Janne Bondi Johannessen (red.), Sonja Erlenkamp,
Jan Terje Faarlund, Elizabeth Lanza
og Arnfinn Muruvik Vonen

På språkjakt

– problemer og utfordringer i språkvitenskapelig datainnsamling

UNIPUB FORLAG
2003

Innsamling av språklige data: Informanter, introspeksjon og korpus

Janne Bondi Johannessen

1 Språkvitenskapen - en empirisk vitenskap

Språkvitenskap er en empirisk vitenskap, og ofte må språkforskeren selv skaffe de dataene som er nødvendige. Det kan man gjøre ved å bruke informanter, introspeksjon, eller man kan bruke eksisterende materiale. For mange språk er det blitt mulig å bruke elektroniske korpus, hvor man effektivt kan få store mengder språkeksempler. Dette har fått stor verdi for mange typer språkforskning, men kanskje særlig for sosiolingvistikk og diskursanalyse, hvor bruk av introspeksjon alene er umulig.

Uansett innsamlingsmetode, kan man si at språkforskeren må drive feltarbeid, i den forstand at vedkommende må ut i felten og skaffe seg data. Det er bare en forskjell på hvor langt man må reise. For å bruke egen introspeksjon trenger man ikke rikke seg av flekken, for å bruke andres introspeksjon, dvs. informanter, må man kanskje ut av huset, og kanskje til den andre siden av jorda. Og bruk av tekstkorpus forutsetter en datamaskin, hvor den nå måtte befinne seg. I det følgende skal vi se på alle tre metodene, og foreta noen sammenligninger. Vi skal først se på noen problemstillinger knyttet til informanter og introspeksjon, før vi ser på bruk av tekstkorpus som kilde for språklige data.

2 Informanter og introspeksjon

Innsamling ved hjelp av informanter kan brukes for mange formål, og er nødvendig når forskeren ikke selv kjenner studiespråket godt. Vanskelighetene med utvalg av informanter er beskrevet mange steder (som påpekt i de andre kapitlene i denne boka), og er slett ikke bare av nyere dato. Noen slike erfaringer er gjengitt nedenfor. De første er av den store norske språkforskeren Georg Morgenstierne, som utførte feltarbeid i Nordvest-India, Afghanistan og Iran de første tiårene av forrige århundre. Sitatene nedenfor er fra den svært leseverdige boka *På sprogjakt i Hindukush*. Morgenstierne hadde en tjener, Yasin, som var ham til stor nytte:

> «Så har det vært en god del individer som Yasin har fisket opp. Noen viste seg straks å være ubrukbare, andre underkastes en kortere eller lengre eksaminasjon for å få rede på de viktigste eiendommeligheter i deres dialekt. Det er ikke alltid lett å gå den like vei til målet, men med Yasins hjelp går det i alminnelighet. Hans begripelse av sprogvitenskap begynner å vokse: han skjønner det gjelder å få tak i eldst mulige individer fra avsides landsbyer. Men han har en tilbøyelighet til å tro at foruten alder også skavanker som enøyethet og tunghørthet, eller til og med tannløshet, øker den lingvistiske verdi av individet.» (Morgenstierne 1992, s. 12-13)

Men tjeneren hadde også sine egne oppfatninger, som kanskje var litt andre enn språkforskerens:

> « – Som prøve talte jeg inn en Baloshi-fortelling i en litt sprukken grammofonplate. Yasin syntes den ble så meget tydeligere enn dem som de innfødte hadde talt inn, at han foreslo at jeg selv skulle tale inn alle dialekter på denne måten, og slippe bryderiet med å la de innfødte gjøre det.» (Morgenstierne 1992:12-13).

Når man arbeider med levende informanter, må man også være oppmerksom på personlige faktorer:

> «Den yngre av mine ledsagere taler godt Pashto, og er ganske intelligent. Han pumper meg iherdig om Norge, og jeg svarer som best jeg kan. Men efter overveielse sløyfer jeg midnattsolen! Jeg vil ikke forstyrre det vakre tillitsforhold som gror opp mellom oss!» (Morgenstierne 1992:102)

Elizabeth Lorimer, kona til den store lingvisten David Lorimer som skrev om språket burushaski, som ikke er beslektet med noen andre i Kashmir/Afghanistan-regionen, har skrevet om deres erfaringer i boka *Language Hunting in the Karakoram*. Hun har dette rådet til hvordan man kan plukke ut informanter - kanskje ikke så lett å følge i våre dager, når forskerne stort sett har mindre å rutte med:

> «You first select a few intelligent men who speak their language well and clearly, and work with them until you light one of the best suited to your purpose. You then annex him to the household at a fixed, generous salary, so that he will always be available when wanted, whether at home or on tour.» (Lorimer 1989: 248)

Når det gjelder språk som forskeren selv ikke kjenner godt, og hvor det finnes få skriftlige kilder, er det påkrevet å gjøre feltarbeid med informanter.

(1) Situasjon hvor informanter er nyttig:
Språket er ukjent; man trenger data
Språket er kjent, men ens egen intuisjon er vaklevoren
Språket er kjent, men det finnes variasjon man ønsker å kartlegge

Men det kan være et problem, slik sitatene ovenfor viser, at ikke alle informanter er like enkle. De kan ofte ha oppfatninger om sin egen språkbruk som ikke stemmer overens med hvordan de faktisk benytter språket. De kan for eksempel ha normative oppfatninger som det viser seg i praksis at de ikke følger. Selv har jeg for eksempel trengt informanter i samband med forskning på sammensatte ord i norsk (Johannessen 2001). Jeg trengte da eksempler fra andre dialekter enn min egen. I utgangspunktet burde dette være enkelt å få til. Sammensetning er en produktiv ordlagingsprosess i alle norske dialekter, så jeg kunne jo bare spørre informantene om å lage noen ord etter mine anvisninger. Men for å være sikker på å få nylagde sammensetninger, og ikke slike som alt var leksikaliserte, måtte jeg finne på noen som kanskje ikke var så sannsynlige pragmatisk sett. I det gitte tilfellet var jeg interessert i hvordan infinitiv kunne settes sammen med et substantiv. Informanten ga mange fine eksempler på ord som allerede tydelig var leksikalisert (som dialektens ord for *fiskekrok* og *stekepanne*). Men mine mer kreative forslag avfødte kommentarer av typen: «Nei, det har jeg ikke hørt før». Det tok litt tid før jeg fikk gjennomslag for at det ikke var betydningen jeg var ute etter, men formen på det nye sammensatte ordet. Og da gikk det heldigvis fint, jeg fikk det jeg trengte.

Hvis man først har kommet over problemene med å skaffe riktig informant, er det mange fordeler med å benytte informanter. Det viktigste er kanskje at informantene er

mennesker (og ikke bøker eller datamaskiner), slik at man i beste fall kan ha toveis kommunikasjon med dem. Det betyr ikke minst at man kan stille tilleggsspørsmål om det er noe som er uavklart. Det betyr også at man kan stille negative spørsmål, ved f.eks. å konstruere antatt ugrammatiske setninger, og få informantenes reaksjon på dem. Et problem er på den annen side at denne toveiskommunikasjonen kan føre til at man påvirker informanten i en gitt retning som passer med ens hypoteser.

Det vi har sagt, gjelder bare dersom informanten og forskeren utveksler informasjonen gjennom dialog ansikt til ansikt. En annen måte å bruke informanter på, er ved f.eks. å benytte spørreskjemaer. (Det forutsetter et visst utdanningsnivå hos informantene.) Spørreskjemaer kan unngå noen av fallgruvene ovenfor, med personlig påvirkning av informanten. Men man mister samtidig muligheten til å foreta en direkte oppfølging, selv om man selvsagt kanskje kan analysere svarene fra spørreskjemaet og formulere nye spørreskjemaer som informantene kan få. Men den manglende direkte oppfølgingen som spørreskjemaene innebærer, fører til at man ikke får kontrollert om skjemaene tolkes riktig.

La oss ta et eksempel. Denne forfatteren har bl.a. forsket på noe som gjerne kalles *negative polaritetsuttrykk* (Johannessen 1998a), altså ord og uttrykk som svært forenklet sagt bare kan forekomme i sammenheng med ordet *ikke* eller andre ikke-veridikale kontekster. Et par eksempler er vist nedenfor (leseren kan selv bedømme hvorvidt slike setninger er grammatiske dersom man fjerner det som gjør dem ikke-veridikale, som ordet *ikke* nedenfor):

(2) a. Petra hadde ikke sett *noen ting*.
b. Petra forsto ikke *en døyt*.

Dersom jeg hadde vært interessert i å teste hva slags nominalfraser som aksepteres av språkbrukere flest, kunne jeg gitt informantene følgende skjema:

(3)

Er nominalfrasene nedenfor akseptable?	**Sett kryss ved det svaralternativet som passer best.**		
	Ja	**Nei**	**Vet ikke**
tre jenter			
fine leker			
stygge leke			
flott gutter			
noen kopp			
søte unger			

Det er alltid stor fare for at informanter ikke kommer på alle mulige slags kontekster, inkludert negative. Et resultat av spørreskjemaet over kunne faktisk vært at de hadde markert *noen kopp* som ugrammatisk, til tross for at de kanskje daglig bruker negative polaritetsuttrykk av denne typen:

(4) a. Jenta har ikke *noen kopp*
b. Gutten kjøpte ikke *noen dukke*
c. Vi har ikke *noen hytte*

Spørreskjemaer er altså noe man må bruke med den ytterste forsiktighet. Det er jo alltid vanskelig å vite om eller hva informantene har misforstått når man ikke har den personlige kontakten med dem. Dialogsituasjonen er kanskje å foretrekke. Man bør forberede seg så godt på forhånd, og stille spørsmålene på så mange måter, at faren for ubevisst å påvirke informanten, minsker.

Strengt tatt er det man gjør når man bruker informanter, å be personene bruke introspeksjon, altså sin språklige

intuisjon. Slik sett er det mye av det som blir sagt her, som gjelder begge metoder. Men begrepet introspeksjon brukes gjerne om den situasjonen der forskeren spør seg selv hvilke grammatiske konstruksjoner som er akseptable og uakseptable, gjerne ved å lage setninger eller fraser som man så smaker litt på. Denne metoden har, og har i årtier hatt, enorm utbredelse, særlig i den teoretiske lingvistikkforskningen, hvor det lenge var slik at man kun forsket på sitt eget språk (særlig engelsk). Å bruke introspeksjon er likevel svært problematisk. Fra et vitenskapelig synspunkt feiler metoden på nærmest alle punkter: a) Det er umulig for andre forskere å etterprøve dataene, b) siden forskeren strengt tatt bare bruker data fra sin egen idiolekt, blir det meningsløst å argumentere mot riktigheten av dataene, 3) det er vanskelig å si hvor generelle hypotesene og teoriene som er dannet på grunnlag av dataene kan være, 4) forskeren går lett i den fella at han eller hun farger evalueringen av dataene ut fra hypotesen eller sitt teoretiske ståsted, 5) mange, også forskere, har en feilaktig oppfatning av hvordan ens egen språkbruk egentlig er, 6) mangelen på input fra andre gjør at man overser andre relevante data.

Et godt eksempel på faren med bruk av introspeksjon, er dette utsagnet fra en, faktisk erfaren, grammatikkforsker i en diskusjon som gjaldt om det på norsk, i motsetning til engelsk, gikk an med et tomt subjektsspor etter subjunksjonen *at*. Dette sa forskeren:

(5) Nei, et tomt subjekt på plassen etter at, mener jeg at _ ikke går på norsk.

Det er et talende eksempel på gapet mellom ens oppfatning av egen språkbruk og den faktiske språkbruken. Ellers vrimler det av eksempler på bedømmelser basert på introspeksjon i den lingvistiske litteraturen. Her er et nyere eksempel, fra en (svært god) bok om adverbplassering av Øystein Nilsen:

(6) a. ... at Per ikke lenger forstår problemet ALLTID HELT (Nilsen 2000:123c)
b. ... at Per ikke forstår problemet LENGER ALLTID HELT (Nilsen 2000:123d)
c. * ... at Per ikke forstår problemet ALLTID LENGER HELT (Nilsen 2000:123f)
d. * ... at Per ikke lenger forstår problemet HELT ALLTID (Nilsen 2000:123g)

For mange vil nok dette eksemplet (hvor de to første setningene skal være velformet og de to siste ikke velformet) vise at bruk av introspeksjon kanskje ikke gir robuste resultater. På den annen side viser dette eksemplet også noe annet, nemlig at enkelte ganger er problemstillingen av en slik art at det vil være umulig å finne data i eksisterende materiale, både fordi konstruksjonene beveger seg i grenseland for hva som er mulig, og fordi man selvsagt ikke kan finne negative data.

Introspeksjon (enten den er lingvistens egen eller informantens) er altså nødvendig når formålet er å vurdere yttergrensene for hva grammatikken tillater. Et annet punkt hvor man vanskelig kan klare seg uten introspeksjon, er vurderinger av strukturell flertydighet. Mye semantikk-forskning dreier seg om kvantorrekkevidde, for eksempel, og det ville være nesten umulig å finne grensene for hva slags tolkningsmuligheter som finnes, ved utelukkende å bruke korpus. At setningen nedenfor er tvetydig (allkvantoren kan tolkes med lang eller kort rekkevidde i forhold til negasjonen), går ikke fram av formen selv, det må enten kontekst eller introspeksjon til for å avgjøre det:

(7) Alle mennesker trenger ikke legehjelp i dag.

Et korpus kan nok gi nok kontekst til å skaffe en visshet om én av betydningene, men det ville nok være vanskelig å finne kontekstuell støtte for begge betydninger.

3 *Tekstkorpus*

3.1 Bakgrunn

De siste par tiårene har elektroniske korpus blitt en mer og mer brukt metode for å finne lingvistiske data (se f.eks. McEnery and Wilson 1996). En grei beskrivelse av hva et korpus er, finner vi i dette sitatet: «A computer corpus is a body of texts put together in a principled way and prepared for computer processing.» (Johansson 1998:3). Det er flere måter å klassifisere korpus på, etter medium, tekstlig innhold, type annotering, antall språk osv. I oversikten nedenfor skal vi si noe om skriftspråks- og talespråkskorpus, samt flerspråklige korpus.

En egenskap ved korpus som kommer i tillegg til det som er sagt i sitatet, er at korpuset ofte er knyttet til et søkesystem. Det muliggjør søking på ord eller ordstrenger, eventuelt også grammatiske kategorier, og resultatene av søket i korpuset vises gjerne som en konkordans, dvs. en alfabetisk liste over søkeuttrykkene med noe kontekst og gjerne kildehenvisning. Vi skal se flere eksempler på konkordanser senere i kapitlet.

De absolutt fleste tekstkorpus er basert på skriftspråk. En annen type korpus er talespråkskorpus, som gjerne består av transkribert tale. Transkripsjonen kan være rent ortografisk eller mer eller mindre fonetisk. Søkesystemet for et talespråkskorpus vil i tillegg til den skriftlige konkordansen ofte gi muligheten til å gå videre fra en bestemt konkordanselinje og til det korresponderende lydopptaket eller videoopptaket (eksempel på det førstnevnte er Talesøk ved Universitetet i Bergen, og på det sistnevnte Big Brother-

korpuset i Oslo). En tredje type er parallellkorpus, dvs. korpus som består av tekster fra to eller flere språk, hvor tekstene er oversettelser av hverandre. Søkesystemet for et flerspråklig korpus har gjerne et mer avansert søkesystem, hvor man kan søke på ulike måter for ett eller flere av språkene i korpuset (et eksempel er Oslo Multilingual Corpus). I referanselista bak gis en oversikt over de viktigste tale- og skriftspråkskorpusene for norsk, svensk og dansk.

3.2 Skriftspråkskorpus

De langt fleste korpus i verden inneholder tekster hentet fra skriftspråk. Årsaken til det er selvsagt at det er enkelt å få tak i slike tekster, og dermed kostnadseffektivt å utvikle slike korpus. Antallet ord som finnes i ulike korpus varierer veldig, f.eks. har Universitetet i Oslos bosnisk-korpus ca. en million ord, mens Oslo-korpuset av taggede, norske tekster inneholder nærmere 20 millioner ord. Aviskorpuset ved HIT-senteret i Bergen har for lengst passert 200 millioner ord. (Til sammenligning har en gjennomsnittlig roman ca. 100 000 ord.)

Mange korpus er grammatisk tagget (det vil si merket eller annotert), med kategorier som ordklasse, bestemthet, tall og så videre. Det gjelder for eksempel Oslo-korpuset av taggede, norske tekster. Hva slags tekster de ulike korpusene inneholder, varierer. Det kommer vi tilbake til mot slutten av dette kapitlet. Men en ting som de fleste korpus har til felles, er et søkesystem som gjør dataene tilgjengelige. I våre dager er korpusene ofte tilgjengelige på Internett, noe som gjør dem lettere å finne og å bruke, men det er fremdeles enkelte som selges på CD-rom, som oftest for å kontrollere at brukerne betaler for seg, men det kan også være for at korpusutviklerne vil sikre seg at korpuset teknisk sett virker som det skal.

Når et korpus er tagget, øker det mulighetene for hvordan man kan søke i korpuset. Det vil ikke lenger bare være

sekvenser av ord og ordstrenger som er søkbare, men også (sekvenser av) grammatiske kategorier, eller kombinasjoner av grammatisk kategori og ordstrenger. Ofte kan man også spesifisere undergrupper av korpuset, f.eks. bare de tekstene som er skrevet av en bestemt forfatter, eller bare romaner. Når søkemulighetene på denne måten øker, er det viktig for brukeren (språkforskeren) at korpuset er lett å søke i. Når det gjelder norsk, er det foreløpig bare ett korpus som er søkbart på så mange plan, nemlig Oslo-korpuset. Brukervennligheten er der forsøkt ivaretatt (se Johannessen, Nøklestad og Hagen 2000), men det er fremdeles flere ting som kan utsettes på det (se Asmussen 2000).

3.3 Talespråkskorpus

Mens det finnes korpus med skriftspråkstekster for mange språk, er det dessverre langt færre talespråkskorpus. Årsaken til dette er åpenbar. Det er svært mange problemstillinger knyttet til å samle inn taledata. Personene som er med, bør ha gitt sitt samtykke. Ved opptak bør de likevel ikke endre talemålet, dvs. at de ikke bør la seg påvirke av en mikrofon eller et videokamera som står midt i mot. Skal man ha spontantale, er det fristende å gjøre opptak i folks dagligsituasjoner, men da kan det lett bli en konflikt mellom forskerens ønske om fullstendige data og personenes ønske om anonymitet og diskresjon. Enkelte forskere er likevel innstilt på at et talespråkskorpus bør inneholde forskjellige talesituasjoner, ikke bare «naturlige», private kontekster, men også prekener, formelle radio-intervjuer, foredrag, møter osv. I Gøteborg har man f.eks. lagt vekt på å inkludere mange forskjellige aktiviteter i sitt talespråkskorpus, og har så langt taledata fra 25 talesituasjoner i GSLC (The Gothenburg Spoken Language Corpus).

Big Brother-korpuset, som er basert på TV-serien, er under oppbygging ved Universitetet i Oslo. Dette er et

talespråkskorpus som unngår en del av de opplagte problemene som er nevnt ovenfor. Personene har gitt sin tillatelse til opptak, men de har ikke vært klar over at opptakene ville bli brukt til språkforskning. Videre har de blitt så vant til mikrofoner og kameraer at man må anta at talespråket, som resten av oppførselen, etter hvert har blitt mer avslappet. Et korpus av denne typen gir informasjon som man vanskelig kan finne på andre måter, verken ved formelle opptak og intervjuer av enkeltinformanter eller ved spørreskjemaer og skriftpråkskorpus: språk brukt for å uttrykke sinne, kjærlighet, hat, irritasjon, vennskap, flørt, kort sagt følelser av mange slag. Videre vil det være mye dialog med tilhørende mulig interessante språkhandlinger som spørsmål, ordrer, fortellinger, samt turtaking, overlappende tale og andre diskursrelaterte problemstillinger. Selvsagt er det også ulemper knyttet til et korpus som Big Brother-korpuset. For det første er det umulig å bruke det som et dialektkorpus, siden deltagerne kommer fra forskjellige steder og antagelig tilpasser språket sitt til de andre. For det andre er det mange temaer som ikke diskuteres, siden deltagerne ikke har lov å omtale personer utenfor de fire veggene der de oppholder seg. Slik blir det temamessig og dermed også vokabularmessig innsnevret.

Ovenfor nevnte vi at en grunn til at talespråkskorpus er vanskelige å lage, er problemene som er knyttet til deltagerne. Et annet hovedproblem som gjelder talespråkskorpus, er at de er svært dyre å produsere. I tillegg til muntlige opptak (og videoopptak), er det ønskelig at opptakene er transkribert, så man kan lese i dem og søke i dem. Transkripsjon må gjøres manuelt, og blir ekstremt kostbart, selv når man velger standardortografi og ikke fonetisk transkripsjon eller spesialtranskripsjon som viser bestemte diskurstrekk.

3.4 Flerspråklige korpus

Flerspråklige korpus er som regel basert på skrift, selv om det ikke prinsipielt behøver å være slik. Johansson (1998:4-5) regner opp flere typer. Den ene typen kan man kalle sammenlignbare korpus, hvor tekstene er av samme genre, selv om de er hentet fra flere språk. Den andre typen er oversettelseskorpus. Disse er igjen av flere typer: 1) originaltekster på ett språk, og oversettelser fra et annet, 2) originale tekster og oversettelser innenfor samme språk, 3) oversatte tekster på forskjellige språk. Johansson foreslår å entydiggjøre terminologien ytterligere, slik at kategori 2) ovenfor bør benevnes enspråklig sammenlignbart korpus. Oslo Multilingual Corpus er et oversettelseskorpus, som dekker alle kategoriene. I tillegg til søkegrensesnittet har korpuset også blitt bearbeidet med et parallellstillingsprogram, slik at det finnes en lenke mellom hver originale setning og dens oversettelse. Det muliggjør interessante søk av typen: «Finn alle setninger hvor ordet *ikke* finnes i originalspråket, men er oversatt til noe annet enn *not*.»

Flerspråklige korpus gir muligheter til andre typer undersøkelser enn enspråklige korpus. Johansson (1998:14) viser til en undersøkelse gjort av ham selv og Berit Løken om den norske partikkelen *nok* i norske og engelske oversettelser. Det viser seg at det aller vanligste er å la denne partikkelen forbli uoversatt i engelsk. Men enda mer slående er det at når man finner *nok* i en oversettelse fra engelsk, viser det seg at det ikke er noe den er oversatt fra! Dette sier mye om partiklenes rolle i ulike språk, og er et godt utgangspunkt for videre studier i for eksempel informasjonsstruktur. Det er ikke tvil om at flerspråklige korpus vil muliggjøre mye interessant forskning innen språktypologi og kontrastiv lingvistikk, i tillegg til at de selvsagt også kan brukes til å belyse språklige forhold i enkeltspråk.

3.5 Andre tekstsamlinger

Det er ikke alle tekstsamlinger som kommer inn under betegnelsen «korpus», slik det ofte brukes i dag, og slik det er tenkt i dette kapitlet. For eksempel var det tidligere vanlig å kalle enhver stor samling av tekster for et korpus. Definisjonen i Bokmålsordboka (nettutgaven) er typisk: *innsamlet materiale*. Her reserverer vi korpusbetegnelsen for de tekstsamlingene som er bearbeidet for elektronisk tilgang, i tillegg til at de er satt sammen på en gjennomtenkt måte. Hvor tekstene kommer fra, om de er gamle eller nye, på et dødt eller levende språk osv., er ikke avgjørende for om det skal kalles et korpus eller ikke. Således finnes for eksempel alle gammelgreske tekster tilgjengelige i korpuset *Thesaurus Linguae Graecae*. På den annen side er tekster fra eldre varianter av norsk ikke å betrakte som et korpus i vår forstand, fordi de ikke ligger tilgjengelige i et søkbart korpus. Likevel er det, som vi skal se, noen likheter mellom å bruke elektroniske tekstkorpus og å bruke trykte tekster. Likhetene har selvsagt å gjøre med at man ikke har muligheten til å inngå i dialog med noen informant, med de mulighetene og begrensningene dette innebærer.

Videre kan man spørre hva som er forskjellen mellom på den ene siden lydopptak av talespråk eller videoopptak av døves tegnspråk, og på den andre et talespråkskorpus. Igjen blir det definisjonen av elektronisk tilgjengelighet sammen med et søkesystem, som gjør utslaget. Men selvsagt er det slik at video-opptakene som nevnes i Erlenkamps kapittel har mye til felles med både tekstene som nevnes i Faarlunds kapittel og korpusene i det herværende kapitlet. Igjen er det slik at opptak skaper en distanse til språkutøverne, og gjør det vanskelig eller umulig å gå inn i dialog med dem. Samtidig er det egenskaper ved opptakene av døve, som gjør dem like informantdata, slik som muligheten for å gå i dialog på et senere tidspunkt.

I det videre skal jeg ikke komme spesielt inn på disse to typene av datamateriale, men leseren vil nok lett se hvor slike data har likheter og ulikheter med de andre typene datagrunnlag som diskuteres i dette kapitlet: introspeksjon, informanter og korpus.

3.6 Noen egenskaper ved korpus

Et problem ved korpus er at det er vanskelig å stille negative spørsmål til det. Dersom man ikke finner belegg for en gitt konstruksjon, skal man da anta at den er ugrammatisk? Det kan være mange grunner til at noe ikke finnes i et korpus. Kanskje er konstruksjonen svært muntlig, eller kanskje er den vanskelig å prosessere, eller kanskje er den så merkelig pragmatisk sett at ingen noen sinne ville finne på å ytre den frivillig? (Se også kapittelet til Faarlund i denne boka for en diskusjon av problemet.)

Bruk av korpus stiller en overfor akutte filosofiske problemstillinger. Hva er det egentlig en som språkforsker prøver å si noe om? Er det, for å si det med Chomsky (1986), i-språk eller e-språk (*internal language* eller *external language*)? Hvis vi nemlig er interessert i universalgrammatikk, er vi egentlig på utkikk etter hvert enkelt menneskes grammatikk (i-språk), og det ville være nærmest katastrofalt om vi begynte å blande sammen flere menneskers språk – særlig hvis de er ulike, og vi betrakter dataene som om de var et uttrykk for samme i-språk. Ut fra dette perspektivet er bruk av korpus faktisk problematisk. Bare i de færreste tilfeller er korpusene organisert på en slik måte at vi vet hvem som er den språklige kilden bak enhver tekst eller ethvert tekstfragment. Det finnes ikke noe korpus som kan sies å være representativt for alle språkbrukere, verken ut fra i-språksperspektiv eller, for den sakens skyld, e-språksperspektiv. Selv i prinsippet ville det være umulig å konstruere et korpus som var representativt i absolutt forstand. For det språket som omgir og uttrykkes av en

filosofistudent fra Trondheim, vil nødvendigvis være et annet enn det som gjelder en baker i Skjeberg eller en bilmekaniker i Alta. For at korpusbruk skal forsvares av en i-språksforsker, er det nødvendig med en god porsjon kritisk sans og helst tilleggsundersøkelser med bruk av informanter og introspeksjon.

Når det er sagt, er det likevel mange korpusutviklere som bestreber seg på at korpuset i det minste skal være balansert med hensyn til teksttype og tekstkilde, slik at det skal representere flere genrer og en del variasjon. Da kan man håpe å finne det som er felles for flest mulig, samt at det blir flere muligheter for å finne eksempler på ulik språkbruk og ulike grammatiske konstruksjoner. Oslo-korpuset av taggede, norske tekster er et eksempel på et korpus hvor det ble gjort noen bestrebelser på balanse: her er det representert både avis- og ukebladtekster, skjønnlitteratur og sakprosa. Dessverre er det ofte praktiske, og særlig juridiske, hindringer i veien for å utvikle balanserte korpus. Problemene er knyttet til forfatterrettigheter og en viss bekymring for at noen skal kunne utnytte tekstene kommersielt. Resultatet av dette er at mange av de største og mest brukte korpusene i verden, er rene «news corpora» med materiale hentet fra avistekster (se katalogen hos Linguistic Data Consortium for en verifikasjon av påstanden). Det sier seg selv at det er uheldig for språkforskningen, og for den sakens skyld for utvikling av språkteknologi for et språk, når det blir avisspråket som utgjør datagrunnlaget for språkforskningen.

(8)

Tekstkorpus

Fordeler:

- Mulighet for store datamengder
- Unngå at datainnsamlingen kan farges av forskerens hypoteser
- Nye og overraskende data kan dukke opp
- Mulighet for å spesifisere detaljerte grammatiske søkeuttrykk

Problemer:

- Vanskelig å formulere spørsmål om negative data
- Kan være vanskelig å følge opp spørsmålene mer inngående
- Kan være vanskelig å få data om marginale konstruksjoner
- Korpuset inneholder en endelig mengde data.
- Dataene er oftest tidsavgrenset
- Man kjenner ikke alltid den språklige bakgrunnen til forfatterne av korpustekstene

4 Bruk av korpus

I dette kapitlet skal jeg illustrere en del mulige bruksområder for korpus, basert på dels min egen og dels andres forskning. Korpus kan benyttes både kvantitativt og kvalitativt. Det er viktig å understreke at bruk av korpus ikke er det samme som å bruke kvantitative metoder. Man behøver slett ikke bruke korpuset kvantitativt, men heller bruke det som et hav man kan fiske ut passende fisk fra.

Gode korpus for lingvistisk bruk har gjerne utstrakt annotering som muliggjør grammatisk formulerte spørsmål. La oss si at jeg er interessert i sammensatte ord i norsk. Som vi vet, er det ikke alltid lett å få passende svar fra levende informanter. Men korpuset, som jo kan betraktes som data fra «tause» informanter, er ikke forutinntatt, og gir oss nøyaktig det vi ber om, i store mengder. Ved å søke etter alle produktivt

sammensatte ord i Oslo-korpuset, fikk jeg straks store mengder resultater. Noen av dem er vist nedenfor:

(9)

Et lite utdrag av treff ved søk etter sammensatte ord i Oslo-korpuset:

AV/Ad96/01: for at resten av de rwandiske **hutu-flyktningene** i Zaire og Tanzania
AV/Ad96/01: ed det amerikanske forsvarets **Europa-hovedkvarter** utenfor Stuttgart i
AV/Ad96/01: og Zaire ved Gisenyi mandag. **Nødhjelparbeidere** mener rundt 350
AV/Ad96/01: steren, tok de med seg andre **stjørdalselever** og overleverte
AV/Ad96/01: tjørdalselever og overleverte **protestunderskrifter** til fylkesordfører
AV/Ad96/01: ansen. I forrige uke troppet **skatvalselevene** opp på Trondheim
AV/Ad96/01: Jagland var på vei hjem fra **samrådsbesøk** på Melhus. Elevene ba
AV/Ad96/01: Med seg har han blant annet **NTNU-professor** Helge Nøsterud, som
AV/Ad96/01: ertid allerede i gang. Under **plasthopprennet** i Granåsen i august, ble
AV/Ad96/01: r Randi sikret tre flasker av **jubileumsakevitten** i går. Hun mente det
AV/Ad96/01: p eksisterer det heller ingen **«slankekontrakter»**, men han kjenner til
AV/Ad96/01: forskning på utviklingen av **hoppteknikk** - vil ikke uten videre
AV/Ad96/01: teen i FIS, et forslag om en **omregningstabell** for lengden på skiene i
AV/Ad96/01: ngden på skiene i forhold til **kroppslengde**. I dag sier regelverket at
AV/Ad96/01: ot Forslaget om regulering av **skilengden** ble tatt opp i FIS-systemet
AV/Ad96/01: Realfagbygget og RiT 2000 : **Milliardkake** til fordeling | ASBJØRN
AV/Ad96/01: illiarder kroner investeres i **gigantprosjektene** Realfagbygget og RiT

Når noe er så frekvent som sammensatte ord, har man virkelig muligheten til å finne massevis av eksempler, slik at man har et godt grunnlag for å gå videre med spørsmål som: Hva slags typer sammensetninger har fuge-s og hva slags har fuge-e, og hva slags har aldri fuge? Hva er det semantiske forholdet mellom forleddet og etterleddet i en sammensetning? Hvilke syntaktiske kategorier forekommer i sammensetninger på norsk? De mange eksemplene på bruk man får ved korpussøking, gjør det mulig å foreta generaliseringer man ellers ville ha problemer med.

Korpus kan også brukes til å teste ens egen intuisjon om grammatiske fenomener. La oss ta et eksempel. I det siste har det kommet ut grammatikkbøker (Faarlund et al. 1997 og Endresen og Simonsen 2000) hvor det sies at forleddet i sammensetninger enkelte ganger kan være bøyd (fordi det finnes eksempler som *kaldtvann*, hvor det ser ut til å være kongruens i tall og kjønn mellom forleddet og etterleddet). De to bøkene er ulike i det de sier om hvorvidt bøying i forledd mer generelt er produktivt, den sistnevnte heller mer mot dette enn den førstnevnte (op. cit. s.117). Men ingen av dem sier at bøying av forleddet er produktivt når det er et adjektiv. Både forfatterne av de nevnte bøkene og undertegnede har nok en intuisjon om at slik samsvarsbøying i norske sammensetninger nettopp ikke er det. Intuisjonen kan en teste ved for eksempel å søke på det bøyde *kaldt* som forledd sammenlignet med det ubøyde *kald*. (Mer om denne diskusjonen av sammensetninger kan man lese i Johannessen 2001.) Nedenfor vises et utdrag av

resultatene etter søk i Oslo-korpuset. Det klart vanligste, og altså produktive, er at forledd ikke bøyes:

(10)

Kaldt **som forledd – 2 treff**

AV/Ad96/01: matkvalitet og det faktum at **kaldtvannsfisken** vokser raskt på lave
SA/Lo84/01: samt isolasjonsmaterialer for **kaldtvannsrør,** behøver ikke være av

(11)

Kald **som forledd – 40 treff:**

AV/Ad96/01: eg husker fremdeles kvalmen, **kaldsvetten** og følelsen av forundring,
AV/Af94/01: g de skal utføres ved SINTEFs **kaldklimalaboratorium** i Trondheim.
AV/Af96/01: dmennene fikk en 20 minutters **kald-dusj** av ivrige og heldige
AV/Af96/01: 23,15 kroner fatet. Det var **kaldvær** i USA og store trekk i lagrene av
AV/BT95/03: Atom Vinter, en av de beste **kaldblodstraverne** som noensinne er satt
AV/BT95/03: port er delt inn i to raser «**kaldblods- og varmblodshesten**.
AV/BT95/05: r dagens gladeste da han vant **kaldblodsoppgjøret** bak Sagi Knut og
AV/BT96/01: ter det prinsippet som kalles **kaldbaking**. Det håndlages eksklusive

Samtidig skal man huske at korpusene er begrenset med hensyn til mengde, teksttype og tekstkilde, samt tidsperiode. Dersom man er ute etter nye bruksmåter, eller ord og ordformer som ikke nødvendigvis er så frekvente, må man lete andre steder enn i et begrenset tekstkorpus. La oss si at vi er

interessert i bøyingen av nylagde sammensatte ord. Et godt eksempel kunne da være ordet *e-post*. Dersom vi tror det kan finnes som tellelig ord, kan det være lurt å søke etter flertallsformen *e-poster*, men i Oslo-korpuset finnes det ikke. (Ordet *e-poster* vil være en nyskapning, vi bruker jo ikke det tilsvarende flertallsordet *poster* om de brevene som ikke er elektroniske.) Dersom det viser seg at et vanlig korpus er for lite, kan man forsøke et vanlig Internett-søk. Som korpus betraktet er jo Internett enormt, også når man begrenser seg til norske sider. Et søk på *e-poster* i søkemotoren Google ga 1180 norske treff, noe som tyder på at nye sammensetninger kan bøyes på en ny måte i forhold til det opprinnelige etterleddet.

(12)

Treff på flertallsformen *e-poster* – 1180 treff

Følgende e-poster er skrevet av dere som kunder.
(http://www.houseofsingles.com/lesebrev.htm)
Jeg har etterhvert mottatt en del E-poster om studielånsrente og om mulighetene til fastlånsrente fra dere.
(http://www.gamco.no/protest/sv_oystein_djupedal.htm)
Du kan motta alle de enkelte e-poster som blir sendt til «medlemsgruppen»
(http://www.swingers.no/email_chat_for_medlemer.htm)
Du kan også motta et resymé av de siste e-poster som har kommet inn.
(http://www.rehabdata.no/diskusjo.htm)

Man skal huske på at det ikke finnes noen grammatisk annotering av sidene på Internett. Derfor kan man ikke spesifisere søkene slik at man bare får det man er ute etter. For eksempel kan jeg ikke sikre meg at alle de 1180 treffene gjelder flertallsordet *e-poster*, og ikke entallsordet *e-poster* (synonymt med *plakat*, snarere enn det flertallsordet jeg var interessert i).

En viss kritisk sans ved bruk av Internett som korpus er derfor nødvendig.

En veldig viktig og god bieffekt av korpus for datainnsamling, er at man kan komme over data man ikke hadde tenkt på. Et eksempel er igjen fra min egen forskning på negative polaritetsuttrykk. Hvordan går man fram for å få en oversikt over hva slags negative polaritetsuttrykk som finnes på norsk? Man kan ikke spørre informanter om de kan ramse opp de polaritetsuttrykkene de bruker. Ei heller kan man søke etter polaritetsuttrykk som sådanne i et korpus, for siden dette har vært et blankt forskningsfelt for norsk, har jo ingen funnet på å tagge korpus med dem heller. Hva med introspeksjon? Ja, her kunne man komme et stykke på vei, men man begrenses jo av sin egen fantasi, og ikke minst av den grammatiske konteksten man som forsker gjerne har sett seg blind på. En utvei blir da å søke på et ord som *ikke* i et korpus, og så undersøke hvorvidt de resulterende setningene ville vært akseptable også uten negasjon. Er de ikke det, må det være et negativt polaritetsuttrykk i dem. Faglitteraturen om slike uttrykk har for øvrig vanligvis vært konsentrert rundt nominalfraser, tilsvarende de norske *noen kopp, en smule, en tråd, en muskel, en døyt*. Men søket i korpuset viser at det også finnes andre grammatiske konstruksjoner som må klassifiseres som negative polaritetsuttrykk. Her får man altså genuint ny viten som et resultat av korpusmetoden. (De negative polaritetsuttrykkene nedenfor er merket med understreking.)

(13)

Eksempler på negative polaritetsuttrykk funnet ved søk på ordet *ikke* i Oslo-korpuset

AV/Ad96/01: k investering. Den dagen det **ikke** lenger er behov for poliovaksine,
AV/Ad96/01: , selv om en tur til Tenerife **ikke** ville være å forakte. - Det er klart j
AV/Ad96/01: essuten råartig, selv om det **ikke** går så bra. Arti' lælI, sier hun. Hi
AV/Ad96/01: opa. Det er ikke få, heller **ikke** på Sørlandet. Agderposten | Betydnings
AV/Ad96/01: da nede i 70 kilo, men hadde **ikke** satt noe mål på hvor langt jeg skulle
AV/Ad96/01: sent stilling, og derfor ser **ikke** vi på den som noen bijobb. Når leger u

Mens vi altså så etter nominalfraser, fant vi eksempler på adverb, adverbialer, og predikater inkludert verb. Når mye av litteraturen har teoretisert rundt emnet med objektsnominal-fraser i tankene, sier det seg selv at analysene står overfor en kraftig revisjon. Korpussøkingen har hjulpet oss til å finne nye og uventede data. (Se Johannessen 1998a og Arne Martinus Lindstad 1999 for mer om negativ polaritet.)

Professor Helge Lødrup, UiO, er også en av dem som har opplevd at datainnsamling med korpus har åpnet øynene. I et arbeid med pseudokoordinering (Lødrup 2001) var han nokså sikker på at det ikke var mulig å koordinere verb etter konstruksjonen *holde på å*, altså at man alltid ville få infinitivsform etter konjunksjonen. Dette viste seg langt fra å holde stikk:

(14)

Bøyd verb etter *holde på x*:

AV/Loka/01: id skikkeleg greie på kva dei **held på og ordnar** med. Til dømes tiltak

SA/GjNi/01: eit anna døme : Ein sildestim **held på og sig** under land ein fin

SA/GjNi/01: t det er nokre arbeidarar som **held på og byggjer** veg i ein berghamar

SA/SS57/01: hop med at Duun i denne tida **heldt på og skreiv** dei forteljingane som

SA/SS57/01: lumsk form for analfabetisme **held på og utviklar** seg - ho grip om seg i

SK/BaEn/01: ugo frå ein appelsinkasse han **held på og opnar** og der står ho. - Ei eske

SK/BeOl/01: sjøla på skyss-stasjona, som **heldt på og vart** litt for varm i greie. Han

SK/BeOl/01: vet og bygdatrollet attåt alt **heldt på og åt** opp Skoleholderen frå båe

SK/BeOl/01: a til - mest som ein ynk - og **heldt på og slo** ut lyset sitt mot dørkarmen.

Lødrup lette videre etter eksempler på *ta og*, for å få bekreftet at verbet etter konjunksjonen måtte ha et transitivt verb med objekt. Her var intuisjonene hans mer i tråd med det han fant i korpuset, men han gikk også til en av søkemotorene for web-sider, og fant faktisk så mange eksempler på også intransitive verb, at han måtte trekke hypotesen:

(15)

Intransitivt verb etter konjunksjonen, funnet på Internett

ta og piss oppover veggen
ta og ring til ham

Av og til er det man er på jakt etter, svært marginalt. Da må man være ekstra på vakt etter hvordan man søker. Et eksempel er pronomenet *denne* (den bruken av *denne* som viser til et ledd foran i setningen eller i en foregående setning, og som ikke har noe substantiv etter seg). Jeg var interessert i å finne ut noe om hvorvidt det var mulig å se om dette pronomenet fulgte bestemte grammatiske regler, og ønsket å bruke korpus for å slippe interferens fra mulige hypoteser i datainnsamlingen. Det viste seg å være vellykket:

(16)

Eksempler på pronomenet *denne*, funnet i Oslo-korpuset:

AV/Ad96/01: r fungert i undervisningen. - Vi har tatt av egen ramme, blant annet vikarressursen, og brukt av **denne** i klassene. Vi har også organisert arbei

AV/Ad96/01: r, med konkrete prosjekter i Trinidad og Venezuela. Opsjon på ei tomt er skaffet på Trinidad, og **denne** går frem til høsten 1998. Tiden skal bruke

AV/Af94/01: inne i en brytningstid. Braathen hadde sin Østenkonsesjon, men det var klart at man ikke ville få **denne** fornyet. Regelverket ga imidlertid muligheter

AV/Af94/01: forøvrig det følgende år. Brinckmann leverte i sine yngre dager enkelte prestasjoner av rang, og **denne** skriver seg fra 1937 : Coggan Brinckmann I Som

AV/BT95/01: nnheringen søkes det etter en person som er villig til å ta et vikariat i svangerskapspermisjon for **denne**. Er det snakk om et vikarierende svanger

AV/BT95/01: , er rene galskapen. 2. JÅ sende hundrevis av Ulriken-elever til Gimle skole for om litt å gjøre **denne** til en mammutskole, er elev- og samfunnsfi

Generaliseringen viste seg å være at pronomenet *denne* har en tendens til å vise til alt annet enn et tidligere subjekt (se Johannessen 1996).

Enkelte fenomener kan være så marginale at det kan være bortimot umulig å bruke korpus. For igjen å ta et eksempel jeg kjenner godt: I doktorgradsarbeidet (publisert som Johannessen 1998) mitt skrev jeg om koordinering, og spesielt var jeg interessert i de tilfellene der konjunktene i en konjunksjonsfrase hadde grammatiske trekk som var uventet i den syntaktiske posisjonen de var i, som i (17a) versus (17b):

(17) a. Per og meg gikk på kino.
b. * Meg gikk på kino.

Jeg visste ut fra enkelte dialektgrammatikker at denne muligheten også finnes på norsk, og min egen bruk av introspeksjon sa meg også at dette var mulig, men det viste seg å være umulig å finne skriftlige belegg i korpus. Antagelig er dette en svært muntlig uttrykksmåte, hvor et talespråkskorpus kunne vært nyttig, men hvor de skriftspråkskorpusene jeg hadde til rådighet, ikke var til noen hjelp. I slike tilfeller må man selvsagt ty til andre metoder, som introspeksjon og informanter, samt, selvsagt, det som måtte være av publiserte grammatikker.

5 En sammenligning av metodene

De tre innsamlingsmetodene er, som vi har sett, ikke jevnbyrdige, verken når det gjelder muligheten for å velge mellom dem, eller hvilke fordeler og ulemper de har. Nedenfor er det gitt en oversikt over hva slags muligheter man har til datainnsamling ut fra den situasjonen som er gitt for språkforskeren:

(18)

Metode / Situasjon	Informant (andres introspeksjon)	Egen introspeksjon	Skriftspråkskorpus	Talespråkskorpus
Språket snakkes i dag	x	x	x	x
Språket snakkes ikke i dag?			x	
Ukjent språk (for lingvisten)	x		x	x
Kjent språk for lingvisten	x	x	x	x
Mangler skriftspråk	x	x		x
Har skriftspråk	x	x	x	x

Man skulle kanskje tro at det var vanntette skott mellom de ulike innsamlingstypene, i den forstand at de representerer prinsipielt ikke. Nedenfor skal vi se på noen viktige enkeltpunkter og vurdere de tre hovudmetodene for datainnsamling opp mot disse.

5.1 Forskerens påvirkning på datainnsamlingen

Vi har allerede vært inne på at både informantbruk og introspeksjon i stor grad er det samme. Den største forskjellen mellom egen introspeksjon og informantintrospeksjon er kanskje at informanten ikke er klar over hva som i detalj er formålet med undersøkelsen, slik at dataene ikke blir farget i samme grad. Her har bruk av informant en utpreget fordel. På den annen side kan man ikke alltid unngå at informanten i samtale med forskeren har en viss vilje (eller uvilje) til å gi forskeren det han vil ha. Her kan forskerens ansiktsuttrykk og øvrige mimikk, stemmebruk o.a. avsløre hva slags data man venter eller ønsker. Forskerens rolle i samkvemmet med informanten eller i samfunnet generelt, vil også avgjøre hva slags forhold informanten har til en, og kan ha en innvirkning på i hvilken grad informanten ønsker å gjøre en til lags. (Disse problemene peker Lanza, Vonen og Erlenkamp på i sine kapitler i denne boka.) Når det gjelder korpusbruk, er det liten fare for at forskeren påvirker datainnsamlingen, med mindre vedkommende bruker helt feilaktige søkemetoder (noe som jo kan skje).

5.2 Påvirkning fra språklige normer

Informanten er ofte lingvistisk naiv, og kan la svarene farges av andre forhold enn rent språklige. For eksempel finnes det i alle språk normer for hva som regnes som god og dårlig språkbruk, og en informant kan lett farges av slike betraktninger. Men problemet kan også lett ramme språkforskeren selv, under bruk av egen introspeksjon. Videre eksisterer dette problemet også for skriftspråkskorpus; som regel er tekstene som finnes i spesielt utviklede korpus hentet fra kvalitetskilder, der tekstene er skrevet av skriveføre mennesker og gjerne språkvasket og korrekturlest. Dermed følger de ikke bare normen til de enkelte skribentene – hva de oppfatter som godt og dårlig språk – men

også normen til den publikasjonen der tekstene er trykket. Når det gjelder norm, er det kanskje bare talespråkskorpus som i prinsippet kan være mer unormert. Dersom talespråkskorpuset består av mye spontan tale fra dialoger og samtaler, er det mindre sannsynlig at språket er hemmet av bevisste normer. Her snakker vi om gradsforskjeller, antagelig er all språkbruk i noen grad farget av ens forhold til hvordan en oppfatter at språket bør være.

Dersom lingvisten kan språket selv, kan informantenes opplysninger sjekkes mot ens egen intuisjon – noe som både kan være en kontrollmetode, men også en ytterligere feilkilde.

5.3 Mulighet for oppfølgingsspørsmål

Som regel oppdager språkforskeren at datagrunnlaget ikke er fullstendig etter første runde med datainnsamling. Man kan oppdage unøyaktigheter i dataene, eller muligheter for misforståelser der informanter har vært inne i bildet. Men aller vanligst er det nok at man etter å ha sammenstilt de første dataene mot de opprinnelige hypotesene, trenger å ytterligere presisere de opprinnelige hypotesene, eventuelt omformulere dem helt. Da trenger man gjerne å utvide datagrunnlaget, siden dataene i første runde var samlet inn ut fra de opprinnelige hypotesene.

Rent prinsipielt kan man stille oppfølgingsspørsmål til alle tre typene av kilder. Det er ikke noe i veien for å stille videre spørsmål til informantene, eller bruke introspeksjon med de nye problemstillingene, eller for den saks skyld formulere søkekriteriene i korpuset annerledes enn man gjorde første gang. Men ofte er man i den situasjonen at man ikke lenger har tilgang til informantene, for eksempel fordi informantene befinner seg et annet sted i verden. Informantene kan også få et annet forhold til innholdet i spørsmålene når de blir konfrontert med materialet igjen, og er kanskje ikke så stabile i forhold til de første dataene lenger. Introspeksjon i runde to er ikke

mindre spekket med feilkilder enn i runde en. Antagelig er det bruk av korpus som gir den største muligheten for oppfølgingsspørsmål, fordi korpuset er der det alltid har vært, tekstene og språket i dem er det samme, og nye spørsmål vil ikke påvirke hvordan korpuset svarer.

5.4 Etterprøving av data

Et vitenskapelig ideal er muligheten for å etterprøve data. Innenfor medisinen er dette svært vanlig, men ikke i språkforskningen. En grunn til den manglende etterprøvingen kan være problemet med datainnsamlingen. Data som har fremkommet ved introspeksjon kan vanskelig etterprøves av andre. Data som har fremkommet gjennom informant-undersøkelser er i prinsippet etterprøvbare, men i praksis er det store vanskeligheter. Det er kostbart og tidkrevende å spore opp gamle informanter, og dersom de er i live, er det ikke sikkert at de har de samme språklige intuisjoner som de engang hadde. Datainnsamling ved hjelp av korpus er i praksis den eneste metoden der dataene kan etterprøves.

5.5 Negative data

Vi har nevnt flere ganger at negative data kan være viktige for lingvisten. Man spør seg ofte: Er denne setningen eller konstruksjonen eller dette ordet mulig? Både bruk av informanter og introspeksjon gir muligheten til å stille direkte negative spørsmål. Derimot er det prinsipielt vanskelig å finne negative data i et korpus. Dersom man ikke finner et eksempel på det man har søkt om, er det ikke gitt hvordan man skal tolke svaret. Finnes ikke fenomenet i korpuset av tilfeldige grunner? Eller er det fordi korpuset ikke er sammensatt på en mest mulig representativ måte (i forhold til nødvendige kriterier i den gitte sammenhengen)? Er fenomenet svært marginalt, men hyppig i visse type tekster? Faarlund viser i sitt kapittel at man likevel ikke behøver å avvise skriftlige kilder helt når det gjelder

negative data. Ved å benytte flere tilnærminger, teoretiske så vel som empiriske, kan man faktisk nærme seg et svar på hvorfor enkelte konstruksjoner ikke er belagt i en tekst eller et tekstkorpus.

5.6 Tolkning av data

Mens lingvisten enkelte ganger trenger å finne eksempler på mulige grammatiske konstruksjoner, er målet andre ganger heller å tolke gitte setninger eller tekstutsnitt. Når man ønsker å teste hva slags rekkevidde kvantorene i en gitt setning har, eller om en bestemt konstruksjon kan være flertydig, er det nødvendig for lingvisten å bruke andre enn bare seg selv. Semantiske tolkningsvurderinger kan være så vanskelige at faren for påvirkning av data er for stor dersom forskeren også skal være sin egen dataleverandør. Selv om korpus kan være brukbart for å finne ulike betydninger av kvantorer i ulike posisjoner, vil det sjelden være mulig å finne tvetydighet for nøyaktig samme setning eller konstruksjon. Da er det lurt å teste tolkningsmulighetene på informanter.

5.7 Store datamengder

Enkelte forskningsformål krever store datamengder. I noen lingvistiske teorier (som kognitiv grammatikk, representert av for eksempel Ronald Langacker 1987) er det helt nødvendig med store datamengder fordi forklaringskraften er nært knyttet til frekvensen til de språklige konstruksjonene. Også en del sosiolingvistiske og dialektlingvistiske formål er avhengige av store datamengder for å kartlegge utbredelsen av ulike fenomener. Korpusundersøkelser er selvsagt uovertrufne når det gjelder å skaffe store mengder data, i hvert fall når man tar tids- og økonomifaktoren med. Men dersom et korpus skal være brukbart for slike kartlegginger og frekvenstellinger, er det viktig at det er balansert mht til de variablene forskeren er interessert i, eventuelt at det generelt er så balansert som det er

praktisk mulig å tenke seg, mht teksttyper og geografi, alder og utdannelse på korpusets «språkleverandører» og så videre. De færreste korpus i dag er så balanserte som noen språkforsker skulle ønske, og man må derfor ta kvantitative undersøkelser basert på korpus med mange store klyper salt. Som et illustrerende eksempel kan vi ta et eksempel på ordfrekvens. I Oslo-korpusets avisdel på bokmål er det nesten ti millioner ord. Navnet *Bergen* er det 74. mest frekvente, mens *Oslo* havner lenger ned, på 101. plass, og *Trondheim* helt ned på 281. plass. Det sier seg selv at fordelingen av disse bynavnene ikke reflekterer alle språkbrukere i Norge, men er en tilfeldighet ved det avisutvalget som er brukt.

5.8 Variasjon i språkbruk

Mange studier innenfor sosiolingvistikk eller diskursteori har som mål å avdekke hvordan språkbruk varierer innenfor ulike situasjoner eller mellom ulike grupper av befolkningen, eventuelt innenfor ulike tidspunkter av en dialogsituasjon. I slike tilfeller må man bruke informanter. Det kan eventuelt tenkes at man kan ta i bruk et korpus som er satt sammen av ulike informanter på en systematisk måte som kan dekke forskerens behov. Informanter kan også brukes i noen grad til å teste ut om de funnene man har gjort, virker riktige. Men dersom variasjonen også avspeiler ulikheter i hva som er sosialt akseptabelt, er det ikke sikkert informanten kan fri seg fra normen når vedkommende blir bedt om å ta stilling til lingvistens funn. Egen introspeksjon er ikke nok for denne typen studier.

5.9 Mulighet for å bruke grammatisk terminologi under innsamlingen

Bruk av grammatisk terminologi kan gjøre datainnsamlingen mye enklere. Har man en grammatisk kunnskapsrik informant, kan man bruke slik terminologi, men det er nok heller sjelden

dette er mulig. (Enkelte vil også hevde at informanter ikke bør være altfor skolerte, slik vi har sett i Vonens kapittel.) Mange korpus inneholder grammatisk taggede tekster. De har som regel et søkesystem som gjør det mulig å utnytte de grammatiske kategoriene. Dermed kan man formulere grammatiske søk som «Gi meg alle verb som er fulgt av preposisjon, men ikke av substantiv», noe som gir uante muligheter for innsamling av data for å belyse et fenomen, for eksempel som et første skritt i studier av partikkelverb. Men selvsagt er det begrensninger også her: Dersom en grammatisk kategori ikke er avmerket i korpuset, kan man selvsagt ikke søke på den heller.

5.10 Mulighet for nye og overraskende data

Det er sjelden at språkforskeren har full oversikt over alle sider ved et fenomen før datainnsamlingen har begynt. De metodene som kan gi en nye og overraskende data, er derfor å foretrekke. Bruker man informanter i en dialogsituasjon, vil det alltid være en mulighet for at nye og overraskende data kan fremkomme, særlig dersom informanten tenker selvstendig og er samarbeidsvillig. En forsker kan bruke introspeksjon og komme fram til enkelte nye data, men som regel vil man være for bundet opp til det man har i tankene fra før, til at man kan gi seg selv store overraskelser. Korpus er en klar kilde til overraskelser. De færreste mennesker er klar over at et bestemt ord eller en bestemt spørsmålsformulering er flertydig, når man selv har én betydning for øyet. Når forskeren formulerer søkekriterier til korpuset, vil det derfor ofte vise seg fra resultatene at man får flere og/eller andre ting enn man har bedt om. Som regel vil slike utilsiktede resultater gi forskeren noe å tenke på, og ikke sjelden vil det føre til at atter nye hypoteser ser dagens lys. Bruk av introspeksjon kan nok i enkelte tilfeller føre til overraskelser for forskeren, men det er i

så fall ingen følge av metoden, men kanskje heller på tross av den.

5.11 Tid og penger

Når det gjelder å få mye data raskt, er korpusmetoden uovertruffen. Ingen mennesker kan i den grad gi resultater og i et slikt omfang. Bare størrelsen på korpuset begrenser hvor mange søkeresultater som finnes. Oslo-korpuset gir eksempelvis 145 655 forekomster etter søk på negasjonsordet *ikke*. Hvor mange informanter ville brukt hvor mye tid på en tilsvarende mengde?

Korpus er også uten tvil den raskeste og enkleste innsamlingsmetoden. Man behøver ikke bevege seg, men kan sitte godt nedsunket i kontorstolen og fiske fram data, eller for den saks skyld ligge hjemme i senga med en bærbar datamaskin, eller sitte i hytteveggen. Korpus kan til og med være raskere enn introspeksjon, fordi korpuset aldri tviler, grunner og nøler, slik mennesker kan gjøre når de skal vurdere setninger eller hente fram gode eksempler fra sitt eget hode.

Korpus og introspeksjon er de eneste metodene som er forholdsvis billige. Når man slipper å betale billetter og forskningsopphold, belønne informanter eller forskningsassistenter, eller selv bruke verdifull arbeidstid på å hente ut informasjon, er det opplagt at metoden blir billig. Av denne grunn kan det være veldig lurt å benytte korpus (dersom det finnes for det aktuelle språket) i de innledende forskningsfaser, selv når man vet at det av ulike grunner kommer til å være nødvendig å benytte informanter i tillegg.

(19) En oppsummering av metodene og deres anvendelse

Innsamlingsmetode / Egenskaper ved metoden*	Informant (andres introspeksjon)	Egen introspeksjon	Korpus
Forskeren påvirker dataene	Kanskje	Ja	Kanskje
Påvirkning fra normer	Ja	Kanskje	Ja
Mulighet for oppfølgingsspørsmål umiddelbart	Ja	Ja	Ja
Mulighet for oppfølgingsspørsmål på et senere tidspunkt	Kanskje	Kanskje	Ja
Resultatet kan sjekkes av andre, etterprøving	Kanskje	Nei	Ja
Man kan få negative data	Ja	Ja	Kanskje
Man kan få semantisk tolkning av data	Ja	Kanskje	Kanskje
Store datamengder	Ja	Nei	Ja
Variasjon i språkbruk	Ja	Nei	Ja
Mulighet for å bruke grammatisk terminologi i datainnsamlingen	Kanskje	Kanskje	Kanskje
Mulighet for nye og overraskende data	Kanskje	Kanskje	Ja
Raskt og effektivt	Nei	Ja	Ja
Billig	Nei	Ja	Ja

* For de tilfellene der en rute i tabellen er fylt med kommentaren «Kanskje», vises det til diskusjon av dette punktet ovenfor.

5.12 Kombinasjon av metodene

Etter denne gjennomgåelsen av de ulike metodene, er det viktig å påpeke at metodene ikke behøver å brukes en og en. Der det

finnes korpus tilgjengelig for et språk, og forskeren selv er en taler av språket, kan man naturligvis kombinere metodene. Dermed unngår man en del av fallgruvene som gjelder den enkelte metode, samtidig som ulike metoder kan virke som kontrollfunksjon for hverandre. Dersom man for eksempel har kommet fram til enkelte svar ut fra introspeksjon, vil det være foruroligende dersom både informantundersøkelser og korpusundersøkelser ikke støtter opp om funnene. Samtidig er korpusmetoden så billig og så sikker at den kanskje alltid bør brukes i det minste i en første fase, en pilotundersøkelse. Ved først å bruke korpus, kan man blant annet få en pekepinn om hvorvidt de spørsmålene man stiller, er generelle nok, presise nok, om de er flertydige eller på siden av det man egentlig ønsker å vite noe om, eller om det finnes andre spørsmål som ville være enda bedre for å finne de dataene man ønsker, og få svar på de hypotesene man stiller.

5.13 Konklusjon

Vi har i dette kapitlet sett på hvordan korpus kan brukes i språkforskningen, og på noen heldige og mindre heldige sider ved korpusbasert forskning. Vi har også sammenlignet korpusbasert forskning med de to andre nærliggende metodene for datainnsamling: informanter og introspeksjon. På en lang rekke punkter fant vi at korpusbasert forskning har mye for seg, av og til mer enn bruk av informanter og introspeksjon. Kanskje ikke overraskende viser denne gjennomgåelsen at ulike mål gir behov for ulike typer datainnsamling. Men siden korpusmetoden er uovertruffen når det gjelder datamengde, tid og penger, er det ingen tvil om at den, dersom det finnes korpus for det språket man vil undersøke, bør brukes også i de tilfellene man vet at man må bruke andre og mer kostbare og tidkrevende metoder, i en første fase.

Litteratur

Asmussen, H. 2000. *Korpus 2000: - En undersøgelse af brugergrupper og korpusværktøjer.* Prosjektoppgave, Institut for Datalingvistik, Handelshøjskolen i København.

Chomsky, N. 1986: *Knowledge of language: its nature, origin, and use.* New York: Praeger.

Endresen, R.T. og H.G. Simonsen. 2000: Morfologi. I Endresen, R.T., H.G. Simonsen og A. Sveen (red.). *Innføring i lingvistikk.* Oslo: Universitetsforlaget.

Faarlund, J.T., S. Lie og K.I. Vannebo. 1997: *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.

Johannessen, J.B., A. Nøklestad og K. Hagen. 2000: A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts. I Gavrilidou, M., G. Carayannis, S. Markantonatou, S. Piperidis og G. Stainhaouer (red.) *Proceedings, Second International Conference on Language Resources and Evaluation (LREC 2000)*, Aten, 1725-1729 .

Johannessen, Janne Bondi. 1996: DENNE. *Norsk Lingvistisk Tidsskrift 14-1*, p. 3-27.

Johannessen, Janne Bondi. 1998a: Negasjonen ikke: Kategori og syntaktisk posisjon. I Faarlund, J.T., Mæhlum, B. og T. Nordgård (red.) *MONS 7*, s. 80-94, Oslo: Novus forlag.

Johannessen, Janne Bondi. 1998b: *Coordination.* New York : Oxford University Press.

Johannessen, Janne Bondi. 2001: Sammensatte ord. *Norsk Lingvistisk Tidsskrift* 19-1, 59-92.

Johansson, Stig. 1998: On the role of corpora in cross-linguistic research. I Johansson, S. og S. Oksefjell: *Corpora and Cross-linguistic Research*, Rodopi, Amsterdam, Atlanta, GA, 3-24.

Langacker, Ronald W. 1987: *Foundations of cognitive grammar.* Stanford, California: Stanford University Press.

Lindstad, Arne Martinus. 1999: *Issues in the Syntax of Negation and Polarity in Norwegian – A Minimalist Analysis.* Hovedoppgave, Universitetet i Oslo.

Lorimer, Elizabeth. 1989 [1938]: *Language Hunting in the Karakoram.* Karachi: Indus Publications.

Lødrup, Helge. 2001: The Syntactic Structures of Norwegian Pseudocoordinations. Paper presentert på The 6th International Lexical Functional Grammar Conference, Hong Kong June 2001.

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

Morgenstierne, Georg. 1992: *På sprogjakt i Hindukush. Dagboksnotater fra Chitral 1929.* Utgitt ved Eva M. Lorentzen i samarbeid med Knut Kristiansen og Fridrik Thordarson. Indo-iransk bibliotek, Universitetet i Oslo.

Nilsen, Øystein. 2000: *The Syntax of Circumstantial Adverbials.* Oslo: Novus forlag.

Korpus

Norske talespråkskorpus

Big Brother: http://www.tekstlab.uio.no/talespraak/bigbrother/

Talesøk: http://www.hf.uib.no/i/Nordisk/talekorpus/Hovedside.htm

Svensk talespråkskorpus

Göteborg Spoken Language Corpus:
http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

Dansk talespråkskorpus

Bysoc – Dansk talesprog: http://www.id.cbs.dk/~pjuel/cgi-bin/BySoc_ID/index.cgi

Norske skriftspråkskorpus

Oslo-korpuset av taggede norske tekster:

http://www.tekstlab.uio.no/norsk/bokmaal/

Oslo Multilingual Corpus:
http://www.hf.uio.no/german/sprik/korpus.shtml

Aviskorpus ved HIT-senteret, UiB:
http://www.hit.uib.no/hit/avis-pro.htm

Svenske skriftspråkskorpus

Språkbanken: http://spraakbanken.gu.se/

SUC-korpuset: http://www.ling.su.se/staff/sofia/suc/suc.html

Danske skriftspråkskorpus

Arboretum (dansk trebank):
http://corp.hum.sdu.dk/arboretum_about.html

Korpus 90/2000: http://corp.hum.ou.dk/corpustop.html

Annet

Tekstressurser for mange språk: Linguistic Data Consortium:
http://www.ldc.upenn.edu/

Bokmålsordboka, nettutgave:
http://www.dokpro.uio.no/ordboksoek.html