NyS

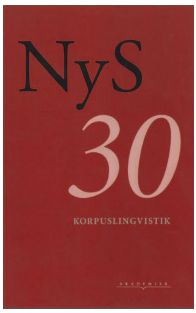
Titel: Fyrre kilometer kryds og bolle.
Metoder til grammatisk opmærkning i største skala

Forfatter: Peter Juel Henrichsen

Kilde: *NyS – Nydanske Sprogstudier* 30. *Korpuslingvistik*, 2002, s. 68-88

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.

Fyrre kilometer kryds og bolle

Metoder til grammatisk opmærkning i største skala

PETER JUEL HENRICHSEN

1. INDLEDNING

Opgaven for den deskriptive lingvistik er at beskrive sproget som det udtrykker sig når det ikke er under videnskabelig observation, det frit strømmende sprog. Som deskriptivist har man ikke meget udbytte af de omhyggeligt tilvirkede sprogprøver som grammatikbøgerne giver, men er henvist til at tappe sine primærdata ved kilden. *Kvantiteten* får dermed betydning i sig selv; når man giver afkald på at redigere sine data, må man til gengæld have rigelige mængder for at sikre forsyningen af eksempler. Den deskriptive sprogvidenskab har derfor fået et stærkt redskab i netværkscomputeren i direkte forbindelse med samfundets store, søgbare tekstbanker. Deskriptivisten er blevet *korpuslingvist*.

Almindelig søgeadgang til et stort tekstkorpus er dog ikke nok. Ud over avanceret fritekstsøgning har korpuslingvisten typisk brug for at søge ud fra grammatiske kriterier – altså ikke blot søge efter konkrete ord og tekstdele, men også efter mønstre og grupperinger beskrevet i den aktuelle undersøgelses egne termer.

Hvordan man målretter en søgning mod et grammatisk fænomen med måske titusinder af forekomster i et korpus der tæller millioner eller milliarder af ord, er emnet for denne artikel.

Vi demonstrerer en metode der bygger på *machine learning*. Teknikken består i at optræne computeren i grammatisk analyse og dernæst bruge den til at annotere korpusteksten – ord for ord – med information om ordklasse og bøjning, med det formål at forsyne tekstmassen med en grammatisk søgedimension.

Artiklen begynder med at præsentere et eksemplarisk korpus (Berling-

ske-99), og vi giver eksempler på spørgsmål som en korpusundersøgelse kunne besvare. Derefter præsenterer vi en træningsalgoritme og viser hvordan dens komponenter kan udvikles i praksis (Eric Brills algoritme *Transformation Based Learning* med PAROLEs annoterede korpus som reference). Vi anvender den trænede applikation på Berlingske-99, og der gives prøve på automatisk analyseret tekst. Det demonstrerede system er funktionsdygtigt – men, må det indrømmes, ikke særlig brugervenligt. Det bygger på en meget rig annotation med omkring 150 forskellige kategorier, og et så finmasket analytisk net er ikke altid hensigtsmæssigt. Et halvt hundrede kategorier eller færre, udvalgt med et specifikt formål, giver ikke blot en mere anskuelig analyse, men i tilgift en lavere fejlprocent (Kesons reducerede Parole-tagset gives som eksempel). Ideelt set burde hver korpuslingvist have mulighed for at definere sin personlige grammatiske taksonomi så den understøttede søgninger præcist rettet mod hans egne mål. Vi undersøger vilkårene for en sådan fleksibel korpussøgetjeneste, og til slut tegner vi omridset af en internetportal der tilbyder automatisk grammatisk opmærkning i største skala – med et personligt tilsnit.

1.1 ET EKSEMPLARISK KORPUS

Den daglige avis er et af samfundets mest levende tekstfora. I avisen mødes alle de tekstarter der spiller en rolle i samfundslivet, og enhver ordbog over det nutidige sprog må forholde sig til sproget som det bruges i avisen. Dagbladenes brødtekst opfylder altså korpuslingvistikkens krav om autencitet, relevans og kvantitet.

Dertil kommer kravet om søgbar repræsentation. Et af de danske bladhuse som tidligst erkendte nytten af aviser på computerlæsbar form, er det Berlingske Hus, som udgiver hver årgang af Berlingske Tidende Morgenavisen og Weekendavisen på cd-rom kort efter årets slutning. Berlingskes årlige opsamlings-cd'er har allerede fået status af standardressource i den danske korpuslingvistik, og derfor har vi som gennemgående eksempel i denne artikel valgt korpus *Berlingske-99*, bestående af al indekseret brødtekst i samtlige 1999-udgaver af Morgenavisen og Weekendavisen.

Berlingske-99 rummer godt 32 millioner løbende ord, fordelt på 52.000 avisartikler. Skrev man hele brødteksten ud som én lang spalte, ville den være på længde med et maraton.

1.2 ORDKLASSER SOM TEKSTINDGANGE

Hvilke spørgsmål kunne man have lyst at stille Berlingske-99? Det kommer naturligvis an på hvem man er.

Sociolingvisten kunne finde det relevant at undersøge om mandlige forfattere oftere end kvindelige anvender passive verbalkonstruktioner, og politiske kommentatorer oftere end sportsjournalister.

Leksikografen kunne have brug for at verificere at der blandt kombinationerne VERBUM-ADJEKTIV-SUBSTANTIV er stor overvægt af "*give* grønt lys" i forhold til "*give* rødt lys" og "*have* grønt lys".

Dansklæreren har måske brug for en række eksempler på en karakteristisk grammatisk form eller konstruktion – i flæng kan nævnes gerundium ("deres *klynken* hørtes tydeligt", "al *løben* og *legen* forbudt"), konjunktiv ("Fødselaren *leve*", "Herren *være* lovet") og objektsprædikat ("eksplosionen *gjorde omegnen ubeboelig*", "vi *valgte Ib til formand*").

Sprogingeniøren kunne have nytte af et materiale af hyppigt forekommende forvekslinger – for eksempel infinitiv for præsens ("han studere lingvistik") eller hyperkorrekt kommatering ("hun plejede, at læse Berlingeren"), med henblik på automatisk tekstkorrektur.

Det er let at fortsætte listen af mulige morfologiske, syntaktiske og leksikalske korpusundersøgelser der refererer til grammatiske kategorier. Her ses en annoteret tekstprøve fra Berlingske-99, læseren selv til inspiration:

EKSEMPEL 1. Tekstprøve fra Berlingske-99 annoteret med de "skolegrammatiske" kategorier.

Den	øgede	trafik	vil	især	kunne	mærkes	omkring	skisportsstederne	.
$PRON_{fkøn+sing.}$	$V_{perf.part.}$	$N_{fkøn+indef.+sing.}$	$V_{præs.aktiv}$	ADV	$V_{inf.}$	$V_{inf.passiv}$	PRÆP	$N_{ikøn+def.+plur.}$	TEGN

2. KOMPONENTERNE I ET AUTOMATISK SYSTEM

Søgning efter grammatiske kriterier kræver, som før nævnt, at hvert ord i objektkorpus er annoteret med ordklasseinformation. Da det naturligvis er praktisk umuligt at opmærke et kæmpekorpus med håndkraft, sætter man en computer i lingvistens sted.

Der er to principielt forskellige måder at opskole computeren. Enten må man formulere en række eksplicitte grammatikregler, et ekstremt res-

sourcekrævende projekt, som samtidig er næsten umuligt at føre til en ende – eller også må man følge den mere overkommelige strategi at lade computeren lære kunsten selv. Der findes i dag effektive metoder til at lade computere træne sig selv op uden supervision, til at opmærke ukendt tekst. Fælles for de metoder der bygger på optræning, er at de kræver adgang til et korpus af mindre størrelse, som til gengæld er omtrent perfekt annoteret. Et sådant *referencekorpus* er typisk stykket sammen af repræsentative tekstarter og omhyggeligt opmærket af lingvister. Træningen består nu i at lade computeren studere referencekorpusset og udvikle regler til udvælgelse af ordklassemærker (herefter kaldet *tags*, udt. på engelsk) i så nær overensstemmelse med referencen som muligt. Når computeren er færdigtrænet, er den i stand til at anvende den udviklede algoritme på ukendt tekst, og skridtet kan dermed tages fra det mellemstore referencekorpus til det vilkårligt store objektkorpus.

Fælles for de mest udbredte systemer til optræning i automatisk ordklasseanalyse er altså de tre grundlæggende komponenter:

- *TAGSET*: Et repertoire af tags der dækker de almindeligste analysekategorier (ordklasser, bøjningsformer etc.)
- *REFERENCEKORPUS*: Et mellemstort korpus (typisk 100.000-1.000.000 ord), manuelt opmærket med tags fra *TAGSET*
- *TRÆNINGSALGORITME*: En applikation der lader computeren studere *REFERENCEKORPUS* og derved træne sig op til at annotere ukendt tekst

Dagens træningsalgoritmer giver ikke computeren samme kompetence som en øvet lingvist (eksempler gives i det følgende), men man kan dog nå resultater som er gode nok til mange praktiske formål.

2.1 *TRÆNINGSALGORITME* TRANSFORMATION-BASED LEARNING

Algoritmen *Transformation-based learning* (TBL, Brill 1993) er en regelbaseret metode der lader computeren udvikle et slags 'kompendium' af grammatiske annotationsregler. Når træningen er gennemført og regelsamlingen er færdigudviklet, kan computen anvende reglerne på vilkårligt store tekstmængder som derved bliver annoteret med grammatiske tags, typisk med en nøjagtighed på 90-95%.

En TBL-træningssession tager udgangspunkt i et perfekt annoteret referencekorpus, *MASTER*. Som optakt til træningen dannes et parallelkorpus, *DUMMY*, bestående af præcis de samme ord som *MASTER*, men nu annoteret ud fra en primitiv initial regel – fx "alle ord er substantiver". Træningen består nu i gradvist at bringe annotationen i *DUMMY* i overensstemmelse med annotationen i *MASTER*. I hvert trin i træningsprocessen udvikles én ny regel som føjes til de øvrige, nemlig den regel der til enhver tid mindsker afstanden fra *DUMMY* til *MASTER* mest effektivt. Typiske TBL-regler kan være:

- Ord der ender på '-ede', er verber i præteritum
- Ord der ender på '-este', er adjektiver i superlativ når de forekommer netop før et substantiv i ubestemt form
- Ordet 'det' er et determinativ, når det forekommer som ord nummer 1 eller 2 før et adjektiv i bestemt form

Det er klart at TBL-regler ikke er ufejlbarlige, men blot fornuftige approksimationer. Typisk vil anvendelsen af en regel medføre fejl som ikke var der før – men hvis blot reglen retter flere fejl end den selv skaber, er skaden ikke stor, for så kan en senere regel korrigere de nye fejl. En TBL-tagger arbejder altså efter parolen to-skridt-frem-og-et-tilbage: en initial, meget grov annotation følges af en lang serie af korrektioner og korrektioner-til-korrektioner.

2.2 *TAGSET* OG *REFERENCEKORPUS:* DET DANSKE PAROLE

Som nævnt skal træningsalgoritmen forsynes med input i form at et grammatisk opmærket referencekorpus på helst 100.000 ord eller mere. I skrivende stund er der kun ét tilgængeligt dansk tekstkorpus som kan fungere som *REFERENCEKORPUS* i en TBL-træningssession som den beskrevne, nemlig det danske PAROLE-korpus. Kun dette korpus har på én gang den fornødne størrelse og kvalitet i opmærkningen.

Det opmærkede PAROLE-korpus består af ca. 290.000 grammatisk annoterede tokens, heraf ca. 250.000 egentlige leksemer (resten er interpunktion og andre ikke-alfabetiske tegn). Korpus består af blandede tekstgenrer, med avistekster som den største del. Opmærkningen er foretaget med halvautomatiske metoder og er efterfølgende verificeret af

lingvister – den skulle dermed være så tæt på perfekt som man i praksis kan komme.

Tagsettet som er benyttet i PAROLE-korpusset, er defineret så det kan gøre rede for (næsten) alle grammatiske dimensioner i den danske morfologi. Substantiver kan således markeres for:

- *subkategori* (proprium/appellativ),
- *genus* (fælleskøn/intetkøn),
- *numerus* (singularis/pluralis),
- *kasus* (neutral/genitiv) og
- *bestemthed* (definit/indefinit)

Tag som eksempel de to former "skisportsstedets" og "skisportsstedernes". PAROLE-tagsettet kan specificere de morfologiske forskelle og ligheder mellem disse to, idet de analyseres som hhv. NCNSG==D og NCNPG==D. Disse tagsymboler er systematiske:

- De første tre segmenter, NCN, bestemmer begge de to former som Noun-Common-Neuter (substantiv-appellativ-intetkøn).
- Det fjerde segment, S hhv. P, bestemmer de to former som Singularis hhv. Pluralis.
- Det femte segment, G, står for Genitiv.
- Det ottende segment, D, betyder Definit (bestemt form).

(Det sjette og syvende segment er ikke anvendt for substantiver, derfor er disse pladser blokeret med tegnene ==).

De andre ordklasser har tilsvarende systematiske tags. I alt omfatter det danske PAROLE-tagset 151 tags (se Dorte Haltrups introduktion til PAROLE-tagsettet og PAROLE-korpusset sidst i dette nummer af NyS).

2.3 DEN TRÆNEDE TAGGER

Et træningsforløb tager typisk to-fem døgn, hvis man anvender Eric Brills originale software og et referencekorpus i størrelsesordenen som det annoterede PAROLE-korpus. I vores konkrete forsøg udviklede taggeren 960 regler (heraf 447 leksikalske regler og 513 kontekstregler, jf. Haltrup 2002).

Den udviklede regelsamling sætter, som før nævnt, den automatiske tagger i stand til at annotere en *vilkårlig* tekst med PAROLE-tags. Vi anvender derfor taggeren på Berlingske-99 – og i løbet af nogle få timer har vi en fuldt opmærket version.

En tekstprøve fra det friskopmærkede Berlingske-99 ses herunder. Sætningen er den samme som i eksempel 1: "Den øgede trafik vil især kunne mærkes omkring skisportsstederne ".

EKSEMPEL 2. Tekstprøve fra Berlingske-99, automatisk annoteret (PAROLE-tags)

Token	PAROLE-tag	Grammatisk kategori
Den	PD-CSU–U	Pronomen (fkøn+sing.)
øgede	VAPA=S[CN]DA-U	Verbum (perf.part.)
trafik	NCCSU==I	Substantiv (fkøn+indef.+sing.)
vil	VADR=– – – –A–	Verbum (præsens aktiv)
især	RGU	Adverbium
kunne	VAF-=– – – –A–	Verbum (infinitiv aktiv)
mærkes	VADR=– – – –P–	Verbum (præsens passiv)
omkring	SP	Præposition (apposit.)
skisportsstederne	NCCPU==D	Substantiv (fkøn+def.+plur.)
	XP	Interpunktion

Som man ser, har taggeren taget fejl to steder. Verbet 'mærkes' er blevet analyseret som en *præsens* passiv frem for det korrekte *infinitiv* passiv. Substantivet 'skisportsstederne' er fejlagtigt rubriceret som *fælleskøn*. I begge tilfælde er fejlene dog moderate. Ordklasserne er korrekte, og den morfologiske placering er kun delvist forkert: bestemmelsen som 'passiv' hhv. 'def.+plur.' er således rigtig nok. Hovedparten af de fejl den automatiske tagger begår, er netop sådan at fejlanalyserede tokens trods alt placeres *i nærheden* af den korrekte kategori.

Som vi skal se herunder, bliver denne type fejl ofte usynlige hvis man skifter til et tagset med færre og større kategorier.

3. KESONS REDUCEREDE PAROLE-TAGSET

PAROLE-taggene er højt strukturerede og rige på morfologisk information. Dette er en stor søgeteknisk fordel. Man kan fx finde alle substanti-

ver ved at søge på tags indledt med 'N'. Søger man på 'NC', får man kun appellativerne (dvs. proprierne udelukkes). 'NCN' giver kun appellativer i intetkøn, mens 'NCNS' af disse kun lader singularisformerne komme igennem, osv. Hvert tilføjet tegn virker som et nyt filter.

I mange praktiske sammenhænge er PAROLE-taggene dog for besværlige at arbejde med. Som man ser i eksempel 2, er tagsymbolerne ikke lette at læse. Derfor kan det være praktisk at afbilde det fulde tagset på en mindre delmængde og på den måde skjule en del af den grammatiske information.

Britt Keson har foreslået en allround reduktion bestående af 38 tags, omtalt i det følgende som *Det Reducerede PAROLE-tagset*, eller blot *RedPAR* (Keson 1999). Keson reducerer alle ordklasser i cirka samme grad, og for substantivernes vedkommende betyder det en sammenlægning af de oprindelige 25 tags til bare fire: {EGEN, EGEN_GEN, N, N_GEN}. Samtidig erstatter Keson de strengt systematiske PAROLE-tags med mere læselige varianter. Symbolet EGEN_GEN kan fx let genkendes som "egennavn i genitiv".

Herunder ses til sammenligning "skisports"-sætningen fra eksempel 2, nu afbildet på *RedPAR*. Som det ses, er *RedPAR*'s tagnavne lette at gennemskue og behøver ikke nærmere beskrivelse.

EKSEMPEL 3. Tekst annoteret med Det Reducerede PAROLE-tagset (RedPAR).

Den	øgede	trafik	vil	især	kunne	mærkes	omkring	skisportsstederne	.
PRON_DEMO	V_PARTC_PAST	N	V_PRES	ADV	V_INF	V_PRES	PRÆP	N	TEGN

Bemærk at den ene annotationsfejl i eksempel 2 er blevet usynlig efter afbildningen på det reducerede tagset, nemlig "skisportsstederne" der nu kun er bestemt som *N*, dvs. uspecificeret appellativ. Den anden fejl ('mærkes' som præsensform) er stadig synlig.

Generelt bliver en del homografi usynlig når tagsettet reduceres. Tag som eksempel formen 'fornemme', der kan være både (i) adjektiv i definit singularis, (ii) adjektiv i pluralis og (iii) verbum i infinitiv. Det fulde PAROLE-tagset har tags for alle disse muligheder. *RedPAR* har kun to adjektiv-tags, {ADJ, ADJ_GEN}, og kan altså ikke beskrive forskellen på adjektiverne i "(den) fornemme (vin)v og "(mange) fornemme (vine)".

Eksemplerne herunder viser hvordan visse homografier stadig er synlige efter opmærkningen med *RedPAR*, mens andre typer reduceres eller forsvinder helt.

EKSEMPEL 4. Usynlige homografer efter opmærkning med RedPAR

Intakt homografi

'tier'	N	V_PRES
'Øst'	N	EGEN
'fortyndes'	V_INF	V_PRES

Reduceret homografi

'blandede'	($V\_PARTC\_PAST_{sing.+def.}$	$V\_PARTC\_PAST_{plur.}$)	V_PAST
'kort'	ADJ	($N_{sing.}$	$N_{plur.}$)
'fornemme'	($ADJ_{sing.+def.}$	$ADJ_{plur.}$)	V_INF

Kollapset homografi

'års'	($N\_GEN_{sing.}$	$N\_GEN_{plur.}$)	
'ægtepar'	($N_{sing.}$	$N_{plur.}$)	
'ting'	($N_{fkøn+sing.}$	$N_{ikøn+sing.}$	$N_{plur.}$)

(I eksempel 4 er de originale PAROLE-tags erstattet med læseligere symboler, og irrelevant information er udeladt; fx. er NCCSU==I erstattet med $N_{fkøn+sing.}$. Homografer der bliver uskelnelige i *RedPAR*-annotation, er sat i parentes).

Afbildningen af den rigere annotation på den fattigere med grovere inddelinger tilslører altså typisk en del taggingfejl. Derfor giver det ikke mening at spørge hvor stor en fejlprocent en given taggingalgoritme har *per se* – præcisionen er uløseligt forbundet til det anvendte tagsets størrelse og art.

4. DET PERSONLIGE *TAGSET*

Kesons reducerede PAROLE-tagset har vundet udbredelse som et pædagogisk udvalg der på anskuelig måde knytter forbindelsen mellem det komplette PAROLE-set og de alment kendte ordklasser – et nyttigt hjælpemiddel for den studerende. Som korpuslingvist får man dog snart brug

for at definere tagset efter sine egne kriterier, uddifferentiere visse kategorier og sammenlægge andre, som dikteret af den undersøgelse man er i gang med. Med andre ord, man har behov for et *personligt* tagset.

Det personlige tagsets mulighed kræver dog lidt reflektion. Går man frem som beskrevet i de foregående afsnit, virker vejen håbløst lang fra definitionen af tagsettet til den trænede applikation er klar til brug:

1. Det personlige tagset $TAGSET_{MY}$ defineres
2. *REFERENCEKORPUS* opmærkes med $TAGSET_{MY}$, hvorved opstår $REFERENCEKORPUS_{MY}$
3. *TRÆNINGSALGORITME* arbejder på $REFERENCEKORPUS_{MY}$ og udvikler taggeren $TAGGER_{MY}$
4. $TAGGER_{MY}$ anvendes på *OBJEKTKORPUS*

Stadium 1, udviklingen af det personlige tagset, kan gennemføres på nogle få timer eller minutter, og arbejdet føles fagligt tilfredsstillende fordi det er relateret direkte til den aktuelle undersøgelse. Stadium 2, opmærkningen af referencekorpusset, kræver derimod adskillige ugers rutinepræget arbejde uden særlig forbindelse til undersøgelsen, det vil sige: spildtid. Stadium 3, selve den ikke-superviserede optræning, tager ofte flere dage, og det samme gør stadium 4, opmærkningen af objektkorpusset (hvis det er i en størrelsesorden som Berlingske-99).

Heldigvis kan der snydes nogle hjørner. Hvis man udvælger sine personlige kategorier skønsomt og definerer $TAGSET_{MY}$ som en mange-til-en afbildning af PAROLE's tagset (eller et andet tilgængeligt *superset*), så kan man genbruge et allerede eksisterende referencekorpus. Tabel 1 giver tre eksempler på sådanne PAROLE-afbildninger med forskellig fokusering. De to første, $TAGSET_{passiv}$ og $TAGSET_{adv}$, er rettet mod henholdsvis undersøgelser af passive verbalkonstruktioner (fx 'blev kørt'/'kørtes') og adverbialer ('ofte'/'oftere'/'oftest'). Den tredje er Kesons *RedPAR* (jf. afsnit 3).

TABEL 1. Mange-til-en afbildninger af PAROLE-tagsettet (uddrag)

Grammatisk kategori	$TAGSET_{PAROLE}$	$TAGSET_{passiv}$	$TAGSET_{adv}$	RedPAR
Verbum (præsens aktiv)	VADR= - - - - A -	V_PRES_AKT	V_FIN	V_PRES
Verbum (præsens passiv)	VADR= - - - - P -	V_PRES_PASS		
Verbum (præteritum aktiv)	VADA= - - - - A -	V_PAST_AKT		V_PAST
Verbum (præteritum passiv)	VADA= - - - - P -	V_PAST_PASS		
Verbum (infinitiv aktiv)	VAF-= - - - - A -	V_INF_AKT	V_INF	V_INF
Verbum (infinitiv passiv)	VAF-= - - - - P -	V_INF_PASS		
Adverbium (posistiv)	RGP	ADV	ADV_POS	ADV
Adverbium (komparativ)	RGC		ADV_KOMP	
Adverbium (superlativ)	RGS		ADV_SUP	
Adverbium (ubøjelig)	RGU		ADV_U	

At der er tale om mange-til-en-afbildninger ses af at de reducerede tagset aldrig underdeler de basale PAROLE-kategorier.

Hvis $TAGSET_{MY}$ defineres som en mange-til-en afbildning af PAROLE-tagsettet, kommer man altså næsten gratis til ressourcen $REFERENCE\text{-}KORPUS_{MY}$, som nu kan afledes direkte af PAROLE-korpuset. Man er blot underlagt den begrænsning ikke at kunne uddifferentiere supersettets basale kategorier.[1] Dermed er den værste hurdle, stadium 2, undgået.

Tilbage er de mange timer eller dage som *computeren* anvender på sagen, til træning og efterfølgende opmærkning. Hvis også de kan undgås, er der ikke langt til at grammatisk fokuserede søgninger kan tilbydes online. Man kunne for eksempel tænke sig en internetbaseret tjeneste der tillod lingvisten at:

1. definere sit personlige $TAGSET_{MY}$ i et tekstområde i sin browsers vindue,
2. udpege et *OBJEKTKORPUS* blandt en række tilbudte muligheder (se fx Kirchmeier-Andersen 2002 i dette nummer af NyS),
3. definere en søgning der refererer til kategorierne i $TAGSET_{MY}$, og
4. trykke SØG,

hvorefter web-tjenesten efter få minutter afleverede et komplet søgeresultat.

Resten af denne artikel er viet overvejelser over en sådan eksprestjenestes mulighed – og muligheder.

5. PERSONLIG TAGGING SOM EKSPRESSERVICE

Skal man undgå de lange ventetider, må man skyde genvej uden om stadium 3 og 4. For det første må man undgå at udvikle personlige taggere og i stedet i alle tilfælde anvende en tagger trænet på et forud givet supertagset og referencekorpus, sådan at stadium 3 kun behøver passeres én gang. For det andet skal der oparbejdes en bank af objektkorpora annoteret i det rige supertagset ved hjælp af den trænede tagger, så at også tidsforbruget i stadium 4 bliver en engangsudgift. Når slutbrugeren har defineret et $TAGSET_{MY}$, udpeget et *OBJEKTKORPUS*, samt beskrevet en søgning, gennemføres søgningen efter en simpel afbildning af *OBJEKTKORPUS*' annotation, som beskrevet i sidste afsnit. Nu kan søgeprocessen gennemløbes på ganske få minutter.[2]

Før man iværksætter Projekt Eksprestagger efter disse retningslinier, er der dog et spørgsmål som må besvares: Sætter man noget til på taggingens *præcision* ved at overgå fra den tidskrævende model Reduce-then-Train-then-Tag (RTT) til genbrugsmodellen Train-then-Tag-then-Reduce (TTR)?

Vi undersøger kvalitetsforholdet mellem de to metoder ved at gennemføre en lille forsøgsrække.

1. Berlingske-99 annoteres af både RTT-taggeren og TTR-taggeren[3]. Hvor stor er afvigelsen i absolutte tal?
2. 1000 afvigende domme vurderes manuelt. Hvilken tagger har oftest ret?
3. Taggingfejl er typisk koncentreret om *homograferne*. Der udskilles en mængde af særligt kritiske 'testhomografer'. Hvor meget afviger taggingen i denne *worst case* gruppe?

Konklusionerne samles op i et afsluttende afsnit.

5.1 RTT VERSUS TTR

Undersøgelsen begynder med at korpus Berlingske-99 tagges to gange, af henholdsvis RTT-taggeren (den dyre) og TTR-taggeren (den billige). I hvert tilfælde er resultatet en individuel version af Berlingske-99 opmærket med *RedPAR*.

Ved at sammenligne de to korpusversioner token-for-token finder man at ca. 94,4% af alle tokens tagges *ens* af de to taggere. I vores sammenhæng samler interessen sig naturligvis om de sidste 5,6%, for det er her man skal søge svaret på hvilken tagger der er den bedste.

I tabellerne herunder ses en opgørelse over taggingens resultater. Resultatet er opgjort for de to delkorpora Morgenavisen og Weekendavisen, og desuden er en enkelt udgave (Morgenavisen 22-02-99) udtaget til nærmere analyse (se næste afsnit).

TABEL 2. Enstaggede tokens.

	Udgaver (=filer)	Artikler	Tokens	Enstaggede	Enighed
Morgenavisen	358	48.427	27.669.109	26.134.305	94.45%
Weekendavisen[4]	57	3.756	4.443.356	4.191.570	94.33%
Mor. 22-02-99	1	96	50.465	47.605	94.33%

TABEL 3. Distribution over RedPAR.

Tag	TTR	RTT
ADJ	2.317.486	2.237.690
ADJ_GEN	1.471	4.298
ADV	1.599.026	1.607.080
EGEN	1.586.210	1.868.969
EGEN_GEN	111.209	93.572
FORK	17.400	33.969
FORM	2.107	29
INTERJ	8.098	9.874
N	5.501.599	5.454.119
NUM	443.813	392.381
NUM_GEN	351	351

NUM_ORD	41.266	35.757
NUM_ORD_GEN	4	4
N_GEN	216.220	218.032
PRON_DEMO	695.491	689.186
PRON_DEMO_GEN	371	371
PRON_INTER_REL	42.659	42.229
PRON_INTER_REL_GEN	1.639	0
PRON_PERS	1.017.220	1.021.256
PRON_POSS	193.315	193.245
PRON_REC	5.351	5.351
PRON_REC_GEN	371	371
PRON_UBST	892.189	876.881
PRON_UBST_GEN	2.234	875
PRÆP	3.233.566	3.227.177
SKONJ	918.925	923.820
SYMBOL	562	7.396
TEGN	3.243.156	3.250.660
UKONJ	526.056	526.737
UL	18.665	12.777
UNIK	917.332	917.698
V_GERUND	1.583	16.112
V_IMP	18.199	46.637
V_INF	789.170	796.676
V_MED_INF	6.082	4.568
V_MED_PARTC_PAST	1.058	905
V_MED_PAST	5.486	4.819
V_MED_PRES	17.134	18.216
V_PARTC_PAST	596.987	576.869
V_PARTC_PRES	72.826	57.326
V_PAST	676.808	637.079
V_PRES	1.897.418	1.831.287
XX	30.996	26.460

De største relative forskelle samler sig (heldigvis) om de sjældnest benyttede tags: ADJ_GEN (adjektiv i genitiv), FORK (forkortelse), FORM (formel), osv (konkrete eksempler på anvendelsen af FORK ses i tabel 5 herunder).

At de store afvigelser ses på de små forekomster er ikke overraskende: Her har det ret lille referencekorpus ikke kunnet levere tilstrækkeligt med eksempler til en egentlig regeldannelse, og de grammatiske domme er følgelig ret tilfældige. Da de bemeldte tags kun anvendes nogle få tusind gange, påvirker de umotiverede domme ikke den overordnede fejlprocent ret meget.

5.2 1000 STIKPRØVER

De 1000 første uens-taggede tokens i en tilfældigt valgt udgave af Berlingske Morgen (22-02-99) er udtaget til manuel kontrol. Det dækkede korpusområde består af 18.599 tokens, svarende til 37% af udgavens samlede brødtekst (0.06% af hele korpus).

Hvert af de 1000 tokens blev afbildet i en 2+2 kontekst og annoteret med sine to afvigende tags (hhv. TTR og RTT). I langt de fleste tilfælde var et sådant 5-ords tekstvindue tilstrækkeligt til en sikker afgørelse; i tvivlstilfælde blev de taggede filer konsulteret.
I hvert tilfælde blev vurderingen af de to uens tags udtrykt med en kode:

0 = begge forkerte
1 = TTR-tagget korrekt, RTT-tagget forkert
2 = RTT-tagget korrekt, TTR-tagget forkert
3 = begge korrekte
12 = TTR-tagget korrekt, RTT-tagget mildt forkert
21 = RTT-tagget korrekt, TTR-tagget mildt forkert
? = vurdering usikker/umulig

I tabel 4 herunder ses de første 10 uens-taggede tokens (vist i **fed** font).

TABEL 4. Uens tags: De første 10 stikprøver.

Nr.	Tekstvindue	TTR-tag	RTT-tag	Vurdering
1.	at det **anerkendte** museum opfører	V_PAST	ADJ	2
2.	, stærkt **omdiskuteret** tilbygning tegnet	V_PARTC_PAST	ADJ	2
3.	kroner skal **Humlebæk-borgerne** op med	N	V_INF	1
4.	de vil **matche** den pris	N	V_INF	2
5.	den nye » Støtteforening for	ADJ	N	0
6.	nye » **Støtteforening** for Gammel	N	EGEN	21
7.	Støtteforening for **Gammel** Humlebæk Havn	EGEN	ADJ	12
8.	. » **Skødet** er jo	N	EGEN	1
9.	er jo **betinget** af ,	ADJ	V_PARTC_PAST	21
10.	Og det **tvivler** jeg stærkt	V_PRES	N	1

Bemærk at token nr. 2 er vurderet som '2' (kun RTT-tagget er korrekt, dvs. ADJ), idet 'omdiskuteret' næppe kan anses for en form af et verbum 'at omdiskutere'. Tokens nr. 6 og 7 forekommer begge som egennavne i den aktuelle kontekst, men da 'Støtteforening' og 'Gammel', *leksikalsk* betragtet er hhv. substantiv og adjektiv, er de alternative tags N og ADJ bedømt som blot 'mildt forkerte' (hhv. kode '21' og '12').

Til sammenligning ses i tabel 5 nogle typiske *kode-3* vurderinger, altså alternative taggings der begge anses for korrekte:

TABEL 5. Uens tags: Godkendte alternativer.

Tekstvindue	TTR-tag	RTT-tag	Vurdering
med sig hjem . (end-of-line)	N	ADV	3
kan præsentere Beograd-styret for et	N	EGEN	3
, professor dr. jur. (end-of-line)	FORK	N	3
følge af Tvind-loven efterbetalt .	N	EGEN	3
med 30-40 tidligere gadebørn ,	ADJ	ADV	3
kendskabet til Stairway-pædagogikken i Danmark	N	EGEN	3
er det kendetegnende , at	ADJ	V_PARTC_PRES	3
fandt politiet afrevne ærmer med	ADJ	V_PARTC_PAST	3
og 40 mm luftværnskanoner .	N	FORK	3

Som det fremgår, er kode-3 ofte brugt hvor to alternative tags kan motiveres ud fra hhv. indholdsmæssige og formmæssige kriterier (fx adjektiv vs. participium, substantiv vs. forkortelse, proprium vs. appellativ).

Kode-0 (begge taggings forkerte) optræder typisk ved stærkt homografe ord ("31 års **tro** tjeneste", "Den **sky** amerikaner"), ord med arkaisk bøjning ("dolket til **døde**", "i går **aftes**", "skulle ske **fyldest**"), ikke-danske former ("et par **XXL-jeans**", "avisen Welt **am** Sonntag") og alle slags ortografiske anomalier ("Man ved **selvfølgeig** aldrig", "**www.dr.dk/p3/singletons**").

Optællingen af de 1000 vurderinger viser at de to taggere deler fejlene næsten ligeligt mellem sig. Blandt de 1000 udvalgte tokens har TTR tagget 473 korrekt, mens RTT har tagget 464 korrekt. Detaljerne ses herunder.

Vurdering	Antal
'?'	2
'0'	138
'1'	377
'12'	19
'2'	346
'21'	41
'3'	77
ialt	1000

I dette forsøg viser RTT-taggeren og TTR-taggeren sig altså praktisk talt jævnbyrdige.

5.3 TESTHOMOGRAFER

Vi udpeger nu en kontrolleret delmængde af de tokens som forekommer i Berlingske-99, nemlig de homografer som giver størst mulighed for afvigende tagging. Man kan forvente af taggingen i dette vanskeligste fragment af korpus afviger mere end de gennemsnitlige 5,6%. Spørgsmålet er hvor meget afvigelsen øges eller, med andre ord, hvor robuste taggerne er over for homografi.

Vi definerer først kategorien *testhomografer*. En ordform *W* udgør en testhomograf hvis de to applikationer RTT og TTR

1. har (mindst) to leksikalske indgange hver for *W*,
2. har identiske indgange for *W*,
3. opmærker *W* forskelligt på (mindst) to forekomster i *OBJEKTKORPUS*.

Leksemet 'fører' er et eksempel på en testhomograf: For det første har RTT- og TTR-taggeren hver to leksikalske indgange for dette leksem:

TTR:

indgang	**tag**
fører	V_PRES
fører	N

RTT:

indgang	**tag**
fører	V_PRES
fører	N

For det andet er indgangene parvist identiske – det vil sige at TTR aldrig kan tagge en forekomst af 'fører' på en måde som er uden for RTTs rækkevidde, og vice versa. For det tredje *anvendes* begge tags i praksis af begge applikationer. Med andre ord: Der findes i *OBJEKTKORPUS* (mindst) to forekomster af ordet 'fører' som TTR-applikationen tagger forskelligt, og tilsvarende med RTT.

Herunder ses en analyse af de to største testhomografer, nemlig 'det' og 'den'. De er begge leksikaliseret af TTR- og RTT-applikationen som såvel personligt pronomen som demonstrativt pronomen. Som det fremgår af tabellen er de to applikationer langt oftest enige i valget af tag, nemlig i hhv. 91% og 96% af tilfældene.

TABEL 6. De to største testhomografer

ordform	RTT-tag	TTR-tag	antal	ens?
det	PRON_PERS	PRON_PERS	248.181	+
	PRON_DEMO	PRON_PERS	19.753	–
	PRON_PERS	PRON_DEMO	16.289	–
	PRON_DEMO	PRON_DEMO	119.291	+
	forekomster=403.514	**enstaggede=367.472**	**enighed=91,06%**	
den	PRON_PERS	PRON_PERS	40.182	+
	PRON_DEMO	PRON_PERS	5.269	–
	PRON_PERS	PRON_DEMO	7.247	–
	PRON_DEMO	PRON_DEMO	257.835	+
	forekomster=310.533	**enstaggede=298.017**	**enighed=95,96%**	

Enighed i denne størrelsesorden viser sig at være normen. Blandt de ti største testhomografer ligger overensstemmelsen for de nis vedkommende i området 91-99% (jf. tabel 7), og kun den femtestørste testhomograf, 'indtil', tagges forskelligt i 19 ud af 100 tilfælde.

TABEL 7. De 10 største testhomografer

rang	ordform	enighed
#1	det	91,06%
#2	den	95,96%
#3	Det	94,03%
#4	Den	97,52%
#5	indtil	81,42%
#6	fører	92,22%
#7	Så	90,52%
#8	dét	93,46%
#9	Med	98,93%
#10	Da	99,05%

De 10 mest frekvente testhomografer dækker ca. 98% af testhomografmassen, og i dette korpusfragment er den samlede enighed ialt på

92,99%. Da testhomograferne er særligt udvalgt til at udstille de to applikationers svagheder, har vi hermed et indtryk af såvel *gennemsnitlig* enighed (94,4%) som *worst case* enighed (93,0%). Den beskedne forskel mellem disse tal viser at taggernes indbyrdes afvigelse kan forventes at være nogenlunde konstant i alle dele af objektkorpus. Dertil kommer at taggingens præcision er uafhængig af den valgte metode (med forbehold for usikkerheden i den lille 1000-ordstest).

Konklusion på undersøgelsen er altså at opmærkningskvaliteten er den samme med de to metoder. Eksprestagging (TTR) synes dermed inden for rækkevidde.

6. AFSLUTTENDE BEMÆRKNINGER

Vi vil anbefale udviklingen af et (internetbaseret) opmærkningsværktøj til annotation af meget store korpora med personlige tagset.

En ekspresservice som den skitserede vil for alvor udløse det potentiale der er i de moderne maskinlærte metoder til grammatisk annotation. På få øjeblikke kan en vilkårlig tekst forsynes med en 'lingvistisk undertekst'. Annotationen kan ske så hurtigt at den vil kunne tilbydes som en standardfeature i en offentlig søgetjeneste.

Fra et teoretisk-lingvistisk synspunkt ville en sådan eksprestjeneste åbne en endnu ukendt analytisk dimension, idet det vil blive praktisk muligt at opfatte *definitionen af tagsettet*, altså selve batteriet af grundlæggende analysekategorier, som en eksperimentel parameter.

Sidst men ikke mindst vil man hermed omgå et problem som i dag står i vejen for den automatiske ordklasseanalyses udbredelse, nemlig den tekniske barriere. I dag kræver det fx indsigt i programmeringssprogene c++ og Perl at udnytte Eric Brills algoritme i personlige træningssessioner. Med udbydelsen af en internetbaseret ekspresservice kunne man holde alle lingvistisk uvedkommende teknikaliteter skjult bag ved en web-side.

Og dermed vinde nye sjæle for korpuslingvistikken.

Peter Juel Henrichsen
Institut for Datalingvistik,
Handelshøjskolen i København
email: pjuel@id.cbs.dk

NOTER

1. PAROLEs tagset blev netop udviklet som et tilbud om et supertagset der er tilstrækkeligt rigt til at den lingvistiske slutbruger altid har de mindstekategorier til rådighed han har brug for; desuden omfattede PAROLE-projektet, som før nævnt, også den manuelle opmærkning af et (mellemstort) korpus.
2. Nok så vigtigt: søgning kan ske i bedre-end-lineær tid i forhold til størrelsen af objektkorpus (forudsat at søgeværktøjet kan håndtere indekseret tekst). Dette sidste er et stort plus, fordi det fremtidssikrer metoden.
3. I de to forsøgsrækker anvendes Berlingske-99 som objektkorpus, mens rollerne som TTR- og RTT-applikation spilles af hhv. $TAGGER_{PAROLE}$ og $TAGGER_{RedPAR}$)
4. Korpus omfatter nogle ekstrasektioner til Weekendavisen, derfor er antal udgaver >52.

LITTERATUR

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

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

Haltrup Hansen, D. (2002): To ressourcer. *NyS* 30.

Hardt, D. (2001): Dansk grammatikkontrol med Transformation-Based Learning. *NyS* 30.

Henrichsen, P.J. (2001): Sidste Års Aviser - grammatisk opmærkning af et stort dansk aviskorpus. *Lambda* 27. Handelshøjskolen i Kbh:

Henrichsen, P.J. (2001): Transformation-Based Learning of Danish Stress Assignment. *EuroSpeech-01.*

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

Kirchmeier-Andersen, S. (2002): Dansk korpusbaseret forskning - hvordan kommer vi videre? *NyS* 30.