


HUMANIORA
KØBENHAVNS UNIVERSITET



Sprogteknologi og formel semantik

Bolette Sandford Pedersen
Center for Sprogteknologi
Københavns Universitet
bolette@cst.dk

HUMANIORA
KØBENHAVNS UNIVERSITET



Overzicht

- Det formelle paradigme
- Vigtige forskningsspørgsmål inden for sprogteknologi og formel semantik
- Formel leksikalsk semantik og leksikologi
- DanNet: et leksikalsk-semantisk ordnet for dansk
- Sprogteknologi og samfundet
- Sprogteknologi i fremtiden generelt og på KU

Lingvistikredsens 75-års jubilæum 2

HUMANIORA
KØBENHAVNS UNIVERSITET




Det formelle paradigme (Spang-Hanssen 2006 & Prebensen 2006)

Det formelle paradigme - at anskue sproget som et produktionssystem:

- Chomskys generative grammatikker: de centrale grammatiske strukturer kan beskrives i form af genskrivningsregler
- Ikke direkte knyttet til computere, men til matematikkens og logikkens udvikling af formelle logiske systemer
- Computeren som et håndgribeligt udtryk for hvordan man kan formalisere et bredt felt af menneskelig tænkning og symbolbehandling

Lingvistikredsens 75-års jubilæum 3

HUMANIORA
KØBENHAVNS UNIVERSITET




Det formelle paradigme

- Regelbaserede systemer giver et billede af kompleksiteten i naturlige sprogs grammatikker
- Fascinerende at se sproget som et matematisk, logisk system
- I dag er sprogteknologi meget mere.. ambitionerne og mulighederne er blevet større

Lingvistikredsens 75-års jubilæum 4

HUMANIORA
KØBENHAVNS UNIVERSITET




Formel semantik (Partee 1996)

Det formelle paradigme har traditionelt stået mere faglende overfor **semantikken**, men

- Montague fik stor betydning for formel semantik i lingvistikken (*English as a Formal Language* 1970, *Universal Grammar* 1970)
- Syntaks og semantik kunne hver især beskrives som en algebra med relationer imellem dem
- Partee introducerede Montague for lingvister, modelteoretisk semantik blev en lingvistisk disciplin

Lingvistikredsens 75-års jubilæum 5

HUMANIORA
KØBENHAVNS UNIVERSITET



Formel semantik: interesseobjekter

'To know the core meaning of a sentence uttered as a statement is to understand the conditions under which it is true' *Cann 1993:15*

Interesseobjekter i den 'klassiske' formelle semantik:

- sandhedsværdi
- reference
- virkefelt
- kausalrelationer
- kompositionalitet (betydningen af helet er en funktion af betydningen af de enkelte dele)

Lingvistikredsens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET

Sprogteknologi og formel semantik

Interesse for de elementer i naturligt sprog som har 'parallel' i logiske sprog, 'funktionsord':

- kvantorer (determinatorer, ubestemte pronominer, *alle, en*)
- operatorer (konjunktioner *og, eller, hvis, så* adverbier *ikke*)

Eks. på formel repræsentation i prædikatslogik:
Someone liked Jo and hated Fiona
 $\exists x [(like'(jo')(x) \& (hate'(fiona'))(x))]$


 Lingvistikredensens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET

Formel semantik: udvikling

Aspekter som gjorde formel semantik anvendelig for lingvister og som 'flyttede' lingvistikken:
Barwise & Cooper 1981: behandling af andre kvantorer end *en* og *alle* som fx *nogle, de fleste, mere end, mange, 2* ved at opfatte NPer som *generaliserede kvantorer* inden for mængdeteorien

$$[two'(student')] = \{X \cap [student'] \mid \geq 2\}$$


 Lingvistikredensens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET

Formel semantik og den omgivende sprogforskning

Kritikpunkter:

- hvad med den leksikalske semantik? Formelt apparat?
- hvor kompositionel er semantik egentlig? (fx. kollokationer - *få skudt noget i skoene* sammensatte ord - *moderkage*)
- hvad med overførte betydninger og metaforer? (Lakoff)
- semantikken og de kommunikative og kognitive aspekter

 Lingvistikredensens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET


Leksikon i det formelle paradigme

Leksikologi og leksikalsk semantik er kommet i fokus i det formelle paradigme pgr. 'leksikalismen'

Før: 'lexicon is a wastebin of irregularities' (Chomsky), 'list of irregularities, appendix to the grammar' (Bloomfield), 'regular variations are not matters for the lexicon, which should only contain idiosyncratic items' (Chomsky & Halle 1968)

Nu: vi taler om 'leksikalisme' inden for formelle paradigme: mere og mere strukturel information indsættes ordbogen, ordet anses for det centrale i sproget (i modsætning til sætningen)

ordbog-----grammatik

 Lingvistikredensens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET


Formel leksikalsk semantik

Fokus på **indholdsordene**: substantiver, verber, adjektiver,

Hvordan kan de beskrives i et formelt sprog?
Hvilken information ønsker vi at kunne beregne på et formelt grundlag? (Copestake&Briscoe, Calzolari, Pustejovsky)

Først og fremmest ønsker vi et formelt apparat der gør det muligt

- at udtrykke forskellige betydningsdimensioner i de leksikalske enheder
- at opløse flertydighed
- at beregne semantisk similaritet
- at udtrykke relationerne i mellem begreber

 Lingvistikredensens 75-års jubilæum

HUMANIORA
ÅRHUS UNIVERSITET


Formel leksikalsk semantik

Pustejovskys Generative Lexicon (1995)
Betydningsdimensioner i form af Qualia Structure:

- formal role: typisk 'is_a'-relation
- agentive role: oprindelse, typisk 'created_by'-relation
- telic role: formål, typisk 'used_for'-relation
- constitutive role: intern struktur, typisk 'has as parts'-relation

Generative mekanismer, fx selektiv binding:
a fast car - hurtig til at køre
a fast typist - hurtig til at skrive

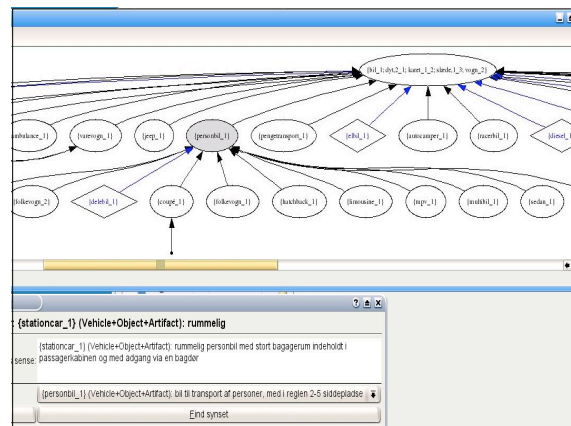
Håndtering af systematisk polysemi vha 'dotted types'
book - Artifact+Semiotic

 Lingvistikredensens 75-års jubilæum

HUMANIORA
KØBENHAVNS UNIVERSITET

DanNet – leksikalsk semantisk ordnet

Lingvistikredens 75-års jubilæum 19



HUMANIORA
KØBENHAVNS UNIVERSITET

Dansk forskning på området

Fokus på udviklingen af *danske* sprogresurser i form af:

- formelle ordbøger
- formelle grammatikker
- termbaser
- korpora: annoterede korpora, træbanker, parallelle korpora

Herudover foregår en del grundforskning bl.a. med brug af sprogresurser

Lingvistikredens 75-års jubilæum 21

HUMANIORA
KØBENHAVNS UNIVERSITET

Sprogteknologi og det omgivende samfund

Sprogteknologi har en anvendelsesside i sprogværktøjer:

- Avanceret tekstbehandling/skrivehjælp
- Informationssøgning/navigering
- Automatisk resumering
- E-læring, indlæringsystemer
- Maskinoversættelse
- Taletjenester (kunstig tale, analyse af talte sprog)

Lingvistikredens 75-års jubilæum 22

HUMANIORA
KØBENHAVNS UNIVERSITET

Fremtiden

- Sprogteknologi udfordrer sprogforskningen bl.a. fordi den opstiller modeller som kan afprøves; derved opstår nye problemstillinger og nye erkendelser
- Sprogteknologi vil påvirke sprogforskningen bl.a. via de værktøjer den stiller til rådighed
- Kravene til omfattende undersøgelse af sproglige data vil stige
- Statistikbaserede beskrivelser og regelbaserede beskrivelser vil gensidigt berige hinanden

Lingvistikredens 75-års jubilæum 23

HUMANIORA
KØBENHAVNS UNIVERSITET

Fremtiden på KU

“Med den centrale plads, som sproget har inden for humaniora, må et humanistisk fakultet uden undervisning og forskning i sprogteknologi i dag betragtes som en anakronisme” Spang-Hanssen 2006

Lingvistikredens 75-års jubilæum 24



Fremtiden på KU?

- Tværfagligt tilvalg i IT og sprog
- Forslag om sprogteknologi som et element i en kandidatuddannelse i IT og kognition

Tre søjler:

- datalogi
- sprogteknologi/lingvistik
- psykologi

Hvis uddannelsen godkendes, kan den igangsættes fra efteråret 2007.