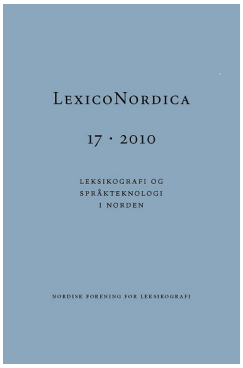
LexicoNordica

Titel: Deepdict - et korpusbaseret relationelt leksikon

Forfatter: Eckhard Bick

Kilde: LexicoNordica 17, 2010, s. 17-34

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

DeepDict – et korpusbaseret relationelt leksikon

Eckhard Bick

DeepDict (at www.gramtrans.com) is a new type of lexical resource, built from grammatically analysed corpus data. Co-occurrence strength between mother-daughter dependency pairs is used to automatically produce dictionary entries of typical complementation patterns and collocations, in the fashion of an instant monolingual usage dictionary. DeepDict is capable of abstracting lemma relations and semantic classes from inflected surface forms, and provides concordances and statistics for the relations found. Entries are supplied to the user in a graphical interface with various thresholds for lexical frequencies as well as absolute and relative co-occurrence frequencies. DeepDict draws its data from Constraint Grammar-analysed corpora, ranging between tens and hundreds of millions of words, covering the major Germanic and Romance languages, among them both Swedish, Danish and Norwegian. Apart from its obvious lexicographical purposes, DeepDict also targets teaching environments and translators.

1. Leksikografisk motivation

I bred leksikografisk forstand vil en korpusbaseret ordbog ikke alene generelt have et bedre dækningspotentiale, men også en større autenticitet end en traditionel ordbog kompileret vha. introspektion og litterære citater. Mange moderne ordbøger gør derfor brug af korpusdata, optimalt set med udgangspunkt i et materiale, der er balanceret mht. domæne, register etc. Alligevel ligner slutproduktet, den publicerede ordbog, som regel stadigvæk en traditionel ordbog, selv i elektroniske udgaver, fordi korpusdata er blevet brugt mere til eksemplificering, eller i bedste fald

frekvensoplysninger, end til egentlige ordbogsopslag. To undtagelser er *Sketch Engine* (Kilgariff et al. 2004), der benytter sig af n-gram-kollokationer og grammatiske relationer på systematisk vis, og *Wortschatz*-projektet ved Universität Leipzig (Biemann et al. 2004), der genererer netværk af semantisk beslægtede ord fra monolingvale korpora.

Men selv hvor der benyttes korpora i det leksikografiske arbejde, det være sig selektivt eller systematisk, kan der være store begrænsninger i tilgængeligheden af den information, der gemmer sig i et korpus, især hvad angår strukturel information, fordi de fleste korpora af den nødvendige størrelse kun foreligger som rene tekstkorpora, uden dybere grammatisk opmærkning. Allerede det mest basale opmærkningsniveau, med lemmatisering og ordklasse-entydiggørelse, vil tillade en bedre udnyttelse af korpusmaterialet, normalisering og optælling svarende til opslagsordets grundform etc.; men først en dyb syntaktisk-funktionel opmærkning med markering af subjekts- og objektsrelationer m.m. tillader ekstraktion af strukturelle relationer mellem ord, der ikke står umiddelbart ved siden af hinanden i teksten (såkaldte n-grammer).

Endelig, selv hvor leksikografen har adgang til et opmærket korpus af tilstrækkelig størrelse, med en brugerflade, der tillader opstilling af konkordanser og ordstatistik, vil det kun være muligt at undersøge ét relationelt mønster ad gangen – en besværlig proces, ikke mindst for verber med et komplekst frasalt og semantisk konstruktionspotentiale. Og ofte kan et givent mønster slet ikke findes i korpusset, enten fordi søgeformalismen ikke er tilstrækkelig finkornet, idet den fx er tekstbaseret snarere end kategoribaseret, eller fordi korpora med den nødvendige opmærkningsdybde (en såkaldt træbank) som regel kun produceres som håndopmærkede korpora med få hundredetusinde ord[1].

1 Karel Kalurand anfører netop begrænsninger af denne type, dvs. dækningsgrad og statistisk prægnans, som problemer i forbindelse med hans

Det leksikografiske redskab, der præsenteres her, DeepDict, forsøger at gå nye veje, både hvad angår den lingvistiske kvalitet i den tilgængelige korpusinformation, og mht. en mere integreret præsentation af de relationelle informationer for det enkelte ord. DeepDict blev udviklet af GrammarSoft Aps og lanceret på internetadressen www.gramtrans.com i september 2007.

I modsætning til en papirordbog har en elektronisk ordbog som DeepDict ingen volumenbegrænsninger, så opslaget for et sjældent ord kan fylde lige så meget som for et højfrekvent ord, og udelukkelsen af sjældne ord og relationer behøver derfor ikke at være absolut, men kan reguleres af brugerstyrede tærskler. Men særlig store bliver fordelene for en produktionsordbog: På papirmediet er det nemlig nemmere at fremstille passive ("definitions-") ordbøger end aktive (produktivt-kontekstuelle) ordbøger, fordi førstnævnte henvender sig til modersmålsbrugere af målsproget (MS), mens sidstnævnte optimalt set skal levere en stor mængde detaljerede brugsinformationer, semantiske restriktioner og kompletteringsmønstre for brugere med MS som fremmedsprog. Fx "A gives x to B" – med A, B som person-variable (+HUM) og x, y som ting-variable (-HUM). En elektronisk ordbog kan derimod rumme et væld af brugsinformation "on demand" og tilbyde ubegrænsede korpuseksempler – eksempler, der ikke optager plads i det primære opslag og først bliver synlige, når brugeren aktiverer et tilsvarende link.

2. Kompileringen af en leksiko-relationel database

For at honorere de krav om robust og detaljeret grammatisk korpusopmærkning, der blev drøftet i kapitel 1, valgte vi Constraint Grammar (CG, Karlsson et al. 1995) som sprogligt analyse- og

deepdict-lister, der bygger på en estisk CG-baseret træbank med 100.000 ord (http://math.ut.ee/~kareel/NLP/Programs/Treebank/DepDict).

opmærkningsparadigme, dels pga. metodens meget lave parsing-fejlprocenter og gode leksikalsk-morfologiske dækningsgrad, dels fordi CG-syntaksen bygger på dependensrelationer, dvs. relationer mellem ord snarere end mellem non-terminale konstituenter, med al syntaktisk information tilgængelig på ordniveau – et forhold, der medfører betydelige lettelser i computer-processeringen af opmærkede data. I det følgende beskrives den valgte fremgangsmåde for opbygningen af en leksiko-relationel database.

2.1. Korpusopmærkning

Det første skridt for hvert sprog bestod i den grammatiske opmærkning af samtlige tilgængelige korpora vha. Constraint Grammar-parsere, efterfulgt af en dependens-analyse med CG-tags (fx @SUBJ = subjekt, @ACC = direkte objekt) som input (Bick 2005). Resultatet kan beskrives som en gigantisk træbank på ca. en milliard ord, med dependensrelationer for samtlige ord i hver sætning[2]. For nogle af vores korpora var det dog kun det sidste trin, der var del af DeepDict-projektet selv, idet materialet allerede forelå som CG-opmærkede korpora inden for CorpusEye-systemet (http://corp.hum.sdu.dk). Tabel 1 giver et overblik over art og omfang af de anvendte korpora.

I det nedenstående opmærkede sætningseksempel har både subjektet *Peter* (ord 1) og objektet *nødder* (ord 6) dependensrelationer (#x→y) til verbet *spiste* (ord 2).

Peter "**Peter**" <hum> PROP @**SUBJ** #1→2
spiste "spise" V IMPF #2→0
en håndfuld ....
nødder "**nød**" <fruit> N P @**ACC** #5→2

2 Dependenstræerne har fuld dybde og er således informationsækvivalente med tilsvarende konstituent-træbanker, CG3-dependenser (beta.visl.sdu.dk/constraint_grammar.html) eller Functional Dependency Grammar (www.connexor.fi).

	Korpus-størrelse[3]	Genre	Parser[4]	Status[5]
Dansk	159 mio.	blandet	DanGram	+
Engelsk	210 mio.	blandet	EngGram	+
Esperanto	58 mio.	blandet	EspGram	+
Fransk	[67 mio.]	Wiki, Europarl	DTT+FrAG	–
Italiensk	46 mio.	Wiki, Europarl	DTT+ItaGram	+
Tysk	44 mio.	Wiki, Europarl	GerGram	+
Norsk	50 mio.	Wiki, kundedata	Obt / NorGram	+
Portugisisk	210 mio.	avis, Europarl	PALAVRAS	+
Spansk	90 mio.	internet, wiki, Europarl	HISPAL	+
Svensk	60 mio.	avis, Europarl	SweGram	+

Tabel 1: Korpora og parsere

2.2. Dependensbigrammer

Det er denne type binære relationer, dvs. dependenspar, der blev "høstet" fra de opmærkede korpora, med informationer om lemma, ordklasse og syntaktisk funktion for både dependenten ("datterordet") og hovedet ("moderordet").

Peter_SUBJ → spise_V
kat_SUBJ → spise_V
nød_ACC → spise_V
mus_ACC → spise_V

For at undgå en eksplosion af informationsløs leksikalsk mangfoldighed blev talord og navne udelukkende gemt uden deres lemma, for sidstnævnte dog med en markering af semantisk klasse,

3 Wiki = Wikipedia (http://www.wikipedia.com), Europarl = the Europarl Corpus (Koehn 2005).

4 Mere information om parserne fås på: http://beta.visl.sdu.dk/constraint_grammar.html.

5 Der er fri adgang til DeepDict for portugisisk, svensk og esperanto, mens der kræves login/abonnement for de øvrige sprog.

fx <hum> (menneske), <org> (organisation) etc. Også præpositioner fik en særbehandling i ekstraktionsprocessen; dels var det styrelsen, dvs. den semantiske kerne, snarere end præpositionen selv, der blev betragtet som hovedet, dels blev der de facto brugt 3-leds-relationer, idet præpositioner blev gemt som en slags kasusmarkør sammen med deres styrelse (fx *tygge* ← *på* ← *problem* giver relationen *tygge* ← *problem**på*).

De fleste af de anvendte parsere leverer foruden den syntaktiske også en semantisk opmærkning med såkaldte semantiske prototyper for substantiverne – i stil med den allerede nævnte navneklassificering, men på et højere distinktionsniveau med ca. 200 prototyper. <fruit> (frugt), for eksempel, er en undertype af <food> (mad), der igen kan være en undertype (<food-c>, <food-m>) af <cc> (tællelige konkreta) eller <cm> (mængdekonkreta). En række hovedkategorier tilføjer semantiske underklasser som små bogstaver efter et stort bogstav for hovedklassen, fx <Vair> (air vehicle), <tool-cut> (skære-redskab) og <Hprof> (human professional).

Lægger man de enkelte lemma-, ordklasse- og prototype-relationer samlet ind under dependenshovedet som opslagsord, får man fx for verbet *eat* ('spise') et summarisk opslag, der viser, hvem der spiser (SUBJ-subjekt, fx PROP-proprium), og hvad der spises (ACC-objekt):

{PROP, *kat*, <hum>, ...} SUBJ → *spise*
spise ← {*nød*, *mus*, <fruit>} ACC

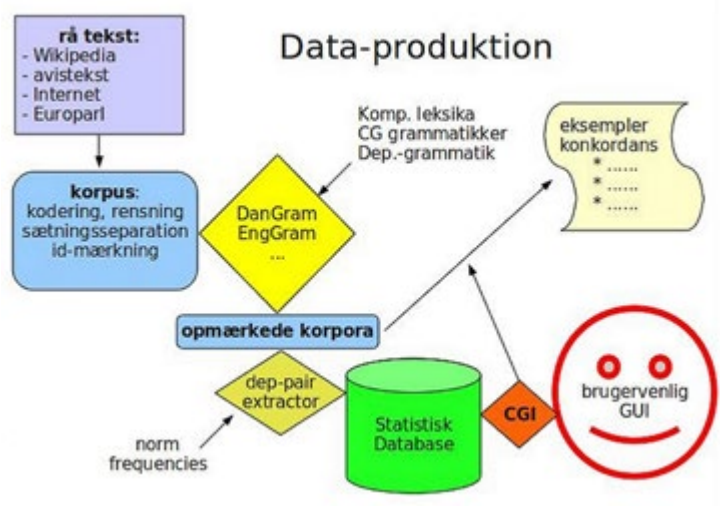
2.3. En database over korrelationsstyrker

Det er åbenlyst, at dependentlisterne i et sådant opslag uden statistisk information hurtigt ville blive reduceret til "leksikalsk støj" af den kombinatoriske mangfoldighed i et stort korpus. Det er med andre ord nødvendigt at skelne mellem typiske komplementer og korrelationer på den ene side og ikke-informativ "støj"-variation

på den anden side. Vi har derfor benyttet et statistisk mål for korrelationsstyrke, dvs. sandsynligheden for samforekomsten af 2 ord i en given syntaktisk relation. For at sondre mellem typiske og ikke-informative korrelationer dividerede vi den absolutte frekvens for samforekomsten med produktet af korpus-normalfrekvenserne for hvert af de 2 ord alene:

$$C * \log(p(a \rightarrow b) \, ^2 / \, (p(a) * p(b)))$$

hvor p() står for frekvenser, og C er en konstant, der sammen med logaritmiseringen blev introduceret for at placere statistisk signifikante værdier mellem 0 og 10. Forskellen mellem vores formel og Church's *Mutual Information*-mål (Church & Hanks 1990) er den øgede vægtning (^2 = kvadratvægtning) af selve samforekomstfrekvensen – en vægtning, vi anså for gavnlig i leksikografisk øjemed, fordi den hindrer stærke men sjældne eller forkerte kollokationer i at udkonkurrere kollokationer bestående af mere almindelige ord (og tilsvarende høje frekvensværdier i brøken).



Figur 1: Data-produktion og GUI (graphical user interface)

Den endelige standardiserede dependensdatabase indeholder for hvert "dep-gram"-ordpar, foruden dets absolutte frekvens og kookkurens-styrke, også et indeks over id'erne på de relevante sætningsforekomster i kildekorpusset.

Selv for et enkelt sprog kan hele processen tage dage eller uger, og databaserne har en størrelse (p.t. 90 GB), der gør det umuligt at benytte sig af almindelige database-programmer, fordi et enkelt opslag ville medføre en for brugeren uacceptabel ventetid på flere minutter, og vores interface-programmør, Tino Didriksen, var nødt til at udvikle særlige opslagsalgoritmer og multiple filstrukturer for at løse problemet.

3. Brugerinterfacet

Opslag i DeepDict er dynamiske "leksikogrammer" – frekvenssorterede, grafisk ordnede lister af kollokater. Præcis hvilke kollokater der vises, er afhængigt dels af opslagsordets ordklasse og dermed typiske funktionelle kompletteringsmønster, dels af en række tærskelværdier, der kan sættes individuelt for at tilgodese forskellige brugerprofiler:

- minimum-forekomst (af dependens-kollokationen) – bruges til at bortfiltrere tekstfejl, opmærkningsfejl og hapaxer
- minimum-kookkurrens-styrke (default > 0) – regulerer typiciteten af kollokaterne
- maksimum antal kollokater, der vises per funktionsfelt
- leksikalsk minimumsfrekvens for kollokat-ordene (4 niveauer) – kan bruges til at sikre, at kun almindelige ord vises som kollokater, fx til skolebrug

Af grammatiske årsager skelnes mellem fx "tale_V" (verbum) og "tale_N" (substantiv/nomen), og hver ordklasse har sin egen leksikogramskabelon. Leksikogrammet for det engelske substantiv

voice, for eksempel, indeholder således ikke bare typiske flerordsudtryk som *voice actor* eller *voice recorder*, men viser også typiske attributter (i feltet "premodifier"), fx *loud*, *deep*, *husky* og det flertydige *passive voice*.

Figur 2: Substantiv-leksikogram

Felterne i DeepDict er placeret på en måde, der understøtter "naturlig læsning". Attributter findes derfor til venstre og hoveder til højre for adjektiviske og substantiviske opslagsord på engelsk, svarende til sprogets normale ordfølge. Tilsvarende placeres subjekter til venstre for et verbum og objekterne til højre. Nogle felter er forsynet med en tekstramme for at skabe illusionen af en "sætning", fx "one can *{recognize, hear, lower, lend, raise}* a voice".

Værdierne for kookkurrensstyrken angives optionelt som røde tal foran det enkelte kollokatord, efterfulgt af en kolonseparator og den duale logaritme af den absolutte forekomst. Som default vises kun kollokationer med en logaritmeklasse på 2 eller højere (4 eller flere forekomster). Rækkefølgen af ordene i et felt er en

samlet funktion af kookkurensstyrke og absolut frekvens, og for yderligere at skelne mellem sikre og usikre kollokater vises høje logaritmeklasser med fed skrift. Når man klikker på et kollokat, åbnes et konkordansvindue der viser sætningseksempler og en fuldformsstatistik for den pågældende lemma-kollokation.

Forms	Abs Freq	Rel Freq
Total	6749	100.00%
vægt -> lægger	3608	53.46%
vægt -> lægge	1270	18.82%
vægt -> lagde	820	12.15%
vægt -> lagt	815	12.08%
vægten -> lægger	100	1.48%
vægten -> lægge	68	1.01%

Show all forms...

word sketch:

lægger . stor mest afgørende meget **vægt** **på** {. / . / , fordi } {bag / på denne / på ved ansættelser / på det}

Concordances for: vægt_N -> lægge_V

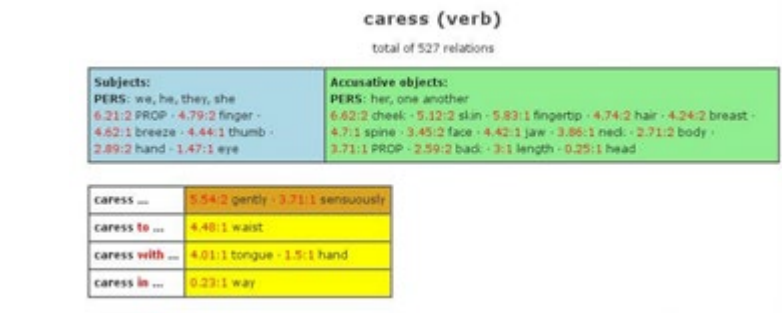
ID	Text
inf30-33978	« Markedsføring på værdier og ansvar skal være en vigtig kilde til at øget konkurrenceevne på verdensmarkedet , fordi den globale forbruger i stadig større udstrækning **lægger vægt** på den slags , når han eller hun køber ind .
inf100-35624	I sidste ende giver den diplomatiske balancegang dog mulighed for , at den tyrkiske regering for en tid kan **lægge** mest **vægt** på sin egen tillægserklæring af indenrigspolitiske årsager , mens EU fokuserer på toldaftalen .
inf40-148330	« I dag **lægger** vi megen **vægt** på , at tilldsmanden skal arbejde med uddannelse og kunne svare på spørgsmål om pension .
c2000-	Netop den pind har vi fået ind , og det **lægger** vi stor **vægt** på , fordi vi netop vil gøre

Figur 3: Konkordansopslag

For støtteverbumskonstruktioner kan det være nødvendigt med en dependensdybde større end 2, dvs. at vise flere komplementer på én gang, som i udtrykket *lægge ... vægt på/bag ngt*. Her fungerer ordet *vægt* som syntaktisk objekt, men indgår i en inkorporation med verbet, hvis egentlige komplement er præpositionssyntagmet *på ...*. Mens DeepDicts primære opslag kun fokuserer på det umiddelbare objekt, vises hele konstruktionen i konkordansopslaget som en såkaldt "word sketch".

Personlige og kvantitative pronominer er så frekvente, at eksakte statistiske værdier her kun har begrænset interesse. Til gengæld kan pronominer levere semantisk information, "abstraheret"

som pronominale prototyper (fx ±human, køn, ±tællelig, sted/retning), og DeepDict viser derfor en ordnet liste af karakteristiske pronominer på subjekts- og objektspladserne. Personlige subjektspronominer kan hjælpe med at klassificere aktiviteter som typisk mandlig ('han') eller kvindelig ('hun'), markere objekter som mængdeord ('meget') eller endda tillade sociolingvistiske deduktioner. Således viser DeepDict-opslaget for det engelske verbum *caress* at mænd ('he') typisk er subjekt og kvinder ('her') typisk objekt i kærtegningsrelationen.



Figur 4: Verbums-leksikogram

Eksemplet viser desuden, at metaforisk brug dækkes ind på samme måde som konkret brug – således vises der ud over objekt-kropsdele, der kærtegnes, og subjekt-kropsdele, der kærtegner (*finger*, *thumb*), også metaforiske agenter som *breeze* og *eye*. Endelig illustreres, hvordan præpositioner (*with tongue/hand*) håndteres i DeepDicts verbalskabelon.

Adverbium-verbums-kollokationer eksisterer i flere funktionelle varianter – (a) ubundne tids-, steds- og mådesadverbier, (b) valensbundne adverbier (feel *how*, go *where*) og (c) verbalintegrerede partikler (give *up*, fall *apart*), og i nogle tilfælde kan det endda være svære at skelne mellem kategorierne (fx cut *out*). Fordi formålet med DeepDict er leksikografisk snarere end syntaktisk, nøjedes vi dog her med kun at fremhæve verbalpartiklerne som

separat klasse (for at understøtte en underlemmatisering af det pågældende verbum) og at samle alt andet adverbielt materiale i en og samme paraplykategori (brunt felt, fx *gently/sensuously* for verbet *caress*).

4. Betydningsnuancer igennem dependens-kollokater

Selvom DeepDict også for polyseme ord viser kollokaterne samlet[6], kan det undertiden hjælpe at udgrænse forskellige kernebetydninger, nemlig igennem det semantiske spektrum af kollokaterne (fx både konkrete og abstrakte prototyper) og igennem den syntaktiske funktion, der knyttes til en given relation. Således fremgår det af leksikogrammet for det portugisiske adjektiv *pesado* ('tung'), at ordet både anvendes konkret (= 'af høj vægt') og abstrakt (= 'betydelig/alvorlig'), og at det i førstnævnte betydning har en tendens til at blive brugt som postmodifikator, mens det foranstilles som præmodifikator ved abstrakte kollokater.

Figur 5: Adjektiv-leksikogram

6 Medmindre distinktionen allerede er en del af den forudgående korpusopmærkning.

Tilsvarende kan DeepDict hjælpe med at fremhæve betydningsnuancerne mellem nære synonymer. Således kan det for en studerende af dansk som fremmedsprog være svært at vide, hvornår han skal benytte hhv. *mistænksom* og *mistænkelig*. De to tilsvarende opslag på DeepDict vil imidlertid gøre det klart, at førstnævnte beskriver et udtryk/indtryk, mens sidstnævnte bruges om handlinger og hændelser:

mistænksom ...	**mistænkelig ...**
stemme, ?forsvindingsnummer, receptionist, øje, grimasse, blik, gemyt, tonefald, ?transaktion	transaktion, person, færden, grad, pengeoverførsel, adfærd, personage, dødsfald, forhold
<expression>	<action, event, situation>

Tabel 2: Betydningsnuancer

Omvendt kan det for en dansker være vanskeligt at anvende de engelske adjektiver *big*, *large* og *high* korrekt, men også her formår DeepDict-korrelaterne implicit at "definere" betydningerne:

high ...	**big ...**	**large ...**
• level • [school] • concentration • speed • proportion • altitude • elevation • temperature	• [bang, band] • hit • problematic • break • difference • brother • star, bird • man, city	• number • quantity • amount • proportion • sum • portion, part • city, island • population
<degree> <measure>	<size> <importance>	<extension> <quantity-mass>

Tabel 3: Semantisk motiverede kollokationsrestriktioner

Samtidigt identificeres visse flerordsudtryk *[big bang, big band]*, engelske komposita med tryk på første led. Men mens sådanne

flerordsudtryk også er tilgængelige for en ren tekstuel kollokationsanalyse, drager de øvrige, funktionelle kollokationer fordel af CG-dependensrelationerne. Således vil relationen *high + temperature* findes, selv hvis der ikke foreligger en eneste sætning, hvor ordene står ved siden af hinanden – fordi relationerne også fanges i *high room temperature* eller i prædikativ brug, *ambient temperature was rather high when ….*

I et bilingvalt perspektiv kan DeepDict advare brugeren om, at en oversættelse, selv mellem nært beslægtede sprog, ikke nødvendigvis matcher ordene en-til-en. Således rummer den svenske oversættelse *smeka* af dansk *kærtegne* også betydninger (fx 'stryge'), der end ikke metaforisk dækkes af det danske ord, og DeepDict-leksikogrammet viser dette igennem de fundne typiske objekter:

kærtegne ...	**smeka ...**
• bryst, krop, kind, hud, balder, mave, inderlår, brystvorte, hår, ansigt, klitoris, lår, sexbombe, nosse, røvhul, nakke, hals, kropsdel, bagdel • silkestof, græsbane • PROP-hum	• kind, könsorgan, bröst, stjärt, klitoris, kropp • PROP-hum • boll, passning, tennisboll • elgitarr • lack, rännil, instrumentpanel, julle, murbrok, vidunder

Tabel 4: Bilingval polysemikontrol

5. DeepDict som arbejdsredskab

Eksemplerne i de forudgående kapitler viser art og omfang af den information, der gemmer sig i DeepDict-opslagene. Men på sin vis er der tale om et uslebent værktøj, hvor mange muligheder nok understøttes, men på den anden side forudsætter en vis grad af nytænkning og tilpasning hos brugeren. Oplagte brugergrupper ud over den almindelige "ordbogs"-bruger er (a) leksikografen

og (b) universitetsunderviseren. Leksikografen kan således finde inspiration mht. kompletteringsmønstre, flerordsudtryk, frasale verber m.m. og uddrage de mest karakteristiske eksempler for en given konstruktion, snarere end bare de mest frekvente. Bl.a. vil en metaforisk kombination ofte udvise en høj korrelationsværdi, netop hvis den ene part ellers er et lavfrekvent ord. Desuden understøttes semantiske subdistinktioner og sammenligninger som vist ved adjektiveksemplerne i sidste afsnit.

For underviseren kan DeepDict, i forbindelse med udarbejdelse af det relevante didaktiske materiale, være et middel til at stimulere de studerendes sproglige nysgerrighed og give undervisningen et mindre teoretisk, men mere empirisk og datanært præg, især når redskabet kombineres med almindelig korpusbrug. Mulighederne strækker sig fra ordfelt-øvelser (fx mad & drikke, via verberne *spise* og *drikke*, sprog- og landenavne etc.), over kombinatoriske undersøgelser (hvilken præposition styres typisk af et givent substantiv eller verbum?) til semantiske (fx metaforer) eller sociolingvistiske øvelser (fx konnotationerne af ordene *udlænding*, *indvandrer* og *flygtning* igennem tilknyttede adjektiver).

6. Konklusion og perspektivering

DeepDict viser, hvordan syntaktisk relaterede ordpar kan "høstes" fra grammatisk opmærkede dependenskorpora til at kompilere en statistisk database, der tillader genereringen af såkaldte "leksikogrammer" – halvgrafiske oversigtssider for monolingvale ordbogsopslag, med information vedrørende hoved- og modifikator-selektionsrestriktioner, verbalkomplettering og frasale kollokationer. DeepDict gør det muligt for leksikografen ikke alene at finde korpuseksempler og -frekvenser for bestemte (kendte) kollokationer og leksikale strukturer, men også at kompilere (nye) lister over sådanne kollokationer og strukturer.

6.1. Bedre parsere

Rent forskningsmæssigt kan de statistiske informationer fra DeepDict-databasen bruges til at forbedre CG-parserne, der så igen kan levere bedre korpora til en ny runde DeepDict-generering. Således har forfatteren udvidet det portugisiske parsingleksikon med tags for sandsynligheden for at en given syntaktisk funktions-"slot" udfyldes af en bestemt semantisk prototype:

- *pensar* ('tænke'): <fSUBJ/H:74>, <FSUBJ/org:25>
  (<fSUBJ/H:74>: f=frekvens, SUBJ=subjekt, H=human, 74=frekvensprocent)
- *competir* ('konkurrere'): <fPRP-com/H:81>, <fPRP-com/A:18>

Denne type information kan så bruges i fx en anaforgrammatik til at human-markere portugisiske personlige pronominer, der ellers kun har grammatisk køn:

ADD (£hum) TARGET PERS + @P<
(p @PIV LINK 0 PRP-COM LINK p (<fPRP-com/H>70>))

(markér PERS som human[£hum], hvis den fungerer som styrelse (@P<) til et præpositionelt objekt (@PIV) 'com' (=med), som så igen har et dependenshoved (p) med verbal-kompletterings-tag der kræver både præpositionen 'com' og trækket H (human) med en sandsynlighed større end 70)

6.2. Framenet

DeepDicts nuværende leksikogrammer fokuserer på én binær relation ad gangen, dvs. at fx subjektsfeltet og objektsfeltet beregnes uafhængigt af hinanden. Mens dette er fuldt tilstrækkeligt til mange anvendelser, kan det i en fuldstændig beskrivelse af verbets potentiale være interessant også at inddrage mulige gensidige

afhængigheder af subjekter og objekter og derfor at arbejde med såkaldte "frames" (http://framenet.icsi.berkeley.edu/), fx <hum> 'læse' <sem-r>, i stedet for dependens-bigrammer (<hum> 'læse' og 'læse' <sem-r> hver for sig). Dette kan imidlertid lade sig gøre med de samme annoterede korpora som udgangspunkt, og forfatterens plan er således at benytte DeepDicts database til at fuldføre det påbegyndte danske framenet på www.framenet.dk.

6.3. Brugertilpasning

Med slutbrugere i tankerne kan DeepDict som integreret eller separat modul kobles til andre leksikale ressourcer – traditionelle definitionsordbøger, ontologier eller bilingvale ordbøger (fx QuickDict-ordbøgerne på www.gramtrans.com), hvor DeepDict kan udfylde rollen som *aktiv* ordbog, dvs. vise brug og brugsrestriktioner for et givet målsprogsord.

Fordi DeepDict-metoden i princippet er anvendelig for alle typer af tekstkorpora, der kan analyseres med en Constraint Grammar-parser, vil det desuden være muligt at forsyne sprogforskere, leksikografer og lærere med individuelle DeepDict-installationer for specifikke brugerkorpora, tilpasset et bestemt domæne, en særlig genre eller forskellige geografiske eller sociale sprogvarianter.

Litteratur

Bick, Eckhard 2005: Turning Constraint Grammar Data into Running Dependency Treebanks. I: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.): *Proceedings of TLT 2005*, Barcelona, December 9th–10th, 2005), 19–27.

Bick, Eckhard 2006: A Constraint Grammar-Based Parser for Spanish. I: *Proceedings of TIL 2006 – 4th Workshop on Information and HLT.*

Biemann, Chris & Stefan Bordag & Uwe Quasthoff & Christian Wolff 2004: Language-Independent Methods for Compiling Monolingual Lexical Data. I: *Comp. Linguistics and Intelligent Text Processing*. Berlin: Springer, 217–228.

Church, Ken & Patrick. Hanks 1990: Word Association Norms, Mutual Information and Lexicography. I: *Computational Linguistics,* vol.16:1, 22–29.

Karlsson, Fred et al. 1995: Constraint Grammar – A Language-Independent System for Parsing Unrestricted Text. I: *Natural Language Processing*, no. 4. Berlin & New York: Mouton de Gruyter.

Kilgarriff, Adam, Rychlý, P., Smrž, P. & Tugwell, D. 2004: The Sketch Engine. I: *Proceedings of Euralex 2004 (Lorient, France)*, 105–116.

Koehn, Philipp 2005: Europarl – A Parallel Corpus for Statistical Machine Translation. I: *MT Summit X* (Sept.12–16, 2005). Phuket, Thailand.

Eckhard Bick
forskningslektor, dr.phil.
Syddansk Universitet
Rugbjergvej 98
DK-8260 Viby J
eckhard.bick@mail.dk