MULINCO- Korpusplatform til sprog- og oversættelsesstudier

Maegaard, Bente; Offersgaard, Lene; Farø, Ken Joensen; Lepetit, Xavier; Navarretta, Costanza; Pedersen, Johan; Povlsen, Claus

Publication date:
2006

Document Version
Også kaldet Forlagets PDF

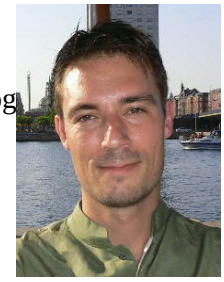
Citation for published version (APA):
Maegaard, B., Offersgaard, L., Farø, K. J., Lepetit, X., Navarretta, C., Pedersen, J., & Povlsen, C. (2006). MULINCO- Korpusplatform til sprog- og oversættelsesstudier. Paper præsenteret ved Tidsskrift for Universiteternes efter- og videreuddannelse, nr. 7: E-læring i sprogfag, Danmark.

MULINCO - Korpusplatform til sprog- og oversættelsesstudier

Bente Maegaard
Direktør
Center for Sprogteknologi
Københavns Universitet
bente@cst.dk
www.cst.dk/bente/index.html

Lene Offersgaard
Seniorudviklingsingeniør
Center for Sprogteknologi
Københavns Universitet
loff@cst.dk
www.cst.dk/loff/index.html

Ken Farø
Ph.d.-studerende
Institut for Engelsk, Germansk og Romansk
Københavns Universitet
kenfaroe@hum.ku.dk
www.engerom.ku.dk



Xavier Lepetit
Postdoc
Institut for Engelsk, Germansk og Romansk
Københavns Universitet
lepetit@hum.ku.dk
www.engerom.ku.dk

Costanza Navarretta
Seniorforsker
Center for Sprogteknologi
Københavns Universitet
costanza@cst.dk
www.cst.dk/costanza/index.html

Johan Pedersen
Lektor
Institut for Engelsk, Germansk og Romansk
Københavns Universitet
jhp@hum.ku.dk
www.engerom.ku.dk

Claus Povlsen
Seniorrådgiver
Center for Sprogteknologi
Københavns Universitet
claus@cst.dk
www.cst.dk/claus/index.html

Forfatterne er ansat ved Center for Sprogteknologi og Institut for Engelsk, Germansk og Romansk, Københavns Universitet. Vores baggrund er altså på den ene side sprogteknologisk og på den anden side sproglig og pædagogisk. Vi har fundet sammen i et udfordrende samarbejde.

1. Indledning

Center for Sprogteknologi og Institut for Engelsk, Germansk og Romansk, Københavns Universitet, har modtaget støtte fra Forskningsrådet for Kultur og Kommunikation til udviklingen af en korpusplatform med parallelkorpora som kan benyttes til at undersøge en række sproglige, herunder kontrastive, problemstillinger. Korpusplatformen skal bruges både i undervisning og forskning. MULINCO står for MUltiLINgual Corpus of the University of COpenhagen.

En korpusplatform består grundlæggende af dels de korpora der tilvejebringes, dels de tekniske muligheder der tilbydes af korpusværktøjerne. Arbejdet i den første periode har derfor bestået i først en behovsanalyse og en kortlægning på både korpussiden og værktøjssiden, og derefter er det tekniske arbejde med at indsamle værktøjer, afprøve dem, installere dem mv. gået i gang, således at vi nu står med den første version af platformen og kan gå i gang med at bruge den.

2. Korpora

Det følger af sig selv at den brug man kan gøre af en korpusplatform, er helt afhængig af de tekster som indgår i korpus. I udgangspunktet fokuserede projektet især på litterære tekster; men af en række grunde er andre typer af tekster også interessante, som det ses nedenfor.

Inden indsamlingen af tekster begyndte, formulerede projektet en række ønsker til korpus: Korpus skal dække de sprog som vi beskæftiger os med i projektet, nemlig dansk, engelsk, tysk, fransk, italiensk, spansk, og hvis muligt også portugisisk og nederlandsk. Den danske del af korpus skal stå stærkt. Eftersom vi gerne vil kunne studere oversættelse, skal i hvert fald en del af korpus bestå af originaler og deres oversættelse. Det er fint at have oversættelse til en lang række sprog, og fx for EU-tekster er dette muligt, mens det kun er muligt at opnå for en begrænset mængde litterære tekster. En del af teksterne skal være korte og afsluttede (det muliggør undersøgelse af tekststrukturer af semantisk, pragmatisk, stilistisk og syntaktisk art). Udover oversatte tekster skal også sammenlignelige tekster (dvs. originaltekster på flere sprog der svarer til hinanden mht. teksttype og indhold) være repræsenteret. Endelig skal korpus være stort nok til at man kan udføre kvantitative undersøgelser.

På grundlag af disse overvejelser blev det besluttet at arbejde sideløbende på:

- At opbygge et "kernekorpus" bestående af litterære tekster, herunder noveller på et af de involverede sprog og oversat til så mange af de andre som muligt. Denne del af korpus (et europæisk novellekorpus) vil dels være et særligt træk ved MULINCO, dels være egnet i pilot-opmærkningsfasen, og vil endelig på længere sigt være meget anvendeligt i undervisningssammenhæng. Dele af romaner, fx 1-2 kapitler, kan også indgå her.
- At opbygge et "kvantitetskorpus" med så mange andre tekster og oversættelser som muligt, især blandt lettere tilgængelige tekster. Denne del af korpus inddrager især EU-tekster, undertekster til film, brugsanvisninger, madopskrifter etc., såvel original med oversættelse som sammenlignelige tekster. Også her vil uddrag af tekster kunne anvendes.

Der har været ret store problemer med at indsamle et moderne litterært korpus, på grund af ophavsret: Forlagene kan ikke videregive tekster til tredjepart, selv ikke til forskning, uden at overtræde de aftaler de har indgået med forfattere og oversættere. Derfor er indsamlingen af tekster ikke gået så hurtigt; men direkte henvendelser til forfattere og oversættere har dog givet nogle gode tekster. En anden følge af ophavsretsproblemerne er at platformen ikke kan gøres åbent tilgængelig for de pågældende tekster, men kun kan benyttes af ansatte og studerende på de to institutter. For ældre tekster gælder ikke samme begrænsning, og dette er blevet udnyttet til at indlemme bl.a. H.C. Andersen, Jules Verne m.fl.

Til den anden del af korpus, som omfatter brugstekster af forskellig slags, har det været nemmere at indsamle tekster. EU stiller bl.a. tekster til rådighed, som projektet udnytter. Dels anvendes udskrifter af forhandlingerne i Europaparlamentet, dels lovteksterne fra Acquis-korpus (Steinberger et al. 2006). Endvidere er der netop i forbindelse med årsskiftet indsamlet en række nytårstaler som kan danne grundlag for sammenlignende studier.

Arbejdet med at indsamle tekster fortsætter, både på den litterære front og for brugsteksterne, og både for oversættelseskorpora og sammenlignelige korpora.

3. Korpusplatformens funktionalitet

3.1 Platformen – behov og valg

En korpusplatform skal give mulighed for at man kan søge i teksterne. Hvis man vil kunne søge på andet end blot bestemte ord eller rækkefølger af bogstaver, må teksterne være opmærket. At en tekst er opmærket betyder at der er indsat oplysninger i den. Fx kan der til hvert ord være knyttet dets ordklasse (POS: part-of-speech). Når sådanne oplysninger er til stede, kan man fx søge på alle ord med en bestemt ordklasse, eller søge efter bestemte rækkefølger af ordklasser. Dette vil blive beskrevet nærmere nedenfor.

Der er i projektet gennemført en behovsanalyse (Farø et al. 2005) for at fastlægge brugernes ønsker til korpusplatformens funktionalitet. Af de overordnede ønsker skal her nævnes:

1. Det skal være muligt at opmærke tekster med forskellige typer oplysninger: morfosyntaktisk (ordklasse, grundform, bøjningsform), syntaktisk og semantisk.
2. Der skal kunne søges på ord og/eller opmærkede oplysninger inden for og uden for periodegrænser og der skal kunne uddrages visse statistiske oplysninger fra korporaene, såsom frekvenslister for ord og specifikke søgninger
3. Der er behov for at platformen kan håndtere tekster der er periodealignerede, sådan at man for en given periode på et sprog kan se den tilsvarende periode på et andet sprog.
4. Platformen skal være internetbaseret, sådan at man let kan få adgang til den ved hjælp af en browser, og søgningerne skal kunne foretages med kort svartid.

Disse krav har ledt frem til at IMS Corpus Workbench (CWB)[1] er blevet udvalgt som korpusplatform[2]. Dette værktøj er meget anvendt i de datalingvistiske forskningsmiljøer. Der er et veldokumenteret og velfungerende søgesprog CQP (Corpus Query Processor Language) og der er i værktøjet lagt vægt på god hastighed ved søgning. Værktøjet er linux-baseret, og det er muligt at programmere en web-grænseflade til selve søgemaskinen, sådan at brugere via web-grænsefladen kan udføre søgninger. Til forskningsformål kan CQP bruges uden licensudgifter.

De ovennævnte brugerønsker leder ikke blot til krav til selve korpusplatformen, men også til behov for værktøjer der automatisk eller semiautomatisk kan foretage opmærkning af teksterne. Dette emne berøres i det følgende afsnit.

3.2 Opbygning af korpora

I projektet har der i den første periode været fokus på opbygning af monolingvale korpora, dette skyldes primært at korpusindsamlingen er gået langsommere end håbet. Selvom et vigtigt mål er parallelle korpora bestående af originaler og tilhørende oversættelser, kan søgning i monolingvale korpora også med stort udbytte benyttes i undervisningssammenhæng.

Før teksterne kan stilles til rådighed i korpusplatformen, skal de naturligvis indsamles og opmærkes. Indsamling af teksterne sker vha. en web-grænseflade[3]. Der indsamles ikke blot den egentlige tekst, men der afleveres også informationer om teksten, fx forfatter, genre, rettigheder, udgivelsestidspunkt[4]. Hvis teksten er en oversættelse, angives dette også, suppleret af yderligere informationer. Disse oplysninger er vigtige når teksterne skal processeres og skal samles i korpora.

Bl.a. er det vigtigt at tekster der benytter gammel retskrivning, dvs. fx benytter *aa* i stedet for *å*, processeres med værktøjer der er tilpasset dette.

Opmærkningen af de indsamlede tekster kan for visse informationstyper foretages automatisk. Det gælder for de morfosyntaktiske oplysninger som kan fastlægges med en POS-tagger og en lemmatiser.

En POS-tagger tildeler ordklasseoplysninger til ord i en tekst, dvs. om det er et substantiv, verbum, osv. Ofte tildeler POS-taggere ikke blot ordklasseoplysninger, men også oplysninger om bøjningsform. Den enkelte POS-tagger har et defineret tagsæt, som benyttes til opmærkningen af ordklasseoplysningerne. Dette tagsæt kan være mere eller mindre detaljeret. CST's tagger for dansk benytter 50 tags der indeholder oplysning om ordklasse og bøjningsform, fx er EGEN og EGEN_GEN de to mulige tags for proprier, enten grundform eller genitiv. Yderligere beskrivelse af taggeren kan ses på http://www.cst.dk/online/pos_tagger.

En lemmatiser er et program der for hver ordform i en tekst finder lemmaformen. En lemmatiser arbejder som regel bedst hvis den har adgang til en ordbog med alle kendte lemmaer. Yderligere beskrivelse af lemmatiseren for dansk kan ses på http://www.cst.dk/online/lemmatiser.

Disse POS- og lemmaopmærkninger knyttes til det enkelte ord. Der findes for alle de sprog vi behandler i projektet, både en POS-tagger og en lemmatiser. For dansk benyttes CST's værktøjer (Jongejan og Hansen 2005) for både nutidigt og ældre dansk, og for de øvrige sprog har vi testet flere tilgængelige værktøjer på primært litterære tekster, og udvalgt dem der giver en god kvalitet af opmærkningen. Det skal bemærkes at POS-taggere for forskellige sprog ofte benytter forskellige tagsæt. På denne måde kan en POS-tagger være særligt tilpasset de fænomener som findes på et givet sprog. I den første fase af MULINCO-projektet er der for hver af de involverede sprog udvalgt en POS-tagger og lemmatiser. Alle værktøjer vil naturligvis udføre opmærkningen med en vis fejlrate. For en god POS-tagger ligger fejlraten ofte på 1,5-5% (Halteren, 1999), og en del af teksterne udvælges derfor til manuel korrektur af POS-opmærkningen og lemmatiseringen, sådan at der for hvert sprog kan opbygges et mindre monolingvalt korpus af tekster, hvor opmærkningen er checket for fejl.

Brugerønsket om opmærkning af syntaksstrukturinformation og semantik vil sandsynligvis på de fleste sprog skulle udføres manuelt eller evt. assisteret af hjælpeværktøjer. På dette forskningsområde findes der nemlig ikke frit tilgængelige værktøjer der kan udføre opmærkningen automatisk, og projektet har derfor på nuværende tidspunkt ikke adgang til den type værktøjer. En manuel opmærkning af syntaktiske strukturer vil også kunne tilpasses de enkelte forskeres fokusområder, sådan at de opnår det optimale forskningsgrundlag, fx ønsker man for spansk at opmærke bevægelsesverber.

Der er udarbejdet procedurer for filarkivering og filhåndtering i den manuelle korrekturfase og den manuelle opmærkningsproces, sådan at teksterne kan håndteres optimalt. Dette sikrer at man på sigt kan udvide opmærkningen i de enkelte tekster med manuelle opmærkninger.

3.3 Søgning i monolingvale korpora

Korpusplatformen skal som nævnt ovenfor både kunne håndtere tekster med opmærkning der knytter sig til ordniveau og også med strukturel opmærkning der bl.a. kan beskrive syntaksstruktur.

Den strukturelle opmærkning er således ikke tilknyttet enkelte ord, men foretages i dokumentet vha. xml-tags.

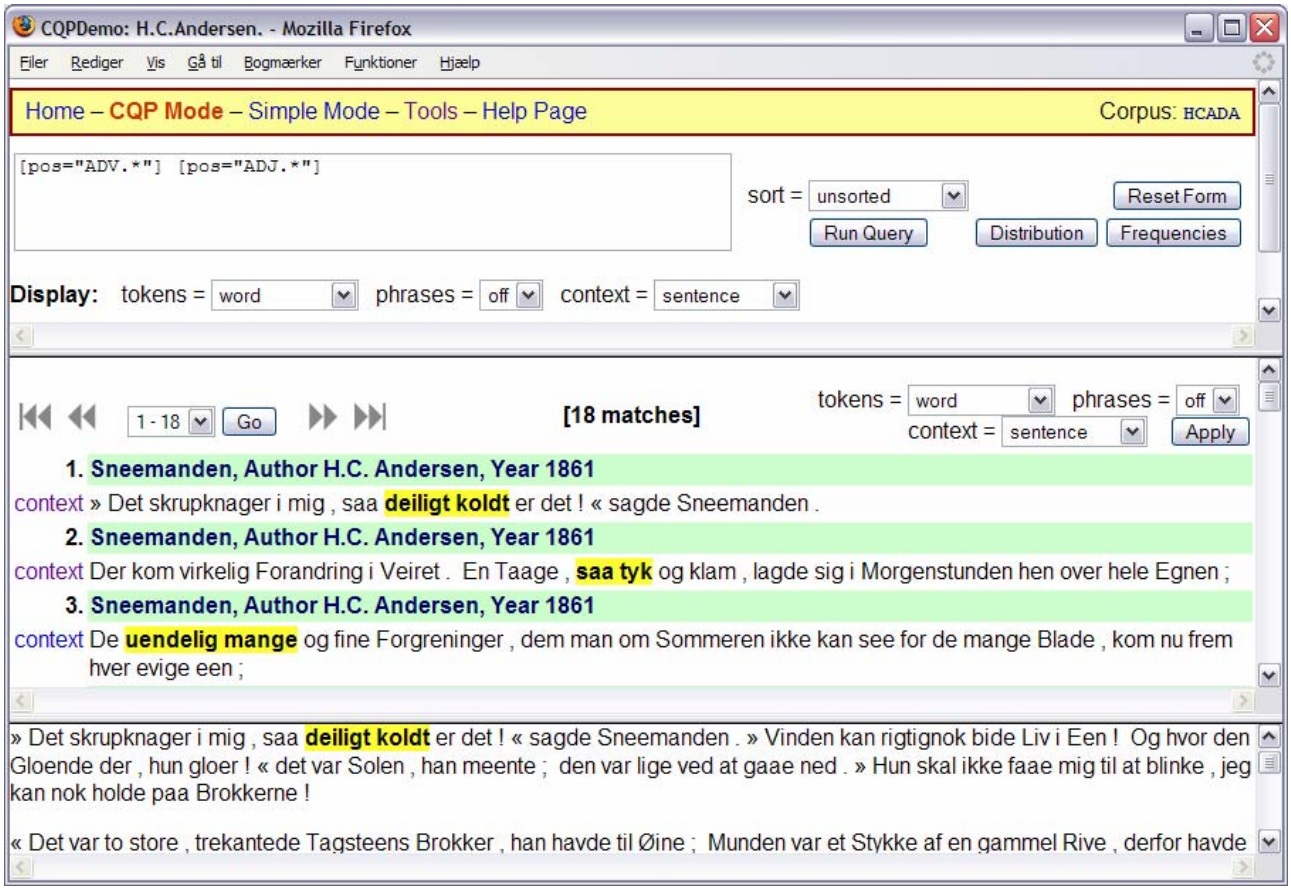
Når man opbygger et monolingvalt korpus i korpusværktøjet, vil dette korpus normalt bestå af en lang række tekster, og man har derfor brug for at kunne opmærke den enkelte tekst sådan at man ved søgningerne kan se netop hvilken tekst der producerede hvilket søgeresultat. Hver tekst opmærkes derfor med oplysninger om titel, forfatter og filnavn. For at man i søgningerne kan referere til et passende segment af teksten, fx en periode eller en paragraf, opmærkes dette også i teksterne. Denne opmærkning af tekstoplysninger og tekststruktur sker vha. xml-tags.

Nedenfor ses et eksempel på tekststrukturelle opmærkninger vha. xml-tags for uddrag af eventyret Sneemanden af H.C. Andersen. Xml-opmærkningen og markerer start og slut på et afsnit (paragraf), <s> og </s> markerer start og slut på en periode. Den morfosyntaktiske analyse af teksten er angivet for et ord pr linje. Oplysninger fra POS-tagger og lemmatiser ses i hhv. 2. og 3. kolonne:

```
<file name="da_Sneemanden.txt">
<titlepage>
<title len="11">Sneemanden.</title>
</titlepage>
…
<p>
<s>
»               TEGN                    »
Seer            V_PRES                  see
en              PRON_UBST               en
Kakkelovn  N_INDEF_SING   kakkelovn
saa             ADV                     saa
deilig          ADJ                     deilig
ud              ADV                     ud
!               TEGN                    !
«               TEGN                    «
</s>
</p>
…
```

Figur 1: Eksempel på tekststrukturelle opmærkninger vha. xml-tags

Et eksempel på den nuværende web-grænseflade for de monolingvale korpora ses i nedenstående figur for det danske HCA-korpus. Søgningen er udtrykt i CQP-søgesproget og angives i tekstvinduet øverst til venstre. Den viste søgning finder tilfælde hvor et adverbium er efterfulgt af et adjektiv. I eventyret Sneemanden er der 18 tilfælde af dette. Det er muligt at vælge at se både ord, POS-tags og lemmaer i søgeresultaterne, i eksemplet er blot ordene vist. Ud for hvert søgeresultat er angivet et link til konteksten, og her viser det nederste vindue konteksten for det første søgeresultat. Der er mulighed for frekvens-oplysninger for den givne søgning ved at klikke på knappen ”Frequencies”. Det er ligeledes muligt at specificere hvilken sortering man ønsker på søgeresultaterne.



Figur 2: Web-grænseflade til monolingval søgning.

Muligheder for anvendelser af søgegrænsefladen for de monolingvale korpora vil bl.a. blive beskrevet i afsnittet ”Eventyr og sprogteknologi i et læringsperspektiv”.

3.4 Flersproglige korpora

De flersproglige korpora opdeles, som nævnt ovenfor, i parallelle og sammenlignelige korpora.

Parallelle korpora

Parallelle korpora består af parallelle tekster på flere sprog, fx eventyr af H.C. Andersen på dansk og oversat til bl.a. engelsk**.** Disse parallelle tekster periodealigneres, sådan at man ved søgninger på et sprog kan se den eller de tilsvarende perioder af teksten på de øvrige sprog. De indledende undersøgelser har vist at litterære tekster kan være vanskeligere at periodealignere end brugstekster eftersom oversætteren i højere grad bearbejder teksten i oversættelsen. Der vil blive udviklet en web-grænseflade til søgning i de parallelle korpora, hvor det er muligt at vælge hvilket sprog man vil udføre søgningen på og hvilke parallelle sprog der skal vises i grænsefladen. Et eksempel på udnyttelse af parallelle korpora kan ses i afsnit ”Parallelkorpora og sprogtypologi” (afsnit 5.1).

Sammenlignelige korpora

De sammenlignelige korpora består af sammenlignelige tekster på forskellige sprog. Et eksempel på et sammenligneligt korpus er de netop indsamlede officielle nytårstaler, hvor der for dansk er samlet nytårstaler afholdt af H.M. Dronningen, mens der på de andre sprog er samlet de tilsvarende

statsoverhoveders taler. Sådanne korpora kan naturligvis ikke periodealigneres. Eksemplet med nytårstalerne uddybes i afsnittet ”Korpusbaseret kontrastiv tekstologi og fraseologi”.

4. Eksempler på hvordan platformen kan anvendes i undervisningen

Nedenfor gennemgås nogle eksempler på brug af korpusplatformen. Her er eksempler på sproglige studier, både monolingvale og kontrastive. Der er også eksempler på litterære overvejelser, samt sproglige følger af samfundsmæssige forhold. Endelig kan de studerende ved at bruge værktøjerne tilegne sig sproglig viden og skærpe forståelsen af det relative i sprogbeskrivelsen. Som eksempel på det sidste kan man bemærke at den inddeling af ord i ordklasser og underklasser som en tagger arbejder med, kan virke fremmed og derfor ”forkert”, fordi den evt. er udtryk for en anden grammatisk tradition end ens egen; men dette kan jo netop bruges som anledning til at diskutere at der ikke findes én sandhed om ordklasser mv. Samtidig kan selve tilskrivningen af ordklasser sættes til diskussion. Er fx det korrekt altid at betragte det italienske *splendente* (glimrende) som et participium eller er det ordenes funktion og gængse brug der bør have indflydelse på den måde de opmærkes på? Endelig er ordklasseinddelingen yderst anvendelig når man diskuterer funktion og betydning af ord som fx kan tilhøre forskellige ordklasser. Ordenes betydning kan kun entydiggøres ud fra en syntaktisk analyse af den kontekst som de optræder i. Dette er fx tilfældet med det italienske *quando* (når/hvornår, dengang), som både kan være en konjunktion og et adverbium og det danske *der* som kan være et adverbium, et formelt subjekt eller et relativt pronomen.

4.1 Eventyr og sprogteknologi i et læringsperspektiv

De første eventyr fra H.C. Andersens hånd blev udgivet i 1835 og blev af den samtidige etablerede elite blandt andet kritiseret for ikke at være tilstrækkeligt belærende - at eventyrene ikke umiddelbart havde en handlingsanvisende morale med det rette pædagogiske sigte. Det mest vægtige kritikpunkt var at det blev anset for helt utilladeligt at lade en litterær tekst være indlejret i en mundtlig fortællestil - at skrive ikke blot om børn, men også som børn var uacceptabelt set fra det litterære parnas.

Det sidste kritikpunkt er af eftertiden blevet vurderet til netop at være det nytænkende og grænseoverskridende ved H.C. Andersens eventyr. Jens Andersen fremhæver således i sin biografi om H.C. Andersen dennes ofte geniale brug af lydord i sin eventyrlige fortællestil (Andersen 2003, pp. 372-75). En sådan mere lingvistisk tilgang til en forståelse af det unikke i H.C. Andersens fortællestil er også af andre forskere blevet anvendt som en konstruktiv fremgangsmåde til en mere fyldestgørende forståelse og fortolkning af H.C. Andersens eventyr og deres virkemåde.

En mere præcis og effektiv analyse af sproget i H.C. Andersens eventyr kan opnås ved at udnytte den ovenfor nævnte sprogteknologiske søgeplatform med resultater fra den automatiske lemmatisering og ordklasse-tagging for dansk[5]. Brugen af de sprogteknologiske værktøjer er selvfølgelig også oplagt i en læringsmæssig proces og sammenhæng.

Eksempler på anvendelser

I en mere fyldestgørende kortlægning af de karakteristiske træk ved H.C. Andersens sprogbrug vil det således være muligt at få lavet en liste over alle de adjektivsyntagmer og adverbiumsyntagmer der forekommer i eventyrene. Her vil hypotesen, som korpusanalysen kunne understøtte eller afkræfte, være: er der gennemsnitligt set flere af disse syntagmer i eventyrene end i eksempelvis almindelig prosa. Forventningen vil være at disse syntagmer (hvor et forholdsvist betydningsfattigt

adverbium (fx *så, ganske* og *rigtignok*) modificerer enten et adjektiv eller et andet adverbium) i eventyrene er signifikant mere frekvente idet de udgør karakteristiske elementer i det talte sprog. Søgemønstret kunne for adjektivsyntagmerne være identisk med eksemplet ovenfor, altså [pos="ADV"] [pos="ADJ"] hvor søgeresultatet vil være enkle adjektivsyntagmer bestående af et gradsadverbium efterfulgt af et adjektiv. Tilsvarende med adverbiumsyntagmet vil søgemønsteret være [pos="ADV"] [pos="ADV"]. En søgning efter enkle adjektivsyntagmer i et meget lille korpus, H.C. Andersens eventyr Sneemanden indikerer at den ovennævnte hypotese vil kunne blive bekræftet af en mere generel korpusanalyse. I eventyret forekommer der således 17 adjektivsyntagmer[6] (ex. *saa dejlig, saa lykkelig* og *ganske underlig*). Dette svarer til en frekvensrate på godt en procent, medens der i et tilsvarende størrelse korpus af skrevet sprog (repræsenteret ved artikler på dagbladet Politikens hjemmeside) overhovedet ikke forekommer nogen adjektivsyntagmer.

De læringsmæssige aspekter er mange og i flere dimensioner. Indlæring af sproglig viden og kompetencer vil således kunne blive udført i forbindelse med en korrekturlæsning af de automatisk foretagne ordklassetilskrivninger. Automatisk ordklasse-tagging er til en vis grad fejlbehæftet og kræver derfor manuel inspektion hvis man skal være sikker på at tilskrivningen er fejlfri. Den påkrævede videntilegnelse af det definerede tagsæt og brugen af denne viden i korrekturlæsningen vil også i en indlæringsproces styrke den sproglige kompetence for de studerende.

Et andet læringsperspektiv er hvordan skriftsprogsnormer har udviklet sig over tid. På H.C. Andersens tid var der ikke hvad man kunne kalde et praktisk anvendt retskrivningsprincip som alle kunne tilslutte sig. I hvert fald to konkurrerende principper var gældende - det fonetiske (hvor ortografien følger lyden) og det etymologiske hvor ortografien mere er baseret på et sproghistorisk grundlag. Disse principper og deres vægt var i 1800-tallet genstand for stor debat hvor meget forskellige holdninger kom til udtryk. Fokus er her hvordan H.C. Andersen i sine skrifter forholdt sig til de samtidige skriftsprogsprincipper, også ud fra en sammenligning med nutidige retskrivningsnormer.

4.2 Sprogundervisning ved hjælp af monolingvale og parallelle korpora

Korpusplatformen giver mulighed for at undersøge anvendelsen af bestemte ord (eller ordklasser) i forskellige værker og er en uvurderlig støtte ved undersøgelse af leksikalske mønstre. Fx kan man finde de kontekster i hvilke adjektivet *god* efterfølges af en præposition. Dette gøres ved at søge på ordet med lemmaet *god* efterfulgt af en præposition (i korpusplatformen udtrykkes dette med søgemønstret [lemma="god"][pos="PRÆP"]) og dernæst sortere adjektivets prædikative valensmønstre fra andre præpositionelle konstruktioner. Figur 3 indeholder eksempler på resultater fra en sådan søgning, hvor søgemønstret er i kursiv:

Kvinder skal blive *bedre til* at score kassen.
Det kan også være , at han er *god mod* dyrene.
Hvad er *godt ved* Nettet ?

Figur 3: Søgeresultater for forespørgsel [lemma="god"][pos="PRÆP"]

I de parallelle korpora kan man undersøge hvordan særlige konstruktioner på et sprog er oversat til andre sprog. Fx kan man undersøge hvordan kombinationer af neksusadverbialer på dansk bliver oversat. Dette gøres ved fx at søge følger af adverbier før og efter verbale former i det danske korpus og dernæst udvælge neksusadverbialerne for at gennemgå disses mulige oversættelser.

Søgningen udtrykkes i korpusplatformen med forespørgslen ”([pos=”V.*”][pos=”ADV”] [pos=”ADV”]+)|([pos=”ADV”] [pos=”ADV”]+ [pos=”V.*”])” som læses: find alle ord opmærket som verber efterfulgt at to eller flere adverbier, eller to eller flere adverbier efterfulgt af en verbalform.

I figur 4 vises to eksempler på brug af to neksusadverbialer i H.C. Andersens Sneemanden og de tilsvarende oversættelser til engelsk og italiensk.

Sneemanden	engelsk oversættelse	italiensk oversættelse
ja hun løb *jo rigtignok* før, da jeg saae stift paa hende, nu lister hun fra en anden Kant!	" Yes, it was running itself, when I saw it a little while ago, and now it comes creeping from the other side.	Lui sì che è corso via prima, quando l' ho guardato fisso, ora sbuca fuori da un' altra parte!"
det er et uskyldigt Ønske, og vore uskyldige Ønsker maae *dog vist* blive opfyldte	It is an innocent wish, and our innocent wishes are certain to be fulfilled.	È un desiderio innocente e i nostri desideri innocenti dovrebbero avverarsi.

Figur 4: Uddrag fra Sneemanden: neksusadverbialer oversat til engelsk og italiensk.

I det første eksempel er de danske neksusadverbialer *jo rigtignok* ikke oversat til engelsk og italiensk, mens neksusadverbialerne *dog vist* som efterfølger det modale *må* (maae) i det andet eksempel bliver oversat med *are certain to* på engelsk og med den modale *dovere* i betinget form på italiensk.

Et andet eksempel er undersøgelse af hvordan aspekt kan udtrykkes på forskellige sprog, ved hjælp af tid, aspektrelaterede konstruktioner, samt brug af adverbier. Dette illustreres i figur 5 med to enkelte uddrag fra Sneemanden og oversættelserne på engelsk og italiensk.

Sneemanden	engelsk oversættelse	italiensk oversættelse
den var lige ved at gaa ned	which was just about to set	stava proprio tramontando
jeg gnavede	I was gnawing	stavo rosicchiando

Figur 5: Uddrag fra Sneemanden: aspekt.

4.3 Parallelkorpora og sprogtypologi

Parallelkorpora kan være et meget anvendeligt redskab i undervisningen i sprogtypologi og oversættelse, i særdeleshed når formålet er at integrere det sprogtypologiske og kontrastive perspektiv direkte i oversættelsesdisciplinen.

I en række emnekursusforløb om sprogtypologi og oversættelse, rettet mod BA-projektskrivning/emneopgaveskrivning på overbygningen, har vi set på hvorledes parallelkorpora kan bruges til at undersøge hvordan man udtrykker spatiale relationer og bevægelsessituationer på hhv. spansk/fransk og dansk. Spatiale relationer er kognitivt set basale og derfor særdeles interessante at undersøge i et sprogtypologisk og kontrastivt perspektiv. Vi har navnlig interesseret

os for hvordan sprogtypologiske forskelle kan have indflydelse på oversættelse af spatiale relationer.

(Talmy 2000) skelner mellem hvad han kalder *verb-framed languages* (fx romanske sprog) og *satellite-framed languages* (fx germanske sprog). I et *verb-framed* sprog som spansk og fransk er bevægelsens retning typisk leksikaliseret i verbet (fx sp. *entrar, salir* [gå ind/ud]; fr. *monter, descendre* [gå op/ned]), mens bevægelsesmåden eventuelt udtrykkes som en adverbiel tilføjelse. I et *satellite-framed* sprog som dansk er det til gengæld bevægelsesmåden der er leksikaliseret i verbet (fx *køre, cykle* osv.), mens retningen typisk tilføjes i en partikel (*satellite*), så som *op, ned, ind, ud, over,* osv.

Vi fandt det interessant at se nærmere på spansk og fransk i forhold til dansk som henholdsvis *verb-framed* og *satellite-framed* sprog, eftersom Talmy ikke refererer specifikt til dansk i sin analyse af problematikken. Vi har navnlig set på hvad det kan betyde for en oversættelse mellem de to romanske sprog og dansk (se fx Andreassen 2005). Forskellige leksikaliseringsmønstre kan nemlig tænkes at have indflydelse på hvilke af bevægelseskonstruktionens informationer der prioriteres højest (Slobin 1997). Det kunne i så fald betyde at de informationer der er leksikaliseret verbalt i et sprog, gives højeste prioritet af brugeren. Dette ville fx betyde at *måde* prioriteres højere i dansk, mens informationer om *retning* tilsvarende prioriteres højest i spansk. Lotte Andreassen har således i sit projekt (Andreassen 2005) undersøgt danske og spanske oversættelser af bevægelseskonstruktioner idet hun stiller følgende spørgsmål:

- Hvilke konsekvenser har forskellene i leksikaliseringsmønstre for oversættelse mellem de to sprog?
- Nedprioriteres nogle typer af information i en oversættelse, mens andre fremhæves?

På basis af paralleltekster kunne hun for det første observere tre grundlæggende forskelle mellem danske og spanske bevægelseskonstruktioner (Andreassen 2005): 1). Komponenten *måde* er i danske bevægelseskonstruktioner, til forskel fra de spanske, typisk leksikaliseret i verbet. 2) Komponenten *retning* er i danske bevægelseskonstruktioner typisk leksikaliseret i en satellit (fx adverbium), mens den i de spanske er leksikaliseret i verbet. 3) Danske bevægelseskonstruktioner kan have begge betydningskomponenter leksikaliseret, mens de spanske typisk kun har den ene leksikaliseret. For det andet kunne hun se at disse forskelle har skabt problemer for oversætterne.

Betydningskomponenten *måde* har givet flest problemer, da de to sprogs leksikaliseringsmønstre har gjort det nødvendigt at tilføje eller udelade information og dermed afvige fra kildetekstens indhold. I nogle tilfælde viste det sig muligt at danne konstruktioner svarende til de originale, men oversætterne er i de fleste tilfælde gået på kompromis og har valgt målsprogets typiske konstruktion. Dette betyder at de danske oversættere har måttet medtage komponenten *måde*, selvom den ikke er en del af den oprindelige spanske konstruktion, og at de spanske oversættere i nogle tilfælde har haft vanskeligt ved at konstruere komponenten måde, og derfor simpelthen har udeladt den, selvom *måde* har været eksplicit udtrykt i den danske kildetekst.

Nedprioriteringen af betydningskomponenten *måde* i de spanske oversættelser fra dansk har selvfølgelig den konsekvens at den nye læser, til forskel fra læseren af kildeteksten, selv må forestille sig hvorledes bevægelsen foregår. Det har i visse tilfælde også betydet at humor er gået tabt i oversættelsen, at personkarakteristika delvist er gået tabt, og at der ikke følges op på informationer der er nævnt tidligere i teksten, således at oversættelsen kan forekomme mindre

sammenhængende end originalen. Komponenten *retning* har ikke på samme måde skabt problemer for oversætterne. Det har her kun været et spørgsmål om hvorledes denne information skulle udtrykkes (verbalt eller i en satellit). Andreassen har således ikke i de undersøgte paralleltekster fundet at sprogenes leksikaliseringsmønstre har betydet at oversætterne har måttet vælge denne information fra eller til.

Hvad angår spansk-dansk har undervisningen, og den efterfølgende opgaveskrivning, været baseret på paralleltekster, dvs. originaltekster (begge sprog) og disses oversættelse i ikke-digitaliseret form (se Andreassen 2005 og Christensen 2005). Eksempler på bevægelsessituationer er taget fra forskellige værker på hvert af de to sprog for at modvirke stilistisk bías i materialet. I udvælgelsen af værkerne har flere faktorer spillet ind. Der er kun anvendt narrative tekster (i modsætning til lyrik og fagsprog) skrevet inden for de sidste 50 år, således at eksemplerne så vidt muligt afspejler det moderne sprog som det tales og skrives. Valget har endvidere været begrænset af at teksterne skulle kunne findes i en tilgængelig oversættelse, og ikke mindst at de indeholdt en tilstrækkelig mængde bevægelseskonstruktioner.

Skulle bevægelseskonstruktionerne bare tilnærmelsesvis have været repræsentative for hvert af de to sprog, ville det selvfølgelig have krævet et betydeligt større korpusmateriale. Ved brug af store korpusenheder bliver muligheden i MULINCO for at søge elektronisk på taggede tekster vigtig, således at man vil kunne begrænse søgningernes output og præcist vil kunne isolere de sproglige størrelser man ønsker at undersøge. Problemet er her at den konkrete problemstilling – bevægelsesverber i et typologisk perspektiv – kræver en manuel opmærkning af korpus hvis en søgning skal give et komplet output. Man vil dog også kunne anvende en række delsøgemønstre der indirekte identificerer de sproglige størrelser der undersøges.

Mht. sammenligningen af franske og danske spatial-partikler giver søgninger på kombinationen "verbal + *op/ned*" i H.C. Andersens eventyr Sneemanden i MULINCO således 10 (relevante) forekomster af konstruktionen. Sammenligner man disse 10 forekomster med deres gengivelse i den franske oversættelse af Sneemanden, kan man klassificere disse i tre kategorier:

1. De tilfælde hvor betydningen i den danske satellit slet ikke er gengivet på fransk, som i *Hun* (solen) *lærer dig nok at løbe* ***ned*** *i voldgraven* som bliver gengivet med *il saura t'apprendre à courir dans le fossé* (egentlig: han lærer dig nok at løbe i voldgraven). Disse tilfælde udgør 5 ud af 10.
2. De tilfælde hvor betydningen i den danske satellit er gengivet på fransk i et syntetisk verbum som ikke eksplicit udtrykker en spatial relation, som i *den var ved at gaae ned* som bliver gengivet med *le soleil disparaissait* (egentlig: solen var ved at forsvinde). Disse tilfælde udgør 4 ud af 10.
3. De tilfælde hvor betydningen i den danske satellit er gengivet på fransk i et syntetisk verbum som udtrykker en spatial relation, som i *Fuldmånen stod op* som bliver gengivet med *la lune monta*. Dette tilfælde udgør 1 ud af 10.

Ud fra denne lille undersøgelse kan man konkludere at de spatiale relationer i dansk enten ikke gengives, eller, når de gengives, leksikaliseres vha. syntetiske verbaler, som ikke nødvendigvis indeholder en eksplicit spatial relation.

Skabelsen af store alignerede elektroniske parallelkorpora indenfor rammerne af MULINCO-projektet vil give mulighed for at lave større, mere dybtgående, og også mangesproglige, sprogtypologiske og oversættelsesmæssige undersøgelser af bevægelseskonstruktioner og spatial-

partikler, om end arbejdet er besværliggjort af problemstillingens semantiske karakter. På den anden side vil mindre parallelkorpora indenfor MULINCO-sprogrækken sagtens kunne bruges i undervisningssammenhæng, hvor kravet om repræsentativitet ikke er så afgørende.
Dette er pilotprojektet Sneemanden (digitaliseret parallelkorpus – 6 sprog) et eksempel på. Som vi har demonstreret i sammenligningen af spatiale relationer i fransk og dansk, kan man lave en række søgemønstre der identificerer konstruktionerne, selvom der selvfølgelig ikke er tale om at en enkelt søgning vil kunne returnere en komplet liste af de ønskede forekomster.

4.4 Korpusbaseret kontrastiv tekstologi og fraseologi

MULINCO rummer muligheder for empirisk-systematisk at inddrage de sprogspecifikke fænomener over ordniveau *kontrastiv tekstologi og fraseologi* i undervisningen. I begge tilfælde er der tale om strukturer, der kun kan erkendes og undersøges funktionelt adækvat ud fra korpusundersøgelser. Samtidig viser erfaringen, at muligheden for empiriske selvstudier af disse fænomener gør sprogvidenskaben mere interessant for studerende på fremmedsprogsfagene. Hidtil har der manglet interlingvale korpora med dansk, som specifikt tager højde for dette. MULINCO udgør derfor et vigtigt empirisk redskab til undervisningen i fremmedsprogslingvistikken.

Korpus kan bl.a. anvendes af studerende til induktivt at opnå indsigt i interlingvale konventioner for specifikke *hverdagstekstgenrer* (jf. Hartmann 1980, Heinemann 2000). En del af MULINCO fokuserer på almensproget, som det forekommer i sammenlignelige tekster og er derfor særligt velegnet til undervisning på universitetet og på ikke-specialiserede niveauer, idet det netop ikke er et fagsprogligt korpus.

Foreløbig er der bl.a. blevet indsamlet sammenlignelige tekster af genren *vittighed* og *officiel nytårstale* fra projektets sprog. Præliminære studier af sidstnævnte (fra 2000-2005) viser fx, at de danske nytårstaler er længere end de tyske, at de danske ikke betjener sig af en indledende formel (jf. *Liebe Mitbürgerinnen und Mitbürger,*), men derimod af en fast afsluttende fraseologisme (*Gud bevare Danmark!*), som oven i købet er fuldstændigt genrespecifik, idet den optræder med en sandsynlighed på 100 %. Dronningetalen indeholder flere formelle elementer end kanslertalen og er i øvrigt sprogligt-stilistisk mere arkaisk (*de danske, på søen, i Danmarks lod*). I begge tilfælde udgør nationalt denoterende substantiver og adjektiver som *Danmark, dansk, danske, Deutschland* og *deutsch* ca. hvert 100. ord i talerne.

En interessant indholdsmæssig strukturel parallel er, at den i de danske nytårstaler obligatoriske nævnelse af de atlantiske dele af riget modsvares af en tilsyneladende fast henvisning til de tyske delstater. I begge tilfælde er der åbenbart tale om en retorisk afsikring af statens særlige struktur (hhv. kongerige og forbundsstat).

Ud over disse tekster er det planlagt bl.a. at indsamle weblogs, kontaktannoncer, kommunepræsentationer på Nettet, brugsanvisninger, madopskrifter og andre kulturelt-sprogligt interessante hverdagstekstgenrer.

Fraseologiforskningen har i de seneste par år været inde i en levende udvikling på tværs af sproglige og teoretiske skel (jf. *Hermes* 35), uden at det tilsyneladende mere end ét sted i Danmark har resulteret i egentlig undervisning på universitetsniveau. Forståelse og undersøgelse af fraseologi, herunder især idiomatik, kræver omfattende empiri, fordi mange frasemer er stærkt ”ikoniske”. Det skaber metodisk problemer for introspektion og praktisk for både interlingval kommunikation og indlæring, hvor den manglende fraseologiske kompetence ofte fører til rent kontekstløs

interpretation af frasemets betydning og andre tegnegenskaber. Disse kan imidlertid ikke udledes af dets udtryksside, men er principielt ususbaserede (Farø, u. udarb.). Med MULINCO får studerende og undervisere nye muligheder for at undersøge dansk og fremmedsproget fraseologi i kontekst, fordi korpus bliver opmærket specifikt med fokus på dette problem. Både brugen af fraseologi interlingvalt og dens oversættelse vil kunne studeres. MULINCO's genremæssigt polytekstuelle koncept gør det desuden muligt at fokusere på fraseologi i forhold til bestemte tekstgenrer, fx reklamer eller opskrifter.

Bl.a. for disse to teoretisk interessante og praktisk vigtige områder af fremmedsprogsundervisningen udgør MULINCO altså et nyt redskab til empiribaseret undervisning.

5. Perspektiver

Som man kan se, er vi netop gået i gang med at udnytte korpusplatformen i undervisningen, og der vil opstå mange nye ideer i det yderligere arbejde. En af de fordele, som vi egentlig ikke havde tænkt så meget på på forhånd, er dette at grammatikbegrebet udfordres af platformen: der kan være mere end én måde at beskrive sprog på, ikke bare mellem sprogene, men også for det enkelte sprog. Dette synliggøres fx meget af de taggere der er anskaffet: De anvender ordklasser på bestemte måder, som naturligvis ikke alle vil være enige i.

Fremover vil vi arbejde mere med søgemønstre der kan fremfinde ønskede konstruktioner mere præcist, og med søgninger i sammenlignelige korpora. Det vil også være muligt at undersøge fx dansk i originale tekster over for dansk i oversatte tekster; her kan sproget undersøges mht. hyppighedsforhold, bestemte ord og vendinger mv. Når forskerne har beriget teksterne med manuelt indførte opmærkninger med forskellige oplysninger (litterære oplysninger, fx om tema, eller semantiske oplysninger osv.), kan de også anvendes til undersøgelser hvor disse oplysninger inddrages.

Visse faciliteter findes endnu ikke i platformen, bl.a. er der planlagt arbejde med alignering, det gælder både periodealignering, hvor tekster ordnes i par af kildeperiode og målsprogsperiode(r), og ordalignering hvor man tilsvarende kobler ord og flerordsenheder på de to sprog med hinanden.

Fremtiden vil altså byde på flere tekster, flere værktøjer og mere manual opmærkning, således at der vil være langt flere muligheder for både undervisning og forskning.

Projektets deltagere

Forfatterne takker alle MULINCO-projektets øvrige deltagere for deres bidrag til opbygningen af platformen: Vibeke Appel, Henrik Gottlieb, Lina Henriksen, Hanne Jansen, Steen Jansen, Ole Jorn, Jens Erik Mogensen, Viggo Hjørnager Pedersen, Lene Schøsler, Erling Strudsholm.

Litteratur

Andersen Jens (2003): *Andersen en Biografi*, Bind I, Gyldendal, København.

Andreassen, Lotte (2005): *Leksikaliseringsmønstre og oversættelse. Bevægelseskonstruktioner i spansk og dansk*. BA-projekt spansk, Romansk Institut, Københavns Universitet.

Christensen, Kurt (2005): *Bevægelsesverber på spansk versus bevægelsesverber på dansk*. Emneopgave, spansk overbygningsuddannelse, Romansk Institut, Københavns Universitet.

Farø, K., L. Henriksen, H. Jansen, S. Jansen, X. Lepetit, B. Maegaard, C. Navarretta, L. Offersgaard, C. Povlsen (2005): *MULINCO Behovsanalyse, Rapport 1,* Københavns Universitet.

Farø, Ken (u. udarb.): *Idiomatizität – Ikonizität – Arbitrarität. Beitrag zu einer Theorie der Idiomäquivalenz.* København: Københavns Universitet [ph.d.-projekt].

Halteren, Hans van (Ed.) (1999): *Syntactic wordclass tagging.* Kluwer Academic Publishers, The Nederlands, 19-21.

Hansen, D.H. (2000): *Træning og brug af Brill-taggeren på danske tekster*. Ontoquery Teknisk Rapport, Center for Sprogteknologi, København.

Harder, P.,L. Heltoft & O.N. Thomsen, (1996). "Danish Directional Adverbs", in E. Engberg-Pedersen, L.F. Jakobsen & L.S. Rasmussen (eds), *Content, Expression and Structure: Studies in Danish Functional Grammar*, John Benjamins, Amsterdam,159 -198.;

Hartmann, R.R.K. (1980): *Contrastive Textology*. Heidelberg: Groos.
Heinemann, M. (2000): Textsorten des Alltags. I: Brinker, Klaus et al. (red.): *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*. Berlin/New York: de Gruyter.

Hermes. Journal of Linguistics, nr. 35 (2005).

Jongejan, B. og D. Haltrup Hansen (2005): *The CST Lemmatiser,* Technical Report, Center for Sprogteknologi, Københavns Universitet.

Lamiroy, B. (1983), *Les verbes de mouvement en français et en espagnol*, Amsterdam, John Benjamins Publishing Company and Leuven University Press.

Pedersen, B. S. (2000): Systematic Verb Polysemi in MT: A study of Danish Motion Verbs with Comparison to Spanish", in: Harold Somers (ed.) *Machine Translation Vol. 14*, Kluver Academic Press, The Netherlands, 35-82.

Pedersen, V. H. (1993): A Wonderful Story of a True Soldier and a Real Princess. Problems in Connection with the Rendition of Hans Andersen's Vocabulary in English, pp. 197-209 i Johan de Mylius, Aage Jørgensen & Viggo Hjørnager Pedersen (red.): *Andersen og Verden. Indlæg fra den første internationale H. C. Andersen-konference, 25.-31. august 1991,* Odense Universitetsforlag, Odense.

Pedersen, V. H. (1999): H. C. Andersen's Fairy Tales in Translation: Prepositions and 'Small Words, In: Johan de Mylius, Aage Jørgensen and Viggo Hjørnager Pedersen (ed.): *Hans Christian Andersen. A Poet in Time. Papers from the Second International Hans Christian Andersen Conference 29 July to 2 August 1996*, Odense Universitetsforlag, Odense.

Slobin, D.I. (1996a). "Two ways to travel: Verbs of Motion in English and Spanish", in: Shibatani and Thompson (eds.) *Grammatical Constructions: Their Forms and meanings*, Clarendon Press, Oxford, 195-219.

Slobin, D. (1997): Typology and rhetoric: Verbs of motion in English and Spanish. I: Masayoshi Shibatani and Sandra A. Thompson (eds.), *Grammatical constructions: their form and meaning*, Oxford: Oxford University Press, s. 195-219.

Steinberger R., B. Pouliquen, C. Ignat, A. Widiger, T. Erjavec (2006): The Acquis Communautaire multilingual parallel corpus and Eurovoc
http://wt.jrc.it/lt/Acquis/doc/README_Acquis-Communautaire-corpus_JRC.html

Talmy, L. (1975). "Semantics and Syntax of Motion", in: Kimbal (ed.): *Syntax and Semantics* vol. 4, Academic Press, London, 181-238.

Talmy, L. (2000): *Toward a cognitive semantics*. Vol.1 and 2. Cambridge, MA: MIT Press.

[1] IMS Corpus Workbench (CWB) er udviklet på Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench

[2] Der er også overvejet andre korpusplatforme: ParaConc, et kommercielt produkt som er en Microsoft Windows-applikation, se http://www.athel.com/para.html og CorpusSearch, java-applikation hvor inddata og uddata specificeres i filer, se http://corpussearch.sourceforge.net.

[3] Online korpusaflevering sker på: http://www.cst.dk/mulinco/php/corpaflev.php

[4] Oplysningerne om teksten gemmes i en særlig headerfil for hver tekstfil. Formatet i headerfilen er baseret på XCES-standarden, se også http://www.xml-ces.org. Der er dog foretaget enkelte ekstra tilføjelser til formatet i forhold til standarden. For en fuldstændig liste over de oplysninger der registreres for hver tekst, se http://cst.dk/mulinco/corpus/headerhelp.html

[5] Af andre sproglige analyser af H.C. Andersens værker i oversættelsesmæssigt perspektiv hvor sprogteknologi er inddraget, skal nævnes (Pedersen 1993 og 1999).

[6] Idet kun 17 af de 18 forekomster af et adverbium efterfulgt af et adjektiv var adjektivsyntagmer.