Peter Widell og Mette Kunøe (udg.):
8. Møde om Udforskningen af Dansk Sprog
Århus 2001

MORFOSYNTAKTISK TAGGING AF DANSKE TEKSTER

Af Britt-Katrin Keson (Det Danske Sprog- og Litteraturselskab)

Indledning

Denne artikel tager udgangspunkt i det danske morfosyntaktisk annoterede (taggede) tekstkorpus, der blev opbygget i forbindelse med PAROLE-projektet. Dette annoterede korpus på ca. 250.000 løbende tekstord udgør en del af et større dansk PAROLE-korpus på 20 mio. løbende tekstord, som snart vil være til rådighed for forskning og industri. Artiklen indledes med en kort beskrivelse af opbygningen og indholdet af det danske annoterede korpus[1], og derefter præsenteres nogle foreløbige resultater af et forsøg med at benytte dette korpus som træningsmateriale til træning af en automatisk tagger.

Præsentation af PAROLE-korpusets opbygning og indhold

Det danske morfosyntaktisk annoterede PAROLE-korpus er et almentsprogligt korpus i elektroniske tekstfiler indeholdende tegn fra iso-8859-1 tegnsættet. Det indeholder i alt 250.209 løbende tekstord (ekskl. interpunktionstegn) fordelt over 16.062 sætninger og 1.553 individuelle tekstuddrag. Dette korpus blev udformet under det EU-støttede projekt PAROLE (Preparatory Action for linguistic Resources Organization for Language Engineering) i perioden april 1996 til april 1998. Baggrunden for PAROLE-projektet er et ønske fra EU-Kommissionen om, at der for alle EU-sprogene skal foreligge større elektroniske tekstkorpora, samt heraf afledte leksika, der alle har en fælles opbygning og opmarkering. Ifølge PAROLEs formålserklæring skal disse ressourcer desuden (i) så vidt muligt være baseret på genbrug af allerede eksisterende maskinlæsbart materiale, (ii) være harmoniseret ifølge en fælles PAROLE-standard, og (iii) være offentligt tilgængelige (evt. mod betaling), dels igennem ELRA (European Language Resources Association), dels igennem et europæisk netværk af PAROLE-partnere. De danske partnere i PAROLE-projektet er Det Danske Sprog- og Litteraturselskab (DSL) og Center for Sprogteknologi (CST).

[1] En mere uddybende beskrivelse af det danske annoterede korpus' opbygning og indhold vil blive gjort tilgængelig sammen med selve korpuset (Keson 1999).

For hvert af de 14 EU-sprog[2] der omfattes af PAROLE-projektet, er der i forløbet således opbygget et almentsprogligt tekstkorpus på ca. 20 mio. løbende tekstord, som vil være tilgængeligt (med diverse restriktioner) hos den respektive PAROLE-partner. Efter anbefaling fra NERC (Network of European Reference Corpora) følger alle 14 PAROLE-korpora (i) en fælles specifikation for korpusets sammensætning ifølge korpusteksternes medium (samt genre og emne) (jf. Norling-Christensen 1996), (ii) en fælles SGML-opmarkeringsstandard for korpora (kaldet Corpus Encoding Standard), (iii) et fælles format for den morfosyntatiske annotering, samt (iv) et fælles dokumentationsformat. Af hvert af disse 14 korpora skal et delkorpus på ca. 3 mio. løbende tekstord desuden være frit tilgængeligt (evt. på en cd-rom) igennem ELRA. Til sidst skal (mindst) 250.000 af disse 3 mio. løbende tekstord være morfosyntaktisk annoterede ifølge et fælles PAROLE-format og -tagsæt.

PAROLE-korpusets opmarkering

Hele det annoterede PAROLE-korpus - inkl. hvert eneste af de 1.553 tekstuddrag - er opmarkeret med SGML-koder ifølge reglerne i PAROLEs fælles Corpus Encoding Standard eller CES (jf. Ridings 1996). At et tekstkorpus er opmarkeret ifølge en CES betyder, at korpusets overordnede struktur er fastlagt på forhånd af reglerne i CES'en. Denne struktur udtrykkes eksplicit vha. SGML-koder, og CES'en bestemmer først og fremmest, hvilke SGML-koder der er obligatoriske og hvilke SGML-koder det anbefales at anvende under opmarkeringen af korpuset. For eksempel bestemmer PAROLE CES'en hvilke meta-oplysninger om deres korpustekster, de 14 PAROLE-korpora bør indeholde (som f.eks. teksttitel, forfatternavn, antal løbende tekstord, angivelser om tekstens medium, genre, emne, osv.) SGML-koderne består (næsten altid) af en startkode (**<kode>**) og en slutkode (**</kode>**), der står hhv. før og efter den del af teksten, de er fælles om at afgrænse og beskrive. CES'en fastlægger desuden detaljeringsgraden for de SGML-koder der optræder inde i de løbende korpustekster for at angive deres interne struktur. Således er det ifølge PAROLE CES'en obligatorisk at inddele korpusteksterne i afsnit (** ... **), mens det kun er anbefalet at inddele korpusteksterne i sætninger (**<s>...</s>**).

Som illustreret i Figur 1 omgives hele PAROLE-korpuset af SGML-startkoden **<teicorpus.2>** og slutkoden **</teicorpus.2>**. Korpusets overordnede struktur består af to dele: (i) en 'corpus header' (korpushoved), som er omgivet af SGML-koderne **<teiHeader type=corpus>** og **</teiHeader>** og indeholder meta-oplysninger om PAROLE-korpuset som helhed, samt (ii) en korpuskrop, som

[2] PAROLE-projektet omfatter følgende EU-sprog: dansk, engelsk, finsk, fransk (samt belgisk fransk), græsk, irsk, italiensk, katalansk, nederlandsk, portugisisk. spansk, svensk og tysk.

indeholder de individuelle korpustekster. Alle korpustekster er også opmarkeret på samme måde, dvs. de består alle af et teksthoved og en tekstkrop. Teksthovedet er omgivet af **<teiHeader type=text>** og **</teiHeader>** og indeholder meta-oplysninger om den individuelle korpustekst. Tekstkroppen er omgivet af **<text>** og **</text>** og indeholder selve tekstuddraget (mellem **<body>** og **</body>**). I det morfosyntaktisk annoterede PAROLE-korpus er hvert eneste løbende tekstord (samt interpunktionstegn) desuden omgivet af SGML-startkoden **<W>** og slutkoden **</W>**.

PAROLE-korpusets morfosyntaktiske annotation

Det danske PAROLE-korpus blev annoteret med morfosyntaktiske oplysninger igennem en halv-automatisk proces i samarbejde med stud. ph.d. Thomas Bilgram (Aarhus Universitet)[3]. Disse oplysninger blev derefter automatisk konverteret til PAROLEs eget format og 'tagsæt' (dvs. analyseinventar). Som vist i Figur 2 består PAROLE-tagsættet af et fast antal fælles morfosyntaktiske træk samt et antal sprog-specifikke træk for hvert PAROLE-sprog. De morfosyntaktiske træk, der er fælles for alle PAROLE-sprogene, findes på pladserne 1 til 7 i tabellen i Figur 2, mens de sprog-specifikke (dvs. danske) træk findes på plads 8 og opefter. De hvide felter i tabellen repræsenterer de træk der faktisk anvendes i det danske annoterede PAROLE-korpus. De grå felter repræsenterer de fælles PAROLE-træk, som ikke anvendes i det danske korpus.

Første plads i PAROLE-tagsættet i Figur 2 er reserveret til en angivelse af tekstordets ordklasse (kaldet *CatGram*), og anden plads er reserveret til en yderligere underinddeling af denne ordklasse (kaldet *SsCatGram*). Hver plads udfyldes normalt med et bogstav eller et tal, der repræsenterer en bestemt værdi for det givne træk. I de tre korpuseksempler i Figur 3 nedenfor er tekstordene *russiske*, *historikere* og *Andronik* omgivet af SGML-koderne **<W>** og **</W>**. Disse koder indeholder desuden to attributter, der angiver hhv. tekstordets lemma (*lemma=*) og morfosyntaktiske analyse (*msd=*). Af første plads i msd-analyserne nedenfor fremgår det, at *russiske* er et adjektiv (**A** = adjective), mens *historikere* og *Andronik* begge er substantiver (**N** = noun). På plads nr. to i msd-analysen af *historikere* og *Andronik* kan det desuden ses, at *historikere* er et appellativ (**C** = common), mens *Andronik* er et proprium (**P** = proper). Appellativet er også markeret for genus (**C** = common), numerus (**P** = plural), kasus (**U** = unmarked, dvs. ikke-genitiv) samt bestemthed (**I** = indefinite), mens propriet kun er markeret for kasus (**U** = unmarked, dvs. ikke-genitiv).

[3] Dette samarbejde beskrives mere udførligt i Bilgram & Keson (1998).

Figur 3: Tre eksempler på annoterede korpusord

```
<W lemma="russisk" msd="ANP[CN|PU=|DI|U">russiske</W>
<W lemma="historiker" msd="NCCPU==I">historikere</W>
<W lemma="Andronik" msd="NP--U==">Andronik</W>
```

Anvendelse af det danske PAROLE-korpus i automatisk korpustagging

Det annoterede danske korpus blev anvendt i et forsøg med at træne en automatisk korpustagger, dvs. et program, der automatisk kan tildele morfosyntaktiske analyser til løbende tekstord. Fordelen ved at træne en automatisk tagger er, at det derefter vil være muligt - uden større ressourceomkostning - at tagge en (i princippet) ubegrænset mængde af løbende tekstord. Formålet med dette forsøg var at sammenligne resultatet af at træne to forskellige korpustaggere på det samme annoterede korpus ved at afprøve de trænede taggeres færdigheder på den samme ikke-annoterede tekst[4]. Endvidere blev begge taggere trænet på to forskellige versioner af det danske annoterede PAROLE-korpus: (a) en version med de fulde analyser (dvs. inkl. træk fra alle hvide felter i tabellen i Figur 2), hvilket giver i alt 151 mulige analyser, og (b) en version med forkortede analyser (dvs. kun inkl. træk fra de to første pladser i tabellen i Figur 2[5]), hvilket giver i alt 38 mulige analyser. Umiddelbart forekommer det indlysende, at et mindre tagsæt vil give bedre resultater, men undersøgelser af forskellige tagsæt (Elworthy 1995) har vist, at den optimale størrelse for et tagsæt kan variere fra sprog til sprog samt fra tagger til tagger. Store tagsæt indeholdende mange morfosyntaktiske oplysninger kan for visse sprog (og taggere) føre til bedre resultater, da de ekstra morfosyntatiske oplysninger kan bidrage til at skabe større deltaljering i taggerens disambigueringsregler.

Korpustaggerne

Q-taggeren er en offentligt tilgængelig, sprog-uafhængig tagger udviklet af Oliver Mason (University of Birmingham)[6]. Den repræsenterer en ren statistisk tilgang til korpustagging og er baseret på Markov-modellering. På baggrund af det annoterede træningsmateriale bygger Q-taggeren en ordbog med alle forekommende tekstord og deres tildelte analyser. Denne ordbog suppleres desuden med statistiske oplysninger om fordelingen af disse analyser på tekstordene, samt en algoritme til at gætte

[4] I dette indledende forsøg blev taggernes relative hastigheder ikke sammenlignet, men alene antallet af korrekt tildelte analyser.
[5] Samt, for verberne, træk fra de første fire pladser i tabellen.
[6] Q-taggeren (med dokumentation) kan findes på flg. URL: **http://www-clg.bham.ac.uk/QTAG**

'ukendte' ord (dvs. ord, der ikke forekommer i træningsmaterialet). Q-taggerens statistik er baseret af en beregning af følgende tre sandsynligheder for hvert forekommende tekstord: (i) $\mathbf{P_w}$ = **P(analyse|tekstord$_3$)**, dvs. sandsynligheden for, at dette tekstord har fået en given analyse (uafhængig af konteksten), (ii) $\mathbf{P_c}$ = **P(analyse|tekstord$_1$,tekstord$_2$)**, dvs. sandsynligheden for, at denne analyse følger efter analyserne for **tekstord$_1$** og **tekstord$_2$** (de to foregående tekstord), samt (iii) $\mathbf{P_{w,c}} = \mathbf{P_w} * \mathbf{P_c}$, dvs. sandsynligheden for, at en given analyse er den korrekte analyse, baseret på dens kontekst-uafhængige sandsynlighed ($\mathbf{P_w}$) samt dens kontekstuelle sandsynlighed ($\mathbf{P_c}$). En lignende beregning for den samme kontekst udføres derefter for **tekstord$_1$** og **tekstord$_2$**. Q-taggeren har for nylig udvist lovende resultater i et forsøg med at tagge rumænske tekster (jf. Tufis og Mason 1998).

Brill-taggeren er ligeledes en offentlig tilgængelig, sprog-uafhængig tagger, udviklet af Eric Brill (Johns Hopkins University)[7], men den repræsenterer en regel-baseret tilgang til automatisk korpustagging. Denne tagger benytter sig af en indlæringsmetode, hvormed den 'underviser sig selv' i at generere (i) leksikalske regler til at analysere ukendte tekstord, samt (ii) kontekstuelle regler til syntaktisk disambiguering. Udgangspunktet for Brill-taggeren er to versioner af det samme korpus: (i) den oprindelige annoterede version, samt (ii) en 'nøgen' version, hvor alle de morfosyntaktiske analyser er fjernet. Taggeren tildeler i første omgang vilkårlige analyser til det ikke-annoterede korpus. Derefter begynder den at generere tusindvis af transformationsregler, som systematisk ændrer analyserne for de forskellige tekstord. Enhver regel, som får dette korpus til at nærme sig det oprindelige annoterede korpus, får en højere vægtning, mens enhver regel, der får dette korpus til at fjerne sig fra det oprindelige, bliver smidt væk. På denne måde bygger Brill-taggeren en ordnet liste af transformationsregler op. Fordelen ved Brill-taggerens regler er, at de er læselige for et menneske, og således er det muligt for brugeren at redigere i dem ved at tilføje/fjerne regler efter behov for at hjælpe taggerens læreproces.

Tabellen i Figur 4 nedenfor viser eksempler på de ordnede leksikalske og kontekstuelle transformationsregler, Brill-taggeren har genereret på baggrund af det danske annoterede korpus. Den første leksikalske regel til venstre i tabellen ændrer analysen af ethvert tekstord til et bestemt appellativ (singularis, fælleskøn), dvs. **NCCSU==D**, hvis ordet ender på de to bogstaver *en*. Næste leksikalske regel ændrer analysen for et tekstord fra at være et ubestemt appellativ (singularis, fælleskøn), dvs. **NCCSU==I**, til en analyse som ubestemt appellativ (pluralis, fælleskøn), dvs. **NCCPU==I**, hvis ordet ender på de to bogstaver *er*. Den sidste leksikalske regel til venstre i tabellen

[7] Brill-taggeren (og dokumentation) kan findes på følgende URL: **ftp://ftp.cs.jhu.edu/pub/brill/Programs/**

har en meget lav vægtning og er blot taget med som eksempel på hvor deltaljeret Brill-taggerens leksikalske regler kan blive (fortolkningen af denne meget 'danske' regel overlades til læseren!). Ligeledes er sidste kontekstuelle regel til højre i tabellen taget med som eksempel på en kontekstuel regel med meget lav vægtning (konteksten for denne regel er ordet *Korsør*), som man burde overveje at fjerne fra taggerens endelige regelsæt.

Figur 4: Eksempler på Brill-taggerens transformationsregler

Leksikalske regler (til ukendte ord)	Kontekstuelle regler (til syntaktisk disambiguering)
en hassuf 2 NCCSU==D 3003 NCCSU==I er fhassuf 2 NCCPU==I 1437 et hassuf 2 NCNSU==D 1240 ne hassuf 2 NCCPU==D 942 NCCSU==I at fgoodright VAF-----A- 697 et goodright NCNSU==I 473 NCCSU==I ede fhassuf 3 VADA-----A- 459 NCCSU==I 1 fchar AC---U--- 455 ens hassuf 3 NCCSG==D 397 r deletesuf 1 VADR----A- 343 ... NCCSU==D sen faddsuf 3 NP--U---- 5	U= CS PREV1OR2WD , PP3NSU-NU PD-NSU--U NEXT1OR2OR3TAG NCNSU==I ANP[CN]PU=[DI]U ANP[CN]SU=DU NEXT1OR2TAG NCCSU==I PD-[CN]PU--U PP3[CN]PN-NU PREV1OR2WD , SP RGU NEXTTAG SP ANP[CN]SU=DU ANP[CN]PU=[DI]U NEXT1OR2OR3TAG NCCPU==I ANP[CN]PU=[DI]U ANP[CN]SU=DU PREV1OR2WD det PP3NSU-NU PD-NSU--U NEXTTAG ANP[CN]SU=DU PD-CSU--U PP3CSU-NU NEXTTAG VADR----A- PP3[CN]PN-NU PD-[CN]PU--U NEXT1OR2OR3TAG NCCPU==I ... NPU--U--- NCCSU==I PREV1OR2WD Korsør

Foreløbige resultater

Resultaterne af behandlingen af 250 vilkårligt udvalgte danske sætninger fra det ikke-annoterede PAROLE-korpus med de to 'trænede' versioner af Brill-taggeren og Q-taggeren vises i Figur 5 nedenfor. Tallene mellem parenteser angiver procent korrekt tildelte analyser **efter** at de analyser, der er tildelt interpunktionstegn, er trukket fra, da interpunktionstegn altid analyseres entydigt og korrekt (ifølge dette tagsæt). Tallene mellem parenteserne giver derfor et måske mere nøjagtigt billede af taggernes færdigheder. Af resultaterne i Figur 5 nedenfor fremgår det klart, at (i) Brill-taggeren klarer sig signifikant bedre end Q-taggeren, og at (ii) resultaterne for de forkortede analyser generelt er bedre end for de fulde analyser for begge taggere. Disse resultater dækker både antallet af korrekt-analyserede ukendte ord samt korrekte syntaktiske disambigueringer. I forbindelse med det fulde tagsæt klarede begge taggere sig forholdsvis godt mht. syntaktisk disambiguering, men det endelige tal for korrekt tildelte analyser blev for begge taggere reduceret pga. deres forholdvis dårligere håndtering af ukendte tekstord iføge det fulde tagsæt.

Figur 5: Procent korrekt tildelte analyser

Brill-tagger + fulde analyser:	97,6 %	(97,2 %)
Brill-tagger + forkortede analyser:	98,3 %	(98,0 %)
Q-tagger + fulde analyser:	92,4 %	(91,3 %)
Q-tagger + forkortede analyser:	94,0 %	(93,1 %)

Eksempler på taggerenes resultater

Følgende fire eksempelsætninger er taget fra de ovennævnte 250 vilkårligt udvalgte sætninger, og de er vist i formatet: tekstord/analyse. Af pladshensyn vises hhv. Brill-taggeren og Q-taggerens analyser kun hver for sig (mellem skarpe parenteser og adskilt af kolon) i de tilfælde, hvor deres respektive analyser **ikke** falder sammen. Sætning (61a) nedenfor demonstrerer to typiske kontekstuelle fejl i Q-taggerens resultater: den har tildelt en analyse som possessiv pronomen (**PO**) til *Hans*, og en analyse som infinitiv-markør (**U=**) til *at*. Brill-taggeren har derimod entydiggjort begge disse tekstord korrekt på baggrund af sætningens kontekst.

(61a) Men/CC det/PP mener/VADR Hans/[NP:PO] Engell/NP ikke/RG ,/XP at/[CS:U=] man/PI skal/VADR lægge/VAF- så/RG meget/AN i/SP ./XP

Følgende to versioner (med hhv. fulde og forkortede analyser) af sætning (17) viser andre typiske fejlmønstre for de to taggere. I (17a) tager Q-taggeren fejl af de to ukendte ord *asylsøgere* og *lynbehandlet*, mens Brill-taggeren tildeler *asylsøgere* den korrekte analyse som appellativ (og *lynbehandlet* den forkerte). I (17b) tildeler begge taggere den korrekte analyse til *asylsøgere*, og Brill-taggeren tildeler også den korrekte analyse til *lynbehandlet*, mens Q-taggeren igen tager fejl af dette ord. Som det fremgår af disse to eksempler, har brugen af fulde analyser i træningsmaterialet ikke nødvendigvis hindret den korrekte tildeling af morfosyntaktiske analyser (tværtimod).

(17a) Asylsøgere/[NC:AN] fra/SP lande/NC uden/SP politisk/AN forfølgelse/NC skal/VADR have/VAF- deres/PO sag/NC lynbehandlet/NC ./XP

(17b) Asylsøgere/NCCPU==I fra/SP lande/NCNPU==I uden/SP politisk/ANP[CN]SU=IU forfølgelse/NCCSU==I skal/VADR=----A- have/VAF-=----A- deres/PO3[CN][SP]UPNU sag/NCCSU==I lynbehandlet/[VAPA=S[CN][ARU]-U:NCNSU==D] ./XP

Til sidst er sætning (33a) et eksempel på en kontekst, hvori begge taggere begår den samme fejl ved at tildele *spiller* en analyse som finit verbum. Forholdsvise komplekse sætninger som denne vil dog

altid kunne optræde i almentsproglige korpustekster og dermed medvirke til, at en automatisk tagger aldrig vil opnå 100% korrekte resultater.

(33a) Fra/SP en/PI stjernetilværelse/NC som/U= spiller/VADR i/SP bl.a./RG Borussia/NP Mönchengladbach/NP i/SP Vesttyskland/NP og/CC FC/NP Barcelona/NP står/VADR Allan/NP Simonsen/NP nu/RG udenfor/RG som/U= træner/NC for/SP sin/PO barndomsklub/NC ,/XP Vejle/NP Boldklub/NC ./XP

Konklusion

På baggrund af de opnåede resultater i forsøget med at benytte det danske morfosyntaktisk annoterede PAROLE-korpus til at træne to forskellige automatiske taggere kan det konkluderes, at (i) korpuset er velegnet som træningsmateriale til automatiske taggere, og (ii) især Brill-taggerens resultater virker lovende både med det fulde og det forkortede tagsæt.

Litteratur

Bilgram, Thomas & Britt Keson. 1998. The Construction of a Tagged Danish Corpus. *NODALIDA '98 Proceedings*, København:129-139.

Brill, Eric. 1992. A Simple Rule-based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL.

Brill, Eric. 1994. Some Advances in Transformation-Based Part of Speech Tagging. *Proceedings of the 12th National Conference on Artificial Intelligence*, Seattle:722-727.

Elworthy, David. 1995. Tagset Design and Inflected Languages. *Proceedings of the ACL SIGDAT Workshop - From Texts to Tags: Issues in Multilingual Language Analysis*, Dublin:1-9.

Keson, Britt. 1999. *Vejledning til PAROLEs morfosyntaktisk taggede korpus* (under udgivelse).

Norling-Christensen, Ole. 1996. *Design and Composition of Reusable Harmonized Written Language Reference Corpora for European Languages*. MLAP PAROLE 63-386 WP 4.1.1.

Ridings, Daniel. 1996. *Text Representation in PAROLE*. MLAP PAROLE 63-386 WP 4.1.3.

Tufis, Dan & Oliver Mason. 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada:589-596.