enten en vil forklare det ut fra en teori om grammatisk kompetanse (jf. Hammarberg & Viberg 1979), en teori om informasjonsprosessering (jf. Håkansson 2001) eller en teori om kommunikativ funksjonalitet (jf. Lund 1997), så er det vårt håp at ASK skal være nyttig for alle andrespråksforskere *uansett teoretisk orientering*. Korpuset har som mål å skaffe til veie eksempler på selve fenomenet, uten å si noe om hva det skyldes, og bruker derfor ikke termer som er spesielle for noen av disse tre teoretiske tilnærmingene i sin beskrivelse av avviket.

Det epistomologiske grunnlaget for vår tilnærming

Prinsippet om teorinøytral beskrivelse er et metodisk grunnprinsipp for ASK, og de begrepene vi bruker er empirisk snarere enn teoretisk forankret, noe som gjør at det blir meningsløst å diskutere om korpuset er "riktig" eller "galt". Det som det derimot er meningsfullt å spørre om, er om den gir data som er relevante eller ikke for en gitt problemstilling. Siden teorinøytral beskrivelse av dataene er et så viktig grunnlag for oss, ønsker vi å gi vår argumentasjon en mer generell vitenskapsteoretisk forankring.

Vår grunnleggende vitenskapsteoretiske premiss er følgende: For at et hvilket som helst observert fenomen kan forklares, må det identifiseres og beskrives så presist som mulig uavhengig av teoretiske fordommer, og selve det begrepsmessige rammeverket for observasjonen (*observasjonsteorien* i Poppers terminologi) må være uten teoretisk problematiske begreper. Dette popperianske prinsippet er etter vårt syn en grunnleggende forutsetning for å teste ut en hvilken som helst teori i et hvilken som helst vitenskapsgren (jf. Popper 2002), og når vi bruker målspråksbeskrivelsen og en generalisert

form for feilklassifisering for å beskrive tekstene, blir det mulig å identifisere fenomener som enhver gangbar teori om andrespråkslæring bør kunne forklare. Vårt deskriptive apparat utgjør i så måte ikke en teori, men en *modell* i Poppers forstand, dvs. et begrepsapparat generert av observasjonsteorien.

Av disse grunnene har det vært veldig viktig for oss at vi så langt som mulig kunne beskrive språket i tekstene med et begrepsapparat som er empirisk definerbart og teorinøytralt. Vårt eneste mål har vært å tilby en objektiv og preanalytisk beskrivelse av informantenes språkbruk, og for dette har vi funnet målspråksgrammatikken som velegnet, både for kategorisering av språklige kategorier og for beskrivelse av avvik fra målspråksnormen. I korpuset er dette den *eneste* rollen målspråksgrammatikken innehar.

Vi deler Poppers oppfatning rent prinsipielt, at med mindre en eller annen form for empirisk identifikasjon og rimelig presist definert preteoretisk beskrivelse av fenomener som trenger forklaringer er mulig, så blir empirisk validering av en hvilken som helst teori om dem ikke bare umulig, men kanskje til og med meningsløs. Så hvis det ikke skulle være mulig å identifisere språkbruk som avviker fra målspråksnormen, enten dette skjer i form av "feil" eller på annen måte, hva slags grunnlag skulle en da ha for å evaluere teorier om andrespråkslæring? Det må være mulig, iallfall i prinsippet, å observere, beskrive, klassifisere og referere til det en hvilken som helst teori om andrespråkslæring/-utvikling/-bruk bør kunne forklare, uten dermed å binde seg til noen aprioriske forestillinger om innholdet i en slik teori.

Målspråksgrammatikken som modell for empirisk observasjon.

Det å kunne beskrive fakta i en teorinøytral ramme er således ikke bare en mulighet, det er en forutsetning at det som skal forklares, lar seg identifisere uavhengig av den teori som måtte ha som mål å forklare det. Vi mener at målspråksgrammatikken er det opplagte valg for rollen som deskriptiv modell. Som nevnt ovenfor, forutsetter vi at datakilden, altså mellomspråkstekstene stort sett lar seg tolke semantisk og pragmatisk med innfødt kompetanse i norsk som eneste språklige ressurs. Vi mener at dette ikke bare er et tilgjengelig, men også et nødvendig og vitenskapsteoretisk velmotivert valg.

En kan derfor si at våre feilkodinger er en systematisk målspråksbasert beskrivelse av observerbare overflatefenomener, inklusive identifikasjon

og klassifikasjon av entiteter og strukturer som avviker fra målspråksnormen, uten at vi tar noen som helst stilling til hva som måtte være forklaringen på de avvik vi registrerer. Vårt mål er derfor ene og alene å tilrettelegge data, dvs. potensielle informasjonskilder, som gjør utprøving av teorier innenfor vårt fagfelt enklere og mer effektivt enn det som kanskje ellers ville vært mulig.

Bruk av målspråksgrammatikken som begrepsmodell er imidlertid ikke den eneste tenkelige løsningen. En alternativ tilnærming kunne være å utføre en uttømmende mellomspråksanalyse av alle tekstene i henhold til en eller annen teoretisk modell, for så å kode tekstene i samsvar med de enhetene og strukturelle begrepene som en slik teori motiverer og forutsetter. Vi nekter ikke for at et slikt prosjekt kunne være teoretisk mulig, men det ville for det første av mange grunner kreve større ressurser enn det som det overhodet ville være mulig å skaffe finansiering til i all overskuelig framtid. Men en mer interessant innvending er at anvendeligheten av en slik modell etter all sannsynlighet ville ha begrenset nytteverdi: Ulike teorier innenfor andrespråksforskningen er så ulike, så foranderlige, så usammenliknbare og så heterogene at et korpus bygd på grunnlag av en enkelt kontroversiell teori ville bli foreldet som rammeverk, og ville være mindre fleksibel enn vår av en rekke forskjellige grunner.

Vi mener at den store fordelen med vår tilnæring er at vår beskrivelsesmodell, målspråksgrammatikkens begreper, er velkjent, ikke bare som en skisse, men som en rik, detaljert og rimelig entydig norm. Dette gjør modellen velkjent for alle som har jobbet med feilkodingen, siden alle har utdannelse i norsk som andrespråk, noe som ikke bare er praktisk, men som også sikrer en enhetlig referanseramme for alle koderne og for alle tekstene, noe som er av stor verdi, både for feilkodingens validitet og reliabilitet.

Hvorfor ASK-databasen ikke er en eneste stor sammenlikningsfelle

Siden vår selvforståelse altså går på at utvikling av korpuset ikke er et teoretisk men et preteoretisk anliggende, erklærer vi oss *ikke skyldige* i en anklage om at vi begår den metodefeilen det er å sammenlikne innlærerspråket med målspråket. Vi mener at vi ikke går i sammenlikningsfella! Vi mener at Bley-Vromans og andres metodologiske problematisering av alle andrespråksteorier som ser mellomspråket med målspråksøyne, er en viktig kritisk påpekning. Men man har fullstendig misforstått hvis man inspirert av denne advarselen om gangbar andrespråks*teori* bannlyser enhver

sammenlikning med målspråket som *forskningsmetode*. Det ville ikke bare være å kaste ut barnet med badevannet, det ville være å begå en fundamental kategorifeil på begrepsnivå! Siden ASK med sine ulike kodinger er en datakilde, og ikke en teoretisk avhandling om andrespråkslæring, kan den ikke avvises med Bley-Vroman argumentasjon.

Dette manglende skillet mellom teori på den ene side og forskningsmetode på den andre, mener vi er en begrepsfeil som har ridd andrespråksfeltet som en mare i mange år. *Kontrastiv analyse*, for eksempel, framstilles ofte både som en teori og som en forskningsmetode. Som metodisk redskap har jo språksammenlikning slett ikke utspilt sin rolle, selv om den teoretiske forestillingsverden som gjorde det særlig nærliggende å bruke denne metoden ikke lenger lever. Det som har utspilt sin rolle er naturligvis kontrastiv analyse som *teori*. På samme måte kan det til og med godt hende at de såkalte *morfemstudiene* rapportert i en serie publikasjoner av Dulay & Burt på 70-tallet (jf. f.eks. Dulay & Burt 1974) i prinsippet fremdeles kan være en relevant metode for andrespråksforskning, fortutsatt at de resultatene denne framskaffer, kan tolkes produktivt av en moderne teori om andrespråksforskning, noe som den naive universalismen som i sin tid motiverte denne metoden, *ikke* var.

Poenget vårt er at den meget rettkomne kritikken mot de teoretiske forestillingene som metodisk kontrastiv analyse, morfemstuder og feilanalyse var motivert og inspirert av, ikke må forveksles med den potensielle nytten disse metodene har som forskningshjelpemidler i seg selv. Da Schachter (1986) påviste *"an error in error analysis",* beviste hun ikke at feilanalyse var "galt", men at den som metode hadde visse begrensninger i utforskning av hennes teori om unnvikelsesstrategisk språkadferd, noe som i sin tid var en meget viktig innsikt. Poenget er at denne påpekningen av feilanalysens uanvendelighet for dette spesifikke formålet ikke på noen som helst måte diskvalifiserer feilanalyse som metode på et mer generelt plan.

Vi synes at denne tendensen til å forveksle teori og metode i SLA har forårsaket mye unødig kontrovers. Derfor er det viktig å understreke at ASK essensielt er et metodeverktøy som ikke forutsetter noen spesielle teoretiske føringer. Derfor er det å utnytte korpuset som *kilde* til potensiell viten om andrespråksfenomener ikke noen *comparative fallacy*. Metodisk sammenlikning mellom innlærerspråk og målspråk er ikke i seg selv å gå i sammenlikningsfella - med mindre man skulle mene at *enhver* referanse til målspråket i andrespråksforskningen er en metodefeil, en posisjon vi mener er fullstendig uholdbar.

Som det framgår av vår redegjørelse for strukturen i korpuset, har vi både i dens arkitektur og måten vi har arbeidet på søkt å unngå dette helt fra grunnen av, i det fromme håp at korpuset skal bli nyttig for forskere som arbeider med norsk eller andre nordiske språk som andrespråk i Norge og verden for øvrig.

Søkemuligheter i ASK

Siden artikkelens hovedhensikt har vært å diskutere bruk av feilkoder i andrespråkskorpus har dette fått mye oppmerksomhet i artikkelen. Søk på feil er imidlertid ikke noe hovedpoeng for mange forskere, og vi vil derfor også rette fokus mot andre søkemuligheter som korpuset gir. Det er nemlig fullt mulig å bruke ASK, uten å ta utgangpunkt i feil og feilkoder. Leech har utviklet sju maksimer for annotering av korpus (Leech's Maxims of Annotation), og den første maksimen lyder: "It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus" (Leech 1993 parafrasert i McEnery & Wilson 1996:25). ASK oppfyller dette kriteriet, da korpusets datamateriale lett kan brukes uten å utføre søk på feilkoder, uten at feilkodene kommer til syne eller på noen måte forvansker tilgangen til rådataene, dvs. tekstene i sin opprinnelige form. For eksempel kan man søke etter forekomster av bestemt ordklasser, enkeltord eller lemma, eller strenger av ord, for eksempel *verb + preposisjon, substantiv + possesiv* eller omvendt: *possesiv + substantiv*, *ordet det + lemma av verbet være*. Vi kan også søke på bestemte uttrykk som *i hvert fall*, *i dag* osv. I tillegg er det mulig å bruke regulære søkeuttrykk, dvs. søkeuttrykk som ikke er lagt inn i korpuset søkemeny. Alle søk kan vises i konkordansformat, eller man kan få frem bruksfrekvensen. Frekvensen kan sorteres etter ulike personopplysninger som morsmål, alder, oppholdstid i Norge osv. Men kan også begrense søkene til å gjelde bestemte morsmål, bare en av testtypene, eller etter andre personvariabler, slik at man ikke søker i alle 1700 tekstene. Se vedlegg 2-7 for illustrasjoner av ulike søk og visningsformat.

Koblingen mellom språkdata, persondata og et effektivt brukergrensesnitt gjør det mulig å bruke ASK til å belyse et bredt spekter av problemstillinger innen andrespråksbruk, andrespråksutvikling, andrespråkslæring og testing av språkkompetanse, ikke minst slike problemstillinger som er generert gjennom tidligere studier i norsk og nordisk andrespråksforskning.

Vi håper imidlertid også at korpuset skal åpne for testing av nyere teorier innen feltet, og ikke minst håper vi at det skal gi rom for spennende eksplorative studier, både av grammatiske, leksikalske, pragmatiske og tekstlige fenomener, i tillegg til undersøkelser av språkeksterne faktorers betydning for språklæringsprosessen; som informantenes alder, kjønn, antall år i Norge, opprinnelsesland, morsmål, utdanningsbakgrunn osv.

Noen avsluttende merknader

Selv om korpuset har rike muligheter til søk uten å bruke feilkodene i det hele tatt, vil vi likevel avslutte med at feilkodingen også er en viktig ressurs i ASK. Vi vil dessuten minne om at referanse til målspråket i andrespråksforskningen er omfattende. Flere av de veletablerte termene i forskningsfeltet vårt forutsetter faktisk en form for målspråkssammenlikning som i høy grad kan diskuteres: både begrepene *tilegnelse*, *fossilisering*, *unnvikelse* og *overgeneralisering* tar utgangspunkt i en form for metaforikk som forutsetter at målspråket er en gjenstand for tilegnelse, selv om mellomspråk i dag betraktes som noe som *skapes* snarere enn tilegnes. Det er altså ikke bare begrepet *feil* som er problematisk. Når vi tenker oss om, finner vi kanskje at hele metaforstrukturen selv i moderne andrespråksteori i større eller mindre grad forutsetter sammenlikning med målspråket, ikke minst selve termen *mellomspråk* gjør det, siden den antyder en reise fra et utgangspunkt til et mål.

Noter

1. Denne artikkelen baserer seg i stor grad på en tidligere artikkel (Tenfjord, Hagen og Johansen 2006) publisert i Rivista di Psicolinguistica Applicata (RiPLA). Korpuset er også presentert i Tenfjord (2004, 2007) og Tenfjord, Meurer og Hofland (2006).
2. Tilsvarer Norskprøve 3 etter den nye læreplanen fra 2005.
3. ASKeladden-prosjektet studerer transfer fra morsmålet i innlærerspråk. Studien er korpusbasert og ASK er den viktigste datakilden. Prosjektleder er Kari Tenfjord, og Norges forskningsråd finansierer prosjektet for perioden 2008–2012. Videreutvikling av ASK er et viktig delmål i ASKeladden, bl.a. blir 1200 av tekstene vurdert på nytt for å få en sikrere og mer detaljert plassering på grunnlag av beskrivelsene i CEFR. Dette delprosjektet ledes av post.doc Cecilie Carlsen ved ASKeladden.
4. Oslo-Bergen-taggeren er en automatisk tagger for norsk bokmål og nynorsk, og en beskrivelse av den finnes på følgende adresse: http://omilia.uio.no/obt/les.html
5. Feilkoding er brukt i de største innlærerkorpusene vi kjenner til, for eksempel *The*

Internationale Corpus of Learner English (ICLE) og *The Cambridge Learners' Corpus (CLC)*.

Litteratur

Bley-Vroman, Robert 1983. The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity. *Language Learning. A Journal of Applied Linguistics*, 2006, volume 33,1, s. 1 – 17.

Corder, Stephen Pit 1973. *Introducing Applied Linguistics.* Harmondsworth: Penguin Education.

Council of Europe 2002. *The Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Dulay, Heidi & Burt, Marina 1974. Natural sequences in Second Language Acquisition. *Language Learning* 24, s. 37-53.

Gass, Susan & Selinker, Larry 1992. *Second Language Acquisition: An Introductory Course.* Hillsdale NJ: Erlbaum.

Granger, Sylviane 2002. A bird's-eye view of learner corpus research. I: S. Granger, J. Hung og S. Petch-Tyson (red.). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam and Philadelphia: John Benjamins.

Hammarberg, Bjørn & Åke Viberg 1979. The place holder constraint, language typology, and the teaching of Swedish to immigrants. *Studia Linguistica* 31 s. 106 – 133.

Håkansson, Gisela 2001. Tense Morphology and verb-second in Swedish L1 children, L2 children and childern with SLI. *Bilingualism: Language and Cognition* 4,1 s. 85 – 99.

Jarvis, Scott 2007. Theoretical and Methodological Issues in the Investigation of Conceptual Transfer. *Vigo International Journal of Applied Linguistics* 4, s. 43 –71.

Jarvis, Scott & Aneta Pavlenko 2008. *Crosslinguistic Influence in Language and Cognition*. New York and London: Routledge.

Lakshmanan, Usha & Larry Selinker 2001. Analysing Interlanguage: How do we know what learners know? *Second Language Research* 17, 4, s. 393 – 420.

Larsen-Freeman, Diane & Michael H. Long 1991. *An introduction to second language acquisition research.* London and New York, Longman.

Leech, Geoffrey 1993. Corpus Annotation Schemes. *Literary and Linguistic*

Computing 8, 4, s. 275 – 281.

Lund, Karen 1997. *Lærer alle dansk på samme måde? En lengdeundersøgelse af voksnes tilegnelse af dansk som andetsprog.* København: Special-pædagogisk forlag.

McEnery, Tony & Andrew Wilson 1996. *Corpus Linguistics*. Edinburgh University Press.

Pienemann, Manfred 1998. *Language Processing and Second Language development: Processability Theory.* Studies in Bilingualism 15, Amsterdam/Philadelphia, Benjamins.

Popper, Karl R. 2002. *The Logic of Scientific Discovery*. Revised edition: Original 1959, London: Routledge.

Schachter, Jacquelyn 1974. An error in error analysis *Language Learning* 27, s. 205 – 214.

Tenfjord, Kari 2004. ASK – A Computer Learner Corpus. *CALL for the Nordic Languages*, s. 147 – 158. Copenhagen Studies in Languages 30, Samfundslitteratur.

Tenfjord, Kari 2007. "ASK and you will find what you seek". I Cecilie Carlsen & Eli Moe (red.). *A Human Touch to Language Testing*. *A collection of essays in honour of Reidun Oanes Andersen on the occasion of her retirement.* Oslo: Novus Press.

Tenfjord, Kari, Jon Erik Hagen & Hilde Johansen 2006. The hows and whys of coding categories in a learner corpus (or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy). *Rivista di Psicolinguistica Applicata (RiPLA)* VI.3, s. 198 – 208.

Tenfjord, Kari, Paul Meurer & Knut Hofland 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC),* s. 198 – 208.

White, Lydia 2003. On the Nature of Interlanguage Representation. I Catherine J. Doughty & Michael H. Long (eds.): *The Handbook of Second Language Acquisition,* s. 19 – 42. Blackwell.

Year, JungEun 2004. Instances of the Comparative Fallacy. I *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* 4,1.

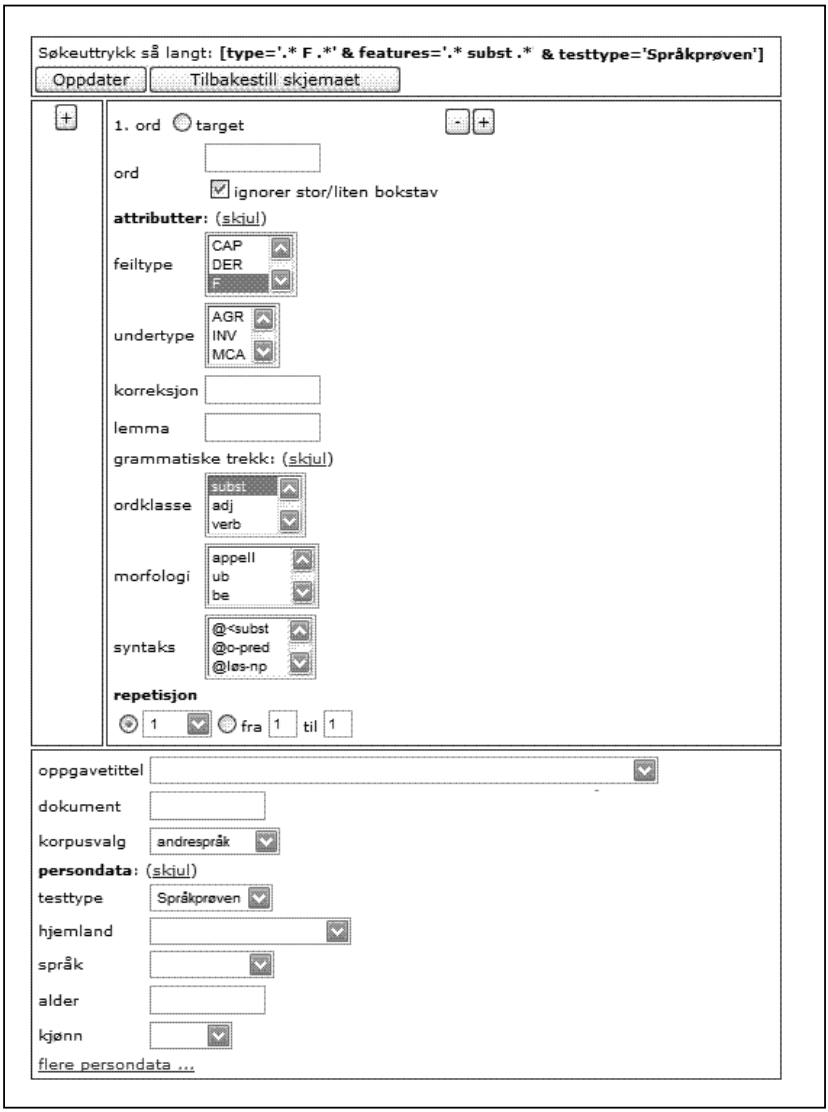
Vedlegg

Vedlegg 1: Eksempel på feilkoding og korreksjon i programmet Oxygen.
Vedlegg 2: ASKs søkegrensesnitt med eksempel på søk med feilkode.
Vedlegg 3: Visning av treff i Kwic-konkordans (Key Word In Context-konkordans)
Vedlegg 4: Eksempel på søk etter sekvens av ord.
Vedlegg 5: Visning av treff i Kwic-konkordans.
Vedlegg 6: De mest frekvente verbene i ASK (kolligasjonsøk).
Vedlegg 7: Visning av treff i setningsparallellstilt konkordans.

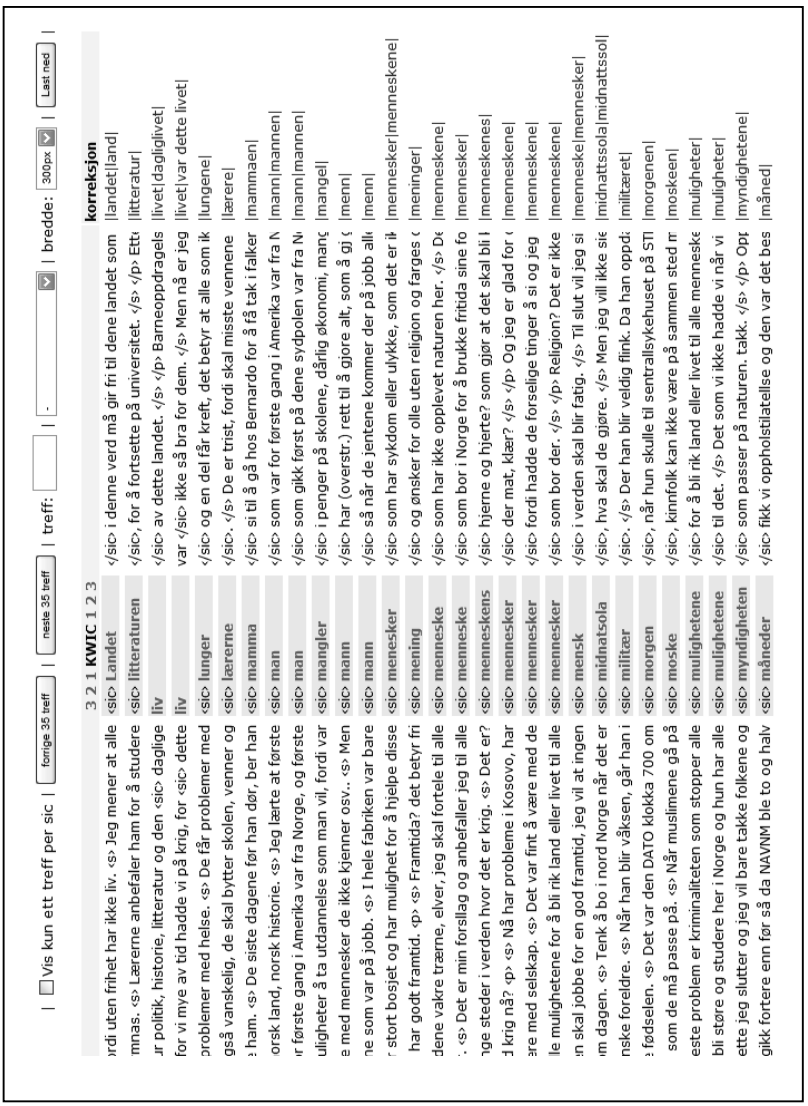
Vedlegg 1: Innlærertekstene er skrevet inn i programmet Oxygen, og ved hjelp av HTML-koding er språk som avviker fra norsk norm kodet med en feilkode og et forslag til korreksjon. Utdraget her viser en feil av typen SPL – særskrivingsfeil.

```
<p>Jeg bor i fjerde etasje. Vi er glade i <sic type="SPL" corr="panoramautsikt">panorama utsikt</sic>. Vi ser hele &sted1;
gjennom vinduet. Jeg har to barn og jeg har alltid ønsket meg å bo på et sånt sted. To skoler og mange barnehager
befinner seg i området. Jeg synes at det er veldig viktig for folk som har barn. Hvis jeg ville kjøpe en leilighet, så skulle jeg
legge vekt på alle disse punktene.</p>
```

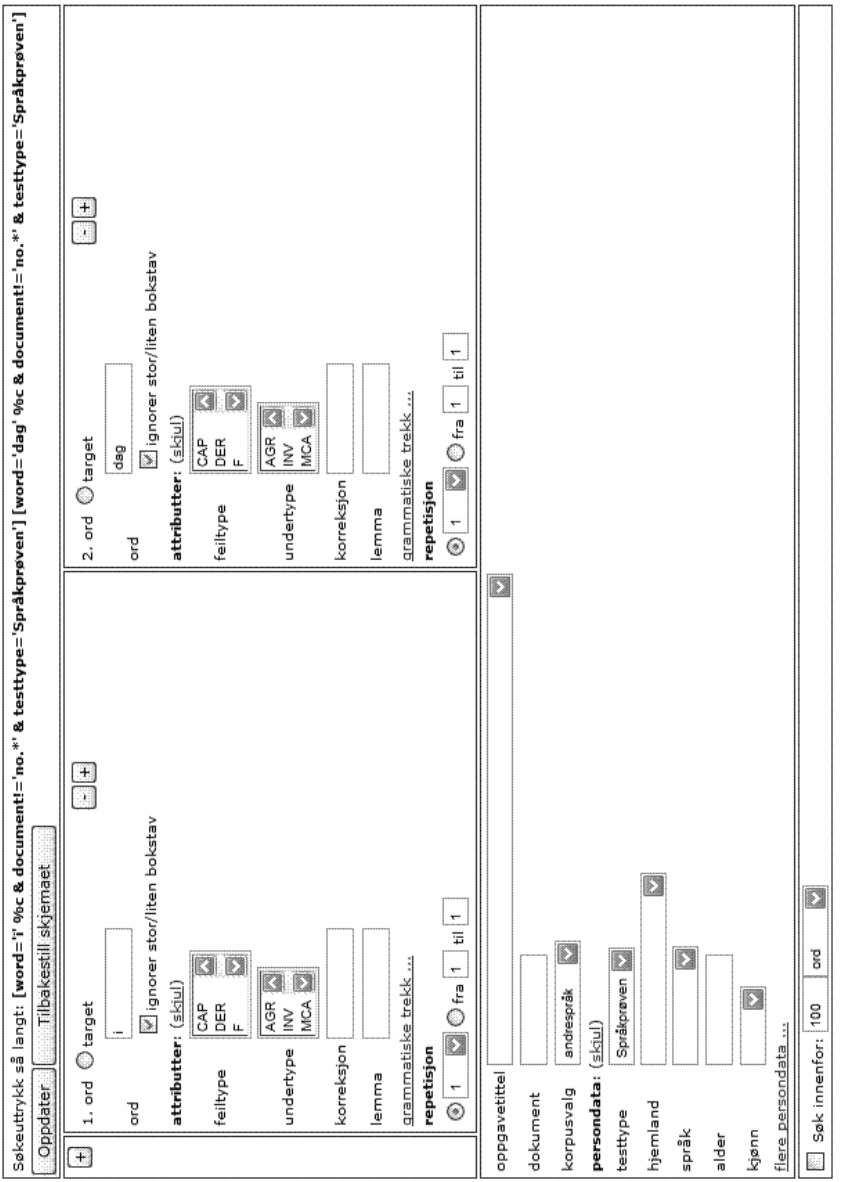
Vedlegg 2. Søk på substantiv med feilkoden F (feil morfosyntaktisk kategori) i Språkprøvetekstene.



Vedlegg 3: Utdrag fra Kwic-konkordansen for søket i vedlegg 2.



Vedlegg 4: viser søk etter forekomster av 'i dag' i tekstene hentet fra Språkprøven i norsk for voksne innvandrere



Vedlegg 5: Viser KWIC-konkordans for søket i vedlegg 4

Vedlegg 6: Liste over de 21 mest frekvente verbene i tekstene hentet fra Språkprøven i norsk for voksne innvandrere.

Søket ga 55754 treff. Det ble funnet 1042 forskjellige kolligasjoner.

match lemma	absolutt frekvens	relativ frekvens
være	11683	0.20955
ha	4956	0.08889
kunne	3014	0.05406
måtte	1543	0.02768
bli	1462	0.02622
skulle	1247	0.02237
gå	1089	0.01953
bo	1082	0.01941
få	1029	0.01846
gjøre	973	0.01745
komme	928	0.01664
ville	877	0.01573
se	867	0.01555
lære	830	0.01489
synes	737	0.01322
like	674	0.01209
si	601	0.01078
ta	566	0.01015
snakke	552	0.00990
tro	525	0.00942
tenke	507	0.00909

Vedlegg 7: En setningsparallellstilt visning av søk på ordstrengen ’jeg er’, hvor verbet ’er’ skal erstattes med et annet verb, gjør det mulig å se både originalsetningen og den rekonstruerte utgaven av den.