

STO - EN SPROGTEKNOLOGISK DATABASE

Af Sussi Olsen (Center for Sprogteknologi)

1. Indledning

Denne artikel beskriver STO-ordbasen, den sprogteknologiske ordbase, som er resultatet af et treårigt projekt støttet af Ministeriet for Videnskab, Teknologi og Udvikling. Udover at beskrive ordbasens ordforråd, dækningsgrad og struktureringen af de lingvistiske oplysninger, vil der blive givet nogle eksempler på hvordan den meget strukturerede og detaljerede beskrivelse afdækker træk ved de enkelte ord, her nogle morfologiske træk, som ikke har været belyst før. Desuden gives der en kort beskrivelse af STO's brugergrænseflade på internettet.

2. STO-projektet, formål og baggrund

STO-projektet var et 3-årigt projekt, der løb fra 2001-2004 støttet af Ministeriet for Videnskab, Teknologi og Udvikling med det formål at skabe en sprogteknologisk ordbog der kunne afhjælpe den flaskehals som danske sprogteknologiske projekter i årevis er løbet ind pga. det store arbejde der ligger i at skabe sprogteknologiske ordbogsdata af en vis mængde. På sin vis startede projektet dog før idet Center for Sprogteknologi havde gjort forstudier til ordbasen via et mindre pilotprojekt. Ordbasen bygger delvist videre på lingvistiske og strukturelle beskrivelser fra de europæiske ordbogsprojekter Parole (LE_PAROLE 1998) og Simple (Lenci et al. 2000), og data fra disse projekter indgår i modificeret form i den nuværende STO-ordbase.

Formålet med STO-projektet var at producere en dansk sprogteknologisk ordbase på 45.000 lemmaer med morfologiske og syntaktiske oplysninger, hvoraf ca. 10.000 lemmaer skulle forsynes med semantisk information af forskellig detaljeringsgrad. Ordbasen er beregnet til maskinel anvendelse i alle slags datalingvistiske og sprogteknologiske projekter, såvel af forskningsmæssig som af udviklingsmæssig karakter.

Projektet var et samarbejdsprojekt der ud over Center for Sprogteknologi som projektleder inkluderede følgende samarbejdspartnere: Institut for Datalingvistik, HHK, Institut for Almen og Anvendt Lingvistik, KU samt Institut for Erhvervsinformatik, Kolding.

3. Ordforråd og korpus

Det blev tidligt besluttet at STO skulle være korpusbaseret. Dette betød at vi i samtlige processer prioriterede hvad vi fandt i korpus over leksikografernes introspektion, naturligvis under hensyntagen til antallet af forekomster i de anvendte korpusser. Korpus blev inddraget i alle processer, både ved lemmaselektion og ved morfologisk og syntaktisk og semantisk kodning. At fastlægge retningslinjer for korpusbrug og at fastholde tolkningen af disse retningslinjer blandt en flok leksikografer spredt på flere institutioner viste sig at være en af de vanskeligere opgaver i projektet.

Ordforrådet i STO består dels af et almensprogligt ordforråd og dels af ordforråd fra seks domæner. Det almensproglige ordforråd er udvalgt fra lister udarbejdet af Den Danske Ordbog med hyppige ord i dansk. En liste over de hyppigste blev verificeret via et aviskorpus og Korpus2000.

Følgende skema viser ordforrådet i STO-basen fordelt på ordklasser, med oplysninger om deres lingvistiske beskrivelsesniveau:

Leksikalsk kategori	Antal lemmaer	Morfologi	Syntaks	Semantik
Substantiv	64.740	100 %	53 %	12 %
Adjektiv	9770	100 %	68 %	13 %
Verbum	5775	100 %	98 %	17 %
Adverbium	770	100 %	19 %	
Interjektion	160	100 %	0 %	
Præposition	80	100 %	0 %	
Konjunktion	60	100 %	0 %	
Pronomen	45	100 %	0 %	
Diverse	130	100 %	0 %	
Total	81.530			

Som det ses blev det totale ordforråd med morfologiske oplysninger langt større end planlagt da det viste sig at den morfologiske kodning gik langt hurtigere end ventet mens det ikke var muligt at overskride antallet med syntaks og semantik. Samtidig blev vi opmærksomme på at der var stor interesse for store mængder ord med morfologiske oplysninger. De hyppigste ord fra morfologien blev valgt til syntakskodning, mens ordforrådet med semantik dels består af 7000 almensproglige ord dels af ordforrådet fra domænet sundhed/helse på ca. 2500 ord.

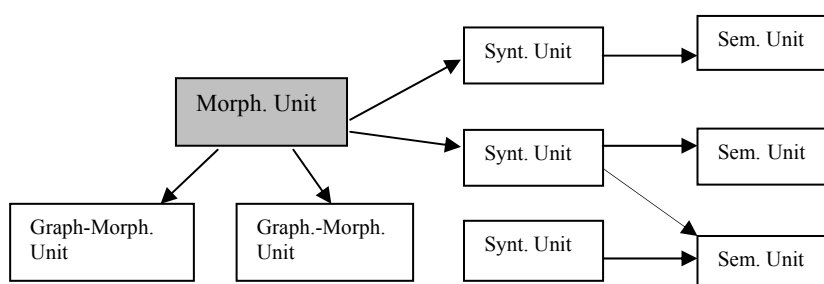
Det fagsproglige ordforråd blev udvalgt via tekster indsamlet fra nettet. Ordforrådet skulle ikke være faglige termer, men derimod ord der ligger tæt op ad det almene ordforråd, men er særligt relevante for det aktuelle domæne, såkaldte gråzone-ord. Derfor blev der valgt tekster skrevet af eksperter til lægmand og disse blev samlet i korpusser på mellem 1 og 2 mio. løbende ord (se Jørgensen 2001). De seks domæner blev: IT, miljø, handel/erhverv, forvaltning, sundhed/helse samt finans.

Det fagsproglige ordforråd fordeler sig på de forskellige domæner som det fremgår af følgende skema:

Domæne	Substantiv	Verbum	Adjektiv	Total
IT	1730	160	115	2005
Miljø	1770	50	300	2120
Handel/Erhverv	1800	60	160	2020
Forvaltning	2430	25	220	2675
Sundhed/Helse	2285	40	250	2575
Finans	1880	30	160	2070
Total				13.465

4. Lingvistisk beskrivelsesmodel

Et af de karakteristiske træk ved STO er den skarpe opdeling af det morfologiske, syntaktiske og semantiske niveau, samlet i hhv. morfologiske, syntaktiske og semantiske enheder som det ses nedenfor:



Den morfologiske enhed (med tilhørende grafisk morfologisk enhed) indeholder ordklasseangivelse, stavemåder (fx hæfte/hefte), bøjningsmønstre (ofte mere end et), dekomposition (for sammensatte substantiver), fugeelement (for substantiver), autonomi (angiver om ord kun kan forekomme i kollokationer) samt om ordet er godkendt i denne stavemåde af Retsskrivningsordbogen. Desuden indeholder den oplysninger om status, oprindelse og frekvens.

Den syntaktiske enhed indeholder oplysninger om ordets valensramme med angivelse af aritet, den kanoniske rækkefølge af de styrede led samt om et led er obligatorisk eller optionelt. Desuden oplyses om refleksivitet, partikelverbum, leddenes syntaktiske kategori (NP, PP, at-sætning, infinitiv og spørgende sætning), leddenes syntaktiske funktion (subjekt, objekt, indirekte objekt), samt for de fleste ord et eksempel (til brug for mennesker).

Nedenfor følger to eksempler på syntaktiske enheder:

Lemma: **indsprøjte**

Description: **Dv3NP_a0-i** → Trivalent verbum, subjekt (ikke udtrykt), objekt=NP, adverbialt PP-objekt (optionelt) indledt af præp. 'i'.

Eksempel: Lægen **indsprøjter** stoffet i patienten

Lemma: **præference**

Description: **Dn2GP_{ni}-for** → Divalent substantiv, valensbunden genitiv, optionelt PP-objekt indledt af præp. 'for', hvor styrelsen kan være et NP eller en infinitiv.

Eksempel: Frischs **præference** for arbejderklassen

Den semantiske enhed er ikke en entydig størrelse. Da projektet i opstartsfasen blev beskåret på netop semantikdelen, blev målet at integrere semantikdataene fra SIMPLE-projektet og derudover at give ordene fra et af domænerne mere begrænset semantisk information end det var tilfældet i Simple-projektet. Ordforrådet er derfor nu beskrevet med semantik i tre forskellige detaljeringsgrader: **Niveau 1:** angivelse af domæne som eneste semantiske information (alle ord), **Niveau 2:** angivelse af domæne, betydningsopdeling, ontologisk tilhørsforhold ud fra SIMPLE-ontologien, argumentstruktur og selektionsrestriktioner (2500 sundhedsord) **Niveau 3:** informationer fra niveau 2 + qualiaroller samt polysemi- og synonymirelationer på basis af SIMPLE-modellen (7000 almensproglige ord).

Følgende eksempel viser en semantisk enhed på niveau 2:

Korpuseksempel: Læger **indsprøjter** stoffet N. i narkoofre.

Ontologisk tilhørsforhold: PurposeAct

Event Act **Relational act** Purpose act

Syntaktisk konstruktion: Dv3NP_a0-i

Verbum trivalent, styrede led: subjekt (default), NP objekt, optionel modifikator med præposition 'i'

Argumentstruktur: arg12_ADJi

subjekt=arg1, objekt=arg2, præpositionssyntagme=adjunct

Selektionsrestriktioner: ARG12ADJhum_sub_bpahum

subjekt: human, objekt: substance, adjunct: body part/hum

5. Bøjningsmønstre – remove/add-metoden

STOs bøjningssystem er baseret på en automatisk beregningsmetode, den såkaldte remove/add-metode. Der arbejdes ikke med regler og dermed heller ikke med regelmæssige/uregelmæssige bøjninger, og der er ikke tale om at skelne mellem stamme og fleksiv, kun om ren beregning, som det fremgår nedenfor:

Lemma		removed	added	resultat
Stol	Ubestemt sing	0	0	stol
	Bestemt sing	0	en	stolen
	Ubestemt plur	0	e	stole
	Bestemt plur	0	ene	stolene

Hvert unikt sæt af fjernede og tilsatte 'tekststreng' samles i et bøjningsmønster, og dette mønster gælder så for en lang række ord som i eksemplet ovenfor, eller for kun ét enkelt ord som i eksemplet nedenfor:

Lemma		removed	added	resultat
barnebarn	Ubestemt sing	0	0	barnebarn
	Bestemt sing	0	et	barnebarnet
	Ubestemt plur	arnebarn	ørnebørn	børnebørn
	Bestemt plur	arnebarn	ørnebørnene	børnebørnene

Dette system er yderst velegnet og hurtigt til maskinel behandling af data, men pga. de manglende generaliseringer og det store antal mønstre (ca. 700), er det vanskeligt at håndtere for leksikografen, så vi var nødt til at udarbejde et værktøj hvor mønstrene var grupperet efter almindelige morfologiske principper såsom konsonantfordobling, synkope, vokalskifte etc. til brug ved kodningen af ord.

6. Eksempler på detaljeret morfologisk information

Det faktum at STO er korpusbaseret, og at et bøjningsmønster afspejler nøjagtig de former af et ord som kan findes i korpus, har vist sig at føre til nye interessante opdagelser om sprogets opførsel. Et eksempel på dette er verbet *give* og de mange sammensætninger der dannes heraf.

Et verbs bøjningsmønster indeholder temmelig mange former som det fremgår af nedenstående skema, bl.a. fordi vi også medtager participiernes forskellige bøjningsformer når disse bruges adjektivisk.

Ordform	Genus	Numerus	Bestemthed	Tid	Modus	Diatesse	Transcat
give					infinitiv	aktiv	
gives					infinitiv	passiv	
giver				præsens	indikativ	aktiv	
gives				præsens	indikativ	passiv	
gav				præterit	indikativ	aktiv	
gaves				præterit	indikativ	passiv	
giv					imperativ		
givende				præsens	particip		
givet				præterit	particip		
given	fælles	singularis	umarkeret		(gerund)		transnominal
givende	umarkeret	umarkeret	umarkeret	præsens	particip		transadjectival
given	fælles	singularis	ubestemt	præterit	particip		transadjectival
givet	neutrum	singularis	ubestemt	præterit	particip		transadjectival
givne	umarkeret	singularis	bestemt	præterit	particip		transadjectival
givne	umarkeret	pluralis	umarkeret	præterit	particip		transadjectival

Det er en velkendt at nogle verber ikke har passiv, dette gælder fx de obligatorisk refleksive (*skammes), samt mange uregelmæssige verber der ikke bruges i præteritum passiv (*ådes). Der findes imidlertid ikke en gennemgang af præcis hvilke verber det sidste fænomen angår. Ved kodningen af *give* og dets sammensætninger viste det sig at *give* stadig tillader brugen af præteritum passiv (*der gaves...*) om end formen er sjælden på nudansk. Men det samme gælder ikke for alle sammensætninger af *give*, som det ses af nedenstående eksempler:

- Alle passive former bruges: *tilkendegive*, *afgive*...

- Ingen passive former overhovedet: *begive, hengive* (refleksivitet)
- Kun præsens passiv: *medgive, foregive, opgive...*

For at kunne tildele det rigtige bøjningsmønster er det nødvendigt at undersøge hvert eneste ord i detaljer.

Samtidig spiller netop de adjektivisk brugte participier også en vigtig rolle når der skal tildeles entydigt bøjningsmønster til disse verber da participier af uregelmæssige verber kan ende på –n eller –t når de modificerer et ubestemt substantiv af fælleskøn, og for nogle af verberne lever begge former side om side

- en forsvunden person 30 forekomster i korpus
- en forsvundet person 7 forekomster i korpus

og begge former skal naturligvis være afspejlet i ordbasen.

Tildelingen af bøjningsmønstre til verbet *give* og sammensætningerne deraf fik derfor endnu en komplikation da der også skulle tages højde for participiumsformen.

n-form: *give, hengive*

t-form: *udgive, tilgive, gengive...*

begge former: *frigive, navngive, angive...*

Give og dets sammensætninger involverer dermed seks forskellige bøjningsparadigmer, og de af verberne der har begge participiumsformer, skal have tildelt to mønstre.

Eksempler på opdagelser af denne art er der mange af også for andre ordklasser. Eksempelvis registreres syntetisk komparation i adjektivernes bøjningsmønstre, hvorfor det skal undersøges for hvert eneste adjektiv om det kan kompareres syntetisk, en oplysning der ikke normalt optræder i en ordbog og kun behandles på regelbasis i grammatikkerne.

Den morfologiske tilgang i STO medfører altså at man kan finde bøjningsdetaljer i STO som normalt ikke findes registreret i en ordbog, heller ikke i Retskrivningsordbogen.

7. Anvendelsesmuligheder for STO

Det var målet at ordbasen skulle kunne udnyttes som basisprodukt i så mange forskellige anvendelser som muligt, fx i online informationssøgning, hjælpeværktøjer til at oversætte tekster, talegenkendelse og -syntese, sproglige hjælpemidler for handicappede og undervisningsprogrammer.

Der har allerede i projektets løbetid været adskillige henvendelser fra både forskere og udviklere der ville bruge STO-data i såvel sprogteknologiske applikationer, såsom udviklingen af en lemmatiser, semantisk-baserede søgemaskiner og maskinoversættelse som i mere alment lingvistiske forskningsprojekter, datalingvistiske forskningsprojekter, ph.d.-projekter

og projektopgaver, samt til ordbogsudvikling. Og interessen fortsætter her efter projektets afslutning. Også projektets brugergrænseflade på internettet (se nedenfor) har stor søgning af alle mulige professionelle brugere af dansk

8. STOs grænseflade

STO fik i 2002 en bevilling fra Danmarks Elektroniske Forskningsbibliotek, DEF, til udvikling af en internetbaseret brugergrænseflade til STO. Grænsefladen har hovedsagelig interesse for menneskelige brugere, især studerende og forskere i dansk sprog, hvor sidstnævnte, og det gælder både datalingvister, almene lingvister og dansklærere, bruger grænsefladen og dermed STO-data i kurser og undervisning.

Brugergrænsefladen giver adgang til at lave ordsøgninger hvor et ords morfologiske og syntaktiske oplysninger vises. Der kan også søges på ordsammensætninger hvor søgeordet indgår, og der kan foretages opslag med det pågældende ord i et aviskorpus. Endelig kan man søge på særlige mønstre af morfologisk eller syntaktisk karakter, fx en særlig type bøjning eller en bestemt syntaktisk ramme, og få en liste på max 30 ord fra databasen der matcher de stillede kriterier. Grænsefladen er tilgængelig på <http://www.cst.dk>.

9. Opsummering og fremtidsperspektiver

Med STO har Danmark fået den forholdsvis store ordbase som ofte har været flaskehals for udviklingen af sprogteknologiske applikationer. En udvidelse på flere fronter ville være ønskelig, men dette kræver finansiering ligesom vedligeholdelse af databasen kræver finansiering. For tiden bliver STO udvidet med udtaleinformation via et projekt på Handelshøjskolen i København, og derudover står en udvidelse med kollokationer samt semantik på en større del af ordforrådet meget højt på ønskesedlen. Fremtidsvisionen er naturligvis at STO kan sælges til forskellige kommercielle projekter og derigennem blive vedligeholdt, opdateret og udvidet som ovenfor nævnt og også gerne med et større ordforråd.

Henvendelse angående erhvervelse af STO-data foregår til

- Center for Sprogteknologi

Anna Braasch, anna@cst.dk

Hjemmeside for STO-projektet og søgninger via grænsefladen: www.cst.dk

Litteratur

Braasch, A., C. Navarretta, S. Nimb, S. Olsen, B. Pedersen, C. Povlsen (2004) *Lingvistiske Specifikationer for STO*. Under udgivelse.

- Braasch, A. (2002) Current Developments of STO – the Danish Lexicon Project for NLP and HLT Applications. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 986-992). Las Palmas.
- Braasch, A. & B. S. Pedersen (2002) Recent Work in the Danish Computational Lexicon Project „STO“. In: *Proceedings from the Tenth Euralex International Congress* (pp.301-314), CST, Copenhagen.
- Jørgensen, S. W., C. Hansen, J. Drost, D. Haltrup, A. Braasch, S. Olsen (2003) Domain specific corpus building and lemma selection in a computational lexicon. In: *Corpus Linguistics 2003 Proceedings* (pp. 374-383). Lancaster.
- Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruini, M. Villegas, A. Zampolli (2000) SIMPLE: A General Framework for the Development of Multilingual Lexicons. In: *International Journal of Lexicography*, Vol. 13, no. 4, (pp.249-263). OUP.
- LE-PAROLE, (1998) *Danish Lexicon Documentation*. Internal report, Center for Sprogteknologi, Copenhagen.
- Nimb, S. (2004) A corpus-based syntactic lexicon for adverbs. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon.
- Olsen, S. (2002) Lemma selection in domain specific computational lexica – Some specific problems. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp.1904-1908). Las Palmas.