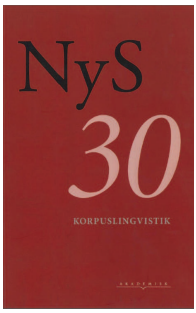
NyS

Titel:	Dansk korpusbaseret forskning. *Hvordan kommer vi videre?*
Forfatter:	Sabine Kirchmeier-Andersen
Kilde:	*NyS – Nydanske Sprogstudier* 30. *Korpuslingvistik*, 2002, s. 11-26
Udgivet af:	Akademisk Forlag A/S
URL:	www.nys.dk



Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Dansk korpusbaseret forskning

Hvordan kommer vi videre?

SABINE KIRCHMEIER-ANDERSEN

Abstract

Artiklen beskriver situationen for dansk korpusbaseret forskning. Situationen er alvorlig, idet der mangler koordination og standardisering på området. Der argumenteres for oprettelsen af en Dansk Sprogbank som skal koordinere og standardisere indsatsen på området.

1. KORPUSLINGVISTIK I DANMARK

Korpuslingvistikken i Danmark har været i langsom fremmarch fra slutningen af 1970'erne med Ruus & Maegaards (1978) ordhyppighedsundersøgelser til i dag, hvor langt de fleste sproglige forskningsmiljøer anvender eller ønsker at anvende korpora og korpuslingvistiske metoder i større eller mindre omfang.

Selvom der sandsynligvis har været en rimelig kontinuerlig aktivitet på området i de sidste 30 år, opleves det alligevel som om udviklingen er forløbet i spring med nogle ganske få højdepunkter – styret af de store korpusudgivelser der har fundet sted undervejs. Således har Bergenholtz' DK87-90 domineret anvendelsen af almensproglige korpora i de forløbne 10 år, for først nu for alvor at blive afløst af DSLs Den Danske Ordbogs korpus som her i foråret er blevet gjort tilgængeligt via internettet[1]. Inden for det fagsproglige område har især det aftaleretlige Dyrberg et al. (1991) og det genteknologiske korpus Lauridsen & Andersen (1993) været vigtige milepæle i korpusudviklingen. Disse store og gode indsamlingsinitiativer har medført en del korpusbaserede undersøgelser, men antallet af tilgængelige korpora er langt fra tilstrækkeligt[2], og der savnes stadig en koordineret indsats for at indsamle og tilgængeliggøre danske tekstkorpora.

Et stort problem for korpusbrugerne har været og er stadig tilgængeligheden af teksterne. De fleste tekster er beskyttet i hht. lov om ophavsret, og det betyder at de fleste korpora i deres helhed kun må anvendes af en snæver kreds af forskere.

Et andet problem er at korpora som rene tekstsamlinger er ved at være utilstrækkelige, og at der derfor i stigende grad arbejdes på at berige teksterne med forskellige former for opmærkning.

Et tredje problem er at der i dag udvikles korpusdatabaser og søgeværktøjer i næsten alle involverede miljøer, men at anvendelsen af teksterne på tværs af miljøerne vanskeliggøres af at teksterne er lagret og opmærket i forskellige formater, og at de derfor ofte kun kan bearbejdes med særligt udviklede værktøjer.

Alligevel er interessen for tekstkorpora stigende også uden for den snævre kreds af forlag og sprogforskere. Det sker i takt med at flere og flere virksomheder og offentlige institutioner bruger dokumentdatabaser og elektronisk arkivering, og at sprogteknologiske virksomheder og især teleselskaber udvikler deres produkter primært på basis af statistiske undersøgelser i korpusmateriale. Inden for især sprogteknologi vil vi i de kommende år opleve et stigende behov for korpora i mange forskellige domæner og inden for forskellige genrer, således at nye sprogteknologiske programmer hurtigt vil kunne tilpasses til nye anvendelsesområder.

Spørgsmålet er om alt dette tekstmateriale ikke burde gøres mere frit tilgængeligt til glæde for både forskning og erhvervsliv.

2. HVAD ER ET KORPUS

Et tekstkorpus er en afgrænset mængde af tekster som er indsamlet efter klart definerede kriterier, beskrevet efter en given standard og som er tilgængelige i elektronisk form. I princippet ville enhver tekstsamling hvad enten den er elektronisk eller trykt, kunne kaldes et korpus, men betegnelsen korpus bruges primært om tekstsamlinger som opfylder de ovennævnte kriterier. Der knytter sig omfattende problemstillinger til hvert af de nævnte kriterier, og nogle af disse har været genstand for diskussion i korpuskredse igennem flere årtier, før der er opnået en nogenlunde konsensus om håndteringen af dem. Jeg vil derfor kun i oversigtsform opridse de centrale problemstillinger for korpusindsamling.

2.1 AFGRÆNSNING AF TEKSTMÆNGDE

Hvor meget tekst et korpus bør omfatte, afhænger i høj grad af hvilke undersøgelser man ønsker at foretage. Til kvalitative undersøgelser kan et korpus på få tusinde ord være tilstrækkeligt, hvorimod kvantitative undersøgelser kræver korpora af en vis størrelse – typisk over en million ord. Efterhånden som flere og flere tekster er blevet tilgængelige i elektronisk form er ordmængden i de indsamlede korpora vokset nærmest eksponentielt. DanWord (1978): 250.000 ord, DK87-90 (1987-90): 4 mio ord, DSL (2001): 40 mio ord. Den indsamlede ordmængde har været betinget af de ressourcer som har været til rådighed til korpusindsamling. Efterhånden er indsamling og bearbejdning af teksterne dog blevet betydeligt forenklet eftersom mange tekster foreligger i elektronisk form, således at kostbar tid til indscanning eller indtastning af teksterne i dag ofte kan undgås.

Korpora giver normalt statiske øjebliksbilleder af en tekstmængde og dermed det sprog der bruges i indsamlingsperioden[3]. Nogle forlag bruger imidlertid åbne korpora i deres ordbogsproduktion. Det betyder at korpusset er under konstant udvikling, og at der hele tiden tilføjes nye tekster. COBUILD har således gennem flere år anvendt betegnelsen *Monitor Corpus* om deres kontinuerligt voksende tekstdatabase. Dette åbner igen for interessante forskningsperspektiver, idet man får mulighed for ikke blot at studere statiske øjebliksbilleder af sproget på et givet tidspunkt, men også for at undersøge udviklingstendenser over en årrække. Denne type tekstsamlinger bliver dog i reglen mindre fokuseret, og ansvaret for at en sproglig undersøgelse er baseret på et forsvarligt sammensat delkorpus bliver således lagt i hænderne på f.eks. forskeren eller ordbogsredaktøren.

Diskussionen om hvor mange ord et korpus skal indeholde for at resultater af søgninger skal kunne kaldes repræsentative, og for at påstande som er baseret på korpusundersøgelser, kan fremføres med en vis vægt, er efterhånden forstummet til fordel for en mere nuanceret beskrivelse af korpussets beskaffenhed og en vurdering af søgeresultatet relativt til korpussets sammensætning og omfang. Selv i et meget stort korpus kan man komme ud for at man kun finder meget få eller ingen eksempler på et sprogligt fænomen. Ligeledes kan hyppige forekomster af et ord opstå mere eller mindre tilfældigt.

Set ud fra et videnskabeligt synspunkt kan korpora således danne et glimrende udgangspunkt og inspiration for empiriske undersøgelser, men de er næppe egnet til på betryggende vis at verificere eller falsificere generelle sprogvidenskabelige påstande. Man har derfor bl.a. inden for den generative grammatiske tradition (Chomsky 1964) udtrykt betydelig skepsis overfor anvendeligheden af tekstkorpora, idet man med rette hævdede at introspektion og udgangspunkt i 'Competence' var et langt bedre grundlag for undersøgelsen af sprogets struktur end den 'Performance' som kommer til udtryk i et korpus.[4] At korpuslingvistikken trods denne skepsis har haft en betydelig fremgang viser nedenstående skema.

Udvikling i antallet af korpusbaserede studier af engelsk i perioden 1965-1991 (Johansson 1991):

- - til 1965: 10
- - til 1975: 50
- - til 1985: 240
- - til 1991: 320
- - til 2001 : ???

2.2 PRINCIPPER FOR TEKSTINDSAMLING

Ligesom det er vanskeligt at give klare mål for korpussets omfang, er det vanskeligt at give klare kriterier for hvilke tekster der bør indgå i et korpus. Dette afhænger igen af hvilken type undersøgelse man ønsker at foretage. Et korpus som skal bruges til at afdække forekomsten af anglicismer i danske aviser, vil nødvendigvis være anderledes sammensat end et korpus som skal danne udgangspunkt for undersøgelsen af den juridiske diskurs i danske domme. Og et korpus der skal danne udgangspunkt for ordbogsproduktion, skal nødvendigvis have en helt anden sammensætning end et korpus der skal bruges til at udvikle en talegrænseflade til en mobiltelefon eller et dikteringsprogram.

Men når korpusset først er indsamlet, beskrevet og klar til brug, udgør den en nærmest uopslidelig ressource for efterfølgende forskning og udvikling af sprogteknologisk software, sålænge korpussets sammensætning er klart dokumenteret.

2.3 STANDARDER FOR BESKRIVELSE AF TEKSTER

Det fremgår af ovenstående at både kriterier for korpussets omfang og teksternes sammensætning er meget vagt definerede, og at holdbarheden af de konklusioner der træffes på grundlag af et korpus i høj grad beror på et skøn. Anvendeligheden af resultatet af en korpusundersøgelse står og falder således med om det er muligt at vurdere hvilke tekster der er indgået i undersøgelsen, og efter hvilke søgekriterier oplysningerne er trukket ud. Dette kan kun lade sig gøre, hvis man forsyner råteksten med yderligere information – den såkaldte opmærkning. Der findes to typer af opmærkning:

1. Den ekstra-lingvistiske opmærkning. En såkaldt 'header' eller teksthoved, dvs. information som knytter sig til tekstens omfang (antal ord), oprindelse (de klassiske bibliografiske oplysninger som forfatter, årstal, forlag osv.) og indhold (sprog, medium, genre og emne).
2. Den lingvistiske opmærkning. Såkaldte tags eller meta-tekst, dvs. information som knytter sig til de enkelte enheder i teksten: grammatiske oplysninger (f.eks. ordklasser og syntagmegrænser), semantiske oplysninger (f.eks. ordbetydninger og relationer mellem ord) og diskursoplysning (f.eks. taleture i en dialog).

Begge former for opmærkning er illustreret i figur 1 for en lille tekststump taget fra en avis. Eksemplet er hentet fra EU-projektet PAROLE, som er noget af det tætteste vi kommer på en EU-standard for korpusopmærkning[5].

FIGUR 1

```
<tei.2>
 <teiHeader type=text>
  <fileDesc>
   <titleStmt>
    <title>Tagged sample of: 'Jeltsins skæbnetime'</title>
   </titleStmt>
   <extent words=158>158 running words</extent>
   <publicationStmt>
    <distributor>PAROLE-DK</distributor>
    <address><addrline>Christians Brygge 1,1., DK-1219
Copenhagen K.</address>
    <date>1998-03-20</date>
    <availability status=restricted><p>by agreement with distribu-
tor</availability>
   </publicationStmt>
   <sourceDesc>
     <biblStruct>
      <analytic>
       <title>Jeltsins skæbnetime</title>
       <author gender=m born=1925>Nikulin, Leon</author>
      </analytic>
      <monogr>
       <imprint><pubPlace>Denmark</pubPlace>
        <publisher>Det Fri Aktuelt</publisher>
        <date>1992-12-01</date>
       </imprint>
      </monogr>
     </biblStruct>
   </sourceDesc>
  </fileDesc>
<profileDesc>
    <creation>1992-12-01</creation>
    <langUsage><language>Danish</langUsage>
    <textClass>
     <catRef target="P.M2">
     <catRef target="P.G4.8">
     <catRef target="P.T9.3">
    </textClass>
   </profileDesc>
  </teiHeader>
<text id=AJK>
<body>
<div1 type=main>
<p>
<W lemma="to" msd="AC---U=--">To</W>
<W lemma="kendt" msd="ANP[CN]PU=[DI]-">kendte</W>
<W lemma="russisk" msd="ANP[CN]PU=[DI]-">russiske</W>
<W lemma="historiker" msd="NCCPU==I">historikere</W>
<W lemma="Andronik" msd="NP--U==-">Andronik</W>
<W lemma="Mirganjan" msd="NP--U==-">Mirganjan</W>
                    ...
<W lemma="diktaroriske" msd="XX">diktaroriske</W>
<W lemma="befølelser" msd="XX">befølelser</W>
<W lemma="." msd="XP">.</W>
</p>
</div1>
</body>
</text>
</tei.2>
```

Hver oplysning består af navnet på en oplysningstype <title> og selve oplysningen "Jeltsins skæbnetime" og den afsluttes med en markering af at oplysningen er afsluttet </title>.[6]

Den ekstralingvistiske opmærkning står i begyndelsen af teksteksemplet og indledes med koderne <tei.2> <teiHeader type=text>, som fortæller at opmærkningen følger TEI-standarden[7]. Herefter følger en masse forskellige ekstralingvistiske oplysninger som anvendes til entydigt at identificere og klassificere den pågældende tekst heriblandt titel <title>Jeltsins skæbnetime</title> og forfatter <author gender=m born=1925>Nikulin, Leon</author>.

Den lingvistiske opmærkning er omsluttet af start og slutkoderne <body> </body> og ser i første omgang fuldstændig uoverskuelig ud, men det

skal man ikke lade sig slå ud af. Teksten er splittet op således at der står et ord med tilhørende opmærkningskoder på hver linje. Det er muligt at gennemskue koderne efter en kort gennemgang af opmærkningen af ordet "kendte" i syntagmet "to kendte russiske historikere".

<W ....>kendte</W> markerer ordets form sådan som det optræder i teksten. lemma="kendt" angiver ordets grundform. msd="ANP[CN]PU=[DI]-" angiver at ordet er et Adjektiv i en bestemt grammatisk form, nemlig Normal, Positiv, Commune eller Neutrum, Pluralis, Umarkeret for kasus, Definit eller Indefinit.

Der er tale om en meget høj specificeringsgrad som til gengæld er meget fleksibel, så man kan trække oplysninger ud af teksten alt efter ens behov. I et forskningsprojekt vil man ofte ønske så præcise og detaljerede oplysninger som muligt, hvorimod man til undervisningsformål i gymnasiet vil kunne nøjes med oplysningen om ordklassen. Opmærkningsstandarden gør det endvidere muligt automatisk at konvertere teksten, så den præsenterer sig i et mere læservenligt format, som f.eks. i figur 2.

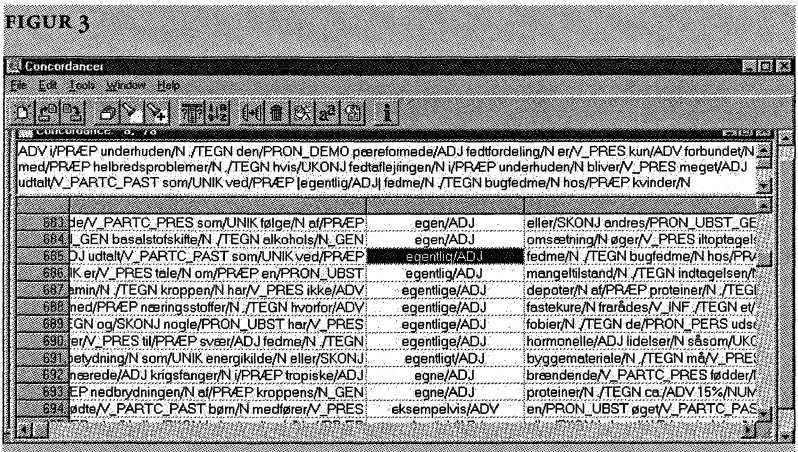
FIGUR 2

to	*kendte*	*russiske*	*historikere*
to/A	*kendte/A*	*russiske/A*	*historikere/N*

Eksempel på simpel meta-tekst til ordklasseidentifikation konverteret fra Parole-koderne

Både teksthovedet og meta-teksten kan anvendes af korpusprogrammer til at filtrere de ønskede informationer fra tekstsuppen. Ved hjælp af hovedet kan man således filtrere korpusset, således at det bliver muligt kun at søge i tekster fra en bestemt forfatter, eller fra et bestemt årstal eller en bestemt genre. Ved hjælp af meta-teksten kan man koncentrere søgningen om bestemte typer af ord eller ordforbindelser, som i figur 3 hvor der er foretaget en søgning på alle adjektiver[8]. De anvendte koder er en tredje variant af Parole-koderne.

FIGUR 3



Desværre kan de færreste korpusprogrammer på tilfredsstillende vis håndtere både lingvistiske og ekstralingvistiske oplysninger, og der ligger et stort arbejde i at udvikle bedre programmer til korpushåndtering.

For at både hovedet og metateksten imidlertid kan anvendes bredt (dvs. på tværs af faggrænser) er det nødvendigt med en vis standardisering af både opmærkningsformen og oplysningstyper. Det ville være særdeles hensigtsmæssigt, hvis der kunne opnås enighed om at fremtidige korpusindsamlinger og opmærkninger overholdt en international eller europæisk standard, f.eks. Parole-standarden som er anvendt i dette eksempel.

2.4 TILGÆNGELIGHED

Mange tekster er ophavsretsligt beskyttet, og det sætter grænser for både indsamling og distribution af korpora. Det mest tydelige eksempel på hvad disse grænser betyder er DSL's nylige distribution af Den Danske Ordbogs korpus med søgeprogrammet Semascope. Programmet gør det muligt at søge i de ca. 40 mio. danske ord som er indsamlet til brug for redaktørerne af Den Store Danske Ordbog. Men der gives kun adgang til en ganske lille del af konteksten omkring de fundne ord, og det er ikke muligt at arbejde med teksterne i deres helhed. Selvom der altså er tale om et yderst prisværdigt initiativ som vil være et nyttigt og anvendeligt redskab i f.eks. sprogundervisningen på næsten alle niveauer, er DSLs korpus ikke anvendeligt i sprogteknologiske sammenhænge og i mange andre sammenhænge, da det ikke kan lade sig gøre at foretage undersø-

gelser af større dele af de tekster som korpusset indeholder. Dermed bliver det umuligt at lave de teksttypologiske og genrespecifikke undersøgelser som er nødvendige f.eks. til forskning inden for fagsprog eller for at finjustere sprogteknologiske værktøjer.

Det er derfor af afgørende betydning at tekstkorpora i videst muligt omfang bliver frit tilgængelige. Det ville være nærliggende at håndtere dem efter Open-Source-princippet, dvs. at de gøres frit tilgængelige på betingelse af at evt. modifikationer og forbedringer som andre brugere måtte foretage, ligeledes gøres frit tilgængelige[9]. Man kunne f.eks. forestille sig at nogen samlede et korpus og andre sørgede for opmærkning, hvorefter det opmærkede korpus blev gjort alment tilgængeligt. En anden mulighed er at distribuere korpora via det europæiske initiativ til distribution af sproglige ressourcer ELRA[10].

Et af de største problemer for en fri udveksling af tekster er imidlertid den gældende lov om ophavsret, der giver ophavsmanden til et givet værk eneret til at råde over værket herunder at distribuere det[11]. Det betyder at det i praksis er umuligt at distribuere et korpus uden samtykke fra alle involverede forfattere. Dette problem må og skal løses, hvis der skal komme skred i anvendelsen af korpora.

3. TYPER AF KORPORA OG DERES ANVENDELSESMULIGHEDER

Nedenstående skema beskriver de kendteste danske korpussamlinger. Der findes utvivlsomt mange flere, men de er enten ikke tilgængelige eller ikke tilstrækkeligt dokumenteret, hvilket understreger at der er brug for en koordinerende indsats.

TABEL 1

Korpus	Udgiver	Indhold	Omfang (antal ord)	Årstal	Tilgængelighed
DANwORD	Maegaard/Ruus	børnebøger romaner aviser ugeblade populære fagblade	250.000 pr. delkorpus i alt 1, 25 mio.	1978	ikke tilgængeligt
DK87-90	Bergenholtz	romaner/noveller aviser ugeblade	2 mio. 1 mio. 1 mio.	1990	kun til forskning
Det aftale-retslige korpus	Dyrberg, Faber, Hansen og Tournay	love/lovforarbejder domme kontrakter lærebøger artikler	1,1 mio. 0,8 mio. 2,3 mio. 2,4 mio. 3,2 mio.	1990	kun til forskning
Det danske gentekno-logiske korpus	Riiber/Kaas Andersen/Lauridsen Søndergård	avisartikler kronikker/læserbreve forfattet af hhv. lægmænd og fageksperter juridiske tekster	1 mio	1991	kun til forskning
PAROLE	Keson	almensprog	250.000	1998	frit
DSLs	DSL	almensprog	40 mio	2000	kun til forsknings-formål
Korpus 2000	DSL	almensprog	?	2001	begrænset søgning/frit

Det er efterhånden kun fantasien der sætter grænser for hvem der kan have glæde af korpora, så nedenstående liste er kun et udsnit. Korpora anvendes i dag inden for

- lingvistisk deskriptiv grundforskning
- datalingvistik - historisk sprogforskning
- leksikologi/leksikografi
- terminologi/terminografi
- dialektologi
- sociolingvistik
- sprogpsykologi
- fremmedsprog og oversættelse

- fremmedsprogspædagogik
- taleteknologi - videnshåndtering og kvalitetsstyring

Og der kommer stadig flere til.

4. EN DANSK SPROGBANK?

I de nordiske lande har der i løbet af de seneste år været øget fokus på sprogteknologi og deriblandt korpusteknologi. Dette har medført at der er blevet afsat betydelige summer til bl.a. korpusindsamling og -forskning.

I Norge har man anbefalet bl.a. udviklingen af en Norsk Språkbank som et paraplyprojekt for udvikling af tekst- og talekorpora. I Sverige har man med hjælp fra erhvervs- og teknikudviklingsrådet (NUTEK) bl.a. opbygget et Svensk OrdNet og arbejdet med tekstanalyse for informationssøgning, korpusbaserede leksika og analyseværktøjer.

Selvom man altså har taget væsentlige initiativer til indsamling af elektroniske tekster i de nordiske lande, foregår der p.t. ikke nogen tilsvarende koordinerede aktiviteter i Danmark.

Forskningsministeriets arbejdsgruppe for IT på dansk har derfor i juni 2001 foreslået at der oprettes en Dansk Sprogbank[12].

4.1 FORMÅL

Sprogbankens hovedformål skal være at koordinere indsamling, opmærkning og udnyttelse af tekstkorpora for dansk. Sprogbanken skal støtte og igangsætte initiativer til indsamling af talt og skrevet sprog af almensproglig karakter og af fagsproglig karakter inden for mange forskellige genrer og domæner.

Sprogbanken skal forske i metoder og teknikker til berigelse af de indsamlede korpora med f.eks. morfologiske, syntaktiske og semantiske oplysninger både på det leksikalske og på det strukturelle plan. Det er tanken at sprogbanken skal huse den første danske *treebank*, en samling af sætninger og deres syntaktiske struktur vist som en træstruktur.

Sprogbanken skal forske i metoder og teknikker til udnyttelse af det indsamlede datamateriale herunder værktøjer til statistisk bearbejdning af store tekstmængder.

De indsamlede korpora skal stilles til rådighed for forskningsmiljøer og virksomheder f.eks. via internettet efter Open-Source-princippet.

Sprogbanken skal i nært samarbejde med internationale korpusinitiativer (f.eks. Text Encoding Initiative/PAROLE) sikre at internationale standarder overholdes.

Sprogbanken vil kunne skabe et bedre arbejdsgrundlag for sprogteknologiske forskningsprojekter og for udviklingsprojekter og er velegnet til at bygge bro mellem forskning og erhvervsliv.

4.1.1 Indsamling af korpora

Det er ønskeligt at sprogbanken spiller en central rolle ved indsamlingen af tekstkorpora med henblik på udvikling af sprogteknologi hvilket indebærer:

- fokus på fagsproglige korpora
- indsamling af korpora på andre sprog (f.eks. hovedsprogene) i tilsvarende fagsproglige domæner
- indsamling af parallelle (oversatte) tekster
- indsamling af dialoger
- indsamling af (transskriberet) talesprog

Orienteringen imod bestemte fagsproglige domæner er betinget af at danske virksomheder i stigende grad anvender elektronisk dokumenthåndtering og knowledge management. Der må forudses et stigende behov for sprogteknologiske produkter der kan håndtere og ekstrahere data fra disse kilder.

Orientering imod fremmedsprogede korpora er betinget af et stigende behov for flersprogligt terminologiarbejde, samt et øget fokus på oversættelse og interkulturel kommunikation.

Orientering mod parallelle (oversatte) tekster er betinget af et stigende behov for udvikling af værktøjer til maskinstøttet oversættelse. Parallelle tekster er endvidere også velegnede til systematisk bearbejdning af flersproglig terminologi samt studier af oversættelsesstrategi.

Orientering mod dialoger skal sikre et tilstrækkeligt empirisk materiale til at muliggøre forskning i dialogsystemer, dvs. natursproglige grænseflader.

Orientering mod talesprog skal muliggøre videreudvikling af taleteknologisk software samt intensivere forskningen inden for talesprog generelt.

Det forudsættes at sprogbanken i forbindelse med indsamlingsarbejdet spiller en koordinerende og rådgivende rolle, men at hovedparten af indsamlingen foretages af de miljøer som ønsker at anvende de indsamlede data. Sprogbanken forventes at bidrage med specialiseret viden om registrering, lagring og dokumentation, mens de respektive miljøer spiller en central rolle ved udvælgelsen af de relevante tekster.

4.1.2 Opmærkning af tekster

Opmærkning af teksterne kan foretages på forskellige niveauer:

- morfologi
- syntaks
- semantik
- pragmatik
- prosodi
- samtalesegmenter

Opmærkningen består i at der tilføjes yderligere information til teksten om f.eks. ordenes ordklasser, den aktuelle bøjningsform, ordets betydning eller relation til andre ord i teksten. Der er efterhånden udviklet teknikker til automatisk opmærkning af tekster på f.eks. det morfologiske niveau, mens der på de andre områder stadig kræves en intensiv forskningsindsats for at en automatisk opmærkning kan lykkes. I udlandet øges for tiden aktiviteterne i retning af semantisk opmærkning[13].

Sprogbanken vil kunne yde et vigtigt bidrag til udvikling af dansk sprogteknologi ved at videreudvikle og tilpasse opmærkningsprogrammer til dansk.

4.1.3 Værktøjer til analyse og bearbejdning af korpusdata

For at de rå eller opmærkede korpora kan bringes i anvendelse i forsknings- og udviklingssammenhænge, kræves der programmer som kan trække data ud af teksterne, præsentere dem på en overskuelig måde og bearbejde dem statistisk. Der findes allerede i dag et antal værktøjer til

disse opgaver. Det vil imidlertid blive nødvendigt med udvidelse og videreudvikling af disse programmer, samt en kobling til programkomponenter til statistisk analyse.

Det skal fremhæves at de metoder og teknikker til korpushåndtering som sprogbanken vil kunne stille til rådighed, med stort udbytte også vil kunne anvendes inden for andre forskningsfelter, jf. afsnit 3.

- historisk sprogforskning
- forskning inden for sprogpsykologi
- forskning inden for fremmedsprog og fremmedsprogspædagogik - forskning inden for videnshåndtering generelt

Sprogbanken vil med andre ord være gavnlig for mange andre forskningsområder end det sprogteknologiske.

Sabine Kirchmeier-Andersen
Institut for Datalingvistik,
Handelshøjskolen i København
email: ska.id@cbs.dk

NOTER

1. www.dsl.dk
2. Michael Barlows korpusside ved Department of Linguistics på Rice University indeholder et væld af links til korpora på alverdens sprog. Dansk glimrer ved sit fravær. http://www.ruf.rice.edu/~barlow/corpus.htm
3. Dette er kun i begrænset omfang tilfældet for DK 87-90 idet indsamlingskriteriet har været udgivelsesåret, hvilket har medført at nyudgivelser af ældre tekster er kommet med i samlingen.
4. Chomsky har dog selv modereret sit synspunkt (Chomsky (1964)) og har ikke helt afvist performance-tilgangen i forbindelse med studier af sprogtilegnelse.
5. EU-projektet PAROLE havde til formål at indsamle korpora på alle EU-sprog, at udvikle en fælles standard for beskrivelse og morfosyntaktisk opmærkning samt at forsyne en mindre del af teksterne (250.000 ord) med opmærkningskoder. (Keeson 1998)
6. Opmærkningsformalismen i dette eksempel er den internationale standard SGML (Standard Generalized Mark-up Language). Standarden er videreført i den nye standard XML (Extended Mark-up Language) som ydermere har den fordel at den kan anvendes direkte i applikationer på internettet. Opmærkning af nye korpora bør derfor beskrives efter XML-standarden.
7. TEI (Text Encoding Initiative)
8. Konkordansen (men ikke opmærkningen) er produceret med programmet Concordancer for Windows, et sharewareprogram, der er meget velegnet til analyse af tekster med ingen eller kun meget simpel opmærkning. Programmet og en kort vejledning kan hentes fra: http://www.id.cbs.dk/~ska/korpus/korpus.htm
9. Se www.opensource.org
10. European Language Resources Association. Se www.icp.inpg.fr/ELRA/home.html
11. Lov om ophavsret LOVBEKENDTGØRELSE NR. 706 AF 29. SEPTEMBER 1998
12. http://www.fsk.dk/cgi-bin /doc-show.cgi?doc_id=81431&doc_type=29&leftmenu=NYHEDER
13. Dansk er for første gang repræsenteret i et internationalt projekt til semantisk opmærkning: Senseval2. Se http://www.itri.bton.ac.uk/events/senseval/

LITTERATUR

Bergenholtz, H. (1991): DK87-DK90: Dansk korpus med almensproglige tekster. M. Kunøe & E.V. Larsen (red.): *3. Møde om Udforskningen af Dansk Sprog*: 32-42. Århus Universitet 11-12-oktober 1990. Århus.

Chomsky, N. (1964): Formal Discussion. U. Bellugi & R. Brown (red.): *The Acquisition of Language. Monographs of the Society for Research in Child Development* 29: 37-9.

Chomsky, N. (1965) : *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Dyrberg, G. et al (1991): Oprettelse af fagsproglige tekstkorpora: engelsk-fransk-dansk juridisk sprog (aftaleret). In: Ark nr. 60, København.

Johansson, S. (1991): Times change and so do corpora. Aijmer & Altenburg (red.): *English corpus linguistics: studies in Honour of Jan Svartvik*: 305-14. London: Longman.

Keson, B. (1998): *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus.* Det Danske Sprog- og Litteraturselskab (DSL). Kbh.: Lauridsen & Andersen.

Lauridsen, O., Riiber, T. & Søndergård, H. (1991): Erstellung eines dänischen und eines deutschen Textkorpus – Fachsprache Genetik. *Hermes* 6. Århus.

Maegaard, B. & Hanne Ruus (1978): DANwORD: Hyppighedsundersøgelser i moderne dansk: Baggrund og materiale. *Danske studier* 1978: 42-70.

Sperberg-McQueen, C.M. & Burnard, L. (1994): *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* Chicago & Oxford: Text Encoding Initiative.