Peter Widell og Mette Kunøe (udg.):
10. Møde om Udforskningen af Dansk Sprog
Århus 2004

CorpusEye - Et brugervenligt web-interface for grammatisk opmærkede korpora

Af Eckhard Bick (Syddansk Universitet)

1. Indledning

I de senere år har teknologiske og datalingvistiske fremskridt gjort det muligt at kompilere og annotere stadigt større tekstsamlinger, til gavn for empirisk, korpus-baseret sprogforskning. For flere af de store sprog foreligger der i dag korpora med over 100 millioner ord, og internettet kan i princippet betragtes som eet stort, multilingualt korpus. For dansk er de største offentligt tilgængelige tekstkorpora Korpus90 og Korpus2000 (kompileret af DSL), med hver 26 millioner ord, mens der på talesprogsområdet findes fx BySoc (Henriksen 1998) og transskriberede parlamentsdiskussioner fra Folketinget og Europaparlamentet (Europarl). Imidlertid er brugsværdien af et korpus ikke kun afhængig af design-parametre som størrelse, genre, tidsperiode m.m., men også af eksistensen og kvaliteten af tilføjet grammatisk meta-information, samt tilgængeligheden og acceptansen blandt lingvistiske forsker, lærere, leksikografer og andre. Jeg har tidligere præsenteret et automatisk korpus-opmærkningssystem for dansk (DanGram, MUDS-8), samt et projekt til manuel lingvistisk revision af det opmærkede Korpus90/2000 (MUDS-9), og vil denne gang fokusere på det sidste aspekt - vejen fra korpus til bruger.

2. Et integreret, internetbaseret søgeinterface

Her er det afgørende, om der foreligger brugervenlige redskaber til korpussøgning - dvs. redskaber, der (a) ikke kræver køb og installering af specialiseret software, og (b) ikke forudsætter, at brugeren tilegner sig et korpusspecifikt kodesprog. En elegant løsning på (a) er internetbaseret korpusadgang, idet kompatibiliteten med brugerens komputersystem her sikres igennem browseren, - og hjemmesiderne af både BySoc og Korpus2000 er gode eksempler herpå. Begge systemer har dog visse begrænsninger. For det første er de tilpasset

til eet bestemt korpus og tillader ikke kombination og sammenligning med andre korpora i samme søgning og samme interface. For det andet har brugeren ingen mulighed for at anvende morfologiske eller syntaktiske kategorier i sin søgning.

CorpusEye-projektet (http://corp.hum.sdu.dk) på Syddansk Universitet er et forsøg på at designe og programmere et internetbaseret søgeinterface, der dels tilbyder eens redskaber og samme formalisme på tværs af flere korpus-typer og på tværs af flere sprog, dels tillader at udnytte den grammatiske information i opmærkede korpora på en brugervenlig og menubaseret måde.

3. Et stort og voksende korpusudvalg for flere sprog

Samtlige korpora i CorpusEye er blevet forsynet med morfologiske og syntaktiske tags vha. VISL's Constraint Grammar-baserede parsere, for træbankernes vedkommende med et efterfølgende PSG-modul (Bick 2003-1), der i stedet for ordformer benytter syntaktiske funktioner som terminaler i sine genskrivningsregler. På nuværende tidspunkt er der tale om følgende materiale:

- **Dansk:** 5 korpora, ca. 50 millioner ord (bl.a. Korpus90/2000, Skalk, Europarl[1], Folketingsdebatter og Arboretum-træbanken[2])
- **Portugisisk:** 4 korpora, ca. 250 millioner ord (bl.a. Público, Folha de São Paulo[3] og Floresta-Sintá(c)tica-træbanken)
- **Engelsk:** 4 korpora, ca. 120 millioner ord (herunder BNC, KEMPE[4] or Europarl)
- **Tysk:** 4 korpora, ca. 50 millioner ord (herunder MAK, BZK[5] og Europarl)

---

[1] *Europarl* er et stort, frit parallelkorpus med debatudskrifter fra Europaparlamentet, der dækker i alt 11 sprog med 20-30 millioner ord hver, fra perioden 1996-2003. Korpuset er oprindeligt kompileret af Philip Koehn.

[2] Arboretum er stadigt under opbygning, og indeholder nu reviderede analyser af ca. 15.000 sætninger (ca. 300.000 ord) i både Constraint Grammar- og træbank-format.

[3] *Público* og *Folha de São Paulo* er store dagblade i hhv. Portugal og Brasilien. Korpuserne er kompileret af Linguateca-projektet (v/ Diana Santos), og opmærket med forfatterens PALAVRAS-parser. Et uddrag fra begge tekstsamlinger underkastes løbende lingvistisk revision på træbank-niveau *(Floresta Sintá(c)tica).*

[4] *Kempe*, 'Korpus of Early Modern Playtexts in English', er kompileret af Lene B. Petersen and Marcus X. Dahl og opmærket i samarbejde med VISL. *BNC* (British National Corpus) indeholder både skrift- og talesprog.

[5] Både *MAK* (Mannheimer Korpus) og *BZK* (Bonner Zeitungskorpus), samt det spanske *El Diario Sur* og det franske *Le Monde*-Korpus stammer fra *European Corpus Initiative*, og er siden opmærket med forfatterens parsere, der for fransk og tysk få leveret morfologisk input fra hhv. Achim Stein og Helmut Schmid's DTT-tagger og Lingsofts GERCG.

- **Fransk:** 3 korpora, ca. 35 millioner ord (herunder Le Monde, Europarl og Arboratoire[6])
- **Spansk:** 3 korpora, ca. 30 millioner ord (herunder El Diario Sur og Europarl)
- **Esperanto:** 5 korpora, ca. 17 millioner ord (bl.a. Monato, Eventoj[7], klassisk litteratur)
- **Estisk:** 1 korpus, 3.500 ord (Arborest-træbanken[8])

4. Søgeformalisme og datastruktur

Den interne søge-database benytter IMS' *Corpus Query Protocol* (Christ 1994, http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/) for CG-korpora og linux-værktøjet tgrep2 (http://tedlab.mit.edu/~dr/Tgrep2/) for træbankerne. Internet-brugen af CQP er inspireret af andre søge-interfaces der tidligere er programmeret af hhv. Paul Meurer for norsk (Oslo Universitet) og Diana Santos for portugisisk (Linguateca). Selvom CorpusEye tillader direkte brug af CQP[9] samt såkaldte regular expressions[10] (joker-søgninger, sets og matematiske operatorer), henvender projektet sig primært til den humanistiske bruger uden formalistiske forkundskaber. Det er derfor muligt at komme i gang med simple tekstsøgninger, der præsenteres i konkordans-format. Med et enkelt klik kan brugeren producere statistiske, absolutte eller frekvensnormerede, oversigter over bestemte positioner i søgningen eller dens kontekst. Ønsker man at søge på fx leksemer

[6] *Arboratoire*-træbanken er en del af *Freebank*-initiativet, et samarbejde mellem VISL og ATILF/Loria (Susanne Salmon Alt, Nancy). Teksterne opmærkes med forfatterens FrAG-parser, efterfulgt af uddragsvis revision.

[7] *Monato* er et internationalt nyhedsmagasin, *Eventoj* er et internetbaseret nyhedsbrev.

[8] *Arborest* er blevet til i forbindelse med det NorFa-støttede Nordiske Træbank-Netværk, og bygger på CG-analysen af estisk avismateriale, der underkastes en efterfølgende PSG-analyse (revideret af Heli Uibo).

[9] Her formuleres søgekriterier for hvert token (ord) for sig, fx. [morph="PR AKT] [pos="N" & func="<SUBJ"] for to på hinanden følgende ord, det første (et verbum) i præsens aktiv, det andet et nomen (substantiv) med subjekt-funktion og venstrevendt dependens (<), som fx i .... *siger talskvinden, ... Her har hjorten ligget.*

[10] *Regular expressions* tillades både i rene tekst-søgestrenge, og i CQP-udtryk. De vigtigste operatorer er '?' (ingen eller een), '*' (ingen eller flere) og '+' (een eller flere), og kan tillægges bogstaver (fx 'korpus+er' = korpuser, korpusser), sæt (fx 'k[æøå]be[rn]?e?' = kæber, køben, kåbe, ...) og joker-tegn (fx. .* for en vilkårlig streng).

(bøjningsformsneutraliseret), ordklasse (fx. nominativ-substantiver efter "hendes") eller syntaktiske funktioner (fx. frontstillede objekter), kan dette gøres igennem kategorimenuer, der knytter sig til den enkelte søgeposition (det enkelte ord), der så - usynligt for brugeren - oversættes til CQP-udtryk og CG-tags. Tilsvarende er det muligt grafisk at knytte såkaldte operatorer (repetition, negation, optionalitet m.m.) til en given søgeposition. Knapperne er i høj grad selvforklarende igennem popup-vinduer, og indgangssiden byder på en introducerende flash-film, der guider brugeren igennem systemet.

X. TextPainter: Korpora "on the fly"

Mange grammatiske CALL-øvelser *(Computer Aided Language Learning)* fokuserer på eet, snævert emne ad gangen, såsom ordklasser, et bøjningsproblem eller kommatering af relativsætninger, og hvis læreren ikke kan finde en eksisterende øvelse der passer ind i undervisningsstoffet, vil der som regel ikke være mulighed for at ændre i eller tilpasse eksisterende CALL-øvelser. Problemet er særligt relevant i faget "Almen Sprogforståelse", der på relativ kort tid søger at dække en lang række emner og øge elevernes sproglige bevidsthed som sådan. Her vil strategien ofte være at lade eleven selv "opdage", hvilke karakteristika, distribution og brugsregler der knytter sig til bestem grammatisk kategori. Et brugervenligt korpusinterface kan hjælpe eleven at finde relevante eksempelsætninger og ændre i søgningerne på en fleksibel og inkrementel måde.

I forbindelse med URKAS- og VISL-SEM-projekterne har forfatteren forsøgt at integrere korpusopmærkning, "text grading" og grammatiske øvelser i et nyt redskab, *TextPainter*[11], der tillader emne/kategori-specifik opmærkning af brugertekster, der løbende underkastes en automatisk grammatisk analyse. *TextPainter* accepterer således cut-and-paste-tekst på 7 sprog, og fremhæver ord med en ønsket grammatisk kategori eller kategorikombination, fx *subjekter, objekter, verber* eller *prædikativt brugte adjektiver*.

En overordnet øvelse kan bestå i genrebestemmelsen af en tekst: Ved fx at farve verber rød og adjektiver blå, vil man kunne skelne mellem en mere handlingspræget action-fortælling og en mere deskriptiv landskabsskildring.

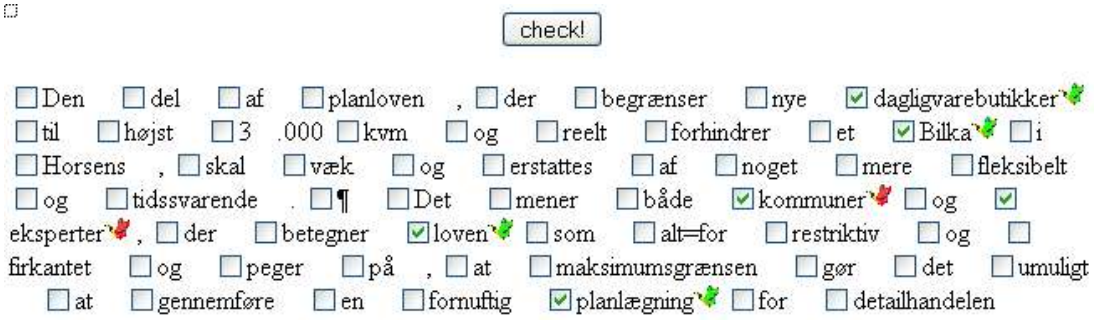
[11] http://beta.visl.sdu.dk/visl2/texttyping.htm

For at opnå en robust analyse og en lav fejlprocent, arbejdes der videst muligt med regel-baserrede Constraint Grammar[12] parsere, og al grammatisk information markeres på ordniveau. Komplekse syntaktiske funktioner repræsenteres således på konstituentens kerneled, i en dependensgrammatisk tradition. Ledsætningsfunktion, som fx "relativsætning" (@CL-N<), knyttes således til den pågældende sætnings første verbum:

categories: @CL-N< ... *OR* ... NONE

Den del af planloven , der begrænser nye dagligvarebutikker til højst 3.000 kvm og reelt forhindrer et Bilka i Horsens , skal væk og erstattes af noget mere fleksibelt og tidssvarende . ¶ Det mener både kommuner og eksperter , der betegner loven som alt=for restriktiv og firkantet og peger på , at maksimumsgrænsen gør det umuligt at gennemføre en fornuftig planlægning for detailhandelen

I interaktiv modus skal brugeren selv *finde* alle ord med en bestemt kategori, fx. *direkte objekt*. Feedback gives i form a røde og grønne Grammy-bævere, og performansen evalueres med udgangspunkt i en vægtning af falsk positive og falsk negative svar, den såkaldte F-score.



All in all, there were 60 words in the text.
For the category/categories in question, you found 4 out of 5 possible words, and had 2 false positives. This equals a recall of 80 % and a precision of 66.66 %, combining into an **F-score of 72.72 %**. To compare your own with the computer's opinion, check the highlights below!

Den del af planloven , der begrænser nye dagligvarebutikker til højst 3.000 kvm og reelt forhindrer et Bilka i Horsens , skal væk og erstattes af noget mere fleksibelt og tidssvarende . ¶ Det mener både kommuner og eksperter , der betegner loven som alt=for restriktiv og firkantet og peger på , at maksimumsgrænsen gør det umuligt at gennemføre en fornuftig planlægning for detailhandelen

X. Perspektiv

Selvom udviklingen på ingen måde er afsluttet, er systemet på nuværende tidspunkt fuldt funktionelt, og der afholdes introducerende workshops ved ISK. Der planlægges tilbud om opmærkning og tilgængeliggørelse af brugerens egne tekstsamlinger, samt lancering af joint ventures til linguistisk revision af benchmark-korpora.

[12] http://beta.visl.sdu.dk/visl2/constraint_grammar.html

Litteratur

Bick, Eckhard (2003-1), A CG & PSG Hybrid Approach to Automatic Corpus Annotation, In: Kiril Simow & Petya Osenova (eds.), "Proceedings of SProLaC2003" (at Corpus Linguistics 2003, Lancaster), pp. 1-12

Bick, Eckhard (2003-2). "Morfosyntaktisk opmærkede korpora for dansk". I: Peter Widell & Mette Kunøe (udg.), "9. Møde om Udforskningen af Dansk Sprog", pp. 43-54. Århus Universitet.

Oli Christ (1994). "A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest. .ps.gz

Henrichsen, P.J. (1998). Peeking Into the Danish Living Room – Internet access to a large speech corpus, 11th NODALIDA pp.109-119