Korpus og korpusanalyse i språkdidaktisk forskning og praksis

Anne-Line Graedler, Thomas Egan og Susan Nacey

1. Innledning

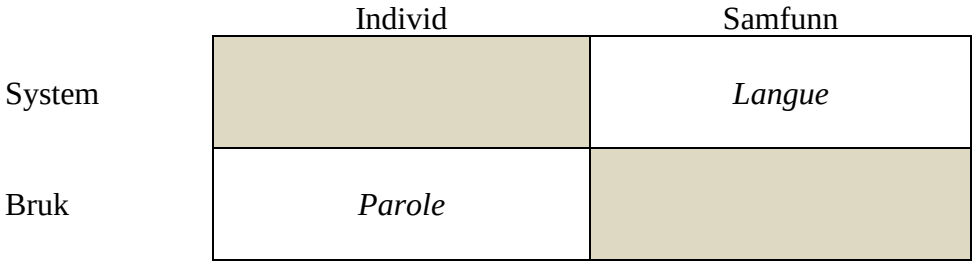
I denne artikkelen vil vi diskutere digitale språkkorpus og korpuslingvistiske metoder spesielt med tanke på anvendelse innenfor fremmedspråksdidaktikk. Etter en kort generell drøfting av språkets vesen og ulike typer evidens, hvor korpuslingvistikken plasseres i det språkvitenskapelige landskapet (del 2), gir vi i del 3 en kort oversikt over bakgrunnen for og utviklingen av digitale korpus, med vekt på ulike typer engelskspråklige korpus. Hovedfokuset ligger på korpus som kan være aktuelle i forbindelse med fremmedspråkslæring og -undervisning, eller forskning på disse områdene, dvs. korpus over elev- og studenttekster, og oversettelses- og parallellkorpus. Relevante innsikter fra korpusanalyse er tema i del 4. Her presenteres eksempler fra vårt eget arbeid med korpusbaserte prosjekter, som viser for eksempel hvordan korpusanalyse gir innsikt i elevers språklæring, eller hvordan korpusanalyse som metode kan bidra til en økt forståelse for tverrspråklige problemstillinger. I del 5 diskuterer vi bruksområder for korpus i fremmedspråkundervisningen i skolen. Her ser vi på praktisk anvendelse av korpusdata i klasserommet, for eksempel selvstyrt læring gjennom utforsking og oppdagelse, og hvordan disse tilnærmingene kan forankres i data gjennom korpusbasert undervisning.

2. Korpuslingvistikkens plass i moderne språkforskning

Den moderne språkvitenskapens fødsel dateres gjerne til utgivelsen av Ferdinand de Saussures *Cours de linguistique générale* (1916). Saussure trekker opp flere skillelinjer som har hatt stor gjennomslagskraft, ikke bare i den språkvitenskapelige diskursen, men også tildels innenfor andre humanvitenskaper, slik som forholdet mellom den syntagmatiske og den paradigmatiske dimensjonen, og mellom synkrone

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

og diakrone språklige fenomener. Den kanskje viktigste av Saussures distinksjoner går mellom språket som system, som et strukturert sett av regelmessigheter som språkbrukere legger til grunn for sine uttrykk, og selve ytringene som individet produserer med utgangspunkt i disse regelmessighetene. Med Saussures terminologi kalles det systemet som deles av språkbrukerne i et språksamfunn for *langue* (språket), og de individuelle ytringene for *parole* (språkbruk). Forholdet mellom dem kan illustreres som i Figur 1.



Figur 1: Saussures *langue* og *parole*

Som Figur 1 viser, var ikke Saussure spesielt opptatt av *idiolekt* (språksystem på individnivå), men fokuserte på regler som deles av alle medlemmene i et språksamfunn. Han hadde også lite å si om de samlede ytringer som produseres i et språksamfunn. Denne innfallsvinkelen til språkstudier deles av Noam Chomsky, grunnleggeren av generativ språkvitenskap, og en av de mest innflytelsesrike lingvister etter annen verdenskrig. I *Aspects of the Theory of Syntax* (1965) skiller Chomsky mellom individets "interne" språksystem, *competence*, og dets *performance*, som er hans navn for *parole*. Chomsky's kompetansebegrep tar utgangspunkt i en idealisert språkbruker som behersker alle sidene av språket i det språksamfunnet vedkommende tilhører. På den måten kan det sidestilles med Saussure's *langue*.

Parallelt med framveksten av den generative lingvistikken ble det utviklet flere andre teorier som tar utgangspunkt i ulike aspekter ved språklige ytringer, for eksempel teorier om pragmatikk, og tekst- og diskursanalytiske tilnærminger.[1] Ulike

[1] Slik som Austins (1962) talehandlingsteori, Grices (1975) teori om konversasjonelle implikaturer, og flere arbeider av Gumperz (for eksempel Gumperz 1982).

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

varianter av det som med en samlebetegnelse kan kalles funksjonell lingvistikk oppsto i kjølvannet av Prag-skolen, og fikk stor innflytelse på europeisk språkforskning. Sterkt forenklet kan man si at funksjonelle tilnærminger har de enkelte språkytringene som sitt metodologiske utgangpunkt, dvs. uttrykk som hører hjemme i *Parole*-cellen i Figur 1. Disse danner så grunnlag for teorier om individets språklige interaksjon (cellen for Individ/System i Figur 1), og videre om det felles systemet som muliggjør slik interaksjon, *langue*.

En del lingvister valgte også å fokusere på språksystemet på individnivå som sitt primære forskningsobjekt. Det ble utført en lang rekke psykolingvistiske studier av de kognitive mekanismene som muliggjør individets produksjon og forståelse av språklige ytringer. På 1980-tallet ga interessen for kognitive prosesser seg bl.a. utslag i utviklingen av kognitiv lingvistikk, en retning innen språkvitenskapen som i dag er i sterk vekst (se Geeraerts & Cuyckens 2008).

Selv om lingvister kan ha svært forskjellige utgangspunkt for sin forskning, både teoretisk og metodologisk, kan man si at de aller fleste lingvister er interessert i språkets struktur, *langue*.[2] Det hersker imidlertid uenighet både om språkets gestalt, og om hvordan man best kan erverve seg kunnskap om språk, både når det gjelder ontologiske og epistemologiske spørsmål. Uenigheten går med andre ord både på teori og metode. Og med tanke på metodologiske utfordringer utgjør *korpuslingvistikken* – språkforskning med utgangspunkt i tekstkorpus – et interessant bidrag. Framveksten av digitale korpus gjør at forskere kan nærme seg cellen i nederste høyre hjørne i Figur 1; ikke i den forstand at man har tilgang til alle språkytringene i et samfunn, men i den forstand at man kan frambringe noe som representerer et tverrsnitt av alle ytringer. Dette gjør at spekulasjoner omkring en del språklige fenomener ikke lenger er verken aktuelt eller nødvendig. Som Geoffrey Pullum sier det: “Looking back at the syntax published a couple of decades ago makes it rather clear that much of it is going to have to be redone from the ground up just to reach minimal levels of empirical accuracy” (Pullum 2007: 36).

[2] Det finnes selvsagt også lingvister som forsker på individets språksystem, idiolekten. For eksempel vil lingvister som arbeider med barns språkinnlæring, eller språktap hos hjerneskadde, ofte gjøre bruk av korpus som består av ett eller noen få menneskers språkproduksjon. Det samme gjelder innenfor kriminalteknisk lingvistikk, hvor man sammenlikner ytringer fra forskjellige individer.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Det er ikke allmenn enighet om hvorvidt korpuslingvistikken kan sies å utgjøre en fullverdig teori om språk, eller om den best kan betraktes som en metode som kan benyttes av lingvister med ulike teoretiske utgangspunkt, som kilde til grunnlagsmateriell for språkforskning. Faktisk brukes korpus i dag som datakilder av lingvister med svært forskjellige teoretiske ståsteder, inkludert enkelte generative grammatikere.

I tillegg til korpus, opererer man gjerne med to andre hovedkilder til språklige data: introspeksjon og eksperimentering. Introspeksjon bygger på forskerens eller informanters språklige intuisjon, og har vist seg, som Pullum har påpekt, å være høyst feilbarlig. I språklig eksperimentering utsetter forskere sine objekter for språklige eller ikke-språklige stimuli, for så å kartlegge deres respons. Eksperimenter kan være svært verdifulle, men er ofte kostbare å gjennomføre. Digitale språkkorpus, derimot, er lett tilgjengelige, ettersom alt som behøves er tilgang til data og en datamaskin.

3. Utviklingen av digitale korpus

Spørsmål om språk har opptatt mennesker siden antikken, og det å bruke korpus – i betydningen innsamlet tekstmateriale – for å beskrive språk, går antakelig like langt tilbake som det å bedrive forskning på språk (Tognini-Bonelli 2001: 50). Språkhistoriske studier, for eksempel, er i sin natur korpusbaserte, ettersom den eneste måten å få tilgang til fortidens språk på, er gjennom bevarte tekster (Leech 2007a: 3; Tognini-Bonelli 2001: 51). Men også for språkvitere som er interessert i synkron språkbruk vil den mest nærliggende form for data bestå av innsamlet språkmateriale – skriftlige og muntlige tekster (Leech 2007b: 316).

I språkvitenskapelig sammenheng innebærer imidlertid en definisjon av korpus spesielle krav til lagringsformat, representativitet og autentisitet:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair 2005).

De tidligste korpusbaserte språkstudiene dateres gjerne til 1960-tallet, som et resultat av utgivelsen av Brown-korpuset (*the Brown University Corpus of American English*, 1964), som omfatter 500 tekstutdrag på ca. 2 000 ord hver, innsamlet fra et systematisk selektert utvalg av amerikanske publikasjoner fra 1961. Et tiår senere ble Brown-korpuset etterfulgt av et tilsvarende britisk korpus, LOB-korpuset (*the Lancaster-Oslo/Bergen corpus*), også på en million ord, og basert på et liknende utvalg britiske publikasjoner fra 1961 (Bowker 2007: 304; Leech 2007a: 4). Brown-korpuset representerer en milepæl i den forstand at det uten tvil er det mest kjente, om ikke det aller første, eksempel på et *maskinlesbart* (elektronisk) språkkorpus. Maskinlesbarhet anses i dag for å være et av de essensielle trekkene ved et korpus, og tas som regel for gitt (Taylor 2008: 195).

En av grunnleggerne av Brown-korpuset, W.N. Francis, minner oss imidlertid om at korpus ikke var noe som plutselig så dagens lys i 1961. Leksikografiske korpus har for eksempel lenge vært brukt i arbeidet med å skille ut ulike betydningsnyanser i ord, og med å finne gode eksempler på hvordan forskjellige ord brukes i kontekst. Kanskje det mest kjente eksemplet fra før dataalderen er grunnlagskorpuset for *The Oxford English Dictionary* (OED). Da tanken om OED først ble lansert i 1879, så man for seg en ny type ordbok – et komplett leksikon over det engelske språk, med kronologisk ordnede sitater for hver atskilt betydning. For å muliggjøre dette, ble fem millioner sitater nedtegnet på sedler og samlet inn i løpet av de omlag femti årene som gikk fra starten på OED-prosjektet til utgivelsen av den første innbundne utgaven i 1928 (Landau 2001: 80). Francis beskriver OED-materialet som “one of the largest bodies of material collected for any linguistic project” (Francis 1992: 21).

Brown-korpuset hadde også minst én direkte forløper, nemlig *The Survey of English Usage* (SEU), innsamlet av Sir Randolph Quirk i siste del av 1950-tallet. SEU ble opprettet for å skaffe grunnlagsdata til en engelsk grammatikk. Det innsamlede materialet ble transkribert på sedler som ble systematisert med utgangspunkt i ulike språklige registre, fra ”formell tale” til ”hverdagssamtale”, og videre arkivert slik at brukere av materialet kunne søke på og sortere sedlene etter spesifikke typer informasjon (Francis 1992: 29). Hvis man for eksempel var interessert i data som kunne belyse en spesiell grammatisk konstruksjon, kunne SEU-arkivet bidra med

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

opplysninger – forutsatt at man fysisk befant seg i London for å utforske materialet! SEU-arkivet er for øvrig også et av de aller tidligste maskinlesbare korpusene, ettersom det ble digitalisert i 1980 (Léon 2007: 39).

3.1 Hva gjør et korpus til et korpus, og ikke bare en samling tekster?

Til tross for Sinclairs definisjon ovenfor, eksisterer ingen fullstendig konsensus om hva et elektronisk korpus er eller bør være. Leechs "a helluva lot of text, stored on a computer" (Leech 1992: 106) dekker de mest grunnleggende krav til et korpus, nemlig at det må bestå av *tekst* – skriftlig eller muntlig, enten i form av komplette tekster eller tekstutdrag. Men for de fleste korpuslingvister er denne definisjonen for vid. Francis hevder for eksempel at et korpus ikke bare består av en samling tekster, men at en bestemt hensikt må ligge til grunn for innsamlingen, dvs. at språkdataene i et korpus må være satt sammen med det siktemål å bedrive *språklig analyse* (Francis 1992: 17).

Videre vil de fleste korpuslingvister hevde at et korpus må være *representativt* for den språklige varieteten det skal si noe om. Ideelt sett skal man kunne trekke generelle konklusjoner på grunnlag av evidensen i et korpus. Spørsmålet om representativitet er imidlertid ikke problemfritt: kan et (avgrenset) tekstkorpus i det hele tatt sies å representere noe annet enn selve tekstmengden i korpuset? Og hvis så er tilfelle, hvordan kan man sikre representativitet? Dessuten bør et korpus inneholde *autentisk* språk. Dersom et korpus utgir seg for å representere for eksempel språket i amerikanske publikasjoner fra 1961, er det nettopp denne typen språk brukerne av korpuset vil forvente å finne. Språk som er produsert som resultat av tester og eksperimentering, og som ikke er naturlig forekommende, bør derfor utstyres med tydelig merking.

For å oppfylle krav til representativitet, autentisitet, og andre viktige hensyn som angår hvilke tekster som skal inkluderes, bygges de fleste korpus opp etter spesielle kriterier. Som eksempel kan vi ta *The British National Corpus* (BNC) fra 1994, på 100 millioner ord. Ideen bak BNC er at korpuset skal representere allmenn, nåtidig britisk-engelsk språkbruk, bade skriftlig og muntlig, dvs. at det er et

allmennspråklig, synkront, blandet ettspråklig korpus. Ifølge retningslinjene for BNC omfatter korpuset tekstutdrag på rundt 45 000 ord hver (*tekstlengde* er med andre ord en faktor), hvorav 10 prosent består av taledata og de resterende 90 prosent av skriftlige tekster. Det finnes også mer spesifikke seleksjonskriterier for tekstutdragene, for eksempel kommer 75 prosent av skriftmaterialet fra informative tekster, mens resten består av såkalte "imaginative texts" (se Burnard 1995 for detaljert informasjon om BNC). Med slike detaljerte kriterier for sammensetning av datamaterialet, skal man ideelt sett kunne oppnå et *balansert* korpus, dvs. et korpus som ikke har slagside mot spesielle typer tekst, kilder, osv.

Tekstene i et elektronisk korpus er gjerne utstyrt med koder som gir brukerne ekstra informasjon. Ulike former for markering (*mark-up*) ivaretar informasjon som ellers ville forsvinne når teksten overføres til digitalt format (for eksempel informasjon om skriftstørrelse og -typer, kursivering og utheving i skriftlige tekster, og informasjon om nøling, gjentakelser, prosodiske detaljer osv. for taledata). *Annotasjon* er informasjon som ikke finnes i originalteksten, men som legges inn av forskere. Den vanligste typen er ordklassetagging (*POS-tagging*), hvor hvert enkelt leksikalsk element i en tekst utstyres med en tagg som angir ordklassetilhørighet (Sinclair 2007: 424-427).

Som en følge av de ulike kriteriene som anvendes når man bygger opp et korpus, finnes det flere ulike korpustyper. De fleste korpus er ettspråklige, og konsentrerer seg gjerne om en spesiell språkvariant, for eksempel amerikansk engelsk (Brown-korpuset) eller norsk bokmål (Leksikografisk bokmålskorpus). Det er også betydelig variasjon når det gjelder korpusstørrelse. Brown-korpuset inneholder en million ord, noe som i sin tid var en anselig mengde, mens det nyere BNC har 100 millioner ord. Det er ingen selvfølge at et korpus har en avgrenset mengde tekst; *monitorkorpus* er åpne korpus som konstant utvides for å fange opp nye ord eller endrede bruksmåter av gamle ord (McEnery & Wilson 2001: 30). Et eksempel på denne typen er Norsk aviskorpus, et selvekspanderende korpus som automatisk samler inn og bearbeider tekst fra et stort utvalg norske aviser, noe som bl.a. muliggjør kontinuerlig overvåkning av nyorddanning i norsk avisspråk. I oktober 2010 omfattet korpuset over 850 millioner ord, og gjennomsnittlig vekst per døgn er

på 200 000-250 000 ord (Aksis, s.a.). Til og med nettet – the World Wide Web – kan anses som et korpus, selv om det her kommer inn spørsmål om repliserbarhet (såkalt *"brittleness"*), i tillegg til hensyn som representativitet og balanse. Ettersom søkemotorer på internett endrer indekserings- og søkestrategier, er det ofte umulig å replisere resultater basert på data fra nettet, noe som i sin tur får alvorlige konsekvenser for resultatenes validitet (se Kilgarriff & Grefenstette 2003; Lüdeling, Evert & Baroni 2005; Renouf, Kehoe & Banerjee 2007).

Per i dag er de fleste korpus basert på skriftlige tekster. Dette er ikke overraskende sett på bakgrunn av at det finnes en rekke praktiske hensyn som påvirker oppbyggingen av et korpus. Tid og energi er to faktorer som alltid spiller inn, og som også henger direkte sammen med de økonomiske rammene for et prosjekt. Juridiske hensyn som opphavs- og eiendomsrett kan også spille en rolle; likeså etiske hensyn. Er det for eksempel etisk forsvarlig å publisere andres personlige korrespondanse som del av et korpus?

Denne typen utfordringer blir gjerne enda mer merkbare i forbindelse med oppbyggingen av et talespråkskorpus. Innsamlingsmetodene må vurderes nøye; de fleste land har for eksempel lover og regler som påbyr at folk må gi samtykke til opptak og bruk av tale. Det såkalte *observer's paradox*, dvs. spørsmålet om informanter oppfører seg på samme måte når de ikke er under observasjon, gjelder også for innsamling av taledata. For å gjøre dataene søkbare, må muntlig språk dessuten transkriberes til skriftlige symboler før det legges inn i et korpus. Uten avanserte talegjenkjenningsverktøy er dette en ekstremt tidkrevende prosess (Bowker 2007: 315; Francis 2007: 287). En tommelfingerregel for prosjektdeltakerne i LINDSEI-korpuset (beskrevet i del 3.2 nedenfor) er at ca. fem timer går med til transkripsjon av 15 minutter tale.

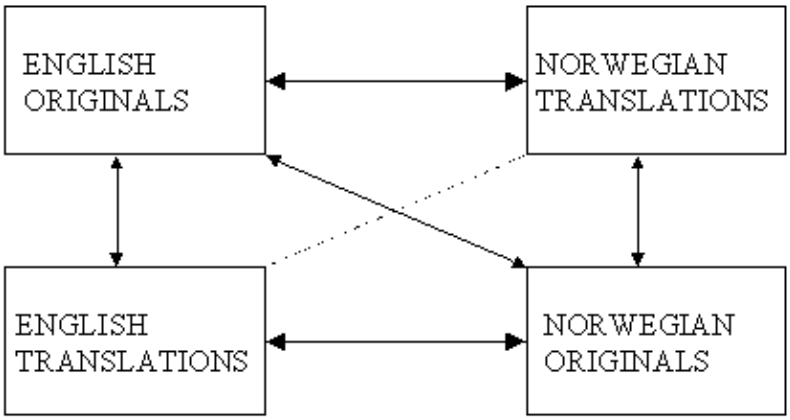
I tillegg til å være tidkrevende, vil transkripsjon av muntlige data alltid innebære en form for fortolkning, i og med at den som transkriberer må gjøre en rekke valg når lyd (og i mange tilfeller også ikke-verbal kommunikasjon) skal overføres til skrift. Dette vil i sin tur si at maskinelle søk i talespråkskorpus ikke er søk i primærdata i egentlig forstand. Det er derfor viktig at resultater fra det transkriberte

korpuset kan kontrolleres mot de opprinnelige lyd- eller videoopptakene, som i størst mulig grad bør foreligge som supplement til transkripsjonen.

3.2 Korpus som pedagogisk ressurs

Korpus inneholder store mengder språklig informasjon, noe som åpner mange muligheter for de som arbeider med språklæring og -undervisning, det være seg forskere eller lærere. To hovedtyper av korpus kan være spesielt verdifulle i denne sammenheng. Den første typen er *oversettelseskorpus*, som består av originaler på ett språk som er parallellstilt med oversettelser av de samme tekstene til et annet språk. Et eksempel på denne typen er *Engelsk-norsk parallellkorpus* (ENPC), et tospråklig korpus bestående av engelske originaltekster med norske oversettelser, og norske originaltekster med engelske oversettelser. Originaltekster fra begge språk er satt sammen i kategorier basert på teksttype og publiseringstidspunkt, slik at man i tillegg til oversettelseskorpuset har et *sammenliknbart korpus*. Som pilene i Figur 2 antyder, åpner denne kombinasjonen for en rekke muligheter: kontrastive studier av originaltekster og oversettelser, andre typer oversettelsesstudier, og tverrspråklige studier basert på parallelle og sammenliknbare originaltekster. Gjennom å legge til flere språk kan man utvide mulighetene i et parallellkorpus ytterligere. *Oslo Multilingual Corpus* (OMC) følger for eksempel samme struktur som ENPC, men inneholder tyske og franske tekster i tillegg til norsk og engelsk, samt en mindre andel nederlandske og portugisiske tekster (både ENPC og OMC er beskrevet i Johansson 2007: 10-21).

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)



Figur 2: Strukturen i ENPC (Johansson 2007: 11).

En annen type korpus som utgjør en verdifull pedagogisk ressurs, er korpus av innlærerspråk (*Computer Learner Corpora, CLC)*. Utviklingen av denne relativt nye typen korpus fant sted på slutten av 1980- og tidlig på 1990-tallet. Flere store forlag har satt sammen kommersielle korpus av innlærerspråk som benyttes som databaser i ordboksproduksjon, men det finnes også en rekke forskningskorpus bygd opp ved akademiske institusjoner. Mange av forskningskorpusene er private korpus, innsamlet og satt sammen av enkeltforskere som datagrunnlag for spesielle prosjekter, mens andre korpus er åpne for forskningsformål (Granger 2004: 129-130). Tilgang til denne typen korpus frigjør forskere fra byrden ved å bygge opp egne korpus, og gir også mulighet for flere ulike studier av samme sett med språkdata.

Korpus av innlærerspråk inneholder tekst produsert av språkbrukere som er i ferd med å lære seg det aktuelle språket, som oftest et andre- eller fremmedspråk (L2), og bare sjelden førstespråket (L1). De er satt sammen med det formål å danne seg et bilde av innlærerens mellomspråk (*interlanguage*) eller *innlærerspråk* (*learner language*), dvs. en idiolekt med trekk fra både L1 og L2, men også trekk som verken finnes i L1 eller L2. Innlærerspråket er ustabilt, i den forstand at det representerer et overgangssystem som endrer seg ettersom innlæreren får stadig bedre kjennskap til og kompetanse i målspråket (Corder 1981: 17, 85). De fleste innlærerkorpus består av skriftlige tekster på engelsk produsert av innlærere med samme førstespråk, og som

ligger på et ferdighetsnivå fra middels til avansert i fremmedspråket (Granger 2004: 130; Nesselhauf 2004: 129).

Dette bildet er imidlertid i ferd med å endre seg. *Norsk andrespråkskorpus* (ASK, s.a.), for eksempel, inneholder norske eksamenstekster fra *Språkprøven i norsk for voksne innvandrere* og *Test i norsk – høyere nivå*, skrevet av innvandrere med svært variert språkbakgrunn. En annen mulighet er å kombinere oversettelses- og innlærerkorpus. Et eksempel på dette er NEST-korpuset (*Norwegian-English Student Translations*) som inneholder norske originaltekster, og flere studentoversettelser per originaltekst, produsert av engelskstudenter ved norske universiteter og høgskoler (se Graedler, under utgivelse).

Et annet unntak fra det typiske innlærerkorpuset med data fra innlærere med samme språkbakgrunn, er *The International Corpus of Learner English* (ICLE), resultatet av et internasjonalt samarbeid under ledelse av Centre for English Corpus Linguistics (CECL) i Belgia. ICLE inneholder ca. 6 000 engelske essay (3,6 millioner ord) fordelt etter L1 på 16 nasjonale delkorpus, fra bulgarsk til tswana. Hovedmålsettingen for initiativtakerne har vært å undersøke hva som gjør at innlærerspråket til viderekomne elever føles fremmed, til tross for at de kan ha en brukbar kommunikativ kompetanse. CECL har også laget et samenlikningskorpus, *The Louvain Corpus of Native English Essays* (LOCNESS), som inneholder essay skrevet av skoleelever og studenter med engelsk som førstespråk, men som ikke er profesjonelle skribenter (dvs. såkalte *novice writers*). Dette korpuset er laget spesielt for at alle som undersøker innlærerspråk i ICLE skal ha enkel tilgang til et kontrollkorpus. I tillegg har CECL initiert et søsterprosjekt til ICLE, *The Louvain Database of Spoken English Interlanguage* (LINDSEI), med delkorpus hvor informanter med ulike førstespråk intervjues av en morsmålstaler. Hvert delkorpus inneholder transkripsjoner av femti intervjuer (Granger, Dagneaux, Meunier & Paquot 2009; Granger & de Cock, s.a.; LINDSEI 2010). Innsamling av data og oppbygging av den norske komponenten av LINDSEI foregår i løpet av 2010 ved Høgskolen i Hedmark

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

4. Hva kan fremmedspråkslærere og -forskere lære av korpus?

Innlærerkorpus og oversettelseskorpus egner seg godt til tverrspråklige studier, spesielt med tanke på å øke vår forståelse av aspekter ved språkstilegnelse. Det meste av forskningen på innlærerkorpus så langt har vært knyttet til metoder assosiert med kontrastiv analyse av innlærerspråk (*Contrastive Interlanguage Analysis, CIA)*. Metoden springer ut av den eldre metoden *kontrastiv analyse*, hvor trekk fra innlæreres L1 og L2 sammenliknes for å avdekke likheter og ulikheter mellom de to språkene, og dermed kunne forutsi hvilke vanskeligheter innlæreren vil møte. CIA innebærer til forskjell fra den tradisjonelle kontrastive analysen at man sammenlikner data fra morsmålstalere og ikke morsmålstalere som ytrer seg på samme språk. Dermed tilrettelegger man gjennom CIA for forskning på to svært viktige områder: sammenlikning av tekster på L2 og L1 for å belyse ulike trekk ved innlærerspråket som avviker fra morsmålsbruk, og sammenlikning av innlærerspråk produsert av språkbrukere med ulik L1-bakgrunn (for eksempel "fransk-engelsk" kontra "norsk-engelsk"), for å kunne avdekke effekten av ulike L1-variabler på L2-data (Granger 2007b: 175-176).

Korpusbasert forskning på innlærerspråk dekker et vidt spekter av emner og områder, og bare toppen av isfjellet er berørt i denne artikkelen (for bibliografier over oversettelseskorpus og innlærerkorpus, se for eksempel VARIENG s.a., og Centre for English Corpus Linguistics, Université catholique de Louvain 2009). Eksempler på områder hvor det er publisert mye forskning, er kollokasjoner, høyfrekvent vokabular, og modalitet i innlærerspråk. Innlærerkorpus avdekker også over- eller underhyppighet av spesielle språklige elementer eller strukturer i innlærerspråk, noe tidligere studier innen tradisjonell kontrastiv analyse ikke var i stand til. Oversettelseskorpus gir et tilleggsperspektiv på de aktuelle språkene, og gir mulighet for å belyse for eksempel områder hvor studentene kan støte på problemer pga. språklig transfer (overføring av elementer eller strukturer fra ett språk til et annet, eller spor av slik overføring i innlærerspråk). Som illustrasjon på den forskningen som pågår, vil vi i det følgende presentere eksempler fra vår egen forskning på temaer med tilknytning til språkinnlæringsproblematikk.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

4.1 Metaforbruk hos norske innlærere av engelsk

Et område som ofte anses å være vanskelig for innlærere å tilegne seg på et fremmedspråk, er metaforbruk; en antakelse er at innlærere produserer tekster som er mer "bokstavelige" i stilen enn tilsvarende tekster skrevet av morsmålstalere (Danesi 1993, 1994). Nacey (2010) undersøker metaforbruk i en komparativ studie av to grupper argumentative essay skrevet på engelsk. Den første gruppen essay er fra den norske delen av ICLE-korpuset (omtalt i del 3.2), og omfatter ca. 20 000 løpeord produsert av norske engelskstudenter på universitets- og høgskolenivå. Den andre gruppen, fra LOCNESS-korpuset (omtalt i del 3.2), teller også ca. 20 000 ord, og er essay skrevet av britiske *A-level*-studenter.

Hvert av de 40 000 ordene fra de to korpusene ble identifisert og analysert ved hjelp av et nyutviklet analyseverktøy, *the Metaphor Identification Procedure*, MIP (Pragglejaz Group 2007; Steen et al., i trykk; Steen et al. 2010). Gjennom denne metoden bestemmes "metaforisiteten" til hver leksikalske enhet i et tekstkorpus. En leksikalsk enhet vil normalt være det samme som et ord, selv om det gjøres noen unntak for partikkelverb, sammensetninger, og flerordslemma som *of course, a lot*¸ osv. Grunnbetydningen til hver leksikalske enhet bestemmes ved referanse til ordbøker, den kontekstuelle betydningen ved referanse til den aktuelle teksten, og disse to betydningene blir deretter sammenliknet. Dersom betydningene er forskjellige, men likevel kan relateres til hverandre gjennom ulike typer sammenlikning (i motsetning til, for eksempel, spesifisering, generalisering, hyperbole, osv.), markeres den leksikalske enheten som et ord som brukes metaforisk.

Før vi ser på noen av funnene i studien, er det nødvendig med en presisering av hva som ligger i termen *metafor*. MIP springer ut fra konseptuell metaforteori, som definerer metafor som resultatet av en overføring fra ett domene til et annet, hvor begrepet i måldomenet forstås i lys av begrepet i kildedomenet. Ifølge konseptuell metaforteori opererer metaforer på to nivåer samtidig – det språklige og det begrepsmessige – og slike metaforer gjennomsyrer vårt språk og vår begrepsverden, og dermed vår tenkning.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

MIP-prosedyren kan identifisere alle språklige metaforer, men ikke de underliggende konseptuelle metaforene. Kort sagt omfatter språklige metaforer de faktiske ordene og frasene som brukes i en tekst, dvs. språkets metaforer, mens konseptuelle metaforer utgjør den underliggende motiveringen for språklige metaforer, dvs. tankens metaforer. Som et eksempel kan vi ta setningen *we spend time wisely*. Den eneste leksikalske enheten som identifiseres som en metafor her, ifølge MIP, er *spend*¸ som et resultat av forholdet mellom de ulike betydningsdistinksjonene dette verbet har i ordbøkene. Den konseptuelle metaforen (TID ER PENGER) som ligger under den språklige metaforen identifiseres ikke gjennom MIP.

I den omtalte undersøkelsen finner Nacey (2010) at 17,8 prosent av ordene i NICLE-materialet er metaforiske, mot 16,8 prosent i LOCNESS-tekstene. Sagt på en annen måte vil det si at ca. ett ord for hvert hvert fem og et halvt ord (5,62) i engelsken til de norske studentene er relatert til en metafor, mens det samme gjelder for ca. hvert sjette ord (5,95) i det britiske materialet. Resultatet av observasjonene vises i tabell 1.

Tabell 1: Foredeling av metaforer, ikke-metaforer og forkastede enheter i NICLE og LOCNESS.

	NICLE	**LOCNESS**	**NICLE + LOCNESS**
Metafor	3 677	3 401	7 078
Ikke metafor	16 998	16 790	33 788
Forkastet	0	52	52
Totalt antall ord	20 675	20 243	40 918

Analysen viser tydelig at metaforisk språk er utbredt i begge varianter av engelsk, selv om ikke-metaforisk bruk er det vanlige. Faktisk inneholder NICLE-materialet med norske innlæreres produksjon av engelsk en noe høyere andel av språklige metaforer enn morsmålsmaterialet i LOCNESS, noe som tilsynelatende motbeviser antakelsen om at tekster skrevet av L2-innlærere har en mer ”bokstavelig” stil enn tilsvarende L1-tekster. Forskjellene i forholdet metafor – ikke-metafor i de to datasettene er statistisk signifikant på nivå p=0.05 ($\chi 2$= 6.31 (df=1), p=0.012). Med LOCNESS-

korpuset som referanse blir konklusjonen at den høye andelen av metaforiske ord i NICLE-tekstene ikke er tilfeldig.

For å avdekke mulige årsaker til den relativt høye forekomsten av metaforer i NICLE, ble alle de identifiserte metaforene fordelt på to kategorier etter grad av konvensjonalitet (”konvensjonell” eller ”ny”), og videre delt inn etter ordklasse, dvs. leksikalske ord (substantiv, verb, osv.) og funksjonsord (preposisjoner, konjunksjoner, osv.). Konvensjonelle språklige metaforer gjenkjennes typisk på den semantiske kodingen de er gitt i engelske standardordbøker. Leksemet SPEND i frasen *spending time*, for eksempel, kategoriseres som en konvensjonell metafor fordi den kontekstuelle betydningen er oppført i ordbøkene. Nye språklige metaforer, derimot, er ikke oppført på denne måten, og er dermed språklige metaforer brukt på en atypisk måte – kanskje som resultat av L1-transfer, kreativitet, eller uheldig ordvalg. Som eksempel kan vi se på bruken av ordet *life-pattern* (istedenfor det konvensjonelle alternativet *lifestyle*) i eksempel (1), sannsynligvis dannet etter modell av det norske ordet *livsmønster*, og dermed et resultat av L1-transfer.

(1) I mentioned earlier that I don’t think that the *life-pattern* of people today gives less room for dreams and imaginations. (ICLE-NO-BU-0002.1)

Tabell 2: Frekvensen av konvensjonelle og nye metaforer i NICLE og LOCNESS.

		NICLE (tot. 20 675 ord)	**LOCNESS** (tot. 20 243 ord)
Konvensjonelle	Leksikalske ord	1 911	1 534
	Funksjonsord	1 586	1 774
Nye	Leksikalske ord	95	54
	Funksjonsord	85	39

Tabell 2 viser inndelingen av språklige metaforer etter grad av konvensjonalitet. Vi ser at metaforisk språk i begge korpus generelt sett er av den konvensjonelle typen. Ca. 95 prosent av NICLE-metaforene er konvensjonelle, og bare fem prosent er nye. LOCNESS-tekstene har både færre konvensjonelle og færre nye metaforer; frekvensene her tilsvarer ca. 97 prosent konvensjonelle og tre prosent nye metaforer. Imidlertid viser det seg at nye metaforer, som er relativt sjeldne i begge korpusene, er

nesten dobbelt så frekvente i tekstene som er produsert av norske innlærere – en forskjell som er høyst signifikant ($\chi 2$= 22.25 (df=1), p=0.0000). Forutsatt at det er en høyere andel konvensjonelle leksikalske metaforer i NICLE, er ikke antallet nye leksikalske metaforer uventet høyt hvis man sammenlikner med de tilsvarende tallene for LOCNESS-materialet. Men forholdstallet mellom konvensjonelle og nye funksjonsord i de to korpusene viser forskjeller på nivå p=0.0005 ($\chi 2$= 23.55 (df=1), p=0.0000). Alt i alt er det 1 663 metaforiske funksjonsord i NICLE, og 94,8 prosent av disse er konvensjonelle, mens 5,2 prosent er nye. I LOCNESS er det 1 813 metaforiske funksjonsord; 97,8 prosent er konvensjonelle, og 2,2 prosent er nye.

Dette vil med andre ord si at dersom man bruker LOCNESS-dataene som referanse, har tekstene i NICLE en overhyppighet av nye metaforiske funksjonsord (nærmere bestemt preposisjoner) i forhold til frekvensen av konvensjonelle funksjonsord. Eksempel (2) er en illustrasjon på dette:

(2) Trevor chooses to do the things he likes *on* his spare time. (ICLE-NO-AC-0021.1)

Her har innlæreren valgt preposisjonen *on*, muligvis som et resultat av transfer fra kollokasjonen *på fritiden*, som inneholder den nærmeste norske semantiske ekvivalenten til preposisjonen *on,* nemlig *på*, mens *in* ville anses som det naturlige valget i standard engelsk. Dette eksemplet må forstås i lys av den underliggende konseptuelle metaforen TID ER ROM, som i denne konteksten realiseres språklig på ulike måter i de to språkene.

Sammenlikningen av metaforer i tekstproduksjonen til to ulike grupper av språkbrukere viser altså at engelsken til norske innlærere er mer metaforisk enn tilsvarende tekster skrevet av britiske L1-studenter, i motsetning til hva man kunne forvente. Størstedelen av metaforene til de norske studentene er konvensjonelle, dvs. de produserer metaforisk språk som er i henhold til de språklige normene. Det mest iøyenfallende avviket ligger i deres (metaforiske) bruk av preposisjoner. Preposisjonsvalg er ofte motivert av metaforiske overføringer, og er ikke tilfeldig. En forståelse av dette, samt kunnskap om hvordan metaforiske utvidelser i ett språk

stemmer (eller ikke stemmer) overens med et annet, kan hjelpe innlærere til å produsere mer idiomatisk språk.

4.2 Tverrspråklig analyse av preposisjoner

En sammenlikning av metaforiske utvidelser i to språk fordrer en grundig analyse av preposisjonsuttrykk i begge språkene. Egan (2010) tar for seg de to engelske preposisjonene *through* og *between*, sammen med deres semantiske ekvivalenter i norsk (*gjennom* og *mellom*), tysk (*durch* og *zwischen*), og fransk (*à travers* og *entre*). Noe av bakgrunnen for undersøkelsen er Kennedys (1991) analyse av forekomster av *through* og *between* i LOB-korpuset (omtalt i del 3 over). Han nevner innledningsvis at "*Between* and *through* are among the hundred or so most frequently used words in English. Like most other structural words, they are semantically complex" (Kennedy 1991: 95), og konkluderer med at

> The analysis of the semantic functions in which *between* and *through* occur in the LOB corpus […] suggests why these words may be difficult to learn or use. For example, both are associated with movement, time and a variety of other relationships (Kennedy 1991: 109).

Egan (2010) undersøker hvorvidt de to preposisjonene representerer semantisk overlapping i en slik grad at det er sannsynlig at de kan forveksles av innlærere. Undersøkelsen ser på de to gruppene med preposisjoner på tvers av språk, dvs. én gruppe som innkoder "gjennom-het" og én som innkoder "mellom-het", både i ENPC og i OMC (begge korpus er omtalt i del 3.2). Oversettelser av uttrykk som innkoder mellom-het har mye mer til felles med hverandre enn oversettelser av uttrykk som innkoder gjennom-het. Dette tyder på at preposisjonen *through* vil kunne by på større utfordringer enn preposisjonen *between* for innlærere av engelsk. Når det gjelder spørsmålet om semantisk overlapping mellom de to begrepene, viser resultatene bl.a. at

- 13 av 690 forekomster av *through* i ENPC er oversatt med *mellom*. Åtte av disse innkoder forflytting i rom, og de resterende fire innkoder persepsjon.

- Ingen av de 486 forekomstene av *between* i ENPC er oversatt med *gjennom.*
- Av 419 forekomster av *mellom* i OMC, er ni oversatt med *through.* Åtte av disse innkoder forflytting i rom, den siste plassering i rom.
- Av 321 forekomster med *gjennom* i OMC, er ingen oversatt med *between.*

Bare når det gjelder ett semantisk forhold, skiller norsk seg vesentlig fra engelsk med hensyn til disse to preposisjonene, nemlig når man innkoder bevegelse mellom flere objekter. En slik bevegelse innkodes ofte som en ”gjennom-handling” i engelsk og en ”mellom-handling” i norsk, som i eksemplene (3) fra ENPC og (4) fra OMC.

(3) Diana picked her way through the women and answered her front door. (ST1)
Diana smøg seg ut mellom kvinnene og lukket opp døren for ham.

(4) Piken gikk *mellom* trestammene, og kom til en lysning. (NF1)
The girl was sauntering *through* the trees and came to a clearing.

Bortsett fra i tilfeller som (3) og (4), bør ikke disse preposisjonene by på noen vanskeligheter for norske morsmålstalere som lærer engelsk. Dette tilsier at språklærere ikke behøver å vie dem særskilt oppmerksomhet, og at det i de fleste tilfeller vil være tilstrekkelig for elevene å lære at *through* tilsvarer *gjennom*, og *between* tilsvarer *mellom*.

4.3 Leksikalske valg i studentoversettelser fra norsk til engelsk

Mens del 4.1 over rapporterer fra en studie med utgangspunkt i et innlærerkorpus, og studien beskrevet i del 4.2 benytter et oversettelseskorpus, vil vi i dette siste eksemplet vår egen forskning kombinere de to perspektivene gjennom å se på noen data fra et innlærer-oversettelseskorpus.

Ordvalg anses gjerne for å gi en god indikasjon på hvor godt en innlærer mestrer målspråket (såkalt *conceptual fluency*). Gjennom å undersøke flere oversettelser av samme originaltekst, kan man få et innblikk i leksikalske valg og

variasjon, både hos profesjonelle oversettere (se omtale av *The Multiple-translation project*, Johansson 2007: 197-198), og hos innlærere. Korpuset NEST (*Norwegian English Student Translations*), som er under utvikling ved Høgskolen i Hedmark (Graedler, under utgivelse), inneholder materiale som kan gi informasjon om leksikalske valg hos innlærere. Korpuset består av oversettelser fra norsk til engelsk gjort av engelskstudenter ved norske universiteter og høgskoler. Selv om NEST ennå ikke fungerer som et fullt utviklet og søkbart korpus, er det mulig å finne eksempler som illustrerer ulike aspekter ved ordvalg i multiple oversettelser av samme tekst til engelsk.

Fra et forskerståsted kan studentoversettelse ses som en type oppgave som nærmer seg et elisiteringseksperiment, i og med at studenten "tvinges" til å gjøre språklige valg som kan omgås eller unngås i friere typer tekstproduksjon. Et eksempel på dette er oversettelsen av det norske ubestemte pronomenet *man.* Oversettelsen av leksemet *man* representerer en utfordring for norske innlærere, ettersom det norske pronomenet er både mer frekvent og forekommer i et bredere stilistisk register enn dets nærmeste leksikalske ekvivalent i engelsk, *one* (Hasselgård, Johansson & Lysvåg 1998: 139). *Stor engelsk ordbok* (Henriksen & Haslerud 2001) foreslår følgende oversettelser for *man:* 1 (som omfatter den du snakker til) *you*; 2 (som omfatter den som snakker) *one*; 3 (folk i sin alminnelighet) *they*, *people*.

Flere av de norske kildetekstene i NEST har forekomster av *man*; eksempel (5)-(8) er tatt fra en tekst som inneholder en eldre oppskrift på krumkaker, og illustrerer en instruktiv teksttype hvor norsk bruker andre språklige virkemidler enn det man vanligvis finner i tilsvarende tekster på engelsk, nemlig passiv form av verbet (forekommer ikke i eksemplene under), og det ubestemte pronomenet *man*. Engelske oppskrifter inneholder typisk verb i imperativ form, uten uttrykt subjekt. Betydningen av *man* i eksempel (5) - (8) tilsvarer hovedsakelig betydning 1 nevnt over (men merk at ordboka mangler pragmatisk informasjon om bruken av *man* i denne spesielle tekstsjangeren). De engelske oversettelsene i (5) - (8) er representative for den vanligste oversettelsesekvivalenten i korpuset, slik det foreligger:

(5) Deretter tar *man* i ca. 5 dl mel, ¼ ts ingefær eller kanel […]

Then add approx. 2 cups flour, ¼ tea spoon ginger or cinnamon […]

(6) Så legger *man* i en full spiseskje av røren, jernet trykkes sammen, […]
Pour a full tablespoon of the mixture on the iron, close it, […]

(7) *Man* kan også ha mer mel i røren så kakene blir tykkere.
You can also add more flour to the batter so that the cakes will be thicker.

(8) Da ruller *man* dem ikke, og slike kaker kalles ”avletter”.
Then *you* do not roll them onto a cone, and such cakes are called “avletter”.

Studentoversetternes ordvalg er oppsummert i Tabell 3. Det er en gjennomgående preferanse for samme type oversettelse i (5) og (6), nemlig imperative verbformer uten uttrykt subjekt. Eksempel (7) inneholder et modalt hjelpeverb som ikke uten videre lar seg oversette til en imperativkonstruksjon, og den negative formen i originalen i (8) vil kanskje føles for mye som en advarsel dersom man velger en oversettelse med imperativ (for eksempel *Do not roll them …*).

Tabell 3: Oppsummering av valg av oversettelsesekvivalenter for *man* i eksempel (5)-(8).

	eksempel (5)	eksempel (6)	eksempel (7)	eksempel (8)
you	2		6	8
imperativ	14	14	2	2
one			3	
they				2
passiv			1	3
andre			2	1

Resultater av den typen som vises i tabell 3 tydeliggjør hvilke språklige kontekster som skaper problemer for norske innlærere, enten ved at det foreligger mange alternative oversettelser uten noen klar indikasjon på hva som er den “beste” eller mest idiomatiske uttrykksmåten på engelsk, eller ved at originalteksten inneholder konstruksjoner som ikke uten videre lar seg realisere på engelsk, og på den måten setter kontrastive forskjeller i fokus.

5. Bruk av korpus i undervisningen

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Korpuslingvistikken er solid forankret i den deskriptive tradisjonen innenfor lingvistikken, og har som mål å beskrive faktisk språkbruk. Avvik fra etablerte normer framstilles gjerne som variasjon, og regelmessigheter i språket ofte som tendenser med ulik styrkegrad, snarere enn absolutte regler. I en fremmedspråklæringskontekst er det imidlertid vanlig å beskrive innlærerspråk med utgangspunkt i målspråkets normer (jf. omtalen av konstrastiv analyse i del 4 over). Ulike typer terminologi benyttes for å identifisere spesielle trekk ved innlærerspråket, som *negativ transfer, fossilisering*, osv. (Granger 2008: 346), men den tradisjonelle samlebetegnelsen på kreative nydanninger i innlærerspråk er *språkfeil*. Her er terminologien i seg selv preskriptiv, ettersom målet med språklæring som regel er at innlæreren skal oppnå en økende grad av kompetanse i målspråket. Når vi i det følgende bruker termen *feil* om innlærerspråk, er det derfor viktig å holde fast ved at formålet fra et forskersperspektiv først og fremst er å beskrive trekk ved innlærerspråket som gjør det til et eget språksystem som er i utvikling, og som kan nærme seg, men ikke er identisk med, målspråket.

Som eksemplene i del 4 over viser, er korpusanalyse egnet til å avdekke både hvilke områder innenfor fremmedspråkstilegnelse som fortjener økt oppmerksomhet, og hvilke som ikke representerer spesielle problemer for innlærere. Videre kan elektroniske korpus gi viktige bidrag til utviklingen av konkrete metoder og undervisningsmateriell til bruk i fremmedspråksundervisningen. Såkalt dataassistert feilanalyse (*Computer-aided Error Analysis*, *CEA*) er spesielt verdifull i så måte. CEA har sitt utspring i tradisjonell feilanalyse (*Error Analysis*) fra 1970-tallet. CEA innebærer annotering av feil eller avvik i innlærertekster, enten ved hjelp av spesiell programvare, eller ved manuell gjennomgang av tekstene i et korpus, og forslag til rettelser gis for hver feil som identifiseres. Sluttresultatet blir et elektronisk korpus med feiltagger, som i sin tur kan få praktisk anvendelse i klasserommet. For eksempel kan man ved hjelp av denne typen korpus få et godt bilde av en bestemt innlærerpopulasjon med tanke på hvilke feiltyper som er vanlige. Basert på slik informasjon kan man så videreutvikle både undervisningsmetoder og -materiell, fra

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

rene læreverk til språkkommentarer i ordbøker (se Dagneaux, Denness, & Granger 1998; Granger 2007a; Nesselhauf 2004).

Kunnskap fra CEA kan også brukes i forbindelse med dataassistert språklæring (*Computer-aided Language Learning*, *CALL*) til å lage praktiske hjelpemidler for språkinnlærere. CALL-metoden innebærer automatisk generering av et variert utvalg oppgaver direkte tilpasset en spesiell gruppe eller en individuell innlærer (Granger 2003). Et eksempel på et slikt hjelpemiddel er læringsplattformen *IWiLL* (*Intelligent Web-based Interactive Language Learning*), utviklet og i bruk ved skoler i Taiwan (se Wible, Kuo, Chien, Liu, & Tsao 2001). Elevene leverer skriftlige oppgaver direkte i IWiLL, og lærerne gir tilbakemelding på for eksempel grammatikk, stilnivå, tekststruktur, osv. Elevene kan deretter søke i dataene for å identifsere hvilke problemer de selv må arbeide med, og lærerne kan bruke de samme dataene for å avdekke problemer, både for individuelle elever, og for gruppen som helhet. Systemet kan brukes til å gi automatisk tilbakemelding til eleven, for eksempel hvis det er spesielle feil eller feiltyper som forekommer hyppig. Etter hvert vil en slik samling av tekster og tilbakemeldinger også kunne gi god innsikt i elevenes utvikling over tid. I tillegg vil denne typen tekster og lærerkommentarer kunne danne utgangspunkt for et ekspanderende korpus som kan gi forskere unike data om både innlærerspråk, læreres vurderingspraksis, og sammenhengen mellom disse.

Korpus kan også utnyttes direkte i klasseromsundervisningen, for å stimulere elevenes læring gjennom oppdagelse (*learning by discovery*), eller såkalt datadrevet læring. Ifølge Nesselhauf (2004: 139) kan datadrevet læring "be attempted straight away by anybody who has access to a learner corpus or is willing to create one." Gjennom en slik tilnærmingsmåte kan data fra innlærerkorpus som viser forskjeller mellom L1 og L2 øke elevenes bevissthet, både om direkte feil, og om under- eller overhyppighet av visse konstruksjoner eller leksikalske elementer. Et eksempel er Hasselgreens (1994) undersøkelse av norske elevers mestring av engelsk vokabular, hvor konklusjonen er at de overforbruker enkelte ord med generell betydning og et vidt bruksområde, sammenliknet med elever med engelsk som L1. Gjennom å se spesifikt på hvilke forsterkende ord og uttrykk som brukes til å modifisere verbet *apologise* finner Hasselgreen at "Norwegian learners are much more likely than

native speakers to use core items like *very (much), a lot (of)*, and *extreme(ly)*" (Hasselgreen 1994: 255), på bekostning av forsterkere som har en mer begrenset bruk, for eksempel *profusely*. Lærere kan selvsagt formidle denne typen kunnskap til elevene ved å fortelle dem om problemet (ofte kalt *lexical teddy bears*). Men å la elevene oppdage slike forskjeller selv, for eksempel gjennom å studere konkordanselister (i dette tilfellet setninger som inneholder leksemet *apologise/apologize*) fra korpus som den norske komponenten av ICLE og søsterkorpuset LOCNESS, kan bidra til mer effektiv læring.

Oversettelseskorpus kan også gi verdifull informasjon til språkeleven ved at man fokuserer på semantiske og pragmatiske forskjeller mellom leksikalske elementer som ellers kan oppfattes som likeverdige. Et søk i ENPC etter engelske oversettelsesekvivalenter for frasene *god morgen* og *god kveld*, sammenholdt med de norske oversettelsesekvivalentene for *good morning* og *good evening*, gir for eksempel nyansert informasjon om måten de to språkene deler inn dagen på – forskjeller en språkstudent ikke alltid er klar over, og som man ikke nødvendigvis finner informasjon om i en vanlig ordbok, men som kan observeres ved hjelp av et korpus. En engelsk morsmålstaler som lærer norsk trenger å vite når hun skal slutte å bruke *god morgen* og gå over til å si *god dag*. Og en nordmann som lærer engelsk trenger å vite på hvilket tidspunkt hun skal gå over fra å si *good afternoon* til å si *good evening*. Som et alternativ til å bruke innlærer- og oversettelseskorpus i klasserommet, er det selvsagt mulig å la elevene få undersøke data fra ettspråklige korpus med tekster av morsmålstalere, for å avdekke grammatiske regler eller underliggende mønstre gjennom egen observasjon.

Ifølge Nesselhauf er datadrevet læring egnet til å utvikle elevautonomi, oppøve elevene i å legge merke til forskjeller mellom morsmålstekst og innlærertekst, og fremme en mer positiv holdning til ulike typer feil (de to første momentene nevnes for øvrig eksplisitt som læringsmål i *Kunnskapsløftet*). Feil i denne sammenhengen bør ses som en mulighet for læring, noe man kan oppdage og lære av, snarere enn et problem som skal korrigeres. Samtidig innrømmer Nesselhauf at denne typen læringsmetodikk er svært sjelden i bruk i klasserommet (Nesselhauf 2004: 140). Flere faktorer bidrar sannsynligvis til dette. For det første er fremmedspråkslærere ofte

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

ukjente både med korpus og med potensialet for språklæring som ligger i denne datakilden. For det andre kan lærere vegre seg for å gi elevene for mye informasjon – autentiske språkdata er verken systematiske eller "rene", og de fleste språkregler har unntak. Dessuten kan morsmålstalere også gjøre feil, som deretter blir en del av språkkorpuset. Språkinnlærere får dermed en stor utfordring i å skille klinten fra hveten gjennom å gå gjennom utallige konkordanselinjer på leting etter et mønster eller en regelmessighet. Sist men ikke minst er tidsaspektet en viktig faktor i klasserommet, og en deduktiv presentasjon av lærestoffet er selvsagt tidsbesparende på kort sikt, i forhold til oppdagelsesbaserte metoder som fremmer induktiv læring. Granger (2004: 136) konkluderer med at "the number of concrete corpus-informed achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice."

6. Oppsummering

I løpet av de siste 30 årene har bruk av korpus ført til grunnleggende endringer innen språkvitenskapen, både når det gjelder teori og metode. Som nevnt i del 2 er ikke språkviteren lenger henvist til intuisjon som kilde til kunnskap om en del språklige fenomener. Bruk av korpusdata har også ført til at flere tidligere etablerte "sannheter", som var basert på forskerens egne eller informanters intuisjoner, er blitt kraftig modifisert eller forkastet. De siste tiårene har resultert i mye ny kunnskap om språk, men det gjenstår fremdeles mange uutforskede muligheter for å bruke korpus til å belyse språklige fenomener, både når det gjelder språk generelt, og – av særskilt interesse for oss – innlærerspråk spesielt. For eksempel har man bare så vidt kommet i gang med å utforske fleroversettelseskorpus. Korpusprosjektet NEST (*Norwegian English Student Translations*) er fortsatt under utvikling, men kan forventes å gi oss data som bidrar til å identifisere områder hvor innlærere møter på spesielle problemer. Når transkriberingen av de norske LINDSEI-dataene er fullført, vil det i tillegg være mulig å sammenlikne det muntlige språket til disse studentene med innlærerspråk fra studenter med andre morsmål gjennom det internasjonale LINDSEI-prosjektet.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Når det gjelder den praktiske anvendelsen av korpus i undervisningen i det fremmedspråklige klasserommet, som vi diskuterer i del 5, finnes det også fortsatt mye upløyd mark. Det mangler for eksempel praktiske håndbøker med nyttige tips for lærere som vil bruke korpus i språkundervisningen. Et annet område nevnt i del 5, som sannsynligvis knapt har vært utnyttet i Norge, er bruk av dataassistert språklæring (CALL) til å lage fortløpende oppdaterte korpus av student- eller elevarbeider. Det er med andre ord ingen mangel på framtidige prosjekter for anvendelse av korpus innenfor fremmedspråksdidaktisk forskning.

Referanser

Aksis (s.a.). Norsk aviskorpus: Statistikk. Lokalisert 14. november 2010, på http://avis.uib.no/om-aviskorpuset/statistikk

ASK (s.a.). ASK - Norsk andrespråkskorpus. Lokalisert 14. november 2010, på http://ask.uib.no/index.page.

Austin, J.L. (1962). *How to do things with words*. Oxford: Oxford University press.

Bowker, L. (2007). Towards a Corpus-Based Approach to Terminography. I: Teubert & Krishnamurthy (2007), bind 3, s. 303-324.

Burnard, L. (red.) (1995). Users Reference Guide for the British National Corpus. Published for the British National Corpus Consortium by Oxford University Computing Services, Oxford. Lokalisert 14. november 2010, på http://www.csd.abdn.ac.uk/~cmellish/teaching/NLP/practicals/bnc-doc.pdf.

Centre for English Corpus Linguistics, Université catholique de Louvain (2009). Learner corpus bibliography. Lokalisert 14. november 2010, på http://www.uclouvain.be/en-cecl-lcBiblio.html.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: The M.I.T. Press.

Corder, S. P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-Aided Error Analysis. *System* 26, s. 163-174.

Danesi, M. (1993). Metaphorical Competence in Second Language Acquisition and Second Language Teaching: The Neglected Dimension. I: J. E. Alatis (red.),

Language, Communication and Social Meaning (s. 489-500). Washington DC: Georgetown University Press.

Danesi, M. (1994). Recent Research on Metaphor and the Teaching of Italian. *Italica: Bulletin of the American Association of Teachers of Italian* 71, s. 453-464.

Egan, T. (2011). Through seen through the looking glass of translation equivalence. I: Hoffman, S., Rayon, P. & G. Leech (red.), *Corpus linguistics: Looking back - moving forward*. Amsterdam: Rodopi.

Egan, T. (Under utgivelse.). Between and through revisited. *VARIENG Studies in Variation, Contacts and Change in English*. http://www.helsinki.fi/varieng/journal/

Francis, W. N. (1992). Language Corpora B.C. I: Svartvik (1992), s. 17-32.

Francis, W. N. (2007). Problems of Assembling and Computerizing Large Corpora. I: Teubert & Krishnamurthy (2007), bind 1, s. 285-298.

Geeraerts, D. & Cuyckens, H. (red.) (2008). *Handbook of Cognitive Linguistics.* Oxford: Oxford University Press.

Graedler, A.-L. (Under utgivelse.). NEST – a corpus in the brooding box. *VARIENG Studies in Variation, Contacts and Change in English*. http://www.helsinki.fi/varieng/journal/

Granger, S. (2003). Error-Tagged Learner Corpora and Call: A Promising Synergy. *CALICO Journal* 20 (3), s. 1-16.

Granger, S. (2004). Computer Learner Corpus Research: Current Status and Future Prospects. I: U. Connor & T. A. Upton (red.), *Applied Corpus Linguistics: A Multidimensional Perspective* (s. 123-145). Amsterdam/New York: Rodopi.

Granger, S. (2007a). A Bird's-Eye View of Learner Corpus Research. I: Teubert & Krishnamurthy (2007), bind 6, s. 44-72.

Granger, S. (2007b). The computer learner corpus: A versatile new source of data for SLA research. I: Teubert & Krishnamurty (2007), bind 2, s. 166-182.

Granger, S. (2008). Learner Corpora in Foregin Language Education. I: Van Deusen-Scholl, N. & Hornberger, N.H. (red.), *Encyclopedia of Language and Education, vol. 4: Second and Foreign Language Education* (s. 337-351). New York: Springer.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (red.). (2009). *International Corpus of Learner English, Version 2*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Granger, S., & de Cock, S. (s.a.). Locness: Louvain Corpus of Native English Essays Lokalisert 24. august 2007, på http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm

Grice, H. P. (1975). Logic and conversation. I: Cole, P. & Morgan (red.), *Syntax and semantics: Speech acts* (bind 3, s. 41–58). New York: Academic.

Gumperz, J. (1982). *Discourse Strategies*. Cambridge, UK: Cambridge University Press.

Hasselgren, A. (1994). Lexical Teddy Bears and Advanced Learners: A Study into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics*, 4 (2), s. 237-260.

Hasselgård, H., Johansson, S. & Lysvåg, P. (1998). *English Grammar: Theory and Use.* Oslo: Universitetsforlaget.

Henriksen, P. & Haslerud, V. (red.) (2001). *Engelsk stor ordbok.* Oslo: Kunnskapsforlaget.

Johansson, S. (2007). *Seeing through Multilingual Corpora.* Amsterdam/Philadelphia: John Benjamins.

Kennedy, G. (1991). Between and through: The Company They Keep and the Functions They Serve. I: Aijmer, K. & Altenberg, B. (red). *English Corpus Linguistics* (s. 95–110). London: Longman.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29 (3), s. 333–347.

Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. I: Svartvik (1992), s. 105-122.

Leech, G. (2007a). Corpora. I: Teubert & Krishnamurthy (2007), bind 2, s. 3-17.

Leech, G. (2007b). The Value of a Corpus in English Language Research. I: Teubert & Krishnamurthy (2007), bind 1, s. 315-325.

Léon, J. (2007). Claimed and Unclaimed Sources of Corpus Linguistics. I: I: Teubert & Krishnamurthy (2007), bind 1, s. 326-341.

LINDSEI (2010). Lokalisert 1. oktober 2010, på http://www.uclouvain.be/en-cecl-lindsei.html.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Lüdeling, A., Evert, S. & Baroni, M. (2005). Using web data for linguistic purposes. Lokalisert 19. september 2007, på http://sslmit.unibo.it/~baroni/publications/WAC-LuedelingEvertBaroni.pdf.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Nacey, S. (2011). Comparing linguistic metaphors in L1 and L2 English, Oslo: Det humanistiske fakultetet, Universitet i Oslo.

Nesselhauf, N. (2004). Learner Corpora and Their Potential for Language Teaching. I: J. M. Sinclair (red.), *How to Use Corpora in Language Teaching* (s. 125-152). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Oxford English Dictionary (1928). Oxford: Oxford University Press.

Pragglejaz Group (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol* 22(1), s. 1-39.

Pullum, G. (2007). Ungrammaticality, rarity, and corpus use. *Corpus Linguistics and LinguisticTheory* 3, s. 33-47.

Renouf, A., Kehoe, A. & Banerjee, J. (2007). WebCorp: an integrated system for web text search. I: Nesselhauf, C., Hundt, M. & Biewer, C. (red.), *Corpus Linguistics and the Web* (s. 47-68). Amsterdam: Rodopi.

Saussure, Ferdinand de. (1968 [1916]). *Cours de linguistique générale*. (R. Engler, red.). Wiesbaden: Harrassowitz.

Sinclair, J. (2005). Corpus and Text - Basic Principles. I: M. Wynne (red.), *Developing Linguistic Corpora: a Guide to Good Practice* (s. 1-16). Oxford: Oxbow Books. Lokalisert 7. oktober 2010, på http://ahds.ac.uk/linguistic-corpora/.

Sinclair, J. (2007). Intuition and Annotation. I: Teubert & Krishnamurthy (2007), bind 2, s. 415-435.

Steen, G., Biernacka, E., Dorst, L., Kaal, A., López-Rodríguez, I., & Pasma, T. (i trykk). Pragglejaz in Practice: Finding Metaphorically Used Words in Natural Discourse. I: G. Low, L. Cameron, A. Deignan & Z. Todd (red.), *Metaphor in the Real World.* Amsterdam/Philadelphia: John Benjamins.

Steen, G., Dorst, L., Herrmann, B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A Method for Linguistic Metaphor Identification: From MIP to MIPVU.* Amsterdam/Philadelphia: John Benjamins.

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

Svartvik, J. (red.). (1992). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin, New York: Mouton de Gruyer.

Taylor, C. (2008). What Is *Corpus Linguistics*? What the Data Says. *ICAME Journal: Computers in English Linguistics* (32), s. 179-200.

Teubert, W. & Krishnamurthy, R. (red.). (2007). *Corpus Linguistics: Critical Concepts in Linguistics*. London: Routledge.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: Benjamins.

VARIENG (s.a.). The English-Norwegian Parallel Corpus Bibliography. Lokalisert 14. november 2010, på http://www.helsinki.fi/varieng/CoRD/corpora/ENPC/bibliography.html.

Wible, D., Kuo, C.-H., Chien, F.-y., Liu, A., & Tsao, N.-L. (2001). A Web-Based EFL Writing Environment: Integrating Information for Learners, Teachers, and Researchers. *Computers & Education* 37, s. 297-315.

Omtalte korpus

AKS – Norsk andrespråkskorpus http://ask.uib.no/

BNC – The British National Corpus http://www.natcorp.ox.ac.uk/

ENPC – Engelsk-norsk parallellkorpus http://www.hf.uio.no/ilos/english/services/omc/enpc/

ICLE – The International Corpus of Learner English http://www.uclouvain.be/en-cecl-icle.html

LBK – Leksikografisk bokmålskorpus http://www.hf.uio.no/iln/tjenester/sprak/korpus/skriftsprakskorpus/lbk/

LINDSEI – The Louvain Database of Spoken English Interlanguagehttp://www.uclouvain.be/en-cecl-lindsei.html

LOB – The Lancaster-Oslo/Bergen corpus http://khnt.hit.uib.no/icame/manuals/lob/index.htm

LOCNESS – The Louvain Corpus of Native English Essays http://www.uclouvain.be/en-cecl-locness.html

Norsk aviskorpus http://www.hit.uib.no/aviskorpus/

OMC – Oslo Multilingual Corpus http://www.hf.uio.no/ilos/tjenester/sprak/omc/

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)

SEU – The Survey of English Usage http://www.ucl.ac.uk/english-usage/

The Brown University Corpus of American English
http://icame.uib.no/brown/bcm.html

Graedler, A-L., Egan, Thomas & Nacey, S. (2011). Korpus og korpusananlyse i språkdidaktisk forskning og praksis [Corpora and corpus analysis in language didactic research and practice]. In P. Dyndal, T.O. Engen & L.I. Kulbrandstad (Eds.), *Lærerutdanningsfag, forskning og forskerutdanning. Bidrag til kunnskapsområder i endring* (pp 101-127). Vallset: Opplandske bokforlag. (Author's copy)