NyS

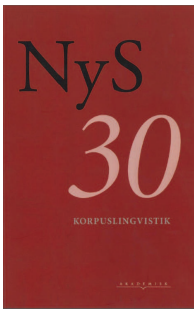
Titel: Danske resurser til automatisk opmærkning

Forfatter: Dorte Haltrup

Kilde: *NyS – Nydanske Sprogstudier* 30. *Korpuslingvistik,* 2002, s. 59-67

Udgivet af: Akademisk Forlag A/S

URL: www.nys.dk



Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Danske resurser til automatisk opmærkning

DORTE HALTRUP

INDLEDNING

At opbygge sproglige resurser (såsom korpora, leksika og sprogteknologiske værktøjer) er meget tidskrævende og derfor meget bekosteligt. Når man opbygger resurser, er det derfor vigtigt at de kan bruges af mange og genbruges til forskellige formål, ikke mindst for at få dem finansieret.

Genbrug af sproglige resurser er særlig aktuelt i korpuslingvistikken, hvor der ofte stilles store krav til tekstresursernes størrelse og grammatiske opmærkning. Denne artikel præsenterer to vigtige emner i den danske korpuslingvistik, nemlig PAROLE-korpuset og Eric Brills algoritme til maskinlæring, Transformation-based Learning (TBL). Tilsammen danner disse to resurser udgangspunkt for mange af de danske taggere (programmer til automatisk korpusopmærkning med grammatiske kategorier) der er i brug i dag.

PAROLE og TBL er omtalt i flere af de øvrige artikler i dette nummer af NyS.

1. PAROLEKORPUSSET OG DETS TAGSET

Med udgangspunkt i genbrugstanken nedsatte EU i 1993 et udvalg kaldet EAGLES[1] (the Expert Advisory Group for Language Engineering Standards) der har til formål at definere standarder for sproglige resurser. Gruppen har bl.a. udarbejdet en anbefaling om morfosyntaktisk opmærkning af korpora der er udformet som et sprogneutralt tagset. Ideen har været at lave en ramme der er så finmasket at den kan rumme morfosyntaktiske værdier for alle de involverede europæiske sprog. Men der er langt fra teoretisk beskrivelse af et tagset til et konkret annoteret korpus.

Skridtet fra det teoretiske til det praktiske blev taget af det europæiske LE-PAROLE-projekt i årene 1996-1998. Projektet bestod af tre dele for hvert af de 14 involverede sprog: i) opbygning af et almensprogligt tekstkorpus på 20 mio. ord, ii) opbygning af et leksikon på 20.000 lemmaer samt iii) opbygning af et morfosyntaktisk annoteret korpus på 250.000 løbende ord.

I det følgende vil jeg skitsere hvad det danske annoterede PAROLE-korpus er, hvordan det er blevet skabt og derefter se lidt nærmere på dets tagset.

1.1 DET DANSKE MORFOSYNTAKTISK ANNOTEREDE PAROLE-KORPUS[2]

Det annoterede PAROLE-korpus består at 250.209 tekstord der er fordelt på 1553 tekstuddrag. Teksterne dækker 8 forskellige genrer, dog er ca. 70% avistekster (se Keson 99).

Det rå tekstkorpus er først blevet morfologisk analyseret med DAN-TWOL-algoritmen[3], som giver en eller flere alternative analyser til hvert ord. Derefter er den korrekte analyse manuelt markeret <correct!>. Fx:

```
"<*samtlige>"
        "samtlige" <*> A POS UK UT UB NOM <correct!>
"<partier>"
        "parti" N INT PL UBEST NOM <correct!>
"<i>"
        "i" U <adv>
        "i" U <prep> <correct!>
        "i" U <adv>
        "i" U <prep>
        "i" NUM <roman>
"<*folketinget>"
        "folke#ting" <*> N INT SG BEST NOM <correct!>
```

Den manuelle udvælgelse af de korrekte analyser er fortaget parallelt af to personer for at sikre at resultatet er så korrekt som muligt. De steder hvor annotørerne har vært uenige, er der gennem diskussion opnået et

fælles resultat; men til trods for denne omhyggelige fremgangsmåde forekommer der stadig et antal fejlanalyser i korpusset.

Analyserne der er markeret <correct!>, er trukket ud automatisk, hvorefter analyserne er konverteret til det fælles PAROLE-format. Tekststumpen der blev vist i DAN-TWOL-format ovenfor, ser i PAROLE-formatet ud på flg. måde:

```
<W lemma="samtlige" msd="ANP[CN][SP]U=[DI]U">Samtlige</W>
<W lemma="parti" msd="NCNPU==I">partier</W>
<W lemma="i" msd="SP">i</W>
<W lemma="folketing" msd="NCNSU==D">Folketinget</W>
```

– hvor ordet efter lemma er tekstordets lemma, bogstaver og tegn efter msd er tekstordets analyse, og tekstordet står sidst før </W>.

1.2 DET DANSKE PAROLE-TAGSET

Af eksemplet ovenfor fremgår det vist tydeligt at analyser i PAROLE-formatet ikke umiddelbart er let læselige. I dette afsnit vil jeg beskrive hvordan analyserne (taggene) er bygget op ud fra den fælles ramme for alle PAROLE-tagset.

Den generelle ramme er et ordnet sæt af træk hvilket vil sige at hvert træk (eller rettere dets værdi) har sin bestemte plads. Fx står ordklassen (Kat) altid på første plads. Rammens indhold udgør således det maksimale tagset ud fra hvilket hvert sprog kan definere sit eget.

I rammen der er illustreret nedenfor, skal man dels bemærke at de hvide felter er træk der anvendes i det danske PAROLE-korpus; mens de grå felter er træk der *ikke* er anvendt i det danske tagsæt. Og dels skal man bemærke at felterne 1-7 er de træk der er fælles for alle PAROLE-sprogene; mens trækkene i felterne 8-11 er sprogspecifikke (de træk der er vist her, er for dansk).

Pladsnummer

1	2	3	4	5	6	7	8	9	10	11
Kat.	**Subkat**									
Noun		**Genus**	**Num.**	**Kasus**			**Best.**			
Verb		**Modus**	**Tempus**	Person	**Num.**	**Genus**	**Best**	**Tr-kat**	**Diatese**	**Kasus**
Adj		**Grad**	**Genus**	**Num.**	**Kasus**		**Best.**	**Tr-kat**		
Pron		**Person**	**Genus**	**Num.**	**Kasus**	**Poss.**	**Reflek.**	**Regist.**		
Det		Person	Genus	Num	Kasus	Poss				
Art		Genus	Num.	Kasus						
Adv		**Grad**	Funk.	Wh-ness						
Adpos		Format	Genus	Num.						
Conj		C-type	C-pos.							
Num		Genus	Num.	Kasus						
Interj										
Residual										
Unique										

For bedre at forstå hvad trækkene dækker over, kan man fx se på et ord som *partier* der i PAROLE-format ser ud som følgende:

<W lemma="parti" msd=" NCNPU==I ">partier</W>

hvor tagget er NCNPU==I.
Hvis man går ind i skemaet og ser på hvilke danske værdier de forskellige træk for et substantiv kan have, finder man følgende[4]:

Pladsnummer

1	2	3	4	5	6	7	8	9	10	11
Kat	Sub-kat									
Noun		**Genus**	**Num.**	**Kasus**			**Best.**			
N	*proprium= **P***	*fælleskøn= **C***	*singularis= **S***	*genitiv= **G***			*bestemt= **D***			
	*appellativ= **C***	*intetkøn= **N***	*pluralis= **P***	*umark. = **U***			*ubestemt= **I***			

Ud fra skemaet kan man altså se at ordet *partier* er:

N et *substantiv* (noun)
C af *formen appellativ*
N er *intetkøn*
P i *pluralis*
U er *umarkeret for kasus*
= (ikke eksisterende)
= (ikke eksisterende)
I i *ubestemt* form

I alt er der 151 forskellige kombinationer af værdier i det danske PAROLE-tagset, dvs. at der findes 151 forskellige tags.

Nu er det ikke sikkert at man til alle formål har brug for så detaljeret et tagset. Nedenfor kan man se forskellige muligheder for at reducere det. Den mest ekstreme reduktion der indebærer kun at have ordklasserne med, fører til et tagset på 10 forskellige tags. Med dette minimale tagset vil man kunne lave en grov, basal analyse af en tekst; men fx tal og bestemt/ubestemt artikel vil forsvinde i analysen fordi disse størrelser hhv. tilhører ordklassen adjektiv og pronomen. Udvider man tagsettet med ordklassernes undertyper, fås et tagset på 25, medtages fx modus for verberne, fås et tagset på 34 osv. Tagsettet på 38 tags er det Britt Kesons anbefaler[5], og som i resten af artiklen kaldes "det reducerede tagset".

	Pladsnummer									
1	2	3	4	5	6	7	8	9	10	11
Kategori	**Sub. kategori**									
Noun	proprium, appellativ									
Verb	alm., medial	indikativ, imperativ, infinitivform, gerundium, participium	præsens, præteritum							
Adj	alm, kardinal, ordinal.									
Pron	personligt, demonstrativt, ubestemt, interrog/rel, reciprokt, possessivt									
Adv	generel									
Adpos	præposition									
Conj	sideord. underord.									
Interj										
Residual	forkortelse, udenlandske ord, tegn, formler, symboler, andet									
Unique										
10 tags	25 tags	34 tags	38 tags	I alt 151 PAROLE-tags						

Det store spørgsmål er selvfølgelig om det er fordelagtigt at reducere tagsettet, og i givet fald hvad man skal reducere det til (denne diskussion tages op i artiklen af Juel Henrichsen i dette nummer af NyS).

2. BRILL-TAGGEREN

Automatisk tagging består generelt af tre faser:

i. Tekstordene slås op i en ordbog for at finde deres kategori.

ii. For ordformer der ikke findes i ordbogen, skal systemet gætte kategorien.
iii. For ordformer der kan have flere kategorier, skal systemet vælge hvilken der er den korrekte i konteksten.

Hvilke metoder man vælger til at løse disse tre typer problemer, afhænger af ens teoretiske udgangspunkt. I dette afsnit vil vi beskrive Eric Brills taggingmetode der hedder *transformation based learning* (Brill 1993).

Transformation based learning er en algoritme der tager udgangspunkt i at systemet (taggeren) skal lære regler om tekstords kategorier automatisk ved at blive trænet på et allerede tagget korpus. Under træningen arbejdes med to versioner af samme korpus: den oprindelige taggede version samt en version hvor alle taggene er fjernet. Først tildeles ordene i det nøgne korpus et tilfældigt tag. Derefter ændres taggene ved hjælp af transformationer på en måde så den transformationsregel der får det "nøgne" korpus til at nærme sig det oprindelige, får en højere vægtning, mens de regler der får korpus til at fjerne sig fra det oprindelige, bliver smidt væk. På den måde opbygges lister af ordnede regler: leksikalske regler og kontekstuelle regler. Træningen stopper når der ikke kan findes flere regler, eller hvis systemet når en prædefineret grænse.

I de leksikalske regler ses bl.a. på tekstordenes præ- og suffikser hvorved der opbygges information der kan bruges til at gætte ukendte ords kategori. En leksikalsk regel kan fx se således ud:

```
ede hassuf 3 V_PAST 316.266946778711
```

hvilket betyder:

"Hvis ordet har suffikset *–ede* skal tagget (hvad det end er) ændres til V_PAST"

Dvs. at hvis et ord ender på *–ede*, er det ifølge taggeren datidsformen af et verbum. Tallet efter reglen er en form for vægtning af reglen.

De kontekstuelle regler derimod ser på tekstordets omgivelser hvorved de kan bruges til at vælge mellem en række alternative kategorier, altså til at fjerne syntaktisk flertydighed. En kontekstuel regel kan se således ud:

V_PAST V_INF PREWD at

hvilket betyder:

"Ændr V_PAST til V_INF hvis det foregående ord var *at*"

Dvs. at der ifølge taggeren er tale om infinitivsformen af et verbum hvis det foregående tekstord var *at*.

Gennem træningen har taggeren altså opbygget en ordbog og de to sæt transformationsregler. Med disse er den nu i stand til at tagge ny og ukendt tekst. Har man fx følgende tekststump: "Samtlige partier i Folketinget står i dag sammen om at bevilge 50 millioner kroner om året til samfundets svageste", vil den efter tagging se således ud:

Samtlige/ADJ partier/N i/PRÆP Folketinget/N står/V_PRES i/PRÆP dag/N sammen/ADV om/PRÆP at/UNIK bevilge/V_INF 50/NUM millioner/N kroner/N om/PRÆP året/N til/PRÆP samfundets/N_GEN svageste/ADJ

Den taggede tekststump ovenfor er fejlfri. Generelt har Brills tagger og enhver anden tagger en fejlrate på mellem 1,5 og 10 %. Forsøg har vist at Brill taggeren trænet med det reducerede tagsæt har en fejlrate på ca. 4%, hvilket er ganske pænt. Man skal dog være forsigtig med at lægge alt for meget i det pæne resultat. For det første er resultatet opnået på samme teksttype som træningen er foretaget på. Man kunne forestille sig at tagging af en anden teksttype eller et andet domæne ville give et dårligere resultat. For det andet siger tallet kun at der forekommer en række fejl, ikke hvilken type de er, hvor eller hvorfor de er opstået.

Haltrup (2000) samt artiklerne af Hardt og Juel Henrichsen (dette nummer af NyS) giver eksempler på anvendelser af de resurser der er gennemgået her i teksten.

Dorte Haltrup Hansen
Center for Sprogteknologi
email: dorte@cst.dk

NOTER:

1. http.//www.ilc.pi.cnr.it/EAGLES96/home.html
2. Korpusset er udarbejdet på Dansk Sprog- og Litteraturselskab under ledelse af Britt Keson og kan downloades fra: http://korpus.dsl.dk/e-resurser/parole-korpus.html
3. Algoritmen er udviklet til dansk af Thomas Bilgram (jf. Keson 1999)
4. Se en detaljeret gennemgang af trækkenes mulige værdier i :"Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus", der kan downloades sammen med korpusset.
5. Britt Keson, 1999: "Morfosyntaktisk tagging af danske tekster" i 7. Møde om Udforskning af Dansk Sprog (MUDS), red. af Peter Widell og Mette Kunøe, Århus 1999.

LITTERATUR:

Brill, E. (1993): *A Corpus-Based Approach to Language Learning.* Ph.D. thesis, Dpt. of Computer and Information Science, Univ. of Pennsylvania; [hent computerprogrammet på http://www.cs.jhu.edu/~Brill]

Haltrup Hansen, D. (2000): *Evaluering af NP-genkendere.* M.Sc. thesis, Kbh. Universitet; (unpubl.)

Keson, B.-K. (1999): *Vejledning til det Danske Morfosyntaktisk Taggede PAROLE-korpus.* Det Danske Sprog- og Litteraturselskab.