

Sprogteknologiske redskaber til undervisningsbrug:

CALL og faget almen sprogforståelse

Eckhard Bick

University of Southern Denmark

1. Baggrund

Sprogteknologi udvikles i det frugtbare grænseområde mellem humanistisk-lingvistisk og teknologisk forskning, med de finansieringsproblemer og paradigmekonflikter, en sådan forskningsmæssig hjemløshed medfører (for en nærmere diskussion se Mægaard et.al. 2004). Men sprogteknologi er i kraft af sin dobbeltposition også prædestineret til at slå bro mellem menneske og maskine, mellem universitet og samfund, og et af de steder hvor dette potentiale bliver tydeligst, er *computer-assisted language learning* (CALL) - et forskningsområde, der dels udvikler og evaluerer edb-baserede undervisningsredskaber, dels søger at udnytte og tilgængeliggøre generel sprogteknologi (fx sproganalyseredskaber) og generelle sprogdata (fx korpora og leksika) i undervisningsøjemed. I Danmark har VISL-projektet ved Syddansk Universitet (<http://visl.sdu.dk>), med forfatteren som projektleder, siden 1996 arbejdet med sprogteknologi i almindelighed og CALL i særdeleshed.

VISL er et tværsprogligt forskningsprojekt, der satser på en parallel udvikling af på den ene side almene sprogteknologiske redskaber (taggere, parsere, korpora) og på den anden side CALL¹-baseret grammatikundervisning (internet-applikationer, evaluerings-redskaber, tekst-databaser). Med udgangspunkt i et énsartet begrebsapparat for sproglig analyse (Dienhart 2000 og Bick 2002) på tværs af sproggrænser og undervisningsniveauer har VISL altid forsøgt at integrere sine NLP²-værktøjer i et bredt vifte af undervisningsprogrammer (Bick 1997, 2004). Materialet er frit tilgængeligt via Internet, men distribueres også på cd-rom. Nærværende artikel præsenterer konkrete tiltag i forbindelse med det nye gymnasiums-fag, *Almen Sprogforståelse*, og i særdeleshed integrationen af korpora og korpus-værktøjer i en didaktik, der sigter på øget sproglig bevidsthed og en empirisk-afkodende tilgang til sprogforståelse.

2. Almen Sprogforståelse

Faget Almen Sprogforståelse undervises i 1.g med typisk 48-60 lektioner i løbet af et halvt år, og pensum omfatter, foruden en introduktion til Latin og europæisk sproghistorie, sociolingvistik, psykolingvistik, sproglige læringsteknikker og ikke mindst kernegrammatikken, herunder sætningsanalyse. Det åbenlyse politisk-didaktiske dilemma ligger så dels i, at de færreste lærere videnskabsmæssigt er rustet til et så stort spektrum af emner, dels i, at det med almindelige undervisningsmetoder forekommer umuligt at gennemgå stoffet i nogen dybde i den tid der er til rådighed. Førstnævnte kan søges løst gennem videreuddannelseskurser, mens sidstnævnte kræver en eklektisk tilgang, hvor den enkelte lærer fokuserer på udvalgte emner. Med økonomisk støtte fra ELU³ lancerede VISL derfor i 2003 det såkaldte URKAS-projekt (<http://visl.sdu.dk/urkas.html>), en kombination af udvikling af CALL-materiale til faget og et kursustilbud til gymnasielærere, hvor lærerne inspireres til en integration af lingvistik og CALL. Kongstanken ved projektet var og er at den CALL-baserede metode

¹ Computer-Assisted Language Learning

² Natural Language Parsing

³ Efteruddannelsesudvalg for Længerevarende Uddannede

- optimerer udnyttelsen af undervisningstiden, idet deltagerne parallelt interagerer med en computer, og ikke behøver at kappes om interaktionstid med en enkelt lærer
- tillader undervisningsdifferentiering både emnemæssigt og på det individuelle plan
- tillader, ved hjælp af korpusarbejde, en empirisk-afkodende tilgang til sproget
- som nyt medie tilfører undervisningen et element af spil og adspredelse
- tillader evaluering og kvantificering, dels af elevfærdigheder, dels af sproglige forhold

Herudover gør satsningen på et tværsporgligt begrebsapparat det muligt at udnytte en synergieffekt på tværs af sprogfagene og undervisningsårene, således at viden fra og til faget Almen Sprogforståelse kan udbygges og udnyttes fortløbende også i timer, der primært sigter på fremmedsprogsundervisning.

3. Sprogspiralen

For at ordne og didaktisk tilgængeliggøre eksisterende og nyudviklede VISL-CALL-redskaber til faget Almen Sprogforståelse, og for at tilgodese et behov for en anbefalet emne- og metodeprogression, blev der i løbet af projektet udviklet et specielt indgangsportal, Sprogspiralen (http://visl.sdu.dk/urkas_pealrs.html).

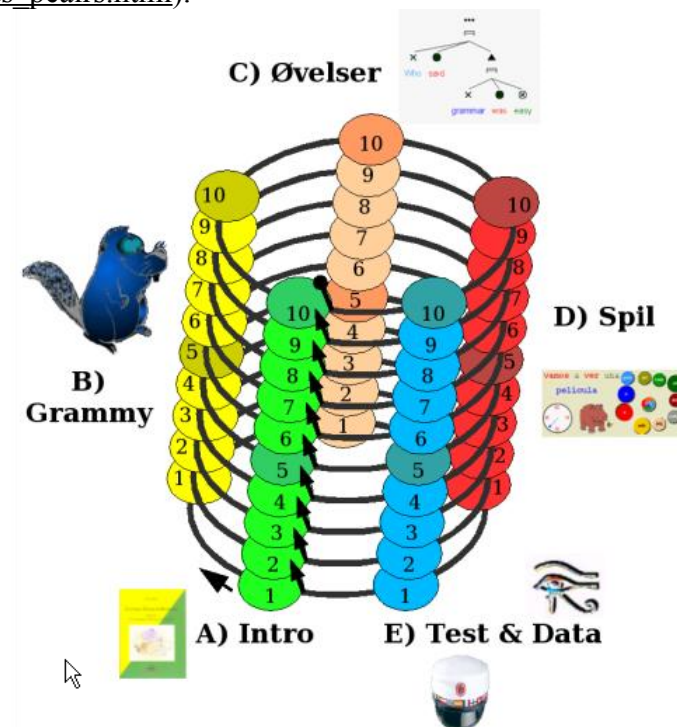


Fig. 1: Sprogspiralen

Brugeren bliver ledt igennem en række emner med stigende sværhedsgrad. For hvert emne inddrages fem elementer: (a - grøn) en introducerende, teoretisk tekst, (b - gul) et afsnit af *Grammy*-historien (Bick 2002), (c - brun) en eller flere øvelser (d - rød) et eller flere relevante VISL-spil og (e - blå) korpus-baserede empiriske øvelser og supplerende materiale. (a) - (e) gennemgås successivt i denne rækkefølge for en progression af emner (de 10 niveauer i spiralen), fra f.eks. ordklasser som *substantiv* og *verbum* over funktioner som *subjekt* og *objekt* til strukturelle kategorier som *ledsætninger* og *sideordning*. Orienteringen i systemet sker vha. ikonbaserede navigationsbjælker, der linker direkte ned til de relevante sider/programmer på VISL's website. Det var desuden fra de deltagende amtscentralers side intentionen, at en skole/lærer ville kunne

integrere portalen eller dele heraf i et LMS-system med supplerende materiale, uddybende forklaringer samt valgfrie links, således at et konkret kursusforløb ville kunne fremstå som et strømlinet forløb, mens det på den anden side rummer dybere videnslag til dem, der måtte have særlige interesser eller behov.

2. KillerFiller: Sætningskorpora som indsætningsøvelser

Blandt de mest kendte CALL-applikationer er indsætningsøvelser, bl.a. fordi øvelserne er nemme at evaluere og egner sig til screening ved holdopdeling, afdækning af individuelle svagheder m.m. CALL-øvelserne er dog som regel ikke andet end (datamæssigt statiske) genoplæg af tilsvarende øvelser fra lærebøger og testark, eller produceret én for én med et snævert formål eller et bestemt kursusforløb i sigte, ligesom evaluering foretages for det meste af den enkelte lærer.

Et nyt VISL-redskab, KillerFiller, er et forsøg på dynamisk at generere indsætningsøvelser ud fra opmærkede korpora, og tilbyder en systematisk statistisk og komparativ elevevaluering. Ved at registrere brugeren og føre "logbog" over resultaterne fra forskellige indsætningsøvelser, kan KillerFiller på én gang tilgodese:

(a) forskerens ønske om evaluering af CALL-redskabet selv (VISL er sjov, men gør det en forskel?)

(b) lærerens behov for feed-back: Viser en given elev fremskridt på et bestemt grammatisk område eller ikke? Er der muligt at observere tendenser og problemer for holdet som helhed?

Fig. 2a: Brugerevaluering Please login to your VISL-game account If you do not have an account, create a new one by clicking here! Username learner Password Login	Fig. 2b: tværsproglighed Which language do you want to train?  	Fig. 2c: brugerstyrede øvelser og datasæt Sentence collection Grammy 1 Word class v Show sentence 
---	---	--

Programmet som sådan er et IT-baseret pandæmonium af indsætningsøvelser, hvor der udtrækkes randomiserede, grammatisk opmærkede sætninger ikke kun fra VISL's egentlige undervisningskorpora (p.t. 25 forskellige sprog), men også fra VISL's forskningskorpora (7 sprog, Bick 2003). KillerFiller erstatter ord af en af brugeren defineret ordklasse (fx aller præpositioner eller alle verber) med tomme felter, der så udfyldes af den studerende eller elev.

Fig. 3: korpus-afledte indsætningsøvelser



Kasparow zu (besiegen) (müssen -pr-) für den
Computer ein Genuß (sein) (sein)

Alt efter hvilket sprog der vælges, oplyses der for visse ordklasser grundform og/eller bøjningskategori (fx tempus eller numerus) for at undgå flertydigheder i de jo ellers kontekstløse sætninger, som det fremgår af ovenstående, tyske eksempel (Fig. 3). Indtil videre anvendes

udelukkende manuelt kontrollerede data, men takket være den meget lave ordklasse-fejlprocent af modne Constraint Grammar-taggere vil det i fremtiden være muligt - i det mindste for VISL's hovedsprog⁴ - at generere indsætningsøvelser også udfra ad hoc-tekster der leveres eller produceres af den enkelte lærer selv.

En server-baseret og password-beskyttet database registrerer resultaterne for hver gang, en øvelse udføres, sorteret efter sprog og øvelsestype. Efter hver runde (defineret som 10 sætninger) vises en graf over præstationen og fremskridtet. Lærere og evaluators har adgang til historiske oversigtssider over relevante sprog-emne-konstellationer. Lærerne kan oprette sprog- eller niveauspecifikke grupper, og eleverne tilknyttes en eller flere grupper, hvorved det bliver muligt at vurdere den gennemsnitlige forbedring (eller forbedringsrate) i fx et 2.g's fransk-hold hhv. før og efter et VISL-baseret grammatikkursus.

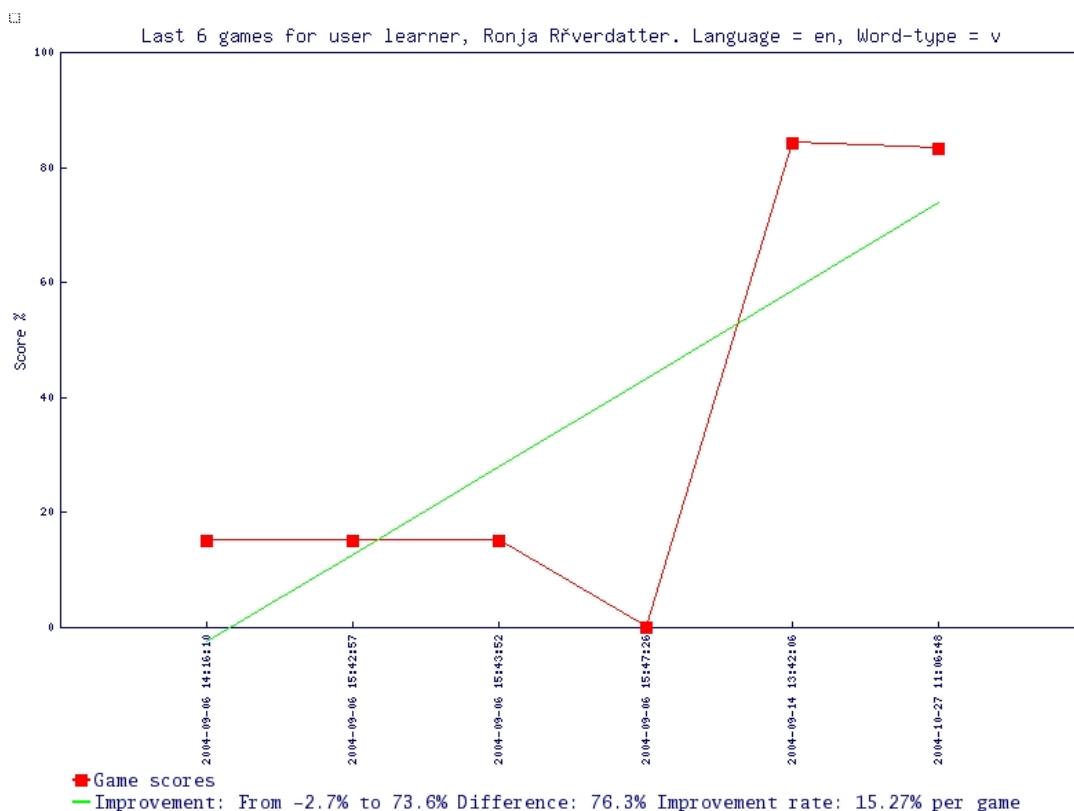


Fig. 4: *KillerFiller*-statistik

3. *TextPainter*: Grammatisk evaluering af brugertekster

Mange grammatiske CALL-øvelser, bl.a. i kommercielle cd-systemer, fokuserer på ét grammatisk emne eller ét grammatisk træk ad gangen (fx genitiv-s, relativsætninger, objektspronominer), og hvis en lærer ikke umiddelbart kan finde det emne han søger, vil han som regel ikke have mulighed for at tilpasse eller ændre en eksisterende øvelse. På den anden side vil et fuldt sætningsanalyse-modul som VISL's *tree builder* være for kompleks og for lidt fokuseret på et bestemt undervisningsdelmål eller et bestemt stadie i et undervisningsforløb. Problemet er særligt nærværende, hvor formålet er at opbygge sproglig bevidsthed og at få eleverne til selv at opdage distributionen af et

⁴ Der er tale om CG-systemer (http://visl.sdu.dk/constraint_grammar.html), der foretager en fuld automatisk sætningsanalyse: DanGram for dansk, ENGCG for engelsk, GerGram for tysk, FrAG for fransk, samt PALAVRAS og PALAVRAS-HIS for portugisisk og spansk. Systemerne har typisk en ordklasse-korrekthedsprocent på 99%, og 95-96% for syntaktiske funktioner.

givent grammatisk træk, dens formelle karakteristika og regler for anvendelse. Her kan *korpuslingvistikken* tilbyde lovende redskaber og nem adgang til fleksible eksempler, såkaldte *konkordanser* (jf. senere kapitel).

I en nyskabende tilgang har forfatteren integreret korpusopmærkning, tekstevaluering og grammatiske øvelser i ét enkelt redskab, *TextPainter*, der muliggør opmærkning af brugertekster for bestemte træk og karakteristika ”on the fly”. Programmet leverer således live analyser for almindelig klippebordstekst på syv sprog. Resultaterne kan farvemærkes for en bestemt kategori, eller kombination af kategorier, såsom objekt, subjekt, verber eller adjektiver. En mulig anvendelse er fx, at typologisere en novelle som en verbs-tung action-tekst, eller som adjektiv-præget beskrivende, malende.

Fig. 5: *TextPainter*

Robusticitet, dvs. fejlfrie eller i det mindste fejlfattige analyser (ikke kun af bestemte sætningsstrukturer, men af hele spektret af muligheder i et naturligt sprog), er afgørende i et analyseredskab, hvor der ikke er mulighed for hverken korrekturlæsning eller restriktioner hvad angår ordforråd og teksttype. Analyserne håndteres derfor vha. Constraint Grammar-parsere, hvor al grammatisk information markeres vha. ordbaserede *tags* (“etiketter”), og hvor komplekse syntaktiske funktioner markeres som dependenser mellem syntagmekerner. Relativsætninger, for eksempel, markeres med et tag på det første finite verbum i ledsætningen:

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

Fig 6: *farvemærkning af gramatiske kategorier*

Ordbaserede tags er nemme at ekscerpere, søge og filtrere med simple datalingvistiske edb-kommandoer, og vises i *TextPainter* ved at farvelægge det pågældende ord⁵. De samme tags forbliver usynlige i interaktiv modus, hvor brugeren sættes til at finde, for eksempel, samtlige direkte objekter i en tekst. Afkrydsningen krydscheckes mod computerens beregnede tags, og feedback gives i form af røde (= forkert) og grønne (= korrekt) bæversymboler (Fig. 7). Endelig beregnes et tal for den procentvise andel af falsk positive og falsk negative svar, den såkaldte *F-*

⁵ Ved successiv opmærkning af 2 forskellige kategorier benyttes 2 forskellige farver, således at det er muligt, fx at undervise/synliggøre forskellen på nominelle subjekter på den ene side og pronominaler på den anden.

score⁶ (Fig. 8).

☐ 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

Fig. 7: Interaktiv version af TextPainter

☐ 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

Fig 8: Bruger-feedback i TextPainter

4. Sproglig bevidsthed: Den korpusbaserede metode

Sproglig bevidsthed er blevet et nøgleord i den danske gymnasieskole, og har fået tildelt en central rolle i studieplanen for faget *Almen Sprogforståelse*:

*En nyskabelse i gymnasireformen er et samarbejde mellem de almene gymnasiers sprogfag i forløbet Almen Sprogforståelse. Som en del af den nye gymnasieuddannelses grundforløb skal Almen Sprogforståelse styrke elevernes teoretiske sprogforståelse, **samspillet mellem sprogene** og studiekompetencen ... Forløbet Almen Sprogforståelse vil indgå i grundforløbet for nye elever på det almene gymnasium fra sommeren 2005. Formålet er at **vække og styrke elevernes sproglige viden, bevidsthed og opmærksomhed** ... (Undervisningsministeriet - Nyhedsbrev nr. 9 - 2004)*

I noget højere grad end det er tilfælde i traditionel sprogundervisning, lægger disse målsætninger op til en empirisk metode og brugen af autentiske sprogdata - med andre ord, et for skolerne tilgængelig korpus-søgeinterface. VISL har igennem flere år udviklet et sådant søgeinterface, *Corpus Eye* (<http://corp.hum.sdu.dk> - fig. 9), der på den ene side tillader simple tekstsøgninger og ordstatistikker, men på den anden side åbner for komplekse, kategori-og kontekstbaserede søgninger, hvor brugeren måtte have behov for det. Tanken er, at man ikke behøver at sætte sig ind i en bestemt deskriptiv eller søgnings-formalisme for at kunne benytte interfacet, men på den anden side efterhånden kan få menubaserede valg og forklaringer til at udføre mere komplekse søgninger. Som resten af VISL's værktøjer er også korpussøgeinterfacet tværsproglig udlagt, og understøtter således, som en sidegevinst, Undervisningsministeriets sekundære mål, *samspillet mellem sprogene* (jf. citat). Korpussiden har derfor spillet en fremtrædende funktion også i VISL's forømtalte videreuddannelseskurser for gymnasielærere. I det følgende gives en række didaktiske eksempler

⁶ F-Score-målinger benyttes også i evalueringen af sprogteknologiske redskaber i almindelighed, såsom taggere og parsere. Tallet beregnes som et vægtet gennemsnit ud fra *recall* (hvor mange af alle mulige rigtige svar blev fundet?) og *precision* (hvor mange af de faktisk givne svar var rigtige?).

på hvordan korpus søgning kan bruges til at realisere målene med *Almen Sprogforståelse*.



Guided tour

Fig. 9: *Corpus Eye*, VISL's søgegrænseflade for opmærkede korpora

4.1 Dyremetaforer

Distinktionen mellem kerne-semantisk og metaforisk brug af et givent ord har betydning både i lyset af (bevidst) stilistisk variation (sproglig bevidsthed) og som et potentielt oversættelsesproblem i et tværspørgsmål perspektiv.

En mulig øvelse på dette område tager udgangspunkt i et af de mere frekvente metafortyper, dyremetaforer: Eleverne bedes her om at undersøge, i hvilken grad dyr-relaterede ord anvendes metaforisk i moderne dansk avissprog. Fx kan man udnytte VISL's opmærkning af ordklasser og semantiske substantivklasser til at formulere en søgning⁷, der fokuserer på adjektiver efterfulgt af dyre-substantiv. Kategorien *adjektiv* kan - som ordklasse - vælges i et menu, mens den semantiske - sekundære - kategori, *dyr*, eller her, "*jord-dyr*" (*Azo*), skrives ind i et åbent søgefelt.

Fig. 10: *Menubaseret korpus søgning*

Ca. halvdelen af af alle forekomster i Korpus90/2000 viser sig faktisk at være metaforiske:

Et spil der appellerer til den **politiske ræv**, fordi det kan spilles uden altid at tænke i strategi
 havde juryen indført en ekstra jurypris med **tilhørende sølvbjørn** til den tyskfødte instruktør Wim_Wenders
 ikke kan få den **russiske bjørn** til at gungre med
 Gys, gru og store **stygge ulv**.

⁷ Bag kulisserne benytter søgemaskinen sig af en *Corpus Query Processor* (CQP), udviklet ved Institut für Maschinelle Sprachenverarbeitung, Stuttgart (Christ 2004, <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>). CQP har sin egen formelle søgesyntaks, der også kan benyttes direkte som et mere fleksibelt, men mindre brugervenligt alternativ til den menubaserede søgning.

styrken ved at være i kontakt med den **indre abe** er i hvert_fald til at få øje på.
 og prøve at følge med de **unge løver**, der vil køre stærkt.
 Kun de allerbedste er i_stand til at tæmme den **olympiske vildhest** 49er.
 de **unge skakløvers** forslag blev fulgt af flertallet
 til forveksling kan ligne dem fra de mest ideologiske **unge løver** i Venstre

Som næste trin vil det være oplagt at sammenligne de fundne dyreudtryk med deres ækvivalenter på andre undervisningssprog, og herigennem synliggøre metaforicitet som sådan eller mulige tværsproglige "universalismer" og variationer i hvilke adjektiver der associeres med hvilke dyr. I det tyske ECI-corpus⁸, for eksempel, kan man vha. VISL's korpusinterface sandsynliggøre ikke alene en tilsvarende frekvens, men også et lignende kernespektrum af metaforisk betydning (jf. nedenfor). Eleven kan her - på en kvantitativt noget uvidenskabelig, men absolut diskussionsværdig og sprogbevidstheds-skabende måde - argumentere for, at dansk deler visse dyremetaforer med tysk. Imidlertid har ikke alle eksempler så transparente oversættelses-ækvivalenter som *snu ræv* og *stærk bjørn*, og hvis øvelsen gennemføres i fremmedsprogsundervisning snarere end i faget Almen Sprogforståelse, vil det være nærliggende, at tage udfordringen op med oversættelser til de danske *unge løver* eller til den tyske *party-bjørn* (2. eksempel).

Der **Berliner Bär** lächelnte nur einen Tag
 Ansonsten war dort der **sprichwörtliche Bär** los: Jubel, Trubel, Heiterkeit
 Damit hat sich der **schlaue Fuchs** das Monopol der Regierung für die Veranstaltung von ...
 Gies, der **alte Fuchs**, steigt in den AGF-Verwaltungsrat auf
 48 Stunden vor der Oberhauswahl berief der **finanzpolitische Fuchs** für Freitagabende eine ...
 den Herkules aus Pennsylvania beendete der **schwarzhaarige Bär** aus Teheran
 Und auch die türkischen Neofaschisten um den **grauen Wolf** Alparslan Türkeş, ...
 Sollte sich einmal mehr Döring als **cleverer Fuchs** erwiesen und ...
 schliesslich muss er eigentlich den **einsamen Wolf** mimen ...

Mens metafor-øvelsen inviterer den studerende til selv og interaktivt at surfe rundt i korpusbrugerfladen, er det også muligt at konstruere færdigstøbte sprogbevidsthedsøvelser ud fra korpusdata, som i det følgende eksempel, hvor en lærer først ekscerperer, vha. interfacet, den venstre attributive kontekst af 3 nær-synonyme ord (her: *indvandrere*, *udlænding*, *flygtning*) og efterfølgende bruger de færdige lister eller statistikker i undervisningen. Idéen i det konkrete didaktiske eksempel er at lade eleverne gætter/argumentere for, hvilken adjektivliste bedst passer til hvilket af de tre synonyme.

For at compilere adjektivlisterne bruges lemma-feltet, der grupperer fx *udlænding*, *udlændinge*, *udlændingen*, *udlændingene* etc. i én søgning, og de resulterende konkordanser blev frekvenssorteret efter *relativ frekvens*, hvor frekvensen i *konteksten* divideres med den såkaldte *leksikalske frekvens* (i sproget som sådan).

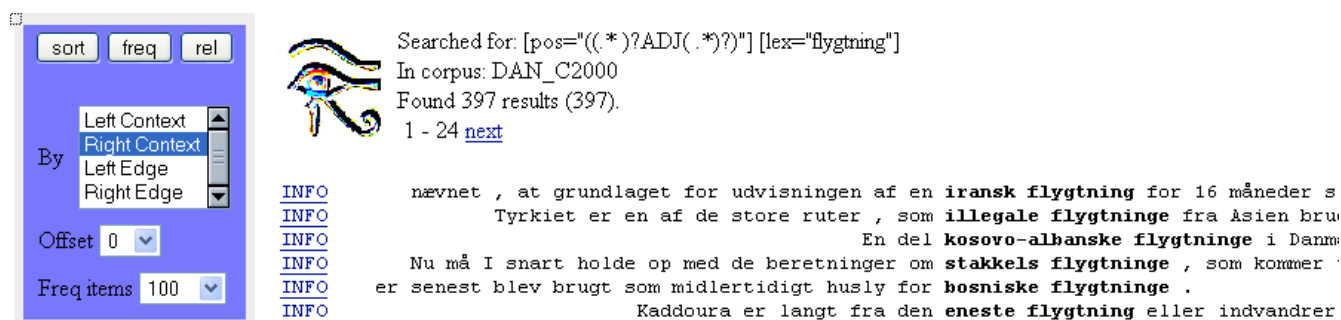


Fig. 11: Konkordanser og kontekststatistikker

⁸ European Corpus Initiative, data kompileret og udgivet af EACL (European chapter of the Association of Computational Linguistics)

I en sammenlignende tabel bliver det så åbenlyst for elevens spirende empiriske sprogbevidsthed, at dansk klassificerer *flygtninge* ud fra deres nationalitet og *immigranter* ud fra (il)legalitet og alder, mens *udlændinge* bedømmes efter deres eventuelle kriminalitet og uddannelsesmæssige baggrund.

Fig 12a: <i>indvandrere/e/ne</i> DAN_C2000 (315)	Fig. 12b: <i>udlænding/e/ne</i> DAN_C2000 (97)	Fig. 12c: <i>flygtning/e/ne</i> DAN_C2000 (397)
<i>frequencies:</i> <i>rel freq num</i>	<i>frequencies:</i> <i>rel freq num</i>	<i>frequencies:</i> <i>rel freq num</i>
illegale 4941110 18.7 [59]	narkodømte 9565309 3 [3]	somaliske 2052303 7.3 [29]
unge 238271 29.2 [92]	velkvalificerede 1366472 3 [3]	bosniske 1705722 8.5 [34]
nordafrikanske 113378 0.9 [3]	herboende 781479 5.1 [5]	kosovo-albanske 1015170 1 [4]
illegal 100781 1.5 [5]	kriminelle 571063 12.3 [12]	rwandiske 518159 1.7 [7]
højtuddannede 41228 0.9 [3]	uønskede 298333 4.1 [4]	palæstinensiske 407880 7.5 [30]
kriminelle 30459 2.8 [9]	voksne 12961 3 [3]	hjemvendte 333937 2.5 [10]
uønskede 28289 1.2 [4]	sådanne 12753 3 [3]	Kosovo-albanske 285516 0.7 [3]
muslimske 11944 1.2 [4]	unge 10687 6.1 [6]	illegale 258260 4.2 [17]
arbejdsløse 10586 1.5 [5]	dollarstærke 10628.12 1 [1]	albanske 238863 4 [16]
tyrkiske 9655 1.2 [4]	narkodømt 10628.12 1 [1]	kosovaalbanske 142758 0.7 [3]
ikke-franske 8062.48 0.6 [2]	Tossede 10628.12 1 [1]	tibetanske 79310 1.2 [5]
Veluddannede 8062.48 0.6 [2]	danskfødt 10628.12 1 [1]	nyankomne 78090 1 [4]

Fig. 12: *Synonymdifferentiering vha. kontekstkorrelationer*

I stedet for et sociolingvistisk fokus kan man konstruere lignende korpusbaserede korrelationsøvelser med det formål at undervise/synliggøre forskelle i brugen af semantisk beslægtede ordpar som *høj* - *stor*, eller *stærk* - *kraftig*. I faget *Dansk for Udlændinge* vil de relevante korpuseksempler måske kun blive brugt til simple ja/nej-multiple choice-øvelser, mens en lærer på gymnasie- eller universitetsniveau måske vil bruge de samme data til at øge opmærksomheden for lingvistiske kategorier, såsom *selektionsrestriktioner* og *semantiske kategorier*. Således sandsynliggør de anførte eksempeldata, at *høj* måler virkeligheden vha. af en akse a diskrete *enheder* (meter, kroner, år ...), mens *stor* forekommer mere beskrivende end målende, idet det lægger sig til *objekter* og *ikke-tællelige abstrakta*. Absolutte frekvenser fra Korpus2000 er anført i parentes:

høj/t ... grad (434), kvalitet (148), niveau (141), pris (57), prioritet (47), tempo (42), fart (40), indhold (36), humør (35), alder (34), jurs (33), hastighed (31), klasse (28), arbejdsløshed (24)

stor/t ... del (1214), betydning (432), succes (297), forskel (267), antal (245), interesse (205), vægt (201), problem (196), flertal (156), indflydelse (154), gruppe (142), rolle (141), glæde (141)

Hvad angår det andet eksempelpar, synes animater (mennesker, organisationer, dyr) prototypisk at korrelere med adjektivet *stærk*, mens handlinger og hændelser korrelerer med *kraftig*. Bemærk i øvrigt den metaforiske "humanicitet" af *position*, *ønske* og *økonomi*.

stærk/t ... hold (28), pres (22), mand (21), vilje (15), position (15), ønske (15), vækst (15), kontrast (14), modstander (14), leder (14), økonomi (13)

kraftig/t ... stigning (31), vækst (30), kritik (22), jordskælv (13), afstand (9), advarsel (9), pres (8), opfordring (8), forbedring (8), fald (8), slag (7), mistanke (7)

Ordet *vækst*, der ellers forekommer i begge lister, viser, at *relative* frekvenser kan være nyttige. I en relativ frekvens-sorteret søgning på ADJ + *vækst* er der således stor forskel på de 2 adjektiver, idet *kraftig* havner på plads 2, mens *stærk* degraderes til plads 11.

5. Propædeutisk korpusbrug: På vej til et dannet komma

I den senere tid har der været en betydelig interesse i at anvende VISL's CALL-værktøjer for indirekte formål - enten ved at kæde udvalgte eksisterende værktøjer og spil sammen med supplerende tekster og link-strukturer i helstøbte kursusforløb, eller ved eklektisk at bruge dem for at introducere og træne grammatiske kategorier og analytiske færdigheder, der måtte være instrumentelle for det egentlige, senere undervisningsemne. Således kan viden om visse sætningsafgrænsende ordklasser (konjunktioner, interrogativer, relative), samt analytiske færdigheder vedrørende subjekt-prædikat-strukturen betragtes som en forberedende instrumentalisering for undervisningen af ledsætnings-komma (Dansk Sprognævn 2004). Denne form for "propædeutisk viden" kan gøres eksplicit med bl.a. følgende VISL-baserede CALL-øvelser:

- ekstrahere korpus-sætninger med ordet "at", og lade eleverne vælge mellem konjunktions-'at' (med komma) og infinitiv-at (uden komma) ved kun at klikke på fx sidstnævnte.
- benytte en KillerFiller-indsætningsøvelse for at gøre eleverne fortrolige med ordklassen konjunktion.
- sammenligne korpuseksempler af sekvensen *subjekt* - "*ikke*" - *verbum* med et andet sæt eksempler på *subjekt* - *verbum* - "*ikke*". For dansk vil førstnævnte markere en ledsætning (med obligatorisk kommer efter, og fakultativ komma før sætningen). En elaborerende sprogbevidsthedsøvelse kunne bestå i empirisk at afgrænse hvilke andre adverbier der på dansk typisk optræder mellem subjekt og verbum (prædikator). Øvelsen kan udføres vha. det forbeskrevne menubaserede interface i *CorpusEye*.
- lære at skelne mellem de forskellige betydninger af ordet "*som*" - komparativ "præpositionel konjunktion" (*stærk som en bjørn*), komparativ "subordinerende konjunktion" (*som jeg tidligere har sagt*) og relativt pronomen (*som ikke kan vente*), hvor kun de sidste to typer medfører (sætnings-) komma. Eksemplerne kan hentes fra og visualiseres i VISL's syntaktiske undervisningskorpora.
- benytte et træbankkorpus som Arboretum (<http://beta.visl.sdu.dk/visl2/treebanks.html>) for at ekstrahere (a) adverbielle ledsætninger (fA:fcl)⁹ og (b) relative ledsætninger (DN:fcl), hvor førstnævnte altid medfører sætningskomma, mens sidstnævnte er "komma-ambiguøs", alt efter om der er tale om parentetiske ledsætninger eller ej.

De sidste to punkter beskæftiger sig med forholdsvis komplekse strukturelle forhold, og det kan være en pædagogisk fordel at benytte en grafisk fremstilling af den involverede syntaks. Faktisk tillader både VISL's didaktiske sætningssamlinger (25 sprog) og de store egentlige træbanker (7 sprog) først at finde alle sætninger med et givent træk (som fx "DN:fcl", Fig. 12), herefter grafisk at visualisere den enkelte sætning, og endelig interaktivt at inspicere (og genopbygge) den (Fig. 13).

⁹ Søgestrengene benytter her VISL's form- & funktion-konvention, med en syntaktisk kategori (majuskler) og en ordklasse-kategori (minuskler) adskilt af et kolon. Herudover kan en søgning indeholde tekststreng (ordformer eller leksemer) eller dele heraf, såkaldte regulære udtryk, samt tgrep2-relationer mellem sætningskonstituent (moder, datter, søster m.m., jf. <http://tedlab.mit.edu/~dr/Tgrep2/>).

□

search results for 'DN+fcl << /som/' in citat2000.t2c :

For graphical tree inspection, click on ID-code

- #10 ID=[adrvbmig](#) Japansk virksomhed overtager alle rettigheder til det industrielle spildevandsanlæg, **som** EnvoTech i Sønderborg har udviklet. #A1/2
- #18 ID=[aphzklum](#) Naturligvis vil han blive spurgt om, hvordan han vil øremærke den danske støtte til ganske bestemte formål, når den nu engang skal kanaliseres gennem et enevældigt regimes hænder og derefter - måske - videre til politi og dommere, **som er dele af et effektivt kinesisk undertrykkelsessystem**. #A12/12
- #24 ID=[agsshzhzw](#) "Men ud af ca. 500 møbelproducenter er der fortsat kun omkring 50, **som har en omsætning på over 100 mio. kr.** #A1/1
- #28 ID=[aovcphzg](#) Tivoli præsenterer "den originale amerikanske version", **som ingen til pressemodet kunne oplyse, hvornår havde haft premiere i USA**. #A1/3
- #42 ID=[agrbqagc](#) Russernes udbredte jødehad rammer netop en mand som Beresovskij, **som er rig, snedig og svarer at stole på**. #A1/2

Fig. 13: Søgning i træbank-korpora

□

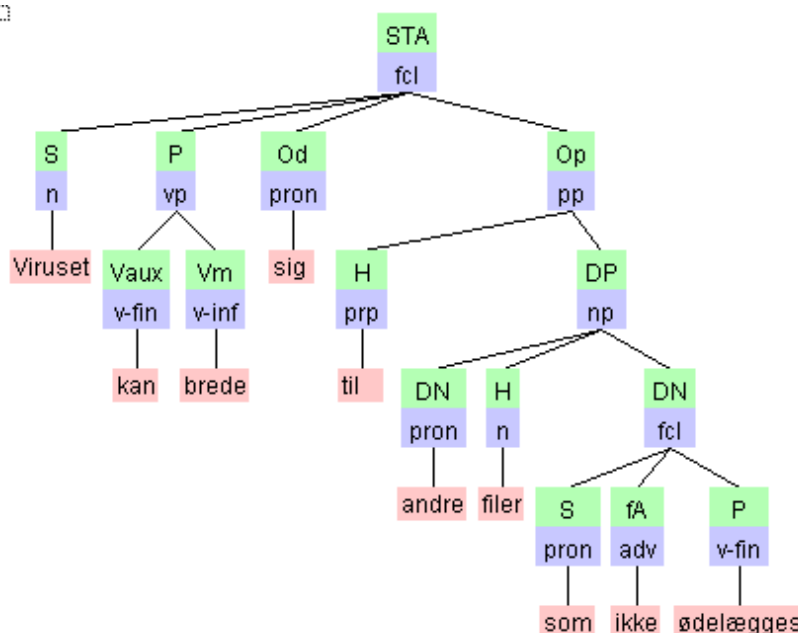


Fig. 14: Grafiske træstrukturer

6. Fejlevaluering

Det er en udbredt misforståelse blandt deltagende lærere på VSIL-seminarer, at sprogteknologiske redskaber, der er udviklet til at beskrive korrekt sprog, også umiddelbart kan anvendes til at korrigere fejlbehæftet tekst - en lærende funktion, mange af lærerne efterlyser i forbindelse med CALL, og hvor der ikke er den korrekte analyse, men det korrekte sprog, der står i forgrunden.

Men konfronteret med en fejlbehæftet sætning, vil en almindelig parser faktisk - alt efter anvendt teknologi - enten afvise sætningen som ugrammatisk (generative systemer) eller forsøge at tildele den den mest sandsynlige (probabilistiske systemer) eller mindst "stødende" analyse (Constraint Grammar-systemer). Når der forventes mere end en ren ordlistebaseret stavekontrol, altså herudover fx rettelse af ellers eksisterende, men i konteksten forkerte ord, bøjningsformer, ordfølge og kongruens etc., kræver det en særlig fejldreven grammatik til at fange fejlene. VISL-systemets danske tekstrettelsesfunktion (<http://beta.visl.sdu.dk/visl/da/tools/spelling.html>) benytter således en variant af den kommercielle stavekontrol OrdRet (Jandorf 2005 og <http://www.ordret.com>), der ud over almindelige stavefejl klassificerer ca. 30 forskellige

grammatiske fejltypes med rettelsesforslag. Den lærende funktion, der ligger i den eksplicite fejltypespecificering, muliggøres ved at det ordinære syntaktiske modul af DanGrams CG-system erstattes med en specialudviklet *fejlmapper*, dvs. et sæt af kontekstbaserede CG-regler, der "mapper" fejltypes-tags på ugrammatiske ord. Desuden bruges de almindelige CG-regler til at prioritere blandt rettelsesforslagene¹⁰. I et CALL-baseret undervisningsforløb kan læreren udvælge bestemte fejltypes og forsyne holdet med en to-do-liste af VISL-øvelser i netop det grammatiske emne, hvis beherskelse der kan afbøde fejlen.

I VISL's aktuelle samarbejde med lærerseminarierne (VISL-SEM, <http://beta.visl.sdu.dk/vislsem.html>) ligger fejlevalueringsfokus et andet sted. Her er det de lærerstuderendes evne selv at afdække og klassificere fejlene, der står i forgrunden - man kunne tale om en meta-fejlevaluering. Projektet har derfor som mål dels at sætte deltagerne i stand til selv at designe forømtalte fejltypesfokuserende CALL-forløb, dels at udvikle en interaktiv grænseflade, hvor brugeren præsenteres med skjult fejltypes-opmærkede *learner korpora*¹¹, og så skal finde og klassificere fejlene igennem menuvalg o.l., mens computeren vurderer de trufne valg og evaluerer den samlede præstation.

I fremtiden vil denne kvantitative evalueringsmulighed, i forbindelse med en unik bruger-id (jf. pilotanvendelse af bruger-login i KillerFiller), gøre det muligt at benytte dette og evt. andre af VISL's CALL-værktøjer i eksamensøjemed.

Bibliografi

- Bick, Eckhard, "Internet Based Grammar Teaching", in *Datalogvistisk Forenings Årsmøde 1997 i Kolding, Proceedings*, Ellen Christoffersen & Bradley Music (red.), pp. 86-106. Kolding: 1997 Institut for Erhvervssprog og Sproglig Informatik, Handelshøjskole Syd.
- Bick, Eckhard, *Grammy i Klostermølleskoven - "VISL light": Tværspørglig sætningsanalyse for begyndere*. Århus: 2002, Forlaget Mnemo
- Bick, Eckhard, "A CG & PSG Hybrid Approach to Automatic Corpus Annotation", in *Proceedings of SProLaC2003* (Corpus Linguistics 2003, Lancaster), Kiril Simow & Petya Osenova (red.), pp. 1-12, Lancaster: 2003
- Bick, Eckhard, "Grammatik for sjov: IT-baseret grammatik-læring med VISL", in *Call for the Nordic Languages: Tools and methods* (Proceedings of NorFa CALL Net Symposium Sept. 30. - Oct. 1. 2004), Peter Juel Henriksen (red.), København: 2004
- Bick, Eckhard, "CorpusEye: Et brugervenligt web-interface for grammatisk opmærkede korpora", in *10. Møde om Udforskningen af Dansk Sprog 7.-8.okt.2004, Proceedings*, Peter Widell & Mette Kunøe (red.), pp.46-57, Århus: 2005, Århus Universitet
- Christ, Oli, "A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest: 1994
- Dansk Sprognævn, "Kommaregler". Copenhagen: Dansk Sprognævn, pp. 17-30, København: 2004
- Dienhart, John, "VISL-projektet: Om anvendelse af IT i sprogundervisning og -forskning", in *At undervise med IKT*, pp. 51-70. Gylling: 2000, Narayana Press

¹⁰ Constraint Grammar er en reduktionistisk metode, der bygger på disambiguering af flertydigheder. Rettelsesforslag med forskellige ordklasser kan således kontekstuel disambigueres på samme vis som det sker for almindelige flertydige ord (fx 'køber' som hhv. substantiv eller verbum) - forudsat en approkimeringsalgoritme eller listesammenligning først har fundet de mulige retteleskandidater, og forudsat DanGram's morfologiske modul har tildelt disse deres mulige ordklasselæsninger.

¹¹ Det interne format lægger sig op ad det kommercielle program *Markin*, således at eksisterende lærer-rettede stile m.m. i Markin-format vil kunne importeres i VISL's internetbaserede retteværktøj.

Jandorf, Birgit Dilling (red.), *Rapport om OrdRet - en it-baseret stavekontrol*, København: 2005

Maegaard, Bente et.al., “Strategisk Satsning på Dansk Sprogteknologi”, København: 2004, Statens Humanistiske Forskningsråd