Korpus i forskning og undervisning

Korpus i forskning og undervisning

Hvor står vi i Norge?

Rapport fra en konferanse på Høgskolen i Agder høsten 2004

Sissel Rike (red.)

Sammendrag

Artiklene er basert på foredrag som ble holdt på konferansen ”Korpus i forskning og undervisning. Hvor står vi i Norge?”, arrangert av Høgskolen i Agder, Fakultet for humanistiske fag. Siktemålet med konferansen var å presentere et tverrsnitt av aktuelt korpuslingvistisk arbeid innenfor fagspråk og sakprosa i Norge, og på dette grunnlaget stimulere til intensivert forskning. Artiklene er skrevet av professor Stig Johansson, Universitetet i Oslo, professor Cathrine Fabricius Hansen, Universitetet i Oslo, professor Kjersti Fløttum, Universitetet i Bergen og professor Magnar Brekke, Norge Handelshøyskole.

Skriftserien nr. 124

84 sider
110,- NOK
ISSN: 1503-5174 (elektronisk utg.)
ISBN: 82-7117-573-4 (elektronisk utg.)

Serviceboks 422, N-4604 Kristiansand

Design: Høgskolen i Agder

Emneord:

Korpus
Lingvistikk
Terminologi
Korpusbasert forskning

Innholdsfortegnelse

Forord 1

Å bruke korpus er å se
Stig Johansson, Universitetet i Oslo 3

Lingvistiske og diskursive studier i KIAP-korpuset. Om utvikling og bruk av et flerspråklig og flerdisiplinært korpus av vitenskapelige artikler
Kjersti Fløttum, Universitetet i Bergen 25

Parallellkorpora som innfallsport til språksammenligning på tekstnivå – med særlig henblikk på (norsk/tysk/engelsk) sakprosa
Cathrine Fabricius-Hansen, Universitetet i Oslo 53

Økonomisk-administrativ kunnskapsbank: eit korpusbasert terminologiprosjekt
Magnar Brekke og Kai Innselset, Senter for fagspråkforsking, Institutt for fagspråk og interkulturell kommunikasjon, Norges Handelshøyskole 67

Forord

”Korpus i forskning og undervisning. Hvor står vi i Norge i dag?” var tema for en konferanse arrangert høsten 2004 av Høgskolen i Agder, Fakultet for humanistiske fag. Konferansen var ledd i et prosjekt innenfor Institutt for oversetting og fagspråk finansiert av Sørlandets Kompetansefond.

Siktemålet med konferansen var å presentere et tverrsnitt av aktuelt korpuslingvistisk arbeid innenfor fagspråk og sakprosa i Norge, og på dette grunnlaget stimulere til intensivert forskning. Man ønsket å sette søkelyset på den nytteverdien og det potensialet som korpuslingvistisk forskning kan ha for undervisning og fagtilbud.

Foredragsholdere var professor Stig Johansson og professor Cathrine Fabricius Hansen fra Universitetet i Oslo, professor Kjersti Fløttum fra Universitetet i Bergen og professor Magnar Brekke fra Norge Handelshøyskole. De står alle i spissen for sentrale norske korpusbaserte forskningsprosjekter.

Professor Stig Johansson viser i sin artikkel at å bruke korpus er ’å se’. Han definerer hva et korpus er, og gir en rekke eksempler på områder hvor bruk av forskjellige typer korpus, både enspråklige og tospråklige, har hatt stor betydning. Videre peker han på det potensialet som bruk av korpus representerer både innenfor forskning og undervisning. En vesentlig forutsetning for å se, er imidlertid at man har noe som er godt egnet til å kikke på. Johansson understreker at det er viktig å satse energisk og systematisk på å bygge opp det empiriske grunnlaget. Enkelte andre land har sine nasjonalkorpus, mens Norge her ligger langt etter utviklingen.

Tittelen på professor Kjersti Fløttums artikkel er ”Lingvistiske og diskursive studier i KIAP-korpuset. Om utvikling og bruk av et flerspråklig og flerdisiplinært korpus av vitenskapelige artikler”. KIAP står for *Kulturell Identitet i Akademisk Prosa: nasjonal versus disiplinavhengig.* Professor Fløttum beskriver prosjektet og orienterer om korpuset før hun presenterer utvalgte resultater. Avslutningsvis sier hun at erfaringene

med korpusstudiene i KIAP-prosjektet har vært svært positive. Korpuset utvikler stadig nye problemstillinger og vekker lingvistisk nysgjerrighet. Fremdeles er det mye som er uutforsket, som det kan arbeides videre med når prosjektet er formelt avsluttet i 2005.

Professor Cathrine Fabricius Hansens artikkel omhandler parallellkorpora som innfallsport til språksammenligning på tekstnivå – med særlig henblikk på (norsk/tysk/engelsk) sakprosa. Hennes betraktninger er først og fremst av metodologisk art. Hun presenterer forskjellige typer parallellkorpora til forskjellige formål og gir eksempler på hvordan disse kan brukes som oppdagelsesreise og middel til hypotesedannelse når det gjelder språkspesifikke prinsipper for hvordan informasjon struktureres i skrevet tekst. Særlig viser hun at oversettelseskorpora ikke bare er nyttige som ledd i oversettelsesrelatert undersøkelser, men kan danne grunnlag for kontrastive problemstillinger på tekstnivå og dermed utgangspunkt for videre interessante analyser.

I artikkelen "Økonomisk-administrativ kunnskapsbank: eit korpusbasert terminologiprosjekt" presenterer professor Magnar Brekke KB-N (KunnskapsBank for Norsk økonomisk-administrativt domene). Han setter KB-N-prosjektet inn i et teoretisk perspektiv, og beskriver deretter systemarkitektur og metodologisk tilnærming. Programverktøyet og brukergrensesnittet blir presentert, blant annet med bilder av forskjellige elektroniske vinduer. Artikkelen avsluttes med en redegjørelse om aktuelle bruksområder for KB-N.

Samlet gir disse artiklene et variert bilde av den korpusbaserte forskningen som foregår i Norge i dag. Innfallsvinklene er mange, og potensialet er stort. Ikke minst viser det seg at forskning med basis i korpus stadig avdekker nye problemstillinger og interessante perspektiver, som gir grunnlag for videre forskning.

Kristiansand, 2005

Sissel Rike (red.)

Å bruke korpus er å se

Stig Johansson, Universitetet i Oslo

1. Å se gjennom korpus

Tittelen på mitt foredrag er inspirert av noen ord som tilskrives Henrik Ibsen: ”Å dikte er å se”. Det forutsettes at det som er sett, også må uttrykkes. Men først må det bli sett. Det samme gjelder forskning og bruken av korpus i språkforskningen. Det er ofte blitt sagt at korpus gjør det mulig for oss å se nye mønstre, eller å se tydeligere det vi bare hadde en vag forestilling om.

> Corpus methods can organize huge masses of data, and make visible patterns which were only, if at all, dimly suspected. In giving access to new data, the technology opens up research topics which were previously inconceivable. We now have facts about language use which no amount of introspection or manual analysis could discover […] . (Stubbs 2002: 221)

Det er dette jeg vil prøve å vise. Men hva mener vi egentlig med et korpus?

2. Hva er et korpus?

Et korpus betyr helt enkelt en samling tekster. Det er ikke noe nytt at tekster blir brukt språkforskningen. Store ordbøker som *Oxford English Dictionary* og grammatikkbøker som Jespersens *Modern English Grammar* var basert på store samlinger med eksempler på språkbruk i tekst. Det samme gjelder mange tradisjonelle avhandlinger i språkvitenskap. Det som er nytt nå, er at vi har korpus i elektronisk form som vi kan søke i og analysere ved hjelp av dataprogrammer. Et moderne korpus er også strukturert og bygd opp på en prinsipiell måte.

Ved tradisjonell ekserpering fra tekst var det umulig å få med alt, og det var en tendens til å ta med det som var spesielt og avvikende. James Murray, den berømte første redaktøren av *Oxford English Dictionary*, klager i et brev:

> […] the editor or his assistants have to search for precious hours for examples of common words, which readers passed by […]. Thus of *Abusion*, we found in the slips about 50 instances: of *Abuse* not five. (James Murray, sitert fra K. M. Elizabeth Murray, barnebarn til redaktøren, i boken *Caught in the Web of Words*, 1979: 178)

I et elektronisk korpus har vi tilgang til store mengder løpende tekst, og vi kan bruke det både til å studere språkets *capriccio* og dess *ostinato*, både det som er avvikende og det som er typisk (jf. Allén 1992: 1).

I løpet av de siste tiårene er det blitt en enorm utvikling når det gjelder tilgang til og bruk av tekstkorpus. Det har å gjøre både med den teknologiske utviklingen og med en økende interesse blant lingvister for å studere språk i bruk. Det klassiske Brown Corpus fra 1960-tallet besto av en million ord. Det var svært mye på den tiden og kostet mye arbeid. De første korpusene var også vanskelige å bruke for en lingvist som ikke hadde spesielle datakunnskaper. Nå har vi datakorpus som består av hundre millioner ord eller mer, for eksempel *British National Corpus* fra 1990-tallet, med brukervennlig programvare som ikke krever noen særlig opplæring. Og foruten godt planlagte korpus har vi den store verdensveven, med en nesten uendelig tekstmasse.

Noen taler om *World Wide Web* som korpus. Her er jeg ikke helt enig. Jeg ville heller si at verdensveven er et stort tekstarkiv, som vi kan bruke for å bygge opp korpus for spesielle formål. Slik jeg tenker meg et korpus for bruk i språkforskningen, er det en elektronisk tekstsamling som er systematisk sammensatt med tanke på hvilke spørsmål som språkforskeren vil stille. Det finnes en lang rekke mulige spørsmål, og derfor også mange forskjellige typer korpus. Dermed er vi inne på neste spørsmål.

3. Hvorfor bruke korpus?

Som utgangspunkt kan vi se på noen punkter, de fleste basert på en liste laget av Jan Svartvik (1992: 8-10):

- Bruk av korpus er mer objektivt enn introspeksjon. Studiet er ikke begrenset av den enkelte lingvistens erfaring og intuisjon.
- Resultatene er verifiserbare.
- Korpus er nødvendig i historisk lingvistikk.
- Korpus er nødvendig i kvantitative studier.

- Korpus gjør det mulig å studere språkvariasjon systematisk: tale vs. skrift, regionale og sosiale forskjeller, stil osv.
- Korpus gjør det mulig å studere språk i kontekst.
- Korpus kan brukes som en ressurs i teoretisk lingvistikk, for eksempel for å finne relevante eksempler.
- Korpus er et viktig grunnlag i anvendt lingvistikk: språkundervisning, språknormering, oversettelsesstudier, leksikografi, språkteknologi osv.
- Tradisjonelt – før vi fikk korpus i elektronisk form – brukte hver enkelt forsker vanligvis sitt eget tekstmateriale. Tilgangen til korpus i elektronisk form kan danne en felles basis for studier av forskjellige aspekter av det samme materialet, slik at forskere lettere kan bygge videre på hverandres arbeid. Bruken av korpus i engelsk har ført til et verdensomfattende samarbeid innenfor rammen av *International Computer Archive of Modern and Medieval English* (ICAME).
- Tilgjengeligheten av korpus har gjort det lettere å studere tekster som det er svært vanskelig for den enkelte forskeren å samle og behandle, f.eks. forskjellige typer talespråk.
- Mange forskere studerer språk som de ikke har som morsmål, og de er derfor avhengige av andre datakilder enn introspeksjon.
- Bruken av korpus gjør det mulig å oppnå *total accountability*, dvs. en totalbeskrivelse, og er ikke begrenset til det som den enkelte lingvisten har oppfattet som interessant eller relevant. *Korpuset tvinger lingvistene til å se det de ellers lett kunne overse*.

Det finnes altså mange gode grunner til å bruke korpus, og det er blitt mer og mer akseptert. Derfor kunne Jan Svartvik bruke tittelen 'Corpora are becoming mainstream' i en artikkel fra 1996.

Vi har fått en ny term, *corpus linguistics* eller korpuslingvistikk, først brukt i begynnelsen av 1980-tallet; se f.eks. publikasjonen etter ICAME-konferansen i Nijmegen 1983 (Aarts and Meijs 1984). Termen er innarbeidet, man den kan lett bli misforstått. Korpuslingvistikk er ikke en type lingvistikk på linje med f.eks. psykolingvistikk eller sosiolingvistikk. Det er heller ikke knyttet til en spesiell språkteori, som f.eks. kognitiv eller generativ lingvistikk. Det er helt enkelt lingvistikk hvor vi bruker korpus, og korpus er anvendelig i svært mange typer lingvistikk. Korpus er heller ikke det eneste saliggjørende. Her vil jeg gjerne sitere hva Wallace Chafe sier om korpuslingvisten:

> What, then is a 'corpus linguist'? I would like to think that it is a linguist who tries to understand language, and behind language the mind, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations. […] But I continue to believe that one should not characterize linguists, or researchers of any kind, in terms of a single favorite tie to reality. […] I would like to see the day when we will all be more versatile in our methodologies, skilled at integrating all the techniques we will be able to discover for understanding this most basic, most fascinating, but also most elusive manifestation of the human mind. (Chafe 1992: 96)

Bruken av korpus betyr ikke at vi skal avvise andre måter å studere språket på. Jeg går ikke inn på det her, men viser til en liten tekst jeg skrev for et år siden (Johansson 2003). Hensikten nå er å vise noe av det vi kan oppnå gjennom korpusstudier.

4. Bruk av enspråklig korpus

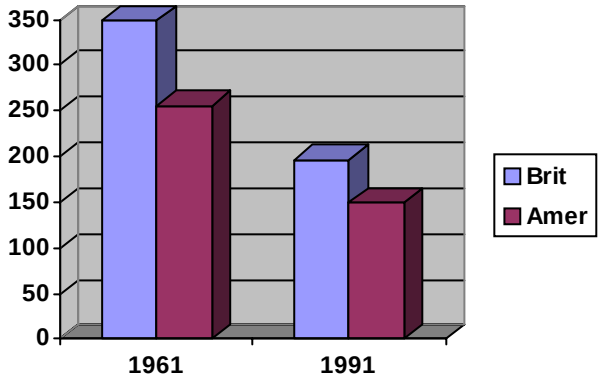
De første korpusene var enspråklige, det er de største korpusene og de som er blitt mest brukt. Derfor begynner jeg med dem. På grunn av at jeg har arbeidet mest med engelsk, konsentrerer jeg meg om engelske korpus. Engelsk er også det språket hvor korpuslingvistikken slo igjennom først og har fått mest gjennomslag.

Jeg begrenser meg til å ta opp noen få eksempler og begynner med to områder hvor bruken av korpus har hatt stor betydning: variasjonsstudier og leksikalske studier. Det finnes mange andre viktige områder, ikke minst historisk lingvistikk, hvor Universitetet i Helsingfors har spillet en ledende rolle (under ledelse av Matti Rissanen og Tertu Nevalainen). Interessant nok kobler prosjektet i Helsingfors sammen språkhistorie og språkvariasjon.

4.1 Språkvariasjon

Et område hvor korpus har hatt stor betydning er studiet av språkvariasjon. Brown Corpus ble satt sammen for å representere viktige typer av amerikansk engelsk skriftspråk fra begynnelsen av 1960-tallet. Senere har vi fått sammenlignbare korpus fra andre deler av den engelsktalende verden, f.eks. *Lancaster-Oslo/Bergen (LOB) Corpus* for britisk engelsk og *Kolhapur Corpus* for indisk engelsk. Vi kan kalle disse Brown-familien. Vi har nå fått en ny generasjon av Brown-familien: to korpus som

representerer britisk og amerikansk engelsk fra begynnelsen av 1990-tallet. De har fått navnene *Frown* og *FLOB*, etter Universitetet i Freiburg hvor de ble bygd opp under ledelse av Christian Mair. Et enkelt eksempel viser hvordan vi kan bruke Brown-familien. Hyppigheten av *shall* i de fire korpusene er sammenfattet i figur 1.



Figur 1 Frekvensen av *shall* i fire korpus: LOB (britisk engelsk, 1961), FLOB (britisk engelsk, 1991), Brown (amerikansk engelsk, 1961), Frown (amerikansk engelsk)

Vi ser med én gang at det er en forskjell mellom de britiske og de amerikanske korpusene. *Shall* er vanligere i britisk enn i amerikansk engelsk. Det var forventet. Det er mer interessant å konstatere at hjelpeverbet er blitt mindre hyppig både i amerikansk og i britisk engelsk i løpet av de tre tiårene mellom Brown/LOB och Frown/FLOB.

Slike kvantitative forskjeller må selvfølgelig undersøkes nærmere. Vi kan f.eks. se på forholdet mellom typen tekst og bruken av *shall*. Vi ser da at forskjellen mellom britisk og amerikansk engelsk var størst i skjønnlitteratur i 1961. I alle fire korpusene forekommer *shall* hyppigst i offisielle dokumenter – og omtrent like ofte i det britiske og det amerikanske materialet – men det er blitt en halvering i Frown og FLOB sammenlignet med Brown og LOB.

Det er åpenbart skjedd en betydelig forandring i bruken av *shall* i de siste tiårene. Slike resultater må tolkes. Har det å gjøre med en forandring i retning mot en mindre formell språkbruk? Er det skjedd en tilnærming mellom britisk og amerikansk engelsk skriftspråk? Hva er forholdet til andre modale uttrykk? Slike spørsmål er blitt tatt opp i en ny artikkel av Geoffrey Leech (2004).

Vi har også en annen korpusfamilie som er særlig egnet for å studere språkvariasjon, *International Corpus of English (ICE)*, et prosjekt som ble initiert av Sidney Greenbaum, University College London. Målet var å bygge opp korpus fra forskjellige deler av den engelsktalende verden, både engelsk som morsmål og som andrespråk. I motsetning til Brown-familien inneholder disse også talespråk (omtrent

60 %). I Oslo har vi nå tilgang til flere slike korpus: fra Storbritannia, India, Østafrika, Singapore og Filippinene. Her er noe jeg nylig fant i det indiske korpuset:

(1) And these women *are not knowing* anything about it.

(2) Now you *were knowing* this accused by name Varma?

(3) Now tell me this Mr Angale since how long you *had been knowing* Chandrahas Chipre?

(4) Now you *must be knowing* you see during the colonial days Englishmen considered Indian English as a subvariety of British English.

Det er ikke ukjent at mange regionale varianter av engelsk bruker den progressive formen ved statiske verb hvor den vanligvis er utelukket i standardengelsk. Det er bl.a. rapportert for indisk og afrikansk engelsk. Men det er påfallende hvor mye jeg fant i det indiske ICE-korpuset: 24 *BE knowing* i det indiske korpuset mot 2 i det østafrikanske korpuset. Derfor har en av mine studenter akkurat fått som oppgave å undersøke bruken av den progressive formen i indisk engelsk.

De viktigste typene variasjonsstudier gjelder ikke regional variasjon, men forskjeller mellom teksttyper. Særlig er det blitt gjort store fremskritt i studiet av talespråk. En viktig milepæl var London-Lund-korpuset (se Greenbaum & Svartvik 1990), som er blitt gjenstand for svært mange studier: syntaks, ordforråd, fraseologi, diskurs, prosodi osv. Allerede 1986 ble det publisert en bok basert på en sammenligning av LOB-korpuset og London-Lund-korpuset (Tottie & Bäcklund 1986). Et par år senere publiserte Douglas Biber sitt innflytelsesrike arbeid *Variation Across Speech and Writing* (Biber 1988), også den basert på LOB-korpuset og London-Lund-korpuset.

For noen år siden kom den korpusbaserte *Longman Grammar of Spoken and Written English* (Biber et al. 1999), hvor sammenligningen mellom konversasjon og tre typer skriftlige teksttyper går som en rød tråd gjennom hele boken. Det ville være enkelt å ramse opp mye mer, både bøker og artikler, men jeg nøyer meg med å vise til to nye tekster som fokuserer på betydningen av korpus i studiet av talespråket (og språket generelt): en artikkel av Michael Halliday (2004) og Jan Svartviks foredrag ved ICAME 25 (Svartvik 2004).

4.2 Ordstudier

Ett annet område hvor korpus har hatt stor betydning er ordstudier. Pioneren er John Sinclair, som brukte korpus for ordstudier allerede rundt 1970 (Sinclair et al. 1970, 1972). For ordstudier blir et korpus med en million ord fort altfor lite. I en artikkel fra begynnelsen av 1980-tallet understreker Sinclair at vi trenger store korpus (Sinclair 1982). Her introduserer han også begrepet *monitor corpus*, dvs. et dynamisk korpus som følger språkutviklingen.

Dette ble begynnelsen på arbeidet med COBUILD-prosjektet ved Universitetet i Birmingham som etter flere år førte til en ny ordbok som var nyskapende på flere måter (se Sinclair 1987). Utvalget av ord og eksempler var basert på et stort korpus. Hensikten var at beskrivelsen skulle gjenspeile språkbruken i korpuset. Definisjonene ble skrevet på en ny måte. COBUILD-ordboken innledet en ny epoke i leksikografien. Det er nå standard at nye ordbøker for engelsk bruker korpus. Vi kan endog tale om en revolusjon i leksikografien.

4.3 Kollokasjoner

Det er ikke bare metoden som er ny, men også måten ord blir beskrevet på. Et sentralt begrep er kollokasjon, dvs. måten ord kombinerer med hverandre. Begrepet går tilbake til den engelske lingvisten John Rupert Firth (1957), som skjelnet mellom *colligation* (forbindelsen mellom ord og den grammatiske konteksten) og *collocation* (forbindelsen mellom enkeltord). 'You shall know a word by the company it keeps,' sier Firth. Ordbøker har tradisjonelt behandlet sekvenser i større eller mindre grad, men ved hjelp av korpus kan det gjøres på en mye mer systematisk måte.

I en tankevekkende artikkel skriver Göran Kjellmer (1993) om det engelske ordet *sanction*, som kan ha to betydninger som virker helt motsatt: 'godkjenne/godkjenning' eller 'straff/straffe'. Det ene betegner en positiv, det andre en negativ reaksjon. Hvordan er dette mulig uten at det blir et alvorlig sammenbrudd i kommunikasjonen? Kjellmer undersøker språkbruken i korpus og viser hvordan semantiske indikatorer i konteksten peker i den ene eller andre retningen. I en nyere artikkel tar han opp et par andre eksempler på polysemi (Kjellmer 2003). Disse eksemplene lærer oss noe ikke bare om enkeltord, men om hvordan språket fungerer. Gjennom at ord kan ha flere betydninger blir språket mer økonomisk. Vi kan uttrykke

mer med mindre midler. Men budskapet kommer frem allikevel, på grunn av kollokasjoner og andre ledetråder fra konteksten.

Den som har vært mest innflytelsesrik i studiet av kollokasjoner, er nok John Sinclair (se for eksempel Sinclair 1991, 2003). Utgangspunktet er mønstre slik de kommer frem i konkordanser og i statistiske beregninger av kollokasjoner. La oss ta et eksempel fra en av hans artikler, substantivet *brink* (Sinclair 1999). Figur 2 viser ord som forekommer mest typisk sammen med *brink*. Jeg har brukt *British National Corpus* (BNC), som inneholder 100 millioner ord og er systematisk sammensatt for å representere forskjellige typer skrift og tale. Sinclair selv baserte sin analyse på et annet korpus, men det er godt samsvar mellom resultatene.

No.	Word	Total No. in the whole BNC	As collocate	In No. of texts	Mutual information value
1	teetering	65	15	15	13.34926905
2	teetered	43	9	9	13.20840629
3	poised	659	9	9	9.27053699
4	hovering	389	5	5	9.18304812
5	starvation	459	5	4	8.94432393
6	extinction	555	5	5	8.67033017
7	bankruptcy	998	7	7	8.30920519
8	collapse	2526	16	16	8.16210739
9	disaster	2770	13	12	7.72951619
10	revolution	4569	5	4	5.62901217

Figur 2 Ord i BNC som forekommer mest typisk sammen med *brink*

Hva kan vi se? Substantivet *brink* har en negativ *semantisk prosodi*, dvs. det forekommer mest typisk sammen med negative ord og uttrykk (et lignende norsk ord er *rand*). Styrken i forholdet er målt ved hjelp av *mutual information*, et statistisk mål som sammenligner hvor ofte de individuelle ordene er brukt totalt i korpuset, med hvor mange ganger de forekommer sammen med det ordet vi tar utgangspunkt i. For

eksempel: *teetering* forekommer totalt 65 ganger i BNC, og i 15 av forekomstene (i 15 forskjellige tekster) er det brukt sammen med *brink*. Det gir en høy verdi.

I listen finner vi verb som betegner en usikker posisjon (*teetering, teeter, poised, hovering*) og substantiver som betegner noe negativt (*starvation, extinction, bankrupcy, collapse, disaster, destruction*). Hvis vi kikker nærmere på eksemplene, ser vi at det vanligvis er snakk om å være på randen av noe forferdelig. Interessant nok finner vi enkelte eksempler som avviker fra hovedmønstret. Her er et par fra BNC:

(5) As Bosnia pulls back from the *brink of peace*, we look at why the Bosnian Serbs rejected the Vance-Owen peace plan, at what the West might do next, and at the possibility now raised of a wider Balkan war.

(6) And Sam is a lively girl on the *brink of adulthood*. She's entitled to find out what it means, don't you think.

(7) Malle establishes an authentic mood of unspoken suspicions and everyday secrecy, drawing upon performances, decor, even nature itself, to paint a portrait of childhood on the *brink of discovery*.

Gjennom å velge slike uttrykk kan skribenten angi at det ligger noe mer bak ordene, for eksempel at freden på Balkan i eksempel (5) er usikker.

John Sinclair mener at mye av betydningen i språket ligger i sekvenser, med noe innbygget variasjon. Det er disse sekvensene som er *lexical items*, ikke enkeltordene. Han trekker den radikale konklusjonen at vi må tenke om hele vår måte å beskrive språket på. Vi må se på språket med nye øyne og ikke være bundet av forutfattede meninger. *Trust the text* er tittelen på hans siste bok (Sinclair 2004). Denne måten å nærme seg språket på kaller Sinclair *corpus-driven*. Grensen mellom leksikon og grammatikk blir utvisket. Grammatikken blir en form for *pattern grammar*, slik den er beskrevet i Hunston & Francis (2000). Altså: studier av korpus kan tvinge oss till å stille grunnleggende spørsmål om språket og vår syn på språket.

4.4 Syntaks

Her vil jeg bare nevne et korpus som er særlig godt egnet for å studere syntaks av den mer tradisjonelle typen. Jeg sikter til den britiske delen av ICE-prosjektet, eller *ICE-GB*. Korpuset er unikt på mange måter. Det er systematisk sammensatt slik at forskjellige typer tale og skrift er representert. Hoveddelen av tekstene består av transkripsjon av tale, og lydopptakene er også digitalisert, slik at det er mulig å gå fra

transkripsjonen til det opprinnelige lydopptaket. Korpuset er syntaktisk merket ved hjelp av et system for syntaktisk merking utviklet under ledelse av Jan Aarts, Universitetet i Nijmegen, og merkingen er nøye kontrollert. Totalt er det over 83,000 syntaktiske trær. Det er utviklet et søkeprogram, *ICECUP*, for å utforske korpuset.

Ved hjelp av ICECUP kan vi søke etter ting som det er umulig, eller svært vanskelig, å finne automatisk i et vanlig korpus, for eksempel foranstilte objekter eller relativsetninger uten relativmarkør (*zero relatives*). Vi kan også bygge opp trær og søke etter alt i korpuset som samsvarer med disse. Med ICE-GB har vi svært gode muligheter til å stille interessante spørsmål, ikke bare om syntaks men også om forholdet mellom grammatikk og sosiolingvistiske og tekstlingvistiske variabler (alder, kjønn, teksttype, situasjonstype osv.). For eksempel: er det riktig, slik det ofte er blitt hevdet, at halespørsmål (av typen *isn't it*) blir brukt særlig av kvinner? Se videre Nelson et al. (2002).

5. Bruk av flerspråklig korpus

I de siste 10-15 årene er det blitt en stadig økende interesse for flerspråklige korpus. I Oslo laget vi *English-Norwegian Parallel Corpus* (ENPC), et engelsk-norsk parallellkorpus, for circa 10 år siden. Nå pågår det arbeid med flere språk innenfor rammen av *Oslo Multilingual Corpus* (OMC). Jeg viser her til Cathrine Fabricius-Hansens innlegg.

Termen parallellkorpus er ikke helt entydig. Den kan bety korpus med originaltekster og oversettelser. Slik blir termen brukt i språkteknologien. Andre bruker termen om korpus med sammenlignbare originaltekster i to eller flere språk. ENPC er et parallellkorpus i begge betydningene. Det består av originaltekster på både norsk og engelsk og oversettelser til det andre språket. Derfor er det velegnet både for kontrastive studier og oversettelsesstudier. Det håper jeg å kunne vise ved hjelp av noen eksempler. Presentasjonen blir nødvendigvis ganske kort. Se videre Johansson (1998, 2004).

5.1 Oversettelsesparadigmer

Med oversettelsesparadigme mener jeg her ord eller uttrykk i kildespråket og tilsvarende former i målspråket. Vi pleier å bruke termen *korrespondanser*. Tabell 1

viser oversettelser av den norske konsessive markøren *likevel* til engelsk (basert på Fretheim & Johansson 2002). Her er det en rekke ting vi kan observere:

- Det er mange oversettelser, vanligvis mye mer variasjon enn vi finner i en ordbok.
- Oversettelsene varierer avhengig av posisjonen i setningen. De vanligste oversettelsene i initial posisjon er: *all the same, even so, nevertheless*, *still*, og *yet*. I final posisjon dominerer *after all* og *anyway*, som ikke (*after all*) eller bare unntaksvis (*anyway*) blir brukt som oversettelse av *likevel* i initial posisjon.
- I medial posisjon er det mest variasjon. Her er det også mange nullkorrespondanser, dvs. tilfeller hvor det ikke er registrert noen form i oversettelsen som svarer til *likevel* i originalteksten.

Slike oversettelsesparadigmer må også tolkes, og det er det vi gjør i vår artikkel, men det ville føre altfor langt å gå inn på dette her. La oss bare konstatere at vi kan sette opp slike paradigmer med utgangspunkt i et tospråklig korpus.

Tabell 1 Oversettelser av *likevel* i forskjellige posisjoner (ENPC, skjønnlitteratur)

Form	Initial	Medial	Final
after all	0	3	**8**
all the same	**7**	6	1
anyway	1	7	**20**
but	2	5	2
even so	**8**	5	0
however	0	1	0
just the same	1	0	0
nevertheless	**8**	9	1
nonetheless	3	3	0
really	0	2	4
still	**15**	11	0
yet	**8**	3	0
Ø	5	**17**	6
Andre	9	7	5
Totalt	67	79	47

Ved hjelp av et tospråklig korpus med oversettelser i begge retningene kan vi også beregne *mutual correspondence*, eller gjensidig korrespondanse mellom ord og uttrykk i de to språkene, slik Bengt Altenberg viser i en artikkel om bindeadverbialer i engelsk og svensk (Altenberg 1999). Korrespondansen varierer mellom 0 % (ingen korrespondanse) og 100 % (full korrespondanse). I praksis er det nesten aldri full korrespondanse, men forskjellige grader avhengig av hvilke ord eller uttrykk vi sammenligner. Korrespondansene kan også være asymmetriske og avhenge av oversettelsesretningen (*translation bias*). Ta for eksempel svensk *emellertid* och engelsk *however*. Hvis vi utgår fra *emellertid* blir korrespondansen 81 %, dvs. fire av fem eksempler bli oversatt med *however*. Utgår vi derimot fra *however* blir korrespondansen 47 %, dvs.under halvparten blir gjengitt med *emellertid*. Gjennom å arbeide på denne måten kan Altenberg vise ikke bare hvordan enkelte former korresponderer, men relasjonene mellom systemene av konnektorer i de to språkene.

5.2 Semantiske speil

Korrespondansene viser ikke bare hvilke ord og uttrykk i de to språkene som svarer til hverandre, de viser også noe om de enkelte språkene. Det tilsynelatende uskyldige ordet *likevel* viser seg å kunne uttrykke ganske forskjellige betydninger hvis vi ser på korrespondansene i engelsk. I et enspråklig korpus kan vi enkelt studere former og formelle mønstre, men betydningene er det ikke så lett å få øye på. Noe av det som er mest interessant med oversettelseskorpus, er at de kan vise betydninger, slik de fremgår av oversetternes valg. Vaghet og tvetydighet blir avslørt gjennom oversettelsen. Men speilbildet som oversettelsen gir, er ikke perfekt. Som alltid i bruk av korpus må vi vurdere våre data.

Et eksempel på bruken av oversettelse for å studere betydning finner vi i en artikkel av Karin Aijmer om adverb som betegner sikkerhet og usikkerhet (Aijmer 2002). I oversettelsens speil ser vi for eksempel at svensk *säkert* kan svare både til *certainly* og *probably*. Aijmer kobler dette sammen med en utvikling fra sikkerhet til usikkerhet som vi finner også ved andre uttrykk, både i svensk og i andre språk.

I et arbeid som kom ut for et år siden, viser Dirk Noël hvordan ‘translators, through the linguistic choices they make, inadvertently supply evidence of the meanings of the forms they are receiving and producing’ (2003: 757). På grunnlag av oversettelser i det kanadiske Hansard-korpuset (engelsk-fransk) viser han at former

som *BE said to* og *BE reported to* fungerer som en form for hjelpeverb. Han tolker dette som et resultat av en grammatikaliseringsprosess.

Andre eksempler på bruken av flerspråklig korpus for å studere betydning er Aijmer og Simon-Vandenbergen's (2003) undersøkelse av den engelske diskursmarkøren *well* og Åke Viberg's mange verbstudier, for eksempel sammenligningen av engelsk *go* og svensk *gå* (Viberg 1996). Det mest radikale forsøket på å studere semantikk på grunnlag av et tospråklig korpus er Helge Dyvik's 'semantic mirrors'-prosjekt (Dyvik 1998). Jeg regner med at dette er kjent, og går derfor ikke nærmere inn på det. Hovedmålet med prosjektet er å bygge et ordnett og å legge et nytt grunnlag for semantikken.

5.3 Nullkorrespondanse

Det er en vanlig erfaring at ord eller uttrykk i kildeteksten ikke har noen formell motsvarighet i målteksten. Vi har allerede sett et eksempel på slik nullkorrespondanse (se 5.1). I undersøkelsen av *well* fant Aijmer og Simon-Vandenbergen 21 % nullkorrespondanse ved engelsk-svensk oversettelse og 7 % ved oversettelse mellom engelsk og nederlandsk. Selv fant jeg 16 % nullkorrespondanse i oversettelse til norsk (Johansson, under utgivelse).

Nullkorrespondanse kan også gå i den andre retningen, dvs. en oversetter kan legge til noe uten at det finnes en formell korrespondanse i originalteksten. Jeg fant 21 % nullkorrespondanse ved *well* i oversettelser fra norsk til engelsk, dvs. hver femte forekomst av *well* hadde ikke noen formell motsvarighet i den norske teksten.

Grunnen til nullkorrespondanse kan være at det finnes en form for kompensasjon i teksten. Eksempler på nullkorrespondanse ved oversettelse fra engelsk til norsk:

(8) "But you have at some point to understand that your father is not prepared any longer to share his ill-gotten gains with Jasper and all his friends."
"*Well, at least* he is prepared to see they are ill-gotten," said Alice earnestly. (DL2, orig)
"Men før eller senere vil du likevel bli nødt til å innse at din far ikke lenger er innstilt på å dele sine urettmessig ervervede rikdommer med Jasper og alle vennene hans."
"Han er *i det minste* i stand til å innrømme at de er urettmessig ervervet," sa Alice alvorlig.

(9) Sonia wasn't daft. *Well, not then, anyway*, unless that's what the abandoning of your own life for your children can be called. (FW1, orig)
Sonia var ingen tosk. *Ikke da i hvert fall*, bortsett fra hva det at du kaster bort livet ditt for barnas skyld, kan bli kalt.

Bortfallet av en oversettelse for *well* blir her kompensert gjennom *i det minste* og *i hvert fall*. Aijmer og Simon-Vandenbergen fant lignende eksempler.

I mange tilfeller er det ikke mulig å finne en god grunn til at det er nullkorrespondanse, og det er tydelig at teksten uttrykker oversetterens tolkning av situasjonen. I de neste to eksemplene er *well* lagt til av oversetteren, uten at det finnes noen parallell i den norske originalteksten:

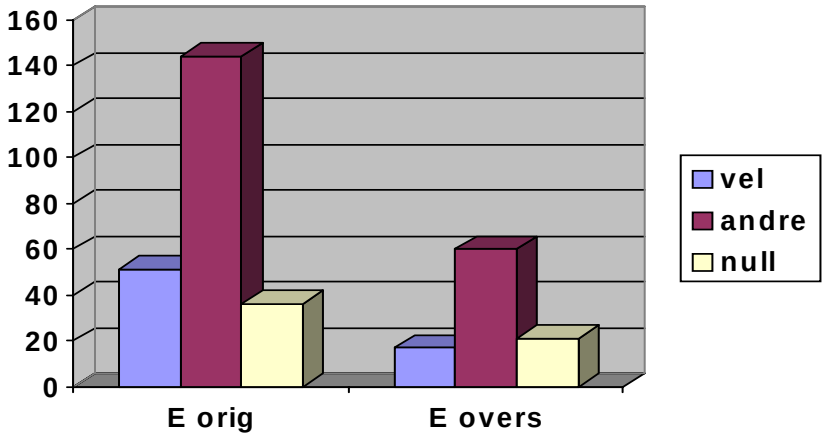
(10) "Jaså, forsvunnet fra et aldershjem?
Hvem meldte henne savnet?
Javel.
Vi snakkes ved." (EG1, orig)
"Reported missing from an old people's home?
Who reported it?
Oh?
Well, we'll talk about that later."

(11) "Har dere tatt knekkebrød igjen?"
Hildegun himlet lidende mot taket og svarte med uforskammet høflighet:
"Neida, mor. Vi drikker bare nypete."
"Den koster også penger. Dere har brukt strøm." (BV1, orig)
"Have you been at the crisp-bread again?"
Hildegun rolled her eyes in suffering towards the ceiling and answered with brazen politeness.
"No, mother, we haven't. We're just drinking rose-hip tea."
"*Well* that costs money too. You've been using electricity."

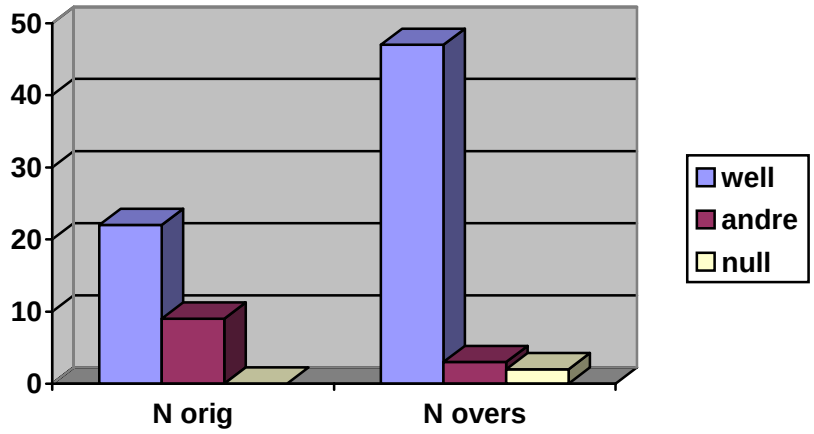
I (10) blir *well* brukt for å avslutte en telefonsamtale. Eksempel (11) begynner med et bebreidende spørsmål. Hildegun prøver å forsvare seg gjennom å bruke *bare* (tilsvarende *just* i den engelske teksten). Moren fortsetter med en annen kritisk bemerkning, men legger til et dempende *well*, uten noen parallell i den norske teksten. Slik viser oversettelsen noe om forskjeller i språkbruk mellom engelsk og norsk.

5.4 Oversettelsesspråk

Siden ENPC inneholder både originaltekster og oversettelser og parallelle originaltekster, er det godt egnet for å studere avvik mellom oversettelser og originaltekster på det samme språket. Som eksempler kan vi bruke engelsk *well* og norsk *vel*. Figur 3 og 4 gir en oversikt over ordene når de blir brukt som diskursmarkører, sammen med korrespondansene i det andre språket.
Figur 3 Engelsk *well* og norske korrespondanser (ENPC, skjønnlitteratur)



Figur 3 Engelsk *well* og norske korrespondanser (ENPC, skjønnlitteratur)



Figur 4 Norsk *vel* og engelsk korrespondanser (ENPC, skjønnlitteratur)

Well er mye mindre brukt i oversettelser enn i engelske originaltekster, for å være eksakt: 98 kontra 231 eksempler. Siden korpuset er balansert, kan vi

sammenligne absolutte tall. For *vel* er forholdet det omvendte: 52 eksempler i oversettelsene mot 31 i de norske originaltekstene. *Well* svarer ofte til *vel*, men vanligvis til noe annet, og nullkorrespondanse er ikke uvanlig (se figur 3). *Vel* derimot svarer for det aller meste til *well*, og det er nesten ikke noen nullkorrespondanse (se figur 4). Vi kan tolke dette slik:

- *Well* har en videre bruk enn *vel*. Fordi det ikke finnes noe i norsk som svarer direkte til *well*, blir ordet underrepresentert i oversettelser. Det blir også mange korresponderende former og ganske mye nullkorrespondanse.
- *Vel* overlapper i bruk med *well*, og bruken blir utvidet i oversettelser sammenlignet med norske originaltekster. Det er lite variasjon i oversettelsene og nesten ingen nullkorrespondanse.

Se videre Johansson (under utgivelse).

Slike oversettelseseffekter har vi også registrert i mange andre tilfeller, og ikke bare for enkeltord. For eksempel har Hasselgård (2000) vist hvordan ordstillingen i oversettelser kan bli påvirket av originalteksten, og avvik i valg av tidsformer er påvist i en artikkel av Elsness (2003).

5.5 Lag ditt eget korpus

I de siste årene har det vært en økende interesse for bruken av korpus i opplæringen av oversettere. Det finnes endog internasjonale konferanser om 'Corpus Use and Learning to Translate', den siste i Barcelona i januar 2004. En ny tendens er at vordende oversettere får opplæring i å lage egne korpus for spesielle formål (se for eksempel Varantola 2003). Her ligger fokuset på fagspråk.

6. Innlærerkorpus

Den siste typen korpus som jeg vil nevne, er *International Corpus of Learner English* (ICLE), et prosjekt ledet av Sylviane Granger, Louvain-la-Neuve, Belgia (se Granger 1998). Vi kan se på dette som en utvikling fra ICE-korpuset, men i motsetning til ICE består ICLE av delkorpus som representerer engelsk som fremmedspråk slik det er brukt av studenter i forskjellige deler av verden: Finland, Italia, Polen, Spania,

Sverige, Tyskland, osv. I det siste er det også bygd et norsk delkorpus, som vi kaller NICLE.

Det er ikke mulig her å gå inn på hvordan korpusene ser ut og hvordan de blir brukt (et godt eksempel er Gilquin 2003). Jeg vil bare understreke to ting. Det første er at oppbyggingen av disse korpusene er et resultat av samarbeid mellom forskergrupper i mange land. Slik passer ICLE-prosjektet godt inn i rammen for ICAME. Det viktigste jeg vil si, er at det er spørsmålene som har styrt planleggingen og utviklingen av prosjektet. I hvilken utstrekning er språkbruken avhengig av brukernes morsmål? I hvilken utstrekning finner vi felles trekk uavhengig av morsmålet? Slik kan vi også stille grunnleggende spørsmål om språkinnlæring. Korpusbygging er ikke en tanketom virksomhet. Kanskje kunne vi si på hjemmesnekret latin: *quale caput, tale corpus*.

7. En fremtidsvisjon

Før jeg går over til min konklusjon, vil jeg nevne at korpus med fordel kan brukes ikke bare for studier av forskjellige språklige nivåer, men også i tekstanalyse. Her vil jeg særlig nevne en bok av Michael Stubbs (1996) og den nye boken av John Sinclair (2004). Og så en kort konklusjon. I de siste tiårene har det vært en rivende utvikling innenfor korpuslingvistikken – hvis vi skal bruke denne termen. Jeg tror dette bare er begynnelsen. Vi har sett mye gjennom korpus. Men alle har ikke vært villige til å se. Noam Chomsky uttalte i et intervju for noen år sedan:

> Bas Aarts: What about modern corpus linguistics?
> Chomsky: It doesn't exist. If you have nothing, or if you are stuck, or if you're worried about Gothic, then you have no choice. […] You don't take a corpus, you ask questions. (Bas Aarts, intervju med Noam Chomsky, jf. Aarts 2000: 5)

Jeg protesterer for all verden. Korpuslingvistikken eksisterer. Gjennom intelligent korpusbruk kan vi stille mange typer spørsmål og se nye ting. Men det er krevende å lage og bruke korpus. Det krever kunnskap om metode, det krever innsikt i språk, det krever fantasi, det krever arbeid.

Noe av det som er mest interessant med korpus, er at det er godt egnet ikke bare for forskning men også for undervisning, slik det fremgår av bidragene til konferansene om 'Teaching and Language Corpora', den siste i Granada i juli 2004. På engelsk snakker vi om *discovery learning*, hvor grensen mellom læring og forskning blir utvisket. Det betyr ikke at vi bare kan sette studentene foran

datamaskinen. De må få opplæring i hvordan de formulerer et spørsmål, hvordan de kan velge et egnet korpus, hvordan de kan analysere materialet og hvordan det de ser i korpuset kan få dem til å formulere nye spørsmål. Slik oppstår det en syklisk prosess, en type samtale mellom forskeren/studenten og korpuset.

I begynnelsen refererte jeg til Ibsens 'Å dikte er å se'. Den kjente litteraturforskeren Francis Bull legger til: 'Men man må se energisk' (Bull 1966). Om vi er villige til å se energisk, er det nesten ikke grenser for hva vi kan se. Her tillater jeg meg å sitere fra et foredrag som jeg holdt ved en konferanse tidligere i år (og mye av det jeg har sagt i dag, kommer fra dette foredraget). Emnet var 'Seeing through multilingual corpora':

> If we are prepared to look energetically into multilingual corpora, we can see correspondences across languages, we can see individual languages in a new light, we can pinpoint characteristics of translation, we can see meanings, we can see grammaticalisation, we can see collocations, we can see the intimate relationship between lexis and grammar. Seeing through corpora we can see through language.
> (Stig Johansson, foredrag ved ICAME 25, Verona, mai 2004)

Til sist en vesentlig ting: Man må ha noe som er godt egnet til å kikke i. Det er viktig å satse energisk og systematisk på å bygge opp det empiriske grunnlaget. For engelsk har vi mange typer korpus. I ti år har vi hatt et britisk nasjonalkorpus, med et bredt utvalg av tekster, både fra skrift- og talespråket. Tsjekkia har sitt nasjonalkorpus. Norge ligger langt etter i utviklingen. Men det kan skje mer også i Norge, kanskje også her i Kristiansand!

Referanser

Aarts, Bas. 2000. Corpus linguistics, Chomsky and fuzzy tree fragments. I: Christian Mair & Marianne Hundt (utg), *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, 5-13. Amsterdam: Rodopi.

Aarts, Jan & W. Meijs (utg). 1984. *Corpus linguistics. Recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.

Aijmer, Karin. 2002. Modal adverbs of certainty and uncertainty in an English-Swedish perspective. I: Hasselgård et al. (2002), 97-112.

Aijmer, Karin, Bengt Altenberg, & Mats Johansson (utg). 1996. *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*, Lund Studies in English 88. Lund: Lund University Press.

Aijmer, Karin & Bengt Altenberg (utg). 2004. *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*. Amsterdam: Rodopi.

Aijmer, Karin & Anne-Marie Simon-Vandenbergen. 2003. The discourse particle *well* and its equivalents in Swedish and Dutch, *Linguistics* 41:1123-1161.

Allén, Sture. 1992. Opening address. I: Svartvik (1992), 1-3.

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. I: Hilde Hasselgård & Signe Oksefjell (utg), *Out of corpora. Studies in honour of Stig Johansson*, 249-268. Amsterdam: Rodopi.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoeffrey Leech, Susan Conrad, & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bull, Francis. 1966. Å dikte er å se. I: *Vildanden og andre essays*, 29-37. Oslo: Gyldendal.

Chafe, Wallace. 1992. The importance of corpus linguistics to understanding the nature of language. I: Jan Svartvik (utg), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, 4-8 August 1991*, 79-103. Berlin: Mouton de Gruyter.

Dyvik, Helge. 1998. A translational basis fort semantics. I: Johansson & Oksefjell (1998), 51-86.

Elsness, Johan. 2003. A contrastive look at the present/preterite opposition in English and Norwegian, *Languages in Contrast* 3 (2000/2001): 3-40.

Firth, J.R. 1957. *Papers in linguistics 1934-1951*. London: Oxford University Press.

Fretheim, Thorstein & Stig Johansson. 2002. The semantics and pragmatics of the Norwegian concessive marker *likevel*: Evidence from the English-Norwegian Parallel Corpus. I: Leiv Egil Breivik & Angela Hasselgren (utg), *From the COLT's mouth … and others'. Language corpora studies in honour of Anna-Brita Stenström*, 81-101. Amsterdam: Rodopi.

Gilquin, Gaëtanelle. 2003. The Integrated Contrastive Model: spicing up your data, *Languages in Contrast* 3 (2000/2001): 95-123.

Granger, Sylviane. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. I: Aijmer *et al*. (1996), 37-51.

Granger, Sylviane (utg). 1998. *Learner English on computer*. London & New York: Longman.

Greenbaum, Sidney & Jan Svartvik. 1990. *The London Corpus of Spoken English: Description and research.* Lund Studies in English 82. Lund: Lund University Press.

Greenbaum, Sidney (utg). 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon.

Halliday, M.A.K. 2004. The spoken language corpus: A foundation for grammatical theory. I: Aijmer & Altenberg (2004), 11-38.

Hasselgård, Hilde. 2000. English multiple themes in translation. I: A. Klinge (utg), *Contrastive studies in semantics*, 11-38. Copenhagen Studies in Language 25. Frederiksberg: Samfundslitteratur.

Hasselgård, Hilde, Stig Johansson, Bergljot Behrens, & Cathrine Fabricius-Hansen (utg). 2002. *Information structure in a cross-linguistic perspective*. Amsterdam: Rodopi.

Hunston, Susan and Gill Francis. 2002. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Johansson, Stig (utg). 1982. *Computer corpora in English language research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig. 1998. On the role of corpora in cross-linguistic research. I: Johansson & Oksefjell (1998), 3-24.

Johansson, Stig. 2003. Noen tanker om datatyper i språkforskningen. Institutt for britiske og amerikanske studier, Universitetet i Oslo.
Johansson, Stig. 2004a. What is a person in English and Norwegian? I: K. Aijmer & H. Hasselgård (eds.), *Translation and corpora*. Gothenburg: Acta Universitatis Gothoburgensis.
Johansson, Stig. 2004b. Seeing through multilingual corpora. Foredrag ved 25th International Conference on English Language Research on Computerized Corpora (ICAME 25), Verona, mai 2004.
Johansson, Stig. Under utgivelse. How well can *well* be translated? On the discourse particle *well* and its correspondences in Norwegian and German. I: Karin Aijmer & Anne-Marie Simon-Vandenbergen (utg), *Pragmatic markers in contrast*. Amsterdam: Benjamins.
Johansson, Stig & Signe Oksefjell (utg). 1998. *Corpora and cross-linguistic research: Theory, method, and case studies*. Amsterdam: Rodopi.
Kjellmer, Göran. 1982. Multiple meaning and interpretation: The case of *sanction*, *Zeitschrift für Anglistik und Amerikanistik* 41: 115-123.
Kjellmer, Göran. 2003. Polysemy and ambiguity. I: Karin Aijmer & Britta Olinder (utg), *Proceedings from the 8th Nordic Conference on English Studies*, 9-24. Göteborg University: Department of English.
Leech, Geoffrey. 2004. Recent grammatical change in English: Data, description, theory. I: Aijmer & Altenberg (2004), 61-84.
Murray, K.M. Elizabeth. 1979. *Caught in the web of words*. Oxford: Oxford University Press.
Nelson, Gerald, Sean Wallis, & Bas Aarts. 2002. *Exploring natural language. Working with the British component of the International Corpus of English*. Amsterdam: Benjamins.
Noël, Dirk. 2003. Translations as evidence for semantics: An illustration, *Linguistics* 41: 757-785.
Sinclair, John. 1982. Reflections on computer corpora in English language research. I: Johansson (1982), 1-6.
Sinclair, John (utg). 1987. *Looking up. An account of the COBUILD Project in lexical computing*. London & Glasgow: Collins ELT.
Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, John. 1999. The computer, the corpus and the theory of language. I: Gabriele Azzaro & Margherita Ulrych (utg), *Transiti linguistici e culturali. Atti del XVIII Congresso nazionale dell'A.I.A. (Genova, 30 settembre – 2 ottobre 1996)*, 1-15. Trieste: E.U.T.
Sinclair, John. 2003. *Reading concordances*. London: Longman.
Sinclair, John. 2004. *Trust the text: Language, corpus, and discourse*. London & New York: Routledge.
Sinclair, John, Susan Jones, & Robert Daley. 1970, 1972. *English lexical studies*. Report to the Office of Scientific and Technical Information. Ny utgave, sammen med et intervju med John Sinclair (av Wolfgang Teubert): Krishnamurthy, Ramesh (utg), 2004. *English collocation studies: The OSTI Report*. London: Continuum Books.
Stubbs, Michael. 1996. *Text and corpus analysis*. Oxford: Blackwell.
Stubbs, Michael. 2002. *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.
Svartvik, Jan (utg). 1992. *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin & New York. Mouton de Gruyter.
Svartvik, Jan. 1996. Corpora are becoming mainstream. I: Jenny Thomas & Mick Short (utg), *Using corpora for language research*, 3-13. London: Longman.

Svartvik, Jan. 2004. Corpus linguistics 25+ years on. Foredrag ved 25th International Conference on English Language Research on Computerized Corpora (ICAME 25), Verona, mai 2004.
Tottie, Gunnel & Ingegerd Bäcklund (utg). 1986. *English in speech and writing: A symposium*. Studia Anglistica Upsaliensia 60. Stockholm: Almqvist & Wiksell International.
TradTerm: Revista de Centro Interdepartamental de Tradução e Terminologia, 10 (2004). Special issue on 'Translation and corpora'.
Varantola, Krista. 2003. Translators and disposable corpora. I: Federico Zanettin, Silvia Bernadini, & Dominic Stewart (utg), *Corpora in translator education*, 55-70. Manchester: St. Jerome Publishing.
Viberg, Åke. 1996. Cross-linguistic lexicology. The case of English *go* and Swedish *go*. I: Aijmer *et al*. (1996), 151-182.

Nettsteder: korpus

Bank of English: http://titania.cobuild.collins.co.uk/boe_info.html
British National Corpus (BNC): http://info.ox.ac.uk/bnc/
English-Norwegian Parallel Corpus (ENPC): http://www.hf.uio.no/iba/prosjekt/index.html
Oslo Multilingual Corpus (OMC): http://www.hf.uio.no/german/sprik/english/corpus.shtml
International Computer Archive of Modern and Medieval English (ICAME): http://helmer.aksis.uib.no/icame.html
International Corpus of English (ICE): http://www.ucl.ac.uk/english-usage/ice/
International Corpus of English, Great Britain (ICE-GB): http://www.ucl.ac.uk/english-usage/ice-gb/
International Corpus of Learner English (ICLE): http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl.html
Oslo Corpus: http://www.tekstlab.uio.no/norsk/bokmaal/
Research Unit for Variation and Change in English, Helsinki: http://www.eng.helsinki.fi/varieng/
Webcorp: http://www.webcorp.org.uk
World Wide Web Access to Corpora Project (W3-Corpora). http://clwww.essex.ac.uk/w3c/

Nettsteder: kurs, informasjon

Catherine N. Ball, Georgetown University: http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html
Michael Barlow, Rice University, Corpus linguistics: http://www.ruf.rice.edu/~barlow/corpus.html
Yvonne Breyer: http://www.corpus-linguistics.de/corpora/corp_parallel.html
David Lee, Bookmarks for corpus linguists: http://devoted.to/corpora
World Wide Web Access to Corpora Project (W3-Corpora). http://clwww.essex.ac.uk/w3c/

Konferanser

Corpus Use and Learning to Translate (CULT)
International Computer Archive of Modern and Medieval English (ICAME)
Practical Applications in Language Corpora (PALC)
Teaching and Language Corpora (TaLC)
Mfl.

Se informasjon på nettet.

Lingvistiske og diskursive studier i KIAP-korpuset. Om utvikling og bruk av et flerspråklig og flerdisiplinært korpus av vitenskapelige artikler

Kjersti Fløttum, Universitetet i Bergen

1. Innledning: Hvorfor et KIAP-korpus?

I en norsk lingvistisk artikkel, hentet fra KIAP-korpuset[1], om *ja* og *nei* som emneinnledere, heter det:

(1) Jeg vil hevde at "nei" i større grad enn "ja" markerer grenser, […]. (noling19)

I en annen som handler om norsk språk i USA, finner vi følgende utsagn:

(2) At det [antall informanter] vel er et noe spinkelt grunnlag for ei språkvitenskapelig avhandling, får så være. Det overraskende er at forfatteren føler det nødvendig å gjøre leseren oppmerksom på at […]. Det veit vi jo! (noling42)

Og i en tredje, som handler om den pragmatiske partikkelen *bare*, uttrykker artikkelforfatteren seg slik:

(3) Man kunne kanskje tenke seg at en versjon uten svarord, men med *bare* i det påfølgende forklarende utsagnet ville bli oppfattet som et tydeligere negativt svar enn (17) B, men slik er det ikke. (noling21)

Disse eksemplene illustrerer noe av årsaken til at vi i KIAP-prosjektet (se del 2) interesserer oss særlig for *stemmene* som kommer til uttrykk i vitenskapelig diskurs (eller akademisk prosa). Diskursen er langt fra nøytral, objektiv og berøvet for personlige innslag, slik en tradisjonell oppfatning kanskje ville hevde. Som eksemplene ovenfor viser er også den individuelle variasjonen stor. I KIAP har vi som følge av blant annet slike observasjoner, valgt å konsentrere oss om forskningsspørsmål som disse:

1) Hvordan manifesterer artikkelforfatter(ne) seg?

(jfr. *Jeg vil hevde at…* (1), *Det veit vi jo!* (2), *Man kunne kanskje tenke seg at …* (3))

2) Hvordan kommer andre enn forfatter(ne) til uttrykk?
(jfr. *Det overraskende er at forfatteren føler det nødvendig å …* (2))
3) Hvordan uttrykker forfatter(ne) holdninger, gjennom presentasjon av egen forskning?
(jfr. *…, men slik er det ikke*. (3)).

Jeg innleder på denne måten fordi jeg ønsker å understreke at det er bestemte lingvistiske og diskursive problemstillinger, knyttet til sjangeren vitenskapelig artikkel, som danner utgangspunktet for KIAP-prosjektet, og at korpusstudier ble valgt som *metodisk tilnærming* for å finne svar på de spørsmålene vi reiser. Vi startet ikke med å utvikle et korpus for deretter å formulere passende problemstillinger til studier av dette. Korpuset er altså en konstituerende del av vår metode. En slik korpustilnærming synes å stemme med hvordan korpusstudier blir brukt i mange sammenhenger i dag. Dersom vi skal plassere oss innenfor korpuslingvistikk, vil det være i tråd med Stig Johanssons utsagn i en artikkel allerede fra 1995 om korpuslingvistikkens studieobjekt:

”[…] the object of corpus linguistics is *not* the study of corpora. It is rather the study of language through corpora, […].” (Johansson 1995:19)

I sin diskusjon av framveksten av korpusstudier, karakteriserer Tony McEnery & Andrew Wilson ([1996]/2003: 1) likeledes korpuslingvistikk som ”an increasingly prevalent methodology in linguistics”.

For å utføre våre studier, innenfor en bestemt sjanger, nemlig den vitenskapelige artikkelen, ønsket vi tilgang til et større korpus. Dermed ble opprettelsen av KIAP-korpuset en integrert del av prosjektet.

Min disposisjon for resten av artikkelen blir som følger: Jeg presenterer i del 2 noen av hovedtrekkene ved KIAP-prosjektet (formelle karakteristika, teoretisk utgangspunkt); deretter redegjør jeg i del 3 for oppbygging og utvikling av KIAP-korpuset, og til slutt presenterer jeg i del 4 noen enkeltresultater fra våre studier.

---

[1] KIAP er et akronym for prosjekttittelen *Kulturell Identitet i Akademisk Prosa: nasjonal versus disiplinavhengig*. Se del 2 av artikkelen.

Referansene er ordnet i to deler. Den første inneholder referanser brukt i denne artikkelen til andre enn KIAP-publikasjoner; den andre er en komplett liste over KIAP-publikasjoner (som også ligger på KIAPs nettside http://kiap.aksis.uib.no); det er referert til enkelte av disse i artikkelen.

2. Om KIAP-prosjektet

2.1 Generelt

KIAP står for *Kulturell Identitet i Akademisk Prosa: nasjonal versus disiplinavhengig*. Prosjektet er NFR-finansiert i perioden 2002–2005 og har sin adresse ved Romansk institutt, Universitetet i Bergen. Administrasjonen av prosjektet samt det informasjonsteknologiske arbeidet med korpuset ligger ved Aksis (Avdeling for kultur, språk og informasjonsteknologi, UNIFOB as). I tillegg til en timelønnet forskningsassistent, cand. philol. Anje Müller Gjesdal, er vi tre hovedmedlemmer: Førsteamanuensis og anglisist Trine Dahl, NHH, postdok og nordist Torodd Kinn (opprinnelig fra Seksjon for lingvistiske fag) samt Kjersti Fløttum, leder for prosjektet, professor i fransk språkvitenskap ved Romansk institutt. Under pilotprosjektet (2001) og i første fase av hovedprosjektet var postdok og nordist Kjersti R. Breivega med i prosjektet.

Vi har ulike samarbeidspartnere: Prosjektmiljøet *Norsk Sakprosa* i Oslo, ledet av K. L. Berge, prosjektgruppen *Equipe Linguistique des textes* ledet av F. Rastier, Paris, prosjektet *Representing Specialized Knowledge*, ledet av M. Koskela, Vaasa; og en forskergruppe ledet av Agnès Tutin og Francis Grossmann, LIDILEM, Grenoble, som studerer vitenskapelig diskurs i et mer didaktisk perspektiv. Det finnes også flere satelittprosjekter som er/har vært knyttet til KIAP (10 fordelt på masterstudenter og doktorgradskandidater). Jeg viser til nettsiden for ytterligere informasjon om prosjektet, deltakere, satelittprosjekter, aktiviteter, publikasjoner m.m.: http://kiap.aksis.uib.no/ (for pilotprosjektet, se også Breivega, Dahl & Fløttum 2002).

2.2 Problematikk

Vår overordnete problemstilling er formulert i spørsmålet om det finnes noe man kan kalle *kulturelle identiteter* i vitenskapelig diskurs, og i hvilken grad disse eventuelt er knyttet til nasjonalspråk og/eller til disiplin. Målet er å beskrive sjangeren *vitenskapelig artikkel* ut fra bestemte språklige og diskursive virkemidler, som i hovedsak kan knyttes til argumentasjon, og som kan peke på likheter og forskjeller mellom artikler skrevet på ulike språk og innen ulike fag/disipliner. I dette dobbelt komparative prosjektet studerer vi artikler hentet fra tre språk (engelsk, fransk og norsk) og fra tre disipliner (samfunnsøkonomi, språkvitenskap og medisin).

I KIAP er vi særlig interessert i individene som befinner seg bak selve forskningsaktiviteten og deres stemmer. Vi mener en undersøkelse av hvordan disse individene manifesterer seg, hvordan de uttrykker sine meninger, holdninger og verdier, og hvordan de interagerer med hverandre og omkringliggende diskurssamfunn, kan gi oss en god pekepinn på de eventuelle kulturelle identitetene vi er ute etter (se Dahl 2004c). Denne fokuseringen gjenspeiles i de tre forskningsspørsmålene nevnt ovenfor (se 1).

Jeg skal komme tilbake til resultater (se 4), men nevner allerede nå at våre første analyser stemte godt overens med vår første hypotese (basert på en pilotstudie, se Breivega, Dahl & Fløttum 2002), nemlig at disiplin er viktigere enn språk for utviklingen av identiteter i vitenskapssamfunnet. Vi foreslo tre karikerte profiler av medisinerforfatteren som fraværende, godt skjult bak teksten; av økonomen som tydeligere tilstede, men på en beskjeden måte; og av lingvisten som den klarest polemiske og mest tilstedeværende i teksten.

2.3 Teoretisk ramme

Vi har ikke én teoretisk ramme, men forholder oss til et sett av teoretiske tilnærminger, som vi håper ikke virker sammensatt på en ukontrollert eklektisk måte. Vi ønsker å bidra med språklige studier av en bestemt sjanger; det medfører at vi bør ha et blikk for den sosioprofesjonelle praksisen som de aktuelle tekstene – de vitenskapelige artiklene – skapes i. Selv om vi også i hovedsak arbeider fra mikro- til makronivå (i oppadstigende retning fra språklige manifestasjoner til den større konteksten), skal jeg

her raskt gjennomgå våre viktigste teoretiske innfallsvinkler i en nedadstigende retning, det vil si fra makro- til mikronivå.

For det første, vitenskapelige artikler er retoriske i den forstand at de representerer en diskurs som blir konstruert for å påkalle kooperative holdninger og handlinger men også for at forfatteren kan posisjonere seg i det aktuelle diskurssamfunnet. Det endelige retoriske mål for en vitenskapelig artikkel er å skape effekter som overbeviser mottakerne i den grad at artikkelen blir ansett som en del av den aktuelle disiplinens litteratur. I KIAP henter vi de overordnede vitenskapsretoriske observasjonene i hovedsak fra Lawrence J. Prellis bok *A Rhetoric of Science: Inventing Scientific Discourse* (1989). Når man som Prelli og mange med ham er kommet forbi stadiet hvor man ser på retorikk og vitenskap som motstridende begreper, og altså aksepterer at vitenskapelig virksomhet og særlig vitenskapelig diskurs er retorisk, blir det viktig å se nærmere på hva denne retorikken består i.

Når vi slik har definert den type diskurs vi arbeider med som retorisk, må vi gå et hakk ned og se nærmere på den spesifikke sjangeren vi studerer. For å komme fram til en forklaring på de lingvistiske observasjonene vi har gjort, er det selvfølgelig nødvendig å trekke inn den konteksten den vitenskapelige artikkelen er laget i. Sjangeren utgjør en kontekst i seg selv. Den vitenskapelige artikkelen har utviklet seg i et sosioprofesjonelt samfunn med egne normer og regler. Uten at jeg skal gå i detalj om sjangerbegrepet her, betrakter jeg sjangeren som en sosial praksis som gjentar seg og som realiseres språklig etter mer eller mindre faste mønstre. Tekstene som utgjør denne realiseringen inneholder normalt bestemte språklige og tekstlige trekk, uten at disse utgjør et fast og ufravikelig mønster. Jeg slutter meg til Bakhtins oppfatning (1986) om at sjangeren er et diskursobjekt som på samme tid er dynamisk (som kan forandre seg og til og med forsvinne) og relativt stabilt (som karakteriseres av bestemte ”regler” som kan være av vidt forskjellig natur – fra lingvistiske til kulturelle). For øvrig er selvfølgelig John Swales’ nå klassiske studie fra 1990 et viktig utgangspunkt for KIAP-prosjektet; likeledes sjangerstudiene utført innenfor prosjektmiljøet *Norsk Sakprosa* (se for eksempel Berge 2001, 2003, Tønnesson 2003; samt de franske innfallsvinklene til sjanger utviklet av Adam 1999 og Rastier 2001. Sjangeren er en viktig forklaringsfaktor for oss, men vi mener også at våre studier

bidrar til en bedre utviklet språklig karakteristikk av den sjangeren det dreier seg om.

Vi nærmer oss nå mikronivået. Det er her KIAP prøver å samle seg om en relativt enhetlig teoretisk innfallsvinkel. Med utgangspunkt i det syn at en vitenskapelig tekst blir skapt i en særegen, flerstemmig utsigelsessituasjon, finner vi det interessant å anvende den franskinspirerte enonsiative (eller utsigelsesmessige) tilnærmingen. De fleste som arbeider "enonsiativt", eller med utsigelsen som utgangspunkt, fokuserer på språket i bruk og dets relasjon til brukere og kontekst (med Ducrot 1984 som et klart unntak; han har alltid gjort det klart at hans studieobjekt er *la langue*). Benveniste (1966) definerte utsigelsen som "la mise en fonctionnement de la langue par un acte individuel d'utilisation" og vektla slik utsigelse som en individuell handling (og en historisk hendelse) hvor språksystemet blir satt i funksjon. Resultatet av denne handlingen er *ytringen,* eller *l'énoncé*, (betraktet som en aktualisering av den abstrakte syntaktiske setningen).

Denne teoretiske innfallsvinkelen er basert på det syn at en ytring nødvendigvis inneholder spor etter handlingen som produserer den, det være seg spor etter konteksten den blir produsert i (tid og sted; og eventuelt andre utsigelser), og/eller spor etter utsigelsens protagonister, nemlig sender- og mottakerinstans (lokutør og allokutør).[2]

En videreutvikling og spesifisering av det utsigelsesbaserte perspektivet er representert i den lingvistiske polyfoniteorien slik den kommer til uttrykk i ScaPoLine (forkortelse for "théorie SCAndinave de la Polyphonie LINguistiquE"; se Nølke, Fløttum & Norén 2004).

Denne har vi anvendt i ulike sammenhenger hvor flerstemmighet er et relevant perspektiv (særlig i analyse av nektingskonstruksjoner, av kontrastive konnektorer (som *but/however, men, mais)* og av epistemiske uttrykk) og med interessante

[2] De fleste av utsigelsens bestanddeler kan relateres til språksystemet og blir slik av mer generell interesse (for eksempel personlige pronomen, verbtider, adverbialer). Sophie Marnette (2001: 244) sier det på denne måten: "There is a crucial distinction to be made between each individual enunciation (seen as a single historical event) and the general phenomenon of enunciation, namely a stable system which emerges from the multiplicity of all the individual acts of enunciation. To study enunciation is, thus, to study a set of specific mechanisms through which the locutor converts the abstract system of *langue* into *discours*. These mechanisms can be studied via the traces they leave in their products, the utterances."
Språket inneholder en lang rekke uttrykk som utgjør spor etter selve utsigelsen; rekkevidden av dette perspektivet blir omfattende og dermed interessant som overordnet teoretisk perspektiv.

resultater, som jeg ikke kan gå videre inn på her (se Fløttum 2004 a, f, under trykking a, b, f).

For å oppsummere: Valg av teoretisk ramme er avhengig av hvilke spørsmål man stiller. Når man velger korpusstudium som metode, må disse spørsmålene i sin tur være formulert slik at man ved operasjonelle søkemåter kan finne svar på spørsmålene. Både den sjangerteoretiske og den semantisk-pragmatiske innfallsvinkelen (som utsigelsen og det polyfoniske utgjør) har vært styrende for vår koding av tekster så vel som for utarbeiding av søkeprogrammet vi bruker på KIAP-korpuset.

Endelig må nevnes at på tvers av den teoretiske rammen presentert ovenfor, anlegger vi en kontrastiv eller komparativ innfallsvinkel på studiene våre. Dette gjør til sammen vår teoretisk-metodiske ramme relativt kompleks, noe det ikke er plass til å gå nærmere inn på her. Jeg nøyer meg med å nevne at vi stadig må minne oss selv om at når vi sammenligner den diskursive vitenskapelige språkbruken i tre ulike språk som engelsk, fransk og norsk, må vi sørge for at vi studerer språklige størrelser som er sammenlignbare (for relevante betraktninger rundt disse spørsmålene, se Fabricius-Hansen 1991.) Men ved å støtte oss på et korpus som er sammensatt av eksempler av samme type tekstsjanger, mener vi å ha sikret et viktig sammenligningsgrunnlag.

3. Om KIAP-korpuset

Denne redegjørelsen er basert på Fløttum (2002) samt en beskrivelse under arbeid av Torodd Kinn.

3.1 Valg av språk, disiplin og tekster

Den valgte komparative innfallsvinkelen er begrunnet med den åpenbare mangelen på denne typen studier av vitenskapelig diskurs. Svært mange av studiene innenfor området er relatert til ett språk, i hovedsak engelsk (se blant annet Bazerman 1988, Swales 1990, Hyland 2000 samt referanser i disse til andre), og ofte til kun ett fag (se for eksempel Salager-Meyer 1998; for andre referanser, se bibliografi på KIAPs nettsted). Situasjonen har endret seg noe siden KIAP startet (med pilotprosjekt i 2001), men det er fortsatt et åpenbart fravær av studier hvor *både* språklige og disiplinmessige forskjeller undersøkes *i sammenheng*.

Innenfor det store feltet *sakprosa* (Berge 2001, Svensson 1999) er *vitenskapelig diskurs* eller *akademisk prosa* et delområde som igjen kan dekke ulike sjangrer som monografien, avhandlingen, den vitenskapelige artikkelen, rapporten, og andre. KIAP har avgrenset sitt studieobjekt til sjangeren *vitenskapelig artikkel.* Denne avgrensingen er basert på særlig ett forhold, nemlig den nøkkelstilling den vitenskapelige artikkelen har blant de akademiske sjangrene (Swales 1990).

Korpuset består av totalt 450 artikler. De tre språkene – engelsk, fransk og norsk – er representert med 150 artikler hver; disse er igjen fordelt med 50 artikler på hver disiplin, samfunnsøkonomi, språkvitenskap og medisin. Av de norske artiklene er 17 skrevet på nynorsk (1 fra økonomi, 12 fra lingvistikk og 4 fra medisin).

I tillegg til å vise til egne faglige interesser, begrunner vi valget av de tre språkene slik: Det engelske språks posisjon innenfor akademia gjør at det blir et selvfølgelig referansepunkt for kontrastive studier av akademisk språkbruk. Norsk er interessant fordi det er et lite språk, og fordi det knapt er gjennomført kontrastive studier av vitenskapsspråk der norsk inngår (det er imidlertid utført viktige studier av norsk eller andre nordiske språk som vi støtter oss på; se bl.a. Laurén, C. & J. Myking (red.) 1999, Breivega 2003, Tønnesson 2003). Fransk er i motsetning til norsk et stort verdensspråk med en mye større vitenskaplig produksjon, men det har på langt nær den utbredelse som engelsk har. I tillegg blir fransk gjerne betraktet som et språk som ikke så lett lar seg påvirke utenfra, for eksempel av engelsk. Endelig er det i liten grad gjennomført kontrastive studier mellom fransk og andre språk (se likevel Vassileva 2000 og Salager-Meyer & Zambrano 2001).

Valget av de tre disiplinene fortjener en mer utførlig diskusjon, som jeg ikke kan gå i dybden på her (se Fløttum 2002 og Dahl 2004c). Hovedbegrunnelsen er imidlertid at de tre disiplinene er hentet fra tre forskjellige vitenskapelige hovedområder (naturvitenskap, samfunnsvitenskap og humaniora); men det er selvfølgelig grunn til å diskutere i hvilken grad de er representative for disse hovedområdene. En annen begrunnelse er at prosjektdeltakerne hadde en viss kjennskap til disse disiplinene, ved tidligere forskning, egen fagkompetanse eller institusjonstilhørighet. Endelig, gjennom tidligere forskning, vet vi at det er rimelig å anta at det er diskursive forskjeller mellom dem (se for eksempel Breivega 2003).

Artikler er hentet fra anerkjente tidsskrifter med fagfellevurderinger (for hvilke tidskrifter vi har hentet artiklene fra, viser jeg til KIAPs nettside). Vi har bestrebet oss på å velge ut "virkelige" vitenskapelige artikler, i den forstand at de gjengir og diskuterer forskning utført av forfatterne selv, av empirisk eller teoretisk art. Vi har med andre ord prøvd å unngå oversiktsartikler og andre typer refererende artikler. Vi kan ikke si at vi er sikre på å ha greidd å holde fast på dette kriteriet fullt ut. Når det gjelder norske tidskrifter, for eksempel, og særlig innen medisin og økonomi, er det bare to tidsskrifter vi kan velge fra, henholdsvis *Tidsskrift for Den Norske Lægeforening* og *Norsk Økonomisk Tidsskrift*. Men samtidig kommer legeforeningens tidsskrift relativt hyppig ut og det blir mange artikler å velge mellom. I norsk lingvistikk har vi hentet størsteparten av artiklene fra *Norsk Lingvistisk Tidsskrift* men også noen fra *Maal og Minne.* Når det gjelder lingvistikk og økonomi, er det ikke alltid noe klart skille mellom det man kunne kalle undersjangrene "forskningsartikler" og "oversiktsartikler".

Generelt er det slik at de fleste medisinske artiklene vi har valgt, omhandler forskning med betydelig empirisk grunnlag, men noen av dem er kasus-studier (av en eller et lite antall pasienter). I økonomi har vi foretrukket artikler som inneholder relativt mye tekst i forhold til matematiske formler, ligninger, o.l., noe som er karakteristisk for mange tekster innen deler av samfunnsøkonomien. Innen språkvitenskap har vi lagt vekt på å velge ut artikler som omhandler et fenomen innen det aktuelle språket foran artikler av mer allmennlingvistisk art. De førstnevnte artiklene kan i større grad være "humanistiske" enn de sistnevnte.

Når det gjelder tilgjengelighet, er det store forskjeller mellom språkene. Det er selvfølgelig innen engelsk vi har hatt mest å velge mellom; her finnes det dessuten en rekke elektroniske versjoner som gjør innhentingen av tekster til et elektronisk korpus mye lettere. Det er også relativt mange franskspråklige tidsskrifter, men her er den elektroniske tilgjengeligheten dårligere. Når det gjelder norskspråklige, er situasjonen, som nevnt, vanskelig, fordi det er så få aktuelle tidsskrifter.

Vi har ønsket å inkludere både enforfatter- og flerforfatterartikler. For de første både kvinnelige og mannlige forfattere. Tilgjengeligheten på dette området er sterkt avhengig av disiplin; for eksempel er det svært sjelden medisinere skriver alene; mens

i økonomi, og særlig i lingvistikk, er enforfatterartikler snarere det vanlige. Dette gjenspeiler seg i korpuset.
Endelig har vi forsøkt å velge artikler hvor forfatterne (i alle fall én av dem når flere skriver sammen) har det språket som artikkelen er skrevet på, som sitt morsmål.

Spørsmålet om representativitet og størrelse er viktig i oppbyggingen av et korpus. Vi vet selvfølgelig at når man velger ut tekster slik vi har gjort, vil utvalget alltid være skjevt i forhold til en eller flere faktorer (se McEnery & Wilson [1996]/2003: 77-81). Våre kriterier er av mer kvalitativ enn kvantitativ art; for eksempel har det vært viktig å inkorporere hele tekster (og ikke bare utdrag). Samtidig mener vi å ha et tilstrekkelig stort korpus til å utføre de analysene vi ønsker; dette begrunner vi delvis med hva andre har brukt som korpus i tidligere studier av vitenskapelige sjangrer (se for eksempel Biber 1988, Hyland 2000).

Alt i alt inneholder korpuset vårt vel 3 millioner ord (en god del figurer, tabeller, o.l. er tatt vekk fra originalene, men all tekst er bevart). I prosjektet arbeider vi mest med "tekstkroppen" (body), det vil si den "rene" hovedteksten, uten sammendrag, titler/overskrifter, noter, referanser, eksempler, ligninger, og lignende. Denne delen av korpuset utgjør 2 250 868 ord:

KIAP-korpus – Antall ord

Korpus	Hel artikkel	Tekstkropp ("body")
engecon	407430	298319
engling	627014	437798
engmed	240579	163663
engall	**1275023**	**899780**
frecon	403702	295859
frling	341749	231296
frmed	200840	138510
frall	**946291**	**665665**
noecon	398682	312850
noling	371716	269913
nomed	160310	102660
noall	**930708**	**685423**
Totalt	**3152022**	**2250868**

Figur 1: *KIAP-korpus – Antall ord*
Forkortelsene står for følgende: eng = engelsk; fr = fransk; no = norsk; econ = samfunnsøkonomi; ling = lingvistikk; med = medisin; all = alle disiplinene samlet.

3.2 Tekstformat og koding

Tekstene som er valgt ut for å bli integrert i korpuset har vært i txt-format. Dette betyr at enkelte formateringstrekk som skrifttype, størrelse, kursiv og lignende er falt ut. Som tidligere nevnt, har vi tatt ut det meste av figurer, tabeller, og lignende. Disse utelatelsene har liten eller ingen betydning for våre forskningsspørsmål, selv om det selvfølgelig er synd at vi har mistet kursiver og lignende former for understreking i tekstene. Når det gjelder det som gjenstår av figurer og tabeller, er dette blitt kodet slik at det ikke kommer med i søkeresultatene.

Vår grad av koding (i hovedsak XML) står i forhold til våre forskningsspørsmål, som i hovedsak kan undersøkes ved søking på enkeltord eller kombinasjoner av slike. Grovt sagt har vi brukt tre former for koding: metakoding, strukturkoding og tekstkoding (de generelle prinisippene for kodingen stemmer godt overens med TEI-koding, men korresponderer likevel ikke med denne standarden, særlig fordi vår koding er mye mindre detaljert).

Metakodingen inneholder et sett av merkelapper med tilhørende informasjon som er plassert først i artikkelen: forfatternavn, bibliografisk referanse, artikkelklassifisering i forhold til de tre parametrene forfatterskap, språk og disiplin.

Strukturkodingen gjelder artikkelens tekstdeler eller seksjoner, hvor "tekstkroppen" (<body>) er den overordnede merkelappen. Utenfor tekstkroppen, koder vi for seg sammendrag (<abstract>), noter (<notes) og bibliografi (<references>). Vi har også en restkategori, <misc> (for *miscellaneous)*, hvor vi putter appendix, takksigelser, adresser, o.l.

Innenfor tekstkroppen deler vi inn i introduksjon (<intro>), midtparti (<mid>) og konklusjon (<concl>), (hvis en slik finnes). For artiklene som er bygd opp etter IMRAD-strukturen (Swales 1990), deler vi midtpartiet i en metode- og resultatdel (<mmr>) og en diskusjonsdel (<disc>). En skisse av kodingen for en typisk medisinsk IMRAD-artikkel er gjengitt i figur 2:

```
<title type="native"> Artikkeltittel </title>
<abstract type="native"> Abstract </abstract>
<body>
        <intro> Introduksjon </intro>
        <mmr> Materiale, metode, resultat </mmr>
        <disc> Diskusjon </disc>
        <concl> Konklusjon </concl>
</body>
<notes type="end"> Noter </notes>
<references> Bibliografiske referanser </references>
```

Figur 2: Skisse av koding for IMRAD-artikkel

Strukturkodingen er særlig viktig for analysene i KIAP: Vi er opptatt av å finne ut hvordan bestemte språkfenomener er distribuert gjennom teksten. Slike studier baserer seg på hypoteser vi har formulert om plasseringen av ulike fenomener i ulike tekstdeler.

Tekstkodingen gjelder noen relativt få merkelapper som vi bruker på konstituenter inne i teksten, men som vi ikke vil ha med når vi søker i tekstkroppen. Vi merker og tar dermed ut av tekstkroppen følgende: <subtitle>, som dekker alle former for avsnitts- og undertitler; <quote>, som dekker direkte gjengitt tale, med en minimumsgrense satt til tre ord; <example>, som dekker de eksempler som er trukket ut i egne avsnitt (gjerne nummerert) eller eksempler inne i teksten som består av mer enn ett ord, i lingvistiske artikler; endelig er tabeller, figurer, matematiske formler og lignende markert med <table>.

I tillegg til xml-kodingen er et par annotasjoner brukt: "*n*RRR" foran tall som angir bibliografiske referanser i de medisinske artiklene; og "NNN" foran notenummer. Dette ble gjort for at det skulle være mulig å søke på henholdsvis referanse- og noteangivelser i tekstene.

3.3 Søkeprogram

Det er medarbeidere ved Aksis (se 2.1) som har generert korpuset fra våre kodete tekster og som også har skreddersydd et søkeprogram for KIAP-prosjektet, som vi er svært fornøyde med.

Vi kan søke på 1, 2 eller 3 enkeltord samtidig. Det er også mulig å bruke trunkering (foranstilt og etterstilt) og ”wild cards” (uspesifiserte bokstaver; dvs at en søkestreng som "analy.e" anvendt på engelsk vil finne eksempler av både *analyse* og *analyze*). Når vi søker på kollokasjoner av opptil 3 ord, kan mellomrommene spesifiseres i antall ord. Konteksten foran og etter søkeordet kan spesifiseres i antall tegn. Det er også mulig å ordne søkene alfabetisk på ulike måter.

Vi kan søke i hele korpuset eller i deler av det (språk, disiplin); vi kan søke i enkeltartikler, i utvalgte deler av artikler, i en-forfatter eller flerforfatterartikler.

Endelig gir søkeprogrammet oss den relative frekvens for hvert søk (prosentdel i forhold til det totale antall ord i den tekstdelen det er søkt i). Aksis har også utarbeidet ordfrekvenslister for hvert enkelt ord i hver enkelt tekst; ordnet alfabetisk og etter frekvens.

Samlet sett er dette et godt redskap for våre forskningsspørsmål. Det er selvfølgelig mulig å gå mye lengre med kodingen, men siden vi har vært mer opptatt av utsigelse og semantisk-pragmatiske spørsmål enn morfosyntaktiske, har vi til nå prioritert som vist ovenfor.

4. Utvalgte resultater

4.1 Generelt

I KIAP har vi stort sett fulgt en overordnet metodologisk framgangsmåte som følger: Etter pilotprosjektet (se Breivega *et al.* 2002) startet vi med en eksplorativ undersøkelse av et utvalg på 180 artikler (i den tidlige fasen da bare 180 artikler var ferdigkodet) som ble gjennomgått både manuelt og maskinelt med det formål å undersøke videre de språklige fenomener vi hadde identifisert som relevante i pilotprosjektet. De første og foreløpige kvantitative resultater ble framstilt (se KIAP-

artikler av Dahl og Fløttum i Fløttum & Rastier 2003). På bakgrunn av disse kunne vi formulere de første hypoteser angående viktigheten av språk og disiplin for kulturell identitet. I det videre har det vært en alternering mellom kvantitative og kvalitative analyser; og etter at det komplette korpuset var ferdig kodet og generert (sommer 2004), har vi gjennomført større kvantitative og kontrastive analyser, som etter hvert blir publiseringsklare og delvis lagt ut på KIAPs nettside.

Vi er nå i stand til å presentere resultater som i noen grad stemmer med vår opprinnelige hypotese om at disiplin er viktigere enn nasjonalspråk. Dette er resultater oppnådd gjennom studier av bruken av fenomener listet opp nedenfor. Resultatene er relaterte til de tre forskningsspørsmålene nevnt i innledningen ovenfor (jeg nevner bare norske eksempler i listen, men til alle korresponderer det uttrykk i engelsk og fransk som er studert parallelt med de norske):

1) Førstepersonspronomen og indefinitte pronomen (henholdsvis *jeg, eg, vi, me* og *man, en, ein*).

2) Verb som kombineres med disse pronomenene (og disses semantisk-pragmatiske art).

3) Konstruksjonen *la oss/la meg* + infinitiv.

3) Metatekstuelle uttrykk (som *i denne artikkelen, jeg skal her analysere*).

4) Bibliografiske referanser, deres form, plassering og frekvens.

5) Polyfoniske konstruksjoner (som polemisk nekting *ikke*, konsessive konnektiver som *men, imidlertid*).

6) Markører av epistemisk modalitet (som *kan, kunne; det synes…*).

7) Utvalgte leksemer som *resultat* (og ord tilhørende samme familie).

For publikasjoner hvor disse (og andre) resultater er gjengitt, viser jeg til den samlede listen *KIAP- publikasjoner* til slutt i artikkelen.

I en generell og grov oppsummering av våre resultater, kan vi altså fortsatt si at *disiplin vinner over språk***,** i den forstand at det er tydeligere forskjeller mellom disiplinene (på tvers av språkene) enn mellom språkene. Dette kan igjen spesifiseres slik:

a) Lingvister er de som manifesterer seg mest og sterkest polemisk (mine uthevinger):

(4) **As I have argued** in this article, **many of these experimental studies conflate** aspects of literal and nonliteral meanings **and often confuse** what occurs during processing of lexical meaning with what occurs when entire utterances are interpreted. **For these reasons, I claim, once again**, that little empirical evidence exists to support the idea that […]. (engling22, concl)

(5) **Berrendonner (1997**), [...], **envisageait même** que [...]. **Je ferais plutôt l'hypothèse** qu'il s'agit, dans les trois cas cités, d'une seule et même structure syntaxique clivée, [...]. (frling01, concl)

(6) **Hvor kommer en slik middel-tolkning av verbet *gråte* fra?** Denne betydningskomponenten kan **ikke** spores tilbake til noe trekk i dette verbets "normale" leksikonrepresentasjon, **og jeg vil mene** at det **heller ikk**e er ønskelig å prøve å legge den inn der. **Jeg vil tvert imot hevde** at denne betydningsvrien må henføres til selve konstruksjonen som verbet her opptrer i, […]. (noling49, mid)

b) Økonomene er også klart til stede, men helst i ”ufarlige” eller ikke-polemiske sammenhenger, som for eksempel når de skal gjøre rede for strukturen i artikkelen eller ta med leseren direkte på et resonnement:

(7) In this article, **I will propose** changes to the conventional empirical approach, which might allow us to arrive at more precise estimates of the growth effect of education. (engecon15, intro)

(8) Dans les sections suivantes, **nous présentons et analysons les résultats**. **Nous essaierons de comprendre** en quoi ils reflètent la spécificité du dispositif français [...]. (frecon21, intro)

(9) **Som vi ser** av tabell 6, fikk en person som var arbeidsløs hele 1992 en god del høyere anslått inntektsreduksjon. (noecon38, mid)

c) Medisinerne manifesterer seg relativet sjelden direkte (for eksempel ved personlige konstruksjoner), og om de gjør det, er det gjerne i en ”nøytral” redegjørelse om materiale, metode eller resultater:

(10) To assess the potential effect of this assumption, **we repeated our analysis** after excluding any respondent who had missing information on any adverse childhood experience **and found** no substantial difference in the results. (engmed24, mmr)

(11) **Nous avons également constaté** que la durée maximale de l'hypothyroïdie transitoire est de 6 mois. (frmed12, disc)

(12) **Vi registrerte** 85 pasienter med subaraknoidalblødning i perioden. (nomed33, mmr)

Men gjennom analyser av polyfoniske konstruksjoner kan vi påpeke en ikke ubetydelig tilstedeværelse og implisitt argumentasjon i deres artikler også, som i dette eksemplet med et klart kontrastivt og argumentativt tilsnitt:

(13) Kvinners og menns ulike måter å finansiere misbruket på vil også kunne medføre ulikheter i sykdomsrisiko. Sosialhjelp/trygd er den inntektskilden som er hyppigst rapportert, **men dette utgjør likevel bare** en mindre andel av den totale inntekten for misbrukerne. (nomed08, disc)

Selv om vi til en viss grad kan opprettеholde vår opprinnelige disiplinorienterte hypotese, har våre studier gjort det nødvendig å modifisere den. Det er nemlig *betydelige forskjeller mellom artiklene skrevet på engelsk, fransk og norsk* (tydeligere forskjeller enn først antatt). De engelske vitenskapelige forfatterne er klart mest direkte til stede; men de norske henger ikke så langt etter. Derimot er de franske de minst tilstedeværende. De franske er likeledes generelt sett mer forsiktige og mindre eksplisitte når de skal annonsere sine resultater enn de norske og engelske. Dette kommer til uttrykk når vi for eksempel ser på hvordan en fransk, en norsk og en engelsk lingvist annonserer sine resultater i artiklenes introduksjon (alle er enforfatterartikler):

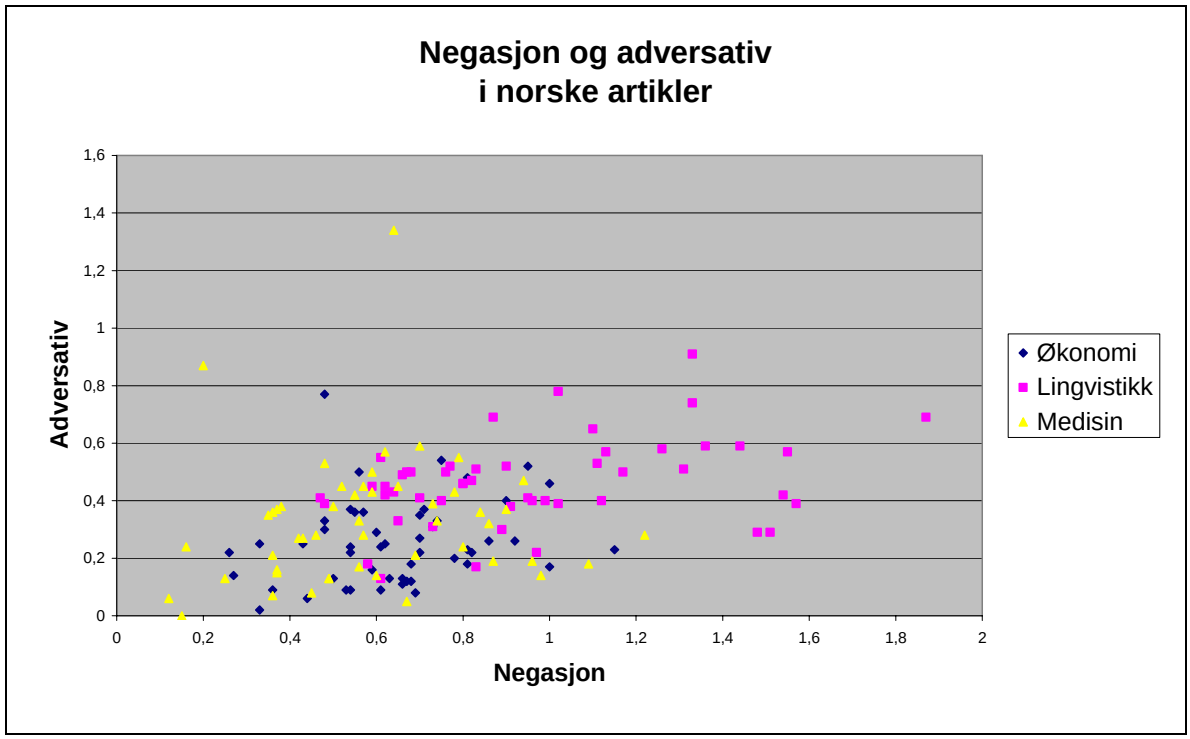
(14) Les prépositions spécifiques des valences verbales sont généralement considérées comme [...].Cette particularité est mentionnée depuis longtemps par les grammairiens français [...]. Comme d'autres langues semblent présenter des faits analogues […], **on peut être tenté d'y voir** un phénomène plus général, qui ne serait pas spécifiquement français. **Je propose de rappeler les principales circonstances** dans lesquelles s'observe ce phénomène de la "préposition à éclipses", et **d'en envisager quelques interprétations**. (frling01, intro)

(15) I denne artikkelen skal jeg belyse flere sider ved pronomenet *denne* […]. I denne artikkelen **skal vi se på** hvilke bruksbetingelser som finnes. **Vi skal se at** begrensningene på pronomenets referanse ikke kan beskrives som en slags språklig deiksis […]. **Tvert imot viser det seg at** hovedbegrensningen ser ut til å være at det aldri kan være koreferent med et subjekt. (noling01, intro)

(16) Recent work on the syntax of tense shows that […]. **I contribute to this discussion**
here by arguing that syntactic locality constrains the interpretation of temporal relations ; temporal dependency between times requires the times to be in a local syntactic configuration at LF. […]. **I argue that** gerundive relatives are temporally dependent on the main clause tense and thus are required to be local with a matrix time at LF. (engling01, intro)

Her samvirker valg av pronomen, verb og (i den franske) modale uttrykk og utgjør viktige indikatorer på forskjelligheten i forfatters tilstedeværelse.

Endelig har vi i det siste året gjort nye observasjoner som blir viktige for våre sluttkonklusjoner. Det er nemlig slik at, selv om det er mulig å peke på viktige disiplinære og nasjonalspråklige forskjeller, viser KIAPs resultater også at *de individuelle forskjellene innenfor hver av de ni undergruppene er betydelige*. Retorikk og argumentasjon er til stede i alle artikler, men forskjellene innenfor både disiplin og språk er store. Dette kan vi vise blant annet ved kvantitative studier, illustrert i grafiske framstillinger som i følgende spredningsdiagram (utarbeidet av Torodd Kinn) over bruken av nektelsesordet *ikke* og den kontrastive (eller adversative) og konsessive polyfonimarkøren *men* i de norske artiklene:



Figur 3: Negasjon og adversativ i norske artikler

Samlet sett mener vi KIAP-resultatene bidrar med viktige karakteristikker av sjangeren vitenskapelig artikkel: Den er retorisk, men i ulik grad innenfor ulike språk og ulike disipliner. Slik gir vi støtte, på samme måte som observasjoner fra andre forskningsmiljøer, for at den tradisjonelle oppfatningen av vitenskapelig diskurs som såkalt nøytral og objektiv, med fravær av eksplisitte spor etter forfatter, kun er en myte. Vi mener imidlertid at vi også bidrar med noe vesentlig nytt gjennom resultater fra studiene hvor språk og disiplin er sett i sammenheng.

4.2 Nye hypoteser utviklet gjennom korpusstudier

En innvending til korpusstudier som gjennomføres etter forhåndsformulerte hypoteser, kan være at man bare finner det man leter etter. Slik har det ikke vært i KIAP-prosjektet. Vi har oppdaget flere interessante fenomener gjennom våre søk og konkordanser og våre semantisk-pragmatiske analyser av disse. I tillegg kommer den praktiske (om enn møysommelige) fordelen at vi alle selv har deltatt i kodingsarbeidet og slik fått god innsikt i tekstenes beskaffenhet – utover de fenomenene vi var interesserte i fra starten av. (Flere av studentene som har vært knyttet til KIAP, har

måttet studere tekster manuelt før de har fått adgang til det elektroniske korpuset.) Nedenfor gir jeg noen eksempler på nye hypoteser som er kommet til, etter hvert som korpusanalysene er blitt gjennomført.

a) *Fra den generelle hypotesen om forfatternærvær ved førstepersonspronomen til hypotesen om forfatternærvær ved ubestemt pronomen i franske artikler.*
(Se for eksempel Fløttum 2003c, i, 2004d.)

Vi er fortsatt overbevist om at førstepersonspronomen er en god og opplagt eksplisitt indikator på forfatternærvær i artiklene. Men forskjellene mellom de engelske og norske artiklene på den ene siden og de franske på den andre, er så store at vi har sett det som nødvendig å undersøke i hvilken grad andre mulige personlige konstruksjoner er brukt i de franske. Det har vist seg at det ubestemte pronomenet *on* (som delvis svarer til de norske *man, en, ein*) er svært frekvent i disse. (Om indefinitte pronomener i engelsk, norsk og tysk, se Johansson 2002). Analysene våre peker på en svært variert bruk av dette "upresise" og samtidig fleksible pronomenet. Det kan i prinsippet ha verdier som svarer til alle de personlige pronomenene; men det mest interessante i vår sammenheng er at det blir brukt med referanse til forfatteren, og slik svarer til *jeg/eg* i enforfatterartikler og *vi* i flerforfatterartikler:

(17) Dans un premier temps (paragraphe 1), **on** présentera une liste de caractéristiques, dans l'ensemble bien connues, qui distinguent les deux emplois [...]. (frling06, intro)

b) *Fra den generelle hypotesen om forfatternærvær ved førstepersonspronomen til hypotesen*
om ulike forfatterroller manifestert ved ulike typer verb kombinert med pronomenet.
(Se for eksempel Fløttum 2003g, 2004d, Kinn 2004.)

En viktig hensikt med våre kvantitative analyser har vært å skaffe belegg for vår hypotese om forfatternærvær ved førstepersonspronomen. Når man søker for eksempel på et enkelt ord som det førstepersonspronomenet er, får man konkordanser som på en

svært iøynefallende måte viser pronomenets umiddelbare kotekst. Det er opplagt at verbet som pronomenet kombineres med, kan si mye mer om forfatternærværets natur enn pronomenet alene. En av våre hypoteser som ble utviklet med basis i pronomensøkene, postulerer at det er snakk om minst tre forfatterroller i den vitenskapelige artikkelen: forfatter som skriver eller redigerer (*jeg starter med*), som forsker (*jeg analyserer*) og som aktør med posisjoneringsbehov (*jeg hevder*).

De helt siste undersøkelsene som Torodd Kinn har gjort viser at det er påfallende store forskjeller mellom delkorpusene når det gjelder typer verb som blir kombinert med førstepersonspronomen (Kinn har særlig sett på flertallspronomenene); og fransk skiller seg særlig fra de to andre språkene.

c) *Fra hypotesen om den fraværende mottaker til hypotesen om invitasjoner til mottakeren om å delta i strukturering og resonnement*
(Se for eksempel Dahl 2003, 2004 c, Kinn under trykking b og c.)

Elektroniske søk etter opplagte mottakerspor, som andrepersonspronomen, har gitt svært beskjedne resultater, om noen i det hele tatt. Den vitenskapelige artikkel er da heller ikke tradisjonelt betraktet som særlig dialogisk. Men studiene av ulike metatekstuelle og metadiskursive uttrykk ga oss nye spor i jakten på den først antatte fraværende mottaker. I uttrykk som *I denne artikkelen skal vi først se på* er det klart at det ikke bare er en redegjørelse fra forfatters side om hva artikkelen skal handle om og i hvilken rekkefølge, men også en invitasjon til mottakeren om å delta (gjennom et inkluderende *vi* og et verb som refererer til en handling man godt kan være to om). Dette blir enda tydeligere i imperativ-konstruksjoner med verbet *la*, som i *La oss undersøke* … Disse studiene har gitt interessante resultater med hensyn til hvordan interaksjon mellom mottaker og forfatter kan foregå.

d) *Fra hypotesen om eksplisitt tilstedeværelse av andre stemmers enn forfatterens til hypotesen om andre stemmers implisitte tilstedeværelse gjennom polyfoniske uttrykk*
(Se for eksempel Fløttum 2003d, e, 2004 f, under trykking a, b, c, f.)

Den opplagte og eksplisitte tilstedeværelsen av andres stemmer i den vitenskapelige artikkelen manifesterer seg ved direkte eller indirekte gjengivelse av resultater og observasjoner hentet fra andre forskeres arbeider, gjerne ved en eller annen form for gjengitt tale. Dette har vi studert systematisk, og kommet fram til resultater som peker på forskjeller mellom språk og disipliner. Men vi har også gjennom først å studere utvalgte artikler manuelt funnet ulike polyfoniske konstruksjoner interessante i denne sammenheng. Andre forskeres stemmer og synspunkter trekkes inn i teksten på en implisitt måte ved blant annet polemisk nekting *ikke* og kontrastive og konsessive *men* og *imidlertid.* Dette er en form for subtil interaksjon som vi ved kvantitative analyser har kunne hevde utgjør et felles og frekvent trekk i alle de ni underkorpusene.

I en mer dyptgående analyse av den syntaktiske nektingen *ikke*s kotekst, har vi også funnet grunn til å sette spørsmål ved det tradisjonelle og absolutte skillet mellom deskriptiv og polemisk nekting. Det viser seg at det dreier seg om en skala av større eller mindre grad av polemisitet, som gir nektingen ulike verdier, delvis bestemt av den sjangeren den forekommer i.

Til slutt nevner jeg en hypotese om polyfoniske (polemiske) konstruksjoner som mulige signaler for kotekster med verdiladete uttrykk. I følge denne skulle det være slik at det for eksempel i en polemisk nektings umiddelbare nærhet typisk skulle finnes verdiladete, subjektive leksikalske uttrykk. Denne hypotesen (som er spennende) trenger å bli undersøkt ytterligere før den eventuelt kan framsettes som holdbar (se Fløttum under trykking b).

5. Avsluttende bemerkninger

Erfaringene med korpusstudiene i KIAP-prosjektet har vært svært positive. Vi er likevel ikke mer naive enn at vi minner oss selv om de kjente advarslene mot bruk av korpus, som for eksempel Stig Johansson nevner i sin artikkel fra 1995. Jeg går ikke inn i disse her, men vi er selvfølgelig klare over det åpenbare at et korpus aldri vil være noe annet enn én måte å få tilgang til språkbruk på. Men til tross for en varsomhetsplakat som finnes i våre hoder, velger vi å vektlegge det positive når våre erfaringer skal oppsummeres:

- Korpuset utvikler stadig nye problemstillinger

Som tidligere nevnt, ved nøye studier av de funn som søk i korpuset har gitt oss, er nye hypoteser og problemstillinger blitt utviklet.

- Korpuset vekker lingvistisk nysgjerrighet

Dette er en kvalitet som ikke bare er viktig i forskningssammenheng, men også i undervisningssammenheng. Det er gledelig å se hvordan masterstudenter har fattet interesse for nye lingvistiske og diskursive problemstillinger gjennom bruk av korpuset. Muligheten til å kunne gå inn i et språkmateriale på så mange måter som vårt søkeprogram gir anledning til, har skapt stor entusiasme.

Korpuset inneholder selvfølgelig mye som fortsatt er uutforsket. Det gjemmer problemstillinger som vi ikke har valgt eller ikke har hatt tid til å gå inn i til nå, men som vi kan gå videre med også etter at KIAP-prosjektet er formelt avsluttet. Eksempler på dette kan være syntaktiske studier (noe som vil kreve mer koding, men som er gjørbart), mer typiske tekstlingvistiske studier knyttet til bruk av konnektiver eller bindeord, studier av informasjonsstruktur, studier av anaforiske uttrykk (som i alle fall lar seg søke på automatisk til en viss grad gjennom pronomener, bestemte og possessive artikler), flere studier knyttet til argumentasjon i utvalgte artikkeldeler (for eksempel introduksjon og konklusjon), forholdet mellom vitenskapelig språkbruk og "allmennspråkbruk" (ved sammenligning av kvantitative data fra KIAP-korpuset med andre og større korpus).

Referanser

Adam, J-M. 1999. Linguistique textuelle. Des genres de discours aux textes. Paris: Nathan.

Bakhtin, M. (1986): *Speech Genres & other late Essays.* Austin: University of Texas Press.

Bazerman, C. 1988. *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison, WI: The University of Wisconsin Press.

Benveniste, E. 1966. *Problèmes de linguistique générale*. Paris: Gallimard.

Berge, K.L. 2001. Det vitenskapelige studiet av sakprosa. Om tekstvitenskapelige utfordringer og løsninger i norsk og svensk sakprosaforskning. K.L. Berge *et al.* (red): *Fire blikk på sakprosaen. Teori og praktisk analyse.* Skrifter fra Prosjektmiljøet Norsk sakprosa 1, 9-71.

Berge, K.L. 2003. The scientific text genres as social actions: text theoretical reflections on the relations between context and text in scientific writing. K. Fløttum and F. Rastier (eds.): *Academic discourse. Multidisciplinary approaches*. Oslo: Novus. 141-157.
Berge, K.L., and P. Ledin. 2001. Perspektiv på genre. *Rhetorica Scandinavica* 18, 4-16.
Biber, D. 1988. *Variations across speech and writing*. Cambridge: CUP.
Breivega, K. R. 2003. *Vitskaplege argumentasjonsstrategiar. Ein komparativ analyse av superstrukturelle konfigurasjonar i medisinske, historiske og språkvitskaplege artiklar*. Sakprosa 8. Oslo: Universitetet i Oslo.
Breivega, K.R., T. Dahl and K. Fløttum. 2002. Traces of self and others in research articles. A comparative pilot study of English, French and Norwegian research articles in medicine, economics and linguistics. *International Journal of Applied Linguistics* 12 (2), 218-239.
Brekke, M., Andersen, Ø., Dahl, T. and J. Myking (Eds.). 1994. *Applications and implications of current LSP research. Proceedings of the 9th European Symposium on LSP, Bergen, August 1993*. Bergen: Fagbokforlaget.
Ducrot, O. 1984. *Le dire et le dit*. Paris: Minuit.
Fabricius-Hansen, C. 1991. Contrastive Stylistics. Outline of a Research Project on German and Norwegian Non-fictional Prose. K. M. Lauridsen & O. Lauridsen (eds.), *Contrastive Linguistics*. Aarhus: The Aarhus School of Business, 51-75.
Hyland, K. 2000. *Disciplinary discourses. Social interactions in academic writing*. Harlow, England/New York: Longman.
Johansson, S. 1995. *Mens sana in corpore sano*: On the Role of Corpora in Linguistic Research. *The European English Messenger*, IV, 2, 19-25.
Johannsson, S. 2002. Viewing language through multilingual corpora, with special reference to the generic person in English, German and Norwegian. L. I. Rábade & S. M. Doval Suarez (eds.), *Studies in Contrastive Linguistics*, Universidade de Santiago de Compostela, 515-554.
Marnette, S. 2001. The French *théorie de l'énonciation* and the study of speech and thought presentation. *Language and Literature* 10 (3), 243-262.
McEnery, T. & Wilson, A. [1996]/2003. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.
Laurén, C. & J. Myking (red.) 1999, *Nordica Bergensia*, no. 20.
Nølke, H., K. Fløttum & C. Norén. 2004. *ScaPoLine. La théorie de la polyphonie linguistique*. Paris: Kimé.
Prelli, L.J. 1989. *A Rhetoric of Science: Inventing Scientific Discourse*. Columbia: University of South Carolina Press.
Rastier, F. 2001. *Arts et sciences du texte*. Paris: PUF.
Salager-Meyer, F. 1998. Reference Patterns in Medical English Discourse. In L. Lundquist, H. Picht and J. Qvistgaard (eds.): *LSP. Identity, and Interface. Research, Knowledge and Society. Proceedings of the 11th European Symposium on Language for Special Purposes*. København: Copenhagen Business School, Vol I, 495–504.
Salager-Meyer, F. and N. Zambrano. 2001. The discourse of competing knowledge claims in academic prose. In: F. Meyer (ed.), *Language for special purposes: Perspectives for the new millennium*, vol. 2. Tübingen: Gunter Narr, 474–479.
Svensson, J. (red.) (1999). *Svensk sakprosa. Summaries from a research project.* Lund: Universitetet i Lund.
Swales, J.M. 1990. *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.
Tønnesson, J. L. 2003: *Tekst som partitur eller Historievitenskap som kommunikasjon.* Oslo: Universitetet i Oslo (Dr.art.-avhandling).

Vassileva, I. 2000. *Who is the Author? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian academic discourse*. Sankt Augustin: Asgard Verlag.

KIAP-publikasjoner (Oktober 2004)

Andersen, P. L. W. 2003. *Le titre de l'article de recherche.* Hovedfagsoppgave. Romansk institutt,
Universitetet i Bergen.
Breivega, K. R, Dahl, T. & Fløttum, K. 2002. Traces of self and others in research articles. A comparative pilot study of English, French and Norwegian research articles in medicine, economics and linguistics. *International Journal of Applied Linguistics* 12 (2), 218-239.
Dahl, T. 2002. Author abstracts and computer-generated summaries: similarities and differences. M. Koskela *et al.* (Eds.), *Porta Scientiae II. Lingua specialis.* Proceedings of the University of Vaasa. Reports 96. Vaasa, 546-556.
Dahl, T. 2003. Metadiscourse in research articles. K. Fløttum & F. Rastier (Eds.), *Academic discourse.*
Multidisciplinary approaches. Oslo: Novus, 120-138.
Dahl, T. 2004a. Absent doctors, shy economists and polemic linguists? Writer manifestation in academic texts. *Synaps* 14, 1-4. Instituttserie, Institutt for fagspråk og interkulturell kommuinkasjon, NHH.
Dahl, T. 2004b. Some characteristics of argumentative abstracts. *Akademisk Prosa* 2, 49-67. *Skrifter for KIAP.*
Romansk institutt, Universitetet i Bergen.
Dahl, T. 2004c. Textual metadiscourse in research articles: a marker of national culture or of academic
discipline? *Journal of Pragmatics* 36 (10), 1807-1825.
Dahl, T. (under trykking a). Accommodating the reader: metatext in research articles. A report from the KIAP
project. *Selected Conference Proceedings, 14th European Symposium on Language for Special Purposes.*
Dahl, T. (under trykking b). *Cultural identity in academic texts. Akademisk Prosa 3. Skrifter for KIAP.* Romansk
institutt, Universitetet i Bergen.
Didriksen, A. A. 2004a. Val av artiklar ved hjelp av kvantitative metodar. *Akademisk Prosa* 2, 99-115. *Skrifter*
for KIAP. Romansk institutt, Universitetet i Bergen.
Didriksen, A. A. 2004b. Konnektorar i franske vitskaplege artiklar. *TRIBUNE* 15, 31-38.
Fløttum, K. 2001. Kulturelle likheter og forskjeller i akademisk prosa. *TRIBUNE* 12, 27–34.
Fløttum, K. 2002. Corpus description and methodological design.
http://www.helmer.aksis.uib.no/kiap/om_prosjektet.htm
Fløttum, K. 2003a. Y a-t-il des identités culturelles dans le discourse scientifique ? *Estudios de lengua y literatura francesas,* 14, 83-93.
Fløttum, K. 2003b. Om forfatternærvær i vitenskapelige artikler. Rapport fra *Nasjonalt fagråd i romanske språk*
og litteraturer 2002. Tromsø: Universitetet i Tromsø, 30-47.
Fløttum, K. 2003c. Personal English, indefinite French and plural Norwegian scientific authors?

Pronominal author manifestation in research articles. *Norsk Lingvistisk Tidsskrift* 21(1), 21-55.
Fløttum, K. 2003d. Polyphonie dans les textes scientifiques. Etude de deux cas français. M. Olsen (Éd.),
Polyphonie – linguistique et littéraire, VII. Roskilde: Samfundslitteratur Roskilde, 111-129.
Fløttum, K. 2003e. Bibliographical references and polyphony in research articles. K. Fløttum & F. Rastier (Eds.),
Academic discourse. Multidisciplinary approaches. Oslo: Novus, 97-119.
Fløttum, K. 2003f. Forskerne sier ... Kronikk i *Bergens Tidende* 03.04.03
http://www.bt.no/meninger/kronikk/article149215
Fløttum, K. 2003g. « JE » et le verbe. *TRIBUNE* 14, 7-14.
Fløttum, K. 2003h. Forskning og formidling. Kronikk i *Bergens Tidende* 17.09.03 (Forskningsdagene 2003).
http://www.bt.no/meninger/article193379.
Fløttum, K. 2003i. The French pronoun 'on' in academic discourse – indefinite versus personal.
In: Hajičová, E., Kotěšovcová, A., Mírovský, J. (ed.), Proceedings of CIL17, CD-ROM. Matfyzpress, MFF UK. Prague, 2003. ISBN: 80-86732-21-5. (Paper marked: S10_KjerstiFlottum).
Fløttum, K. 2004a. Enonsiativ teori – et perspektiv for studiet av vitenskapelige tekster. *Akademisk Prosa* 2, 7-
28. *Skrifter for KIAP.* Romansk institutt, Universitetet i Bergen.
Fløttum, K. 2004b. Thèmes, topiques et marqueurs de cadres discursifs dans les articles scientifiques
français : la présence de personne(s). *Scolia* (Strasbourg) 18, 167-202.
Fløttum, K. 2004 c. Om fagtradisjoner, posisjonering og identitet i vitenskapelig skriving. D. F. Simonsen (red.), *Språk i kunnskapssamfunnet. Engelsk – elitenes nye latin eller uunnværlig redskap for fri kunnskapsutvikling?* Oslo: Gyldendal, 185-191.
Fløttum, K. 2004 d. La présence de l'auteur dans les articles scientifiques : étude des pronoms *je, nous*
et *on*. In: A. Auchlin *et al.* (eds), *Structures et discours.* 2004. Québec: Ed. Nota Bene, 401-416.
Fløttum, K. 2004 e. The expression of "results" in scientific discourse – a cross-linguistic and cross-disciplinary
analysis. Paper given at Societas Linguistica Europaea 37th International Meeting, Kristiansand 29.07-01.08.2005.
Fløttum, K. 2004 f. Polyfonisk interaksjon via IKKE i vitenskapelig diskurs. *Rhetorica Scandinavica 31,* 23-40.
Fløttum, K. 2004g. Den franske vitenskapelige forfatter – *je, nous* eller *on* ? *TRIBUNE* 15, 56-69.
Fløttum, K. (under trykking a). The self and the others – polyphonic visibility in research articles.
International Journal of Applied Linguistics 2005, 15 (1), 29-44.
Fløttum, K. (under trykking b). Polyphonic constructions as epistemic and evaluative qualifications in
research articles. Gabriella Del Lungo & Elena Tognini Bonelli (Eds.), *Evaluation in academic*
discourse. Benjamins.
Fløttum, K. (under trykking c). Traces of others in research articles: the citation cluster. *Selected Conference*

Proceedings, 14th European Symposium on Language for Special Purposes.
Fløttum, K. (under trykking d). Om formulering av forskningsresultater i vitenskapelige artikler. Trykkes i
antologien *Kommunikasjon*, Høgskolen i Østfold.
Fløttum, K. (under trykking e). MOI et AUTRUI dans le discours scientifique : l'exemple de la négation
NE...PAS. G. Bres *et al.* (eds.), *Dialogisme et polyphonie.*
Fløttum, K. (under trykking f). *Individual linguistic variation and possibilities for academic self-hood.*
Akademisk Prosa 3. Skrifter for KIAP. Romansk institutt, Universitetet i Bergen.
Fløttum, K. & Breivega, K. 2002. Cultural identity in academic prose: national versus discipline-specific. M. Koskela et al. (Eds), *Porta Scientiae II. Lingua specialis.* Proceedings of the University of Vaasa. Reports 96. Vaasa, 533-545.
Fløttum, K. & Rastier, F. (Eds.) 2003. *Academic discourse. Multdisciplinary approaches.* Oslo: Novus.
Gjesdal, A. M. 2003. *L'emploi du pronom "on" dans les articles de recherche. Une étude diachronique et*
qualitative. Hovedfagsoppgave. Romansk institutt, Universitetet i Bergen.
Gjesdal, A. M. 2004a. Om bruken av *on* i vitenskapelige artikler. *Akademisk Prosa* 2, 87-97. *Skrifter for KIAP.*
Romansk institutt, Universitetet i Bergen.
Gjesdal, A. M. 2004b. Bruken av pronomenet "on" i franske lingvistiske artikler. En diakronisk
studie 1980-2000. *TRIBUNE* 15, 121-128.
Grinde, L. 2003. *Les choix lexicaux dans des articles de recherche français.* Hovedfagsoppgave. Romansk
institutt, Universitetet i Bergen.
Kinn, T. 2004. Cognitive research agents in academic prose. *Akademisk Prosa* 2, 137-149. *Skrifter for KIAP.*
Romansk institutt, Universitetet i Bergen.
Kinn, T. (under trykking a)."Denne artikkelen analyserer ...". Den tenkjande forskaren i norske forskingsartiklar.
To appear in the proceedings of MONS 10, 2004.
Kinn, T. (under trykking b). Tilbod og innbydingar: imperativ med *la* i forskingsartiklar.
Kinn, T. (under trykking c). *Plays of we-hood: authorial presence and reader engagement in research articles.*
Akademisk Prosa 3. Skrifter for KIAP. Romansk institutt, Universitetet i Bergen.
Skiple, J. 2003. *L'emploi de la négation dans les articles scientifiques français.* Hovedfagsoppgave. Romansk
institutt, Universitetet i Bergen.
Vold, E.T. 2004a. Vers une conception philosophico-logique de la modalité en linguistique. *Akademisk Prosa* 2,
117-135. *Skrifter for KIAP.* Romansk institutt, Universitetet i Bergen.
Vold, E. T. 2004 b. Comment sélectionner ses observables dans l'analyse d'une catégorie sémantique ?
Paper given at *Young Researchers' Conference*, 29 – 30 April, 2004. Paris-X, Nanterre.
Vold, E. T. 2004c. Epistemisk *pouvoir*/*kunne* i akademiske tekster. *TRIBUNE* 15, 229-240.

KIAPs skrifter:

Akademisk prosa, no 1, juni 2003. Romansk institutt, Universitetet i Bergen.

Akademisk prosa, no 2, april 2004. Romansk institutt, Universitetet i Bergen.

Akademisk prosa, no 3, (under trykking), februar 2005. Romansk institutt, Universitetet i Bergen.

Parallellkorpora som innfallsport til språksammenligning på tekstnivå – med særlig henblikk på (norsk/tysk/engelsk) sakprosa

Cathrine Fabricius-Hansen, Universitetet i Oslo

1. Innledning

De følgende betraktninger er først og fremst av metodologisk art. Jeg skal først (avsnitt 2) skissere en typologi for parallellkorpora som er presentert i håndboka *Übersetzung – Translation – Traduction* (Fabricius-Hansen 2004). Deretter vil jeg gi noen konkrete eksempler på hvordan parallellkorpora kan brukes som oppdagelsesreise og middel til hypotesedannelse – og dermed også som forskningsstyrende instrument – når det gjelder språkspesifikke prinsipper for hvordan informasjon struktureres eller "porsjoneres" i skrevet tekst – dvs. hvordan parallellkorpora kan brukes som innfallsport til kontrastiv stilistikk i vid forstand (avsnitt 3). Objektspråkene er tysk og norsk, til dels også engelsk. Flere av eksemplene har jeg diskutert også i andre publikasjoner (se Fabricius-Hansen 1998, 1999, 2003, 2005), stort sett innenfor rammen av prosjektet SPRIK *Språk i kontrast* (se http://www.hf.uio.no/forskningsprosjekter/sprik).

2. Forskjellige typer parallellkorpora (i vid forstand) til forskjellige formål

Uttrykket parallellkorpus er ikke helt entydig. Brukt i snever forstand refererer det til elektroniske tekstsamlinger som består av (utdrag av) originaltekster på et gitt språk og tilsvarende oversatte tekster. Slike parallellkorpora kalles også **oversettelseskorpora** ('translation corpora') I spesialtilfeller kan det være tale om tekster med identisk innhold som er forfattet parallelt på to eller flere forskjellige språk, uten at den ene teksten kan sies å være en oversettelse av den annen – f.eks. EU-dokumenter, avtaletekster i flerspråklig sammenheng etc. I videre forstand kan uttrykket parallelkorpus brukes om en samling originaltekster (eller tekstutdrag) på et gitt språk og en tilsvarende samling originaltekster (resp. tekstutdrag) på et annet språk eller flere andre språk, dvs. to eller flere sett originaltekster som ideelt sett er bygd opp parallelt i henhold til bestemte kriterier – omfang, genre, stilnivå, tema etc.

Sammenlignbare korpora (’comparable corpora’) er en annen, mer entydig betegnelse for parallellkorpora av denne typen.

I tillegg til den grunnleggende sondringen mellom oversettelseskorpora og sammenlignbare korpora er det relevant å til å skille mellom korpora som består av større eller mindre tekstutdrag og korpora som består av hele tekster. Sistnevnte vil gjerne inneholder færre enkelttekster (’samples’) enn førstnevnte.

Kombinerer vi de to inndelingskriterier, så får vi følgende typologi:

	A. Tekstutdrag	B. Hele tekster (gjerne forholdsvis lite antall)
I. Oversettelseskorpora (’translation corpora)	IA	IB
II. Sammenlignbare korpora av originaltekster (‘comparable corpora’)	IIA	IIB

Tab. 1 Parallellkorpora i vid forstand – typologi

Parallellkorpora anvendes i forbindelse med språk(bruks)sammenligning. Men bruksområdet varierer etter hvordan korpus er bygd opp.

Parallellkorpora av type A egner seg bl.a. til (kvantitative) undersøkelser på forskjellige nivåer under øverste tekstnivå, avhengig hvordan tekstene er annotert og hvor stort antall tekstutdrag (’samples’) det er tale om; se f.eks. Biber (1995). Korpora av type B kan danne grunnlag for studiet av av diskursstrukturering og tekstkonvensjoner på øverste nivå (cf. Rastier/Fløttum 2003, Clyne 1991, Oldenburg 1992, Stahlheber 1992 o.a.). Vanligvis vil antallet tekster imidlertid være for lite til å tillate vidtrekkende statistisk velbegrunnede generaliseringer slik at det snarere blir tale om eksemplariske analyser.

Oversettelseskorpora (type I) egner seg selvfølgelig først og fremst til oversettelsesstudier, men kan også gi opphav til spesifikke hypoteser om forskjeller mellom de aktuelle språkene på forskjellige nivåer (leksikon, syntaks, stil, tekstkonvensjoner - dvs. både språksystem og språkbruk); se nedenfor. Når det gjelder språkbruk, må slike hypoteser imidlertid etterprøves på grunnlag av sammenlignbare korpora av autentiske tekster, dvs. parallellkorpora av type II. Et bra alternativ når det gjelder å avdekke ’normalisation’ og ’shining through’ i oversettelse er å kombinere

oversettelsekorpora med bruk av store referansekorpora ('reference corpora') for de respektive språk; jf. Teich 2003.

Det tospråklige parallellkorpus English-Norwegian Parallel Corpus[1] (ENPC) er et eksempel på en kombinasjon av IA og IIA. Det inneholder utdrag (ca. 10000 ord) av engelske og norske originaltekster og deres oversettelser, med samme fordeling mellom skjønnlitterære tekster og sakprosatekster i de to delkorpora. Modellen for ENPC er søkt videreført i Oslo Multilingual Corpus[2] (OMC), som i tillegg bl.a. omfatter originale og oversatte tekster på tysk. Men på grunn av praktiske forhindringer (copyright, oversettelsespraksis) er det vanskelig å gjennomføre sammenlignbarhetsprinsippet helt når man har å gjøre med mer enn to språk.

De fleste sammenlignbare korpora (type I) består av et relativ lite antall autentiske tekster, dvs. de er av type IB. Noen eksempler av denne art er omtalt i Baumann/Kalverkämper (1992); jf. også Stahlheber (1992), Biber et al. (1998).

Et interessant eksempel på et trespråklig korpus av type IIB finner man i prosjektet KIAP *Kulturell Identitet i Akademisk Prosa: nasjonal versus disiplinavhengig*[3], som undersøker tekstkonvensjoner i lingvistiske og økonomiske tidsskriftartikler på hhv. norsk, engelsk og fransk.

3. Konkretisering 1: Informasjonsstruktur og –'porsjonering' på tekstnivå

3.1 Sententialisering i oversettelse (tysk > norsk)

I avhandlingen *Sententialität, Nominalität und Übersetzung* undersøker Solfjeld (2000) et oversettelseskorpus omfattende utdrag på ca. 5000 ord av 30 tyske sakprosatekster og deres norske oversettelser. Han påviser at de norske måltekstene gjennomgående inneholder et signifikant større antall finitte verb enn utgangstekstene; jf. tabell 2. For så vidt som flere finitte verb betyr flere finitte verbalfraser eller enkle setninger, kan en si at de norske måltekstene er preget av **sententialisering** i vid forstand i forhold til de tyske originalene. Denne formen for sententialisering er illustrert i eksempel (1) og (2).

[1] Se http://www.hf.uio.no/iba/prosjekt/.
[2] Se http://www.hf.uio.no/iba/OMC/index.html.
[3] Se http://kiap.aksis.uib.no/.

	Antall tekstpar	% av tekstpar
Mer enn 25% flere finitte verb i no. måltekst enn i tysk ugangstekst	1	3.3
20-25% flere finitte verb i no. måltekst	8	26.7
15-20% flere finitte verb i no. måltekst	7	23.3
10-15% flere finitte verb i no. måltekst	9	30.0
5-10% flere finitte verb i no. måltekst	4	13.3
Negativ differens	1	3.3
I alt	30	100

Tabell 2. Sententialisering i sakprosatekster (jf. Solfjeld 2000: 89)

(1) a Sie *fühlen* sich wohl in ihrem Element. (Fra Solfjeld 2000)
b … de *er* i sitt eget element **og** *føler* seg vel derved.

(2) a [Nach dem kleinsten Imbiß fühlte er sich voll und schwer,] *aber morgens beim Ankleiden* merkte *er*,[daß er offenbar dünner wurde]. (OMC)
b [Han følte seg tung og overmett etter den minste matbit,] *men* ***når*** *han* kledte *seg om morgenen,* la *han merke til* [at han var blitt tynnere].

I norske oversettelser av tyske (sakprosa)tekster kan man også observere en tendens til **setningsdeling –** eller sententialisering i snevrere forstand – i forhold til originalen, dvs. at en uavhengig setning (’helsetning’) forholdsvis ofte oversettes med en sekvens av uavhengige setninger (’helsetninger’), skilt ved punktum (eller annet stort skilletegn). Eksempler på sententialisering i denne snevrere forstand ses i (3) og (4).

(3) a [Das Verhalten des anderen wird imitiert,] weil dieser als Vorbild angesehen wird, oder weil das gezeigte Verhalten belohnt wird, ***oder*** *weil gar keine andere Verhaltensmöglichkeit angeboten ist*. (Fra Solfjeld 2000)
b [En annens atferd blir imitert] fordi vedkommende blir betraktet som et forbilde, eller fordi denne atferden blir belønnet. *Det kan også være at det ikke finnes andre atferdsmuligheter*.

(4) a Der Ethnologe wird sich bei meinen Erzählungen an das sogenannte "magische Denken" vieler Naturvölker gemahnt fühlen, *das auch beim Zivilisationsmenschen noch durchaus lebendig ist und das die meisten von uns zu allerlei entwürdigenden kleinen Zaubereien zwingt*, … (OMC)
b Det jeg her har fortalt, vil få etnologer til å tenke på den såkalte "magiske tenkning" hos mange naturfolk. *Den slags overtro er fremdeles spill levende*

også hos siviliserte mennesker, og tvinger de fleste av oss til alle slags nedverdigende småtrolldomskunster

Konrad Lorenz' bok *Das sogenannte Böse* (fra 1963) er et illustrativt eksempel: Her inneholder de første fem kapitler av den norske oversettelsen (fra 1968) ca. 25% flere uavhengige setninger enn originalens første fem kapitler; jf. tabell 3.[4] Men en liknende tendens er også dokumentert i andre tilfeller (Ramm 2004). Setningsdeling synes ikke å forekomme i tilsvarende omfang i tyske oversettelser fra norsk (jf. Ramm 2004). Det er hva en skulle vente dersom en antar at (i) sjansene for setningsdeling øker med omfanget og kompleksiteten til setningene i originalteksten (jf. Fabricius-Hansen 1998) og (ii) setningene i autentiske norske (sak)prosatekster har en tendens til å være kortere og mindre komplekse enn i autentiske tyske (sak)prosatekster (jf. Fabricius-Hansen/Solfjeld 1994). Den første hypotesen må etterprøves på omfangsrike tysk/norske og norsk/tyske oversettelseskorpora, den andre på sammenlignbare korpora av tyske og norske (sak)prosatekster.[5]

	KoLo	**KoLo-TN**	**KoLo-TE**
A. s-units			
Kap. 1	89	131	81
Kap. 2	114	151	101
Kap. 3	284	361	277
Kap. 4	74	85	74
Kap. 5	290	345	291
I alt	**851** (100%)	**1073**	**824**
Differenz		+222 (+26.1%)	-27 (-3.2%)
B. Kolon--setninger	**50**	**35**	**50**
C. Semikolon-setninger	**22**	**1**	**71**

fortsetter neste side…

[4] Såkalte "s-units" er skilt med punktum.
[5] Fabricius-Hansen/Solfjeld (1994) er en forholdsvis enkel pilotundersøkelse av denne art.

D. Tankestreg- S.	**5**	**39**	**3**
Tankestreger	57	239	3
E. 'Setninger', sum A+B+C+D	**928** (100%)	**1148**	**948**
Differanse		+220 (+23.7%)	+20 (+2.2%)
F. Ord	**26042** (=100%)	**27965** (+7.4%)	**25898** (-7.9%)
Ord/s-unit	30.6	26.1	31.4
Ord/'setning'	28.1	24.4	27.3

Tabell 3. Setningsdeling og setningslengde i de første fem kapitler av Konrad Lorenz' *Das sogenannte Böse* (KOLO), den norske oversettelse (KOLO-TN) og den engelske oversettelse (KOLO-TE) (cf. Fabricius-Hansen 1998)

Sententialisering i vid forstand og setningsdeling (sententialisering i snever forstand) er prinsipielt uavhengige av hverandre, dvs. de kan opptre hver for seg eller i kombinasjon med hverandre.

(i) Sententialisering (i vid forstand) uten setningsdeling finner en i (1), der en setning med et enkelt finitt verb er oversatt med koordinerte finitte verbalfraser (koordinativ sententialisering), og i (2) der en enkelt setning er oversatt med en kompleks setning (subordinativ sententialisering).

(ii) Sætningsdeling uden sententialisering (i vid forstand) finner en når koordinerte finitte verbalfraser eller (del)setninger oversettes med en sekvens av selvstendige setninger skilt med stort skilletegn, som i (3) og (4) ovenfor.

(iii) Sententialisering (i vid forstand) kombinert med setningsdeling innebærer at en enkelt setning inneholdende et enkelt finitt verb gjengis som to eller flere selvstendige setninger; jf. (5).

(5) a Bei Prof. Dements "Traumentzugs-Experimenten" *reagierten* die Versuchspersonen auf den Traumentzug mehrerer Nächte mit Reizbarkeit, Unentschlossenheit und Feindseligkeit. (Fra Solfjeld 2004)
b Professor Dement *hindret* sine forsøkspersoner i å drømme. Etter noen netter *reagerte* forsøkspersonene med irritabilitet, ubesluttsomhet og aggresjoner.

Koordinativ og subordinativ sententialisering (vid forstand) kan også opptrede sammen som i (6). – (7) og (8) er eksempler på lidt lengre utdrag av parallelltekster med tysk original der ovennevnte teknikker er kombinert på forskjellig vis i de norske oversettelser. (7) – som er fra tredje kapittel av *Das sogenannte Böse* – illustrerer også

forskjellene mellom den norske og den engelske oversettelsen når det gjelder setningsdeling (sml. tab. 3).

(6) a Das nächste historisch gewordene Experiment des Traumforschers *bewies* endgültig, dass das Träumen für uns so selbstverständlich *ist* wie ... (cit. Solfjeld 2004)
b Drømmeforskeren *gjorde* så et nytt eksperiment **som** *er* blitt epokegjørende **og** *gav* det endelige bevis for at det *er* like naturlig for oss å drømme som ...

(7)

KoLoTN3.s69-74	<KoLo.3.s59-61>	KoLoTE s52-53
Våre tamme kuer og svin har fremdeles bevart så mye av det sosiale angrepsinstinktet overfor ulven at det kan oppstå alvorlige situasjoner for et menneske som sammen med en engstelig unghund begir seg ut på et jorde der det beiter et større antall storfe.	Unseren Hausrindern und -schweinen liegt der soziale Angriff gegen den Wolf noch so sehr im Blut, daß man durch sie in ernste Gefahr geraten kann, wenn man eine von einer größeren Herde bevölkerte Weide in Begleitung eines ängstlichen jungen Hundes betritt,	The reaction of social attack against the wolf is still so ingrained in domestic cattle and pigs that one can sometimes land oneself in danger by going through a field of cows with a nervous dog
En slik hund vil nemlig gjerne søke beskyttelse mellom sin herres ben, istedenfor å møte angriperne med gjøing eller selv ta flukten.	der, anstatt die Angreifer zu verbellen oder selbständig zu fliehen, zwischen den Beinen des Herrn Schutz sucht.	which, instead of barking at them or at least fleeing independently, seeks refuge between the legs of its owner.
En gang måtte jeg selv hoppe i vannet sammen med en tispe ved navn Stasi, og komme meg i sikkerhet ved å legge på svøm, etter at en flokk kviger hadde dannet en halvsirkel rundt oss og rykket truende nærmere.	Ich selbst mußte einmal samt meiner Hündin Stasi in einen See springen und schwimmend mein Heil suchen, als eine Herde von Jungrindern einen Halbkreis um uns gebildet hatte und drohend vorrückte.	Once, when I was out with my bitch Stasi, I was obliged to jump into a lake and swim for safety when a herd of young cattle half encircled us and advanced threateningly;
Under første verdenskrig tilbrakte min bror en behagelig ettermiddag i et topphugget piletre i det sørlige Ungarn	Mein Bruder hat im ersten Weltkrieg in Südungarn einen angenehmen Nachmittag auf einer Kopfweide verbracht,	and when he was in Southern Hungary during the first world war my brother spent a pleasant afternoon up a tree with his Scotch terrier under his arm,

Med en skotsk terrier under armen hadde han vært nødt til å klatre opp i treet etter å være blitt omringet av en hjord halvville ungarske svin, som beitet fritt i skogen	auf die er mit seinem Scotchterrier unter dem Arm geklettert war, weil eine Herde der frei im Walde weidenden, halbwilden ungarischen Schweine	because a herd of half-wild Hungarian swine, disturbed while grazing in the wood, encircled them,
De trengte seg stadig tettere sammen omkring ham og hunden, mens de blottet huggtennene og tydelig viste hva de hadde i sinne	die beiden eingekreist hatte und den Kreis, in unverkennbarer Absicht die Hauer entblößend, immer enger zog.	and with bared tusks and unmistakable intentions began to close in on them.

(8) (Sitert etter Solfjeld 2004)

Auf der nächtlichen Schaubühne der Menschen von heute kreuzt natürlich ständig das Auto auf.	Bilen dukker naturligvis stadig opp på et moderne menneskes nattscene.
Grund dieser „Verkehrsdichte" ist jedoch weder die Übermotorisierung der Städte, noch die Freud an dem fahrbaren Untersatz.	Men årsaken til "trafikktettheten" er verken at bilene har overtatt byene, eller at drømmeren er bilgal.
Durch die außergewöhnlich starke symbolische Audruckskraft des Autos *benützt* der Traum das Auto, um eine bestimmte Lebenssituation aufzuzeigen.	Bilen *har* en uvanlig sterk symbolsk uttrykkskraft. Derfor *benytter* drømmen seg av bilen for å skissere en bestemt livssituasjon.
Auf eine interessante Deutungsmöglichkeit der Autoträume weist der Schweizer Traumanalytiker Ernst Aeppli hin.	Den sveitsiske drømmeanalytikeren Ernst Aeppli har påpekt en interessant tydningsmulighet for bildrømmene.
Er meint, dass der Begriff „Auto" nicht zufällig mit dem Begriff „Ich" und „Selbst" (auto = griechisch "selbst") übereinstimmt.	Han mener at det ikke er tilfeldig at begrepet "automobil" faller sammen med begrepet "jeg" og "selv" ("auto" er gresk for "selv").
Das Traum-Auto kann die eigene Person symbolisieren!	Drømmebilen kan symbolisere drømmeren selv!
Als die einundvierzigjährige Hausfrau Gerda W. eine Midlife–crisis *durchmachte*, ein neues Leben anfangen und als Entwicklungshelferin nach Afrika gehen wollte, *hatte* sie einen typischen Autotraum.	Gerda W., husmor og 41 år, *gjennomgikk* en "middelalderkrise". Hun *ville* begynne et nytt liv og reise til Afrika for å drive u-hjelp. Da *hadde* hun en typisk bildrøm.

3.2 Relevante spørsmål

Iagttakelser som ovennevnte gir anledning til bl.a. følgende **oversettelsesrelaterte** spørgsmål:

(a) Hva er det i utgangsteksten som utløser sententialisering og/eller setningsdeling, dvs. hvilke typer **kildestrukturer** (Solfjeld 2000) har en å gjøre med?
(b) Hvilke oversettelsesstrategier sættes typisk inn overfor hvilke kildestrukturer?
(c) Hvilken effekt har setningsdeling i den konteksten den foretas i, dvs. på tekstnivå?

Disse og beslektede problemer blir tatt opp av Solfjeld (2000, 2003, 2004), spesielt (a) og (b); Ramm (2004), spesielt (a) og (c); Fabricius-Hansen (1998, 1999), særlig spesielt (c). Det har i den sammenhengen vist seg viktig å skille mellom informasjonsenheter knyttet til setnings- og verbalfrasenivå (jf. frie adverbialer) på den ene siden og informasjonsenheter knyttet til nominalfrasenivå (attributive adjektiver, relativsetninger, genitivattributter, etterstilte attributive preposisjonalfraser) på den annen siden. På tysk kan en ha omfangsrige, komplekse pre- og postnominale attributter på nominalfrasenivå samtidig som samtidig som verbalfrasen/setningen utbygges med diverse fri adverbialer; jf. eksemplene ovenfor. En slik opphoping av informasjon innen for den enkelte setningen utnytter spesifikke strukturelle muligheter i tysk og er i samsvar med stilistiske konvensjoner som gjelder for moderne tysk skriftspråk (sakprosa). Spørsmålet er da: Hvordan takles slike strukturer i oversettelse til et språk (som norsk) som ikke tillater noe tilsvarende, enten det nå skyldes strukturelle forhold eller stilistiske normer som er sterkere orientert mot muntlig språk? Solfjeld sier til dette:

> Bei gleichzeitiger Expansion auf mehreren Niveaus würde die Übersetzung schnell nicht mehr überschau- oder handhabbar bleiben. Anzunehmen ist eine gewisse Tendenz zum Ausgleich zwischen expandierenden, Lexikalisierung auslösenden Strategien auf einer übergeordneten Ebene und eher 'reduzierenden' Strategien – wie etwa Tilgung und analoge Zielstrukturen – auf einer untergeordneten Ebene. (Solfjeld 2003: 23)

De spørsmål som er reist ovenfor flere trekker andre spørsmål etter seg som angår **autentisk språkbruk** i de respektive språkene og som derfor krever en sammenligning av originaltekster (sammenlignbare korpora):

(d) Er de endringer (bl.a. sententialisering) som skjer i oversettelsene strukturelt betinget eller skyldes de snarere forskjeller i stilkonvensjoner? Hvordan er

forholdet mellom systematiske muligheter (språklige ressurser) og stil-/tekstkonvensjoner (språkbruk)?

(e) Er oversettelsene i det hele tatt 'naturlige' norske (resp. engelske) tekster eller er de mer eller mindre sterkt preget av 'shining through' (Teich 2003)?

(f) Hvordan 'porsjoneres' informasjon i norske (resp. engelske) originaltekster? Dersom de enkelte setninger inneholder mindre informasjon enn i tysk, eller informasjonen er anrettet mindre hierarkisk, hvilke konsekvenser har det da for diskursstrukturen på overordnet nivå? Hvilke muligheter finnes det for å kompensere sterkt hierarkiske syntaktiske strukturer slik de kjennes fra tysk sakprosa?

Når det gjelder norsk, virker bl.a. følgende **hypoteser** plausible og interessante som utgangspunkt for en undersøkelse av sammenlignbare norske og tyske sakprosatekster (jf. Fabricius-Hansen 1996, 1999):

(g) Norske sakprosatekster er mindre eksplisitte og refererer ikke så hyppig direkte til hva som er sagt tidligere i teksten.

(h) Norske tekster tenderer mer mot en 'høyrefokal' struktur på tekstnivå, dvs. a bakgrunn og premisser for ny fokussert informasjon plasseres i selvstendige setninger som går forut for den setningen som inneholder den sentrale informasjonen.

(i) Norske tekster er sterkere preget av diskursstrukturell underbestemthet på det overordnede tekstnivå.

4. Konkretisering 2: Funksjonsord

Oversettelsekorpora supplert med sammenlignbare korpora er et glimrende utgangspunkt for å

- avdekke forskjellige betydnings- eller bruksvarianter av frekvente ord, herunder ikke minst funksjonsord som konnektiver (Altenberg 1999, Fretheim & Johansson 2002, Meier 2001)
- belyse konstruksjoner som er holder på å bli grammatikalisert (*være ved å, holde på å* etc.) (jf. Tonne 2001).
- Har en å gjøre med korpora som er grammatisk annotert, vil det også være mulig å avdekke bruksforskjeller for 'samme' grammatiske kategori i forskjellige språk, som f.eks. preteritum eller bestemhet i tysk og norsk.

Jeg skal her – avsluttende – nøye meg med å nevne det tyske pronominaladverb *dabei* som eksempel på et konnektiv som er interessant i kontrastiv sammenheng. Sett fra norsk er ordet en 'falsk venn': Ordrett oversatt svarer det til norsk *derved*, men det har ikke den samme instrumentelle betydningen – i så måte er det *derved* og *dadurch* som motsvarer hverandre. I realiteten har *dabei* ikke noen leksikalsk ekvivalent verken i norsk eller engelsk. I oversettelsessammenheng (OMC, begge veier) finner en derfor ganske stor spredning, som det fremgår av tabell 4. (9) og (10) er eksempler på typisk setningsinitialt bruk av *dabei*.

Category	Examples	
Simple clause-initial connective / conjunction (≠ *and/og*)	*(and) yet, but, however* *thus* *moreover*	*men* *allikevel* *selv om, skjønt*
Spatial connective adverb (clause initial or non-initial)	*here*	*her*
Temporal connective adverb (clause initial or non-initial)		*nå, da , samtidig*
Instrumental connective adverb[6]		*dermed, derved* *derigjennom*
PP containing an abstract anaphor or definite description	*in (all) this, at this activity,* *in this situation,* *in the process*	*i denne forbindelse* *under denne processen* *i den anledning*
Prepositional (*in/by*) *ing*-construction with anaphoric predicate	*in doing so, in so doing* *by doing so*	[No structural equivalent to ing-constructions in Norwegian]
Prepositional (in/by) ing-construction with 'full' predicate	*in hurrying to the end of the bridge*	
Temporal clause *(as/when/while)* with anaphoric predicate	*as I did so, when he did it* *while he is at it*	*mens så skjer*
Temporal clause *(as, while)* with 'full' predicate	*and as I breathed* *while the boy was talking*	*mens unggutten la ut*

Tabell 4. En. og no. 'oversettelsesbilleder' av tysk *dabei* (OMC) (fra Fabricius-Hansen 2005)

(9) a. They were subjected to tests of courage and had to demonstrate their fighting skills. ***In all this*** *a strict code of honour was enforced*.
b. Dem werdenden Macho wurden Mutproben und Schaukämpfe abverlangt. ***Dabei*** *mußte ein strikter Ehrenkodex eingehalten werden.* (HME1)
c. Av den kommende macho ble det krevd prøver på mot og oppvisningskamper. ***Her*** *måtte en strikt æreskodeks overholdes.*

(10)a. Mein Problem ist es oft, nicht fragen zu können.
Dabei *bestehe ich fast nur aus Fragen.* (PH1)
b. Inability to ask questions is often my problem.
And yet *I 'm made up almost entirely of questions.*
c. Det er et problem for meg, dette at jeg ofte ikke er istand til å stille spørsmål.
Og så *jeg, som i grunnen ikke har stort annet enn spørsmål i meg!*

Påfallende ofte har *dabei* (i original eller oversatt tekst) ikke noen eksplisitt motsvarighet i den norske eller engelske paralleltekst (mål- eller utgangstekst) i det hele tatt; jf. tabell 5. Eksempler på slik zero-korrespondens ses i (11) og (12).

	Explicit	Zero	Total
1. → En	52 (47.67%)	57 (52.3%)	109 (100%)
2. ←En	31 (28.2%)	79 (71.8%)	110 (100%)

	Explicit	Zero	Total
1. → No	75 (62.5%)	45 (37.5%)	120 (100%)
2. ←No	19 (45.2%)	23 (54.8%)	42 (100%)

Tabell 5. Zero-frekvens for *dabei* mht En/No (OMC) (fra Fabricius-Hansen 2005)

(11)a. […] ein schwarzäugiger, braunhäutiger Halbwüchsiger kam in Begleitung eines ihm ähnlichen Kindes zur Tür herein und tauschte an der Theke eine große leere Weinflasche gegen eine volle um; ***dabei*** *stellte er das Kind als seinen Onkel vor*. (PH1)
b. A black-eyed, brown-skinned adolescent came in with a child who looked like him, and went to the bar, where he exchanged a large empty wine bottle for a full one. *He introduced the child as his uncle* […].
c. Det bor en del mennesker fra sydligere egner her også: en sortøyet, mørkhudet fremslenging kom inn i følge med en guttunge som var ganske lik ham; borte ved disken fikk de en stor flaske vin i bytte for tomflasken; *den største gutten fortalte at den minste var hans onkel*.

[6] Som nevnt har *dabei* ikke primært instrumentell eller kausal betydning. Oversettelse til/fra *derved* etc. er ikke helt adekvat eller eller sterkt avhengig av konteksten.

(12)a. Sind Sie schon einmal im Wald […] ausgerutscht und *haben **dabei** durch die Laubschicht am Boden in einen vermoderten Baumstrunk gegriffen*? (PH1)
b. Did you ever lose your footing in the woods […] and *reach through the underbrush to grab a rotting tree trunk*?
c. Har De noen gang vært på skogstur […] og glidd og *tatt Dem for og fått tak i en råtten trelegg under løvet på bakken?*

Som Altenberg (1999) gjør oppmerksom på, kan høy zero-andel i oversettelsessammenheng til dels henge sammen med manglende leksikalsk korrespondanse. Men det er ikke bare *dabei* som påfaldende ofte ikke oversettes eller dukker opp i oversatt tekst uten noe eksplisitt forlegg i utgangsteksten – det gjelder også f.eks. et adverb som *wieder*, som har leksikalske motsvarigheter i form av *igjen* og *again* (Fabricius-Hansen 2005, Podut 2005). Det er bl.a. slike oversettelsesmønstre som danner bakgrunnen for – og underbygger – hypotesen om at tendensen til å markere koherens eksplisitt er sterkere i tysk skriftspråk enn i norsk (jf. 3.2).

5. Avsluttende bemerkninger

Som nevnt i innledningen, skulle dette bidraget først og fremst tjene som metodologisk appetittvekker når det gjelder bruken av forskjellige typer parallellkorpora. Hovedformålet har vært å vise at oversettelseskorpora ikke bare er nyttige som ledd i oversettelsrelaterte undersøkelser. De mønstre som avdekkes gjennom oversettelseskorpora kan danne utgangspunkt for presise kontrastive problemstillinger når det gjelder språkbruk og stil på tekstnivå som i sin tur kan danne utgangspunkt for svært interessante analyser av sammenlignbare parallelkorpora.

Litteratur

Altenberg, Bengt (1999). Adverbial connectors in English and Swedish. I *Out of Corpora. Studies in Honour of Stig Johansson,* Hilde Hasselgård & Signe Oksefjell (eds.), 249-268. Amsterdam: Rodopi.

Biber, Douglas (1995). *Dimensions of register variation: a cross-linguistic comparison.* Cambridge: CUP.

Biber, Douglas, Susan Conrad & Ranid Reppen (1998). *Corpus linguistics : investigating language structure and use*. Cambridge: CUP.

Clyne, Michael (1991). Zu kulturellen Unterschieden in der Produktion und Wahnehmung englischer und deutscher wissenschaftlicher Texte. *Info DaF* 18, 4: 376-383.

Engen, Janne Hagen (2002). "og" isn't always "and": From coordination to subordination in English translations of Norwegian texts. SPRIKreports No. 17.

Fabricius-Hansen, Cathrine (1996). Informational Density - A Problem for Translation and Translation Theory. *Linguistics* 34, 521-565.

Fabricius-Hansen, Cathrine (1998). Information density and translation, with special reference to German - Norwegian - English. I *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies*, Stig Johansson & Signe Oksefjell (eds.). Amsterdam: Rodopi. 197-234.

Fabricius-Hansen, Cathrine (1999). Information packaging and translation: Aspects of translational sentence splitting (German – English/Norwegian. I *Sprachspezifische Aspekte der Informationsverteilung,* Monika Doherty (ed.). Berlin. 1999. 175-214.

Fabricius-Hansen, Cathrine (2004). Paralleltext und Übersetzung in sprachwissenschaftlicher Sicht. I HSK *Übersetzung -Translation – Traduction*, Harald Kittel et al. (eds.). Berlin: de Gruyter. S. 322-329.

Fabricius-Hansen, Cathrine (2005). Elusive connectives. A case study on the explicitness dimension of discourse coherence. *Linguistics* 43-1: 17-48.

Fretheim, Thorstein & Stig Johansson (2002). The semantics and pragmatics of the Norwegian concessive marker *likevel*: Evidence from the English-Norwegian Parallel Corpus. I *From the COLT's mouth and .. others'. Language corpora studies in honour of Anna-Brita Stenström,* Breivik, L.E. og Hasselgren, A. (eds.). Amsterdam: Rodopi. 81-101.

Johansson, Stig et al. (eds.) (2002). *Information Structure in a Cross-Linguistic Perspective.* Amsterdam: Rodopi.

Kvam, Sigmund (2002). Kontraktive Konstruktionen al Textgestaltungsmittel. Eine Fallstudie am Beispiel eines deutsch-norwegischen fachsprachlichen Paralleltextes. HERMES Skriftserie. Handelshøjskolen i Århus.

Meier, Einar (2001). *"Since you mention it". A contrastive study of causal subordination in English and Norwegian.* SPRIKreports 22.

Oldenburg, Hermann (1992): Zusammenfassungen und *Conclusions* im Vergleich: Empirische Ergebnisse und praktische Perspektiven. I *Kontrastive Fachsprachenforschung*, K.-D. Baumann & H. Kalverkämper (eds.). Tübingen: Narr, s. 123-134.

Ramm, Wiebke (2005). Sentence boundary adjustments in Norwegian-German and German-Norwegian translations: First results of a corpus-based study. I *Translation and Corpora. Gothenburg Studies in English* 89, Karin Aijmer Karin & Hilde Hasselgård (eds). Göteborg Universitet. 71-86..

Rastier, François & Fløttum, Kjersti (2003). *Academic discourse : multidisciplinary approaches*. Oslo: Novus.

Solfjeld, Kåre (2000)*: Sententialität, Nominalität und Übersetzung. Eine empirische Untersuchung deutscher Sachprosatexte und ihrer norwegischen Übersetzungen.* Frankfurt a.M.: Lang.

Solfjeld, Kåre (2003): *Die Wiedergabe deutscher erweiterter Attribute in auhtentischen norwegischen Übersetzungen.* SPRIKreports No. 19. (http://www.hf.uio.no/german/sprik/reporter.sthml)

Solfjeld, Kåre (2004): *Informationsspaltung nach links in Sachprosaübersetzungen Deutsch-Norwegisch,* SPRIKreports No. 21. (http://www.hf.uio.no/german/sprik/reporter.sthml)

Stahlheber, Eva M. (1992). Die Fachtextsorte Zeitschriftenartikel im Deutschen und Address/Article im Amerikanischen: Popularisierungsgrad und Diachornie von Funktionen und Strukturen. *Kontrastive Fachsprachenforschung*, IK.-D. Baumann &H. Kalverkämper (eds.). Tübingen: Narr, s. 162-189.

Teich, Elke (2003). *Cross-linguistic variation in system and text : a methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.

Tonne, Ingebjørg (2001). *Progressives in Norwegian and the theory of aspectuality*. **(**Acta humaniora 104.) Oslo: Unipub.

Økonomisk-administrativ kunnskapsbank: eit korpusbasert terminologiprosjekt

Magnar Brekke og Kai Innselset, Senter for fagspråkforsking, Institutt for fagspråk og interkulturell kommunikasjon, Norges Handelshøyskole

1. Teoretisk perspektiv

KB-N (KunnskapsBank for Norsk økonomisk-administrativt domene)[1] er eit treårig prosjekt (2004-2006) innanfor NFRs KUNSTI-program[2] i samfinansiering med NHH og i samarbeid med Aksis (Avdeling for kultur, språk og informasjonsteknologi) ved UiB.[3]

KB-N-konseptet kombinerer eit sett av noko ulike modular og subsystem, med røter innanfor ulike vitskapshistoriske tradisjonar og med ulike bruk innanfor moderne språkteknologi og kunnskapsrepresentasjon. KB-N utfordrar den dikotomien som lenge har dominert forskingstradisjonane i kryssingsfeltet mellom språk og teknologi, som i stikkords form kan karakteriserast som motsetninga mellom teoretisk uinformert samleaktivitet og empirisk uinformert parametrisering. I eit overordna komplementært perspektiv søkjer KB-N å gi sitt tilskot til den tilnærminga som i seinare år har skjedd mellom empiriske og teoretiske innfallsvinklar nettopp gjennom kombinasjonen og integrasjonen av aktuelle metodar og resultat.

Datamaskinen er blitt for tekstbasert empiri (herundar korpuslingvistikken) det mikroskopet lenge har vore for t.d. biologisk empiri, men tilgangen til nærmast uendelege datamengdar har tydeleggjort behovet for vidare teori- og metodeutvikling. "Knowledge management", som hadde ei viss blomstringstid på 90-tallet, er no igjen blitt eit sentralt tema for mange verksemder, men overgangen frå "data" via "informasjon" til "kunnskap" har vist seg å vera ganske enkelt uhandterleg utan automatiserte prosedyrar.

Det teoretiske fundamentet for fagspråklingvistikken ligg i den påstanden at eit identifiserbart utsnitt av "røyndomen" utgjer eit definerbart fagkunnskapsdomene, og vidare at den essensielle teoretiske kunnskapen innanfor fagdomenet lar seg representera i språkleg/tekstleg form. Den språklege/tekstlege

[1] http://www.nhh.no/fsk/sff/kbn/
[2] http://program.forskingsradet.no/kunsti/om/
[3] http://www.aksis.uib.no/

kunnskapsrepresentasjonen utkrystalliserar seg i særleg grad i det som av profesjonsutøvarar blir rekna som fagterminologi, og her spesielt i form av komplekse nominalfraser.

Ein av våre utgangshypotesar var at det langs desse linjene skulle vera mogeleg å etterprøva kor pålitelege eksisterande fagdomenekategoriar med tilhøyrande kriterium er, og freista å etablera ei matrise som i større grad fangar opp prototypiske språklege korrelat for det som i hovudsak ofte er pragmatiske eller administrative faginndelingar. Her trudde vi det skulle vera råd å avsvekkja de Beaugrandes påstand om at "there are indefinitely many 'degrees of freedom' between a 'reality' and its discourse representations", men dei førebels røynslene i så måte kan snarare tena til å stadfesta påstanden.

Økonomisk-administrative fagdomene er som kjent ikkje i særleg grad ”reine” fag, men består ofte av hybridar avleidde frå fag som matematikk, statistikk, organisasjonsteori, åtferdspsykologi osb. Mange av dei aktuelle faga rører seg nettopp i dette skjeringsfeltet mellom "opphavlege" omgrep og termar som er blitt overtekne av andre fag og gitt eit til dels avvikande omgrepsinnhald. Ofte endrar domenekunnskapen seg i prosessen, særleg dersom den i tillegg blir omsett til eit anna språk der det "same" fagdomenet har utvikla ganske andre språklege konvensjonar.

Kvar einaste språklege representasjon av eit kunnskapsunivers krev presis handtering også av innhaldssida. Innanfor den klassiske wüsterske "Allgemeine Terminologielehre"[4] har **omgrepet** ei sentral stilling, men det har synt seg vanskelegere å gi det eit presist innhald innanfor mindre eksakte fagdomene som t.d. innan samfunnsvitskapane og humaniora. Her har det føregått ei interessant teoriutvikling både innanfor nordiske miljø og den romanskspråklege verda. Dynamisk omgrepssystematisering er alt ein integrert funksjon i den språkteknologiske plattforma vår. Dette vil utgjera ein heilt grunnleggjande onomasiologisk dimensjon innanfor kunnskapsbasen og bidra til å sikra både validiteten, reliabiliteten og den fleirspråklege integriteten av dei data som skal registrerast, lagrast og gjenfinnast.

2. Systemarkitektur

Kunnskapsbank for norsk økonomisk-administrativt domene (KB-N) har som ambisjon å etablera

[4]Wüster, Eugen (1974). Die allgemeine Terminologielehre - ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. Linguistics 199, January 1, 61-106.

a) *ei språkteknologisk plattform* for å utvikla og utnytta teoretiske og metodiske innsikter i tekstbasert kunnskapsrepresentasjon, med særleg tanke på å kunna utnyttast i maskinomsetjing (MT) for profesjonell kommunikasjon mellom norsk og engelsk.

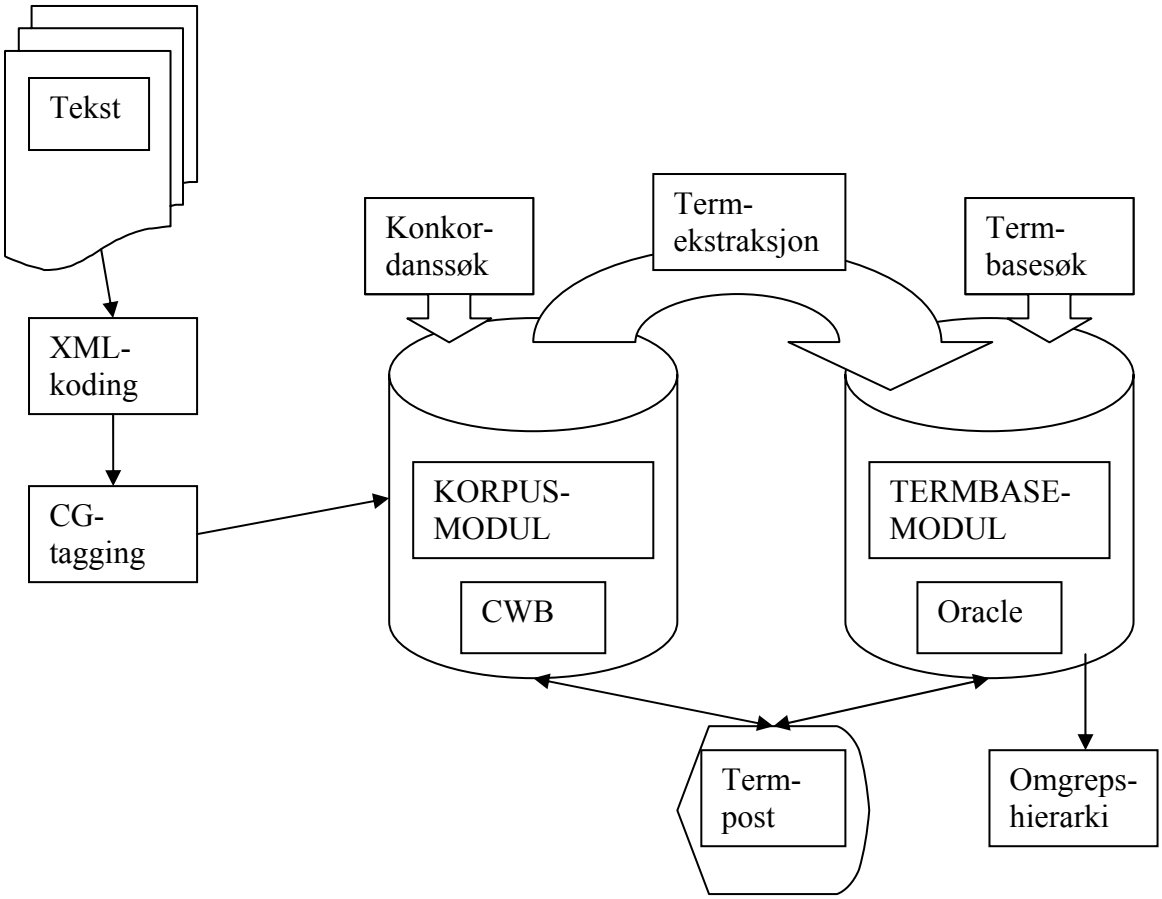
b) *ein språkressursbank*, bygd på den språkteknologiske plattforma, som vidareutviklar, tar i bruk og kvalitetssikrar den fagkompetansen, dei elektroniske verktøya og dei språkressursane som skal til for at norsk fagspråk kan sikrast vidare eksistens og utvikling på fagdomene som har grunnleggjande verdi for norsk samfunns- og næringsliv.

c) eit grensesnitt mot *IKT-baserte applikasjonar* som utnyttar språkressursbanken for effektiv handtering av skriftbasert fagkommunikasjon mellom norsk og engelsk.

Den generelle systemarkitekturen går fram av figur 1 nedanfor, eit prinsippdiagram for KB-Ns verktøysuite.

Med utgangspunkt i programmodular utvikla primært av Paul Meurar og Knut Hofland ved Aksis, og i NHHs spesialkompetanse på domenekunnskap, fagspråkforsking og terminologiske databasar (ved Marita Kristiansen, Kari Øvsthus, Magnar Brekke og Kai Innselset) har KB-N under utvikling ei integrert programsuite med funksjonar for

- tekst/korpus- handtering, ordklassemerking og generering av søkbar konkordans med lenking mot aktuell tekstførekomst og termpost
- XML-koding av makrostrukturen i teksten
- parallellstilling (alignment)
- termekstraksjon: 1) automatisk 2) førehandsekserpert
- hierarkisk omgrepssystematisering:
  interaktiv bearbeiding av omgrepssystem og termhierarki
- termbankregistrering: tekstbasert fleirspråkleg kunnskapsrepresentasjon med eigne felt for domene, kollokasjon, grammatisk informasjon o.l. i termposten
- termbanksøking



Figur 1: KB-N Verktøysuite, prinsippdiagram

3. Metodologisk tilnærming

3.1 Innhenting av tekstmateriale

KB-Ns prosjektplan legg til grunn materiale frå tre ulike tekstlege kommunikasjonstypar som vi reknar med representerer ulike faglegheitsnivå manifestert gjennom ekspositoriske, didaktiske, og populariserande tekstfunksjonar. Desse kan eksemplifiserast som høvesvis forskingsartikkel, lærebok, og avis/tidsskriftartikkel. Tilgang til domenespesifikk tekst frå økonomisk-administrative subdomene byr på svært ulike utfordringar når det gjeld engelsk og norsk, spesielt på høgarere faglegheitsnivå. M.a. gjennom NHHs biblioteksabonnement har vi rikeleg tilgang til aktuell engelsk tekst, men lite på norsk, i og med at dei fleste fagfolk publiserer på engelsk. For didaktisk stoff stiller saka seg litt annleis, og gjennom spesifikke avtalar med norske forlag tar vi sikte på å innhenta dei kvanta norsk tekst som vi treng. Spesielt Fagbokforlaget har stilt seg svært imøtekomande i så måte. Opphavsrettsproblemet er under drøfting med NFF, som også har signalisert interesse for å finne fram til smidige løysingar som ikkje truar forfattarane sine rettar. Bruken av

WWW som kyberkorpus byr på spesielle utfordringer, ikkje minst når det gjeld kvalitetssikring av form og innhald.

3.2 Termekstraksjon

Tradisjonelt har dette skjedd gjennom s.k. ekserpering, ved at ein terminolog les gjennom eit aktuelt dokument og markerer sekvensar som det er grunn til å tru representerer fagtermar. Kontroll av dette skjer normalt i konsultasjon med domeneekspert før innskriving i termbase. Nyare utvikling innan korpuslingvistikken har lagt det metodologiske grunnlaget for termekstraksjon direkte frå eit maskinleseleg tekstkorpus, som på førehand er kompilert av utvalde og (i ein ellar annan forstand) representative e-tekstar. Tekstane gjennomgår ofte ordklassemerking og ein grunnleggjande statistisk analyse (frekvenslister, konkordansar), som dannar basis for avdekking av signifikante førekomstar og kompilering av liste med termkandidatar. Kvaliteten av termkandidatlista avheng av mange ting, men ikkje minst graden av filtrering/sanering av ikkje-termar og annan støy.

I praksis fungerer ”ekserpering” og ”ekstrahering” som komplementære tilnærmingsmåtar. I KB-N blir fagtekstar valde ut i nøye konsultasjon med fageksperter før innlegging i dokumentbasen. Eit sett av data-algoritmar ekstraherar termkandidatar frå korpus (sjå nedanfor), og terminolog/domeneekspert vel termar frå kandidatlista, etablerer ein omgrepsstruktur og identifiserer manglande termar, noko som sjølvsagt ikkje lar seg automatisera.

Automatisk termutvinning frå elektronisk tekst foregår i KB-N langs to hovudlinjer. Den eine er ekvivalensfinning i parallellstilte fleirspråklege tekstversjonar, som ein også utnyttar med stort hell i det nordiske NorNa-prosjektet[5] der gruppa er med. Innan dei fleste subdomene vil dette vera ein grunnleggjande men svært avgrensa metode. Den andre er automatisert termfinning i frittståande tekstar på norsk eller engelsk, ei vesentleg større metodologisk utfordring. Desse tekstane er samanliknbare i den forstand at dei representerer same kunnskapsdomene og kommunikasjonsform, men er ikkje omsette tekstar. Relativt enkle algoritmar for å identifisera lågfrekvente, men fagleg signifikante førekomster, har gitt brukbare resultat for engelsk (System Quirks s.k. ”Weirdness”-funksjon og ”Ferret”-modul), men for norsk er dette tilnærma jomfrueleg mark. Automatisk ordklassemerking (tagging) opnar opp for å utnytta generaliserbare mønster av ordklassesekvensar. Planen er å raffinera metodane i

[5] http://www.norna.dk/

betydeleg grad, men Paul Meurers prosjektinnsats har alt nådd lovande resultat i å vidareutvikla og prøva ut ymse kriteriesett og algoritmar. Nedanfor vil vi gi døme på bruk av denne metoden.

Som første ledd i arbeidet med å fanga opp termkandidatar blir det generert ei liste over nominalfrasar i materialet som etter visse kriterium kan representera moglege fagomgrep. Tabell 1 viser typar av kriterium som ligg til grunn for filtreringa. KB-N-systemet vil her kunna fungera som ein prøvebenk for utvikling og raffinering av slike ekstraksjonsalgoritmar. Ei termkandidatliste vil også vera ein peikepinn for kva eit dokument omhandlar, og følgjeleg for kva subdomene dokumentet skal tilordnast. Her er talet på førekomstar i dokumentet pr. ekstrahert streng særleg nyttig, saman med spreiing på tvers av tekstar og domene.

3.3 Omgrepssystematisering

Omgrepssystematisering uttrykkjer tradisjonelt hovudsakleg generiske/partitive omgrepsrelasjonar innanfor eit avgrensa fagdomene. Men tekstar henta frå økonomisk-administrative fagdomene syner seg som nemnt ofte å vera konglomerat av ulike subdomene og sjangrar, slik at analysen kan enda med eit heilt sett av fragment av ulike omgrepshierarki, ei lite fruktbar framstillingsform. Her har pilotstudium synt at det kan vera meir sakssvarande å leggja til grunn teksten sitt "verksemdsområde" og søkja å gi dette eit tesaurus-liknande omgrepshierarki. KB-N representerer her eit høve til å få prøva ut ein slik metode i stor skala, både med omsyn til strukturering og til grafisk vising av underliggjande element og relasjonar.

Termkandidatfilter
1. lingvistisk filter: (adj. i positiv form)* + subst (minus genitivform) adj +”og/eller” + adj + subst subst + ”-” + ”og/eller” + subst 2. Navnegjenkjennar: Denne overstyrer det lingvistiske filteret etter visse kriterium (blir utvikla i eit eige prosjekt uavhengig av KB-N)

3. Signifikanskalkyle (”Weirdness”) For kvar ordform blir førekomsten i ein relevant fagtekst samanlikna med førekomsten i eit stort allmennkorpus (Hoflands norske aviskorpus ved Aksis, ca. 350 mill. ord). ”Sjeldne” ord i ein fagtekst er ofte fagtermar.

Tabell 1: Kriterium for automatisk ekstraksjon av norske termkandidatar

Med utgangspunkt i dei termpostane som er registrerte i basen kjem no den innleiande omgrepssystematiseringa. Nødvendig fagkunnskap vil koma inn i konsultasjon mellom terminolog og domeneekspert etter at omgrepa er grovsystematiserte av prosjektmedarbeidarane. Hovudverktøyet i denne prosessen er eit dynamisk omgrepshierarki som avspeglar den aktuelle kunnskapsstrukturen, der nye omgrep fortløpande blir føyde inn i hovudhierarkiet.

Domeneeksperten vil på dette punktet lett kunna slå fast korvidt viktige omgrep og termar faktisk manglar, noko som ein automatisk ekstraksjonsalgoritme i prinsippet ikkje kan dersom desse ikkje er representerte i korpus-utvalet. Poenget er at språkteknologiske løysingar alltid må søkja ein optimal interaksjon mellom menneske og maskin.

4. KB-Ns programverktøy og brukargrensesnitt: ein kort presentasjon

4.1 Innleiing

Programvara som skal presenterast nedanfor, er utvikla av forskar Paul Meurer ved AKSIS, UiB i samband med prosjektet ”Norsk språkbank - eit verktøy for korpussøk og -administrering”. I samarbeid med KB-N-prosjektet har Meurer gjort ei hel rekkje spesialtilpassingar for KB-N. Omtalen av grensesnittet nedanfor er knytt opp mot konkrete døme i samband med automatisk termekstraksjon, term- og ekvivalentregistrering, omgrepssystematisering, og vising av konkordansar og kollokasjonar. Desse termane blir forklarte undervegs.

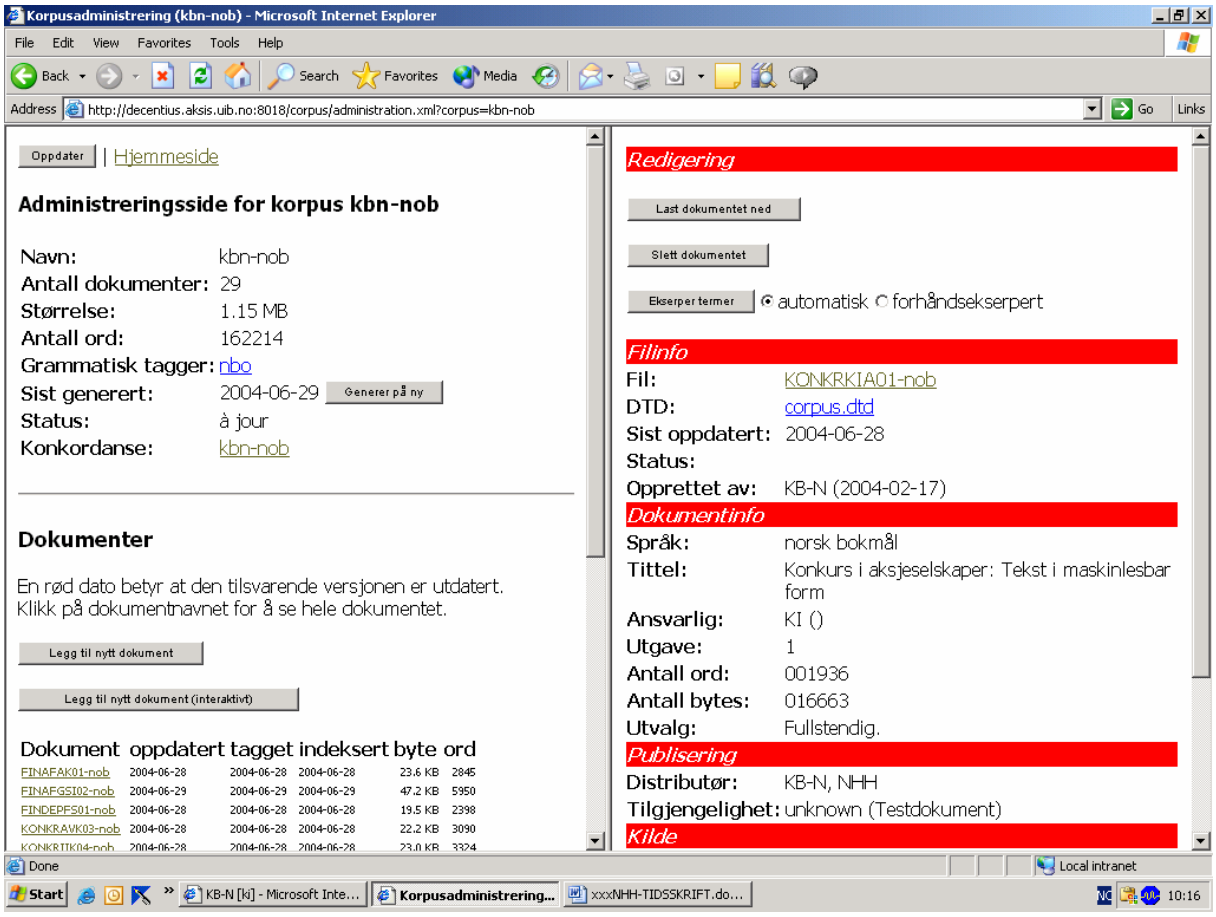
4.2 Tekstbasen

Ved dette stadiet har KB-N-prosjektet liggjande inne to basar, ein norsk base og ein engelsk. Desse basane er bygde opp av parallelltekstar i strengaste forstand. Kvar enkelt dokumentfil i den norske basen har sitt motsvar i ein fil i den engelske basen som representerer enten måltekst eller kjeldetekst i omsettingssamanheng. Base-suiten vil seinare bli utvida med basar for norske og engelske separatfiler, dvs. samanliknbare filar som ikkje er omsetjingar av andre. Bilde 1 nedanfor viser Korpus-basen sitt

administreringsvindu. Vinduet har to rammer; den til venstre inneheld overordna opplysningar om basen. Øvst finn vi m.a. opplysningar om talet på dokument (filar) som er lagde inn og den totale storleiken på dette materialet uttrykt i megabyte og tal på ord. I den nedste delen avramma ligg filkatalogen med tilsvarande storleiksopplysningar for kvar enkelt fil. Det er også via denne ramma ein legg inn nye filar i basen. Desse er førehandskoda i XML i samsvar med XCES-kodestandarden (Corpus Encoding Standard for XML[6]). Ved innlegging i korpusbasen blir materialet automatisk grammatisk annotert ved hjelp av Oslo-Bergen-taggaren[7].

Plasserer vi peikaren på eit filnamn i filkatalogen, får vi fram detaljopplysningar om denne filen i høgre ramme. Øvst i ramma ser vi tre knappar. Den tredje knappen ovanfrå aktiverer termekstraksjon frå den valde filen. Her kan ein velja mellom automatisk ekstraksjon og ekstraksjon basert på manuell førehandskoding av termar. Vi skal i det følgjande visa arbeidsgangen ved termekstraksjon og påfølgjande termregistrering i Term-basen. Dette arbeidet føregår fullt ut i Term-basen sitt tredelte vindu, som blir omtala nedanfor, men val av fil og iverksetjing av ekstraksjonen skjer altså frå Korpus-basen sitt administreringsvindu.

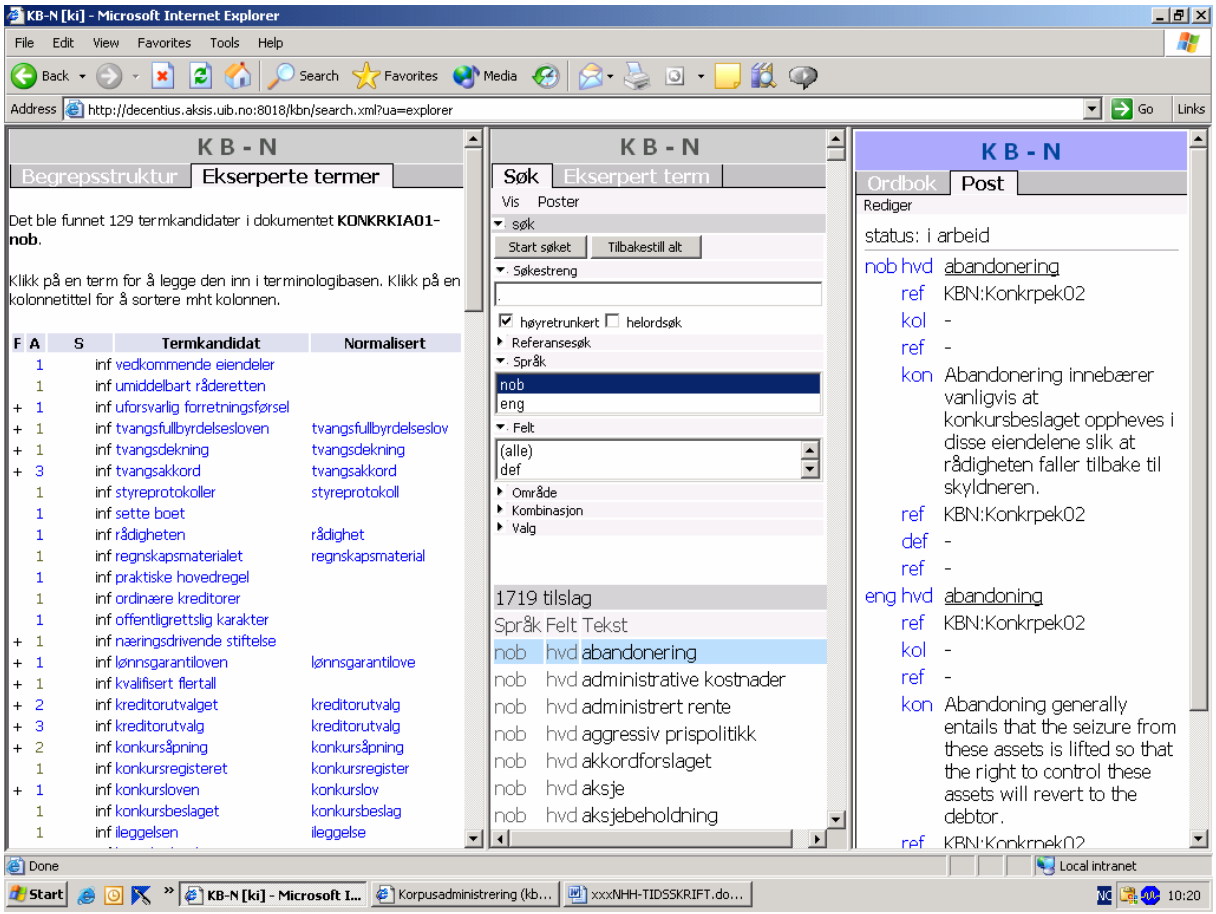
[6] http://xml.coverpages.org/xces.html
[7] http://www.hf.uio.no/tekstlab/tagger.html#Oslo-B



Bilde 1: Korpus-basen sitt administreringsvindu

4.3 Termbasen

Frå Korpus-basen sitt administreringsvindu (bilde 1 ovanfor) valde vi automatisk termekstraksjon. Resultatet kjem fram i venstre ramme i bilde 2 nedanfor. Vi er no ikkje lenger i Korpus-basen, men i Term-basen, og før vi går vidare med å demonstrera termregistrering med utgangspunkt i automatisk termekstraksjon, skal vi gi ein omtale av Term-basen. Denne modulen har berre eitt vindu som til gjengjeld har tre rammer, kvar med to modi. Vi skal ta for oss rammene enkeltvis, knytt opp mot det konkrete dømet vi er i gang med.



Bilde 2: Termbase-vinduet med termkandidatar i venstre ramme

4.3.1 Venstre ramme

Venstre ramme i bilde 2 har eit Termekstraksjons-modus og eit Omgrepsstruktur-modus. Vi er no i Termekstraksjons-modus, og ramma viser difor dei termkandidatane som er automatisk ekstraherte frå den valde filen i korpuset i samsvar med kriteria omtala ovanfor i tabell 1. Vi må igjen understreka at det er termkandidatar som blir ekstraherte; ikkje termar.

Kandidatlista inneheld følgjeleg strenger som Oslo-Bergen-taggaren tolkar som førekomstar av eit sett med førehandsdefinerte typar av nominalfrasar (lingvistisk filter) der førekomstane ligg over ein definert terskelverdi for frekvens i eit allmennspråkleg referansekorpus (”weirdness”-filter). Vi arbeider også med å skreddarsy ei namnegjenkjennar for prosjektet for å kunna ekstrahera namn på ulike institusjonar og organisasjonar med relevans for våre domene. Denne namnegjenkjennaren vil overstyra de ovannemnde filtera. I tillegg vil det bli utarbeidd ei stoppliste primært for å skrella bort uønskte leksem (særleg adjektiv) i fleirords-termkandidatar. T.d. ønskjer vi å få ut ”vedkommende” og ”umiddelbar” i bilde 2, men å ta med ”uforsvarlig” og ”ordinær” inn i registreringa.

4.3.2 Midtramma

Midtramma har modiane Søke-modus og Ekvivalens-modus. Bilde 2 viser ramma i Søke-modus. Her har vi søkt fram alle termar som alt ligg inne i Term-basen og ser altså toppen av termlista nedst i ramma. Ved å klikka på ein term i denne lista kjem tilsvarande omgrepspost opp i høgre ramme der den kan studerast og eventuelt redigerast.

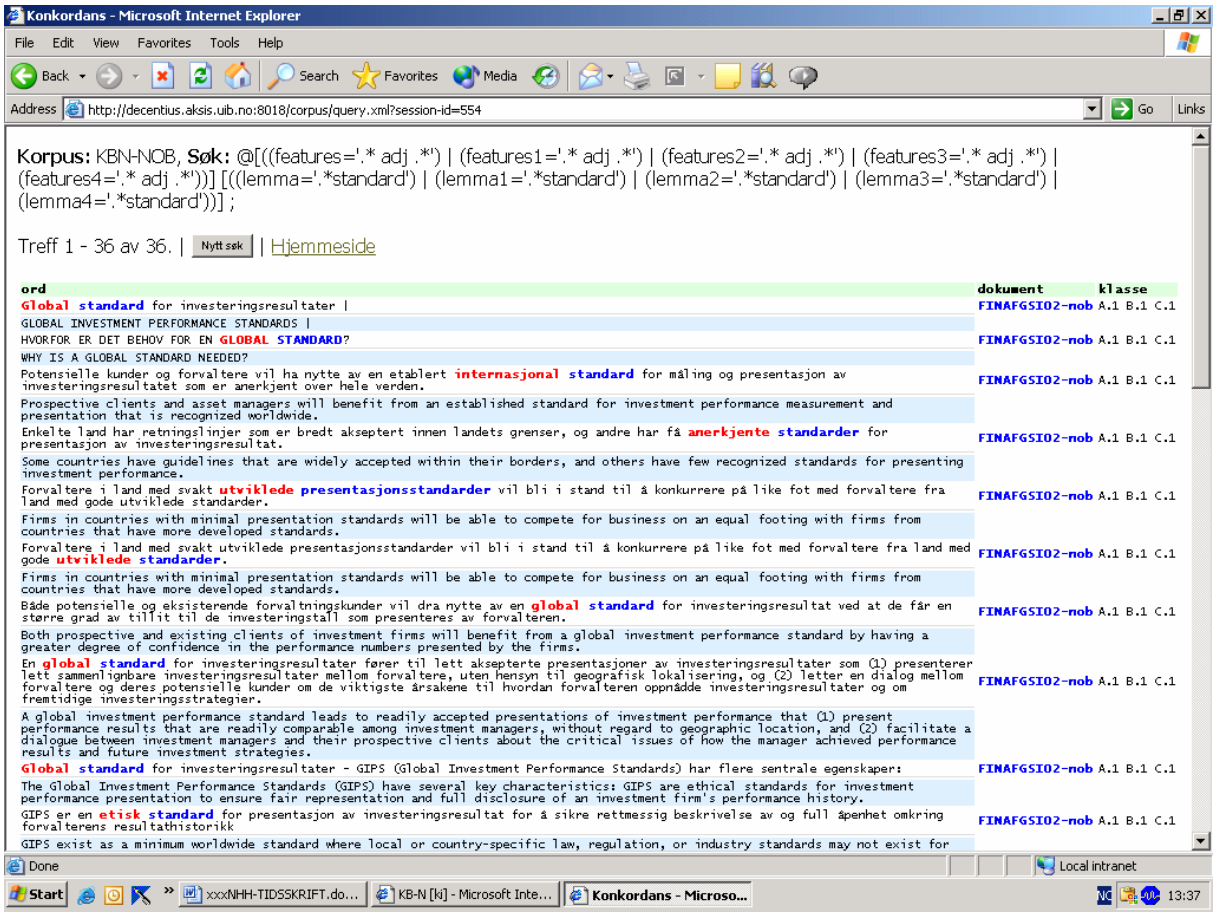
4.3.3 Høgre ramme

Denne ramma viser omgrepspostane i to ulike modi, høvesvis Registrerings-modus og Ordboks-modus. I terminologiarbeidet er det berre Registrerings-modus som er relevant, og det er denne som er vist i bilde 2. Ordboks-modus er ein visingsmodus for sluttbrukarar.

Alle relevante opplysningar om eit gitt omgrep blir samla i éin omgrepspost som er bygd opp av eit sjølvvald sett av førehandsdefinerte felttypar, de fleste repeterbare. Nye opplysningar blir lagde til etter kvart som arbeidet med omgrepet skrid fram. Felta i posten som visest ovanfor er baserte på ein fast mal som blir brukt i samband med automatisk termekstraksjon. I likskap med sjølve feltinnhaldet kan felttypane redigerast etter behov (sletting, nye felt frå lukka liste, repetisjonar mm.).

5.4.4 Vising av konkordansar og kollokasjonar

Ovanfor har vi demonstrert korleis vi registrerer termar i Term-basen med utgangspunkt i automatisk termekstraksjon frå filar i Korpus-basen, deretter korleis vi kan byggja hierarki av registrerte termar.



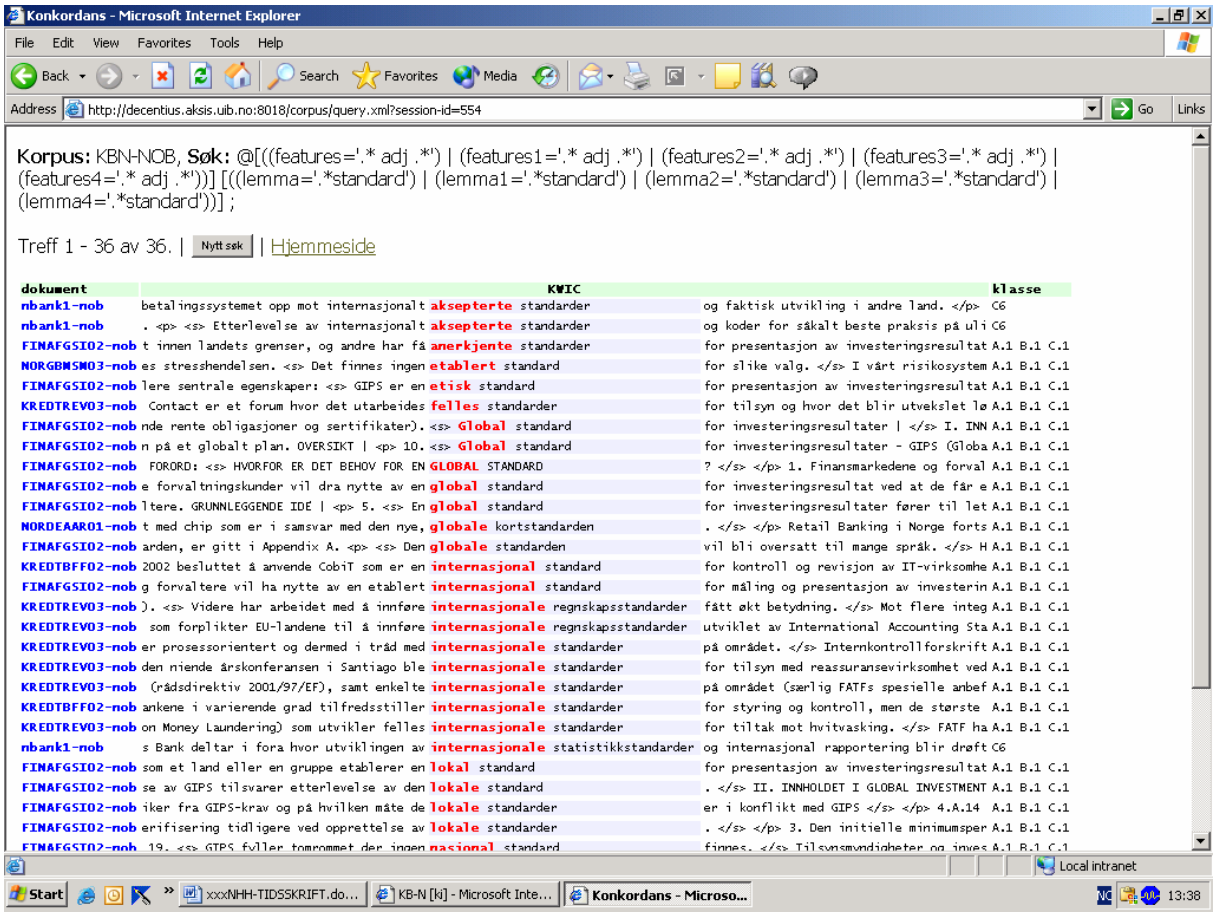
Bilde3: Korpus-basen sitt konkordansvindu i parallell-modus

Vi forlet no Term-basen og går tilbake til Korpus-basen for å visa nokre av søke- og visingsmåtane som er tilgjengelege der. Ein **konkordans** er kort fortalt ei samanstilling av alle tekstlinjer eller setningar som inneheld ein gitt søkestreng. Som døme kan vi halda oss til søk på ulike typar av standardar. Vi vel eit søk som skal fanga opp alle førekomstar av fleirordstermar som inneheld ”standard” som kjerne, altså eit adjektiv etterfølgd av ”standard” åleine eller eit samansett ord med ”standard” som kjerne.

Resultatet av søket ser vi som ein parallell-konkordans i bilde 3. Programmet går gjennom kvar enkelt fil i den norske Korpus-basen, plukkar ut alle setningar som inneheld førekomstar som tilfredsstiller søket, og samanstiller desse setningane ein etter ein med den tilsvarande omsetjinga i den engelske Korpus-basen.

I tillegg til å kunna studera kva norske termar som er brukt for ulike standardar i dokumenta i Korpus-basen, har ein her lett tilgang til dei engelske ekvivalentane slik at ein m.a. kan avdekka eventuelle synonym. Dette skjermbildet er også eit framifrå utgangspunkt for å studera fraseologien som knyter seg til termane på begge språk.

Som nemnt tidlegare finst det i tillegg til Parallell-modus også to andre modi, nemleg Kontekst-modus og KWIC-modus (Key-Word In Context, sjå bilde 4). Begge desse modiane viser konkordans berre for den basen ein søkjer i (altså norsk eller engelsk). Skjermbildet for Kontekst-modus svarar til ei einspråkleg vising av Parallell-modus, og vil difor ikkje bli nærare omtala her.



Bilde 4: Korpus-basen sitt konkordansvindu i KWIC-modus

Frå Korpus-basen sitt søkevindu kan vi også velja vising av **kollokasjonar**, dvs. leksem som hyppig opptrer nær kvarandre, vanlegvis innanfor eit "vindu" på 3-4 ord til høgre eller venstre. Ei klassisk kollokasjonsvising er eit tabellarisk oversyn over dei ulike kontekstane eit gitt ord førekjem i, saman med statistisk informasjon om frekvens (absolutt og relativ) og relevans ("Mutual Infomation" (MI)) av kvar kontekst. MI framhevar leksem som opptrer signifikant oftare saman enn enkeltelementa gjer kvar for seg.

Treffa kan sorterast enten etter frekvens, etter MI, eller alfabetisk, og innstillingane kan lagrast for seinare bruk. Slike konkordans- og kollokasjonsstudiar kan hjelpa til med å identifisere viktige element av domenespesifikk fagspråkleg fleirordsterminologi vsa. synonymval og fraseologi, og dermed gi viktige kriterium for

”word sense disambiguation”. Strukturen i den enkelte omgrepsposten tar høgde for at essensiell informasjon av denne typen kan registrerast og gjerast tilgjengeleg for m.a. analyse og omsetjing.

5. “Kan det brukast til noko?”

Når innhaldet i Korpus-basen og Term-basen, dei to substansielle hovedmodulane i språkressursbanken, har nådd ei kritisk masse og saman med tilhøyrande funksjonar er integrerte på den språkteknologiske plattforma, er KB-N-konseptet i prinsippet realisert. Vi vil då ha eit omgrepsorientert tekst- og termbasert kunnskapshandteringssystem innretta mot språkteknologiske applikasjonar, primært innan omsetjing, dokumentasjon, publisering, men med e-læring, undervising og formidling som viktige utvidingar.

Innanfor ramma av KUNSTI-programmet er det éin hovudbruksarena som peikar seg ut som særleg aktuell: maskinomsetjing (MT). Gjennom LOGON-prosjektet[8] føregår det her ei storstilt satsing på utvikling av norsk-til-engelsk MT, og ei kobling mellom KB-Ns kunnskapsbase (spesielt termdatabasen) og LOGONs framtidige MT-system vil kunna utgjera ein interessant og realistisk utprøvingsarena.

Men ved sida av og i utviding av dette perspektivet ser vi sjølvsagt for oss ei brei vifte av distribusjonskanalar. Som tradisjonelle trykksaker kan det lett produserast domeneorienterte einspråklege termsamlingar så vel som to/fleirspråklege fag-glossar for studiar eller omsetjing. I større format kan stoffet trykkjast i form av ei (fag)ordbok, med CD-ROM eller “USB-drive” som opplagte alternative distribusjonsmedium. Men hovudkanalen vil utan tvil bli ein WWW-tilgjengeleg termbase/kunnskapsbank tilgjengeleg via standard nettlesarar.

Ut over den spesifikke MT-applikasjonen som er nemnt, kva behov er det vi har ambisjonar om å imøtekoma? Planen er å bli ei økonomisk-administrativ kunnskapsbank til dømes for Ekspert som søker faguttrykk, for Lekmann som søker definisjon, for Språkstudent som søker språkbrukseksempel, for Forfattar som søker bruksomfang (scope), eller for Translatør som søker ekvivalent og kontekst. Eit bruksområde som er spesielt aktuelt innanfor NHH-miljøet, er direkte kobling til e-læring.

Skal slike ambisjonar kunne oppfyllast, stillest det store krav til kvalitetssikring av så vel Korpus-base som Term-base. Systemet må leggjast til rette for både maskinoppslag

[8] http://www.norskdok. uib.no/projects/?logon

og menneskeoppslag. Kunnskapsinnhaldet krev kontinuerleg validering, standardisering og normering i samsvar med ISOs standardar og prosedyrar, med profesjonell ivaretaking av referansar, dokumentasjon, sitering, og ikkje minst opphavsrettslege spørsmål.

Endeleg kjem dei økonomiske og praktiske spørsmåla omkring drift, vedlikehald og vidare utvikling av databasesystem og distribusjonsnett, rutinemessig oppdatering av innhald, representasjonsform og grensesnitt. Dette krev presis kontroll med termens livssyklus for å unngå opphoping av forelda materiale. Som ein gjennomgåande dimensjon vil vi leggja til rette for aktiv brukarmedverknad. KB-N-basen skal etter planen ferdigstillast 01.01.07. Norsk Språkbank vil då vera ein realitet når det gjeld økonomisk-administrativt domene.