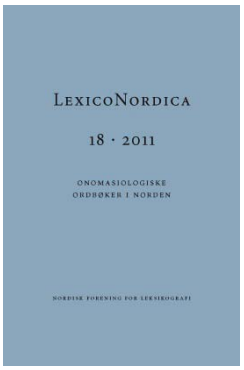
LexicoNordica

Titel: Ontologi, begreppssystem och WordNet

Forfatter: Klaas Ruppel

Kilde: LexicoNordica 18, 2011, s. 157-182

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



Ontologi, begreppssystem och WordNet

Klaas Ruppel

In this article I try to point out similarities and differences between ontologies, dictionaries of synonyms and word nets. The focus lies on ontologies in relation to the other two types. The universal applicability and language-independentness claimed for top hierarchies of ontologies is also discussed. As examples I look at two ontologies (Suggested Upper Merged Ontology and EuroWordNet Top Ontology), a dictionary of synonyms (Lagercrantz 1939) and the Princeton WordNet and some of its offsprings (EuroWordNet, DanNet). I emphasize particularly the content and the weight of ontological work and do not discuss datatechnology.

I denna artikel diskuterar jag likheter och skillnader mellan ontologier, synonymordböcker och ordnät. Som exempel använder jag två ontologier (Suggested Upper Merged Ontology och EuroWordNet Top Ontology), en samisk synonomymordbok samt Princeton WordNet och två andra ordnät som haft WordNet som incitament (EuroWordNet, DanNet).

1. Ontologin

Ontologi är ett begrepp inom teoretisk filosofi. Ordet är bildat av de grekiska orden för 'varande' och för 'lära'. Ontologi är alltså läran om varandet eller om tillvaron. Ordet uppträder första gången i början av 1600-talet (Lorhard 1606), trots att man hade funderat mycket över varandet redan under antiken, då främst inom metafysiken. Ända från början har kategorisering spelat en stor roll vid behandling av varandet.

Inom modern informationsteknik är begreppet ”ontologi” mycket mer praktiskt till sitt innehåll. Det är ”en explicit specifikation av en delad konceptualisering.” (Gruber 1993:199). Med hjälp av en ontologi presenteras alltså begreppsapparaten inom ett ämnesområde på ett sådant vis att den kan utnyttjas för informationstekniska syften.

I en ontologi är idealtillståndet att varje entitet[1] har en och endast en entitet som närmast övergripande entitet. I de två ontologierna som jag granskar här är detta dock inte helt och hållet genomfört. I den ena, Suggested Upper Merged Ontology (SUMO), har vissa av entiteterna fler än en övre entitet. Den andra, toppontologin för EuroWordNet, är å sin sida bara delvis hierarkisk, eftersom entiteterna definieras också via komponentanalys.

I denna artikel ska jag fokusera på den högsta delen av ontologierna, på *toppontologierna* (***Upper** Ontology, **Top** Ontology*), dvs. på den del av ontologierna som är planerade att vara universella och språkoberoende.

1.1. Suggested Upper Merged Ontology

SUMO är en ontologi som numera upprätthålls av Adam Pease, som arbetar på ett företag som producerar webbapplikationer. Anledningen till att jag i detta sammanhang granskar just den ontologin, är att den har tillämpats på WordNet (Niles & Pease 2003; Index of WordNetMappings). SUMO-systemet är uppdelat i tre delar. Den lägsta nivån utgörs av terminologier inom vissa fackområden. Den högsta nivån, själva SUMO, utgör en toppontologi baserad på logiska entiteter. Toppontologin samlar begreppen på de lägre nivåerna hierarkiskt och för till sist ihop dem till en enda topp. Mellan den översta och den nedersta nivån befinner sig det

1 För tydlighets skull använder jag termen *entitet* när jag talar om de hierarkiska noderna i ontologierna. Motsvarande term i begreppsordböcker och ordnät är *begrepp*.

hierarkiska mellanskiktet som alltså kopplar samman toppen och botten (se SUMO).

De terminologier som finns samlade på den nedersta nivån utgör ett mycket begränsat urval av verklighetens alla områden (t.ex. *Communications, Economy, Military, World Airports*). SUMO påstås emellertid vara oberoende av språk och kultur: ”SUMO is language independent.” Och vidare: ”… there is no deep-seated linguistic or cultural bias …” (SUMOFAQ). Hur ser då den översta, allmängiltiga delen av SUMO ut? Den högsta noden i ontologin är ”Entity”, som förgrenar sig i de två noderna ”Physical” och ”Abstract”. Dessa delar i sin tur upp sig på många egna entiteter (figur 1):

Figur 1: SUMO:s högsta noder.

SUMO-systemets alla logiska entiteter och deras inbördes hierarkiska relationer finns på internet (SUMOclasses). När man granskar entiteterna ser man lätt att påståendet om universell giltighet för toppontologin inte håller streck. Argumentet verkar snarare gälla för kommersiell reklam än för vetenskap. Jag ska ge några exempel. Ontologin innehåller bl.a. följande hypernymkedjor (uppifrån och ner) (figur 2):

0 Entity
1 Abstract
2 Quantity
3 PhysicalQuantity
4 ConstantQuantity

```
5            TimeMeasure
6              TimePosition
7                TimeInterval
8                  Month
9                    Januar … December
```

Figur 2: SUMO:s hypernymkedja för månader.

Entiteten ”månad” förgrenar sig alltså enligt det system som används i västvärlden med månaderna januari till december. Några andra månadssystem beaktas inte i ontologin. Likaså finns det utöver ”liter” bara anglosaxiska mått under noden ”VolumeMeasure”, och under ”CurrencyMeasure” förekommer bara dollar och euro. Exemplen av det här slaget är många. I praktiken är det på inget vis fatalt, för de saknade måtten, valutaenheterna osv. kan enkelt läggas till i samband med att nya terminologier läggs in i SUMO. Ur teoretisk synvinkel är det dock ytterst tveksamt om månaders namn och enskilda valutor ska ingå i en toppontologi som är avsedd att vara allmängiltig.

Vad gäller den påstådda allmängiltigheten är följande exempel mycket allvarligare på grund av att den ifrågasatta entiteten befinner sig betydligt högre upp i hierarkin, på kortare avstånd från toppnoden ”Entity” (figur 3):

```
0  Entity
1    Physical
2      Process
3        Motion
4          BodyMotion
5            Ambulating
```

Figur 3: SUMO:s hypernymkedja för ’förflytta sig till fots’.

Den nedersta entiteten ”Ambulating” ’det att förflytta sig till fots’ förgrenar sig vidare på entiteterna ”Running” och ”Walking”. Om springande och promenerande över huvud taget ska anses höra till de entiteter som ska ingå i en toppontologi så borde flera andra förflyttningssätt som exempelvis krypande och galopperande läggas till för att täcka in benförsedda djurs olika sätt att ta sig fram. Begreppen kan förvisso ordna in sig just på det sätt som anges i SUMO på många språk, men exempelvis i samiska ser det annorlunda ut. Där återfinns skillnaden mellan tvåbentas och fyrbentas varelsers förflyttning på begreppsnivå. På nordsamiska heter det exempelvis *olmmoš vázzii* ’människan går’ men *boazu oahkuu* ’renen går’. Om man vill att ontologin ska lämpa sig som översta hierarki också för de samiska språken borde antingen entiteterna ”tvåbenta varelsers förflyttning” och ”fyrbenta varelsers förflyttning” läggas till under noden ”BodyMotion” eller så borde entiteten ”Ambulating” tas bort ur toppontologin.

Sammantaget verkar SUMO förgrena sig ner till alltför låg nivå för att utgöra en toppontologi. Många av de nedre noderna måste snarare anses representera språk- och kulturbundna begrepp än logiska entiteter. Det här är en återspegling av de terminologier som är knutna till SUMO. De ovan nämnda valutabenämningarna ansluter sig till terminologierna *Economy* och *Finance*, och måttsorden ingår i terminologin *North American Industrial Classification System*. Eftersom ontologin har byggts upp genom att vissa terminologier kombinerats är resultatet i sig inte förvånande, men däremot förvånar påståendet om att ontologin skulle vara universellt giltig.

1.2. EuroWordNet Top Ontology

EuroWordNet är ett projekt som syftade till att översätta det engelskspråkiga WordNet till andra språk. Det ursprungliga EuroWordNet-projektet, som omfattade engelska, spanska, italienska

och nederländska, avslutades för över tio år sedan. Basen i EuroWordNet utgörs av en toppontologi, *Top Ontology* (illustration 1; Rodríguez 1998:139; Vossen 1998:22).

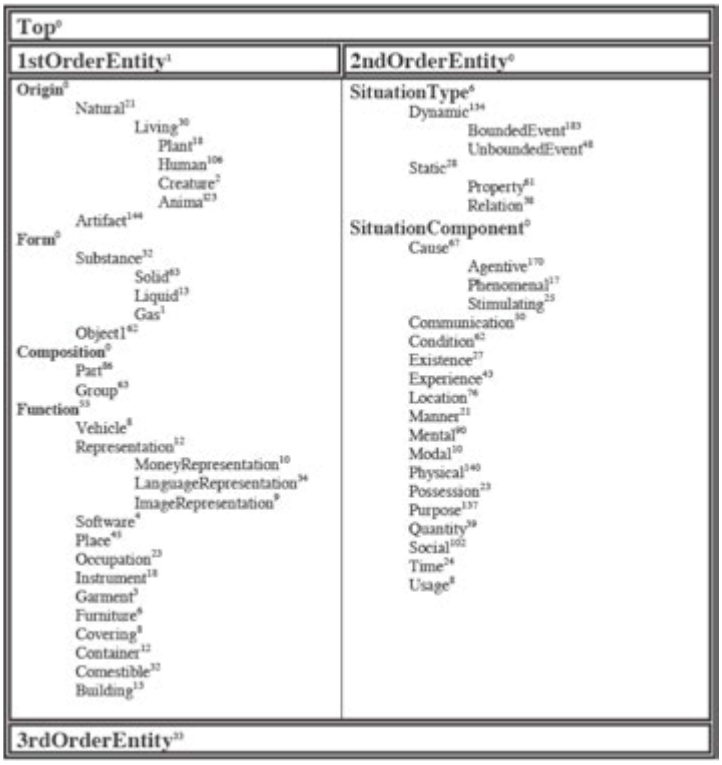


Illustration 1: Toppontologin i EuroWordNet.

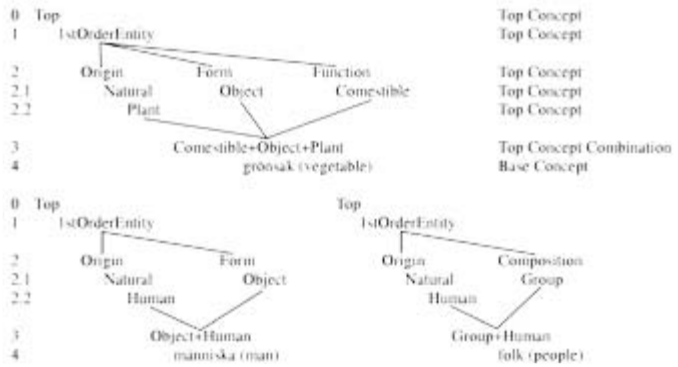
Ontologin består av tre typer av entiteter (Rodríguez 1998:136; Vossen 1998:21):

1stOrderEntity: en konkret entitet som kan uppfattas av sinnen och som har en plats i tid och rum
2ndOrderEntity: en statisk eller dynamisk situation som inte kan uppfattas som ett fysiskt objekt
3rdOrderEntity: ett påstående som är oberoende av tid och rum

Till entiteter av den första ordningen hör konkreta substantiv, till entiteter av andra ordningen substantiv, adjektiv och verb och till entiteter av den tredje ordningen abstrakta substantiv. (Rodríguez 1998:137; Vossen 1998:21) Tills vidare har andra ordklasser än substantiv och verb inte tagits med i ontologin. (Rodríguez 1998:148–149; Vossen 1998:31, Christiansson & Zimmerman 2006:35)

De 63 entiteterna i illustration 1 ingår i en mycket platt hierarki (jfr t.ex. Viberg 2002:136–138). De benämns *toppbegrepp* (*Top Concepts*, TC) och beskrivs som semantiska grundkomponenter (”fundamental semantic distinctions”). Nedanför denna hierarkiska nivå finns ett stort antal *basbegrepp* (*Base Concepts*, BC) som bestäms utifrån de semantiska grundkomponenter som visas i illustration 1. Indexsiffrorna anger hur många basbegrepp som definierats utifrån ett givet toppbegrepp. Basbegreppen kan definieras utifrån ett eller flera toppbegrepp. (En förteckning över alla toppbegrepp finns i Vossen (1998:34–44).) Den ontologiska strukturen är alltså en helt annan än i SUMO.

I figur 4 ges några exempel ur EuroWordNet. På raderna 0–2 i hierarkierna anges vilka semantiska grundkomponenter som använts vid definitionen och vilken deras plats är i hierarkin. På rad 3 syns komponentkombinationen och på rad 4 det basbegrepp som definieras genom kombinationen.



Figur 4: Definitionerna för basbegreppen ”grönsak”, ”människa” och ”folk” i EuroWordNet:s Top Ontology.

Toppontologin utgör EuroWordNets språkoberoende översta hierarkinivå, och med hjälp av den kan begrepp i de ingående språken kopplas till varandra. Ontologin är tudelad och består dels av entitetshierarkin, alltså de semantiska grundkomponenterna, dels av basbegreppen som erhålls genom kombination av grundkomponenterna. I toppontologin förenas alltså strukturalism och ontologi. (För mer information se Rodríguez (1998), Vossen (1998) och Viberg (2002).)

EuroWordNets toppontologi håller sig på ett mycket mer allmänt plan än SUMO, vilket troligen beror på att den från början är gjord för att vara språkoberoende och universell. Den är inte sammanställd utifrån ett smärre antal terminologier utan är avsedd att täcka språkets hela begreppssystem. Ändå påpekar författarna att man också i en ontologi på så här generell nivån konfronteras med språkliga skillnader. Som exempel nämns att det holländska språket inte har ett allmänt begrepp för ”container” (Rodríguez 1998:150; Vossen 1998:31).

1.3. En allmän finsk ontologi

I detta sammanhang bör det nämnas att det också i Finland bedrivs ontologisk forskning. YSO (*Yleinen suomalainen ontologia* ’allmän finsk ontologi’) baserar sig på Nationalbibliotekets ämnesordsindex. Det handlar sålunda om ett ontologiskt projekt av stor omfattning. Den gemensamma, allmänna toppontologin YSO utgör tak för 62 olika ämnesordsindex. Den innehåller för närvarande ca 23 000 begrepp och utgör en del av ett större projekt som syftar till att bygga upp ett nationellt semantiskt nätverk (FinnONTO). Det inbegriper också KulttuuriSampo, som tillhandahåller finländsk kultur via den semantiska webben.

1.4. Ontologier – ett informationstekniskt verktyg

De moderna ontologierna är inte utvecklade för människor utan för maskiner. De behöver därför inte följa intuitivt tänkande. Illustrationerna över SUMO och Top Ontology underlättar visualisering av ontologiernas struktur, men tekniskt består de av logiska fraser skrivna på programmeringsspråk.

Syftet med en ontologi är att göra det möjligt för datorn att ”förstå” betydelser. En dator kan skilja mellan olika teckensträngar. Den identifierar alltså till exempel ”tomat” och ”människa” som två olika saker och vet också att strängen ”potatis” är en tredje sak. Men utan ontologi kan den inte dra slutsatsen att tomat och potatis har mycket gemensamt (de är växter) och att en människa är något helt annat, även om människan har en relation till de här växterna (de odlas och de äts).

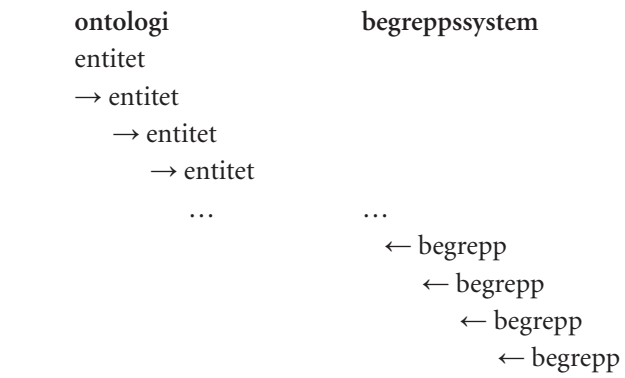
En ontologi försöker hitta likheter.

2. Begreppssystem för ett helt språk (tesaurus)

Begreppsordböcker är hierarkiskt strukturerade tesaurusar över hela språket. De skiljer sig från ontologier och datatekniskt skapade ordnät just genom det faktum att de presenterar hela lexikonet, inklusive slutna ordklasser som pronomen, konjunktioner, pre- och postpositioner etc. De grammatiska elementen har sällan en betydelse som kan fastställas utan kontext och de är därmed självfallet svåra att passa in i system där betydelsen eller begreppet (t.ex. ”huvudbonad”) har sin plats i systemet på basis av sin egen lexikala betydelse. Begrepp som ”när” eller ”att” är svåra att definiera lexikalt. I gränslandet mellan slutna och öppna ordklasser befinner sig pronomen, som å ena sidan utgör en sluten ordklass, å andra sidan har lexikal betydelse (t.ex. ”du”). Begreppsordböcker har en lång historia. Exempel som kan nämnas är Roget

(1856), Dornseiff (1933–1940), Wehrle (1942) och Nagy (1978).

Medan en ontologi byggs upp deduktivt skapas begreppssystemet i en begreppsordbok induktivt. Begreppen i ett språk samlas i synonymgrupper och dessa förses med lämpliga överbegrepp osv. (figur 5).



Figur 5: Ontologiska entiteter och språkliga begrepp.

På någon punkt möter den deduktiva kategoriseringen uppifrån och ner den induktiva kategoriseringen nerifrån upp. Den senare är kopplad till begrepp i språket, den grupperar språkliga begrepp och den är språk- och kulturbunden. Den deduktiva toppontologin använder inte begrepp knutna till språk utan följer logikens lagar. Kategoriseringsenheten är entiteten. Skillnaden framgår direkt av hur begrepp och entiteter namnges. Typiska (språkberoende) begrepp är t.ex.:

byggnad, hus, stenhus; mat, gröt, morgongröt
Gebäude, Haus, Steinbau, Steinhaus; Speise, Brei, Morgenbrei
building, house, stone house; food, porridge

Typiska (logiska) entiteter i SUMO är t.ex.:

Process, IntentionalProcess, SocialInteraction, Contest

EuroWordNets toppontologiska kategorier (topp- och basbegreppen) används – beroende på ontologins strukturalistiska karaktär – som semantiska komponenter, men de är ontologiska element och därmed egentligen entiteter.

Entiteter kan översättas utan att referenten ändras. Däremot ändras referenten ofta i större eller mindre utsträckning om språkliga begrepp översätts från ett språk till ett annat. Organisationen av begrepp i över- och underbegrepp är språk- och kulturspecifik.

2.1. En begreppsordbok för de samiska språken

Jag ska nu undersöka begreppsordbokens natur utifrån en mindre känd språkgrupp, eftersom det på det sättet går bättre att få fram skillnaderna mellan begreppssystem å ena sidan och ontologier och ordnät å andra sidan. År 1939 publicerade Eliel Lagercrantz sitt storverk *Lappischer Wortschatz*, som innehåller ungefär 10 000 uppslagsord och omfattar 1036 sidor. Språken som beskrivs är i huvudsak umesamiska, pitesamiska, lulesamiska och nordsamiska eller de västsamiska språken. Ordboksartiklarna är etymologiskt ordnade så att de olika språkens motsvarigheter har kunnat redovisas tillsammans. Lemmat för en artikel är ett oavlett grundord. Avledningar och sammansättningar återfinns som sublemman, t.ex.:[2]

753. pites. *tjuoj´vat,* lules. *tjuojvvat,* nords. *čuovjat* 'blå, grå'
753.4 pites. *tjuojvuk,* lules. *tjuojvuk,* nords. *čuoivvat* 'grå ren'

Figur 6: Utdrag ur artikel 753 (Lagercrantz 1939:105).

Lagercrantz begreppssystem *System der Bedeutungen* finns beskrivet på sidorna 1037 till 1182. Begreppssystemet togs fram in-

2 För tydlighetens skull presenterar jag bara en del av artikeln och jag använder också modern samisk ortografi.

duktivt. Lagercrantz tog ut alla betydelsangivelser för de samiska orden i sin ordbok och ordnade dem i begreppsgrupper. Det här innebär alltså att begreppssystemet omfattar hela den västsamiska språkgruppen, och inte bara ett språk.

På den översta nivån delas begreppen in i tre grupper (Lagercrantz 1939:1037). Till de metodologiska begreppen (I) hänför Lagercrantz kunskapsteoretiska, perceptionella och språkliga fenomen. Hit hör sinnesförnimmelser och högre varseblivningsområden såsom rum, rörelse, tid och form. Vidare ingår psykologiska begrepp och som fjärde grupp gester och språkstrukturella element såsom konjunktioner, räkneord och pronomen.

När de typerna av begrepp har sammanförts till en grupp är det naturligt att dela in resten av begreppsapparaten i två grupper, kulturella begrepp (II) och begrepp som hänför sig till naturen (III). De kulturella begreppen delas vidare in i samhälleliga och ekonomiska. Naturbegreppen är indelade i undergrupperna organiska och icke-organiska.

Egentligen kan man redan av de tre nivåerna dra slutsatsen att de begrepp som är föremål för klassifikation tillhör ett kultursamhälle med traditionella näringar. Ända upp till den högsta hierarkiska nivån bygger systemet alltså på induktion, på de villkor som gäller för de beskrivna språken och dess kulturgemenskap, även om kategorin metodologiska begrepp som helhet naturligtvis är mer teoretisk till sin natur.

Det var redan ovan tal om en semantisk specialitet hos samiska, nämligen att tvåbenta och fyrbenta varelsers gående hålls isär. Det samiska ordförrådet har också andra särdrag, som återspeglar sig i begreppssystemet.

Många språk har ett ord för (djuret) ”ren” och möjligen därtill ord för renar av honkön och av hankön. I de samiska språken finns det över hundra olika renbenämningar. Renar kan särskiljas enligt en stor mängd olika kriterier, vilket visas i figur 7 (inom parentes

anges antalet begrepp som är inordnade under varje överordnat begrepp, Lagercrantz (1939:1141–1143)):

ren, allmänt (24)
rentjur, renoxe (16)
renko (12)
renkalv (9)
renbeteckningar enligt ålder (25)
renbeteckningar enligt färg (27)
renbeteckningar enligt hornform (13)
renbeteckningar enligt öronmärkning (6)
renhjord, rajd (20)

Figur 7: Synonymgrupperna för 'ren' hos Lagercrantz 1939.

I den samiska kulturen har rennäringen varit en nyckelfaktor som hela storfamiljens välmåga varit beroende av.

Mycket rika är också begreppssystemen för vinter, is och snö (ca 200 begrepp) och för släktskapsförhållanden (ca 150). På de områden som samerna bebor är vintern lång, och ett överlevnadsvillkor är att man är förtrogen med och kan tolka vinternaturen.

Å andra sidan skiljer man inte på samiska mellan växande träd och trä som ämne (liksom inte heller på finska): det nordsamiska ordet *muorra* (och dess finska ekvivalent *puu*) motsvaras i svenska och tyska av två olika begrepp "träd", "Baum" och "trä", "Holz".

En begreppsordbok beskriver en språkgemenskaps förhållande till omvärlden. Begreppssystemet är kopplat till språk- och kulturgemenskapen, och antalet begrepp är vanligen stort på just de områden som är särskilt viktiga för samhället ekonomiskt, psykologiskt, osv. Begreppssystem skiljer sig mest för olika språk just på de punkter där kulturerna skiljer sig.

Induktiva, språkbundna begreppssystem framhäver skillnader.

2.2. Gråzonen mellan ontologi och begreppssystem

Mellan ontologierna som opererar på hög hierarkisk nivå med logiska entiteter och de språkbundna hypernymkedjorna som byggs upp nerifrån finns det en gråzon. Inom SUMO kallas denna zon Mid-Level Ontology. Via den kopplas Upper Ontology, eller den egentliga ontologin samman med terminologier för olika fackområden. Med tanke på helhetsbeskrivningen är kanske just denna gråzon det mest intressanta området. Här möts entiteter och begrepp, logik och intuition, deduktion och induktion, det språkoberoende och det språkberoende. I WordNet (se avsnitt 3) kopplas den övre och den nedre delen av hierarkin, alltså ontologi och begreppssystem, samman så att begreppen längst nere ingår i hypernymkedjor som når ända upp till toppen av ontologin. Det varierar sannolikt från kedja till kedja och från språk till språk var ontologi och begreppssystem möts.

Inom de biologiska taxonomierna finns det däremot nästan inga gråzoner. *Taxonomi* är en biologisk term och syftar på en hierarkisk vetenskaplig klassificering av växter och djur. Också denna term har fått en annan betydelse inom informationsvetenskapen, där en taxonomi avser det hierarkiska systemet för vilken som helst fackområdesterminologi. Här håller jag dock isär termerna *taxonomi*, som avser ett hierarkiskt ordnat system, och *terminologi,* som inte nödvändigtvis avser ett hierarkiskt system. Taxonomier kan å ena sidan anses vara specialfall av begreppssystem, där objektet för systematiseringen inte är hela språket utan ett visst områdes termer. Å andra sidan påminner taxonomier och terminologier också om ontologiska system eftersom termer och taxoner[3] – till skillnad från allmänspråkliga begrepp – är noga definierade.

I fråga om både taxonomier och specialområdesterminologier finns det andra typer av problem. För det första kan nya forsk-

3 En taxon är en taxonomisk enhet (t.ex. ”änder”, ”andfåglar” och ”fåglar” är taxoner).

ningsresultat leda till ändringar i en taxonomi. En förändring sker inte över en natt utan är resultatet av vetenskapliga resonemang. Av detta följer att flera klassificeringar kan existera sida vid sida under en tid, och att den nya taxonomin först småningom vinner insteg. Det mest berömda exemplet på en taxonomi och dess variation torde vara det klassifikationssystem för växter som skapades av Linné (1753) och som sedermera undergått otaliga förändringar och preciseringar.

Terminologier och taxonomier är kopplade till det område där de tillämpas. Morfologin inom filosofi, biologi, medicin osv. är något helt annat än inom lingvistik, trots att det i samtliga fall är fråga om en formlära.

Ett närbesläktat problem är att ett och samma begrepp kan klassificeras i enlighet med olika principer inom olika områden. ”Folkvisa” kan t.ex. vara ett begrepp både inom musik och inom traditionsforskning (hypernymkedjorna i figur 8 är fiktiva):

musik tradition
← musikslag musiktradition →
← folkmusik sång →
← folkvisa →

Figur 8: ”Folkvisa” med två olika överbegrepp.

Om bägge terminologierna samlas under samma ontologiska tak måste man antingen räkna med två olika begrepp ”folkvisa” eller så måste man tillåta att ett begrepp har mer än ett överbegrepp, här ovan ”folkmusik” och ”sång”. I det danska ordnätet DanNet har denna induktiva dikotomi behandlats mycket elegant. Fastän nätet sannolikt innehåller två olika begrepp ”purjolök” ser det för användaren ut som ett enda begrepp med två olika överbegrepp (Pedersen 2010:170–). Jfr DanNet ”porre”.

3. Ordnät

3.1. WordNet

WordNet är ett begreppssystem sammanställt med hjälp av informationstekniska metoder. I likhet med begreppsordböckerna sammanför WordNet synonymerna i synonymgrupper (i WordNet-termer i *synset*) och kopplar dem till överordnade begrepp i hierarkin.

WordNet koncentrerar sig på ord som tillhör de öppna ordklasserna och har lexikal betydelse. Beroende på vilken helhet som ska täckas ingår vanligen bara substantiv, adjektiv, verb och adverb.

WordNet har utvecklat vid Princeton University och föreligger för närvarande i version 3.1. Ordnätet har tillämpats på många olika sätt på andra språk, främst genom översättning. Att WordNet översätts implicerar ett antagande om att det skulle handla om ett universellt system som låter sig översättas. Också EuroWordNet baserar sig på WordNet, men för EuroWordNet har man utvecklat en egen ontologisk högsta nivå, *Top Ontology*. Frågan om i vilken mån det är meningsfullt att översätta WordNet har debatterats flitigt (t.ex. Anderson m.fl. 2010, Lindén & Carlson, 2010). Resultatet av den nyligen gjorda översättningen av WordNet till finska diskuteras av Martola (2011).

Namnet till trots är det inte ord som utgör beskrivningsenheter i WordNet utan begrepp och entiteter. Eftersom ett ord i ett språk kan stå för olika begrepp, dvs. ett ord kan ha flera betydelser, leder detta till att ett och samma ord kan representera flera olika begrepp i ordnätet. Varje begrepp ingår i en egen hypernymkedja. I tryckta ordböcker måste man i sådana fall använda sig av hänvisningar. Som elektronisk källa kan WordNet göras användarvänligare tack vare klickbara länkar.

3.2. Lagercrantz begreppssystem jämfört med WordNet

I det följande ska jag visa på skillnaderna mellan begreppssystemen i Lagercrantz och i WordNet med hjälp av två exempel. Det ena är det astrakta begreppet ”skönhet”, det andra är det för samisk kultur mycket centrala begreppet ”ren”. Lagercrantz klassifierar begreppet ”skönhet” enligt figur 9.

metodologiska begrepp
  sinnesförnimmelser
    modalitet
      allmänbegrepp för modalitet och orsak
        substantiv för modalitet och orsak
          skönhet

Figur 9: Hypernymkedjan för begreppet ”skönhet” i Lagercrantz.

I WordNet ser klassificeringen ut som i figur 10[4].

entitet
  abstraktion, abstrakt entitet
    attribut
      kvalitet
        utseende, visuell aspekt
          skönhet

Figur 10: Hypernymkedjan för begreppet ”skönhet” i WordNet.

De båda klassificeringarna har lika många nivåer men kategorierna ser olika ut. Vissa kategorier är dock jämförbara. Kategorierna ”sinnesförnimmelser” hos Lagercrantz och ”utseende, visuell

4 WordNets hierarki presenteras här vänd upp och ner i jämförelse med WordNets egen presentation, för att de båda systemen ska vara lättare att jämföra.

aspekt" i WordNet motsvarar varandra i stort sett, men i de hierarkiska systemen befinner de sig i motsatta ändar: hos Lagercrantz högt uppe och i WordNet lågt nere som närmast överordnade begrepp till "skönhet".

Tvärtom förhåller det sig med den nivå som hänför sig till grammatiska lösningar. I Lagercrantz särskiljs de olika ordklasserna först i nedre ändan av hierarkin medan WordNet gör indelningen högst uppe. Begreppet "entitet" hänför sig uttryckligen till substantiv.

Begreppssystemen i Lagercrantz och WordNet verkar skilja sig radikalt från varandra. I vissa fall blir över- och underbegrepp närmast diametralt motsatta. Hur är det möjligt? WordNet ska ju vara en allmängiltig och universell beskrivning. En viktig orsak till skillnaderna är hur begreppssystemen kommit till. Lagercrantz utgick från en samling begrepp som han grupperade och kategoriserade. Han sammanställde alltså systemet nerifrån, induktivt. WordNet däremot har skapats uppifrån och ner, alltså deduktivt.

I båda systemen ingår en grammatisk klassificering: i WordNet är entiteterna alltid substantiv, medan Lagercrantz särskiljer ordklasser först på lägsta nivå. På samma nivå som klassen "substantiv för modalitet och orsak" ligger klasserna "adjektiv för modalitet och orsak", "adverb för modalitet och orsak" och "verb för modalitet och orsak". Jämförelsen mellan systemen visar alltså, att de grammatiska kriterierna kan användas på diametralt motsatta ställen i hierarkierna.

Intressant är också att WordNets "utseende, visuell aspekt" är ett underbegrepp medan Lagercrantz "sinnesförnimmelser" är ett överbegrepp för det generella begreppet "modalitet och orsak".

Vad är då utfallet om vi jämför de två systemen med avseende på ett helt annat begrepp, ett kulturbegrepp som är centralt för samerna, nämligen djuret ren? Hierarkin i Lagercrantz ses i figur 11, WordNets mycket djupare begreppsserie[5] framgår av figur 12.

5 Jag har gallrat bort all onödig information och vänt upp och ner på serien så att den går uppifrån och ner i likhet med Lagercrantz hierarki.

naturbegrepp
levande natur
animaliskt liv
människor och djur
djur; husdjur
ren, allmänt

Figur 11: Hypernymkedjan för begreppet ”ren” i Lagercrantz.

entitet
fysisk entitet
objekt, föremål
helhet, enhet
levande varelse
organism
djur
ryggsträngsdjur, kordater
ryggradsdjur, kraniedjur, kraniater, vertebrater
däggdjur
placentadäggdjur
hovdjur, ungulater
partåiga hovdjur, klövdjur
idisslare
hjortdjur, hjortar
ren

Figur 12: Hypernymkedjan för begreppet ”ren” i WordNet.

Där Lagercrantz klarar sig med sex nivåer behöver alltså WordNet hela sexton. Särskilt många gemensamma begrepp innehåller klassificeringarna inte. Gemensamt är självfallet det begrepp som ska klassificeras, eller ”ren”. Bland de överordnade kategorierna är det bara två som är desamma: ”djurriket” och ”levande natur/varelse”.

Utifrån WordNets klassificering vet vi med största exakthet vad en ren är rent naturvetenskapligt. Utifrån Lagercrantz klassificering vet vi något helt annat. Under begreppet ”ren, allmänt” anför Lagercrantz en synonymgrupp med sammanlagt 24 uttryck som syftar på renar av olika slag, till exempel ”avskilt levande ren”, ”utvilad ren (som kan utnyttjas som dragdjur)”, ”ren med benägenhet att dra åt ena sidan”. Se också avsnitt 2.1 figur 7.

3.3. Är WordNet en ontologi?

I hypernymkedjorna för ”tomato” i WordNet (figur 13) ser vi att den vänstra, taxonomiska serien har två noder mera än den högra, icke-vetenskapliga serien (0–9 resp. 0–7). Den ontologiska klassificeringen i entiteter börjar uppifrån (**0–**). Den taxonomiska klassificeringen i taxoner börjar nerifrån (*9–*) liksom den intuitiva klassificeringen i begrepp (*7–*). I den vänstra kedjan utgör noderna 6–9 den biologiska taxonomin medan den översta delen (0–6) bildar ontologin. Nod 6 eller entiteten och taxonen ”växt” utgör länken mellan den ontologiska och den taxonomiska delen. Den högra serien är likartad. Intuitiva vardagsbegrepp är grönsak (nod 5), jordbruksprodukter (4) och livsmedel (3). Däremot är det snarast en taxonomisk uppgift att tomat hör till kärlväxterna (6). Den ontologiska delen av den högra hypernymkedjan består av noderna 0–3, och nod 3 ”föda” utgör länken mellan de ontologiska entiteterna och de språkbundna begreppen.

taxonomiska hypernymkedjans topp	*icke-vetenskapliga hypernymkedjans topp*	
0 entity	entity	**0**
1 physical entity	physical entity	**1**
2 object, physical object	matter	**2**
3 whole, unit	food, solid, food	***3***
4 living thing, animate thing	produce, green goods, green	
5 organism, being	groceries, garden truck	*4*
6 plant, flora, plant life	vegetable, veggie, veg	*5*
7 vascular plant, tracheophyte	solanaceous	
8 herb, herbaceous plant	vegetable	*6*
9 tomato	tomato	*7*

Figur 13: Två (förkortade) hypernymkedjor för begreppen 'tomat' i WordNet.

Utifrån det ovan sagda kan vi konstatera att den översta nivån i WordNet:s hierarki verkligen är ontologisk och baserar sig på logiska entiteter. Nerifrån räknat består hypernymkedjorna däremot snarast av språk- och kulturbegrepp. Den ontologiska delen kan utan vidare översättas till andra språk, eftersom det som ska översättas är benämningar på logiska entiteter vilkas innehåll inte ändras vid översättning. Den begreppsliga delen av WordNet är däremot språkbunden, och att översätta den ord för ord till ett annat språk leder inte till ett rimligt resultat. Orden i WordNet är inte heller ”ord” utan de är begrepp, och översättningen borde därför ske begrepp för begrepp, inte ord för ord. Men det är en omöjlighet, eftersom begreppssystemen i olika språk är olika.

För EuroWordNet har ontologin (Top Ontology) från början utformats som en självständig del, som inte är beroende av de enskilda språken. Begreppen i de ingående språken är kopplade till varandra via den språkoberoende begreppsapparat som i sin tur är kopplad till ontologin. Ett språkoberoende begreppssystem är i ljuset av det ovan sagda en självmotsägelse, och att försöka bygga

upp den delen av EuroWordNet har inneburit stora problem (se Rodríguez 1998; Vossen 1998).

4. Sammanfattning

I denna artikel har jag genom exempel försökt klargöra hur moderna ontologier förhåller sig till WordNet och begreppsordböcker. Begreppsordböcker och begreppssystem har en flerhundraårig historia medan ontologierna trampar i barnskorna. Man kan i alla fall tydligt se att det informationstekniska arbetet bär frukt. Det främjar inte bara datorernas förmåga att hantera semantik – något som möjliggör mer intelligenta webbsökningar och extrahering av viktig information ur stora textmassor – utan det berikar och befrämjar också språkforskning, i synnerhet, semantik, lexikologi och lexikografi.

Det var under upplysningstiden som begreppet encyklopedisk vetenskap uppkom. I encyklopediska publikationer hade man som strävan att samla all information om hela världen. I och med att kunskapsmängden ökar enormt vartefter den vetenskapliga forskningen framskrider har sådana föresatser sedermera visat sig vara omöjliga att uppnå, hur omfattande en redaktion än är. Möjligen kan det visa sig att de nuvarande ansträngningarna att få datorer att via en universell ontologi förstå all existerande och framtida kunskap om världen blir mer framgångsrika.

Litteratur

Ordböcker

Dornseiff, F. 1933–1940: *Der deutsche Wortschatz nach Sachgruppen*. Berlin. [Åtskilliga senare upplagor.]

Nagy, G. 1978: *Magyar szinonimaszótár*. Budapest.

Roget, P. M. 1856: *Thesaurus of English Words and Phrases, Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition*. London. [Åtskilliga senare upplagor.]

Wehrle, H. 1942: *Deutscher Wortschatz. Ein Wegweiser zum treffenden Ausdruck*. Stuttgart. [Åtskilliga senare upplagor.]

Annan litteratur

Anderson, W., Pretorius, L., Kotzé, A. 2010: Base Concepts in the African Languages Compared to Upper Ontologies and the WordNet Top Ontology. I: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10),* 3757–3764. www.lrec-conf.org/proceedings/lrec2010/pdf/247_Paper.pdf

Christiansson, J., Zimmerman, Z. 2006: *Word sense disambiguation med Svenskt OrdNät*. Magisteruppsats i biblioteks- och informationsvetenskap vid Biblioteks- och Informationsvetenskap/Bibliotekshögskolan 2006:34. Högskolan i Borås. bada.hb.se/bitstream/2320/1449/1/06-34.pdf

Gruber, T. R. 1993: A translation approach to portable ontology specification. I: *Knowledge Acquisition 5,2,* 199–220.

Lagercrantz, E. 1939: *Lappischer Wortschatz*. I–II. Lexica Societatis Fenno-Ugricae VI. Helsinki.

Lindén, K., Carlson, L. 2010: FinnWordNet – WordNet på finska via översättning. I: *LexicoNordica 17,* 119–140.

Linné, C. von 1753: *Species plantarum*. Stockholm.

Lorhard, J. (Jacobus Lorhardus) 1606: *Ogdoas Scholastica, continens Diagraphen Typicam artium: Grammatices (Latinae, Graecae), Logices, Rhetorices, Astronomices, Ethices, Physices, Metaphysices, seu Ontologiae, Sangalli*. Apud Georgium Straub.

Martola, N. 2011: FinnWordNet och kulturbundna ord. I: *LexicoNordica 18* (i denna volym).

Niles, I., Pease, A. 2001: Towards a Standard Upper Ontology. I: *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001),* Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001. home.earthlink.net/~adampease/professional/FOIS.pdf

Niles, I., Pease, A. 2003: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. I: *Proceedings of the IEEE International Conference on Information and Knowledge Engineering (IKE 2003),* s. p. home.earthlink.net/~adampease/professional/Niles-IKE.pdf

Pedersen, B.S. 2010: Semantiske sprogressourcer – mellem sprogteknologi og leksikografi. I: *LexicoNordica 17,* 163–180.

Rodríguez, H. m.fl. 1998: The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. I: *Computers and the Humanities 32,* 117–152.

Viberg, Å. 2002: Svenskt OrdNät. I: *Nordisk språkteknologi 2002 – Nordic Language Technology 2002,* 135–144. www2.lingfil.uu.se/personal/viberg/VibergNordSprktekPDF.pdf

Vossen, P. m.fl. 1998: *The EuroWordNet Base Concepts and Top Ontology*. Version 2, Final. January 22, 1998. Contributors: Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Wim Peters. Deliverable D017, D034, D036, WP5. EuroWordNet, LE2-4003. www.vossen.info/docs/1998/D017.pdf

Internethänvisningar (januari 2011)

DanNet = DanNet – Det danske wordnet. wordnet.dk/dannet
DanNet ”porre” = porre – andreord.dk/ord/8057-porre

FinnONTO = National Semantic Web Ontology Project in Finland. www.seco.tkk.fi/projects/finnonto/

Index of WordNetMappings = Index of /KBs/WordNetMappings. sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/

KulttuuriSampo = KulttuuriSampo – suomalainen kulttuuri semanttisessa web 2.0:ssa. www.kulttuurisampo.fi/

SUMO = Suggested Upper Merged Ontology (SUMO) – Ontology Portal. www.ontologyportal.org

SUMOclasses = SUMOclasses.gif. www.ontologyportal.org/images/SUMOclasses.gif

SUMOFAQ = Ontologt Portal – FAQ. www.ontologyportal.org/FAQ.html

WordNet = About WordNet – WordNet – About WordNet. wordnet.princeton.edu/

YSO = ONKI3 | Finnish General Upper Ontology – YSO/ALLSO. www.yso.fi/onto/yso/

Klaas Ruppel
forskare, föreståndare för ordboksavdelningen
Forskningscentralen för de inhemska språken
Berggatan 24
FI-00100 Helsingfors
klaas.ruppel@kotus.fi

Översättning från finska: Nina Martola