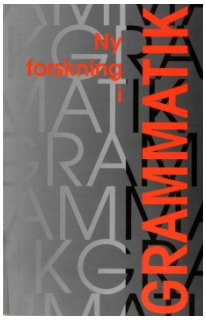
Ny Forskning i Grammatik

Titel: The Copenhagen Dependency Treebanks.
Forskellige niveauer - samme relationer

Forfatter: Iørn Korzen og Henrik Høeg Müller

Kilde: Ny Forskning i Grammatik 18, 2011, s. 173-196

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



The Copenhagen Dependency Treebanks. Forskellige niveauer – samme relationer

Iørn Korzen og Henrik Høeg Müller

1. Indledning

The Copenhagen Dependency Trebanks (CDT) er et igangværende FKK-støttet projekt, der har til formål at annotere 80.000-100.000 ord på de fem sprog dansk, engelsk, tysk, italiensk og spansk mht. morfologisk, syntaktisk, anaforisk og diskursiv struktur. Det korpus vi opmærker, består af små tekstuddrag på 200-240 ord hver, der er oversat fra dansk til de øvrige sprog. Udover opmærkningen af teksterne på de fem sprog mht. de nævnte lingvistiske niveauer foretages der også en alignering mellem dansk og de øvrige sprog. Det går ud på at man manuelt/halvautomatisk forbinder de ord, fraser og udtryk der mere eller mindre korresponderer med hinanden i dansk og det relevante andet sprog. Kort fortalt opbygges der på den måde en multilingval parallel træbank, der kan fungere som et vigtigt værktøj til forbedring af maskinoversættelse, som i øjeblikket hovedsageligt er statistisk baseret, og træning af parsere med henblik på (halv)automatisk annotation af store tekstmængder. Herudover muliggør parallelle træbanker generelt at man kan foretage statistisk funderede undersøgelser af de lingvistiske fænomener der annoteres, både monolingvalt og i et kontrastivt/typologisk perspektiv.

I det følgende vil vi kort beskrive de dependensprincipper som CDT er bygget op efter, men artiklens hovedfokus vil være at redegøre for hvordan Pustejovskys (fx 1995: 85ff.) qualia-relationer er forsøgt anvendt som fælles udgangspunkt for annotation af lingvistisk struktur på de forskellige niveauer CDT arbejder med, syntaks-, morfologi-, anafor- og

diskursniveauet. Heraf artiklens titel. Qualia-strukturen introduceres nedenfor i skema 1.

- FORMAL: That which distinguishes the object within a larger domain (e.g. shape, dimensionality, position, color).
- CONSTITUTIVE: The relation between an object and its constituents, or proper parts (e.g. material, weight, parts and component elements).
- AGENTIVE: Factors involved in the origin or "bringing about" of an object (e.g. creator, artefact, natural kind).
- TELIC: Purpose and function of the object.

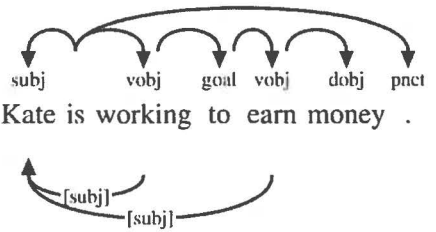
Skema 1. Qualia-strukturens fire komponenter, jf. Pustejovsky (ibid.).

Meget generelt beskrevet kan qualia-strukturen, som er inspireret af Moravcsiks arbejder (1975 og 1990) og de aristoteliske *aitia* – Aristoteles fire forklaringer om entiteters væsen – opfattes som en systematisk fremstilling af de synsvinkler, hvorfra et substantivs denotat kan betragtes. Qualiaens fire komponenter repræsenterer således en del af et substantivs interne semantiske struktur, dvs. en slags semantisk skabelon til beskrivelse af en del af begrebernes anatomi. Den skal ses som almene erkendelsesprincipper i forbindelse med entiteter, den skabelon vores tanker om entiteter er fastlagt i henhold til, og derfor kan den fungere som en model for beskrivelse og præcisering af de relationer der kan etableres fx inden for et NP eller mellem NP'er der er anaforisk knyttet til hinanden. I vores arbejde, som vi skal se i det følgende, har vi som sagt ladet os inspirere af qualia-strukturen og anvender den bredt som en slags overordnet organisatorisk princip som mange semantiske relationer i CDT kan indordnes under.

2. Qualia og syntaks

2.1. Sætningsstruktur

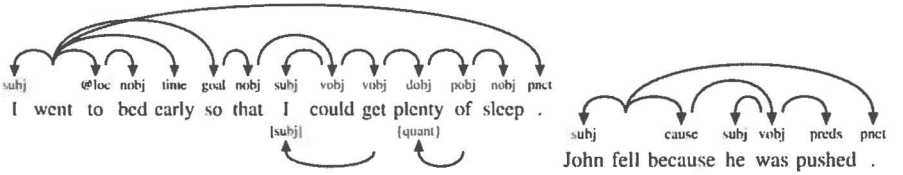
Nedenfor i figur 1 vises hvordan en gængs dependensannotation af en sætning tager sig ud i CDT.



Figur 1. Grundlæggende dependensannotation af sætning i CDT.

Det finitte verbal, *is,* fungerer som sætningens centrum og øverste knude, der på det primære niveau knytter to led til sig i form af et subjekt, *Kate,* og et verbalobjekt, *working to earn money.* Dette er indikeret ved pile over teksten, der udgår fra *is* til de to konstituenter med relationsnavnene skrevet under pilespidsen. Punktum forbindes til det finitte verbum *is* via pilen med relationsnavnet "pnct". Det leksikalske hovedverbum, *working,* etablerer en adverbiel "goal"-relation til *to earn money,* og inden for adverbialet fungerer *earn* som verbalobjekt for infinitivmærket *to,* og *money* som direkte objekt for *earn.* Pilene under teksten specificerer sekundære subjektrelationer. Dependensannotationen på sætningsniveau er beskrevet fx i Buch-Kromann (2006) og Buch-Kromann et al. (2009 og 2010).

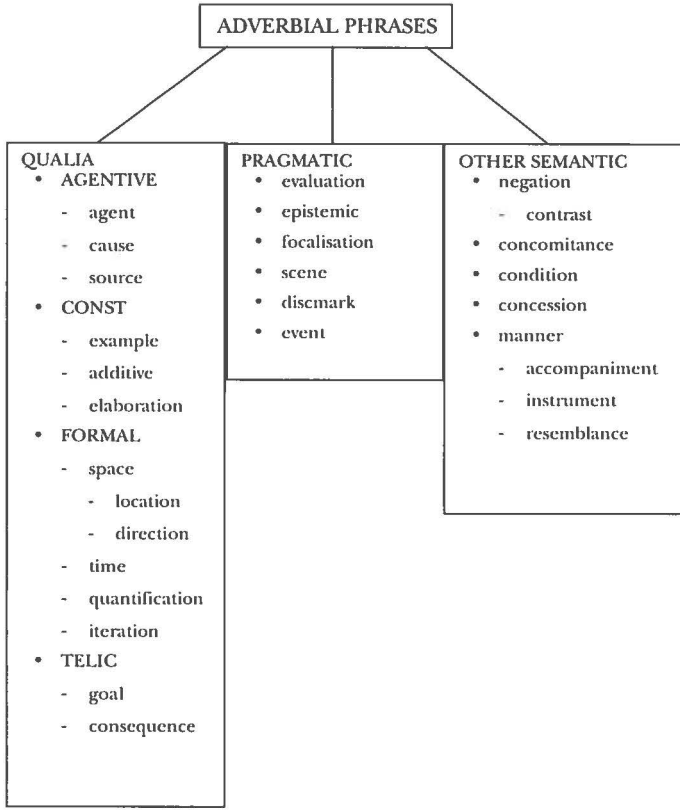
Generelt foretages der ikke semantisk annotation på sætningsniveauet, men kun syntaktisk dependensannotation. Dette gælder dog ikke adverbialerne, som annoteres i henhold til den semantiske relation der etableres mellem kerneleddet og adverbialet. I figur 2 ses to eksempler på annotation af adverbialer.



Figur 2. Annotation af adverbialer der udtrykker goal (venstre) og cause (højre).

I ovenstående sætninger etablerer de adverbielle ledsætninger, *so that I could get plenty of sleep* og *because he was pushed,* betydningsrelationerne "goal" og "cause", som vi jf. skema 2 nedenfor har defineret som hørende

til qualia-inventaret.[1] Som tidligere nævnt aktiveres "goal"-relationen også i figur 1, her af et adverbial i form af en infinitivfrase. Overførslen af qualia-begrebet fra analysen af substantiver til analysen af adverbialer skal forstås på den måde, at ligesom substantiver kan få aktiveret deres qualia-roller af forskellige modifikatorer, i form af fx PP'er, NP'er eller AP'er, kan også adverbielle modifikatorer aktivere forskellige af sætningens qualia-betydningsdimensioner (se også afsnit 5 og note 17). Heri består analogien mellem de forskellige niveauer.



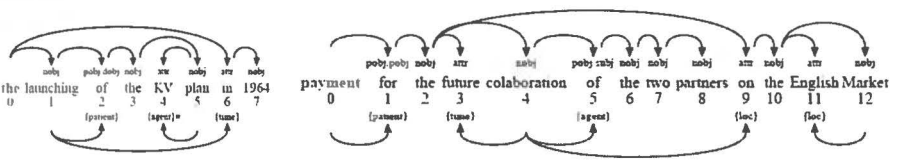
Skema 2. Sematiske og pragmatiske relationer anvendt ved annotation af adverbialer.

1. Af grunde, det vil føre for vidt at komme ind på her, bruger vi i visse tilfælde andre betegnelser end de oprindelige qualia-roller. "Goal" og "cause" svarer således til henholdsvis TELIC og AGENTIVE. Desuden skal det nævnes, at i ledsætninger opfattes ledsætningsindlederen som kerne, hvorfor der i figur 2 går en pil fra henholdsvis *that* og *because* til de pågældende ledsætningers finitte verbaler. Forholdet mellem konstituenterne i den komplekse konjunktion *so that* angives som "nobj" pga. *that*-sætningens nominale karakter.

Det eneste sted semantikken har en plads i analysen af sætningsniveauet, er altså i forbindelse med adverbialerne, og det er også værd at lægge mærke til at qualia naturligvis ikke kan redegøre for en hvilken som helst adverbiel relation, hvilket de øvrige kategorier i skemaet vidner om.

2.2. NP-struktur

Bevæger vi os videre til NP-niveauet, forholder det sig anderledes, idet vi her, jf. figur 3, både foretager en syntaktisk dependensannotation, som er annoteret med pilene over teksten, og tillige udfører en delvis semantisk analyse, som er specificeret med pilene under teksten.[2]



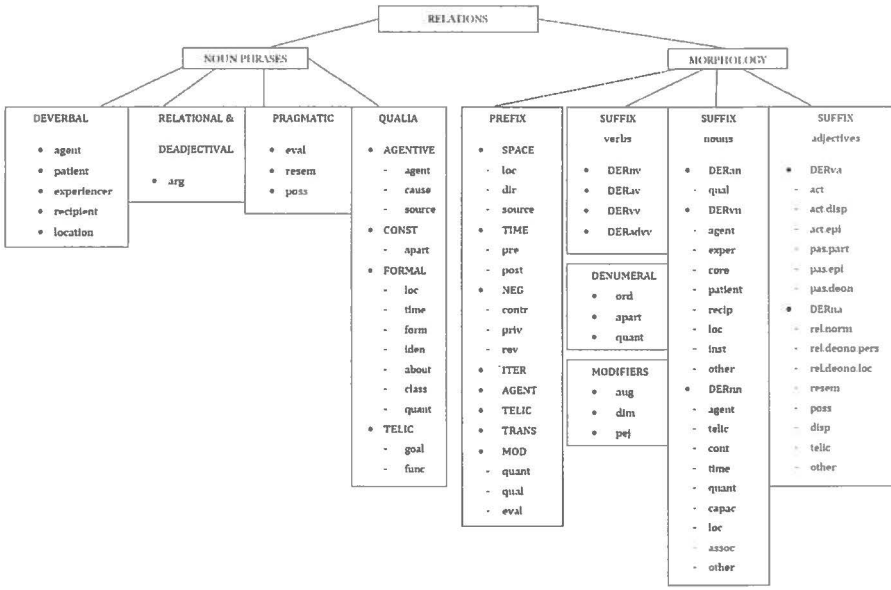
Figur 3. Fuld syntaktisk og semantisk annotation af NP'er.

I den semantiske annotation på NP-niveauet skelner vi mellem komplementer, der impliceres leksikalsk af kernesubstantivet, og adjunkter, der ikke er styret leksikalsk. Eksempler på komplementer ses i begge de ekspanderede NP'er i form af "patient"-rollen, mens "agent"-rollen i *KV plan* (NPet til venstre) og "loc"-rollerne i NPet til højre er eksempler på adjunkter.[3] Også adjunkt-relationerne har vi søgt at definere og strukturere i henhold til qualia, jf. skema 3 nedenfor (venstre del), hvorimod de af deverbale, deadjektivale eller relationelle kernesubstantiver bundne komplementer kategoriseres anderledes. Det skal bemærkes at "agent"-rollen i syntagmet til venstre adskiller sig fra "agent"-rollen i syntagmet til højre, idet konstituenten *of the two partners* fungerer som et

2. Bemærk, at determinativet udgør den øverste knude, såfremt der er et. Optræder der ikke et determinativ, er den nominale konstituent øverste knude.

3. Vi har ladet pilene i den syntaktiske annotation over teksten og den semantiske annotation under teksten følge hinanden i den forstand at fx PP'en *of the KV plan* både fungerer dependenmæssigt som "pobj.dobj" (pilen over teksten) og semantisk som "patient" (pilen under teksten). Her kunne man argumentere for at det ville være mere naturligt, at det var DP'en, der starter ved determinativet *the* (3), der fik tilskrevet "patient"-rollen, fordi det er dens referent der er argument for den semantiske launch-relation.

komplement der er leksikalsk styret af det (sekundære) kernesubstantiv *collaboration*. Endelig indikerer symbolet "#" efter "agent"-rollen i *KV plan* at *KV plan* har visse ligheder med en leksikalsk enhed (et frasalt kompositum). Skema 3's inventar af semantiske relationer skal altså forstås på den måde at vi overordnet sondrer mellem prædikative/relationelle kernesubstantiver på den ene side og absolutte kernesubstantiver på den anden. Er kernen prædikativ og deverbal, etableres der relationer som er identiske med det semantiske rolleinventar vi kender fra sætningsniveauet, mens hvis kernen er relationel eller deadjektival, etableres der blot en argumentrelation der annoteres "arg" uden videre forsøg på underopdeling. Hvis kernen er absolut trækker vi på et udvidet qualia-inventar eller relationerne oplistet i kassen "PRAGMATIC", som dog ikke kommenteres nærmere her.



Skema 3. Semantiske relationer anvendt ved annotation af NP'er og derivationer

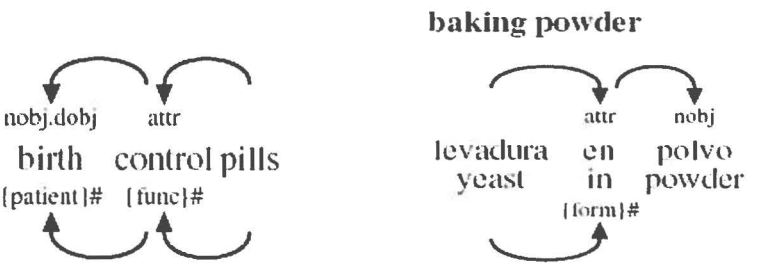
3. Qualia og morfologi

Højre side af skema 3 (MORPHOLOGY) viser de semantiske relationer der bruges ved annotation af derivationsmorfologi, mens kategorierne i venstre side (NOUN PHRASES), udover deres funktion ifm. opmærk-

ning af almindelige NP'er, jf. foregående afsnit, også anvendes ifm. komposita. For at forstå kategorierne og de enkelte relationer er det imidlertid nødvendigt først at forklare hvordan den morfologiske annotation fungerer i praksis.[4]

På det morfologiske niveau anvendes der to former for annotationer: En almindelig dependensannotation udvidet med semantisk annotation, a la den der anvendes i forbindelse med de ekspanderede NP'er i figur 3, og en operatorannotation, hvis der er tale om relationer inden for et ord, som fx mellem affix og base i en derivation, mellem frie morfemer i et sammenskrevet kompositum eller i forbindelse med kombinationer af derivation og komposition (Müller 2010).[5]

Nedenfor i figur 4 gives to eksempler på dependensannotationen på CDT's morfologiske niveau.



Figur 4. Dependensannotation af de frasale komposita birth control pills *(venstre) og* levadura en polvo *[bagepulver] (højre).*[6]

Kernesubstantivet i det frasale kompositum *birth control pills* er *pills*. Relationen mellem kerne og dependent er ikke en argumentrelation, men derimod en funktion vi kalder "attr" (attributive). Sådan forholder det

4. Inventaret af relationer inden for morfologi samt deres anvendelse er inspireret af Rainer (1999) og Varela et al. (1999). Forklaringer på og eksemplificeringer af de enkelte relationer findes på følgende adresse (CDT-manualen): http://code.google.com/p/copenhagen-dependency-treebank/
5. En base defineres jf. Bauer (1983) som "any form to which affixes of any kind can be added".
6. Den engelske oversættelse og glossering af eksemplet *levadura en polvo* (bagepulver) har ingen relevans i denne sammenhæng. Det samme gælder for figur 6.

sig, fordi kernesubstantivet er absolut, dvs. at det ikke er prædikativisk eller relationelt og derfor ikke er i stand til at selektere dependenten leksikalsk. "Attr"-funktionen specificeres af pilen der går fra *pills* til *control* over teksten. Den anden pil over teksten viser at den sekundære kerne *control* fungerer som leksikalsk styrende element i forhold til *birth*, som er et "nobj" (nominal object) med funktionen "dobj" (direct object).

Pilene under teksten viser den semantiske struktur. Dependenten aktiverer kernesubstantivets teliske rolle, her kaldet "func", hvilket bygger på den generelle antagelse at kernesubstantivets qualia kan aktiveres af forskellige modifikatorer, i dette tilfælde et andet NP.[7] Kernen i *birth control* er prædikativisk, og *birth* udfylder "patient"-rollen i kernesubstantivets argumentstruktur. Symbolet "#" efter rollebetegnelserne indikerer at der er tale om et frasalt kompositum. Annotation af *levadura en polvo* til højre følger samme mønster, blot er den spanske konstruktion naturligvis højreekspanderet. *Levadura* er kernen, der syntaktisk forbindes med ikke-kernen via en "attr"-funktion, mens *polvo* styres af præpositionen *en*. Semantisk aktiverer *en polvo* kernesubstantivets "form"-quale, og jf. "#"-symbolet er konstruktionen et frasalt kompositum.

Figur 5 viser to simple eksempler på hvordan operatorannotationen fungerer.

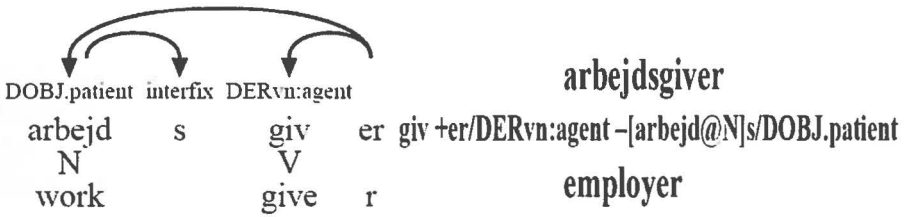
Krigsskib: skib –[krig]s/FUNC Træbord: bord –træ/CONST

Figur 5. Operatorannotation af de sammenskrevne komposita krigsskib *(venstre) og* træbord *(højre).*

Ved analysen af *krigsskib* annoteres først kernen *skib*, der efterfølges af et minustegn, som indikerer ikke-kernen/modifikatorens placering foran kernen. Selve modifikatorens leksikalske materiale står i kantede parenteser, og umiddelbart efter følger fugebogstavet. Endelig angives efter skråstregen den semantiske relation "FUNC" (functional/telic), der aktiveres af modifikatoren og således også følger idéen om qualia. Annotationen af *træbord* er analog, blot er der tale om den semantiske relation "CONST" (constitutive) i stedet for "FUNC" (functional/telic).

7. På det grundlag kunne man argumentere for, at de semantiske pile under teksten skulle vende modsat, men vi har i CDT valgt at lade de semantiske pile under teksten og de syntaktiske dependenspile over teksten følges ad.

For at illustrere at de to annotationsformer, dependensannotationen og operatorannotationen, blot er varianter over den samme grundtanke, har vi nedenfor i figur 6 stillet dem op overfor hinanden. Det er selvfølgelig vigtigt at understrege, at et dansk kompositum som *arbejdsgiver* i praksis altid vil blive annoteret i henhold til operatorannotationen i CDT, idet der er tale om relationer inden for ordgrænsen. Det vil sige, at dependensannotationen af *arbejdsgiver* er konstrueret med henblik på at vise den principielle overensstemmelse mellem de to annotationsformer, men den ville aldrig forekomme i den egentlige annotation. Kort sagt har vi de to annotationsformer af praktiske hensyn, nemlig for at kunne annotere relationer mellem ord (dependensannotationen) og relationer inden for ordgrænsen (operatorannotationen).



Figur 6. Morfologisk analyse af kompositummet arbejdsgiver *annoteret i henhold til dependensannotation (venstre) og operatorannotation (højre).*

Operatorannotationen skal forstås på følgende måde: Kernesubstantivet *giver* er derivationelt afledt. Operatoren "+er/DERvn:agent" indikerer at kernen er en agentnominalisering af verbet *give*, der er aktiveret af suffikset *-er*. Plustegnet angiver at det modificerende element er placeret efter kernen. Annotationen af ikke-kernen, dvs. " –[arbejd@N] s/DOBJ.patient", angiver, kort fortalt, at ikke-kernen er placeret foran kernen, jf. minustegnet, at det leksikalske materiale er et substantiv (N) efterfulgt af fuge-*s*, og at ikke-kernen svarer til et direkte objekt med den semantiske funktion "patient". Indikationen af ordklasse efter snabel-a er valgfri, men hvor der kunne være tvivl, angives ordklassen. Systemet, der angiver et elements position i forhold til kernen vha. plus- eller minustegn, er uafhængigt af elementets semantiske funktion, og samtidig også af om det pågældende element manifesterer sig som affiks eller frit morfem i et kompositum. Dette gør systemet meget fleksibelt og indebærer at det sandsynligvis ville kunne overføres til andre sprog end dem vi arbejder med i CDT.

Dependensannotationen til venstre i figur 6 indeholder faktisk fuldstændig den samme information. De to annotationer er altså indholdsmæssigt identiske, men vi har som sagt brug for begge typer for både at kunne tage hånd om relationer der forekommer mellem separate ord/"tokens" i en konstellation som i figur 4, og så relationer inden for ordgrænsen i form af derivation eller komposition som i figur 5 eller kombinationer heraf som i figur 6.

Desuden skal det nævnes at vi ikke bare analyserer og annoterer substantiver morfologisk, men også adjektiver og verber i henhold til de samme principper. I denne artikel har vi dog valgt udelukkende at fokusere på annotation af substantiver, da fremstillingen ellers ville blive for omfangsrig.

Med udgangspunkt i skema 3 kan man først og fremmest konstatere at qualia spiller en vigtig rolle som semantisk overordnet organisations- og analyseprincip i forbindelse med NP'er, uanset om der er tale om frasale eller sammenskrevne komposita, og dermed morfologiske strukturer, eller almindelige deskriptive NP'er. Hvad angår den egentlige derivationsmorfologi er qualia-strukturens indflydelse måske knap så tydelig, idet systemet af præfikser og suffikser afspejler en semantik som simpelthen er alt for kompleks til at den kan redegøres fyldestgørende for inden for qualia-rammen. Alligevel har vi ladet nogle af qualia-relationerne indgå også i derivationsmorfologien, hvor det har været muligt, fx i forbindelse med visse præfikser, som i *a-callar* (bringe til tavshed) "AGENT", eller substantivsuffikser, *puñal-ada* (knivstik) "DERnn:telic". Markeringen med "telic" bygger på den forståelse af eksemplet, at en knivs formål eller funktion kan være at stikke med den og dermed forårsage et knivstik.

4. Qualia og associative anaforer

Det er naturligt at gå fra N- og NP-analysen over til de nominale anaforer, og her viser qualia-strukturen sig særligt relevant ved de **associative anaforer**, dvs. de anaforer der udpeger en entitet eller et begreb der kan associeres med eller relateres til antecedenten, fx *en bil* [antecedent] ← *motoren, lygterne, chaufføren* [associative anaforer]. Qualia-strukturen kan jo læses som forskellige mulige prædikater i relation til leksikalske

størrelser, og som vi skal se, er det netop sådanne prædikater samt de mulige argumenter i sådanne prædikater der kan optræde som associative anaforer.

Selve termen "associativ anafor" er gængs ikke mindst i den franskprogede almenlingvistiske tradition, hvor denne anafortype er specielt grundigt undersøgt[8]. En anden term, som især har vundet hævd inden for den datalogiske litteratur, er **bridging anaphor**. "Bridging" defineres første gang hos Clark (1975) som de "implicatures" der gør modtager i stand til at "bridge the gap from what he knows to the intended Antecedent" (op.cit. 170), og denne gruppe omfatter normalt de utro koreferentielle anaforer[9] samt associative anaforer[10].

I det seneste årti er der også fremkommet en række konkrete forslag til annotation af de hyppigst forekommende anaforrelationer. Pladsen på dette sted tillader os ikke at gå i dybden hermed; for en nærmere omtale og bibliografiske henvisninger henviser vi til Korzen & Buch-Kromann (2011: 84). De fleste systemer benytter automatisk eller halvautomatisk opmærkning af mulige antecedenter og anaforer (NP- og pronomensyntagmer), hvorimod al anaforannotation i CDT foregår manuelt. CDT adskiller sig også fra andre systemer ved at inkludere samtlige typer af såvel koreferentielle som associative anaforrelationer.

Nedenfor giver vi nogle eksempler på associative anaforrelationer som aktiveres via antecedentens qualia-struktur[11]. I eksemplerne er antece-

8. Se bl.m.a. Kleiber (1997a/b, 2001), Schnedecker et al. (1994), Cornish (1999) og Lundquist (2000). Termen optræder formodentlig første gang hos Guillaume (1919: 162-163) og er også anvendt af Hawkins (1978: 123). For italiensk, se Korzen (2003, 2009).
9. En **utro anafor** er en anafor der er leksikalsk forskellig fra sin antecedent; en **tro anafor** er leksikalsk identisk hermed. Denne skelnen er således kun relevant ved NP-anaforer og ikke ved pronominelle anaforer eller nulformer.
10. Se fx Poesio, Vieira & Teufel (1997: 2), Vieira & Poesio (2000: 558) og Caselli (2009: 73). Hos Clark omfatter "bridging" dog såvel koreferentielle (tro og utro) som associative anaforer, hvorimod Poesio & Vieira (1998) bruger termen synonymt med "associativ".
11. Det er ikke nyt at sammenkæde associative anaforer og qualia-roller, jf. fx Bos, Buitelaar & Mineur (1995), Lundquist (2000), Henry & Bassac (2008), Caselli (2009) og Korzen (2003, 2009). Men det er nyt at qualia-struktur anvendes så konsekvent på flere forskellige analyseniveauer som tilfældet er i CDT.

denterne angivet i kursiv og anaforerne i kursiv og fed; et efterfølgende tal i parentes henviser til CDT-korpusset.

4.1. Formal quale – "ASSOC-FORMAL"

Som sagt i indledningen, udtrykker den formelle qualia-rolle information om elementer der "adskiller den denoterede entitet fra andre i det samme domæne":

> "That which distinguishes the object within a larger domain (shape, dimensionality, position, color)"

Og netop elementer som de i parentesen nævnte kan optræde som associative anaforer, jf. et typisk eksempel som

(1) Den skinke, der skal bruges til retten, skal ikke være alt for salt. Man kan ikke bruge de tynde skiver, som ligger færdigpakkede i en køledisk. *De* er for salte og for våde, og ***smagen*** er ikke god nok. (148)

I CDT opererer vi her med etiketten "ASSOC-FORMAL", som skal læses "association ved aktivering af antecedentens 'formal quale'".

4.2. Constitutive quale – "ASSOC-CONST"

De tre andre qualia-roller beskriver forskellige relationer som entiteten kan indgå i. The "constitutive quale" udtrykker som sagt ovenfor

> "the relation between an object and its constituents, or proper parts (material, weight, parts and component elements)",

så her er der tale om prædikater af typen *har som del, indeholder, er en del af,* og antecedenten er det ene argument i prædikatet, anaforen det andet som i det følgende eksempel:

(2) Ulykken fandt sted i spisetiden ca. 18.45 i aftes, kort efter at *El-Al-flyet* [...] lettede fra Amsterdams Schiphol lufthavn. "Jeg så *flyet* med ***næsen*** pegende nedad, ***venstre vinge*** op og ***højre vinge*** ned ovre bag den flade bygning. Det røg ud af ***motorerne.***[...]" siger øjenvidnet Peter de Neef. (1536)

I det citerede eksempel er anaforerne dele af antecedenten, men forholdet kan også være det omvendte[12]:

(3) DE BEERS CENTENARY åbnede den 8. september kontor i *Moskva*. Til stede var foruden De Beers-topfolk russiske politikere, diplomater og repræsentanter for ***landets*** diamantindustri og -handel. (431)[13]

I begge tilfælde taler vi om en "ASSOC-CONST"-relation, som skal læses "association ved aktivering af antecedentens 'constitutive quale'".

4.3. "ASSOC-AGENTIVE" og "ASSOC-TELIC"

Hvor de to ovenfor nævnte qualia udtrykker statiske beskrivelser, angiver de to sidste, AGENTIVE og TELIC, mere dynamiske relationer eller prædikater, som den denoterede entitet kan indgå i. Her kan såvel de prædikater som kan infereres i forbindelse med hhv. entitetens opståen eller tilblivelse (AGENTIVE) og dens funktion eller formål (TELIC), som de argumenter som kan infereres i sådanne prædikater, optræde som associative anaforer. Selve prædikatet aktiveres i tilfælde som (4)-(5):

(4) *Den biografaktuelle film "Krystalkraniets kongerige"* handler ikke primært om liv i rummet. ... Harrison Ford spiller som sædvanlig Indiana Jones og George Lucas har stået for ***produktionen***. [ASSOC-AGENTIVE] (http://www.liviuniverset.dk/?Nyheder_-_arkiv, fundet 10.10.2010)

(5) *Snyd – Et spil for 2 til 4 deltagere* hvor ***formålet*** [ASSOC-TELIC] er at fange spillere når de snyder ved at spille de forkerte kort ud. (www.gamedesire.com/la-dk,spil-snyt.html, fundet 13.10.2010)

Men her er det dog mindst lige så hyppigt, om ikke hyppigere, at anaforen er en (oftest førsteordens)entitet, der ikke udtrykker selve

12. Jf. også "The CONSTITUTIVE ... quale refers not only to the parts or material of an object, but defines, for an object, what that object is logically part of, if such a relation exists. The relation *part_of* allows for both abstractions". (Pustejovsky 1995: 98).

13. Det kan muligvis diskuteres om antecedenten snarere er *russiske*, men begge kan fungere som antecedent, hvilket kan påvises ved en test hvor den ene eller den anden udelades.

qualia-rollen, men en semantisk rolle heri[14]. I det AGENTIVE tilfælde er vi kun stødt på agent- og instrument-rollen, fx:

(6) Hver gang der laves *en ny film om Jesus*, retter troende nærmest pr. automatik de mest ukristelige beskyldninger mod ***instruktøren*** [ASSOC-AGENTIVE.AGENT]. Mange kristne optræder, som var de på lønningslisten i filmens pr-afdeling ...
(http://braadthomsen.tripod.com/id80.html, fundet 11.10.10)

Her opererer CDT altså med undertyper under de gængse associationer. For at "forstå" og inferere agenten *instruktøren* i (6) må vi først aktivere selve AGENTIVE-qualen.

I forbindelse med den teliske quale kan vi formentlig finde anaforer der udtrykker alle semantiske roller. Fortolkningen af den semantiske rolle kommer her meget an på det prædikat der infereres. Vi beder derfor annotatorerne om at angive dette prædikat i parentes, som det ses i de følgende eksempler:

(7) Ulykken fandt sted i spisetiden ca. 18.45 i aftes, kort efter at *El-Al-flyet* med tre besætningsmedlemmer og en kvindelig passager om bord lettede fra Amsterdams Schiphol lufthavn.
Piloten [ASSOC-TELIC.AGENT/(flyve)] meddelte pludselig kontroltårnet, at han havde motorproblemer og var tvunget til at vende om for at nødlande i Schiphol. (1536)[15]

(8) *To svendeprøver* blev bestået i august. ***Begge lærlinge*** [ASSOC-TELIC.PATIENT/(eksaminere)] er udlært hos Royal Copenhagen A/S Georg Jensen Sølvsmedie. (431)

Hvis man i (7) infererer prædikatet *flyve* som en funktion ved et fly og i (8) prædikatet *eksaminere* som en del af formålet med en (svende)prøve, er *piloten* agent i (7) og *lærlingene* patient i (8). Mange tilfælde tillader mere end én analyse; det gælder bl.a. det følgende:

14. "Første- og andenordensentiteter" er Lyons' (1977: 442ff) termer om hhv. konkrete individer/masser og handlinger/aktiviter/tilstande.

15. En anden mulighed er her at se *piloten* relateret til *tre besætningsmedlemmer* i en ASSOC-CONST-relation (set-membership). Men de to analyser udelukker ikke hinanden.

(9) De venlige filmselskaber og biografer sørger for at præsentere de nye film ved særlige pressevisninger som regel flere dage før premieren ... Hvis der undertiden skulle stå noget vås i *en film-anmeldelse* [sic!], skyldes det altså ikke tidnød, selv om det naturligvis er mest bekvemt for anmelderne, hvis ***læserne*** [ASSOC. TELIC.AGENT/(læse)] / [ASSOC.TELIC.RECIPIENT/(modtage)] tror det. (647).

Læserne vil være agent hvis man infererer prædikatet *læse* som (en del af) en filmanmeldelses funktion, hvorimod de vil være recipient hvis man infererer prædikatet *modtage*.

4.4. Det samlede billede

En samlet oversigt over CDT's associative anaforer baseret på qualia-rollerne – med eller uden undertyper bestemt af evt. argumenters semantiske roller ("semroles") – ser således ud:

assoc-QUALIA (± semrole subtype):

assoc-formal
assoc-const
assoc-agentive
    assoc-agentive.agent
    assoc-agentive.inst(rument)
assoc-telic
    assoc-telic.agent
    assoc-telic.patient
    assoc-telic.exper(iencer)
    assoc-telic.inst(rument)
    assoc-telic.rec(ipient)

Skema 4. Associative anaforrelationer baseret på qualia.

Qualia-struktur kan ikke forklare alle associative anaforer, og omvendt er qualia ikke nødvendig i alle tilfælde: Ved prædikative eller anden-ordensantecedenter kan associative anaforer udtrykke en semantisk rolle direkte tilknyttet antecedenten, jf. fx typer som

(10) *en operation* ← ***kirurgen*** [agent], ***patienten*** [patient]
en trafikulykke ← ***øjenvidnerne*** [experient]

osv.[16] Men der er ikke plads til at uddybe dette her. For et samlet billede af CDT's associative anaforer henviser vi til Buch-Kromann et al. (2010), hvorfra hele CDT-manualen kan downloades, og Korzen & Buch-Kromann (2011). I alle tilfælde går annotationspilene fra øverste knude i antecedent til øverste knude i anafor, jf. figur 7 nedenfor.

5. Qualia og diskurs

Aktiveringen af en qualia-rolle kan også finde sted i relationerne mellem hele tekstsegmenter. Også et helt tekstsegment som fx en periode vil kunne spille en bestemt rolle i relation til det foregående segment, en rolle som kan være – men ofte ikke er – ekspliciteret i form af en konnektor.[17] Tekstrelationer uden ekspliciterende konnektorer er typisk meget vage, og en analyse i semantiske termer (som jo i sig selv typisk tillader forskellige opfattelser eller nuancer) kan være særdeles vanskelig.

CDT opererer derfor med et system af over- og underrelationer, overtyper angives med versaler, undertyper med minuskler efter kolon, og den hierarkiske struktur tillader annotatorer at blive på overtypeniveauet i tilfælde hvor der er tvivl om undertypen. Der vil ofte være tale om relationer mellem perioder eller større enheder, men den grafiske udformning af annotationsrelationerne gør at vi også – modsat andre diskurstræbanker – kan angive relationer mellem sproglige segmenter internt i perioder. I annotationen angiver vi ekspliciterede konnektorer efter etiketten, som */for* i eks. (12) og */eller* i eks. (17) nedenfor, og ikke-eksplicitte men fortolkelige konnektorer sættes i parentes, som */for* i (13) og */fx* i (16).

Det er CDT's mål at etablere en enhedsannotation og en samlet analysemodel af de fire niveauer vi har talt om i denne artikel, og diskursannotationen opfatter vi på samme tid som en semantisk relations-

16. I disse typer er det dog ikke udelukket at operere med en qualia-struktur hos antecedenten, men vi har i CDT valgt at undlade dette "ekstra" analyseniveau, når det nu ikke var nødvendigt.
17. På samme måde kan et tekstsegment som fx en periode parafraseres eller "resumeres" i mindre leksikalske enheder som fx resumptive anaforer, hvis qualia-struktur kan analyseres helt parallelt med alle andre leksikalske enheder.

beskrivelse og en ekstension af den syntaktiske dependensanalyse til tekstniveauet, hvor vi knytter de enkelte tekstsegmenter sammen i en overordnet træstruktur.

Skreven tekst udvikler sig fra venstre mod højre, talt tekst i fysisk tid. I en konstruktion af typen

(11) X C Y

hvor C er en konnektor som forbinder to tekstsegmenter X og Y, vil vi normalt analysere det højrestående tekstsegment, Y, som dependent (dvs. modifikator) til det venstrestående og styrende segment, X[18]. Konnektoren er dependent til Y. I et tilfælde som

(12) Jeg tror Ole er syg. For han plejer altid at komme til alle møderne,

er X = *Jeg tror Ole er syg*, C = *For*, og Y = *han plejer altid at komme til alle møderne*. Her analyserer vi C + Y (dvs. hele sætning 2) som dependent/modifikator til X (sætning 1), og konnektoren *For* som dependent/modifikator til Y. En sådan analyse er problemfri i de tilfælde hvor der ikke er nogen konnektor:

(13) Jeg tror Ole er syg. Han plejer altid at komme til alle møderne,

og den passer godt ind i gængs (strukturalistisk) syntaksanalyse. På det semantiske niveau siger vi at konnektoren medvirker til at disambiguere Y, dvs. Y udvælger sin betydning på basis af konnektoren (hvis den er der). Annotationsmæssigt vil det se således ud:

Skema 5. CDT's analyse af diskursstrukturen tekstsegment – konnektor – tekstsegment[19].

18. Undtaget herfra er overskrifter o.l., som har CDT-etiketten "SCENE" styret af det følgende tekstsegment. Også indrømmende segmenter (med etiketten "CONC(ession)") kan gå forud for deres styrende segment.

19. For flere detaljer og relevante bibliografiske henvisninger henvises til Buch-Kromann & Korzen (2010: 130).

Annotationspilene går i alle tilfælde fra den øverste knude i det styrende segment til den øverste knude i det styrede (dependente) segment, dvs. i X og Y typisk mellem finitte verbaler, jf. analysen knyttet til Figur 1 ovenfor. I visse tilfælde kan pilene forbinde andre led, som vi skal se nedenfor. På denne måde kan diskurs-, syntaks- og morfologisk annotation samlet anskues som én lang dependensrelation gående fra det største sproglige segment, tekstsekvensen, der typisk består af en række forbundne perioder, til det mindste annoterede sproglige segment, morfemet.

På diskursniveauet i CDT fungerer de fire qualia-roller som relationsovertyper med i alt 12 undertyper, hvoraf vi vil eksemplificere visse nedenfor. For et fuldstændigt billede henvises igen til Buch-Kromann et al. (2010). I eksemplerne er det styrede (Y-)segment angivet i kursiv.

5.1. Formal-relationer

I formal-relationerne udtrykker Y-segmentet en beskrivelse, neutral eller positivt/negativt ladet, af X. Disse relationer adskiller sig fra de øvrige ved at det styrende X-segment ofte, jf. (14), men ikke nødvendigvis, jf. (15), er en førsteordensentitet:

FORMAL
    FORMAL:descr: neutral description
    FORMAL:eval: positive/negative evaluation

(14) Så lad os tage hjem til dig, sagde hun.
Hun var anderledes, end jeg troede. I taxien lagde hun sit hoved ind mod min skulder [...]. (602) [FORMAL:descr]

(15) Alligevel skal der kæmpes med næb og kløer, for at få meningsdannerne til at tage "det lette" alvorligt og behandle det som en fuldgyldig del af vort programudbud. *Det er for dårligt.* (102) [FORMAL:eval]

I (14) vil annotationspilen gå fra *hun* i X-segmentet til det finitte verbum *var* i Y, i (15) – på lidt mere "typisk" manér – fra det finitte verbal *skal* i X til det finitte verbal *er* i Y.

5.2. Constitutive-relationer

I constitutive-relationerne gælder det at Y-segmentet angiver dele af det styrende X-segment eller reformulerer dette. Denne qualia-rolle kan blive aktiveret af et *fx* eller et *eller*, ekspliciteret eller fortolkeligt.

CONST
- CONST:exem: exemplification
- CONST:rest: restatement

(16) På de fabrikker vi har i dag, omdannes råvarer til nye produkter. (fx) *Soyabønner bliver til salatolie. Træ bliver til papir.* (204) [CONST:exem/(fx)][20]

(17) Kommende fusions-partnere skal ikke jages væk af skræk for at blive underlagt andres kontrol. *Eller med Preben Nygaards ord: "Hvis Hafnia har en majoritet, kan det få andre til at blive væk."* (150) [CONST:rest/eller]

I (17) vil annotationspilen gå fra det finitte verbal *skal* i X til præpositionen *med* i Y.

5.3. Agentive-relationer

Agentive-rollen bliver på CDT-diskursniveauet fortolket som noget der i bred forstand *forårsager* noget andet, en årsag eller forklaring der typisk kan aktiveres af konnektorer som *for, fordi*:

AGENTIVE
- AGENTIVE:reas: reason
- AGENTIVE:sbj: subjective cause (argumentative)

(18) ATP-direktør Palle Simonsen tager ikke munden for fuld, når han skal kommentere den fremtidige investeringsstrategi. *Han er klar over, at enhver lille bevægelse fra ATP vil blive fulgt nøje af det*

20. En anden mulighed ville være at se eksemplificering som (en del af) beskrivelsen af X-segmentet, dvs. som en del af den formelle qualia-rolle. Men i CDT har vi valgt at henføre det til constitutive-rollen.

danske aktiemarked på grund af ATP's meget store betydning for aktiemarkedet. (781) [AGENTIVE:reas/(for)]

(19) Så vidt vi kan se fra hotellet, har ødelæggelserne hidtil været begrænsede og meget målrettede. Det kunne tyde på, at angrebet skal ses som en alvorlig forskrækkelse til Irak. *Hvis USA ville, kunne man sagtens have bombet det meste af byen i stykker.* [AGENTIVE:sbj/(for)]

I AGENTIVE:reas-typen (18) udgør Y-segmentet en (i afsenders øjne) mere eller mindre neutral årsag til X-segmentet, hvorimod afsender i AGENTIVE:sbj-typen (19) benytter Y-segmentet som et mere personligt/subjektivt argument i en diskussion. Det kan ofte være svært at skelne mellem de to undertyper, og i sådanne tilfælde kan annotatorerne forblive på overtypeniveauet AGENTIVE uden at specificere undertypen.

5.4. Telic-relationer

Telic-relationerne dækker på CDT-diskursniveauet formål og funktion samt heraf følgende, i lidt bredere forstand, konsekvens. De kan aktiveres af konnektorer som *for at, dermed, så*:

TELIC
- TELIC:goal: goal relation
- TELIC:cons.dir: direct physical consequence, result

(20) Og her spiller vedtægterne en afgørende rolle, siger Trangeled, der anbefaler, at man under alle omstændigheder allierer sig med en fagmand, når man skal etablere et bofællesskab. *Også for at sikre sig, at der ikke bliver skattetekniske problemer.* (1035) [TELIC:goal/for at]

(21) Om tre et halvt år må en betydelig del af DSB-færgerne søge nye farvande, når Storebæltsforbindelsen er en realitet. *Og dermed er en spændende historie ved at være historie.* (848) [TELIC:cons.dir/dermed]

I (20) udgør Y-segmentet formålet med X-segmentet, i (21) en direkte fysisk konsekvens eller følge af X[21].

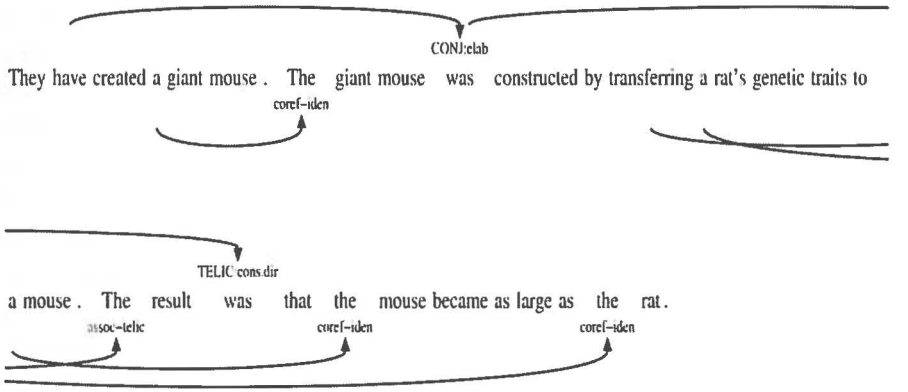
21. Parallelt hermed opererer CDT også med relationen "TELIC:cons.sbj", som bruges om den personlige og/eller argumentative deduktion. Som i AGENTIVE-tilfældet kan det også her være vanskeligt at skelne undertyperne fra hinanden.

5.5. Det samlede billede (diskurs og anaforer)

Figur 7 viser et eksempel på et uddrag af tre perioder af en anafor- og diskursannoteret tekst. Uforkortet lyder de tre perioder således:

> In this way, they have, for example, created a giant mouse. The giant mouse was constructed by transferring part of a rat's genetic "growth traits" to a mouse. The result was that the mouse became almost as large as the rat. (204)

I figur 7 er udeladt tekstelementer der ikke er direkte relevante for anafor- og diskursannotationen.



Figur 7. Diskurs- og anaforannotation af uddrag af tekst 204.

Som det ses, løber anaforpilene under tekstlinjen, og det korte tekstuddrag byder på tre koreferentielle og tro anaforer, *The giant mouse, the mouse* og *the rat*, alle annoteret med etiketten "coref-iden(tity)", samt en associativ telisk anafor, *The result*. Som altid forbinder anaforannotationspilene de øverste knuder i antecedent og anafor, dvs. her determinativerne, jf. også annotationen af *the launching of the KV plan i 1964* i figur 3.

Diskurspilene løber over tekstlinjen, og den teliske anafor, *The result*, ses at korrespondere perfekt med den teliske diskursrelation, "TELIC:cons. dir", mellem de to sidste perioder, jf. diskurspilen fra *(The giant mouse) was* til *(The result) was*. Diskursrelationen mellem de to første perioder er

angivet med etiketten "CONJ:elab", hvor overtypen står for "conjunct" og undertypen for "elaboration". Her tilføjer Y-segmentet detaljer til X-segmentet og videreudvikler det hermed, en umådelig hyppig funktion og relation i de fleste former for diskurs.

6. Afrunding

Vi mener altså at qualia-strukturen er et nyttigt instrument i forbindelse med en semantisk analyse af relationer på flere forskellige sproglige niveauer, klarest måske på NP- og anaforniveau, men også anvendeligt på syntaks- og diskursniveau, i syntaksen især i forbindelse med adverbialrelationerne. En gang imellem sker der sammenfald af qualia-aktiveringerne på de forskellige niveauer, som vi så i figur 7, men dette synes ikke umiddelbart at være et specielt hyppigt fænomen. Fremtidige analyser og statistiske beregninger vil kunne afgøre dette spørgsmål. Det har under alle omstændigheder vist sig frugtbart at lade qualia-strukturen være en form for styrende princip for såvel opbygningen som anvendelsen af vores semantiske inventar, netop i erkendelse af, som jo også titlen på vores artikel indikerer, at vi finder mange af de samme typer relationer på de forskellige annotationsniveauer, som CDT arbejder med. Samtidig står det dog også klart, at ikke alle semantiske relationer kan eller skal redegøres for i regi af qualia-strukturen, hvilket måske specielt annotationen af derivationsmorfologi og visse adverbialer tydeligt viser.

Henvisninger

Bauer, L. (1983). *English Word-formation.* Cambridge: Cambridge University Press.

Bos, J., P. Buitelaar & A.-M. Mineur (1995). Bridging as Coercive Accommodation, i *Workshop on Computational Logic for Natural Language Processing,* Edinburgh.

Buch-Kromann, M. (2006). Discontinuous Grammar. A dependency-based model of human parsing and language learning. Doctoral dissertation. Copenhagen: Copenhagen Business School.

Buch-Kromann, M., I. Korzen & H.H. Müller (2009). Uncovering the 'lost' structure of translations with parallel treebanks, i I.M. Mees, F. Alves & S. Göpferich, (red.), *Methodology, Technology and Innovation*

in Translation Process Research. Copenhagen Studies in Language 38, Copenhagen: Samfundslitteratur, 199-224.

Buch-Kromann, M. et al. (2010). *The inventory of linguistic relations used in the Copenhagen Dependency Treebanks.* Copenhagen Business School. http://copenhagen-dependency-treebank.googlecode.com/svn/trunk/manual/cdt-manual.pdf

Buch-Kromann, M. & I. Korzen (2010). The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks, i *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010,* 127-131.

Caselli, T. (2009). Using a Generative Lexicon Resource to Compute Bridging Anaphora in Italian. *Procesamiento del Lenguaje Natural* 42, 71-78.

Cornish, F. (1999). *Anaphora, Discourse, and Understanding. Evidence from English and French.* Oxford: Clarendon Press.

Clark, H.H. (1975). Bridging, i R.C. Schank & B.L. Nash-Webber (red.), *Theoretical issues in natural language processing.* New York: Association for Computing Machinery.

Guillaume, G. (1919). *Le problème de l'Article e sa solution dans la Langue française.* Paris: Librairie Hachette.

Hawkins, J.A. (1978) *Definiteness and Indefiniteness. A Study in Reference and Grammaticality Prediction.* London: Croom Helm.

Henry, P. & C. Bassac (2008). A toolkit for a Generative Lexicon, i *Fourth International Workshop on Generative Approaches to the Lexicon, Paris 2007.*

Kleiber, G.. (1997a) Des anaphores associatives méronymiques aux anaphores associatives locatives, *Verbum* XIX/1-2, 25-66.

Kleiber, G.. (1997b). Les anaphores associatives actantielles, *Scolia* 10, 89-120.

Kleiber, G. (2001). *L'anaphore associative.* Paris: Presses Universitaires de France.

Korzen, I. (2003). Anafora associativa: aspetti lessicali, testuali e contestuali, i N. Maraschio & T.P. Salani (red.), *Italia linguistica anno Mille, Italia linguistica anno Duemila.* Roma: Bulzoni, 593-607.

Korzen, I. (2009). Anafora associativa: ulteriori associazioni, i F. Venier (red.), *Tra pragmatica e linguistica testuale. Ricordando Maria-Elisabeth Conte. Gli argomenti umani* 13, 307-326.

Korzen, I. & M. Buch-Kromann (2011). Anaphoric relations in the

Copenhagen Dependency Treebanks, i S. Dipper & H. Zinsmeister (red.), *Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena. Bochumer Linguistische Arbeitsberichte* 3, 83-98.

Lyons, J. (1977). *Semantics, 1-2.* Cambridge: Cambridge University Press.

Lundquist, L. (2000). Translating Associative Anaphors. A Linguistic and Psycholinguistic Study of Translation from Danish into French, i I. Korzen & C. Marello (red.), *Argomenti per una linguistica della traduzione. Gli argomenti umani* 4, 111-129.

Moravcsik, J.M. (1975). Aitia as Generative Factor in Aristotle's Philosophy. *Dialogue* 14, 622-636.

Moravcsik, J.M. (1990). *Thought and Language.* London: Routledge.

Müller, H.H. (2010). Annotation of morphology and NP structure in the Copenhagen Dependency Treebanks, i M. Dickinson, K. Müürisep & M. Passaroti (red.), *The Ninth International Workshop on Treebanks and Linguistic Theories (TLT),* 151-162.

Poesio. M. &. R. Vieira (1998). A corpus-based investigation of definite description use. *Computational Linguistics* 24(2), 183-216.

Poesio, M., R. Vieira & S. Teufel (1997). Resolving Bridging References in Unrestricted Text, i Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, 1-6.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge (Mass.). London: MIT Press.

Rainer, F. (1999). La derivación adjectival, i I. Bosque. & V. Demonte (red.), *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa. Volume 3, chapter 70, 4595-4643.

Schnedecker, C. et al. (red.) (1994). *L'anaphore associative. (Aspects linguistiques, psycholinguistiques et automatiques).* Paris: Klincksieck.

Varela, S. & Martín García, J. (1999). La prefijación, i I. Bosque. & V. Demonte (red.), *Gramática Descriptiva de la Lengua Española.* Madrid: Espasa, volume 3, chapter 76, 4993-5040.

Vieira. R. & Poesio, M. (2000). An Empirically-Based System for Processing Definite Descriptions. *Computational Linguistics* 26(4), 539-593.