AUTOMATISK OG MANUEL SAMMENKÆDNING AF TO RESSOURCER: STO OG DANNET

- PRÆCISERING AF METODER OG SPECIFIKATIONER

Denne rapport er struktureret på følgende vis

Sektion I. omhandler generelle spørgsmål i relation til arbejdspakkens målsætning
Sektion II. rapporterer om de beslutninger der er taget i forbindelse med de spørgsmål der er skitseret i Sektion I vedrørende lemmaselektion og sammenkædning

SEKTION I

Arbejdspakkens målsætning

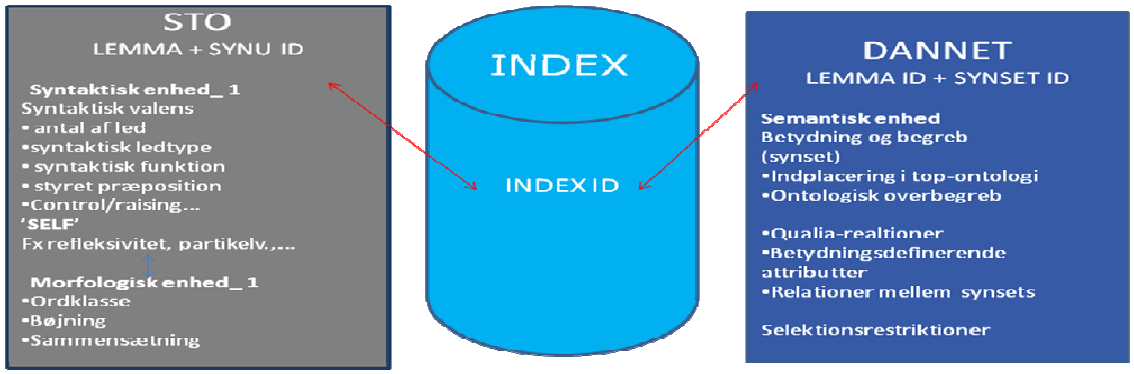
- Sammenføring af STO og DanNet

Arbejdspakke WP4.2.2 bidrager til DK-CLARINs infrastruktur med sammenkædning af to leksikalske ressourcer der er udarbejdet for sprogteknologiske og datalingvistiske anvendelser, for at få det fulde udbytte af disse ressourcer. I praksis betyder det at der etableres link mellem sammenhørende enheder i disse to databaser. Den samlede ressource har den foreløbige betegnelse STO-Net.

Der skal produceres en indeksfil bestående af links der sammenkæder en relevant delmængde af STOs ordforråd med DanNets ordforråd. Udfordringen ligger i at STOs *syntaksbaserede læsningsopdeling* (med udgangspunkt i ords konstruktionsmønstre) skal relateres til DanNets *semantikbaserede læsningsopdeling* (betydning).

STO og DanNet forbliver i hver sin database og struktur, således at for at have adgang til den samlede ressource, dvs. lemmaer med morfologiske, syntaktiske og semantiske oplysninger, har man brug for tre ressourceenheder (jf. Figur 1):

- Indeksfilen
- STO-basen (i DK-CLARIN tilgængelig i XML) se [1]
- DanNet-basen (i DK-CLARIN tilgængelig i RDF/OWL eller kommasepareret fil) se [2]



Figur 1. Sammenkædning af leksikalske beskrivelser af ord i to selvstændige ressourcer

STO-Net vil indgå som leksikalsk ressource i DK-CLARIN.

Diskussion af generelle spørgsmål og problemer

I afsnittet nedenfor beskrives to aspekter der er afgørende for sammenkædningen af STOs syntaks og DanNets semantik. Det første vedrører kompatibiliteten mellem beskrivelsen af et lemmas syntaktiske konstruktioner og dets betydninger. Det andet aspekt er STOs strategi mht. styrede leds valgfrihed i syntaktisk beskrivelse (’brede’ vs. ’smalle’ syntaktiske mønstre), hvilket er en vigtig faktor i præcisionen af sammenkædningen.

Relationen mellem syntaktiske og semantiske læsninger af et lemma

STOs *syntaksbaserede* og DanNets *semantikbaserede læsningsopdeling* er væsensforskellige, og forholdet mellem et lemmas syntaktiske konstruktioner og betydninger kan være ganske komplekst. Dette spørgsmål er behandlet i rapporten om forundersøgelsen af hvordan de to ressourcer principielt set kan sammenkædes [3], jf. afsnittene fra rapporten nedenfor.

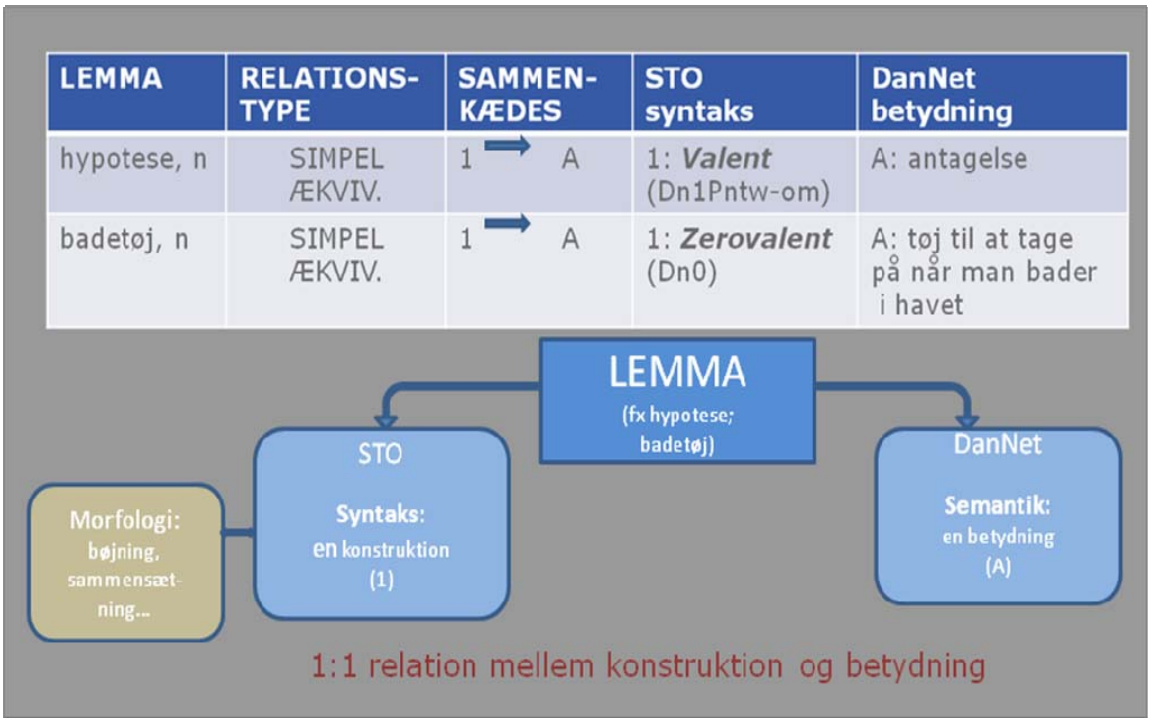
> Sammenkædningen mellem STO og DanNet baseres på sammenkædning af SynU’er/synsets efter to overordnede principper: den *simple* og den *komplekse sammenkædning*. Den simple sammenkædning refererer til en-til-en sammenkædninger, dvs. at der kan etableres en relation mellem én STO-SynU og én DanNet-synset. Den komplekse sammenkædning involverer mere end en SynU/synset fra en eller begge ressourcer. Med andre ord: denne sammenkædning refererer til en-til-mange, mange-til-en eller mange-til-mange relationer.
>
> **Relationstyper i en sammenkædning**
>
> Det faktum at sammenkædningen er baseret på SynU’er/synsets betyder at der mellem et STO-lemma og et DanNet-lemma, kan eksistere flere relationstyper hvis de to lemmaer dækker over flere SynU’er/synsets. Fx kan SynU’erne for et lemma i STO forholde sig på forskellig vis til tilsvarende synsets i DanNet, således at der mellem en SynU og et synset eksisterer fuld ækvivalens, mens et andet SynU for samme STO-lemma kun i begrænset omfang svarer til et andet synset for samme

DanNet-lemma. Det skal kunne udtrykkes at der kan eksistere flere relationstyper mellem et lemma i STO og et lemma i DanNet.

Relationstyper i en simpel sammenkædning:

- *Ækvivalens* – der eksister fuld overensstemmelse mellem de to elementer i sammenkædningen, dvs. en SynU i STO svarer til et synset i DanNet
- *Sammmenlignelighed* – en SynU i STO svarer i hovedtræk til et synset i DanNet, men der er ikke fuld overensstemmelse
- *Simpel delmængde* - STO-SynU dækker en delmængde af DanNet-synset

Figur2 viser det det simpleste relationstype (fuld ækvivalens) mellem en lemmas beskrivelse med syntaks i STO (en syntaktisk konstruktion) og dens semantiske beskrivelse i DanNet (betydning/element i et synset). Arbejdspakkens hovedopgave er at sammenkæde syntaks og semantik inden for denne relationstype.



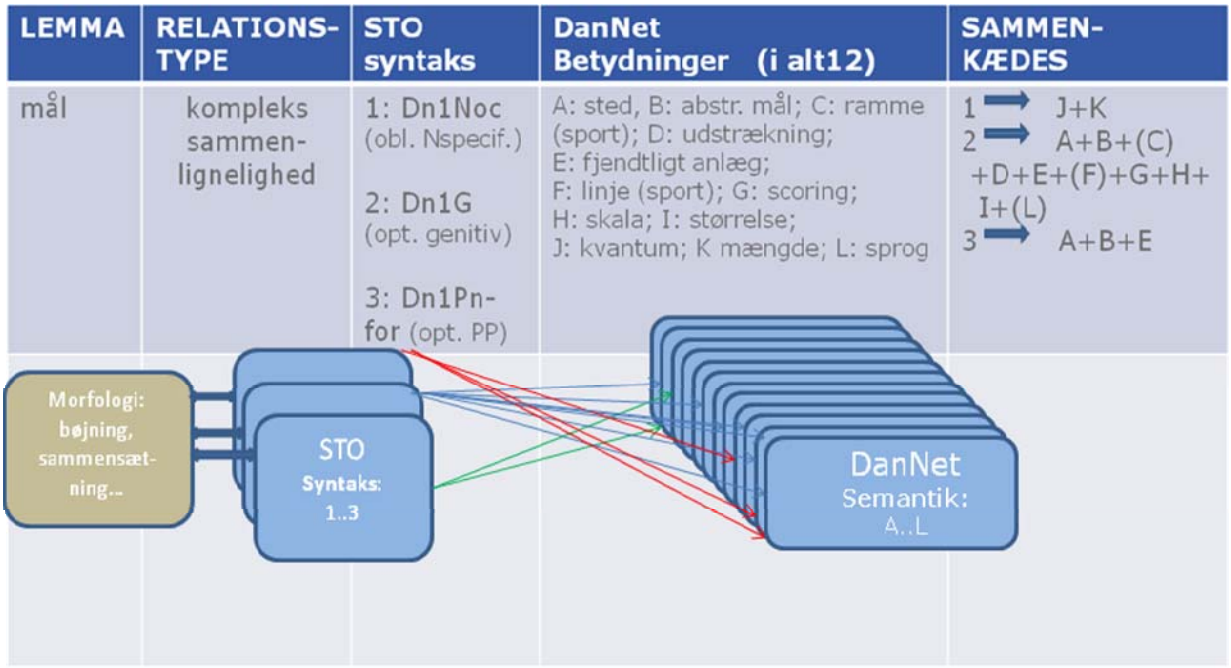
Figur 2. Simpel ækvivalensrelation (1:1) mellem syntaktisk konstruktion og betydning

Homografer samt lemmaer der har flere syntaktiske enheder og flere betydninger hører i stor udstrækning til en anden relationstype, jf. afsnittet fra [3] og *Figur 3.*

Relationstyper i en kompleks sammenkædning
I en kompleks sammenkædning indgår som nævnt flere SynU'er/synsets. De SynU'er/synsets som skal indgå i sammenkædningen, kan endvidere tilhøre en eller flere indgange. Fx findes i DanNet to separate indgange for *pande* mens der i STO kun findes én indgang med de tilsvarende syntaktiske oplysninger for *pande*.

Figur 3 viser et eksempel på *komplekse relationer* mellem et ords syntaktiske konstruktioner og ordets betydninger: en given konstruktion af ordet kan forbindes med mere end en betydning, og en betydning kan udtrykkes med forskellige konstruktioner (henholdsvis 1:n, n:1 og n:n relationer). Kolonnen med DanNet betydninger er i dette eksempel udfyldt med de betydninger der er - eller skal - være kodede i DanNet. DanNet beskriver først og fremmest de prototypiske og hyppigst forekommende betydninger af et ord, mindre kendte, sjældnere betydninger er i den nuværende version endnu ikke medtaget.
I denne rapport fokuseres på de aspekter der kræver flere overvejelser og en revision med henblik på en endelig beslutning.

Til løsningen af problemet med komplekse relationer mellem syntaktisk enhed og betydning(er) skal der udvikles en strategi der sørger for at der linkes mellem syntaktiske og relevante semantiske enheder med en rimelig god præcision. Denne strategi er under udarbejdelse under hensyntagen til detaljeret diskussion af problemet i [3], og problemet vil blive håndteret i en vis udstrækning.



Figur 3. Komplekse relationer mellem syntaktiske konstruktioner og betydninger

STOs strategi for syntaktisk læsningsopdeling: 'Brede' vs. 'smalle' syntakmønstre
Et andet centralt punkt i sammenkædningen af STOs syntaktiske enheder (SynUs) med DanNets tilsvarende 'synsets' er at i den syntaksbaserede læsningsopdeling af lemmaer i syntaktiske enheder er der fulgt forskellige principper i STO, afhængig af lemmaets ordklasse.

Substantiver er kodet med brede syntaksmønstre der dækker flere syntaktiske strukturer med nul og et eller flere komplementer fordi et substantivs komplementer altid betragtes som optionelle, dvs. disse kan være realiseret eller mangle i konstruktionen. Derfor kan et ords lejlighedsvis elliptiske konstruktion og altid komplementløse konstruktion falde sammen, selv om der er tale om to forskellige semantiske læsninger (betydninger) af ordet, jf. *Eks. 1*.

Eks. 1: *'dannelse'*, substantiv

- STO: to syntaksmønstre (konstruktioner)

KONSTRUKTION 1: med optionel genitiv *(STO–kode: Dn1G)*
KONSTRUKTION 2: med et optionelt præpositionsobjekt der består af præposition =af og styrelse =Nominalsyntagme *(STO–kode: Dn1Pn-af)*

- DanNet: to betydninger beskrevet med to forskellige overordnede begreber (GenProx)

BETYDNING A: frembringelse
BETYDNING B: kendskab

Forhold mellem SynU og Synset:
Konstruktion 1 forekommer
- med den optionelle genitiv i
Betydning A "Solsystemets dannelse foregik ved at …" og i
Betydning B: "Thorvaldsens dannelse…"
- uden den optionelle genitiv i
Betydning B: Hvad er *dannelse* i en tid, hvor Google har overhalet Goethe?
(Betydning A: kun meget sjældent, i enkelte, elliptiske konstruktioner hvor 'dannelse af…' er nævnt i sætningen før.)

Konstruktion 2 forekommer kun i
Betydning A: "Dannelsen *af* passiv i dansk…"
(Betydning B: ikke fundet i korpus/på Google)

Denne underspecificerende kodningsstrategi er valgt for at undgå utilsigtet overgenerering ved syntaktisk parsning af tekster. I tilfælde af semantisk entydiggørelse (disambiguering) vil det derimod være fordelagtigt at have mere præcise opdelinger i rammer med og uden komplement (fx *administration* i 'bygning, afdeling' -læsningen er zerovalent, hvorimod i 'handling' -læsning er divalent, men stadig med optionelle led). Endvidere medfører denne kodningsstrategi at det er sværere at etablere præcise links mellem en given syntaktisk konstruktion af ordet og de/n relevante betydning/er.

Vi har overvejet at udfolde (specificere) disse konstruktionsmønstre således at et 'bredt' mønster skulle blive opdelt i flere, mere præcise, 'smalle' mønstre for derved at kunne styre linket mere præcist. Der er stærke argumenter der taler imod en sådan udspecificering. Vi overvejer nøje, hvorvidt en splitning overhovedet ville løse problemet hvis de pågældende substantiver alligevel også skal knyttes til en zerovalent syntaktisk læsning for at kunne dække elliptiske konstruktioner. Desuden ville en sådan splitning indebære en grundlæggende ændring i STOs beskrivelsesstrategi, hvilket er uhensigtsmæssig af flere grunde. Metoden ville have uønskede implikationer fx i form af væsentlig forøgelse af konstruktionsmønstre og syntaktiske enheder.

Ydermere ville en sådan udspecificering kræve en ganske stor arbejdsindsats, hvilket skal stå i forhold til den potentielle gevinst mht. præcision.

Verbernes og adjektivernes syntaksbeskrivelse følger det princip, at alle komplementer er obligatoriske, med mindre de er markeret som optionelle, hvilket er i overensstemmelse med den almene valensteoretiske antagelse om verbers valens, jf. Eks. 2.

Synset: sæt af synonyme betydninger

I forbindelse med etablering af link mellem et ords syntaktiske og semantiske læsning rejser sig et relevant spørgsmål: Hvorvidt/I hvilken udstrækning har en given betydning og dens synonymer ens syntaksmønstre? Dette spørgsmål har forskningsmæssig interesse, men kan først besvares efter en detaljeret empirisk undersøgelse, baseret på et bredt udvalg af ordmateriale, og det falder uden for denne arbejdspakke.

I den foreløbige version af indeksering har vi besluttet at link skal etableres mellem en syntaktisk enhed (STO: SynU) og ordets betydning(er) i den pågældende syntaktiske konstruktion (DanNet). Hvis denne betydning indgår i et synonymsæt (DanNet: synset), så kan denne oplysning og rækken af de pågældende synonymer samt det fælles overbegreb (GenProx) tilgås indirekte over den lænkede betydning, men der bliver ikke linket direkte til synsettets øvrige medlemmer. Man vil på samme måde kunne tilgå en betydnings underordnede begreber (hyponymer)

SEKTION II

Selektion af ordforrådet for sammenkædning

Grundbetingelse for selektionen af et lemma er at det skal være repræsenteret med mindst en syntaktisk beskrivelse i STO og med mindst en semantisk beskrivelse i DanNet.
Mål: 9.000 lemmaer skal selekteres for sammenkædning

Metode for lemmaselektion

For at kunne nå frem til den relevante fællesmængde af ordforrådet der er beskrevet både i STO og i DanNet, skal der således genereres to lemmalister med udvalgte oplysningstyper der beskriver de enkelte lemmaer på det givne niveau. Derefter sammenlignes disse to lemmalister.

STOs lemmaliste skal for hver Mu_id indeholde følgende oplysninger:

- Normaliseret Mu_id ~ lemma
- ordklasse
- SynU-deskriptor(er) 1-n (formaliseret konstruktionsmønster)
- frekvenstal
- SynU id.

DanNets lemmaliste skal i første omgang kun indeholde lemmaet og dets

- synset
- ordklasse
- lemma-id.

For udtrækning af fællesmængden af lemmaer i de to ressourcer udtrækkes ved at se på om STO-lemmaet (+ dets ordklasse) har et tilsvarende lemma med den samme ordklasse i DanNet. Dette udtræk vil indeholde alle lemmaer der er kodet i begge ressourcer. Da arbejdspakken skal fokusere på de hyppigste ord, skal der foretages en selektion baseret på frekvenstallene i STO.

STO indeholder frekvensinformation baseret på Korpus2000 og Korpus90. Det er valgt at basere frekvensselektionen i denne opgave på Korpus2000-frekvenserne, da disse er baseret på det yngste korpus. Frekvensinformationen består både af frekvenstal, der er baseret på disambiguering af en pos-tagger og ordformsfrekvenser, der er tildelt uafhængigt af taggerens forslag.

Da vi i denne opgave ønsker at udvælge lemmaer baseret på frekvens foretager vi to tilnærmelser. STO's mu_id konverteres til en tilnærmet lemmaform og frekvenstallene for de enkelte ordformer for en given mu adderes. Den første tilnærmelse er betinget af at SynU'erne i STO er knyttet til MU_id'er, og at der ikke findes en bedre lemma-specifikation i STO allerede. Den anden tilnærmelse resulterer i at visse indgange som deler ordformer med andre meget hyppigt anvendte indgange får en for stor frekvens. Effekten af denne udvælgelse i forhold til at danne MU-frekvenser baseret på frekvenser på disambiguerede ordformer vil blive vurderet senere.

For den nuværende version af DanNet(v.1.0.1) findes der 14307 indgange i STO hvor STO's lemmaværdi findes i DanNet og hvor indgangen har en K2000-ordformsfrekvens på mindst 20.

Der findes 8581 zerovalente substantiver i STO hvor STO's lemmaværdi findes i DanNet og hvor indgangen har en K2000-ordformsfrekvens på mindst 20. Disse 8581 zerovalente substantiver kan kodes automatisk.

Der frigives snart en ny version af DanNet som har en større mængde ord kodet. Der kan i den anledning uddrages en ekstra mængde ord til automatisk kodning.

Principper for selektion af kandidater der er velegnede til at blive lænket

Fra fællesmængden med den relevante frekvens har vi brug for lemmaer til to typer lænkning hvortil vi anvender de formaliserede syntaktiske deskriptorer i STO som grundlæggende kriterium.

Automatisk lænkning (ca. 8.500)

- zerovalente substantiver og substantiver der kun har én konstruktion (SynU) og én betydning SynU, jf. Figur 2. I sådanne tilfælde er relationen mellem syntaktisk og semantisk læsning 1:1. Det drejer sig for den største dels vedkommende om konkrete substantiver, fx *bord, lampe, badetøj, viskelæder, højttaler.*

Manuel lænkning (ca. 500)
Disse skal danne et testsæt for videre undersøgelser bestående af

- valente substantiver der har 'brede' eller komplekse rammer med henblik på at teste behovet for og effekten af udspecificering af syntaksmønstre (jf. diskussionen vedr. '*dannelse*' med

deskriptorerne Dn1G og Dn1Pn-af). Der findes i øjeblikket 2500 ikke-zerovalente fælles substantiver mellem DanNet og STO med en K2000-ordformsfrekvens på mindst 20.

- verber og adjektiver, da disse ordklasser er mere restriktivt kodede mht. komplementernes valgfrie realisering (optionalitet) end substantiverne. Samtidig har verber mange syntaktiske konstruktionsmuligheder og semantiske læsninger, og disse står oftest i komplekse relationer (jf. Figur 3) til hinanden. Der findes i øjeblikket 2500 fælles verber og 660 fælles adjektiver mellem DanNet og STO med en K2000-ordformsfrekvens på mindst 20.

Formålet med at arbejde med et sådant testsæt af lemmaer er at træffe afgørelse om spørgsmålet vedr. udspecificering af syntaktiske rammer. I denne sammenhæng skal vi

- vurdere om brugen af brede rammer og dermed en 'grov' sammenkædning vil have for store ulemper i forhold til semantikken, dvs. at der vil være mange upræcise eller mangelfulde lænkninger fordi linket enten ikke er præcist nok til at kunne disambiguere flertydige ords læsninger eller der overhovedet ikke kan etableres noget link til enkelte (frekvente) læsninger
- overveje opsplitningen af brede (underspecificerede) rammer som ændring i kodningsstrategien og de lingvistiske implikationer og hvor meget en sådan opsplitning vil betyde for XML-strukturen mht. arbejdsbyrde
- vurdere om en sådan udspecificering vil forbedre fx semantisk opmærkning af tekster og/eller hvorvidt den vil medføre en betydelig (og uønsket) overgenerering.

En anden strategisk mulighed er i første omgang kun at etablere link mellem syntaks og semantik for alle (tilpas frekvente) zerovalente substantiver fra fællesmængde-listen. I dette tilfælde vil der være tale om simple 1:1 eller 1:n relationer. Dette vil give en talmæssigt større ressource (9.000 +?) enten hvis der vælges en frekvensgrænse under 20 eller hvis den kommende version af DanNet inddrages. Denne lænkning vil kunne gøres uden et hjælpeværktøj med brugerinterface og vil være forholdsvis enkelt at programmere.
Fordelen ved denne strategi vil være at vi får en større leksikalsk dækningsgrad, hvorimod bredden af den lingvistiske dækning vil være ganske begrænset.

Referencer:

[1] http://www.cst.ku.dk/sto
[2] http://wordnet.dk/dannet
[3] Pedersen, B.S., Braasch, A., Henriksen, L., Olsen, S., Povlsen, C.: Sammenføring af ordbogsressourcer: STO og DanNet. Rapport fra et pilotprojekt. December 2007.