

Om representation i neurala system

-Vad har neurovetenskapen att säga om hjärnans förmåga att representera omvärlden, och vilka filosofiska implikationer har det?

Av Andreas Chatzopoulos, 2006

D-uppsats i Kognitionsvetenskap

Kollegium SSKKII,

Göteborgs universitet

Handledare: Tomas Kalén

Innehållsförteckning

1) Inledning	s. 3
2) Neurovetenskaplig bakgrund	s. 4
2.1) Kognitiva representationer	s. 4
2.1.1) Cortex	s. 4
2.1.2) Självaktivering och vektorer	s. 5
3) Filosofisk utredning	s. 9
3.1) Symboler och representation	s. 10
3.2) Tankeinhåll	s. 11
3.3) Språklig mening	s. 13
3.4) Meningsteorier	s. 13
3.4.1) Conceptual Role Semantics	s. 14
3.5) Diskussion	s. 15
4) Sammanfattning och slutsats	s. 17
5) Litteratur	s. 19

1) Inledning

Syftet med föreliggande uppsats är att undersöka några traditionella filosofiska frågor om mening och intentionalitet med utgångspunkt i moderna neurovetenskapliga rön. Detta därför att jag är intresserad av vad neurovetenskapen har att säga om hjärnans förmåga att representera och vilka filosofiska konsekvenser det har.

Uppsatsen är i huvudsak baserad på neurovetenskapliga rön beskrivna i boken "The mind-brain continuum", vilket är en essäsamling sammanställd av Rodolfo Llinás och Patricia S. Churchland (1996). Inspiration till de filosofiska vinklingarna på dessa texter kommer från en recension av just denna bok: "From sensory neuroscience to neurophilosophy: reflections on Llinás and Churchland's Mind-Brain Continuum", skriven av John Bickle (1997). Ett återkommande tema i "The mind-brain continuum" är nämligen synen på det centrala nervsystemet som ett stängt system som inte behöver input från omvärlden för att bygga upp kognitiva representationer (en vanlig uppfattning är nämligen att hjärnan arbetar med kognitiva representationer av omvärlden). Om nu kopplingen mellan omgivningen och dessa representationer är svagare än vad man hittills trott, med vilken rätt kan man då kalla dem för just representationer? Bickle skriver angående detta (s. 526):

The role of sensory input is to 'trellis, shape and otherwise sculpt' intrinsic neural activity to produce survival-enhancing representations. Llinás and Paré (Chapter 1) express this idea by claiming that the central nervous system is a 'closed' system whose basic organization is oriented toward generating 'intrinsic' images. Sensory inputs 'specify' internal states and 'modify' activity in this 'closed system' [...] Brains are "self-activating" and capable of generating cognitive representations in the absence of sensory input.

Han ställer sig sedan frågan i vilken mån dessa intergenererade representationer då kan ge information om omvärlden, och därmed i vilken bemärkelse de fungerar som just representationer av omgivningen. Det är denna frågeställning som är utgångspunkten för föreliggande uppsats: *Om hjärnans kognitiva representationer är intergenererade, i vilken mån kan de då sägas representera omgivningen?* Detta skall ses som en del i en undersökning av möjligheten att utarbeta en filosofisk meningsteori som är kompatibel med de aktuella neurovetenskapliga rönen, och hur denna teori i sådana fall bör vara beskaffad.

Mitt tillvägagångssätt är att först redogöra för de aktuella neurovetenskapliga teorierna för att därefter undersöka deras filosofiska konsekvenser. I min kommande filosofiska ut-

redning kommer jag därför att utgå ifrån att de neurovetenskapliga teorierna är korrekta för att se var det leder mig.

2) Neurovetenskaplig bakgrund

2.1) Kognitiva representationer

2.1.1) Cortex

M.M. Merzenich och R.C. deCharms text "Neural Representations, Experience and Change" (1996) behandlar de sätt på vilka intryck från omgivningen representeras i nervsystemet. De talar här mest om cerebrala cortex som de hävdar utgörs av en ansamling neuron vars huvudsakliga funktion består i att just representera omvärlden. Deras åsikt är att dessa neuron utgör ett dynamiskt system som förser våra erfarenheter med både innehåll och kontext (s. 61). De menar att våra varseblivningar, beslut, handlingar med mera, äger rum "against a backdrop of what is going on in our minds at the time" (s. 63). Här talas alltså om en, av nervsystemet skapad, intern kontext som innefattar sådana fenomen som uppmärksamhet, förväntan, känslor, motivation med mera. Detta menar de färgar våra upplevelser och skapar en miljö som påverkar vår uppfattning av objekten: "It is our belief that object and context hold equally important causal roles that lead to perception and onward through action" (s. 64). De menar alltså att kognitiva representationer *upplevs på olika sätt, beroende på kontext*. Förutom att vara representationer har de då alltså även *fenomenologiska* egenskaper, eftersom de upplevs på ett speciellt sätt. Vad gäller frågan om hur dessa representationer implementeras i hjärnan så menar de att det finns många övertygande argument för att cortex representationsenheter inte utgörs av *individuella* neuron. Ett exempel på ett sådant argument är helt enkelt att det inte finns tillräckligt många neuron för alla potentiella representationer. Ett annat indicium på att representationer utgörs av *neuronansamlingar* snarare än ensamma neuron, är alla de experimentella resultat som de hävdar visar att individuella intryck (percepts) "activate cortical systems widely" (s. 65). De syftar här på olika experiment som visar att riktad uppmärksamhet mot ett objekt genererar ökad aktivitet i en hel grupp av neuron (en aktivitet som kan observeras och mätas), och de menar att den ökade aktiviteten äger rum i just de neuron som utgör den kognitiva representationen av objektet i fråga. Enligt deras utsago visar även dessa experiment att neuronmönstret inte är statistiskt: Individens *grad av uppmärksamhet påverkar vilka* neuron som aktiveras. Författarna drar därför slut-

satsen att "the cortical system must use some other strategy to represent the constancy of objects over time than fixed representational patterns" (s. 63).

De hävdar nu att förändringar av dessa representationer i cortex sker med hjälp av *synkroniserade insignaler*. Det vill säga att inkommande signaler som anländer till en given plats i cortex inom en viss tidsram blir en del av samma representation, vilken då lagras i ett sammanhållet område i cortex. Detta innebär att cortex med tiden blir en slags karta över de inkommande signalernas tidsrelationer. Dessa förändringsmekanismer fungerar *konkurrensmässigt*, vilket innebär att olika inkommande signaler konkurrerar om det begränsade cortexområdet, och att endast de *starkaste* intrycken (de insignaler som genererar störst aktivitet hos neuronerna) resulterar i en representationsansamling. En annan viktig egenskap hos denna mekanism är att dessa signaler kan komma såväl *utifrån* cortex som *inifrån*. Vidare menar Merzenich och deCharms att *de flesta* signaler är interna (s. 69-70):

Researchers often think of representational competition in terms of a vying of competing extrinsic inputs, but intrinsic projections from neurons within the local cortical network itself have also been demonstrated to be major players in this competitive mechanism. The majority of input to cortical neurons actually comes from other neurons in the local network neighborhood.

De kognitiva representationerna behöver alltså inte vara direktorsakade av inkommande signaler utifrån, det är vanligare att de skapas av interna signaler inom cerebrala cortex.

2.1.2) Självaktivering och vektorer

R. Llinás utgångspunkt är något annorlunda. Han tar avstamp i den evolutionära funktionen hos hjärnan, och driver tesen att en av hjärnans huvudfunktioner är att *förutsäga händelser* (1987, s. 340). Det vill säga att på basis av inkommande stimuli förutsäga skeenden i omgivningen (vilket ju måste ses som en oerhörd evolutionär fördel då det hjälper organismen att undvika faror). Detta antas då ske genom att "[The brain] incorporate sensory referred properties of the external world *into the internal functional states*" (s. 343). Han tänker sig alltså att omgivningens egenskaper kommer att representeras av olika funktionella tillstånd hos hjärnan. I Llinás och Paré (1996) anknyts det till denna diskussion då det talas om hjärnans funktion som *verklighetsemulator*. "Let us assume that brain function embeds a type of reality emulation that evolved to specify internally the salient aspects of the surrounding world" (s. 3). Llinás och Paré menar här att en dylik emulatorfunktion faktiskt är en *förutsätt-*

ning för koordinerad motorik då den representerar information om kroppens läge i rummet och "creates a predictive image of an event" (s. 3):

The nervous system may implement the transformation of activity patterns obtained by sensory inputs into coherent images. Such an image is considered here to be a premotor template that would serve as a planning platform for behavior, the prerequisite to purposeful action and even to human consciousness.

En tes som de förespråkar är att denna emulatorförmåga är *medfödd* eftersom hjärnans struktur är fixerad redan vid födseln och i princip ser likadan ut hos alla människor, i likhet med andra kroppsfunktioner; "All the muscles, bones, and joints, and most of what they are capable of doing, are in principle inscribed in the geometry of the system at birth" (s. 3-4). Dock har alla individer en viss anpassningsförmåga, kallad *plasticitet*. Ett exempel på detta är den process som äger rum då vi vid uppväxten lär oss ett visst språk; vi föds med förmågan att känna igen alla fonem, men beroende på vilken språklig gemenskap vi växer upp i så utvecklas förmågan att känna igen vissa fonem på bekostnad av förmågan att känna igen de fonem som vi aldrig konfronteras med. Slutsatsen som de drar av dessa överväganden är att många av hjärnans egenskaper är *medfödda* snarare än inlärd, men att en viss anpassning sker.

Författarnas diskussion leder dem nu in på frågan om den övergripande naturen hos det centrala nervsystemets uppbyggnad: Har det en *öppen eller stängd arkitektur*? Den traditionella synen är, hävdar de, att arkitekturen är öppen. Detta innebär att nervsystemet får insignaler från omgivningen, behandlar dessa, och ger feedback i form av reflexer. Detta, menar de, är ett uttryck för en *tabula rasa*-syn på det centrala nervsystemet, en syn enligt vilken hjärnan endast är en "learning machine" som är beroende av inkommande signaler från omgivningen för sin funktion. De anser detta synsätt vara felaktigt då det lämnar många av det centrala nervsystemets förmågor oförklarade och inte heller berör det faktum att dessa förmågor i princip är likadana hos alla individer (exempelvis har vi "alla" förmågan att höra, se färger, lära oss språk, etcetera). De menar därför att det centrala nervsystemet är ett exempel på en stängd arkitektur (s. 4):

Some of us consider the CNS [Central Nervous System] to be a fundamentally closed system with its basic organization geared toward the generation of intrinsic images (thoughts or predictions), where inputs *specify* internal states rather than *inform* "homuncular vernacles".

Deras åsikt är alltså att det centrala nervsystemet i huvudsak är *självaktiverande* och äger förmågan att *generera kognitiva representationer av omgivningen*. De menar också att detta system fungerar på samma sätt *även vid frånvaro av inkommande signaler*, exempelvis då man drömmer. Vakna och drömande tillstånd är väldigt lika enligt Llinás och Paré (s. 6):

As the CNS has only an indirect relationship with external reality, perceptive states may be generated by two distinct mechanisms: through the process of dreaming, or in response to sensory input during wakefulness. In fact, dreaming and wakefulness are so similar from electrophysiological and neurological points of view that wakefulness may be described as a dreamlike state modulated by sensory input.

Deras slutsats är alltså att det centrala nervsystemet från födseln har förmågan att fungera som en verklighetsemulator, och att detta system även är verksamt vid frånvaro av inkommande signaler.

Vad är då relationen mellan de kognitiva representationerna och omgivningen? Llinás och Parés åsikt är att representationerna skapas genom en *internalisering av omvärldens egenskaper i en intern funktionsrymd*, det vill säga väsentligen samma synsätt som har behandlats av Llinás tidigare (1987), och som jag nämnde inledningsvis. Han skriver där om detta internaliseringsbegrepp (s. 344):

By internalization is meant the ability that the nervous system has to fracture external reality into sets of sensory messages (carried to the brain by millions of sensory nerve fibres) and to simulate such reality in brain reference frames.

Ett synsätt som han utvecklar här är att denna internaliseringsprocess formellt kan beskrivas med matematiska metoder. Tanken är att inkommande signaler kommer att representeras av en dynamisk vektor i ett koordinatsystem över det perifera nervsystemets inkommande signaler. Man erhåller därmed ett system där den elektriska aktiviteten från varje inkommande nervbana motsvaras av en komponent hos en sådan vektor (som därmed blir multidimensionell emedan den genereras av impulser från många samtidiga nervbanor). På så vis menar Llinás att dessa "sinnesvektorer" kommer att fungera som representationer av externa objekt. Dock med ett viktigt förbehåll: För att kunna fungera som representationer måste dessa vektorer först *transformeras* till internt meningsfulla vektorer, detta på grund av att hjärnans *interna referensramar* skiljer sig från strukturen hos det externa sinnessystemet, om man får tro Llinás. Som exempel på detta anför han receptorer "which may respond to stimuli such as light, or sound, or angular movement" (s. 344):

Their messages in the form of nerve impulses [...] must, therefore, be decoded by a set of internal neurons that do not share the external physical arrangement or the functional properties that receptor systems have in their periphery.

Ett liknande problem uppkommer av det faktum att antalet inkommande nervbanor till en given plats i hjärnan kan vara större eller mindre än antalet utgående banor. Llinás menar att vad som här behövs är en slags vektortransformation som är oberoende av vilka koordinat-system som används. Dyliga vektortransformationer är inom matematiken tydligen kända som *tensorial transformations*, och Llinás har, på grundval av detta synsätt, tillsammans med Pellionisz utvecklat vad de kallar för en "tensor network theory". Detta som en ansats till en utveckling av ett formellt studium av nervsystemet, en inriktning som "in fact suggests that neuronal networks actually implement tensorial transformations by means of their electrical activity and connectivity", enligt Llinás (s. 344). Jag kommer dock inte att gå närmare in på denna teori här, utan nöja mig med att konstatera att hjärnan, enligt deras teori, använder sig av vektorer för att representera intryck, och att dessa vektorer är ett resultat av en transformering av inkommande signaler.

Alltså: Förutom att, som vi konstaterade innan, det centrala nervsystemet har förmågan att generera kognitiva representationer som inte nödvändigtvis behöver vara resultatet av yttre signaler, så ser vi nu också att dessa representationer *inte heller kommer att likna de signaler som orsakar dem* (på grund av vektortransformationen). Slutsatsen blir då att om Merzenich, deCharms, Llinás och Paré har rätt så arbetar hjärnan med representationer som inte nödvändigtvis måste vara orsakade av omvärlden, och om de nu är ett resultat av signaler utifrån så kan de inte antas likna det de representerar. Detta eftersom de inkommande signalerna transformeras på vägen. En annan viktig slutsats, som tycks följa av allt detta, är att dessa kognitiva representationer även har en *fenomenologisk* karaktär, det vill säga, subjektet *upplever* dem på ett speciellt vis. Detta är tydligt hos Merzenich och deCharms då de anser att kontexten färgar våra upplevelser och därmed påverkar uppfattningen av objekten. Denna fenomenologiska aspekt återfinns även i Llinás och Parés tal om "images" som hjälper en att förutsäga skeenden: Dessa "images" måste rimligtvis upplevas på ett speciellt vis av subjektet.

3) Filosofisk utredning

Om man antar att dessa neurovetenskapliga teorier är korrekta, vilka filosofiska konsekvenser får det då? Låt oss sammanfatta vad vi kom fram till i förra avsnittet:

- Det centrala nervsystemet är självaktiverande och kapabelt att generera kognitiva representationer med en fenomenologisk karaktär.
- Dessa representationer måste inte nödvändigtvis vara direktorsakade av omgivningen, de är resultatet av en komplicerad process där flera neuronlager är inblandade.
- Under denna process så kommer de inkommande signalerna att bli transformerade, vilket innebär att den slutgiltiga representationen inte liknar det den antas representera.

En slutsats man kan dra av detta är att dessa kognitiva representationer som antas vara en så viktig del av våra tankeprocesser, inte logiskt kan kopplas till omgivningen. Detta eftersom vi aldrig kan ha exakt kunskap om den exakta relationen mellan våra kognitiva representationer och omgivningen. Jag vill dock framhålla att detta inte spelar någon roll för frågan om dessa representationers *användbarhet*. De är alltså en viktig del av våra tankeprocesser, oavsett ursprung. Dessa överväganden reser en del frågor, bland annat:

- Om de kognitiva representationerna inte liknar de ursprungliga signalerna som skapat dem, hur kan de då ge någon information om det som de representerar?
- Om vi inte ens kan vara säkra på att representationerna är orsakade av yttre signaler, hur kan vi då överhuvudtaget betrakta dem som representationer?

Detta kan reduceras till tre grundläggande, filosofiska frågor:

1. Vad innebär det att *representera*? Vad är en *symbol*?
2. Hur förklaras *innehållet* hos mentala tillstånd?
3. Hur får språket *mening*, och vad bestäms den av? Interna eller externa faktorer?

Detta är frågor som bör redas ut, och min avsikt är nu att undersöka dessa under antagandet av de neurovetenskapliga teorier som jag talade om. Men först vill jag i tur och ordning fördjupa mig i de filosofiska teorierna bakom frågorna:

3.1) Symboler och representation

Vad innebär det att representera? En enhet som representerar någonting (står för någonting) kallas vanligtvis för en *symbol*. Rent formellt så är en symbol en del av ett *formellt system*. Ett sådant system grundas i sin tur på det väldefinierade logiska begreppet *formellt språk* som utmärks av att det är formulerat utan någon referens till yttre objekt (det vill säga till objekt *utanför* språket självt). Språkets beståndsdelar utgörs av *tecken* (tokens) samt regler som anger hur dessa tecken får kombineras (man skulle kunna se det som ett alfabet med en tillhörande grammatik). Dessa tecken är tänkta att fungera som symboler, men kan inte räknas som sådana förrän de har fått en referens tilldelad. Denna referens specificeras dock aldrig i själva språkdefinitionen (ett formellt språk *måste* vara möjligt att beskriva fullt ut utan referenser till yttre objekt). Det är först vid en *tolkning* som tecknen tilldelas referens och får symbolstatus (mer om det nedan). Om en tolkning av en sats resulterar i att satsen uttrycker någonting som är sant (som överensstämmer med verkligheten) så kallas tolkningen för en *modell* av satsen. När man talar om en *mängd* satser så är en modell en tolkning som gör *alla* satserna sanna.

Ett formellt system är alltså ett system som är definierat i ett formellt språk. Detta system kompletteras med en *deduktionsapparat*. Detta innebär rent konkret att man specificerar vilka formler som skall utgöra axiom, samtidigt som man ger en definition av härledningsregler för dessa axiom. Ett illustrativt sätt att se på dessa formella system är att, som John Haugeland (1996/1997), betrakta dem som ett *spel* som går ut på att manipulera tecken efter givna regler. Dessa tecken måste då, som nämnts, ha en referens för att kunna betraktas som symboler. Haugeland talar i detta sammanhang om *tolkade formella system*: I dylika system är tecknen symboler emedan de *representerar* någonting. De har blivit tilldelade referenser i och med att det formella systemet har blivit tolkat (tolkningen anger alltså vad symbolerna står för). När man konstruerar formella system så utformar man ofta systemets härledningsregler utifrån vad symbolerna är tänkta att representera. Haugeland illustrerar detta faktum med ett exempel som går ut på att han tänker sig en översättning av ett antikt dokument som är författat på ett för oss okänt språk: Om nu någon, som ett försök till översättning, föreslår ett översättningsschema som *inte* genererar någonting vettigt (kanske bara en mängd ord utan sammanhang) så kommer vi förmodligen inte att se att detta översättningsförslag som speciellt plausibelt. Om däremot schemat genererar koherenta meningar, som dessutom verkar vettiga i relation till andra källor, så övertygas vi vanligtvis om riktighe-

ten hos förslaget. Haugeland menar att skälet till detta är att denna översättning är att betrakta som en slags *tolkning*, och skriver apropå detta (s. 18):

Instead of saying what some system think or is up to, a translator says what some string of tokens (symbols) mean. To keep the two species distinct, we can call the former intentional representation, since it attributes intentional states, and the latter (translation) semantic interpretation, since it attributes meaning (=semantics).

Vad han gör här är att han särskiljer mellan intentionala representationer (tankeinhåll) och språklig mening. Istället för att undersöka en talares tankeinhåll så gör en utomstående åhörare en tolkning som tillskriver mening till de uttalade symbolerna. Detta leder oss in på fenomenen tankeinhåll och språklig mening:

3.2) Tankeinhåll

Ett utmärkande drag hos mentala tillstånd är att de har innehåll (content). Med detta menas att de *handlar om* någonting, de *representerar* något. (Exempel: när man ser ett träd så representerar ens perceptuella tillstånd detta träd.) Som Haugeland skriver (s. 4): "intentionality is the character of one thing being 'of' or 'about' something else, for instance by representing it, describing it, referring to it, aiming at it, and so on". Denna intentionalitet kan vara *härledd* eller *ursprunglig*, menar han. Tanken är här att artefakter som böcker, bilder, med mera har härledd intentionalitet. De handlar om någonting *därför att vi som användare anser att de gör det*; det är alltså vår tolkning som ligger till grund för detta, varför deras intentionala egenskaper måste anses härledda (det vill säga härledda från den ursprungliga—den som finns hos våra tankar). Han menar att orsaken till denna skillnad står att finna i det faktum att artefakter med härledd intentionalitet inte *gör* någonting med det de representerar; "they never pursue goals, draw conclusions, make plans, answer questions [...] they just sit there" (s. 7). Våra intentionala representationer, däremot, är ofta orsak till dessa fenomen som här räknas upp, och bör därför räknas som ursprungliga, enligt honom. Detta är alltså Haugelands syn på intentionalitet, ett synsätt som, i likhet med Daniel Dennetts teorier, fokuserar på tillskrivning av intentionalitet utifrån beteende. Det centrala tankesättet är här att man tillskriver mentala tillstånd till ett system som uppvisar tecken på intentionalitet. I andra, mer traditionella filosofiska teorier rörande intentionalitet, är det mest själva de mentala tillstånden som man är intresserad av och inte så mycket det beteende som de ligger till grund för: Det är denna andra, inre, aspekt (de intentionala egenskaperna hos de mentala tillstånden) som jag kommer att intressera mig för här.

Någonting som är viktigt för förståelsen av intentionalitet är analysen av kognitiva representationer, och hur dessa är relaterade till så kallade *propositionella attityder* (propositional attitudes), det vill säga sådana fenomen som manifesteras i trosföreställningar, önskningar med mera. För att exemplifiera kan man använda begreppet "regn". När jag tänker på regn så innebär det (enligt denna uppfattning) att jag har en *kognitiv representation* av regn. Denna representation är nu likadan, oavsett karaktären hos mina tankar (jag kan *tro* att det regnar ute, jag kan *önska* att det ville regna, jag kan vara *rädd för* att det skall börja regna, och så vidare). En bra metafor (som beskrivs av Warfield & Stich, 1994, s. 4) för att illustrera detta förhållande mellan representation och attityd är att tänka sig små lådor inne i huvudet, en för varje attityd. Om jag då tror att det regnar ute så innebär det att jag har en representation av regn i min låda för trosföreställningar, om jag önskar att det börjar regna så återfinns regn-representationen i min önskelåda, och så vidare. Det verkar alltså som att representationer spelar en mycket central roll i vårt tankeliv, precis som vi antog i avsnitt 2. Skillnaden mellan mening (semantik) och innehåll (intentionalitet) kan nu åskådliggöras genom att man urskiljer olika egenskaper hos de mentala tillstånden (representation + propositionell attityd). Exempelvis tanken: "Jag tror att solen skiner ute". Den semantiska egenskapen utgörs då av sanningsvillkoren hos satsen "solen skiner" (som satisfieras av att solen faktiskt skiner), medan den intentionala egenskapen manifesteras i det faktum att det mentala tillståndet handlar om solsken (den mentala representationen *representerar* solsken). Kopplingen mellan innehåll och mening kan nu beskrivas med den grundläggande tanken att *det någon säger är vad han menar*. Det är alltså meningen hos den yttrade satsen visar innehållet hos det mentala tillståndet. Tag som exempel satsen "det finns blommor i vasen". Den som yttrar denna sats *har en trosföreställning vars innehåll motsvaras av satsens mening* (nämligen att det finns blommor i vasen). Detta knyter an till tanken att man, för att förstå meningen hos (låt säga) satsen "det regnar", måste kunna tänka tanken att det regnar. För att detta nu skall vara möjligt är det rimligt att tänka sig att det finns någonting i ens inre som representerar sakförhållandet att det regnar.

Enligt denna uppfattning bör alltså ett mentalt tillstånd ses som en mental representation kombinerat med en propositionell attityd. Detta mentala tillstånd har ett innehåll som bestäms av den mentala representationen och som visas genom den språkliga meningen hos ens yttranden. Denna egenskap, att ha ett innehåll, är vad som kallas för intentionalitet.

3.3) Språklig mening

Ovanstående diskussion är relaterad till filosofiska meningsteorier, det vill säga teorier som tagits fram vid studiet av språkets meningsaspekter. De traditionella frågor som behandlas av dessa teorier är: Hur får ett språk mening? Hur får språkanvändarna kännedom om denna mening? Vad är relationen mellan de språkliga uttrycken och det som de används för att utsäga någonting om? (Det vill säga; vad innebär det att referera?) Michael Dummett (1976) redogör för några traditionella svar på dessa frågor, svar som kan sammanfattas som:

en sats mening = dess sanningsvillkor

ett ords mening = dess bidrag till satsens sanningsvillkor

Konsekvensen av en sådan uppfattning blir då en meningsteori som bygger på sanningsbegreppet: Att förstå en sats är liktydigt med att känna till satsens sanningsvillkor. Något som talar emot en sådan uppfattning är Hilary Putnams tankegångar så som de beskrivs i hans text "The meaning of 'meaning'" (1975/2002). Kärnan i dessa tankegångar är två uppfattningar som Putnam menar inte är förenliga:

1. Ett ords mening bestämmer dess sanningsvillkor.
2. Att känna till ett ords mening är detsamma som att befinna sig i ett visst mentalt tillstånd.

Det Putnam gör är att visa att dessa två påståenden inte kan vara sanna samtidigt. Detta genom att konstruera tankeexperiment där språkanvändaren är okunnig om ordens "egentliga" mening (det vill säga, deras referens) men *ändå* använder dem på samma vis (och därmed befinner sig i samma mentala tillstånd) som en användare med kännedom om den korrekta meningen (referensen). Han visar alltså att det går att föreställa sig situationer där ovanstående påståenden inte är sanna samtidigt, vilket implicerar att man måste förkasta något av dem. Han menar att det är mest rimligt att förneka påstående 2, vilket då resulterar i uppfattningen att mening inte kan finnas "i huvudet" (åtminstone inte helt och hållet). Detta synsätt brukar benämnas *externalism*.

3.4) Meningsteorier

Finns det då någon filosofisk meningsteori som är kompatibel med de neurovetenskapliga teorier som diskuterades i avsnitt 2? Där såg vi ju att även om vi inte logiskt kan koppla de kognitiva representationerna till omgivningen så verkar de spela en viktig roll för våra tanke-

processer. Hur är då detta relaterat med fenomenet mening? Putnams åsikt är ju att språklig mening i allra högsta grad är beroende av omgivningen och dess beskaffenhet. Men vilken roll har då de kognitiva representationerna? Om de nu är så betydelsefulla för våra tankeprocesser, kan de då vara betydelselösa i språklig kommunikation? Det finns dock filosofiska teorier som bejakar bägge aspekter, både objektiv språklig mening och subjektiva inre begrepp. Ett exempel på en sådan teori är *Conceptual Role Semantics*:

3.4.1) *Conceptual Role Semantics*

Conceptual Role Semantics (CRS) förespråkas bland annat av Ned Block i hans text "Advertisement for a semantics for psychology" (1994). Den teori som han här talar om, en så kallad tvåfaktor-variant, är inspirerad av Putnams diskussioner (som ju säger att mening inte både kan finnas "i huvudet" och bestämma referens). På grundval av denna insikt delar Block med sin CRS-teori in meningsbegreppet i två olika komponenter; en "inre" (narrow) komponent som behandlar de inre representationernas inbördes roller, samt en "yttre" (wide) komponent som har med fastställandet av dessa representationers referens att göra. Exakt *hur* den yttre komponenten är beskaffad specificeras inte av Block, och han ägnar det inte mycket uppmärksamhet. Därmed lämnar han här fältet öppet för en kombination med flera olika traditionella meningsteorier; "those who are so inclined could suppose it to be elucidated by a causal theory of reference or by a theory of truth-conditions" (s. 93). Den inre komponenten, däremot, redogör han för mer i detalj. Han kallar den för en "conceptual role component", det vill säga en komponent som är uppbyggd av inre begrepp och deras inbördes roller och som utgör mekanismen bakom våra tankeprocesser. Om detta skriver han (s. 93):

The internal factor, conceptual role, is a matter of the causal role of the expression in reasoning and deliberation and, in general, in the way the expression combines and interacts with other expressions so as to mediate between sensory inputs and behavioral outputs. A crucial component of a sentence's conceptual role is a matter of how it participates in inductive and deductive inferences. A word's conceptual role is a matter of its contribution to the role of sentences.

Det intressanta är här alltså de inre begreppens kausala roller, det vill säga de roller som spelar roll vid härledning (deduktiva och induktiva), beslutsfattande, och dylika processer.

Hur är då detta relaterat till representationernas *fenomenologiska* aspekter som vi talade om i avsnitt 2? Finns det något utrymme för en fenomenologisk komponent inom ramen för CRS? Block redogör för en i sammanhanget intressant variant som går under nam-

net Procedural Semantics (PS). Enligt Block är denna teori lik CRS, och om dess förespråkare skriver han (s. 95):

... procedural semanticists sometimes sound as if they want to take phenomenal terms as primitives whose meaning is given by their "sensory content," while taking other terms as getting their meanings via their computational relations to one another and to the phenomenal terms as well [...] It should be clear that this is a "mixed" conceptual role/phenomenalist theory and not a pure conceptual role theory.

Detta visar då att det finns alternativ som är kompatibla med de neurovetenskapliga rönen, och som förklarar de kognitiva processerna och deras fenomenologiska karaktär samtidigt som språkets objektiva mening bejakas. Det verkar alltså vara möjligt att konstruera en filosofisk meningsteori som gör en skarp distinktion mellan en inre och en yttre komponent, och som beskriver de kognitiva representationerna på ett sådant vis att de logiskt frikopplas från omvärlden (det vill säga, visas vara logiskt oberoende av yttre fenomen). Jag tänker här inte säga mer om exakt *hur* en sådan teori borde vara beskaffad, utan tänker nöja mig med denna skiss som åtminstone visar *att* en sådan teori är möjlig.

3.5) Diskussion

Hur är då CRS relaterat till den föregående diskussionen? Hur skall CRS interna begrepp betraktas? Är de symboler? Något som verkar rimligt är att likställa dessa begrepp med just de kognitiva representationer som det talades om i avsnitt 2. (Det var ju möjligheten till detta som gjorde CRS lockande till att börja med.) Sett ur ett formellt perspektiv så skulle då dessa kognitiva representationer (CRS inre komponent) kunna ses som *tecken* med tillhörande *manipuleringsregler* i ett formellt system. En tolkning av detta formella system skulle då innebära tilldelning av referens till tecknen och (om tolkningen är sann) utgöra en modell av det inre systemet. De kognitiva representationerna skulle då få referenser, och denna modell skulle då kunna sägas motsvara CRS yttre komponent, den som modellerar just relationen mellan den inre komponenten och omgivningen.

Vilken roll har då de kognitiva representationernas *fenomenologiska* aspekter? Som vi såg tidigare så finns det större utrymme för en fenomenologisk komponent i en PS-variant av CRS. Detta är ett skäl till att fokusera på just en sådan variant i sökandet efter en passande teori, eftersom det möjliggör konstruktionen av en teori som tilldelar de kognitiva representationerna associerade fenomenologiska upplevelser (som ju annars lämnas oförklarade). Ett sätt att betrakta en dylik teori (som lämnar utrymme för fenomenologiska upplevelser) är att se den fenomenologiska komponenten som *en egen tolkning av ens egna formel-*

la system. Med detta avses ett synsätt liknande det som anförs av P. N. Johnson-Laird (1983, kap. 11). Han är en förespråkare för en variant av PS, en variant som i hans tappning relaterar språket till en *mental modell*, snarare än till omgivningen. Hans åsikt är även att dessa mentala modeller kan innehålla element som korresponderar med objekt i omgivningen, eller snarare med *objekt i vår uppfattning av omgivningen*. (Elementens egenskaper och deras inbördes relationer motsvarar då vår uppfattning om det sakförhållande som den mentala modellen är tänkt att representera.) Min tanke är nu att analogt med Johnson-Lairds koppling mellan språket och en mental modell införa en liknande relation mellan de kognitiva representationerna och de fenomenologiska upplevelser som de orsakar. Dessa upplevelser skulle då tjäna som en modell av det formella system som utgörs av de kognitiva representationerna och deras inbördes relationer, på samma vis som Johnson-Lairds mentala modeller utgör modeller för ens språk. Det system som hjärnan utgör skulle därmed innefatta mer än bara ett formellt system (eftersom de fenomenologiska upplevelserna då kommer att ligga utanför det rent formella), och just därför skulle man här kunna använda sig av uttrycket "tolka sitt egna formella system" (de fenomenologiska upplevelserna utgör då själva tolkningen, modellen av systemet). I kontrast till den "riktiga" modellen, det vill säga de kognitiva representationernas "korrekta" referenser, så skulle denna upplevda tolkning då utgöras av ens *tänkta* referenser (det man *tror* att ens kognitiva representationer refererar till). Resultatet skulle då bli att ens fenomenologiska upplevelser skulle bestå av *föreställningar om omgivningen*, alltså just sådana simuleringar som Llinás och Paré anser vara evolutionärt viktiga och flitigt använda av hjärnan. Det är alltså vår medfödda verklighetsemulator som genererar dessa upplevda bilder som tjänar som våra privata tolkningar av våra kognitiva representationer. Detta leder oss fram till en tvåfaktorteori, precis som CRS: Dels får vi en yttre faktor som utgörs av de kognitiva representationernas riktiga referens (det är denna objektiva referens som det talas om i traditionella meningsteorier, exempelvis hos Putnam) och dels en inre faktor som utgörs av själva de kognitiva representationerna, deras inbördes roller, samt den egna trosföreställningen om deras referens (den fenomenologiska aspekten). På så vis bibehålls det logiska oberoendet mellan yttre och inre faktorer, och vi får en teori som är kompatibel med Putnams externalism samtidigt som den erkänner vikten av våra inre begrepp och upplevelser.

4) Sammanfattning och slutsats

Ett syfte med uppsatsen var att redogöra för hjärnans förmåga att representera omgivningen genom en undersökning av de kognitiva representationer som den arbetar med. Detta mot bakgrund av neurovetenskapliga rön som hävdar att dessa kognitiva representationer till stor del är internt genererade av hjärnan själv, och därmed logiskt oberoende av omgivningen—frågan var då på vilket sätt de i sådana fall kan anses vara representationer av omgivningen överhuvudtaget. Detta som ett led i en större frågeställning: Är det möjligt att utarbeta en filosofisk teori som är kompatibel med ovan nämnda neurovetenskapliga rön, och hur bör denna teori i sådana fall vara beskaffad?

En genomgång av de aktuella neurovetenskapliga teorierna visade att hjärnan verkar arbeta med hjälp av kognitiva representationer, och att dessa är associerade med fenomenologiska upplevelser som hjärnan genererar utifrån medfödda mekanismer. Filosofiskt viktigt var även att dessa representationer inte med nödvändighet är orsakade av inkommande signaler då hjärnan utgör ett självaktiverande system med inneboende förmåga att på egen hand generera representationer.

I syfte att finna en filosofisk meningsteori som var kompatibel med dessa neurovetenskapliga rön gick jag över till att diskutera tre filosofiska frågor, vilka jag utredde i tur och ordning. Dessa frågor var:

1. Vad innebär det att representera? Vad är en symbol?
2. Hur förklarar man innehållet hos mentala tillstånd?
3. Hur får språket mening, och vad bestäms den av?

Slutsatsen av denna utredning blev att vi bör koncentrera vårt sökande till en teori som beaktar både yttre och inre aspekter, det vill säga en teori som behandlar objektiv språklig mening likväl som inre, kognitiva representationer.

En lovande kandidat tycktes utgöras av Ned Blocks *Conceptual Role Semantics*. Denna teori delar upp meningsbegreppet i en yttre och en inre komponent där den inre utgörs av begrepp och deras inbördes roller medan den yttre hanterar deras relation till omgivningen. Detta verkade passa bra med mitt sökande efter en filosofisk teori som gör yttre och inre fenomen logiskt oberoende av varandra (för att vara kompatibelt med de neurovetenskapliga teorierna), och jag dristade mig nu att likställa teorins 'begrepp' med de kognitiva

representationer jag talat om innan. En ännu bättre kandidat hittades sedan i teorivarianten *Procedural Semantics*, en variant som även gav utrymme för de fenomenologiska aspekterna som de neurovetenskapliga teorierna tycks förutsätta.

Vidare diskussion visade att man kan se dessa meningsteorier som formella system vars tecken och manipuleringsregler utgörs av de kognitiva representationerna och deras inbördes roller. Detta system ges då mening av en tolkning som tilldelar referens till symbolerna. En viktig komponent återstod dock: Var i teoribygget får man rum med de fenomenologiska upplevelserna? Influerad av vissa tankar hos Johnson-Laird införde jag här det något vågade synsättet att betrakta dessa fenomenologiska upplevelser som egna tolkningar av ens egna formella system, det vill säga synsättet att de kognitiva representationerna är relaterade till fenomenologiska upplevelser som tjänar som tolkningar av dessa representationer.

Vilka slutsatser kan nu dras? I vilken mån kan de kognitiva representationerna sägas representera? Som vi har sett så finns det *två* aspekter på deras representation. Om man betraktar dem som delar av ett formellt system så har detta system *dels* en korrekt referens genom en objektiv tolkning av systemet, och *dels* en subjektiv referens genom egna trosföreställningar rörande denna referens. Det man traditionellt menar när man talar om att språkliga uttryck refererar är det som jag benämner objektiv referens, men jag menar att även den subjektiva tolkningen är viktig att ta hänsyn till. Lyckligtvis verkar det inte vara omöjligt att konstruera en filosofisk meningsteori som tar hänsyn till bägge dessa aspekter, både yttre och inre. Här finns olika varianter, men någon form av *Procedural Semantics* verkar vara lämplig att utgå från. Om man är intresserad av att finna en filosofisk teori som förklarar nämnda neurovetenskapliga rön så bör forskningen fokusera på denna typ av meningsteorier.

5) Litteratur

Bickle, J. (1997) From sensory neuroscience to neurophilosophy: reflections on Llinás and Churchland's Mind-Brain Continuum. *Philosophical Psychology*, 10, 523-530.

Block, N. (1994) Advertisement for a semantics for psychology. I: S. P. Stich och T. A. Warfield (red:er) *Mental representation : a reader* (ss. 81-141). Oxford: Basil Blackwell Ltd.

Dennett, D. C. (1981/1997) True believers: the intentional strategy and why it works. I: J. Haugeland (red.) *Mind design II* (ss. 57-79). Cambridge, MA: The MIT Press.

Dummett, M. (1976) What is a theory of meaning? (II) I: G. Evans och J. McDowell (red:er) *Truth and meaning* (ss. 67-137). Oxford: Oxford University Press.

Haugeland, J. (1996/1997) What is mind design? I: J. Haugeland (red:er) *Mind design II* (ss. 1-28). Cambridge, MA: The MIT Press.

Johnson-Laird, P. N. (1983) *Mental models*. Cambridge, UK: Cambridge University Press.

Llinás, R. (1987) 'Mindness' as a functional state of the brain. I: C. Blakemore och S. Greenfield (red:er) *Mindwaves* (ss. 339-358). Oxford: Basil Blackwell Ltd.

Llinás, R. & Paré, D. (1996) The brain as a closed system modulated by the senses. I: R. Llinás och P. S. Churchland (red:er) *The mind-brain continuum* (ss. 1-18). Cambridge, MA: The MIT Press.

Merzenich, M. M. & deCharms, R. C. (1996) Neural representations, experience and change. I: R. Llinás och P. S. Churchland (red:er) *The mind-brain continuum* (ss. 61-81). Cambridge, MA: The MIT Press.

Pagin, P. (1994) Moderna meningsteorier. *Filosofisk tidskrift*, 1, 3-25

Putnam, H. (1975/2002) The meaning of "meaning". I: Chalmers D. J. (red:er) *Philosophy of mind: classical and contemporary readings* (ss. 581-596). Oxford: Oxford University Press.

Warfield, T. A. & Stich, S. P. (1994) Introduction I: S. P. Stich och T. A. Warfield (red:er) *Mental representation : a reader* (ss. 81-141). Oxford: Basil Blackwell Ltd.