Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus

af Britt Keson
Det Danske Sprog- og Litteraturselskab (DSL)

FIGURLISTE:

1. Indledning

Dette er brugervejledningen til det danske morfosyntaktisk "taggede" (dvs. annoterede) tekstkorpus, der er udviklet under det europæiske LE(Language Engineering)-PAROLE-projekt[1]. Vejledningen indeholder en udførlig beskrivelse af det danske korpus' opbygning og indhold. Formålet med denne vejledning er både (*i*) at beskrive den formelle opbygning af korpusset (dvs. strukturen i SGML-opmarkeringen[2] samt i de morfosyntaktiske oplysninger), og (*ii*) at give brugeren et indblik i selve indholdet af korpusset. Vejledningen indeholder således talrige korpuseksempler samt en redegørelse for de lingvistiske beslutninger, der ligger til grund for den morfosyntaktiske tagging.

Det danske morfosyntaktisk taggede PAROLE-korpus er (som de andre europæiske PAROLE-korpora) 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. Korpusset er resultatet af et samarbejde mellem cand. mag. Britt Keson (samt stud. mag. Dorte Haltrup Hansen) fra Det Danske Sprog- og Litteraturselskab (DSL) og stud. Ph.D. Thomas Bilgram fra Institut for Lingvistik, Finsk og Ungarsk, Aarhus Universitet.

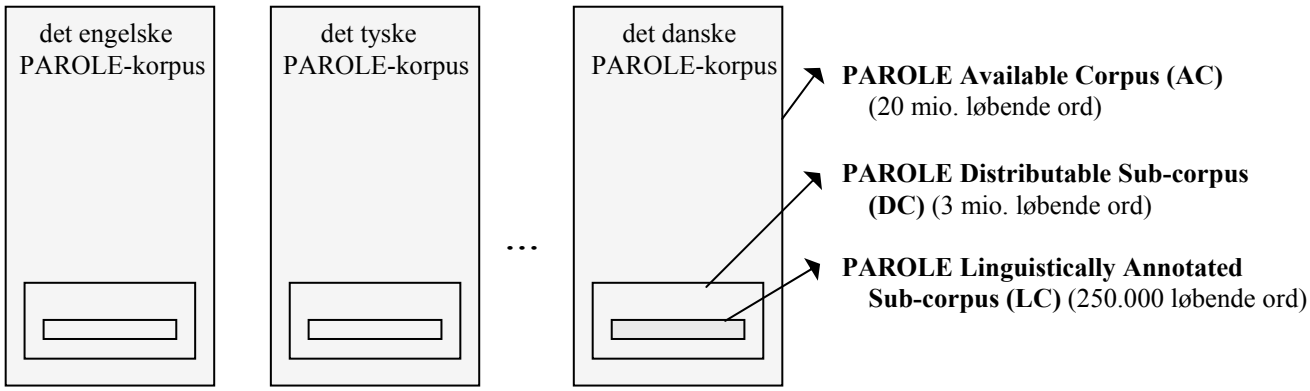
Hvad er PAROLE?
Det danske taggede korpus blev udformet under det EU-støttede projekt, **PAROLE** (**P**reparatory **A**ction for linguistic **R**esources **O**rganization for **L**anguage **E**ngineering), i perioden april 1996 til april 1998. Baggrunden for projektet er EU-Kommissionens ønske om, at der for alle EU-sprogene foreligger skriftsproglige elektroniske tekstsamlinger (korpora) og heraf afledte morfologiske og syntaktiske orddatabaser (leksika), som skal være til rådighed for national og international sprogteknologisk forskning og industri. 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).

Hvad er et PAROLE-korpus?
For hvert af de EU-sprog[3], der omfattes af PAROLE-projektet, er der opbygget et almentsprogligt tekstkorpus på ca. 20 mio. løbende tekstord (kaldet 'Available Corpus'), som er tilgængeligt (med diverse restriktioner) hos den respektive PAROLE-partner. Efter anbefaling fra NERC (Network of European Reference Corpora) følger alle PAROLE-korpora (*i*) en fælles specifikation for korpussets sammensætning ifølge korpusteksternes medium (samt genre og emne)[4], (*ii*) en fælles SGML-opmarkeringsstandard for korpora (kaldet 'Corpus Encoding Standard'), (*iii*) et fælles format til den morfosyntatiske annotering, samt (*iv*) et fælles dokumentationsformat. Af hvert af disse korpora skal et delkorpus på ca. 3 mio. løbende tekstord (kaldet 'Distributable Sub-corpus') desuden være frit tilgængeligt (evt. på en cd-rom)

[1] Jf. mere udførlige beskrivelser af LE-PAROLE-projektets formål og opbygning i LE-PAROLE, 199? samt LE-PAROLE Project Summary, 199?.
[2] 'Standard Generalized Markup Language' (En god indføring i SGML findes i Sperberg-McQueen & Burnard, 1994, kapitel 2: 'A Gentle Introduction to SGML'.)
[3] PAROLE-projektet omfatter følgende EU-sprog: dansk, engelsk, finsk, fransk (samt belgisk fransk), græsk, irsk, italiensk, katalansk, nederlandsk, portugisisk, spansk, svensk og tysk.
[4] Jf. 'Design and Composition of Reusable Harmonized Written Language Reference Corpora for European Languages' (Norling-Christensen, 1996).

igennem ELRA. Til sidst skal (mindst) 250.000 af disse 3 mio. løbende tekstord være morfosyntaktisk taggede ifølge et fælles PAROLE-format og -tagsæt (kaldet 'Linguistically Annotated Sub-corpus'). Denne vejledning omhandler alene det danske morfosyntaktisk taggede delkorpus på ca. 250.000 løbende tekstord, og fremover vil alle henvisninger til "PAROLE-korpusset" i vejledningen være hertil.

Figur 1: Opbygning af de forskellige PAROLE-korpora



Hvad er et morfosyntaktisk tagget korpus?
Hermed forstås et tekstkorpus, hvori de løbende tekstord systematisk er blevet forsynet med en række morfologiske og syntaktiske oplysninger, som f.eks. deres ordklasse og forskellige morfologiske bøjningsoplysninger. Selvom disse morfosyntaktiske oplysninger udtrykkes vha. SGML-koder i de omtalte PAROLE-korpora, skelnes der her i vejledningen mellem (*i*) 'korpustagging', dvs. tildeling af morfosyntaktiske oplysninger til de løbende tekstord i korpusteksterne, og (*ii*) 'korpusopmarkering' eller 'tekstopmarkering', dvs. tilføjelsen af (andre) SGML-koder til korpusset som helhed samt til selve korpusteksterne for at angive deres interne struktur.

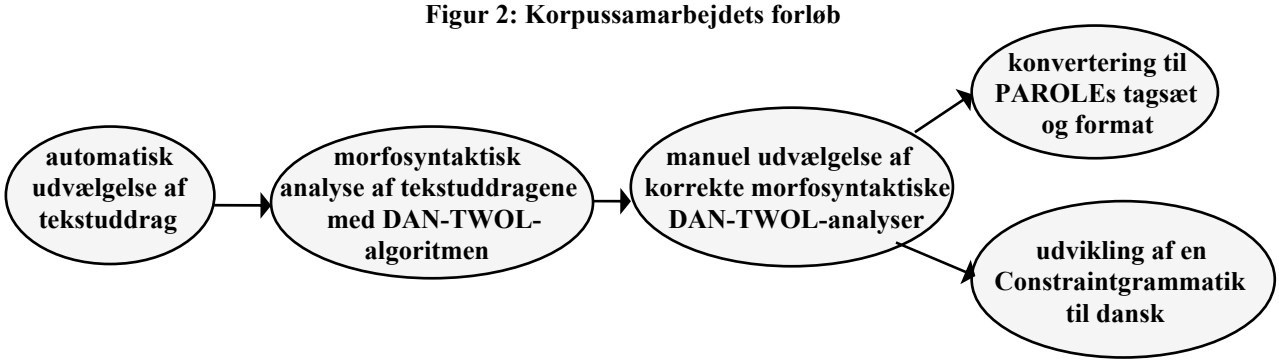
2. Morfosyntaktisk korpustagging

Dette afsnit omhandler forløbet af det samarbejde om korpustagging, der er resulteret i det danske morfosyntaktisk taggede PAROLE-korpus.

2.1 Korpustaggingens forløb

Samarbejdet mellem Britt Keson (samt Dorte Haltrup Hansen) fra Det Danske Sprog- og Litteraturselskab og Thomas Bilgram fra Aarhus Universitet har haft følgende forløb:

Figur 2: Korpussamarbejdets forløb



Automatisk udvælgelse af tekstuddrag
PAROLE-korpussets 1.553 individuelle tekstuddrag indeholder hver især mellem 140 og 180 løbende tekstord (ekskl. interpunktionstegn) og består af et eller flere efterfølgende afsnit, der er automatisk udvalgt fra DSL's elektroniske tekstbibliotek. De første ca. 100.000 løbende tekstord (643 tekstuddrag) er udvalgt fra Den Danske Ordbogs (DDO) SGML-opmarkerede tekstkorpus på 40 mio. løbende tekstord bestående af uddrag af dagblade, bøger, tidsskrifter osv[5]. De sidste ca. 150.000 løbende tekstord (910 tekstuddrag) er udvalgt fra DSL's avistekstbibliotek og er blevet SGML-opmarkeret til PAROLE-korpusset. De fleste tekster i dette korpus er således uddrag af aviskster (jf. appendiks 8.4 om korpusteksternes klassificering ifølge medium). Den oprindelige automatiske udvælgelse af korpustekster blev revideret lidt, idet et par mere eller mindre uegnede tekster (som f.eks. tv-programfortegnelser, strikkeopskrifter osv.) blev fjernet. Tekstuddragene fra ovennævnte to forskellige tekstkilder hos DSL blev desuden harmoniseret mht. obligatoriske SGML-opmarkeringskoder samt et par forskellige tegn (jf. afsnit 4.1.1 om anførselstegn og afsnit 4.1.3 om bindestreger og tankestreger).

Morfosyntaktisk analyse af tekstuddragene med DAN-TWOL-algoritmen
Som en nødvendig del af præproceseringsfasen inden opslag i DAN-TWOLs leksikon blev de udvalgte korpustekster først behandlet af en ordinddeler ('tokeniser'), der er udviklet af Thomas Bilgram til anvendelse i forbindelse med DAN-TWOL-algoritmen. Ordinddeleren er et program, der afgrænser og identificerer løbende tekstord, symboler og interpunktionstegn i korpusteksterne. Dette gør det muligt for DAN-TWOL-algoritmen at tildele de løbende tekstord mm. alle deres mulige morfosyntaktiske analyser.

Manuel udvælgelse af de korrekte morfosyntaktiske DAN-TWOL-analyser
Den manuelle udvælgelse af den korrekte DAN-TWOL-analyse for hvert eneste løbende tekstord — dvs. den "kontekstuelle disambiguering" eller "korpustaggingen" — blev for den største del af tekstmaterialet foretaget parallelt af to personer — "korpustaggerne" — via forbindelse til en server hos Institut for Lingvistik, Finsk og Ungarsk ved Aarhus Universitet. To identiske versioner af korpusteksterne blev "tagget" hver for sig vha. en speciel tagging-mode til **`emacs`**-editoren. Det var således kun muligt for korpustaggerne at bevæge sig igennem teksterne og tilføje eller slette en **<correct!>** markering (samt et par andre markeringer) ud for den korrekte DAN-TWOL-analyse. Formålet med denne fremgangsmåde var at formindske risikoen for, at korpustaggerne ved en fejltagelse ændrede i selve korpusteksterne, samt at muliggøre en automatisk sammenligning af de parallelt taggede tekster med UNIX-kommandoen **`diff`**. Evt. fejl eller uoverensstemmelser i tildelingen af analyser kunne således hurtigt identificeres, diskuteres og behandles undervejs i korpustaggingens forløb.

Konvertering til PAROLEs tagsæt og format
Til PAROLE-projektets korpus var det nødvendigt i en efterredigeringsfase at lave nogle små, systematiske ændringer i korpusannoteringen for at tilpasse den til PAROLEs egne lingvistiske specifikationer[6]. Derefter blev korpusteksterne automatisk konverteret til PAROLEs tagsæt og format. Til sidst blev korpusteksterne SGML-opmarkeret ifølge den fælles PAROLE-opmarkeringsstandard (jf. afsnit 3 om korpus- og tekstopmarkering).

Udvikling af en Constraintgrammatik til dansk
I forbindelse med Thomas Bilgrams Ph.D.-projekt har de taggede korpustekster været direkte anvendelige som træningsmateriale til hans igangværende udvikling af en

[5] Jf. beskrivelsen af sammensætningen af DDOs korpus i Norling-Christensen & Asmussen, 1999.
[6] De danske lingvistiske specifikationer i PAROLE-projektet beskrives mere udførligt bl.a. i Braasch & Norling-Christensen, 1997.

Constraintgrammatik (DAN-CG) til automatisk syntaktisk analyse af dansk (jf. Bilgram & Keson, 1998).

2.2 DAN-TWOL-algoritmen

DAN-TWOL er navnet på en algoritme til automatisk morfologisk analyse af danske tekster. Den er udviklet af stud. Ph.D. Thomas Bilgram og er baseret på to-niveau morfologien, som den er beskrevet i bl.a. Koskenniemi, 1983. To-niveau morfologien er baseret på en behandling af tekstordet som en overfladerepræsentation af en underliggende leksikalsk repræsentation, samt en beskrivelse af forbindelsen mellem disse to repræsentationer. Arbejdet med udviklingen af DAN-TWOL-algoritmen indgik som en indledende fase i Thomas Bilgrams Ph.D.-projekt. Projektets overordnede formål er udviklingen af en automatisk syntaktisk analyse af naturlige danske skriftsprogtekster, baseret på en dansk tilpasning af Constraintgrammatikken, som den er beskrevet i bl.a. Karlsson et al, 1994.

DAN-TWOLs leksikon, der er baseret på oplysninger fra en maskinlæsbar udgave af Retskrivningsordbogen fra 1986 (RO86), består af ca. 46.000 danske opslagsord, og indeholder i en tillempet form oplysninger om ordklasse og bøjning fra RO86. Regelmæssige og enskodede oplysninger i RO86 er automatisk overført til DAN-TWOLs format, mens en restgruppe på ca. 10.000 uregelmæssige ord er delvis manuelt indført i DAN-TWOLs leksikon. Ovennævnte oplysninger er i DAN-TWOL-leksikonet desuden forsynet med (*i*) et subleksikon med fleksive og derivative affikser, (*ii*) regler for behandling af morfotaktisk kontinuation, (*iii*) regler for sammenknytning af rodleksemer til analyse af komposita, samt (*iv*) mekanismer til genkendelse og behandling af allomorfisk variation.

I DAN-TWOL-analysen ”genkendes” og analyseres et ord ved at programmet ”accepterer” ét enkelt bogstav i tekststrengen ad gangen. Når en accepteret bogstavsekvens svarer til et morfem, der er tilladt ifølge DAN-TWOL-algoritmen, tildeles dette morfem en værdi. Værdierne for ordklasse ('N') og genus ('FLS') i eksemplerne i **Figur 3** nedenfor tildeles ordet efter opslag i DAN-TWOLs leksikon, mens værdierne for numerus, bestemthed og kasus først genereres under selve DAN-TWOLs analyse af ordet.

Figur 3: Eksempler på DAN-TWOL-analyser (fra Bilgram & Keson, 1998)

```
bil+Ø+Ø+Ø     = N FLS SG UBEST NOM  bil+Ø+Ø+s     = N FLS SG UBEST GEN
bil+er+Ø+Ø    = N FLS PL UBEST NOM  bil+er+Ø+s    = N FLS PL UBEST GEN
bil+Ø+en+Ø    = N FLS SG BEST  NOM  bil+Ø+en+s    = N FLS SG BEST  GEN
bil+er+ne+Ø   = N FLS PL BEST  NOM  bil+er+ne+s   = N FLS PL BEST  GEN
```

Resultatet af den automatiske analyse af en tekst med DAN-TWOL er, at hvert tekstord — for så vidt at det optræder i DAN-TWOLs leksikon — får tildelt én eller flere morfosyntaktiske analyser. Et sæt analyser for et givent tekstord kaldes ifølge to-niveau terminologien en 'kohorte'. Under korpustaggingen udvælger korpustaggerne den korrekte DAN-TWOL-analyse i kohorten og markerer den med **<correct!>** (jf. afsnit 2.1 ovenfor). Formålet med DAN-CG-projektet er at træne en automatisk Constraintgrammatisk algoritme til at foretage det samme valg blandt kohortens medlemmer. En mere udførlig beskrivelse af DAN-TWOL findes bl.a. i Bilgram & Arndt, 1993, Bilgram, 1994 samt Bilgram & Keson, 1998.

2.3 PAROLEs tagsæt og format

Konverteringen af de taggede tekster til PAROLE-formatet bestod af to opgaver: (*i*) SGML-opmarkering af teksterne ifølge PAROLEs fælles Corpus Encoding Standard (CES) (jf. afsnit 3 nedenfor) samt (*ii*) konvertering af DAN-TWOLs lemmaformer og morfosyntaktiske analyser til PAROLEs eget tagsæt og format, som er specificeret i Volz & Lenz, 1996 efter anbefalinger

fra EAGLES i Monachini, Calzolari et al, 1995. PAROLE-tagsættet består af et fast antal fælles morfosyntaktiske træk, som er blevet udvidet med et antal sprog-specifikke træk for hvert PAROLE-sprog. Baggrunden for de danske morfosyntaktiske træk beskrives mere udførligt i Braasch & Norling-Christensen, 1997.

Tabellen i **Figur 4** nedenfor er baseret på modellen i Volz & Lenz, 1996 og viser indholdet af det danske PAROLE-tagsæt. De grå felter repræsenterer de fælles PAROLE-træk, som ikke anvendes i det danske tagsæt, mens de hvide felter repræsenterer de træk, som anvendes i det danske taggede PAROLE-korpus. De morfosyntaktiske træk, der er fælles for alle PAROLE-sprogene, findes på pladserne 1 til 7 i tabellen, mens de sprog-specifikke træk findes på plads 8 og opefter. Af tabellen fremgår det således, at bestemthed ('Definiteness'), transkategorisering ('TrCat'), diatese ('Voice'), kasus (for participier[7]) ('Case'), refleksivitet ('Reflexive') samt stilleje ('Register') er sprog-specifikke udvidelser i det danske tagsæt.

Figur 4: Det danske PAROLE-tagsæt

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Noun		**Gender**	**Number**	**Case**			**Definiteness**			
Verb		**Mood**	**Tense**	Person	**Number**	**Gender**	**Definiteness**	**TrCat**	**Voice**	**Case**
Adj		**Degree**	**Gender**	**Number**	**Case**		**Definiteness**	**TrCat**		
Pron		**Person**	**Gender**	**Number**	**Case**	**Possessor**	**Reflexive**	**Register**		
Det		Person	Gender	Number	Case	Possessor				
Art		Gender	Number	Case						
Adv		**Degree**	Function	Wh-ness						
Adpos		Formation	Gender	Number						
Conj		Ctype	Coord-posit							
Num		Gender	Number	Case						
Interj										
Residual										
Unique										

Første plads i tagsættet ovenfor er reserveret til en angivelse af ordklassen (kaldet 'CatGram'), og anden plads er reserveret til en yderligere underinddeling af denne ordklasse (kaldet 'SsCatGram'). Hver plads udfyldes normalt med et bogstav eller tal, der repræsenterer en bestemt værdi for det givne træk (jf. også appendiks 8.2, som er en fortegnelse over samtlige værdier i det danske PAROLE-tagsæt). Ifølge PAROLE-specifikationerne skal mindst 50.000 af de 250.000 morfosyntaktisk taggede tekstord være tagget med det fulde tagsæt, mens resten kan nøjes med en angivelse af ordklasse (dvs. 'CatGram') alene. Dog er alle 250.000 løbende tekstord i det danske korpus tagget med det fulde danske tagsæt.

I korpusteksterne angives de morfosyntaktiske analyser i et SGML-opmarkeret format, således at lemmaet og den morfosyntaktiske analyse (’msd’) er fremstillet som SGML-attributter til tekstordet, som vist i de tre eksempler i **Figur 5** nedenfor (Disse tre eksempler er taget fra første sætning i den allerførste korpustekst).

Figur 5: Tre eksempler på morfosyntaktisk taggede tekstord

```
<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>
```

Sekvensen af bogstaver (og tal) i en 'msd-streng', dvs. det SGML-attribut, der repræsenterer den morfosyntaktiske analyse af tekstordet, svarer til værdier på pladserne i tabellen i **Figur 4** ovenfor. For at sikre, at sekvensen i strengen altid svarer til pladserne i tabellen ovenfor, anvendes et par forskellige tegn som “pladsholdere”.

[7] Jf. afsnit 5.2.6. om brug af kasus til participier.

Lighedstegnet (=) anvendes som pladsholder i msd-strengen til at angive, at det pågældende træk, der er knyttet til denne plads, ikke anvendes i hele ordklassen ('CatGram'). Det fremgår f.eks. af ovenstående eksempler i **Figur 5**, at en msd-streng, der indeholder analysen af et appellativ (som f.eks. *historikere*), og en msd-streng, der indeholder analysen af et proprium (som f.eks. *Andronik*), har lige mange pladser (otte). Dog er plads 6 og 7 ikke relevante for hverken appellativer eller proprier, og derfor udfyldes de i begge tilfælde med et lighedstegn. Lighedstegnet repræsenterer således de grå felter i tagsættet.

Bindestregen (-) anvendes som pladsholder i en msd-streng for at angive, at det pågældende træk, der er knyttet til denne plads, ikke er relevant i denne underinddeling af ordklassen ('SsCatGram'), men er relevant for en anden underinddeling af den pågældende ordklasse. Således indeholder f.eks. plads 3, 4 og 8 af en msd-streng for et proprium (som f.eks. *Andronik* ovenfor) altid en bindestreg, fordi de tre pågældende morfologiske træk (hhv. genus, numerus og bestemthed) ikke anvendes i analysen af danske proprier, men kun i analysen af danske appellativer (som f.eks. *historikere* ovenfor).

Til sidst anvendes kantede parenteser ([ ]) til at angive, at mere end én værdi er relevant for et pågældende træk. I DAN-TWOL er f.eks. et adjektiv i pluralis (som f.eks. *russiske* i **Figur 5** ovenfor) ikke eksplicit markeret for bøjning i genus eller bestemthed, da disse (fire potentielle) bøjningsformer falder samme i pluralisformen. Denne type regelbunden "træksammenfald" er heller ikke nærmere udspecificeret i selve PAROLE-korpusset. I msd-strengen er adjektiver i pluralis således underspecificeret ved, at de to mulige værdier for genus (fælleskøn/intetkøn) og de to mulige værdier for bestemthed (bestemt/ubestemt) angives mellem kantede parenteser (som vist i **Figur 5** ovenfor). Når kantede parenteser anvendes i en msd-streng, afgrænser de altid en liste over alle de mulige værdier for den pågældende plads[8]. Således kan der faktisk stå mere end ét tegn på en enkelt plads i msd-strengen.

3. Korpus- og tekstopmarkering

Hele PAROLE-korpusset — inkl. hvert eneste af de 1.553 tekstuddrag — er opmarkeret med SGML-koder ifølge reglerne i PAROLEs fælles Corpus Encoding Standard (CES). PAROLE CES'en er udførligt beskrevet i Ridings, 1996, hvor den sammenlignes med to andre eksisterende CES'er, Text Encoding Initiative (TEI) CES'en (Sperberg-McQueen & Burnard, 1994) og EAGLES CES'en (Ide et al, 1995)[9].

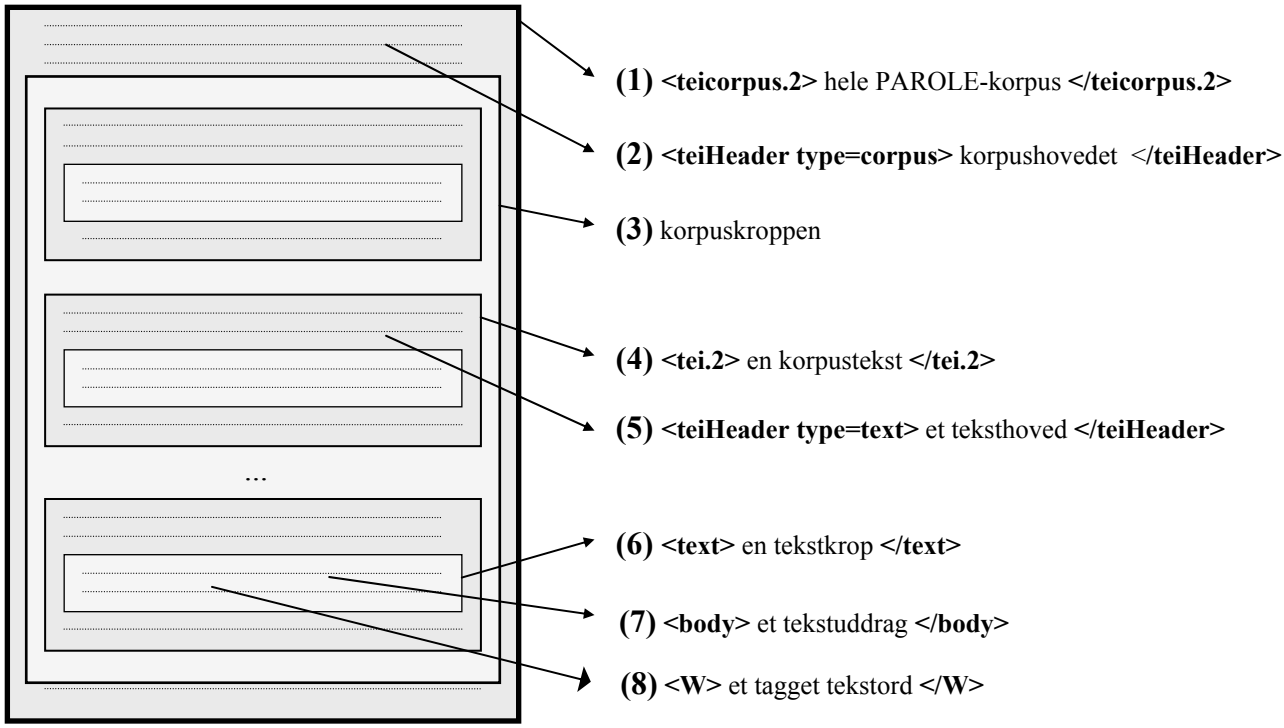
At et tekstkorpus er opmarkeret ifølge en CES betyder, at den overordnede struktur i korpusset er fastlagt på forhånd af reglerne i CES'en. Denne struktur udtrykkes eksplicit vha. SGML-koder, og CES'en bestemmer først og fremmest, hvilke SGML-koder der er obligatoriske og hvilke SGML-koder der er tilladte i opmarkeringen af korpusset. 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 bestemmer desuden hvilke SGML-koder der må (eller skal) optræde inde i de løbende korpustekster for at angive deres interne struktur, samt hvordan disse SGML-koder skal indlejres i hinanden.

[8] Den samme underspecificering kunne også repræsenteres ved at markere den pågældende plads med en "underspecificeret" markør (f.eks. '0') i stedet for en liste, men denne mulighed var ikke tilgængelig ifølge specifikationerne af PAROLE-tagsættet.

[9] I Ridings, 1996, s. 2 står bl.a. "The PAROLE standard follows the recommendation made by EAGLES as far as the detail of information that is to be encoded is concerned, but a text encoded according to the PAROLE standard will parse with the TEI-dtd. When the EAGLES description deviates from TEI with respect to technicalities, PAROLE follows TEI."

Hele PAROLE-korpusset er omgivet af SGML-startkoden **<teicorpus.2>** og slutkoden **</teicorpus.2>** (jf. **Figur 6** nedenfor). Dens overordnede struktur består af to dele: (*i*) et korpushoved ('corpus header'), som er omgivet af SGML-koderne **<teiHeader type=corpus>** og **</teiHeader>** og beskriver PAROLE-korpusset som helhed, samt (*ii*) en korpuskrop, som indeholder de individuelle korpustekster. Korpusteksterne 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 en beskrivelse af den individuelle korpustekst. Tekstkroppen er omgivet af **<text>** og **</text>** og indeholder selve tekstuddraget (mellem **<body>** og **</body>**). I det morfosyntaktisk taggede PAROLE-korpus er hvert eneste løbende tekstord (inkl. interpunktionstegnene) desuden omgivet af SGML-startkoden **<W>** og slutkoden **</W>** (jf. **Figur 5** i afsnit 2.3 ovenfor).

Figur 6: PAROLE-korpussets SGML-opmarkerede struktur



3.1 PAROLE-korpushovedet

PAROLE-korpushovedet optræder allerførst i korpusset efter en kort række ENTITY-beskrivelser. Den begynder med SGML-startkoden **<teiHeader type=corpus**> og afsluttes med SGML-slutkoden **</teiHeader>**. Det danske PAROLE-korpushoved er gengivet i en forkortet udgave i **Figur 7** nedenfor[10], mens en fuldstændig fortegnelse over koderne for medium, genre og emne (oversat til dansk) findes i appendiks 8.4.

Korpushovedet indeholder en overordnet beskrivelse af hele korpusset og består bl.a. af oplysninger om korpussets ophavsmænd (mellem **<respStmt>** og **</respStmt>**), antallet af løbende tekstord i korpusset (mellem **<extent>** og **</extent>**[11]) samt diverse oplysninger om udgivelsen af korpusset (mellem **<publicationStmt>** og **</publicationStmt>**). Korpushovedet

[10] Af pladshensyn vises PAROLE-korpushovedet i to spalter her i vejledningen.
[11] I det morfosyntaktisk taggede korpus henviser antallet af løbende ord i **<extent>** til antallet af ordformer ekskl. interpunktionstegn.

indeholder desuden en fuldstændig fortegnelse (mellem **<classDecl>** og **</classDecl>**) over alle de koder, der anvendes i de individuelle tekstheadere for at angive tekstuddragenes klassifikation ifølge medium, genre og emne.

Figur 7: PAROLE-korpushovedet

```
<!DOCTYPE teicorpus.2 system "tei2.dtd" [
<!ENTITY % TEI.prose 'INCLUDE'>
<!ENTITY % TEI.analysis 'INCLUDE'>
<!ENTITY % TEI.corpus 'INCLUDE'>
<!ENTITY danish SDATA "iso88591.wsd">
<!ENTITY % TEI.extensions.ent SYSTEM "parole.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "parole.dtd">
]>
<teicorpus.2>
 <teiHeader type=corpus>
  <fileDesc>
   <titleStmt>
    <title>PAROLE-DK</title>
    <respStmt>
     <name>Britt-Katrin Keson, Ole Norling-Christensen et al.</name>
     <resp>Collection and Encoding</resp>
    </respStmt>
   </titleStmt>
   <extent words=250,209>250,209 running words</extent>
   <publicationStmt>
    <distributor>Society for Danish Language and Literature (DSL)</distributor>
    <address>
     <addrline>Christians Brygge1, 1., DK-1219 Copenhagen K.
    </address>
    <availability status=restricted>
      <p>by agreement with distributor
    </availability>
    <Date>1998-04-23</Date>
   </publicationStmt>
   <sourceDesc>
    <biblStruct>
     <monogr>
      <author>several</author>
      <title>DSL Text Archive</title>
      <imprint><pubPlace>Copenhagen</pubPlace></imprint>
     </monogr>
    </biblStruct>
   </sourceDesc>
  </fileDesc>
  <encodingDesc>
<projectDesc><p>please refer to the PAROLE documentation </projectDesc>
  <classDecl>
    <taxonomy ID=P>
   <bibl>Parole Corpus</bibl>
    <category ID=P.M>
     <catDesc>Medium</catDesc>
     <category ID=P.M1>
      <catDesc>book</catDesc>
     </category>
     <category ID=P.M2>
      <catDesc>newspaper</catDesc>
     </category>
     <category ID=P.M3>
      <catDesc>periodical</catDesc>
      <category ID=P.M3.1>
       <catDesc>journal</catDesc>
      </category>
      <category ID=P.M3.2>
       <catDesc>local</catDesc>
      </category>
      <category ID=P.M3.3>
       <catDesc>magazine</catDesc>
      </category>
                ...
      <category ID=P.T9.9>
       <catDesc>consumer items</catDesc>
      </category>
      <category ID=P.T9.10>
        <catDesc>traffic</catDesc>
       </category>
     </category>
    </category>
   </category>
     </taxonomy>
  </classDecl>
 </encodingDesc>
 <profileDesc>
  <creation></creation>
  <langUsage>
   <language id=DA>Danish
  </langUsage>
 </profileDesc>
</teiHeader>
          ...
</teicorpus.2>
```

3.2 PAROLE-teksthovederne

Hvert eneste af de 1.553 korpustekster er omgivet af **<tei.2>** og **</tei.2>** og indeholder et teksthoved (mellem **<teiHeader type= text>** og **</textHeader>**) og en tekstkrop (mellem **<text>** og **</text>**). Selve tekstuddraget er omgivet af startkoden **<body>** og slutkoden **</body>**. Inde i startkoden **<text>** findes et **text id**-attribut, der entydigt refererer til tekstens interne id-kode i DSL's tekstbibliotek.

Figur 8: Et PAROLE-teksthoved

```
<tei.2>
 <teiHeader type=text>
  <fileDesc>
   <titleStmt>
    <title>Tagged sample of: 'Jeltsins skæbnetime'</title>
   </titleStmt>
   <extent words=158>158 running words</extent>
   <publicationStmt>
    <distributor>PAROLE-DK</distributor>
    <address><addrline>Christians Brygge 1,1., DK-1219 Copenhagen K.</address>
    <date>1998-03-20</date>
    <availability status=restricted><p>by agreement with distributor</availability>
   </publicationStmt>
   <sourceDesc>
     <biblStruct>
      <analytic>
       <title>Jeltsins skæbnetime</title>
       <author gender=m born=1925>Nikulin, Leon</author>
      </analytic>
      <monogr>
       <imprint><pubPlace>Denmark</pubPlace>
        <publisher>Det Fri Aktuelt</publisher>
        <date>1992-12-01</date>
       </imprint>
      </monogr>
     </biblStruct>
   </sourceDesc>
  </fileDesc>
   <profileDesc>
    <creation>1992-12-01</creation>
    <langUsage><language>Danish</langUsage>
    <textClass>
     <catRef target="P.M2">
     <catRef target="P.G4.8">
     <catRef target="P.T9.3">
    </textClass>
   </profileDesc>
 </teiHeader>
<text id=AJK>
<body>
<div1 type=main>
<p>
<W lemma="to" msd="AC---U=--">To</W>
<W lemma="kendt" msd="ANP[CN]PU=[DI]-">kendte</W>
<W lemma="russisk" msd="ANP[CN]PU=[DI]-">russiske</W>
<W lemma="historiker" msd="NCCPU==I">historikere</W>
<W lemma="Andronik" msd="NP--U==-">Andronik</W>
<W lemma="Mirganjan" msd="NP--U==-">Mirganjan</W>
                    ...
<W lemma="diktaroriske" msd="XX">diktaroriske</W>
<W lemma="befølelser" msd="XX">befølelser</W>
<W lemma="." msd="XP">.</W>
</p>
</div1>
</body>
</text>
</tei.2>
```

Teksthovedet indeholder oplysninger om det individuelle tekstuddrag (jf. eksempelteksthovedet i **Figur 8** ovenfor). Af særlig interesse i teksthovedet er: antallet af løbende tekstord i tekstuddraget (mellem **<extent>** og **</extent**>), titlen på den oprindelige kildetekst, som tekstuddraget stammer fra[12] (mellem **<title>** og **</title**>), navnet på forfatteren/forfatterne til teksten, og — hvis oplysningerne var tilgængelige i DSL's tekstbibliotek — forfatterens/forfatternes køn (**gender**) og fødselsår (**born**) (mellem **<author>** og **</author**>[13]), samt oplysninger om tekstens udgivelsessted og -dato (mellem **<imprint>** og **</imprint>**). Desuden indeholder teksthovedet koder, som henviser til den pågældende teksts klassifikation ifølge medium, genre og emne (mellem **<textClass>** og **</textClass>**). Nøglen til disse koder findes som nævnt i korpushovedet og er også gengivet i fortegnelsen i appendiks 8.4[14]. Heraf fremgår det f.eks., at teksthovedet i **Figur 8** hører til et tekstuddrag fra mediet 'dagblad' (**P.M2**), af genren 'kronik' (**P.G4.8**) og om emnet 'politik' (**P.T9.3**).

3.3 SGML-koder inde i de taggede tekster

Hvert eneste af de 1.553 morfosyntaktisk taggede tekstuddrag omgives af **<body>** og **</body>** og består af løbende tekstord, symboler osv., der hver især er omgivet af koderne **<W>** og **</W>**. Inde i selve tekstuddragene findes der dog desuden et par yderligere SGML-koder, som er anvendt til at angive teksternes interne struktur.

Figur 9: SGML-koder inde i de taggede tekster

[12] Af tekstens titel angives dog højst de første 40 karakterer. Hvis teksten ikke har en titel, eller hvis titlen ikke er registreret i DSL's tekstbibliotek, angives titlen blot som 'UNTITLED' i dette felt (der er 71 korpustekster uden titel i PAROLE-korpusset).

[13] Hvis forfatterens navn ikke er registreret i DSL's tekstbibliotek, angives forfatterens navn blot som 'UNKNOWN' i dette felt (der er 196 korpustekster uden forfatternavn i PAROLE-korpusset).

[14] Disse oplysninger om teksternes medium, genre og emne er automatisk konverteret fra oplysninger i DSL's tekstbibliotek.

Startkode	Slutkode	Beskrivelse
<div1 type=main>	**</div1>**	omgiver en større, ikke nærmere beskrevet tekstenhed
<div1 type=caption>	**</div1>**	omgiver en billedtekst*
<div1 type=external>	**</div1>**	omgiver en tekstenhed der optræder uden for den løbende tekst*
** **	** **	omgiver et afsnit ('paragraph')
<s>	**</s>**	omgiver en sætning [15]
<note resp=auth>	**</note>**	omgiver en note om teksten der er indsat af forfatteren/forfatterne til teksten*
<note resp=comp>	**</note>**	omgiver en note om teksten der er indsat af en DDO redaktør under sammensætningen af DDO's tekstkorpus* (sjælden)
<hi>	**</hi>**	omgiver fremhævet tekst ('highlighted') uden nærmere angivelse af hvordan teksten er fremhævet*

De SGML-koder, der er markeret med en stjerne (*) i **Figur 9** ovenfor, er automatisk konverteret fra allerede eksisterende SGML-koder i DDO's tekstkorpus og optræder således kun i de korpustekster, der svarer til de første 100.000 tekstord i PAROLE-korpusset. De andre SGML-koder kan forefindes i alle korpusteksterne.

4. Ordinddeling

En ordform er et løbende tekstord (eller evt. et symbol eller interpunktionstegn) i den skriftlige udformning, hvormed det optræder i korpusteksten. Det er også denne form, der slås op i DAN-TWOLs morfologiske leksikon, hvor den tildeles en eller flere lemmaformer og morfosyntaktiske analyser. For at opslag i DAN-TWOLs leksikon er mulige, skal DAN-TWOL-tokeniseren først udskille alle SGML-koder samt identificere alle ordformer, symboler og interpunktionstegn i teksten. Set fra DAN-TWOL-tokeniserens synspunkt svarer en ordform stort set til en sammenhængende sekvens af bogstaver, tal eller visse andre tegn (som f.eks. bindestreger), der er omgivet af blanktegn og interpunktionstegn (dog kan der i visse tilfælde optræde blanktegn inde i en ordform). Eksemplet i **Figur 10** nedenfor viser hvordan DAN-TWOL-tokeniseren har behandlet første afsnit af den første korpustekst i PAROLE-korpusset.

Figur 10: Resultat af DAN-TWOL-tokeniserens ordinddeling

To kendte russiske historikere Andronik Mirganjan og Igor Klamkin tror ikke, at Rusland kan udvikles uden en "jernnæve". De hævder, at Ruslands vej til demokrati går gennem diktatur. I en af deres artikler hedder det: "I et autoritært regime lagdel samfundet og forskellige interesser modnes. Og når deres repræsentanter er parate til at gå i struben på hinanden, så stopper en jernnæve" det. På den måde skabes hele tiden betingelserne for en harmonisering af interesser og følgelig for demokratiske reformer".	To kendte russiske historikere Andronik Mirganjan og Igor Klamkin tror ikke , at Rusland kan udvikles uden en " jernnæve " . De hævder , at Ruslands vej til demokrati går gennem diktatur . I en af deres artikler hedder det : " I et autoritært regime lagdel samfundet og forskellige interesser modnes . Og når deres repræsentanter er parate til at gå i struben på hinanden , så stopper en jernnæve " det . På den måde skabes hele tiden betingelserne for en harmonisering af interesser og følgelig for demokratiske reformer " .

Som det fremgår af eksemplet ovenfor, er de ordformer, der er identificeret af DAN-TWOL-tokeniseren og vil blive slået op i DAN-TWOL-leksikonet, normalt alenestående ordformer, der er omgivet af blanktegn og interpunktionstegn. Ordformer er således normalt ikke flerordsforbindelser, der indeholder blanktegn. Det har således i høj grad været DAN-TWOL-tokeniserens afgørelse, hvordan ordformerne i disse tekster er blevet identificeret, evt. samlet og så slået op i DAN-TWOL-leksikonet.

Af ovenstående eksempel i **Figur 10** fremgår det, at f.eks. flerleddede navne (som *Andronik Mirganin* og *Igor Klamkin*) ikke samles af DAN-TWOL-tokeniseren som flerordsforbindelser, men analyseres som selvstændige ordformer. Dette gælder også for andre kendte typer af

[15] Sætningerne er segmenteret automatisk efter en enkel algoritme, hvilket medfører, at der kan optræde enkelte fejlsegmenteringer (især i forbindelse med anførselstegn).

flerordsforbindelser, som f.eks. substantivgrupper (jf. afsnit 5.1), sammensatte verballed og partikelverber (jf. afsnit 5.2 og 5.4.2), *at*-infinitivformen (jf. afsnit 5.2.4 og 5.7.1), sammensatte præpositioner (jf. afsnit 5.4.1 og afsnit 6.3.4), flerleddede konjunktioner (jf. afsnit 5.4.3) og flerleddede pronominer (jf. afsnit 5.5.3). Der eksisterer dog nogle få undtagelser, hvor flerordsforbindelser faktisk analyseres samlet i PAROLE-korpusteksterne, og disse gennemgås mere udførligt i afsnit 4.2 nedenfor. Appendiks 8.5 indeholder desuden en fuldstændig fortegnelse over alle analyserede flerordsforbindelser i PAROLE-korpusset.

4.1 Interpunktionstegn og symboler

DAN-TWOL-tokeniseren har identificeret nedenstående 13 interpunktionstegn ('XP') og 7 symboler ('XS') i PAROLE-korpusteksterne. Mens symboler altid adskilles fra deres omgivelser af DAN-TWOL-tokeniseren, kan interpunktionstegn enten stå alene eller indgå i ordformer, som f.eks. punktummet, der kan indgå i forkortelser som et forkortelsespunktum. Anførselstegn, bindestreger og skråstreger kan ligeledes optræde inde i ordformer, hvilket har medført, at tokeniseren ikke altid har afgrænset ordformerne korrekt (jf. afsnit 6.1.2).

'&'-tegnet kan også stå alene eller indgå i ordformer. Dette tegn tildeles altid en analyse som en konjunktion ('CC'), når tegnet står alene i PAROLE-korpusset (jf. afsnit 5.4.3 om konjunktioner).

Figur 11: Interpunktionstegn og symboler (samt '&'-tegnet)

```
<W lemma="!" msd="XP">!</W>
<W lemma="&amp;quot;" msd="XP">"</W>
<W lemma="(" msd="XP">(</W>
<W lemma=")" msd="XP">)</W>
<W lemma="," msd="XP">,</W>
<W lemma="." msd="XP">.</W>
<W lemma=".." msd="XP">..</W>
<W lemma="..." msd="XP">...</W>
<W lemma=":" msd="XP">:</W>
<W lemma=";" msd="XP">;</W>
<W lemma="?" msd="XP">?</W>
<W lemma=" - " msd="XP"> - </W>
<W lemma="/" msd="XP">/</W>
```

```
<W lemma="$" msd="XS">$</W>
<W lemma="%" msd="XS">%</W>
<W lemma="*" msd="XS">*</W>
<W lemma="+" msd="XS">+</W>
<W lemma="=" msd="XS">=</W>
<W lemma="§" msd="XS">§</W>
<W lemma="°" msd="XS">°</W>

<W lemma="&amp;amp;" msd="CC">&amp;amp;</W>
```

4.1.1 Anførselstegnet og '&'-tegnet

Anvendelsen af SGML-koder i PAROLE-korpusset medfører, at der er to tegn, der ikke kan gengives som sig selv hverken i selve korpusteksterne eller i korpus- og tekstheaderne, uden at det ville føre til problemer for en SGML-parser. Da både det dobbelte anførselstegn (") og '&'-tegnet benyttes i SGML-koderne, skal disse to tegn derfor altid omskrives til hhv. ' " ' og ' & '[16]. Dette er forklaringen på f.eks. følgende kryptiske teksttitel fra PAROLE-korpusset: **<title>'P&T: Betal for andres portofusk'</title>**.

For at entydiggøre anvendelsen af det ufordoblede anførselstegn eller pling ('), der jo også kan have apostrofens funktion (jf. Retskrivningsordbogen fra 1996 - RO96, §6[17]), er alle forekomster af det ufordoblede anførselstegn, der er anvendt som anførselstegn, konverteret til det dobbelte anførselstegn (") inden behandlingen af DAN-TWOL-tokeniseren. Som vist i

[16] Desuden skal de skarpe parenteser ' < ' og ' > ' af indlysende årsager omskrives til hhv. ' < ' og ' &rt; ', dog forekommer disse tegn ikke i PAROLE-korpusteksterne.
[17] Fremover i vejledningen vil der blive henvist til paragraf X.X i RO96 vha. udtrykket "RO96, §X.X".

eksemplerne nedenfor er anførselstegnet tilladt inde i selve ordformen (dvs. mellem ' > ' og ' < '), mens det af indlysende årsager ikke er tilladt i lemmaformen, hvor det derfor erstattes af ' " '. De anførselstegn, der optræde inde i en ordform, som f.eks. i *"Mellem-Os"-læsere*, er dog fjernet helt i lemmaformen.

..Andronik Mirganjan og Igor Klamkin tror ikke, at Rusland kan udvikles uden en "jernnæve"..[18]
<W lemma=""" msd="XP">"</W> .. <W lemma=""" msd="XP">"</W>
..programmet, hvor "Mellem-Os"-læserne også kan glæde sig til den religiøse trommedans..
<W lemma="Mellem-Os-læser" msd="NCCPU==D">"Mellem-Os"-læserne</W>
..vi er vel i bund og grund at sammenligne med Kjeld & Hilda. Jamen, hallo, vi er enige..
<W lemma="&" msd="CC">&</W>
..sammen med brevet ligger et girokort fra P&T, som lakonisk fortæller at T. Hansen skal indbetale..
<W lemma="P&T" msd="NP—U==-">P&T</W>

4.1.2 Forkortelsespunktummet

Når en forkortelse, der afsluttes med et forkortelsespunktum, står sidst i en sætning, sættes der normalt kun ét punktum (jf. RO96, §42.4). Når en sætning i korpusteksterne afsluttes med en forkortelse inkl. forkortelsespunktum, har DAN-TWOL-tokeniseren derfor "adskilt" forkortelsespunktummet og slutpunktummet, således at den morfosyntaktisk taggede sætning indeholder både forkortelsen (inkl. forkortelsespunktummet) og slutpunktummet. Undtagelser er overskrifter mm., som normalt ikke afsluttes med et slutpunktum (RO96, §41.2.a). Hvis brugeren af PAROLE-korpusset evt. ønsker at generere de oprindelige (dvs. "utaggede") tekster på baggrund af de morfosyntaktisk taggede korpustekster, er det således nødvendigt at slå disse to punktummer sammen igen.

..Møllers store malerier solgte som regel for priser mellem 25.000 og 35.000 kr. Han kunne leve af sin kunst..
<W lemma="krone" msd="NCCPU==I">kr.</W> <W lemma="." msd="XP">.</W>
..IG Metall endte med at sige nej til det sidste tilbud, der lød på en lønfremgang på 5,7 pct. </p>..
<W lemma="procent" msd="NCCPU==I">pct.</W> <W lemma="." msd="XP">.</W>

4.1.3 Bindestregen, tankestregen og skråstregen

Bindestregens anvendelse i orddelinger ved linjeskift (RO96, §63.1) forekommer ikke i disse korpustekster, da bindestreger med denne funktion blev fjernet og de pågældende orddele føjet sammen, inden teksterne blev analyseret af DAN-TWOL-tokeniseren. Dog forekommer bindestregen stadig i korpusteksterne i de typer af sammensætninger, hvor bindestreger normalt anvendes (RO96, §63), dvs.: (*i*) sammensætninger med forkortelser, taltegn og andre symboler, (*ii*) "udenlandske" sammensætninger, (*iii*) gruppesammensætninger (jf. afsnit 4.2.2), (*iv*) "usædvanlige" sammensætninger, (*v*) konjunktioner med udeladt fælles orddel (jf. afsnit 6.1.1), (*vi*) sammensætninger med sidestillede led samt (vii) med betydningen 'fra ... til'. I alle disse tilfælde blev bindestregen beholdt i ordformen. I de sidste tre tilfælde var det dog ikke muligt for DAN-TWOL at tildele en morfosyntaktisk analyse (derfor tildeltes 'XX'-analysen som vist i eksemplerne nedenfor).

Tankestregen (–) er også blevet konverteret til en bindestreg (-) inden teksterne blev analyseret af DAN-TWOL-tokeniseren, hvilket betyder, at bindestregen også kan optræde alenestående i korpusteksterne.

Skråstregen anvendes ifølge RO96, §66 typisk: (*i*) til at angive et valg mellem flere muligheder, (*ii*) med betydningen 'pr.', (*iii*) i tal og datoer, (*iv*) i angivelser af tidsrum og (*v*) i

[18] Fremover vises eksempler fra PAROLE-korpusset i dette format. Alle SGML-koder er fjernet her, medmindre de er relevante for eksemplet. Evt. fejl i korpusteksterne (slåfejl osv.) er ikke rettet i disse eksempler.

visse forkortelser (som f.eks. *a/s* eller *t/r*). Det har dog været svært for DAN-TWOL-tokeniseren at adskille skråstregen fra dens omgivelser på korrekt vis, og derfor har en stor del af de ordformer, der indeholder skråstregen, fået tildelt en analyse som tekstfejl ('XX'). Dette gælder dog ikke tal, datoer og andre numeriske tidsangivelser (jf. afsnit 5.3.3 om numeralier og afsnit 6.1 om ordformer, der ikke kunne tildeles en analyse).

..Told- og Skattestyrelsen på jagt efter fejl i oplysninger fra arbejdsgivere..
<W lemma="Told-" msd="XX">Told-</W>
..det ville f.eks. være at gennemføre elektrificeringen af Odense-Padborg..
<W lemma="Odense-Padborg" msd="XX">Odense-Padborg</W>
..vandmanden (21. jan.-18. feb.): Der er en noget urolig indflydelse fra stjernerne..
<W lemma="jan.-18." msd="XX">jan.-18.</W>
..det kunne - som i Hitchcocks Vertigo – handle om den ufrie, besiddende og .. dræbende kærlighed..
<W lemma=" - " msd="XP"> - </W>
..en relativ høj forentning af pensionopsparernes depoter i slutningen af 1980'erne/begyndelsen af 1990'erne..
<W lemma="1980'erne/begyndelsen" msd="XX">1980'erne/begyndelsen</W>
...Lise Andreasen kom til Australien for 4 ½ år siden. Hun var landbrugs-udviklingsstudent i Port Campbell..
<W lemma="1/2" msd="AC---U=--">1/2</W>

4.2 Flerordsforbindelser

Som nævnt i afsnit 4 ovenfor er hovedreglen i det morfosyntaktisk taggede tekstkorpus, at ordformer, der omgives af blanktegn og interpunktionstegn, behandles hver for sig, selv hvis de indgår i forskellige former for flerordsforbindelser. Dette udspringer af en generel tilbageholdenhed med at indføre flerordsforbindelser i DAN-TWOLs leksikon, da det kan være problematisk for tokeniseren at identificere og samle flerordsforbindelser på korrekt vis[19]. Der er dog nogle få undtagelser i PAROLE-korpusset: (*i*) de flerordsforbindelser, der fra starten var opført i DAN-TWOLs leksikon som faste forbindelser (og derfor anerkendes af DAN-TWOL-tokeniseren), samt (*ii*) de flerordsforbindelser, der — af forskellige årsager — er blevet samlet i efterredigeringsfasen. Appendiks 8.5 indeholder en komplet fortegnelse over alle flerordsforbindelser i PAROLE-korpusset.

4.2.1 Faste ordforbindelser

Nogle få hyppige flerordsforbindelser samles med en understregning (_) af DAN-TWOL-tokeniseren og tildeles en morfosyntaktisk analyse ved opslag i DAN-TWOLs leksikon. Disse få faste ordforbindelser er i efterrediggeringsfasen desuden blevet suppleret med et par andre faste vendinger (oftest sammensatte adverbier) på baggrund af en undersøgelse af de hyppigst forekommende flerordsforbindelser fra RO96 i DSL's avistekstbibliotek.

..Bortset fra det kosmiske "lys" dyrkes økologisk vin også uden kunstgødning og sprøjtning..
<W lemma="bortset_fra" msd="SP">Bortset_fra</W>
..vi stiller simpelt hen større krav til idrætsarbejdet ude i klubberne..
<W lemma="simpelt_hen" msd="RGU">simpelt_hen</W>

4.2.2 Gruppesammensætninger

I gruppesammensætninger og -afledninger, hvis første led består af mere end et ord, anvendes bindestregen normalt mellem næstsidste og sidste ord (RO96, §63.7). Da det er uhensigtsmæssigt at behandle disse sammensætningers led hver for sig, er de i PAROLE-korpusset blevet samlet vha. en understregning (_) i efterrediggeringsfasen. Alle de

[19] DAN-TWOL-tokeniseren samler enten alle eller ingen forekomster af en flerordsforbindelse. Den kan således f.eks. ikke skelne mellem *alt for* i *alt for mange mennesker* og *alt for* i *han gjorde alt for hende*.

gruppesammensætninger, der optræder som flerordsforbindelser i PAROLE-korpusset, er opført i appendiks 8.5.1.

..mange er opvokset i fast food-generationen med McDonalds og grillbarer..
<W lemma="fast_food-generation" msd="NCCSU==D">fast_food-generationen</W>
..han [fordrejer] også de klassiske litterære syndromer. Eksempelvis: Øde ø-syndromet..
<W lemma="øde_ø-syndrom" msd="NCNSU==D">Øde_ø-syndromet</W>
..den israelske Bar-Lev forsvarslinie ved Suezkanalen i Yom Kippur-krigen i 1973..
<W lemma="Yom_Kippur-krig" msd="NCCSU==D">Yom_Kippur-krigen</W>
..denne roman er nu blevet den første Helle Stangerup-udgivelse på forfatterens nye forlag..
<W lemma="Helle_Stangerup-udgivelse" msd="NCCSU==I">Helle_Stangerup-udgivelse</W>
..vi har ydet økonomisk støtte til Charta 77-folkene i Tjekkoslovakiet o.s.v...
<W lemma="Charta_77-folk" msd="NCNPU==D">Charta_77-folkene</W>
..Mario Andretti, endnu en Formel 1-veteran, der kørte en omgang med en fart af 373 kilometer i timen..
<W lemma="Formel_1-veteran" msd="NCCSU==I">Formel_1-veteran</W>

4.2.3 Fossilerede kasusendelser

Til sidst har vi samlet de "fossilerede" eller stivnede faste forbindelser, der består af en præposition (som f.eks. *af, i, med, på* eller *til*) og en forældet dativ eller genitiv bøjningsform af et substantiv (som f.eks. *i sinde* eller *til gode*). Desuden er tidsadverbialer med præpositionen *i* (som f.eks. *i aftes* eller *i søndags*), der er levn af en ældre tysk adverbialendelse, blevet samlet som flerordsforbindelser. Disse flerordsforbindelser er opført i appendiks 8.5.4.

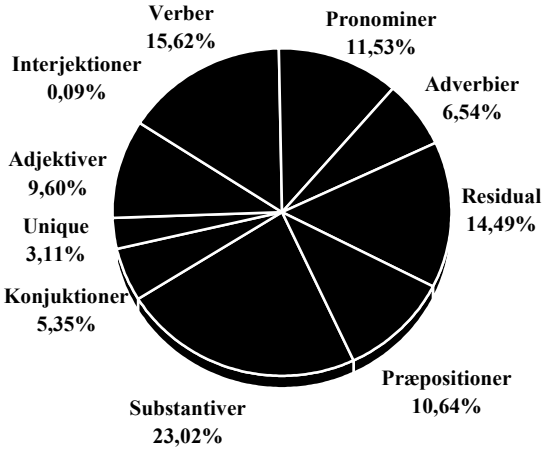
..den 33-årige var meget ilde tilredt, men stadig i live, da han blev smidt ud af vinduet..
<W lemma="i_live" msd="RGU">i_live</W>
..indtil sent i aftes holdt ægteparret deres første møde med deres nye advokat..
<W lemma="i_aftes" msd="RGU">i_aftes</W>

5. Ordklasser

Dette afsnit omhandler de vigtigste lingvistiske beslutninger, der blev truffet af korpustaggerne under korpustaggingen. Det er inddelt efter de forskellige ordklasser, der findes i PAROLE-tagsættet (som vist i **Figur 4** i afsnit 2.3 ovenfor). Hvert kapitel om en ordklasse indledes af et tabeludsnit fra dette PAROLE-tagsæt. Tabeludsnittet viser oftest kun bogstavværdierne for ordklassen ('CatGram') på plads 1 og underinddelingen af ordklassen ('SsCatGram') på plads 2. Bogstavværdier for alle andre pladser findes i oversigten i appendiks 8.1.

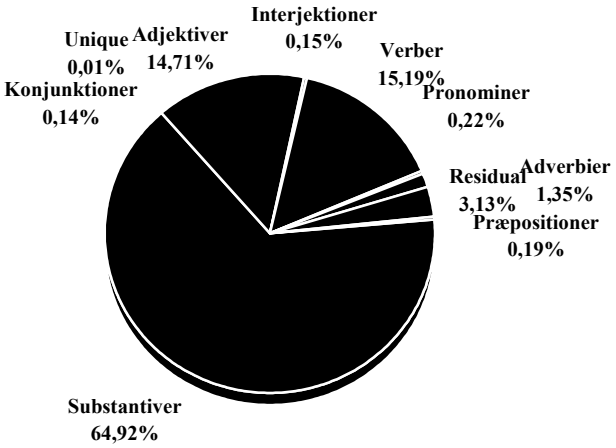
Følgende diagrammer i **Figur 12** viser fordelingen af tekstord ('tokens'), ordtyper ('types') og lemmaformer på de forskellige PAROLE-ordklasser i det taggede korpus. Den første tabel repræsenterer fordelingen af tekstord, sådan som de optræder i korpusteksterne. Her ville f.eks. *Husene, huset, hus, Hus, hus* og *husets* tælle som 6 forskellige forekomster af substantiviske tekstord. Af tabellen fremgår det, at substantiverne er repræsenteret med 66.906 tekstord i alt, mens f.eks. præpositioner forekommer i alt 30.927 gange. Den næste tabel viser fordelingen af grafisk forskellige ordtyper, hvor de grafisk identiske tekstord er slået sammen (og er normaliseret mht. store/små bogstaver, bindestreger og accenttegn osv.). Her ville *Husene, huset, hus, Hus, hus* og *husets* blive talt med som 4 forskellige forekomster af substantiviske ordtyper. Af denne tabel fremgår det, at substantiverne er repræsenteret med 23.990 forskellige ordtyper, mens præpositionerne kun fordeler sig på 71 grafisk forskellige ordtyper. Den sidste tabel repræsenterer fordelingen af lemmaformer i korpusset. Her ville *Husene, huset, hus, Hus, hus* og *husets* tælle som ét eneste substantivisk lemma. Af den sidste tabel fremgår det, at substantiverne er repræsenteret med 18.506 forskellige lemmaformer i PAROLE-korpusset, mens præpositionerne er repræsenteret med 60 forskellige lemmaformer.

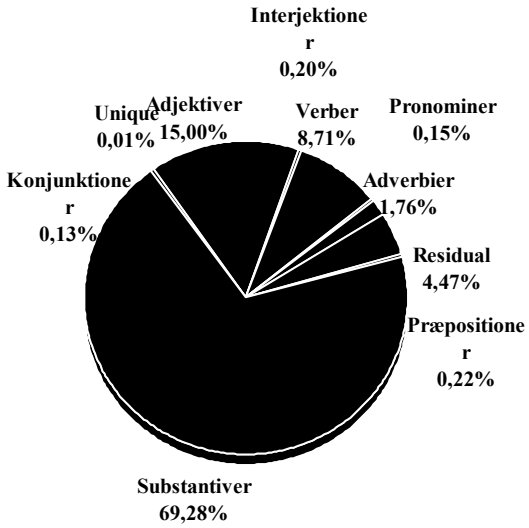
Figur 12: Fordeling af ordklasser i det morfosyntaktisk taggede korpus



Ordklasse	Antal tekstord ('tokens')
Substantiver	66.906
Konjunktioner	15.549
Unique	9.037
Adjektiver	27.900
Interjektioner	259
Verber	45.398
Pronominer	33.493
Adverbier	19.017
Residual	42.114
Præpositioner	30.927
I alt	290.600

Ordklasse	Antal ordtyper ('types')
Substantiver	23.990
Konjunktioner	51
Unique	3
Adjektiver	5.438
Interjektioner	56
Verber	5.612
Pronominer	81
Adverbier	499
Residual	1.157
Præpositioner	71
I alt	36.958





Ordklasse	Antal lemmaformer
Substantiver	18.506
Konjunktioner	49
Unique	3
Adjektiver	4.006
Interjektioner	54
Verber	2.328
Pronominer	41
Adverbier	471
Residual	1.195
Præpositioner	60
I alt	26.713

5.1 Substantiver

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Noun N	Common C	Gender	Number	Case	=	=	Definiteness			
Noun N	Proper P	-	-	Case	=	=	-			

Substantiver (navneord) udgør som bekendt kerneleddet i substantivgruppen; dog kan denne plads stå tom[20] (jf. også afsnit 5.1.7 om "substantivisk" anvendelse). Substantivgruppen kan optræde med forskellige syntaktiske ledfunktioner i sætningen, f.eks. som subjekt, objekt, prædikativ, apposition eller præpositionsobjekt. Substantivkernens adled er enten præled (determinativer eller attributive adjektiver) eller postled (præpositionsforbindelser, relativsætninger eller reducerede relativsætninger). I PAROLE-korpusset er substantiviske kerneled og deres adled dog ikke automatisk samlet som flerordsforbindelser, men derimod er hvert ord, der indgår i substantivgruppen, morfosyntaktisk tagget hvert for sig (jf. afsnit 4.2 om flerordsforbindelser).

Substantiverne som ordklasse inddeles normalt i appellativer (fællesnavne) og proprier (egennavne), og tilsammen udgør de den hyppigst repræsenterede ordklasse i PAROLE-korpusset (med hhv. 52.645/18.960 og 14.261/5.505 tekstord/ordtyper). Der gøres her i brugervejledningen rede for, hvordan der under korpustaggingen er skelnet mellem (*i*) appellativer og proprier, samt mellem (*ii*) substantiver og andre ord, der tilnærmer sig deres funktion. Her kan nævnes f.eks. (*a*) verbalsubstantiver (jf. afsnit 5.2.5 om gerundium), (*b*) participier og adjektiver med "substantivisk" anvendelse, og (*c*) diverse "udenlandske" ord af substantivisk karakter (jf. også afsnit 6.3.1 om ord, der er ukendte for DAN-TWOL).

Mht. valget mellem en analyse som appellativ, proprium eller "udenlandsk" ord for et givent tekstord har tendensen under korpustaggingen været at foretrække en mere informativ analyse (dvs. en analyse, der indeholder flere morfosyntaktiske oplysninger) frem for en mindre informativ analyse. Dette har giver følgende generelle præcedens, som vil blive uddybet i de efterfølgende afsnit:

appellativ > proprium > udenlandsk ord
NC > NP > XF

5.1.1 Appellativ eller proprium?

Ifølge den traditionelle definition på forskellen mellem appellativer og proprier betegner de "beskrivende" appellativer klasser af (mere eller mindre) konkrete eller abstrakte genstande, mens de "benævnende" proprier betegner bestemte individer tilhørende disse klasser (jf. f.eks. Diderichsen, 1968, s. 34-41). Denne semantiske skelnen mellem appellativer og proprier understøttes også i en vis udstrækning af skriftsprogets anvendelse af stort begyndelsesbogstav til proprier (jf. RO96, §12). Dog erkender både Diderichsen, 1968 og RO96, at det i denne forbindelse kan være vanskeligt at afgøre, om et ord faktisk anvendes som proprium eller ej:[21]

> Naar man skriver Appellativer med lille Bogstav (som i Dansk efter Anordning af 1948), er det som Regel Navne (og ikke blot ægte Proprier), der skrives med stort. Men da Begrebet Navne er svært at afgrænse, er det vanskeligt at gennemføre denne Praksis konsekvent. Diderichsen, 1968, s. 34-5

> Proprier skrives med stort begyndelsesbogstav. Denne regel gælder uanset om de pågældende proprier mere eller mindre tydeligt består af ord der også kan bruges som

[20] Jf. f.eks. Diderichsen, 1968, s. 43-5 eller Jensen, 1985, s. 34.
[21] Ifølge Hansen, 1998 fylder reglerne om anvendelse af store og små bogstaver 18% af retskrivningsreglerne i RO96, hvor de kun fyldte 9% i den sidste retskrivningsordbog før reformen i 1948.

> appellativer. I nogle tilfælde er det særlig vanskeligt at afgøre om et ord er brugt som proprium eller ej. RO96, s. 558
>
> Mange ord og ordforbindelser bruges snart som proprier, snart som appellativer.. Det kan ikke altid fastsættes at et bestemt ord i enhver sammenhæng skal skrives med stort eller med lille. RO96, s. 566
>
> Et ord kan godt have propriumskarakter for én sprogbruger eller én gruppe af sprogbrugere uden at have det for andre. RO96, s. 567

At følge anvendelsen af store og små begyndelsesbogstaver til at skelne mellem proprier og appellativer er problematisk på (mindst!) to forskellige måder i korpustagging. For det første varierer anvendelsen af store/små begyndelsesbogstaver en del i korpusteksterne, således at samme individ eller genstand øjensynligt betragtes som enten et appellativ eller et proprium, alt afhængig af forfatterens eget synspunkt[22]:

..Det er begrænset hvad jeg har lov til at referere når statsministeren beder os om at udtale os frit..[23]
..Og Statsministeren satte sig ikke for bordenden som regeringschef, men ved den ene langside som partiformand..
..Mænd er bange for to ting, sagde Lotte Heise i sin kønsrolleprædiken, og det er: aids og tårer..
..På årets Gay & Lesbian Film Festival i København vejer AIDS endnu engang tungt på programmet..
..Arnoldi kritiserer, at ministeriet .. ikke inddrager mennesker med erfaring i, hvad franskmændene ønsker..
..Hvad skal der til, for at Ministeriet kan fungere bedre? "At Ministeriet får en minister, der vælger det af lyst..
..Det er endnu uvist, om enken ønsker at appellere dommen til højesteret..
..sagens udfald er, at de forhold, som revisoren tidligere er dømt for, af Højesteret betragtes som forældede..
..det var Mathias, der var i brugsforeningen for at købe hvidtøl..
..Spurgte de om vej i Brugsforeningen fik de at vide, at jeg vist var flyttet..
..uberørt af den ofte voldsomme trafiklarm i Amaliegade, residerer dronning Margrethes økonomichef..
..I forbindelse med Dronning Margrethe og prins Henrik's sølvbryllup, barslede folketinget med endnu et nyt lovforlig..

For det andet var det ikke muligt for DAN-TWOL-tokeniseren på forhånd at samle flerleddede navne (som f.eks. *Det Konservative Folkeparti, Dansk Sprognævn* eller *Det kgl. Teater*) som substantiviske flerordsforbindelser i korpusteksterne. Ifølge reglen i RO96, §12.2 skal et stort begyndelsesbogstav anvendes i det første ord og i de mere betydningsfulde ord i flerleddede navne. Hvis det første ord er en artikel, er det derimod valgfrit, om det skal skrives med stort eller lille begyndelsesbogstav. Det er dog ikke særligt hensigtsmæssigt at analysere *Det* i *Det konservative Folkeparti* eller i *Det kgl. Teater* som proprier, mens *konservative* og *kgl.* derimod analyseres som adjektiver. Derudover er det også problematisk, hvordan man behandler substantiver med stort begyndelsesbogstav, når de optræder efter et slutpunktum, eller når alle bogstaver i det pågældende substantiv er store, som f.eks. i overskrifter:

..I Det konservative Folkeparti glæder vi os over, at FRPs 16 mandater nu kunne bruges fornuftigt..
..Men også Det Konservative Folkeparti har nu indført central registrering. Socialdemokratiet følger efter i år..
..og .. igen medføre en situation med Det konservative folkeparti uden for indflydelse på sikkerhedspolitikken..
..Arbejdsgiverforeningen og Industrirådet, der især er knyttet til det Konservative Folkeparti..
..Det er en urimelig mistanke at tilsmudse det konservative folkeparti med..
..Derfor blev Dansk Sprognævn bedt om at finde på et mildere ord. Et forslag lød på hospital for lindrende pleje..
..udviklingen af det danske sprog hviler på tre grundpiller: den skrevne presse, radio/tv og Dansk sprognævn..
..min far er belysningsmester på Det kongelige Teater, og min mor er kantinebestyrer..
..at forhandle de sidste småting på plads til operachef-kontrakten med det kongelige teater..
..Her var både figurer, kulisser og bagtæpper og et proscenium inspireret af det daværende kgl. teater..
.."Et vildskud", som måtte "luges ud". Det kostede vort kongelige teater over en halv million kroner..

[22] Jf. også f.eks. RO96, §12.13, 'Særlige problemer', s. 566-9. Af særlig relevans er afsnit §12.13.a om 'proprier eller appellativer?' med eksempler som f.eks. *Højesteret/højesteret, Kommunisterne/kommunisterne, Københavns Havn/Københavns havn* samt afsnit §12.13.b om 'propriumskarakter i snævrere kredse' med eksempler som f.eks. *Museet/museet, Direktionen/direktionen, Ministeren/ministeren.*

[23] Disse eksempelsætninger (og de efterfølgende i dette afsnit) stammer fra DDO's tekstkorpus.

..nationalscenen [vil] kunne sikre .. det vekslende repertoire, der har kendetegnet Det "moderne" Kongelige Teater..
..DET KGL. TEATER er bestandig i krise. Men hvad er årsagen? For få penge og usle lappeløsninger..

En anden måde at skelne mellem appellativer og proprier er at følge deres syntagmatiske relationer, dvs. deres kombinationsmuligheder med andre ordklasser. Appellativer optræder således prototypisk i forbindelse med et eller flere adled, mens proprier normalt optræder som substantivkerner uden adled. Men der findes dog proprier, der syntaktisk set viser en slags overgang til appellativer, f.eks. (*i*) person- eller (geografiske) stednavne, der optræder sammen med et eller flere præled (som f.eks. *det czaristiske Rusland, det højborgerlige København, den vindtørre Søren Kierkegaard*), (*ii*) navne på institutioner, firmaer, produkter m.m., der optræder med et eller flere præled (som f.eks. *en grøn Tuborg, en falmet Berlingske Tidende*) og (*iii*) person- eller stednavne, der optræder med en beskrivende betydning, der minder om appellativernes (som f.eks. *en rigtig Brian, en værre Hitler, et himmelsk Jerusalem*)[24].

I Jensen, 1985, s. 140-2 præsenteres en række mere detaljerede syntaktiske kontekster til at adskille proprier fra tælleligt og utælleligt anvendte appellativer. En lignende morfosyntaktisk distributionel definition fremlægges også i Arndt, 1996, s. 122-3, hvor proprierne defineres som de substantiver, der (kun) kan have genitivsuffiks, og hvis stamform kan forekomme uden adled i en substantivgruppe som subjekt.

På grund af de praktiske problemer, der er forbundet med at skelne mellem appellativer og proprier på en overskuelig og konsekvent måde under korpustaggingen, har vi valgt at anvende en anden, meget enkel morfologisk definition i PAROLE-korpusset. Denne definition, der tilnærmer sig Arndt, 1996 ovenfor, benytter sig først og fremmest af DAN-TWOLs egen tildeling af morfologiske træk. Hvis et substantiv af DAN-TWOL har fået tildelt værdier for de træk, der kendetegner et appellativ (dvs. genus, numerus, kasus og bestemthed), tagges dette substantiv (højst sandsynligt) som et appellativ under korpustaggingen. Hvis substantivet derimod kun kan bøjes i kasus (dvs. genitiv eller ikke-genitiv), tagges substantivet som et proprium. Som det vil fremgå af dette kapitel, findes der dog nogle få undtagelser til denne regel, og disse vil blive gennemgået i de efterfølgende afsnit.

Ifølge denne fremgangsmåde er forskellen mellem appellativer og proprier i høj grad forudbestemt af indholdet af DAN-TWOL-leksikonet. Herved analyseres alle substantiver, der ifølge RO96 kan være appellativer, (næsten) altid som appellativer. Dette er også tilfældet, når de evt. skrives med et stort begyndelsesbogstav, fordi de efter tekstens forfatters mening anvendes som proprier efter den traditionelle definition. Ordklassen 'proprier' omfatter derimod stort set kun de substantiver, der kun kan bøjes i kasus.[25]

Efter den traditionelle definition dækker ordklassen proprier flere forskellige undergrupper: personnavne, geografiske stednavne, karakteriserende navne, navne på offentlige og private institutioner, firmaer, foreninger, sammenslutninger, organisationer, virksomheder, politiske partier, navne på himmellegemer, bygninger, produkter, bøger, film, teaterstykker, malerier, skulpturer, kompositioner, sange, plader, cd'er osv. I PAROLE-korpusset behandles ord tilhørende disse grupper ikke alle automatisk som værende proprier. Her er det oftest kun person- og geografiske stednavne samt nogle firmanavne, produktnavne osv. der får tildelt en analyse som proprium (dvs. netop kun de substantiviske ordformer, der ikke falder sammen med appellativerne). Jf. dog også afsnit 5.1.8 og 5.1.9 om behandlingen af "udenlandske" ord.

[24] Jf. bl.a. Diderichsen, 1968, s. 35.
[25] Denne morfologiske definition på forskellen mellem appellativer og proprier svarer også til den beskrevne fremgangsmåde i Wynne, 1996, som er taggingmanualen til CLAWS 'Constituent Likelihood Automatic Word-tagging System' projektet (jf. afsnit 2.6 'Nouns').

5.1.2 Appellativer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Noun N	Common C	Gender	Number	Case	=	=	Definiteness			

Som vist i tabeludsnittet ovenfor er appellativer (NC) markeret for genus samt bøjning i numerus, kasus (dvs. genitiv eller ikke-genitiv) og bestemthed i PAROLE-korpusset[26]. Her er et appellativ således et substantiv, der i modsætning til et proprium (potentielt) kan tildeles værdier for alle disse morfologiske træk. Denne rent morfologiske definition på appellativer er som nævnt ovenfor en nødvendig følge af, at DAN-TWOL-tokeniseren ikke markerer f.eks. *Det kgl. Teater* som en sammenhængende flerordsforbindelse, men som tre uafhængige ordformer.

.."I må elske musik utroligt meget," råbte Paul Simon til det hårdt prøvede, mudderbrune publikum..
<W lemma="publikum" msd="NCNSU==I">publikum</W>
..MAN må kritisere det kuldslåede fornuftsægteskab med medierne og branchen..
<W lemma="fornuftsægteskab" msd="NCNSU==I">fornuftsægteskab</W>
..to fluers indædte kamp om majsgrødresterne på en uslikket ske og fars hivende åndedræt larmede..
<W lemma="majsgrødrest" msd="NCCPU==D">majsgrødresterne</W>
..i debatten tordnes der løs mod Det kgl. Teaters repertoire..
<W lemma="teater" msd="NCNSG==I">Teaters</W>
..Henrik Andersen og Jan Vedersøe, der leder Klaptræet på Kultorvet, har .. en del navne registreret..
<W lemma="klaptræ" msd="NCNSU==D">Klaptræet</W>

Nogle få (mere eller mindre) faste ordforbindelser, der ikke eksplicit samles som flerordsforbindelser i korpusset, indeholder appellativer, hvor det stort set er umuligt at afgøre, om appellativet er bøjet i singularis eller pluralis, når disse to bøjningsformer falder sammen. I disse tilfælde har ordformen altid fået tildelt en analyse som appellativ i singularis:

..dette sikres ved, at modstanden totalt og i de enkelte organer kan ændres efter behov..
<W lemma="behov" msd="NCNSU==I">behov</W>
..man er i færd med at begå vold mod alle de principper om menneskerettigheder..
<W lemma="færd" msd="NCCSU==I">færd</W>
..Illum's desperate situation fremgår af underskuddets størrelse i forhold til egenkapital..
<W lemma="forhold" msd="NCNSU==I">forhold</W>
..men Annelise Monsen er ikke i tvivl om, at erhvervslivet nok skal tage Jim Leonards ideer til sig..
<W lemma="tvivl" msd="NCCSU==I">tvivl</W>

En del danske person- og geografiske stednavne falder sammen med (teoretisk mulige) appellativer. Når disse substantiver betegner en bestemt person eller et bestemt sted, har vi valgt analysen som proprium frem for en analyse som appellativ (jf. dog også afsnit 5.1.5 om substantiviske sammensætninger).

..det kan kun give en masse rutine," siger amatørspilleren Thomas Bjørn..
<W lemma="Bjørn" msd="NP--U==-">Bjørn</W>
..jeg har gået til jazzballet i mange år," siger Helle, der til hverdag leder Studio "Better Bodies"..
<W lemma="Helle" msd="NP--U==-">Helle</W>
..sammen med Blåbjerg Kommunes borgmester .. tager Arne Toft .. til Christiansborg..
<W lemma="Blåbjerg" msd="NP--U==-">Blåbjerg</W>
..Henrik Andersen og Jan Vedersøe, der leder Klaptræet på Kultorvet..
<W lemma="Kultorvet" msd="NP--U==-">Kultorvet</W>

[26] Bogstavværdierne for disse (og alle andre) morfosyntaktiske træk findes også i tabellen i appendiks 8.2.

I et par tilfælde er appellativer (eller adjektiver) anvendt som øgenavn, kælenavne eller tilnavne på personer i korpusset, og de skrives i denne forbindelse derfor med et stort begyndelsesbogstav (jf. RO96, §12.3). I disse tilfælde beholdes dog analysen som appellativ.

..hun er en smatso sagde Bolle og spyttede, men opdagede at hans nye spidse sko var blevet snavsede..
<W lemma="bolle" msd="NCCSU==I">Bolle</W>
..nu skal de spille Tvebak og Jonathan i dramatiseringen af Astrid Lindgrens eventyr..
<W lemma="tvebak" msd="NCCSU==I">Tvebak</W>

5.1.3 Proprier

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Noun N	Proper P	-	-	Case	=	=	-			

Som det fremgår af tabeludsnittet ovenfor, er et proprium (NP) et substantiv, der kun kan bøjes i kasus (genitiv eller ikke-genitiv) i PAROLE-korpusset. Denne definition dækker dog en stor del af de ord, der også traditionelt opfattes som værende proprier.

..resterne af Jyske Division skal indgå i NATOs Hovedforsvarstyrke, hedder det i rapporten..
<W lemma="division" msd="NCCSU==I">Division</W>
<W lemma="hovedforsvarsstyrke" msd="NCCSU==I">Hovedforsvarsstyrke</W>
..samt tre af de nye Standardflex 300 glasfiber-patruljefartøjer "Havkatten", "Hajen" og "Makrellen"..
<W lemma="havkat" msd="NCCSU==D">Havkatten</W>
<W lemma="haj" msd="NCCSU==D">Hajen</W>
<W lemma="makrel" msd="NCCSU==D">Makrellen</W>
..en anden blomsterhandler, Carsten Ove Dalsø Nielsen fra Carstens Blomster,..
<W lemma="Carsten" msd="NP--U==-">Carsten</W> <W lemma="Ove" msd="NP--U==-">Ove</W>
<W lemma="Dalsø" msd="NP--U==-">Dalsø</W> <W lemma="Nielsen" msd="NP--U==-">Nielsen</W> ..
<W lemma="Carsten" msd="NP--G==-">Carstens</W> < lemma="blomst" msd="NCCPU==I">Blomster</W>
..har Dronningen aldrig anvendt til privat forbrug. Pengene er altid .. overført til Margrethe og Henriks Fond..
<W lemma="dronning" msd="NCCSU==D">Dronningen</W>
<W lemma="Margrethe" msd="NP--U==-">Margrethe</W>
<W lemma="Henrik" msd="NP--G==-">Henriks</W>
<W lemma="fond" msd="NC[CN]SU==I">Fond</W>
..FC København spiller af samme årsag i Parken fordi man er forpligtiget overfor sine sponsorer..
<W lemma="FC" msd="NP--U==-">FC</W>
<W lemma="København" msd="NP--U==-">København</W>
<W lemma="park" msd="NCCSU==D">Parken</W>
..finansminister Henning Dyremose måtte samtidig bøje sig for Socialdemokratiet..
<W lemma="Henning" msd="NP--U==-">Henning</W>
<W lemma="Dyremose" msd="NP--U==-">Dyremose</W>
<W lemma="socialdemokrati" msd="NCNSU==D">Socialdemokratiet</W>

Da denne definition på forskellen mellem appellativer og proprier ikke er afhængig af, om ordet skrives med et stort eller lille begyndelsesbogstav, gælder det, at proprier i PAROLE-korpusset altid er skrevet med et stort begyndelsesbogstav, mens substantiver med et stort begyndelsesbogstav ikke altid er proprier. Nogle få udenlandske person- og geografiske stednavne skrives dog normalt med lille begyndelsesbogstav, og er således de eneste undtagelser til ovennævnte regel: *al, al-Assad, bin, d'Estaing, d'Estaings, de, eks-Jugoslavien, van* og *von*.[27] Andre person- eller geografiske stednavne, der er skrevet med lille begyndelsesbogstav, markeres derimod som tekstfejl (msd="XX") i PAROLE-korpusset.

..og har den saudiske generalløjtnant prins Khalid bin Sultan som øverstbefalende..
<W lemma="bin" msd="NP--U==-">bin</W>

[27] Desuden har sammensætningen *fodbold-EM* også fået tildelt en analyse som proprium (mere om initialforkortelser i afsnit 5.1.6).

..det er en af lederen Giscard d'Estaings egne folk, den tidligere udenrigsminister..
<W lemma="d'Estaing" msd="NP--G==-">d'Estaings</W>
..året 1716-1717 var et godt år for vivaldi her udkom der hele 3 opuser..
<W lemma="vivaldi" msd="XX">vivaldi</W>
..i aften er der dog stadig langt til egyptens smægtende rytmer i Kanonteltet..
<W lemma="egyptens" msd="XX">egyptens</W>

5.1.4 Proprier, der er bøjet i andet end kasus

I PAROLE-korpusset er definitionen på proprier netop, at de som substantiver kun kan optræde med genitivsuffiks. Dog kan nogle få "vaskeægte" proprier — typisk geografiske stednavne som f.eks. *Færøerne* eller *Elfenbenskysten* — alligevel opfattes som havende en anden bøjningsform end den "ubøjede" (dvs. singularis, ubestemte) form[28]. Denne gruppe af proprier har dog stadig fået tildelt analyser som proprier i PAROLE-korpusset. Et par andre proprier (typisk produktnavne) optræder desuden i bestemt og/eller pluralis bøjningsform i PAROLE-korpusset. Disse ordformer har normalt fået tildelt analyser som appellativer på baggrund af deres bøjningsform i korpusteksten.

.."Men Færøerne tilhører dog det danske rigsfælleskab," siger Erik Hjorth Nielsen..
<W lemma="Færøerne" msd="NP--U==-">Færøerne</W>
..på vej mod Bregenz glæder vi os endnu til turen til Alperne..
<W lemma="Alperne" msd="NP--U==-">Alperne</W>
..han havde kurs direkte mod Fiat'en, og jeg så føreren af Fiat'en vride en gang i rattet..
<W lemma="Fiat" msd="NCCSU==D">Fiat'en</W>
..de sidste par nætter har nazisterne brugt parkerede Trabanter til at barrikadere gaderne..
<W lemma="Trabant" msd="NCCPU==I">Trabanter</W>

5.1.5 Substantiviske sammensætninger

De substantiviske sammensætninger, som ikke findes i RO96 og derfor heller ikke findes i DAN-TWOL-leksikonet, får af DAN-TWOL-analyseapparatet automatisk tildelt samme analyse som det sidste sammensætningsled, da overleddet normalt regnes for at være kernen i sammensætningen. Sammensatte appellativer er hyppige i dansk og er oftest sammensat af (*i*) to eller flere appellativer, (*ii*) hhv. et proprium og et appellativ eller (*iii*) hhv. et andet ord og et appellativ. Sammensatte substantiver, hvor det sidste led er et proprium, er derimod sjældne og er tagget som proprier. Bindestreger mellem sammensætningsledene i substantiviske sammensætninger er altid blevet accepteret i PAROLE-korpusset (jf. afsnit 4.1.3 om bindestregen).

..samt tre af de nye Standardflex 300 glasfiber-patruljefartøjer "Havkatten", "Hajen" og "Makrellen"..
<W lemma="glasfiber-patruljefartøj" msd="NCNPU==I">glasfiber-patruljefartøjer</W>
..pisk 2 1/2 del fløde 13 sammen med 1 1/2 dl pskefløde [sic], 2 tsk. lys Dijonsennep, friskrevet muskat..
<W lemma="Dijonsennep" msd="NCCSU==I">Dijonsennep</W>
..de to Næstved-skoler har i tidens løb haft en række udlandsaktiviteter..
<W lemma="Næstved-skole" msd="NCCPU==I">Næstved-skoler</W>
..en kølig brise blæser igennem byen," rapporterede CNN-journalisterne fra Bagdad..
<W lemma="CNN-journalist" msd="NCCPU==D">CNN-journalisterne</W>
..Mini-miniput-spilleren fra Thorning skal nok komme til at spille fodbold igen..
<W lemma="mini-miniput-spiller" msd="NCCSU==D">Mini-miniput-spilleren</W>
..men selvom livet aldrig blev det økonomiske "bummelumliv", han havde drømt om..
<W lemma="bummelumliv" msd="NCNSU==I">bummelumliv</W>
..jeg nikkede; men havde ondt af Astma-Bodil. Hun var slet ikke sådan som hun foregav at være..
<W lemma="Astma-Bodil" msd="NP--U==-">Astma-Bodil</W>

[28] Andre eksempler i PAROLE-korpusset er: *Alperne, Atlanten, Atlanterhavet, Falklandsøerne, Fjernøsten, Gazastriben, Golfen, Ildlandet, Isefjorden, Limfjorden, Mellemøsten, Midtvesten, Mongoliet, Norden, Nordpolen, Stillehavet, Suezkanalen, Sydpolen, Vestbredden, Vesterhavet, Østen* og *Østersøen.* Jf. også RO96, §12.10.a og §12.10.b.

Sammensatte danske personnavne (typisk efternavne) er altid blevet analyseret som proprier, og det samme gælder sammensatte danske geografiske stednavne (typisk bynavne og gadenavne). Overgangen fra sammensat stednavn til sammensat appellativ er dog svær at definere helt præcist, og derfor er de sammensatte stednavne, der har fået tildelt en analyse som proprium, begrænset til kun at omfatte bynavne, gade- og vejnavne, navne på torve, pladser, parker, haver, bjerge, dale, bydele, broer, tunneller osv., også selvom de f.eks. er bøjet i bestemt form[29]. Nogle eksempler på dette fra PAROLE-korpusset er: *Fælledparken, Fritidshaven, Guldbrandsdalen, Halmtorvet, Hareskoven, Holbækmotorvejen, Kvægtorvet, Lillebæltsbroen, Limfjordtunnelen, Lyngbyvejen, Rådhuspladsen, Sjællandsbroen, Sommerfugledalen, Sydhavnen* samt *Ørestaden*. Et par appellativer og adjektiver, der optræder i flerleddede danske geografiske stednavne, har også fået tildelt en analyse som proprium i PAROLE-korpusset: f.eks. *City, Fjord, Gade, Gammel (Gl.), Have, Kongens (Kgs.), Nr., Park, Plads, Sct. (Skt.), Skov, Sønder (Søndre), Torv, Vej* osv.[30]

..Henrik Andersen og Jan Vedersøe, der leder Klaptræet på Kultorvet, har .. en del navne registreret..
<W lemma="Kultorvet" msd="NP--U==-">Kultorvet</W>
..der var langt fra Gentofte, hvor jeg boede, og ind til Gammel Torv..
<W lemma="Gammel" msd="NP--U==-">Gammel</W> <W lemma="Torv" msd="NP--U==-">Torv</W>
..skriv til Månedsmagasinet BILEN, "Brevkassen", Strandboulevarden 130, 2100 København Ø..
<W lemma="Strandboulevarden" msd="NP--U==-">Strandboulevarden</W>
..han kørte 151 km/t ud ad Holbækmotorvejen - og det kostede ham 1600 kr...
<W lemma="Holbækmotorvejen" msd="NP--U==-">Holbækmotorvejen</W>

5.1.6 Initialforkortelser

RO96 indeholder en del substantiviske initialforkortelser (akronymer) samt retskrivningsregler for anvendelsen af store og små bogstaver, når disse forkortelser anvendes som hhv. proprier og appellativer (RO96, §14.2 & §14.3). Heraf fremgår det, at de fleste initialforkortelser skal skrives med store bogstaver, dog kan almindelige initialforkortelser, der er appellativer, også skrives med små bogstaver (som f.eks. *edb, aids, cd* osv.) Behandlingen af initialforkortelser i PAROLE-korpusset har også været afhængig af DAN-TWOL-analyseapparatet. Her analyseres initialforkortelser med store bogstaver oftest som værende proprier, medmindre forkortelsen er angivet som et appellativ i RO96. Hvis en initialforkortelse ifølge RO96 kan skrives med små bogstaver, har den fået tildelt en analyse som appellativ i PAROLE-korpusset, også hvis den faktisk er skrevet med store bogstaver (som f.eks. *AIDS*). Tendensen i PAROLE-korpusset har dog generelt været at behandle initialforkortelser skrevet med store bogstaver som proprier (jf. også afsnit 5.8.1 om andre forkortelser).

..satspuljen stammer fra et tidligere forlig om overførselsindkomster, hvor også SF deltog..
<W lemma="SF" msd="NP--U==-">SF</W>
..resterne af Jyske Division skal indgå i NATOs Hovedforsvarstyrke, hedder det i rapporten..
<W lemma="NATO" msd="NP--G==-">NATOs</W>
..det første, som han fik som fjervægter i 1990 med guld ved JM, DM og NM..
<W lemma="JM" msd="NP--U==-">JM</W>
<W lemma="DM" msd="NP--U==-">DM</W>
<W lemma="NM" msd="NP--U==-">NM</W>
..han har et kontraktudkast fra FC Köln liggende til overvejelse, og i morgen flyver han ned .. til klubben..
<W lemma="FC" msd="NP--U==-">FC</W>
..fem CD'er med ikke mindre en 140 numre indpakket med en 48 siders booklet med over 100 sjældne Elvis-fotos..
<W lemma="CD" msd="NCCPU==I">CD'er</W>

[29] Her afviger vi således lidt fra forskellen mellem sammensatte proprier og appellativer i RO96, §12.10.a og RO96, §12.10.b, idet vi har begrænset de sammensatte proprier til geografiske stednavne og derfor ville analysere f.eks. *Atlantpagten* og *Europaskolen* som appellativer og ikke som proprier (som i RO96, §12.10.a).
[30] I Wynne, 1996, afsnit 2.6.2.3 forefindes en tilsvarende liste med undtagelser ('NNL1/NNL2').

..hans seneste cd "One Day Spent" er indspillet i USA med fornemt akkompagnement ..
<W lemma="cd" msd="NCCSU==I">cd</W>
..mottoet er: "Ka' du læse billeder, ka' du også lære EDB."..
<W lemma="edb" msd="NCCSU==I">EDB</W>
..konkurrencen kommer af, at flere vil udvikle ny teknologi, informationssystemer, edb,..
<W lemma="edb" msd="NCCSU==I">edb</W>

5.1.7 "Substantivisk" anvendelse

I dansk kan det forekomme, at kerneleddets plads i en substantivgruppe står tom, og nogle grammatikker omtaler dette som "substantivisk" anvendelse af det sidste attributive præled i substantivgruppen (Diderichsen, 1968, s. 44, 47-48, 68), (Jensen, 1985, s. 33-34), (Arndt, 1996, s. 125), (ACG, 1996, s. 100-103, 283). I Jensen, 1985 påpeges det dog, at der ikke nødvendigvis er tale om, at præleddet (et adjektiv, numeralie eller participium) er blevet til et substantiv, da disse præled ikke overtager substantivets bøjning i numerus og bestemthed[31]. Selvom PAROLE-tagsættet for adjektiver, numeralier og participier indeholder mulighed for markering af bøjning i kasus (dvs. genitiv eller ikke-genitiv), markeres disse ord ikke eksplicit med "substantivisk" anvendelse i PAROLE-korpusset. Hvis et adjektiv, numeralie eller participium er markeret for 'genitiv' kasus ('G'), kan det naturligvis antages, at ordet er brugt "substantivisk", men hvis det er umarkeret for kasus ('U'), kan intet udledes om ordets evt. "substantiviske" anvendelse.

I PAROLE-korpusset tildeles disse præled de ordklasser, der angives i RO96. Således vil "substantivisk" anvendte adjektiver og numeralier (som f.eks. *de gamle, en arbejdsløs* og *den tredje*) stadig blive analyseret som værende adjektiver og numeralier. Ligeledes vil "substantivisk" anvendte participier (som f.eks. *den rejsende, de medvirkende, en prostitueret* og *de delegerede*) stadig blive analyseret som værende participier (som i RO96[32]).

..den misundelige gnaver på andre, men sårer sig selv..
<W lemma="misundelig" msd="ANP[CN]SU=DU">misundelige</W>
..det var indtil denne uges meninghedsrådsvalg. Nu står det syv-seks. Og det er altså i de helliges disfavør..
<W lemma="hellig" msd="ANP[CN]PG=[DI]-">helliges</W>
..den dag, Sansolo talte som den tolvte i ørkenen, slap vandet op. Kamelerne gav mindre og mindre mælk..
<W lemma="tolvte" msd="AO---U=--">tolvte</W>
..en journalist, der bevægede sig igennem ruinerne i Erzincan efter overlevende..
<W lemma="overleve" msd="VAPR=[SP][CN][DI]A-U">overlevende</W>
.."Jeg har hele tiden kuldegysninger ned ad ryggen. Jeg frygter, der er mange dræbte," sagde en talsmand..
<W lemma="dræbe" msd="VAPA=P[CN][DI]A-U">dræbte</W>

5.1.8 Udenlandske ord

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Foreign F									

Selvom PAROLE-teksterne naturligvis er danske, indeholder de mange ordformer af udenlandsk oprindelse — ordformer som ikke findes i DAN-TWOL-leksikonet. Med betegnelsen "udenlandsk" ord menes ikke de låneord eller fremmedord, der er (mere eller mindre) accepterede i dansk sprogbrug, der kan slås op i RO96 eller evt. en fremmedordbog, og der bliver anvendt syntaktisk korrekt i PAROLE-korpusteksterne, som f.eks. *at booke, at heade, at sample, at taxie, cool, kingsize, macho, offshore, en cheddar, en D-mark, en fight,*

[31] Jensen, 1985, s. 34 tilkendegiver dog også, at det ikke kan afgøres om adjektiverne og participierne bliver bøjet i numerus, fordi bestemthedsbøjningen (*-e*) ikke kan holdes ude fra pluralisendelsen (*-e*). Dog bliver de ubøjelige numeralier ikke mere "bøjelige" af at blive "substantiveret".
[32] Dog kan de evt. være markeret for "adjektivisk" anvendelse ifølge reglerne for transkategorisering af participier i afsnit 5.2.6.1.

defaitisme osv. Hvis sådan et låneord forekommer relativt hyppigt i DDO's tekstkorpus, vil det derimod have fået tildelt en "rigtig" dansk morfosyntaktisk analyse i efterrediggeringsfasen (jf. afsnit 6.3.1 om "ukendte" ord).

Med betegnelsen "udenlandsk" ord menes de ord af udenlandsk oprindelse, der ikke forekommer i DAN-TWOLs leksikon og ikke kan regnes for at være (mere eller mindre) accepterede låneord i det danske sprog. De udenlandske ord, der findes i PAROLE-korpusteksterne, kan inddeles i tre hovedgrupper: (*i*) udenlandske ord, der optræder i udenlandske citater (*Havde Goethe ikke sagt: Dort wo du nicht bist, ist das Glück.*), (*ii*) udenlandske ord, der er navne på danske eller udenlandske institutioner, firmaer, produkter eller (musik)grupper (som f.eks. *Accumulator Invest, Allied-Lion, Cosmopolitan, Coca-Cola, Danisco, Financial Times, Holiday Inn, Kronfågel, Macintosh, Mucomyst, Panodil, Standardflex, Toyota, WordPerfect, Status Quo, Rolling Stones, Nick Cave and the Bad Seeds)* og (*iii*) udenlandske ord, der indgår i udenlandske bog-, plade-, sang- eller filmtitler osv. (som f.eks. *The Natural History of Alcoholism*, *Bridge over Troubled Water, Star Wars, Monsieur Hire*). Langt de fleste "ukendte" udenlandske ord i PAROLE-korpusset tilhører gruppe (*ii*) ovenfor. Følgende afsnit gennemgår behandlingen af disse udenlandske ord i PAROLE-korpusset.

5.1.9 Udenlandsk ord eller (dansk) substantiv?

Beslutningen om, hvordan disse udenlandske ord skulle behandles i PAROLE-korpusset, har været afhængig af to vigtige hensyn: (*i*) vi ønskede ikke at skulle vælge mellem en analyse som dansk proprium og en analyse som udenlandsk ord (er f.eks. *Coloplast* eller *Accumulator Invest* danske proprier eller udenlandske ord?)[33], og (*ii*) vi ville som regel foretrække en morfosyntaktisk analyse som proprium frem for en analyse som udenlandsk ord, da den første analyse er mere informativ. På baggrund af en undersøgelse af de ordformer i PAROLE-korpusset, der tilhører ovenstående tre grupper af udenlandske ord, blev det besluttet at analysere disse ordformer på to forskellige måder. Hvis det udenlandske ord er skrevet med et stort begyndelsesbogstav (langt de fleste tilfælde), var det mest sandsynligt et proprium og fik derfor tildelt en analyse som et (dansk) proprium:

..I går blev Kosan Teknova solgt fra, og tilbage i Kosan Holding er kun Crisplant..
<W lemma="Kosan" msd="NP--U==-">Kosan</W> <W lemma="Teknova" msd="NP--U==-">Teknova</W>
<W lemma="Kosan" msd="NP--U==-">Kosan</W> <W lemma="Holding" msd="NP--U==-">Holding</W>
<W lemma="Crisplant" msd="NP--U==-">Crisplant</W>
..en af de begejstrede kommentarer kommer fra Kvällspostens sportskommentator..
<W lemma="Kvällsposten" msd="NP--G==-">Kvällspostens</W>
..Kræftens Bekæmpelse synes at opfatte bidragyderne .. som en Coca-Cola frabrikant betragter forbrugerne..
<W lemma="Coca-Cola" msd="NP--U==-">Coca-Cola</W>
..dengang Martin og mig drak os fulde i Martini for at undersøge, hvor meget Martin der var i..
<W lemma="Martini" msd="NP--U==-">Martini</W>

Hvis ordformen derimod er skrevet med et lille begyndelsesbogstav, var dens ordklasse sværere at udlede, og den fik derfor tildelt en analyse som udenlandsk ord ('XF'):

..en aftale med .. en nigeriansk forretningsmand om at oplagre gifttønder for 500 naira (650 kr.) om måneden..
<W lemma="naira" msd="XF">naira</W>
..en salat tilberedt af fiskerogn, tacik (yoghurt med agurk, gulerod og hvidløg), dolma (marinerede vinblade)..
<W lemma="tacik" msd="XF">tacik</W>
<W lemma="dolma" msd="XF">dolma</W>
..HOPKINS er ikke kun ond og grusom. Han kan også spille the good guy..

[33] De samme overvejelser nævnes også i Ejerhed et al, 1992, der er brugervejledningen til det svenske Stockholm-Umeå korpus.

<W lemma="the" msd="XF">the</W> <W lemma="good" msd="XF">good</W>
<W lemma="guy" msd="XF">guy</W>

..tilbage til den lille hyggelige østriske alpeby med gemütlichkeit, store værelser og ditto senge..

<W lemma="gemütligchkeit" msd="XF">gemütligchkeit</W>

..fra den dag tilhører han Erlanders "pojkar". 1[sic] 1953 bliver han Erlanders sekretær. Først i sin fritid..

<W lemma="pojkar" msd="XF">pojkar</W>

..mozarellaost skal laves af bøffelmælk. Efterligninger lavet af komælk skal benævnes fior de latte..

<W lemma="fior" msd="XF">fior</W>
<W lemma="di" msd="XF">di</W>
<W lemma="latte" msd="XF">latte</W>

..de kunne ikke .. betegnes som "associate professors", fordi "associate" betyder en midlertidig ansættelse..

<W lemma="associate" msd="XF">associate</W>
<W lemma="professors" msd="XF">professors</W

Denne forholdsvis enkle løsning er naturligvis stadig utilstrækkelig til de (forholdsvis få) udenlandske citater, sangtitler osv., der tilhører grupperne (*i*) og (*iii*) ovenfor, hvor nogle ord skrives med et stort begyndelsesbogstav og andre med et lille begyndelsesbogstav. Udenlandske citater vil dog altid udgøre et problem for korpustagging, og vi har ment, at fordelene ved denne forholdsvis enkle løsning vejer tungere end denne ulempe (især også hvis man vil træne en automatisk korpustagger).

..vi fik "The Boxer", "Cecelia", "Sound of Silence" i smukke versioner, mens mørket sænkede sig..

<W lemma="Sound" msd="NP--U==-">Sound</W> <W lemma="of" msd="XF">of</W>
<W lemma="Silence" msd="NP--U==-">Silence</W>

..Dead Famous People, The House of Love og Nick Cave And The Bad Seeds samt solister som Lloyd Cole..

<W lemma="The" msd="NP--U==-">The</W> <W lemma="House" msd="NP--U==-">House</W>
<W lemma="of" msd="XF">of</W> <W lemma="Love" msd="NP--U==-">Love</W>

Til sidst skal det nævnes, at en "rigtig" dansk morfosyntaktisk analyse under korpustaggingen kan være blevet tilsidesat af hensyn til ordformens "udenlandske" omgivelser. Dette er tilfældet, når et udenlandsk ord, der optræder sammen med andre udenlandske ord, i formen falder sammen med et dansk ord. For eksempel har både *Bridge* og *Over* i sangtitlen *Bridge Over Troubled Water* fået tildelt analyser som proprier på lige fod med *Troubled* og *Water*, selvom disse to ordformer også eksisterer i dansk (dog evt. med en anden ordklasse og/eller betydning).

..disputatsen, som kommer i bogform fra Royal Botanic Gardens..

<W lemma="Royal" msd="NP--U==-">Royal</W>
<W lemma="Botanic" msd="NP--U==-">Botanic</W>
<W lemma="Gardens" msd="NP--U==-">Gardens</W>

..Klaptræet har tidligere også rummet tre biografer - og "Copenhagen Jazz Festival"..

<W lemma="Copenhagen" msd="NP--U==-">Copenhagen</W>
<W lemma="Jazz" msd="NP--U==-">Jazz</W>
<W lemma="Festival" msd="NP--U==-">Festival</W>

..Malin Adelsgård fra Jysk Bäddlager, navnet på Jysk Sengetøjslagers svenske butikker..

<W lemma="Jysk" msd="NP--U==-">Jysk</W>
<W lemma="Bäddlager" msd="NP--U==-">Bäddlager</W>

..de rå rock'n'rollere, som "Tumbleweed", "Love In the Big Town", og "Lies And Rumours"..

<W lemma="Lies" msd="NP--U==-">Lies</W>
<W lemma="And" msd="NP--U==-">And</W>
<W lemma="Rumours" msd="NP--U==-">Rumours</W>

5.2 Verber

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Mood	Tense	=	Number	Gender	Definiteness	TrCat	Voice	Case
Verb V	Medial E	Mood	Tense	=	Number	Gender	Definiteness	TrCat	Voice	Case

I PAROLE-korpusset er alle verber (udsagnsord) inddelt efter de to SsCatGram-typer 'Main' ('A') og 'Medial' ('E'), hvilke omfatter hhv. "almindelige" verber og deponente/reciprokke verber. Verberne er desuden (lidt utraditionelt) inddelt efter modus ('Mood') i følgende verbalformer: indikativ, infinitiv, imperativ, gerundium og participium. Anvendelsen af visse verber som modalverber (mådesudsagnsord) eller hjælpeverber (hjælpeudsagnsord) markeres ikke eksplicit i PAROLE-korpusset. Desuden markeres den konjunktive (optative) verbalform heller ikke eksplicit, da denne falder sammen med infinitivformen.

Som vist i tabeludsnittet ovenfor markeres verberne for bøjning i tempus, numerus, genus, bestemthed, transkategorisering, diatese samt kasus. Det er dog langt fra alle verbalformer, der markeres for alle disse træk, da f.eks. tilstedeværelsen af de nominale træk (numerus, genus, bestemthed og kasus) i ovenstående liste alene skyldes participiernes mulighed for markering for "adjektivisk anvendelse" i korpusteksterne (jf. afsnit 5.2.6.1 om transkategorisering af participier).

Da DAN-TWOL-tokeniseren ikke samler flerleddede verbaler (som f.eks. sammensatte verballed i verbets omskrevne former, partikelverber og refleksive verber), har de individuelle ordformer, der indgår i disse verbaler, fået tildelt analyser uafhængigt af hinandens tilstedeværelse i sætningen.

..Moskva har haft franske statsoverhoveders .. interesse, fordi Frankrig derved har kunnet opnå en særbehandling..
<W lemma="have" msd="VADR=----A-">har</W>
<W lemma="kunne" msd="VAPA=S[CN]I[ARU]-U">kunnet</W>
<W lemma="opnå" msd="VAF-=----A-">opnå</W>
..slagteriarbejderen slæbte hende hen ad asfalten, men det lykkedes hende at vride sig fri og stikke af..
<W lemma="stikke" msd="VAF-=----A-">stikke</W>
<W lemma="af" msd="RGU">af</W>
..de lo alle sammen. Så skyndte de sig at samle tallerkenerne sammen, så de kunne komme af sted..
<W lemma="skynde" msd="VADA=----A-">skyndte</W>
<W lemma="sig" msd="PP3[CN][SP]U-YU">sig</W>

5.2.1 Mediale verber

Deponente og reciprokke verber har en passiv form (dvs. en *-s* endelse), men en aktiv betydning (jf. f.eks. Allan et al, 1996, s. 310-313). Følgende verbalformer har fået tildelt en analyse som medialt verbum i PAROLE-korpusset: *brydes, enedes, enes, fandtes, findes, færdes, følges, kævles, lykkedes, lykkes, længes, længtes, mindes, mislykkedes, mødes, mødtes, nøjedes, nøjes, omgås, ses, skyldes, skyldtes, skændes, skændtes, sloges, slås, stortrives, synes, syntes, trivedes* samt *trives*.

..det var utroligt at se, hvordan kamelerne trivedes - først på saftigt græs og senere på tørre, tornede buske..
<W lemma="trives" msd="VEDA=----A-">trivedes</W>
..en vigtig psykologisk triumf - især hvis det lykkes ham at ramme væsentlige mål i Israel..
<W lemma="lykkes" msd="VEDR=----A-">lykkes</W>
..han .. lærte, at man skal omgås materiellet med en vis forsigtighed for at spare bilen..
<W lemma="omgås" msd="VEF-=----A-">omgås</W>
..det er lykkedes mig nok en gang at lave dit ansigt om til jern..
<W lemma="lykkes" msd="VEPA=[SP][CN][DI][ARV]-U">lykkedes</W>

5.2.2 Indikativformerne

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Indicative D	Present R	=	-	-	-	-	Active A	-
Verb V	Main A	Indicative D	Present R	=	-	-	-	-	Passive P	-

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Indicative D	Past A	=	-	-	-	-	Active A	-
Verb V	Main A	Indicative D	Past A	=	-	-	-	-	Passive P	-
Verb V	Medial E	Indicative D	Present R	=	-	-	-	-	Active A	-
Verb V	Medial E	Indicative D	Past A	=	-	-	-	-	Active A	-

Som vist i tabeludsnittet ovenfor markeres de to indikativformer — præsens indikativ og præteritum indikativ — for bøjning i tempus og diatese i PAROLE-korpusset.

..eller hypnotiserer med løjerlige arytmiske, dub-pulserende basgange..
<W lemma="hypnotisere" msd="VADR=----A-">hypnotiserer</W>
..forudsætningen for Produktregistret er, at oplysningerne hemmeligstemples..
<W lemma="hemmeligstemple" msd="VADR=-----P-">hemmeligstemples</W>
..måske fordi lakkjolen var så stram, at den knirkede, når hun gestikulerede for meget..
<W lemma="gestikulere" msd="VADA=----A-">gestikulerede</W>
..den blå røg fra knallerterne spredtes, og drønet fra bollespisernes motorer fortabte sig i det fjerne..
<W lemma="sprede" msd="VADA=----P-">spredtes</W>
..der er et voksent publikum, som længes tilbage til biografernes illustrerede drømme..
<W lemma="længes" msd="VEDR=----A-">længes</W>
..skillevæggene var så tynde, at jeg tydeligt kunne høre, når de skændtes inde ved siden af..
<W lemma="skændes" msd="VEDA=----A-">skændtes</W>

5.2.3 Imperativformen

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Imperative M	-	=	-	-	-	-	-	-

Den tredje finitte vebalform — imperativformen (bydeformen) — har fået tildelt en analyse som vist i eksemplet nedenfor:

..spring rundt og dans til radioen, køb en børnepose hos slikmutter, og æd den selv..
<W lemma="springe" msd="VAM-=------">Spring</W> ... <W lemma="danse" msd="VAM-=------">dans</W>
<W lemma="købe" msd="VAM-=------">køb</W> ... <W lemma="æde" msd="VAM-=------">æd</W>

5.2.4 Infinitivformen

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Infinitive F	-	=	-	-	-	-	Active A	-
Verb V	Main A	Infinitive F	-	=	-	-	-	-	Passive P	-
Verb V	Medial E	Infinitive F	-	=	-	-	-	-	Active A	-

Som vist i tabeludsnittet ovenfor markeres infinitivformen (navneformen) kun for bøjning i diatese. Den optræder enten som en del af en *at*-infinitiv eller som en nul-infinitiv, men i det første tilfælde samles infinitivmarkøren *at* ikke med infinitivformen som en flerordsforbindelse i PAROLE-korpusset (jf. også afsnit 5.7.1 om infinitivmarkøren). Som nævnt ovenfor skelnes der heller ikke mellem konjunktivformen og infinitivformen, da de begge tildeles en analyse som infinitivform.

..men det vigtigste er ikke at lade sig smitte af depressions-bacillen!..
<W lemma="lade" msd="VAF-=----A-">lade</W>
..Lyt! Gud være lovet, hun trækker vejret og græder ikke. Den gode lille pige..
<W lemma="være" msd="VAF-=----A-">være</W>
..for det beløb kan der godt og vel bygges én Storebæltsforbindelse om året..
<W lemma="bygge" msd="VAF-=----P-">bygges</W>
..at mennesket .. ikke behøver at nøjes med de gener, det blev givet ved undfangelsen..
<W lemma="nøjes" msd="VEF-=----A-">nøjes</W>

5.2.5 Gerundium

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Gerundium G	-	=	Singular S	Common C	Indefinite I	-	-	Unmarked U

Udtrykket 'gerundium' anvendes her om en slags verbalsubstantiv, der også er blevet kaldt 'kentaur'[34] og ifølge Diderichsen, 1968, s. 81 kan dannes af næsten alle verber. I PAROLE-korpusset markeres disse ord med de samme morfologiske træk som et tilsvarende appellativ (fælleskøn, singularis, ubestemt og ikke-genitiv).

..kunst og kunstnerisk skaben handler om modsætninger og deres frugtbare interaktion..
<W lemma="skabe" msd="VAG-=SCI--U">skaben</W>
..ingen andre kan formentlig få Bob Dylan til at sidde og iagttage sig med drenget undren..
<W lemma="undre" msd="VAG-=SCI--U">undren</W>

5.2.6 Participierne

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Participle P	Present R	=	Number	Gender	Definiteness	TrCat	-	Case
Verb V	Main A	Participle P	Past A	=	Number	Gender	Definiteness	TrCat	-	Case
VerbV	Medial E	Participle P	Past A	=	Number	Gender	Definiteness	TrCat	-	Case

Participiumformerne (tillægsformerne) anvendes både i verbets omskrevne former og som attributive eller prædikative beskrivere på samme måde som adjektiver. Som vist i tabeludsnittet ovenfor inddeles participierne ifølge tempus i præsens participium ('PR') og præteritum participium ('PA'). Præsens participiumformen er ubøjelig undtagen *-s* genitivkasusendelsen (jf. afsnit 5.1.7 om "substantivisk" anvendelse). Når præteritum participiumformen anvendes adjektivisk, kan den desuden have bøjningsformer, der angiver bestemthed, numerus og genus (RO96, §31). Anvendelsen af trækket 'transkategorisering' til at markere participiernes "adjektiviske" (eller "adverbielle") syntaktiske anvendelse i PAROLE-korpusset vil blive mere udførligt diskuteret i næste afsnit.

5.2.6.1 Transkategorisering af participier

De fleste participier i PAROLE-korpusset er underspecificeret for trækket 'transkategorisering'. Dette betyder, at der i langt de fleste tilfælde ikke skelnes mellem participiernes "verbale" anvendelse i sammensatte verbalformer og deres "adjektiviske" (evt. "adverbielle") anvendelse som attributiv eller prædikativ. Dette viser sig ved, at både 'A' (for adjektivisk anvendelse), 'R'(for adverbiel anvendelse) og 'U' (for umarkeret - dvs. verbal - anvendelse) optræder mellem kantede parenteser på plads 9 i participiernes morfosyntaktiske analyse. Dette angiver, at det pågældende participium er underspecificeret for hhv. adjektivisk, adverbiel og umarkeret anvendelse (her skal "umarkeret anvendelse" naturligvis forstås som verbal anvendelse). Dette medfører, at de underspecificerede participier er markeret for alle de andre træk der er relevante for både verbal, adjektivisk og adverbiel anvendelse (hvilket således omfatter numerus, genus, bestemthed og kasus). Brugeren af PAROLE-korpusset kan naturligvis vælge enten (*i*) helt at se bort fra disse markeringer for transkategorisering af participier og adjektiver (f.eks. hvis man vil træne en automatisk tagger på korpusset[35]) eller (*ii*) selv at udspecificere de underspecificerede transkategoriseringer.

[34] "De har kentaurstruktur; fortil er de opbygget som nominalhypotagmer .. bagtil som ledsætninger og infinitiver" (Hansen 1992, s. 74).
[35] Et indledende forsøg på at træne to forskellige automatiske korpustaggere på dette PAROLE-korpus er beskrevet i Keson, 1999.

Kun i nedenstående tre tilfælde underspecificeres participiet ikke for transkategorisering med 'ARU', men er til gengæld markeret med enten 'A' eller 'R' på plads 9. Jf. også afsnit 5.3.2 om adjektiver og participier samt afsnit 5.1.7 om "substantivisk" anvendelse.

Figur 13: Ikke-underspecificeret markering af transkategorisering af participier

(1) 'A' ved foranstillet/attributiv adjektivisk anvendelse[36]
(2) 'A' ved bøjning i andet end den "ubøjede" form
(3) 'R' ved foranstillet/attributiv adverbiel anvendelse[37]

..det var ikke et sted, man lod sig se, hvis man ville være velset i byens herskende indremissionske kredse..
<W lemma="herske" msd="VAPR=[SP][CN][DI]A-U">herskende</W>
..alkoholpromiller mellem 2,00 og 2,50 er takseret til den nævnte straf..
<W lemma="nævne" msd="VAPA=S[CN]DA-U">nævnte</W>
..injektioner med genbrugte engangskanyler eller kanyler, der ikke er rensede tilstrækkeligt..
<W lemma="rense" msd="VAPA=P[CN][DI]A-U">rensede</W>
..hun har samlet en stor flok unge, danseglade og overraskende dygtige balletfolk..
<W lemma="overraske" msd="VAPR=---R-U">overraskende</W>
..en undersøgelse af ernærings- og sundhedstilstanden blandt en .. gruppe, formodet raske Østeuropæiske børn..
<W lemma="formode" msd="VAPA=---R-U">formodet</W>

5.2.6.2 Præsens participium

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Participle P	Present R	=	Number	Gender	Definiteness	TrCat	-	Case

I dansk anvendes præsens participiumformen ofte "adjektivisk" eller sågar "substantivisk" — og i det sidste tilfælde kan den optræde med en *-s* genitivkasusendelse (jf. afsnit 5.1.7 om "substantivisk" anvendelse). Som det fremgår af **Figur 13** ovenfor, markeres præsens participiumformen kun for adjektivisk anvendelse ('A'), når den optræder med beskriverfunktion umiddelbart foran et substantiv (også selvom substantivkernens plads er tom). Præsens participiumformen markeres for adverbiel anvendelse ('R'), når den optræder med adverbiel funktion umiddelbart foran et adjektiv eller et adverbium. Ellers underspecificeres anvendelsen af præsens participium som nævnt ovenfor altid med 'ARU' på plads 9 af msd-analysen.

..at fjerne al affald fra arealet og slutdeponere affaldet i overensstemmelse med de gældende regler herfor..
<W lemma="gælde" msd="VAPR=[SP][CN][DI]A-U">gældende</W>
..har en dansk særordning ikke tilstrækkelig juridisk bindende karakter, kan den underkendes af EF-domstolen..
<W lemma="binde" msd="VAPR=[SP][CN][DI]A-U">bindende</W>
..et standpunkt, som .. afsluttende bliver begrundet med, at "jeg ikke længere kan have tillid..
<W lemma="afslutte" msd="VAPR=[SP][CN][DI][ARU]-U">afsluttende</W>
..sociallovgivningen, der omfattede Forsorgsloven - svarende til den nuværende Bistandslov..
<W lemma="svare" msd="VAPR=[SP][CN][DI][ARU]-U">svarende</W>

5.2.6.3 Præteritum participium

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Verb V	Main A	Participle P	Past A	=	Number	Gender	Definiteness	TrCat	-	Case
VerbV	Medial E	Participle P	Past A	=	Number	Gender	Definiteness	TrCat	-	Case

[36] Dette gælder også de tilfælde, hvor substantivkernens plads er tom (her kan participiet desuden optræde med en "substantivisk" *-s* genitivkasusendelse). Jf. afsnit 5.1.7 om "substantivisk anvendelse".
[37] Dette forekommer dog yderst sjældent (hhv. 12 og 2 forekomster af præsens og præteritum participium med attributiv adverbiel anvendelse i hele det PAROLE-korpus).

Ifølge RO96, §32 - §34 optræder præteritum participiumformen (*i*) i forbindelse med et hælpe- eller andet verbum, (*ii*) foran et substantiv eller (*iii*) som frit prædikativ.

Når præteritum participiumformen optræder efter hjælpeverberne *have* eller *få* (eller på anden måde er en del af sætningens verbal), skal den ifølge RO96, §33 anvendes i ubøjet form. Efter *være* og *blive* skal den anvendes i ubøjet form, når et handlingsforløb beskrives, mens den skal "behandles som et adjektiv og bøjes", når en tilstand beskrives. I RO96, §33.2.b tilkendegives det dog, at den ubøjede form også er almindelig sprogbrug her. Pga. den almene usikkerhed omkring bøjning af præteritum participiumformen i dansk skriftsprog (samt potentielle problemer med at træne en automatisk tagger), underspecificeres anvendelsen af præteritum participiumformer i PAROLE-korpusset med 'ARU' på plads 9 (transkategorisering) for alle forekomster af participiet i ubøjet form efter disse hjælpeverber.

Når præteritum participiumformen optræder på samme måde som et attributivt eller prædikativt adjektiv, er den bestemte bøjningsform lig med pluralisformen ifølge RO96, §32 (dvs. *-ede* eller *-te* for svagtbøjede verber, *-ne*, *-ede* eller *-te* for stærktbøjede verber). Bøjning i genus kan desuden forekomme i den ubestemte singularisform af de participier, som i bestemt form og pluralis kan have endelsen *-ne*. Der kan også skelnes mellem en intetkønsform (svarende til den "ubøjede" form) med endelsen *-et* og en fælleskønsform med endelsen *-en*. Når præteritum participiumformen anvendes attributivt foran et substantiv i PAROLE-korpusset angives anvendelsen af participiumformen som værende adjektivisk med 'A' på plads 9 (transkategorisering), også i de tilfælde, hvor substantivkernens plads er tom. I alle de tilfælde, hvor bøjningsformen af præteritum participium ikke falder sammen med den ubøjede form, angives anvendelsen af participiumformen altid som adjektivisk med 'A' på plads 9 uanset syntaktisk placering. Når præteritum participiumformen anvendes attributivt i en adjektivgruppe eller adverbielgruppe, angives anvendelsen af particpiumsformen som adverbiel med 'R' på plads 9 (jf. **Figur 13** ovenfor).

..jeg er kommet til at hade samfundet og øvrigheden for den råddenskab, jeg har været udsat for..
<W lemma="komme" msd="VAPA=S[CN]I[ARU]-U">kommet</W>
..De er i dæmonernes vold, hviskede præsten panisk til Hans, der stod bagbundet .. lige op ad djævlebilledet..
<W lemma="bagbinde" msd="VAPA=S[CN]I[ARU]-U">bagbundet</W>
..den djævlepræst, der nu rettede krumkniven over hans blottede bryst, blev ramt af lynet eller Guds vrede..
<W lemma="blotte" msd="VAPA=S[CN]DA-U">blottede</W>
..bortset fra en samplet optagelse af Dave der drejer tændingsnøglen i sin nyindkøbte Porsche..
<W lemma="sample" msd="VAPA=S[CN]IA-U">samplet</W>

5.3 Adjektiver, numeralier og adverbier

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adj A	Normal N	Degree	Gender	Number	Case	=	Definiteness	TrCat		
Adj A	Cardinal C	-	-	-	Case	=	-	-		
Adj A	Ordinal O	-	-	-	Case	=	-	-		
Adv R	General G	Degree								

Dette afsnit gennemgår behandlingen af adjektiver (tillægsord), numeralier (talord) og adverbier (biord), som er tæt beslægtede i PAROLE-tagsættet.

5.3.1 Adjektiver

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adj A	Normal N	Degree	Gender	Number	Case	=	Definiteness	TrCat		

I PAROLE-tagsættet består ordklassen (dvs. CatGram) 'A' både af "almindelige" adjektiver (tillægsord), "kardinale" adjektiver (kardinaltal) og "ordinale" adjektiver (ordinaltal). De almindelige adjektiver er som vist i tabeludsnittet ovenfor markeret for bøjning i komparation, genus, numerus, kasus, bestemthed samt transkategorisering. I denne vejledning anvendes udtrykket 'adjektiv' kun om "almindelige" adjektiver i traditionel forstand og udtrykket 'numeralie' om talord.

Adjektiver betegner normalt egenskaber hos personer, ting og lignende. De optræder enten som et attributivt præled (vedføjelse) eller som et fast eller frit prædikativ (omsagnsled). Der skelnes ikke eksplicit mellem deres attributive og prædikative anvendelse i PAROLE-korpusset.

..men deres indbyrdes forhold kompliceres ved, at de begge er forelsket i den sammen mand..
<W lemma="indbyrdes" msd="ANP[CN][SP]U=[DI]U">indbyrdes</W>
..venlig, men stille og kontant, betjening af piger med store flæse-forklæder..
<W lemma="stille" msd="ANP[CN][SP]U=[DI]U">stille</W>
..der er desuden bygget et lille køkken ind i præsidentsuiten, så han ingen chance får for at blive sulten..
<W lemma="sulten" msd="ANPCSU=IU">sulten</W>
..vi gjorde os lækre sammen og bestilte en taxa til klokken 21:00..
<W lemma="lækker" msd="ANP[CN]PU=[DI]U">lækre</W>
..alle er klædt om til middagen. Børnene sidder vandkæmmede og trætte ved bordene..
<W lemma="træt" msd="ANP[CN]PU=[DI]U">trætte</W>

5.3.1.1 Komparation, genus, numerus og bestemthed

Alle adjektiver, der ikke optræder gradbøjet i komparativ-, superlativ- eller absolut superlativ-(dvs. *aller*-)form, er markeret som værende positive, selvom (*i*) de ikke kan gradbøjes overhovedet, eller (*ii*) deres gradbøjningen er defektiv. En angivelse af positiv komparation har således kun praktisk betydning for de adjektiver, der kan gradbøjes. Alle adjektiver er desuden markeret for bøjning i genus, numerus og bestemthed. Hvis en eller flere af disse bøjningsformerne falder sammen, vil den morfosyntaktiske analyse indeholde alle de relevante træk mellem kantede parenteser ([ ]), således at adjektiver som f.eks. *tysk* eller *moderne* får hhv. tildelt analyserne: msd="ANP[CN]SU=IU" og msd="ANP[CN][SP]U=[DI]U" (jf. afsnit 2.3 om PAROLE-tagsættet).

..en få måneder gammel ændring i markedsføringsloven giver flyselskaber lov til at tilbyde kunden en tilgift..
<W lemma="gammel" msd="ANPCSU=IU">gammel</W>
..han kiggede nærmere på metalpladen og så, at det var et gammelt reklameskilt for Korsør Løve Margarine..
<W lemma="gammel" msd="ANPNSU=IU">gammelt</W>
..det førte ham det meste af Fyn rundt: til gamle købmænd, kolonihaver, staldlofter..
<W lemma="gammel" msd="ANP[CN]PU=[DI]U">gamle</W>
..det gav et pludseligt ryk i den gamle krop, og hånden hagede sig fast som en ørneklo..
<W lemma="gammel" msd="ANP[CN]SU=DU">gamle</W>
..de er så høflige og flinke, også over for ældre mennesker, som ikke køber deres plader..
<W lemma="gammel" msd="ANC[CN][SP]U=[DI]U">ældre</W>
..selv vil Prince karakterisere hele værket som en moderne opera..
<W lemma="moderne" msd="ANP[CN][SP]U=[DI]U">moderne</W>
..publikumsmæssige magneter, som - hver på sin måde - tegner det moderne Paris..
<W lemma="moderne" msd="ANP[CN][SP]U=[DI]U">moderne</W>
..Moderne biler får mindre og mindre vindmodstand, og det gør det nødvendigt, at trække luft ind..
<W lemma="moderne" msd="ANP[CN][SP]U=[DI]U">Moderne</W>

5.3.1.2 Kasus

Når adjektiver optræder som sidste præled i en substantivgruppe, hvori kerneleddets plads står tom, opfattes det af nogle som "substantivisk" anvendelse af adjektivet (jf. afsnit 5.1.7). Kun i disse tilfælde kan adjektivet optræde med en *-s* genitivendelse, dog indeholder PAROLE-analyserne af adjektiver (samt participier) altid en plads til markering af bøjning i kasus (genitiv 'G' og ikke-genitiv 'U') uanset adjektivets/participiets syntaktiske omgivelser.

..et værdifuldt middel til at kommunikere med de såkaldte "vidtgående fysisk og psykisk handicappede"..
<W lemma="handicappet" msd="ANP[CN]PU=[DI]U">handicappede</W>
..for at give støtte fører en pædagog den handicappedes hånd over stavepladen, indtil klienten peger på et tegn..
<W lemma="handicappet" msd="ANP[CN]SG=DU">handicappedes</W>
..der var spor af en trækiste, men ingen af den døde selv; alle knogler var for længst opløste..
<W lemma="død" msd="ANP[CN]SU=DU">døde</W>
..Sheraton i Luxor. Udsigt nedover svømmepølen og Nilen mod "De dødes rige"..
<W lemma="død" msd="ANP[CN]PG=[DI]U">dødes</W>

5.3.1.3 Transkategorisering af adjektiver

Når et leksikalsk adjektiv anvendes syntaktisk som et adverbium, markeres dette med et 'R' for "adverbiel" anvendelse på plads 9 (transkategorisering). Dette omfatter både (*i*) *t*-adverbialer (som falder sammen med adjektivets intetkønsform), (*ii*) adverbier dannet af adjektiver på *-ig, -lig* og *-vis* samt (*iii*) alle andre adverbialer, der dannes af adjektiver og derfor ikke er opført som selvstændige opslagsord i RO96 i modsætning til "rene" adverbier (RO96, §36 - §39). Jf. også afsnit 5.2.6.1 om transkategorisering af participier og afsnit 5.3.5 om adjektiver og adverbier.

..jeg anser det derfor for risikabelt at tvinge et gulv tilbage til sit oprindelige niveau..
<W lemma="oprindelig" msd="ANP[CN]SU=DU">oprindelige</W>
..hvor langt dette oprindelig har været, er umuligt at sige, da vi ikke ved hvordan den døde har været vendt..
<W lemma="oprindelig" msd="ANP---=-R">oprindelig</W>
..hvilke af aftenens sange, Nat King Cole oprindeligt lancerede, kan diskuteres på samme måde..
<W lemma="oprindelig" msd="ANP---=-R">oprindeligt</W>

5.3.2 Adjektiv eller participium?

I to forskellige tilfælde har det været nødvendigt for korpustaggerne at træffe et valg mellem at tildele en analyse som adjektiv eller en analyse som participium (evt. med adjektivisk anvendelse). For det første har de sammensætninger, der har participier som overled og som ikke optræder i RO96 — og derfor heller ikke i DAN-TWOL-leksikonet — automatisk fået tildelt en analyse som participium af DAN-TWOL-analysen. For det andet findes der adjektiver og participier med sammenfaldende former i både RO96 samt i andre ordbøger. I det første tilfælde er en analyse som participium blevet valgt, hvis det tilsvarende verbum eksisterer i en infinitivform eller indikativform. I det andet tilfælde skelnes der mellem adjektiver og participier i henhold til eksempelsætninger i Becker-Christensen et al, 1996, Vinterberg & Bodelsen, 1990 samt Juul-Jensen et al, 1919.

..et mega-stort træhus med en solopvarmet swimmingpool i baghaven..
<W lemma="solopvarmet" msd="ANP[CN]SU=IU">solopvarmet</W>
..udstyret bestod af en jernkniv i læderskede og med sølvtrådsomviklet træskaft..
<W lemma="sølvtrådsomviklet" msd="ANP[CN]SU=IU">sølvtrådsomviklet</W>
..et internationalt samarbejde, fra forskning til .. fredsbevarende og miljøbevarende opgaver..
<W lemma="miljøbevarende" msd="ANP[CN][SP]U=[DI]U">miljøbevarende</W>
..flertallet lever stadig under plastikstykker eller tæpper, som de har spændt over nogle stokke..
<W lemma="spænde" msd="VAPA=S[CN]I[ARU]-U">spændt</W>
..selvfølgelig var jeg spændt og nervøs for at få en kæmpegæld..

<W lemma="spændt" msd="ANP[CN]SU=IU">spændt</W>
..lensgreve Christian Lerches meget smukke og spændende møntsamling..
<W lemma="spændende" msd="ANP[CN][SP]U=[DI]U">spændende</W>

Mht. valget mellem participium, adjektiv og adverbium som mulig ordklasse har tendensen under korpustaggingen på samme måde som for substantiverne været at foretrække en mere informativ analyse (dvs. en analyse, der indeholder flere morfosyntaktiske oplysninger) frem for en mindre informativ analyse, hvilket giver følgende præcedens:

Participium > Adjektiv > Adverbium
VAP/VEP > AN > RG

5.3.3 Numeralier

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adj A	Cardinal C	-	-	-	Case	=	-	-		
Adj A	Ordinal O	-	-	-	Case	=	-	-		

Numeralier (talord) — dvs. kardinaltal (mængdetal) og ordinaltal (ordenstal) — adskiller sig fra adjektiverne ved stort set at være ubøjelige[38]. Desuden optræder de fleste numeralier af indlysende årsager kun sammen med substantiver i pluralis, og når de optræder som præled i sustantivgrupper, står de foran de adjektiviske præled. Som vist i tabeludsnittet ovenfor har kardinaltallene fået tildelt analysen 'AC', mens ordinaltallene har fået tildelt analysen 'AO'. Langt de fleste numeralier er desuden på samme måde som adjektiverne markeret med 'U' for 'ikke-genitiv' kasus. Lemmaformer for ordinaltal er angivet som ordinaltallene selv (enten som ord, numeriske tal eller romertal) og ikke som det tilsvarende kardinaltal. Ud over punktum og/eller komma kan numeralierne også indeholde en skråstreg eller et kolon (jf. afsnit 4.1.3 om skråstregen og afsnit 5.8.4 om formler).

.."Men der er ingen organisk sammenhæng mellem de to ting," understreger han..
<W lemma="to" msd="AC---U=--">to</W>
..man kan kun tillader sig at gætte, at op imod 5.000 børn bare i det ene land er HIV-smittede..
<W lemma="5.000" msd="AC---U=--">5.000</W>
..men vi gør et nyt forsøg i morgen," sagde Nyrup Rasmussen til TV 2s sene Nyheder..
<W lemma="2" msd="AC---G=--">2s</W>
..dog fik Foreigner et fortjent mindre gennembrud i England med deres fjerde album, "4" (hvor originalt)..
<W lemma="fjerde" msd="AO---U=--">fjerde</W>
..den 24. november 1983 besigtigede Miljø- og Vandinspektoratet Mogenstrup grusgrav..
<W lemma="24." msd="AO---U=--">24.</W>
..tømmeret er det oprindelige fra Christian 4.s renæssanceslot og er altså 350 år gammelt..
<W lemma="4." msd="AO---G=--">4.s</W>

Ordformen *første* klassificeres i RO96 og her i PAROLE-korpusset kun som et ubøjeligt adjektiv. I PAROLE-korpusset analyseres ordformerne *en* og *et* (samt *én*, *een* og *ét*) kun som ubestemte pronominer og aldrig som kardinaltal, da det var upraktisk at skelne mellem dem under korpustaggingen. Derimod analyseres ordformerne *anden* og *andet* enten som ordinaltal[39] eller som ubestemte pronominer.

..mine forældre kaldte hele tiden stedet "midlertidig" .. således at mit første hjem ikke var et rigtigt hjem..
<W lemma="første" msd="ANP[CN][SP]U=[DI]U">første</W>
..det synes Ina de første dage, så havde hun fået nok af Jesus' store familie..
<W lemma="første" msd="ANP[CN][SP]U=[DI]U">første</W>
..Et af vores EDB-lokaler i uddannelsescentret er udstyret med med scanneren Apple Macintosh SE..

[38]Den eneste undtagelse er ordinaltallene *anden* og *andet* (hvis man da ikke regner kardinale/ordinale tal for at være bøjningsformer – og ikke afledninger –af hinanden).
[39] Dog uden markering af bøjning i genus.

<W lemma="en" msd="PI-NSU--U">Et</W>

..den ensomme læge, fortalte, at havde man een gang drukket af Nilens vand, ville man længes tilbage..

<W lemma="en" msd="PI-CSU--U">een</W>

..den kulturkreds, der .. har været involveret i Anden Verdenskrig, krig i Vietnam, på Falklandsøerne..

<W lemma="anden" msd="AO---U=--">Anden</W>

..en anden optagelse viste en far, der græd. "Åh, mine børn," råbte han..

<W lemma="anden" msd="PI-CSU--U">anden</W>

5.3.4 Adverbier

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adverb R	General G	Unmarked U								
Adverb R	General G	Positive P								
Adverb R	General G	Comparative C								
Adverb R	General G	Superlative S								
Adverb R	General G	Abs. Superlative A								

Én kort definition på adverbier (biord) er "ord der ikke hører til de andre ordklasser" (Galberg Jacobsen, 1996, s. 22), hvilket meget passende beskriver den opsamlende funktion denne ordklasse har. Som i RO96 skelnes der i PAROLE-korpusset mellem "rene" adverbier og adverbialer dannet af adjektiver (som *t*-adverbialer osv.). Kun "rene" adverbier har fået tildelt analysen 'RG' som adverbium (jf. afsnit 5.3.1.3 om transkategorisering af adjektiver).

Traditionelt findes der forskellige klassifikationer af adverbier[40], dog er de ikke underinddelt i det danske PAROLE-tagsæt, men er derimod alle markeret med SsCatGram 'G' ('general'). De markeres desuden (som adjektiverne) for komparation. I modsætning til adjektiverne er de fleste adverbier, der ikke er bøjet i komparativ, superlativ eller absolut superlativ (*aller*-) komparation, umarkeret ('U') for komparation, medmindre de kan gradbøjes, i hvilket tilfælde de markeres for positiv gradbøjning ('P'). Nogle få tidsadverbialer som f.eks. *i går*, *for nylig* og *i søndags* samt andre flerordsadverbier som f.eks. *simpelt hen*, *på ny* og *i tusindvis* samles desuden konsekvent som flerordsforbindelser i PAROLE-korpusset (jf. afsnit 4.2 om flerordsforbindelser samt fortegnelsen over flerordsforbindelser i appendiks 8.5).

..det rystede fødselsdagsbarnet fra Kaliningrad i en sådan grad, at det aldrig lykkedes ham at bryde tilbage..

<W lemma="aldrig" msd="RGU">aldrig</W>

..ingen tvivl om, at ranglistens nummer 18, der i søndags var i finalen mod Boris Becker, ikke havde sin gode aften..

<W lemma="i_søndags" msd="RGU">i_søndags</W>

..et sæt, hvor Carlsen er leveringsdygtig i såvel en perlerække af flotte server som gode returneringer..

<W lemma="hvor" msd="RGU">hvor</W>

..foran et først forbavset - så jublende publikum - i KB-hallen formår Kenneth Carlsen at presse verdensstjernen..

<W lemma="først" msd="RGU">først</W>

5.3.5 Adjektiv eller adverbium?

Følgende ordformer, som optræder i PAROLE-korpusset, er i RO96 opført som værende både adjektiver og adverbier. Hvis disse ordformer anvendes med deres adverbielle betydning i PAROLE-korpusteksterne, har de fået tildelt en analyse som et "rent" adverbium, og ikke som transkategoriserede adjektiver: *antagelig, blot, egentlig, endelig, formelig, knap, lige, ligefrem, nok, nær, nødig, pludselig, præcis, ret, rundt, slet, stadig, så, sådan* og *virkelig.*

..et hult melodrama sammensat af lige dele had, pjank, drilleri og afsluttende hjerternes fortrolighed..

<W lemma="lige" msd="ANP[CN][SP]U=[DI]U">lige</W>

..DET har allerede i flere tilfælde medført at danskere er kommet hjem og er røget lige på bistandshjælp..

[40] F.eks. gradsadverbier, holdningsadverbier, konnektive adverbier, modale adverbier, mådesadverbier, stedsadverbier, tidsadverbier, årsagsadverbier osv.

<W lemma="lige" msd="RGU">lige</W>

..som forbruger er .. Peter Gren Larsens budget knapt. For en ung freelancer med købemani er kreditten kort..

<W lemma="knap" msd="ANPNSU=IU">knapt</W>

..den sad i hjernen, men det tog hende kun knap et døgn at komme over den..

<W lemma="knap" msd="RGU">knap</W>

Langt den sværeste beslutning under korpustaggingen var beslutningen om et adjektiv er blevet anvendt adverbielt eller ej i en given kontekst. Sandsynligvis er dette også den største potentielle fejlkilde i de taggede korpustekster. Denne beslutning blev så vidt muligt truffet i overenstemmelse med bøjningsformen af adjektivet. I tvivlstilfælde var der dog en generel præference blandt korpustaggerne for en analyse som adverbiel anvendelse af adjektivet frem for en analyse som (frit) prædikativt adjektiv.

..da hun pludselig blev ramt af en blodprop og faldt bevidstløs om..

<W lemma="bevidstløs" msd="ANPCSU=IU">bevidstløs</W>

..han er nu blandt de tre danskere, der frivilligt er kommet hjem fra Thailand til varetægtsfængsling..

<W lemma="frivillig" msd="ANP---=-R">frivilligt</W>

..Falcks svømmedykkere var hurtigt i vandet, men bilen stod tom på bunden af havnen..

<W lemma="tom" msd="ANPCSU=IU">tom</W>

..hun bøjede dovent det ene knæ og vinkede til ham med tæerne uden at løfte ansigtet..

<W lemma="doven" msd="ANP---=-R">dovent</W>

5.4 Præpositioner og konjunktioner

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adpos S	Preposition P									

Præpositioner og konjunktioner tilhører lukkede ordklasser, hvilket tydeligt fremgår af deres relative fordeling på tekstord og ordtyper i PAROLE-korpusset som vist i **Figur 12** i afsnit 5 ovenfor. Dette afsnit beskriver behandlingen af disse to ordklasser i PAROLE-korpusset.

5.4.1 Præpositioner

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Adpos S	Preposition P									
Conj C	Coordinating C									
Conj C	Subordinating S									

Præpositioner (forholdsord) er småord, der sætter andre ord og led i forhold til hinanden (Galberg Jacobsen, 1996), og sammen med deres styrelse (regimen) danner de en præpositionsforbindelse. Både præpositioner og postpositioner[41] analyseres som præpositioner her i PAROLE-korpusset. Følgende ordformer har fået tildelt en analyse som præposition ('SP'): *ad, af, angående (ang.), apropos, bag, blandt, bortset_fra, efter (e.), for, foran, forbi, foruden, fra, før, førend, gennem, henad, henover, hos, i, iblandt, ifølge (ifølg.), igennem, imellem, imod, in, inden, indenfor, indtil, inklusive (incl.), kontra, langs, med, med_hensyn_til (m.h.t.), mellem, mod, nedover, nær, om, omkring, opad, ovenpå, over, overfor, per (pr, pr.), på, på_grund_af (p.g.a., pga.), siden, til, trods, uanset, uden, udenfor, udfra, udover, under, undtagen, ved, vedrørende* samt *via.*

..og på intet tidspunkt flyver intellektet i historien over hovedet på læseren..

<W lemma="på" msd="SP">på</W>

[41] De er relativt sjældne i dansk (dog findes f.eks. *foruden* i *dem foruden*, *igennem* i *mange år igennem* og *over* i *natten over* eller *hele landet over*).

..i dag har man 10 regulære forretninger, hvor der bliver solgt varer af præcis samme .. kvalitet..
<W lemma="af" msd="SP">af</W>
..grænse-købmand Hans F. Fleggaard sendte i 1990 sin mest ærgerrige medarbejder af sted..
<W lemma="i" msd="SP">i</W>
..ved udgangen af juni beskæftigede de danske industrivirksomheder .. godt 270.000 personer..
<W lemma="ved" msd="SP">Ved</W>

Som det fremgår af listen ovenfor, bliver sammensatte præpositioner (jf. Allan et al, 1995, s. 372-4) normalt ikke samlet af DAN-TWOL-tokeniseren. Dette gælder både (*i*) sammensatte præpositioner af typen **adverbium + præposition** (som f.eks. i *inde i huset, nede på vejen, op ad skrænten, ovre på bakken osv.* — også retningsadverbier som f.eks. i *hen til træet, ud af døren, væk fra barnet, ind i stalden, frem til grænsen*), (*ii*) sammensatte præpositioner af typen **præposition + substantiv + præposition** (som f.eks. *i mangel af, i overensstemmelse med, med henvisning til, til gavn for, til trods for, ved siden af* osv.[42]) og (*iii*) cirkumpositioner (Allan et al, 1995), der analyseres enten som (*a*) **præposition + styrelse + adverbium** (som f.eks. *fra barndommen af, ad helvedet til, for nogle år siden, på ham nær*) eller som (*b*) **præposition + andet + substantiv** (som f.eks. *for hendes skyld, i dit sted, på firmaets vegne*).

I denne forbindelse skal det også nævnes, at forkert sammenskrevne sammensatte præpositioner (jf. RO, §19.1), som f.eks. *indenfor, henad, henover, nedover, opad, ovenpå, overfor, udenfor, udfra* og *udover,* ikke markeres som tekstfejl (msd="XX") i PAROLE-korpusset. Disse ordformer anerkendes i stedet som præpositioner hvis de optræder i forbindelse med en styrelse, ellers analyseres de på normal vis som adverbier (jf. afsnit 6.3.4 om sammenskrivninger).

.."Det er vanskeligt at forklare sig ovenpå en så stor personlig skuffelse," sagde sympatiske Dan Frost..
<W lemma="ovenpå" msd="SP">ovenpå</W>
..kliniske forskere, d.v.s. læger, som står overfor de alvorlige kræftsygdomme..
<W lemma="overfor" msd="SP">overfor</W>

5.4.2 Præposition eller adverbium?

Følgende ordformer er analyseret som værende enten præpositioner ('SP') eller adverbier ('RG') i PAROLE-korpusset : *af, bag, efter, for, foran, forbi, fra, før, i, igennem, imellem, imod, inden, indenfor, langs, med, nær, om, omkring, opad, ovenpå, over, overfor, på, siden, til, uden, udenfor, udover, under* og *ved.* Der er skelnet mellem disse to ordklasser på baggrund af tilstedeværelsen af en styrelse med henvisning til eksempelsætninger i Becker-Christensen et al, 1996, Vinterberg & Bodelsen, 1990 og Juul-Jensen et al, 1919. I spørgesætninger, relativsætninger, sætningskløvninger og sætninger med topikalisering/emfase kan præpositionen stå langt fra sin styrelse, hvilket naturligvis også er en potentiel fejlkilde under korpustaggingen.

..omvendt vil jeg spørge: Hvad skal vi bruge den økonomiske vækst til?..
<W lemma="til" msd="SP">til</W>
.."Vi kender intet til terroristerne," svarer nationalistpartiet Sinn Fein, men det er der ingen, der tror på..
<W lemma="på" msd="SP">på</W>

Da DAN-TWOL-tokeniseren ikke samler partikelverber, har verbalpartikler tilhørende ovennævnte liste fået tildelt en morfosyntaktiske analyse alene på baggrund af tilstedeværelsen af en styrelse, uden hensyn til deres tilknytning til verbet.

..Anders Bircow, 40, slapper i denne tid af efter en succésfuld sæson..

[42] Undtagelserne er *med hensyn til* og *på grund af*, som begge på grund af deres hyppighed samles som flerordsforbindelser i PAROLE-korpusteksterne (jf. appendiks 8.5).

<W lemma="af" msd="RGU">af</W>

..dette får skoleeleverne til at stå af én station før Heureka-centret..

<W lemma="af" msd="RGU">af</W>

..i forkontoret er vi en ryger og en ikke-ryger. Vi må jo se at finde ud af det..

<W lemma="af" msd="SP">af</W>

5.4.3 Konjunktioner

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Conj C	Coordinating C									
Conj C	Subordinating S									

Konjunktionerne (bindeordene) tilhører en lukket ordklasse bestående af småord, der forbinder ord, sætningsled og sætninger (Galberg Jacobsen, 1996). Traditionelt skelnes der ifølge forbindelsesart mellem sideordningskonjunktioner, underordningskonjunktioner og samordningskonjunktioner. Som vist i tabeludsnittet ovenfor er konjunktionerne i PAROLE-tagsættet imidlertid kun underinddelt i sideordningskonjunktioner og underordningskonjunktioner. Samordningskonjunktionerne *for* og *thi* er analyseret som sideordningskonjunktioner, mens alle forekomster af konjunktionen *så* af praktiske hensyn får tildelt en analyse som underordningskonjunktion.

Følgende ordformer har fået tildelt en analyse som sideordningskonjunktion ('CC') i PAROLE-korpusset: *eller, for, jamen, men, og (&), samt* og *thi.* Flerleddede konjunktioner (som f.eks. *hvad enten .. eller, forudsat at, efter at, så at* osv.) samles ikke men tagges hver for sig (jf. Allan et al, 1995, s. 463).

..det er jo ikke sådan, at vi lægger op til konflikt, men i forvejen er udluftningen i ministeriet ikke for god..

<W lemma="men" msd="CC">men</W>

..det er en simpel nødvendighed, for han er alene hjemme. Konen er på arbejde, børnene i skole..

<W lemma="for" msd="CC">for</W>

Følgende ordformer har fået tildelt en analyse som underordningskonjunktion ('CS') i PAROLE-korpusset: *alt_imens (altimens), at, bare, da, dengang, efter, eftersom, end, enten, fordi, fra, før, hvis, hvorvidt, idet, imens, inden, indtil, jo, ligesom, medens, medmindre, mens, når, om, om_end (omend), qua, selv_om, selvom, siden, sidst, skønt, som_om, straks, så, såfremt, såsom, til* og *uden.*

..en sandfærdig beretning om, hvordan en arbejdsløs føler sig til grin, fordi han har lyst til at arbejde..

<W lemma="fordi" msd="CS">fordi</W>

..hvis jeres forældre læser dette, så kan de være helt rolige - det var en flok pæne borgerlige unge mennesker..

<W lemma="hvis" msd="CS">hvis</W>

5.4.4 Konjunktion eller præposition?

Følgende ordformer er blevet analyseret som værende enten en konjunktion ('CC' eller 'CS') eller en præposition ('SP') i PAROLE-korpusset: *efter, for, fra, før, inden, indtil, om, siden, til* og *uden.* For at skelne mellem deres anvendelse som konjunktion og som præposition fulgte vi eksempelsætninger i Becker-Christensen et al, 1996, Vinterberg & Bodelsen, 1990 og Juul-Jensen et al, 1919.

..her er læsernes egen side med spørgsmål om biler og bilfolk..

<W lemma="om" msd="SP">om</W>

..det er umuligt at overskue, om en så stor overførsel er mulig..

<W lemma="om" msd="CS">om</W>

..fire døde inden et år efter operationen, men de øvrige havde det alle meget bedre..

<W lemma="inden" msd="SP">inden</W>
..men inden den 80-årige mand nåede til telefonen, var røveren oppe i stueetagen..
<W lemma="inden" msd="CS">inden</W>

5.5 Pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Demonstrative D	-	Gender	Number	Case	-	-	Register		
Pron P	Indefinite I	-	Gender	Number	Case	-	-	Register		
Pron P	Interrogative/ Relative T	-	Gender	Number	Case	-	-	Register		
Pron P	Personal P	Person	Gender	Number	Case	-	Reflexive	Register		
Pron P	Possessive O	Person	Gender	Number	Case	Possessor	Reflexive	Register		
Pron P	Reciprocal C	-	-	Number	Case	-	-	-		

Ovenstående tabeludsnit[43] viser PAROLE-tagsættet for pronominerne (stedord).

5.5.1 Demonstrative pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Demonstrative D	-	Gender	Number	Case	-	-	Register		

Som vist i tabeludsnittet ovenfor markeres demonstrative (påpegende) pronominer for genus, numerus, kasus samt stilleje ('Register'). Følgende ordformer har fået tildelt en analyse som demonstrativt pronomen ('PD') i PAROLE-korpusset: *begge, de, den (dén, d.), denne, dennes (ds.), det (dét), dette, disse, disses, hin* samt *selv*[44]. Disse ordformer er alle umarkeret for stilleje ('U') undtagen *hin*, der er markeret med 'O' for forældet ('Obsolete') stilleje.

I denne forbindelse er det vigtigt at være opmærksom på, at PAROLE-tagsættet ikke indeholder ordklassen 'artikler'. Derimod analyseres de tre ordformer, der svarer til de bestemte artikler (*den, det* og *de*), som enten (*i*) demonstrative pronominer eller (*ii*) personlige pronominer, og aldrig som bestemte artikler, eller (mht. *det*) som formelt eller foreløbigt subjekt.

Den traditionelle distinktion mellem demonstrative og personlige pronominer følges ikke i PAROLE-korpusset af hensyn til de praktiske problemer med at entyddiggøre disse ordformer under korpustaggingen (samt potentielle problemer med at træne en automatisk korpustagger). *Den, det* og *de* analyseres som demonstrative pronominer, når de optræder som adnominaler i substantivgrupper, hvilket er en typisk placering for både bestemte artikler og demonstrative pronominer. Derimod analyseres *den, det* og *de* som personlige pronominer, når de optræder som pronominaler, hvilket er en typisk placering for demonstrative og personlige pronominer. Mht. *den, det* og *de* svarer PAROLE-analysen som demonstrativt pronomen således til anvendelsen som bestemt artikel eller som adnominalt demonstrativt pronomen, mens PAROLE-analysen som personligt pronomen svarer til anvendelsen som personligt pronomen eller som pronominalt demonstrativt pronomen. *Denne, dennes, dette, disse* og *disses* analyseres derimod altid som demonstrative pronominer uanset deres syntaktisk placering. Ordformerne *dens* og *dets* analyseres derimod altid som possessive pronominer uanset deres syntaktiske placering.

..inden den 80-årige mand nåede til telefonen, var røveren oppe i stueetagen..
<W lemma="den" msd="PD-CSU--U">den</W>

[43] Tabellen for pronominer er ændret lidt i forhold til tabellen i Bilgram & Keson, 1998 (værdierne for hankøn/hunkøn og ejergenus er fjernet her).
[44] *Selv* kaldes i nogle grammatikker for 'et emfatisk pronomen' (jf. f.eks. Allan et al, 1995, s.169).

..torsdag aften .. - den aften, de troede var deres sidste i civilisationen -..
<W lemma="den" msd="PD-CSU--U">den</W>
..brug den i tynde flager som gratinering af en ovnret eller i tern i salatskålen..
<W lemma="den" msd="PP3CSU-NU">den</W>
..sommer i det ny Europa. Ud at se med DSB - ud at lufte den indre sigøjner..
<W lemma="den" msd="PD-NSU--U">det</W>
..er velegnet til .. hunde og mennesker i polarkulden - og smager lige præcis af det smagsstof, der tilsættes..
<W lemma="den" msd="PD-NSU--U">det</W>
..skal man blot huske på, at man ikke må bruge madrester. Det vil kunne tiltrække rotter..
<W lemma="det" msd="PP3NSU-NU">Det</W>

5.5.2 Ubestemte pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Indefinite I	-	Gender	Number	Case	-	-	Register		

Som vist ovenfor markeres ubestemte (indefinitte) pronominer i PAROLE-tagsættet for genus, numerus, kasus samt stilleje ('Register'). Følgende ordformer har fået tildelt en analyse som ubestemt pronomen ('PI') i PAROLE-korpusset: *alting, anden, andens, andet, andre, andres, en (een, én), et (eet, ét, èt) enhver, ens (éns), ethvert, hver, hvert, ingen, ingenting, intet, man, nogen, nogens, noget, nogle* og *somme.* Disse ordformer er alle umarkeret for stilleje ('U') undtagen *somme*, der er markeret med 'O' for 'obsolete' (forældet).

I forbindelse med de ubestemte pronominer er det igen vigtigt at være opmærksom på, at PAROLE-tagsættet ikke indeholder ordklassen 'artikler'. Ordformerne *en* og *et* analyseres derfor altid som ubestemte pronominer og aldrig som ubestemte artikler eller som kardinaltal[45]. Desuden analyseres ordformerne *nogen* og *nogens* (samt *ingen* og *ingens*) altid som ubestemte pronominer (bøjet i singularis fælleskøn) i PAROLE-korpusset, selvom der naturligvis kan være tale om bøjning i pluralis i benægtende, betingende og spørgende sætninger[46] (igen i modsætning til PAROLE-leksikonet).

..sætte gang i regnormenes aktivitet i havens bede. Og det er alfa og omega for en mere muldet jord..
<W lemma="en" msd="PI-CSU--U">en</W>
..i hvert fald ikke, så længe der bare er én, der holder af mig..
<W lemma="en" msd="PI-CSU--U">én</W>
..uden for teltet sidder en lille pige med et spædbarn på skødet..
<W lemma="en" msd="PI-NSU--U">et</W>
..hun stod et øjeblik stille og lyttede. Der manglede et eller andet..
<W lemma="en" msd="PI-NSU--U">et</W>
..efterhånden .. var ingen længere i tvivl om, at her var slagteriets ledelse ude på en straffeekspedition..
<W lemma="ingen" msd="PI-CSU--U">ingen</W>
..politiet har jævnligt besøgt stedet i månedsvis uden anden forklaring end at "de ledte efter nogen"..
<W lemma="nogen" msd="PI-CSU--U">nogen</W>

5.5.3 Interrogative/relative pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Interrogative/Relative T	-	Gender	Number	Case	-	-	Register		

Tabeludsnittet ovenfor viser, at der ikke skelnes mellem interrogative (spørgende) og relative (henførende) pronominer i PAROLE-tagsættet (igen af hensyn til praktiske problemer med at træne en automatisk tagger). Disse interrogative/relative pronominer markeres som vist ovenfor for genus, numerus, kasus samt stilleje ('Register'). Følgende ordformer har fået tildelt

[45] I modsætning til det danske PAROLE-leksikon.
[46] Jf. Hansen, 1975.

en analyse som interrogativt/relativt pronomen ('PT') i PAROLE-korpusset: *hvad (hva, hva'), hvem, hvilke, hvilken, hvilket* samt *hvis* (mere om *som* og *der* i afsnit 5.7.2). Disse ordformer er alle umarkeret for stilleje ('U').[47] Flerleddede interrogative eller relative pronominer (som f.eks. *hvem/hvad/hvilke der, hvad for en/et/nogle/nogen*) samles ikke, men tagges hvert for sig (jf. Allan et al, 1995, s. 192).

I PAROLE-korpusset skelnes der ikke mellem den interrogative og relative anvendelse af disse *hv*-pronominer. Der skelnes dog naturligvis stadig mellem *hvis* som interrogativt/relativt pronomen og som konjunktion. Relativpronominerne *som* og *der* behandles for sig (jf. afsnit 5.7.2), og her skelnes der ikke mellem *som* som relativpronomen og konjunktion, eller mellem *der* som relativpronomen og formelt subjekt.

..det er ligeledes interessant, hvad regeringen vil med redegørelsen i det udenrigspolitiske nævn..
<W lemma="hvad" msd="PT-[CN]SU--U">hvad</W>
..Nørrebro Bibliotek introducerede for et par år siden NU-bøgerne – hvilket vil sige, helt aktuelle bøger..
<W lemma="hvilken" msd="PT-NSU--U">hvilket</W>
..alene de 24.000 tilskuere var noget helt specielt. Hvilken lydkulisse..
<W lemma="hvilken" msd="PT-CSU--U">Hvilken</W>
..på gulvet stod der montrer, hvis glas var knust af det ødelagte tag..
<W lemma="hvis" msd="PT-[CN][SP]G--U">hvis</W>
.."Hvem støtter du?", er teksten over billederne af Mogens Glistrup og en lille jugoslavisk pige..
<W lemma="hvem" msd="PT-C[SP]U--U">Hvem</W>

5.5.4 Personlige pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Personal P	Person	Gender	Number	Case	-	Reflexive	Register		

Tabeludsnittet ovenfor viser, at personlige pronominer markeres for person, genus, numerus, kasus (nominativ, akkusativ eller genitiv), refleksivitet (refleksiv eller ikke-refleksiv) samt stilleje ('Register') i PAROLE-tagsættet. Følgende ordformer har fået tildelt en analyse som personligt pronomen ('PP') i PAROLE-korpusset: *de, De, dem, Dem, den (dén), det (dét), dig, du, ham, han, hende, hun, I, jeg, jer, mig, os, sig* og *vi*. Disse ordformer er alle umarkeret for stilleje ('U') undtagen *De* og *Dem*, der er markeret med 'P' for 'polite' (høflig). Alle personlige pronominer er markeret 'N' for ikke-refleksivitet, undtagen *sig*, der er markeret 'Y' for refleksivitet, og *Dem, jer, dig, mig* og *os*, der er markeret med 'YN', da de kan optræde både refleksivt og ikke-refleksivt.

Som allerede nævnt i afsnit 5.5.1 om demonstrative pronominer analyseres ordformerne *den, det* og *de* altid som personlige pronominer, når de optræder alenestående som pronominaler. *Det* analyseres aldrig som et formelt/foreløbigt subjekt, igen af hensyn til potentielle praktiske problemer med at træne en automatisk tagger.

..når jeg steg ind i den, var jeg stadig den pæne pige. Når jeg steg ud forvandledes jeg til vamp..
<W lemma="den" msd="PP3CSU-NU">den</W>
..Gunnar trak den øverste skuffe ud og lagde brevet på bordet..
<W lemma="den" msd="PD-CSU--U">den</W>
..nu var det alvor. Nu gjorde det ondt. Nej, det vil jeg ikke. Hvorfor det? - Fordi ... nåja..
<W lemma="det" msd="PP3NSU-NU">det</W>
..Antinoos, der altid førte det store ord i hallen, råbte: - Hvad skal vi med alle de tiggere..
<W lemma="den" msd="PD-NSU--U">det</W>
..derimod er det muligt, at firmaet, der havde videomaskinen i sin varetægt, skal erstatte den..

[47] Undtagelsen er *hvo*, som kun optræder i PAROLE-leksikonet, hvor den er markeret med 'O' for 'obsolete' (forældet) stilleje.

<W lemma="det" msd="PP3NSU-NU">det</W>

I PAROLE-korpusset skelnes der ikke eksplicit mellem anvendelsen af de personlige pronominer *Dem, dig, jer, mig* og *os* som refleksive og som ikke-refleksive pronominer, da alle forekomster af disse ordformer markeres med 'YN' for refleksivitet.

..og det er mig, der hedder Anni, sagde Anni og hoppede og dansede på fødderne..
<W lemma="jeg" msd="PP1CSU-[YN]U">mig</W>
..på en måde glæder jeg mig til at kunne give mine patienter en god juleaften..
<W lemma="jeg" msd="PP1CSU-[YN]U">mig</W>

5.5.5 Possessive pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Possessive O	Person	Gender	Number	Case	Possessor	Reflexive	Register		

I PAROLE-korpusset markeres possessive pronominer (ejestedord) for person, genus, numerus, kasus, ejernumerus ('Possessor'), refleksivitet og stilleje ('Register'). Følgende ordformer har fået tildelt en analyse som possessivt pronomen ('PO') i PAROLE-korpusset: *Deres, dens, deres, dets, din, dine, dit, hans, hendes, jeres, min, mine, mit, sin, sine, sit, vor, vore, vores* samt *vort*. Disse ordformer er alle umarkeret for stilleje ('U'). Undtagelserne er *Deres*, der er markeret med 'P' for 'polite' (høflig), samt *vor, vort* og *vore*, der er markeret med 'F' for 'formal' (formel). Alle possessive pronominer er markeret 'N' for ikke-refleksivitet undtagen *sin, sit* og *sine*, der alle er markeret 'Y' for refleksivitet.

Som nævnt i afsnit 5.5.1 om demonstrative pronominer analyseres *dens* og *dets* altid som possessive pronominer og aldrig som demonstrative pronominer.

..jeg skulle møde på min vagt klokken seksten, så da vi havde drukket kaffen .. måtte jeg bryde op..
<W lemma="min" msd="PO1CSUSNU">min</W>
..Heidi sendte sine bedsteforældre et hurtigt blik, men så var det ligesom hun opgav..
<W lemma="sin" msd="PO3[CN]PUSYU">sine</W>
..et værelse fyldt med legetøj og møbler, der passede i hans størrelse..
<W lemma="hans" msd="PO3[CN][SP]USNU">hans</W>
..hunden gjorde i sin vildskab så voldsom modstand, at et led i dens halsjernlænke simpelt hen knækkede..
<W lemma="dens" msd="PO3[CN][SP]USNU">dens</W>

5.5.6 Reciprokke pronominer

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Pron P	Reciprocal C	-	-	Number	Case	-	-	-		

Reciprokke (gensidige) pronominer markeres for numerus og kasus i PAROLE-tagsættet. I PAROLE-korpusset har kun to ordformer fået tildelt en analyse som reciprokt pronomen ('PC'): *hinanden* og *hinandens* (da *hverandre* og *hverandres* ikke optræder i PAROLE-korpusset).

..hendes arbejdsmæssige omgivelser .. kigger på hinanden og tænker: "Nu er hun sgu da for meget"..
<W lemma="hinanden" msd="PC--PU---">hinanden</W>
..søstre kender hinandens inderste på en måde som ingen anden kender dem..
<W lemma="hinanden" msd="PC--PG---">hinandens</W>

5.6 Interjektioner

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Interj	=									

Ordklassen 'interjektioner' dækker både lydord og udråbsord. Lydord og udråbsord er ubøjelige ord, der kan fungere alene som selvstændige ytringer, men ikke har nogen bestemt funktion som led i en sætning (Galberg Jacobsen & Skyum-Nielsen, 1996). Lydord er lydefterlignende og kan opdeles i (*i*) udtryk for følelse (*ak, ih, hmm, ushh* osv.), (*ii*) udtryk for tilskyndelse (som f.eks. *hyp*), (*iii*) lydmalende ord (som f.eks. *bom, tuk-tuk* osv.) samt (*iv*) de såkaldte "rodeord" (som f.eks. *hulk, hvin* og *suk*). Disse ordformer analyseres alle som interjektioner undtagen "rodeordene". Udråbsord er derimod ikke lydefterlignende (*nej, ja, bevares, farvel* osv)[48]. Alle stavevarianter af lydord anerkendes i PAROLE-korpusset (jf. RO96, §68). Følgende ordformer har således fået tildelt en analyse som interjektion ('I='): *aha, ak, alloh, bevares, bom, eh, farvel, fy, godaften, go'da, go'morgen, goddag, godnat, Guuud, hallo, hej, Herregud, hmm, hurra, hva'be'har, hva'beha'r, ih, ja, jah, javel, jo, kors, la, mmm, morn, nej, nuvel, nå, nåja, næ, næh, oh, o.k., ok, pyt, så, såh, tillykke, tja, tjoh, tuk-tuk, ushh, vips, vorherre, værsgo, wau, øv, åh* samt *åååårrr.*

..efter gymnastikken går vi til vore forskellige sysler indtil der bli'r råbt <u>værsgo</u>!..
<W lemma="værsgo" msd="I=">værsgo</W>
..<u>Aha</u> - tilbage til Bikuben. De kunne dog ikke udtale sig om sagen..
<W lemma="aha" msd="I=">Aha</W>
.."<u>Ja</u>, grundlæggende vil de gerne være i god form," svarer Bettina Borg..
<W lemma="ja" msd="I=">Ja</W>
..et endeligt <u>ja</u> til forliget kræver mindst 50 pct. af stemmerne..
<W lemma="ja" msd="NCNSU==I">ja</W>

5.7 Unique

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Unique	=									

De ordformer, der tilhører ordklassen 'unique' ('U='), er alle "unikke" og er ikke klassificeret som tilhørende de andre ordklasser. Her findes infinitivmarkøren *at*, der ikke samles med infinitivformen i PAROLE-korpusset, samt de to underspecificerede ordformer *som* (enten som konjunktion eller som relativpronomen) og *der* (enten som formelt/foreløbigt subjekt eller som relativpronomen).

5.7.1 Infinitivmarkøren

Når ordformen *at* anvendes som infinitivmarkør i forbindelse med et infinitivt verbum, samles den ikke med infinitivformen, men tildeles i stedet en analyse som 'unique' (jf. afsnit 5.2.4 om infinitivformen):

..en lige så barsk og kynisk Harrison Ford, der også tvinges til <u>at</u> forbedre sig..
<W lemma="at" msd="U=">at</W>
..en linie fra og med årgangene 1987-88, som er ønskværdig for Chateau Belgrave <u>at</u> følge fremover..
<W lemma="at" msd="U=">at</W>

5.7.2 *Som* og *der*

[48] Nogle udråbsord (som f.eks. *ja* og *nej*) har desuden i RO96 og i PAROLE-korpusset fået tildelt ordklassen som substantiv.

Som i RO96 skelnes der af praktiske årsager ikke mellem *som* som relativpronomen og som konjunktion (eller evt. som præposition), eller mellem *der* som relativpronomen og som formelt/foreløbigt subjekt. Som en slags anerkendelse af deres særlige status i danske grammatik har disse to ordformer dog fået tildelt en analyse som 'unique' i stedet for hhv. konjunktion og formelt subjekt. Det er således op til brugeren selv, om hun ønsker at udspecificere disse ordformers funktion i korpusteksterne.

I PAROLE-korpusset skelnes der stadig mellem *der* som 'unique' og som adverbium. I tvivlstilfælde mht. analysen af *der* har tendensen under korpustaggingen været at analysere *der* som 'unique' frem for som et adverbium. Som nævnt i afsnit 5.5.1 ovenfor analyseres ordformen *det* enten som demonstrativt pronomen eller som personligt pronomen, men aldrig som formelt/foreløbigt subjekt.

..slagger fra udslukte magmakamre, som engang var ildovne, fulde af smeltede boblende stenmasser..
<W lemma="som" msd="U=">som</W>
..det er marxismen, der er død som en lerdue, og gudskelov for det!..
<W lemma="som" msd="U=">som</W>
..hver gang hun går indendørs, ("under læ", som det hedder her), hiver hun .. den hvide matroshat af..
<W lemma="som" msd="U=">som</W>
..australiereren Vince Jones, der også kan spille lidt trompet med følsom Chet Baker-tone..
<W lemma="der" msd="U=">der</W>
..desuden vil der være lejlighed til at se minelæggeren "Lossen", minestrygerne "Vilsund" og "Grønsund"..
<W lemma="der" msd="U=">der</W>
..Nigeria blev .. kun lige nævnt, .. fordi det ville kræve en hel udsendelse at skildre, hvad der foregår der..
<W lemma="der" msd="U=">der</W>

5.8 Residual

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Abbreviation A									
ResidualX	Foreign F									
Residual X	Punctuation P									
ResidualX	Formulae R									
Residual X	Symbol S									
Residual X	Other X									

Som det fremgår af tabeludsnittet ovenfor dækker 'residual' en broget restgruppe i PAROLE-tagsættet. Dette afsnit gennemgår de forskellige kategorier i 'residual'-gruppen undtagen kategorien 'other', som gennemgås til sidst i et selvstændigt kapitel om 'tekstfejl og sproglige afvigelser' (kapitel 6).

5.8.1 Forkortelser

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Abbreviation A									

De fleste forkortelser i PAROLE-korpusset har ikke fået tildelt denne 'XA'-analyse som forkortelse. De har i stedet fået tildelt en morfosyntaktisk analyse på baggrund af deres fulde form efter opslag i bl.a. Eriksen & Hamburger, 1988. 'XA'-analysen er kun tildelt en mindre restgruppe af forkortelser, der —af forskellige årsager — ikke kunne få en "rigtig" morfosyntaktisk analyse.

..den fartglade forbryder, der fræser ud ad vort motorvejssystem med den formidable hastighed af 120 km/t..

<W lemma="km/t" msd="XA">km/t</W>

..86 lifter giver liftkortet til FF 900,- os ret til at benytte ubegrænset i en uge..

<W lemma="FF" msd="XA">FF</W>

..scenografi Erik Mortensen, ass. af Joakum Zacho Weylandt. Iscenesættelse Piv Bernth..

<W lemma="ass." msd="XA">ass.</W>

..juridiske eksperter i bofællesskabets jungle er ejendomsmægler, cand. jur. Svend Trangeled..

<W lemma="candidatus_juris" msd="XA">cand._jur.</W>

5.8.2 Udenlandske ord

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Foreign F									

Som nævnt i afsnit 5.1.8 om udenlandske ord er det kun de ord, der blev vurderet til at være uegnede til en dansk morfosyntaktisk analyse, der har fået en analyse som udenlandsk ord ('XF'). Dette indbefatter alle "ukendte" (dvs. ukendte for DAN-TWOL) fremmedord, der er skrevet med lille begyndelsesbogstav. "Ukendte" fremmedord skrevet med stort begyndelsesbogstav har derimod fået tildelt en analyse som "rigtigt" proprium (jf. begrundelsen herfor i afsnit 5.1.8).

..tyrkiske mænd drikke næsten altid raki til maden. Det er en brændevin med anissmag..

<W lemma="raki" msd="XF">raki</W>

..vi var de bedste, vi var mestrene, sejren var vor, og den helt nye "Tyskland, Tyskland, alles kaput"..

<W lemma="alles" msd="XF">alles</W>
<W lemma="kaput" msd="XF">kaput</W>

..evnen til at møde disse krav med nye ideer og en produktionskapacitet som er second to none..

<W lemma="second" msd="XF">second</W>
<W lemma="to" msd="XF">to</W>
<W lemma="none" msd="XF">none</W>

..ABBA "Gold" har været massivt annonceret på tv, ligesom AC/DC og andre "greatest hits"..

<W lemma="greatest" msd="XF">greatest</W>
<W lemma="hits" msd="XF">hits</W>

5.8.3 Interpunktionstegn

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Punctuation P									

Følgende tegn har fået tildelt en analyse som interpunktionstegn 'XP' i PAROLE-korpusset (jf. også afsnit 4.1):

Figur 14: Interpunktionstegn

punktum	**.**	14.142	**udråbstegn**	**!**	226
to prikker	**..**	4	**bindestreg (tankestreg)**[49]	**-**	1.755
tre prikker	**...**	135	**venstre parentes**	**(**	299
komma	**,**	17.482	**højre parentes**	**)**	319
kolon	**:**	764	**anførselstegn**	**"**	4.738
semikolon	**;**	54	**skråstreg**	**/**	3
spørgsmålstegn	**?**	470	**I alt**		40.391

..en atmosfære, som indeholder store mængder kuldioxid og svovlsyre. Himlen er orangerød..

<W lemma="." msd="XP">.</W>

..EF giver 3.500 kroner pr. hektar for vårraps og 4500 kroner for vinterrapsen..

[49] Alle tankestreger (–) er konverteret til bindestreger (-) inden behandlingen af DAN-TWOL-tokeniseren.

<W lemma="per" msd="SP">pr.</W>

..vi er sikker på en finaleplads når Henrik Svarrer / Marlene Thomsen møde vores ..bedste konstellation..

<W lemma="/" msd="XS">/</W>

..pladens titel, som ikke er et ord eller en sætning, men et elegant sammensmeltet mand/kvinde-tegn..

<W lemma="mand/kvinde-tegn" msd="NCNSU==I">mand/kvinde-tegn</W>

..med tilbageredt brillantinehår, smøg i kæften og ærmerne smøget op - sindsbilledet på manden..

<W lemma=" - " msd="XS"> - </W>

..måske har de fået den af en mafia-gruppe ude ved grænsen, siger Mehmet..

<W lemma="mafia-gruppe" msd="NCCSU==I">mafia-gruppe</W>

..socialdemokratiet og SF opnår flertal i Folketinget ("Det røde kabinet")..

<W lemma=""" msd="XP">"</W>

..måske har Venus' overflade tidligere været dækket af hav, som nu er fordampet..

<W lemma="Venus" msd="NP--G==-">Venus'</W>

5.8.4 Formler, mm.

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Formulae R									

Betegnelsen 'formler' anvendes her som en betegnelse for de "ukendte" ordformer, der består af en sekvens af bogstaver og tal (og evt. interpunktiontegn eller symboler). Undtagelserne er *B1903, Eleva2ren* og *TV-2*, der alle pga. deres hyppighed er analyseret som "rigtige" proprier i PAROLE-korpusset. Andre kombinationer af tal og forskellige interpunktionstegn[50], der ikke er blevet adskilt på forhånd af DAN-TWOL-tokeniseren, behandles som numeralier (jf. afsnit 5.3.3 om numeralier).

..premierministerens tale kom midt under de store politiske debatter på de to tv-kanaler TF1 og France 2..

<W lemma="TF1" msd="XR">TF1</W>

..Nøgleordene til en ny fremtid er .. Dodge Epic, Intrepid og Viper RT/10, Ford Probe..

<W lemma="RT/10" msd="XR">RT/10</W>

..at han i forårets tre testkampe ikke disponerer over de olympiske spillere fra U-21 holdet..

<W lemma="U-21" msd="XR">U-21</W>

..det var et menneske, som nødlandede passagerflyet SK751 3. juledag, ikke teknologien..

<W lemma="SK751" msd="XR">SK751</W>

..derefter bestilte vi en flaske gin-lime .. Vi gjorde os lækre sammen og bestilte en taxa til kl. 21:00..

<W lemma="21:00" msd="AC---U=--">21:00</W>

..den 15/6 bragte DR-TV en udsendelse i serien "Den offentlige mening..

<W lemma="15/6" msd="AO---U=--">15/6</W>

5.8.5 Symboler

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Symbol S									

Som nævnt i afsnit 4.1 om interpunktionstegn og symboler har følgende tegn fået tildelt en analyse som symbol ('XS') i PAROLE-korpusset: dollartegnet ($), procenttegnet (%), stjernen (*), plustegnet (+), lighedstegnet (=), paragraftegnet (§) og gradtegnet (°). Minustegnet (-) falder sammen med bindestregen.

..Venus er et meget ugæstfrit sted med en overfladetemperatur på omkring 500º C..

<W lemma="°" msd="XS">°</W>

..hans billeder, som Lousiana netop har erhvervet et af, koster nemt op mod 250.000$..

[50] Dvs. punktum, komma, bindestreg, skråstreg eller kolon.

<W lemma="$" msd="XS">$</W>

6. Tekstfejl og sproglige afvigelser

CatGram	SsCatGram	3	4	5	6	7	8	9	10	11
Residual X	Other X									

'Tekstfejl' er den generelle betegnelse, der her anvendes til de ordformer, som af forskellige årsager ikke får tildelt en af de andre "rigtige" morfosyntaktiske analyser – enten af DAN-TWOL-analyseapparatet eller af korpustaggerne under selve korpustaggingen. For disse ordformer er den morfosyntaktiske analyse blot angivet som 'XX'. Dette afsnit beskriver behandlingen af følgende tre typer af ordformer, der har med tekstfejl og sproglige afvigelser at gøre: (*i*) (sprogligt korrekte) ordformer, som ikke kunne analyseres af DAN-TWOL-analyseapparatet, (*ii*) sproglige afvigelser, som ikke blev accepteret under korpustaggingen samt (*iii*) sproglige afvigelser, som blev accepteret i PAROLE-korpusset, og som derfor har fået tildelt en "rigtig" morfosyntaktisk analyse. Ordformer tilhørende typerne (*i*) og (*ii*) har derimod fået tildelt analysen 'XX'.

6.1 Ordformer, der ikke kunne tildeles en analyse af DAN-TWOL

6.1.1 Udeladt fælles orddel

Når to eller flere sammensatte eller afledte ord med fælles orddel optræder sideordnet i et paratagme, er det ifølge RO96, §63.2 korrekt at anvende bindestregen til angive en udeladt orddel (som f.eks. *sundheds-* i *sundheds- og ernæringstilstanden*). Da det dog ikke var muligt for DAN-TWOL at analysere disse ordformer, har de alle fået tildelt 'XX'-analysen, naturligvis ikke fordi der er tale om egentlige tekstfejl, men snarere om analysefejl.

..dels intern uenighed om foreningens struktur. Og almindelig sym- og antipati for fuld udblæsning..
<W lemma="sym-" msd="XX">sym-</W>
..så er der jo ingen grund til at skelne mellem A- eller B-licens..
<W lemma="A-" msd="XX">A-</W>
..og har vundet et hav af store mesterskaber, bl.a. tre VM- og én OL-guldmedalje..
<W lemma="VM-" msd="XX">VM-</W>
..især inden for føde- og drikkevarer, der omfatter såvel sukker- som spritfabrikker...
<W lemma="føde-" msd="XX">føde-</W> .. <W lemma="sukker-" msd="XX">sukker-</W>

6.1.2 Andre ordformer, der ikke har fået tildelt en analyse af DAN-TWOL

Følgende eksempler viser andre typer af ordformer, der har fået tildelt 'XX'-analysen, fordi deres morfologi og ordklasse var uigennemskuelige, enten pga. en ordinddelingsfejl forårsaget af DAN-TWOL-tokeniseren (jf. afsnit 4.1 om interpunktionstegn og symboler), eller fordi den pågældende ordform ikke kunne analyseres af andre årsager (typisk slåfejl eller ordinddelingsfejl i originalteksten).

..de somaliere, der støtter en amerikansk militær aktion i det hungerramte land. (Foto:AP)..
<W lemma="Foto:AP" msd="XX">Foto:AP</W>
..HORSENSPolitiet i Horsens er rasende over Vestre Landsrets løsladelse af to rockere..
<W lemma="HORSENSPolitiet" msd="XX">HORSENSPolitiet</W>
..køber har bestemt sig til et hus/ejerlejlighed til f.eks. 650.000 kr...
<W lemma="hus/ejerlejlighed" msd="XX">hus/ejerlejlighed</W>
..selvom vi har forberedt os, sker det hele nu hu-hej-vilde dyr..

<W lemma="hu-hej-vilde" msd="XX">hu-hej-vilde</W>
..udvalgets indstilling kan indebære merudgifter på op til "x-mill." kr...
<W lemma="x-mill." msd="XX">x-mill.</W>

6.2 Ikke-accepterede sproglige afvigelser

Dette afsnit om ikke-accepterede sproglige afvigelser er inddelt i to dele: Den første del omhandler ikke-accepterede sproglige afvigelser, hvor den pågældende ordform **ikke eksisterer** i det danske sprog. Anden del omhandler ikke-accepterede sproglige afvigelser, hvor den pågældende ordform **eksisterer** i det danske sprog, men af andre årsager alligevel ikke blev accepteret under korpustaggingen.

6.2.1 Ikke-eksisterende ordformer

Nogle slå-, stave-, orddelings- og ordinddelingsfejl osv. i PAROLE-korpusset er resulteret i en ordform, der ikke eksisterer i det danske sprog. Hvis korpustaggerne har vurderet, at der ikke var tale om et mere eller mindre "moderne" dansk slangudtryk, låneord, proprium osv., eller evt. et udenlandsk ord, har ordformen fået tildelt en analyse som tekstfejl ('XX'). Her drejer det sig hovedsageligt om stave- og slåfejl samt forkert brug af bindebogstaver og konsonantfordobling.

6.2.1.1 Stave- og slåfejl, som resulterer i en ikke-eksisterende ordform

Under korpustaggingen er stave- eller slåfejl af samme type som i de følgende eksempler ikke blevet accepteret, da de på forskellige måder afviger fra Retskrivningsordbogens stavenormeringer: (Jf. dog også afsnit 6.3.4 om sammenskrivninger og afsnit 6.3.5 om udenlandske sted- og indbyggerbetegnelser.)

..de mener, at Folkekongressen skal give præsidenten diktaroriske befølelser..
<W lemma="diktaroriske" msd="XX">diktaroriske</W> <W lemma="befølelser" msd="XX">befølelser</W>
..han udsendte en pressemeddelse netop i anledning af Bodil'en..
<W lemma="pressemeddelse" msd="XX">pressemeddelse</W>
..kunstnerkolleger har tidligt påskønnet hans skulpturers plastik og ynde med Akademiets guldmedallie..
<W lemma="guldmedallie" msd="XX">guldmedallie</W>
..de to banker tilhører i parantes bemærket Brøndbys øverste sponsorluftlag..
<W lemma="parantes" msd="XX">parantes</W>
..dermed var Andelsbanken .. reelt sat uden for beslutningerne i toppen af Unibank-hirakiet..
<W lemma="Unibank-hirakiet" msd="XX">Unibank-hirakiet</W>
..og - på korter ellere længere sigt – drager Danmark ind i et tættere samarbejde med udlandet..
<W lemma="korter" msd="XX">korter</W> <W lemma="ellere" msd="XX">ellere</W>

Dette gælder også (gennemskuelige) stave- eller slåfejl i danske proprier og ord af udenlandsk oprindelse. Jf. RO96, §67).

..hvis Martianne Jelved skal voldføre sit skeptiske bagland..
<W lemma="Martianne" msd="XX">Martianne</W>
..partiets politiske ordfører, Mogens Lykketoft, nærmest roste Polul Schlüter..
<W lemma="Polul" msd="XX">Polul</W>
..helt forrygende er Koko Taylors udlægning af "Merry, Merry Christmans"..
<W lemma="Christmans" msd="XX">Christmans</W>
..han .. forsyner os .. med lange lister om aviser, cigaretter og fødevarer," oplyser den danske charge d'affairs..
<W lemma="charge_d'affairs" msd="XX">charge_d'affairs</W>

6.2.1.2 Mangel på bindebogstav eller forkert konsonantfordobling

En væsentlig del af de ikke-eksisterende ordformer i korpusset er resultatet af enten (*i*) manglen på et eller flere korrekte bindebogstaver (enten *e, s* eller *er*) i sammensætninger eller afledninger, eller (*ii*) en forkert dobbeltskrivning eller enkeltskrivning af konsonanter i bøjningsformer (RO96, §10). Disse ordformer er alle markeret som tekstfejl i PAROLE-korpusset.

..Ekstra Bladets ansvarhavende chefredaktør, Bent Falbert: "Det med Cosmopolitan lyder lidt mærkeligt..
<W lemma="ansvarhavende" msd="XX">ansvarhavende</W>
..i oktober 1991 underskrev ministerpræsident Hun Sen .. en freds- og hensigterklæring..
<W lemma="hensigterklæring" msd="XX">hensigterklæring</W>
..straks efter bryllupet kastede prins Andrew sig over sin militære løbebane..
<W lemma="bryllupet" msd="XX">bryllupet</W>
..Isafold satte efter sammenstøddet øjelikkelig [sic] en gummibåd med påhængsmotor i vandet..
<W lemma="sammenstøddet" msd="XX">sammenstøddet</W>
..med ansvar for såvel Danish Paper Packaging som Rackmanns Fabriker..
<W lemma="Fabriker" msd="XX">Fabriker</W>

6.2.2 Eksisterende ordformer

Det er straks et mere interessant — men også mere problematisk — spørgsmål, hvordan man behandler de ordformer i teksten, der svarer til et eksisterende dansk ord, men alligevel kan opfattes som værende en sprogfejl i den pågældende syntaktiske kontekst. Følgende eksempler viser hvilke forskellige kategorier af grammatiske (dvs. syntaksbaserede) slå- og stavefejl er markeret som tekstfejl i PAROLE-korpusset.

6.2.2.1 Særskrivning

Ordformer, der opstår af en ukorrekt særskrivning ifølge RO96, §18-§19, markeres med 'XX'-analysen, også selvom de enkelte orddele svarer til ellers korrekte danske ord. Dette gælder både (*i*) når særskrivningen øjensynligt er resultatet af en slåfejl, og (*ii*) når særskrivningen er et typisk eksempel på en mere udbredt usikkerhed omkring orddeling i det danske skriftsprog. Ukorrekte særskrivninger er ikke blevet accepteret, da de individuelle orddele ikke uproblematisk kunne indføres i DAN-TWOL-leksikonet. Hyppigt forekommende, ukorrekte sammenskrivninger er derimod i højere grad blevet accepteret i PAROLE-korpusset, da mange af disse ordformer problemfrit er indført i DAN-TWOL-leksikonet (jf. afsnit 6.3.4 om sammenskrivning).

..lidt nede af den isglatte vej står et folkevognsrug brød, der har fragtet en håndfuld norske skibumser til St. Anton..
<W lemma="folkevognsrug" msd="XX">folkevognsrug</W> <W lemma="brød" msd="XX">brød</W>
..dansk, engelsk, historie, fransk, spansk, italiensk, portugisisk/brasiliansk, informationsviden skab..
<W lemma="informationsviden" msd="XX">informationsviden</W> <W lemma="skab" msd="XX">skab</W>
..at give ophold til personer som dig, der ikke kan indordne sig de aller simpleste adfærdsnormer..
<W lemma="aller" msd="XX">aller</W> <W lemma="simpleste" msd="XX">simpleste</W>
..det er nemlig i følge amatørreglerne simpelthen forbudt en spiller at reklamere med sit navn eller billede..
<W lemma="i" msd="XX">i</W> <W lemma="følge" msd="XX">følge</W>
..men med mindre manden søger lægebehandling, så er der nok en risiko for, at han får en infektion i såret..
<W lemma="med" msd="XX">med</W> <W lemma="mindre" msd="XX">mindre</W>

6.2.2.2 Et andet lemma

Nogle få slåfejl i korpusteksterne er resulteret i en ordform, der tilhører et andet lemma, end det, der efter al sandsynlighed var ment af tekstens forfatter. Disse ordformer har (så vidt det har været muligt at identificere dem) også fået tildelt 'XX'-analysen i PAROLE-korpusset:

..det værste er at blive forfulgt at teenage-piger. Hvis du bare skal skrive et par autografer, er det fint..
<W lemma="at" msd="XX">at</W>
..samtlige folk, der har pladser oppe bag dem, skal defilere lige fordi den første række..
<W lemma="fordi" msd="XX">fordi</W>
..indklædt i aluminium både på under- og oversiden og formet som en flot fly-vinge, der næste svæver frit i luften..
<W lemma="næste" msd="XX">næste</W>

6.2.2.3 Forkert bøjningsform

En anden gruppe slåfejl er resulteret i en ordform, hvis lemma er korrekt i den pågældende kontekst, men bøjningsformen er forkert (pga. kongruensuoverensstemmelse). I de tilfælde, hvor korpustaggerne har været opmærksomme på disse uoverensstemmelser, har disse ordformer fået tildelt 'XX'-analysen.

..hvad angår Danmarks økonomiske udbyttet af projektdeltagelse i de forskellige EF-forskningsprogrammer..
<W lemma="udbyttet" msd="XX">udbyttet</W>
..yderligere 17 pct. af udgifter går til finansiering af de store teaterscener, egnsteatre og billetstøtte i København..
<W lemma="udgifter" msd="XX">udgifter</W>
..når Henrik Svarrer/Marlene Thomsen møde vores normalt bedste konstellation Thomas Lund/Pernille Dupont..
<W lemma="møde" msd="XX">møde</W>
..af og til er de lige ved at snubler over det store forbillede, Tommy Kenter..
<W lemma="snubler" msd="XX">snubler</W>
..det var imponerede at se hvor hurtigt og professionelt soldaterne gik i gang med førstehjælp..
<W lemma="imponerede" msd="XX">imponerede</W>
..som om solen ikke skinnende dér, som om jeg ikke legede og havde det rart..
<W lemma="skinnende" msd="XX">skinnende</W>
..en samiske kunstner Nils-Aslak Valkeapää modtog 1disk [sic] Råds Litteraturpris..
<W lemma="samiske" msd="XX">samiske</W>
..Yves Saint Laurent meddelte at man desværre ikke kunne bruge fr. Campbell på grund af den meget presseomtale..
<W lemma="meget" msd="XX">meget</W>
..afghanerne er en fantastisk folk. Smukke, stærke og ikke til at knække..
<W lemma="en" msd="XX">en</W>
..Han svenske kone gennem 10 år Christina var rejst med parrets fælles barn Christopher på syv år..
<W lemma="Han" msd="XX">Han</W>
..måske hænger det sammen med, at vi har det koldere, sidder mere ind og bliver deprimerede om vinteren..
<W lemma="ind" msd="XX">ind</W>

6.2.2.4 Et ord for meget

Nogle få slåfejl i PAROLE-korpusset er resulteret i tilstedeværelsen af et overskydende tekstord i korpusteksterne. I de tilfælde, hvor korpustaggerne har været opmærksomme på problemet, blev 'XX'-analysen tildelt det overskydende ord. Slåfejl, der resulterer i, at der mangler et ord, er langt sjældnere i PAROLE-korpusset og blev ikke markeret eksplicit under korpustaggingen.

..den blev i går erklæret konkurs ligesom sin ejer, Klaus Riskær Pedersen, efter begæring af af bobestyrer..
<W lemma="af" msd="XX">af</W>
..som leder af Filmskolen i Danmark har har således været med til at udvirke, at instruktøren Gert Fredholm..
<W lemma="har" msd="XX">har</W>
..specielt i 1. sæt varierede jeg mit spil og godt..
<W lemma="og" msd="XX">og</W>
..flere CD'er mener, at at Den Danske Forening og CD er uforenelige størrelser..
<W lemma="at" msd="XX">at</W>

6.3 Accepterede sproglige afvigelser

Selvom afsnittet ovenfor gennemgår forskellige stave- og slåfejl, der alle har fået tildelt 'XX'-analysen, indeholder PAROLE-korpusset også andre, hyppigt forekommende sproglige afvigelser, der er blevet accepteret under korpustaggingen, selvom de afviger fra Retskrivningsordbogens stavenormeringer og retskrivningsregler. Dette afsnit omhandler således de afvigelser fra Retskrivningsordbogens forskrifter, der ikke er blevet markeret som tekstfejl i PAROLE-korpusset, men derimod har fået tildelt en "rigtig" morfosyntaktisk analyse.

6.3.1 Ord, som er ukendte for DAN-TWOL

Et mindre antal tekstord i PAROLE-korpusset har ikke fået tildelt en morfosyntaktisk analyse af DAN-TWOL, fordi de af forskellige årsager ikke optræder i DAN-TWOLs leksikon. I de fleste tilfælde var der tale om enten (*i*) mere eller mindre etablerede (med henvisning til DDO's korpus) "nyere" udenlandske låneord, eller (*ii*) danske sammensætninger eller afledninger, som ikke kunne håndteres af DAN-TWOL, men alligevel blev opfattet som værende acceptable under korpustaggingen. Mange af disse (for DAN-TWOL) "ukendte" ord har i efterrediggeringsfasen derfor fået tildelt en analyse på baggrund af deres morfologiske bøjningsform og syntaktiske kontekst i korpusset (Jf. også afsnit 5.1.8 om udenlandske ord).

..manden, der ville sidde med en pint og "The Sun" på den lokale pub..
<W lemma="pint" msd="NCCSU==I">pint</W>
..hans beslutning om at droppe sit pitstop er typisk for hans dømmekraft..
<W lemma="pitstop" msd="NCNSU==I">pitstop</W>
..nye "afsløringer" om Dianas mere-end-venskab med en anden af husets venner..
<W lemma="mere-end-venskab" msd="NCNSU==I">mere-end-venskab</W>
..men det kunne have været bedre, hvis min puttning havde fungeret bedre..
<W lemma="puttning" msd="NCCSU==I">puttning</W>
..i samme øjeblik bliver han en "refusnik" - en person der ønsker at gå i landflygtighed..
<W lemma="refusenik" msd=""NCCSU==I">refusenik</W>
..Pia Kjærsgård, der ikke vil finde sig i at blive belært og "nurset" af Kirsten Jacobsen..
<W lemma="nurse" msd="VAPA=S[CN]I[ARU]-U">nurset</W>
..en række flotte ny-ekspressionistiske lærreder, hvor det eksotiske brydes med det moderne..
<W lemma="ny-ekspressionistisk" msd="ANP[CN]PU=[DI]U">ny-ekspressionistiske</W>
.."Hav en fortsat god dag, tusind tak, mange tak, på gensyn" er f.eks. bare en mediumlang sætning..
<W lemma="mediumlang" msd="ANPCSU=IU">mediumlang</W>
..intervalviskere, bagrudevisker, triptæller, fem gear, højdejustérbare S-seler og jernstivere i dørene..
<W lemma="højdejusterbar" msd="ANP[CN]PU=[DI]U">højdejustérbare</W>
..det feminine og romantiske Gug'ske præg er også tydeligt i lejligheden..
<W lemma="Gug'sk" msd="ANP[CN]SU=DU">Gug'ske</W>

6.3.2 Interpunktionstegn og symboler

Alle forekomster af ukorrekt anvendelse af interpunktionstegn ifølge reglerne i RO96, §40 - §66 — som oftest kommafejl — blev accepteret uden kommentar i PAROLE-korpusset. Bindestregens anvendelse i andre typer af sammensætninger end dem, der eksplicit nævnes i RO96, §63, blev også accepteret under korpustaggingen, hvor disse ordformer blev behandlet uden særlig hensyn til bindestregens tilstedeværelse.

..varerummet .. er ikke et bagagerum på samme måde, som et aflåseligt bagagerum på en almindelig personbil..
<W lemma="," msd="XP">,</W>
..filmen, der altså er en engelsk-dansk co-produktion bliver afviklet på en blanding af engelsk og dansk..
<W lemma="co-produktion" msd="NCCSU==I">co-produktion</W>
<W lemma="blive" msd="VADR=----A-">bliver</W>

..det er første gang man prøver at skildre ulands-problemer som en action-serie..
<W lemma="ulands-problem" msd="NCNPU==I">ulands-problemer</W>
..dengang kom vi på samme fodbold-hold i Vedbæk Boldklub..
<W lemma="fodbold-hold" msd="NCNSU==I">fodbold-hold</W>

Accenttegnet 'accent aigu' (´) anvendes ifølge RO96, §5 typisk til at forebygge misforståelser eller fejllæsninger i skriftsproget. I PAROLE-korpusteksterne accepteres dets tilstedeværelse alle steder (og kræves ingen steder).

..røveren [var] oppe i stueetagen, hvor han tvang manden ned at ligge på entrégulvet..
<W lemma="entrégulv" msd="NCNSU==D">entrégulvet</W>
..røveren [trak] et sæt håndjern frem og lænkede parret til en radiator i entreen..
<W lemma="entre" msd="NCCSU==D">entreen</W>

Anvendelsen af apostrofen foreskrives i RO96, §6 til at markere grænsen mellem et ords stamme og dets endelser i f.eks. forkortelser uden forkortelsespunktum, taltegn og symboler, proprier, fremmedord med stumme konsonantendelser, visse genitivendelser samt ved udeladelsen af bogstaver. I PAROLE-korpusteksterne er apostrofens tilstedeværelse accepteret i alle ovennævnte tilfælde, men også i andre sammenhæng, som f.eks. i forbindelse med udeladelse af århundredbetegnelsen i årstal (RO96, §6.7). Dens udeladelse er også accepteret efter taltegn og symboler (RO96, §6.2).

..de tre første episoder foregår i forskellige perioder - henholdsvis 30erne, 60erne og i nutiden..
<W lemma="30er" msd="NCCPU==D">30erne</W> .. <W lemma="60er" msd="NCCPU==D">60erne</W>
..han lod sig i '83 pensionere for at fuldføre sin livsopgave - at skrive dansk revys historie..
<W lemma="'83" msd="AC---U=--">'83</W>
..HVEM KØBER PLADERNE? Fans'ne. Hvordan ser de ud, de 160.000, der har købt en Sko og Torp plade?..
<W lemma="fan" msd="NCCPU==D">Fans'ne</W>
..et virvar af psykologiske moraler, systemkritiske spørgsmål og fremtidsspående epos'er..
<W lemma="epos" msd="NCNPU==I">epos'er</W>
..desværre er mordet på Ole Ernst nødvendig for at få spænding på krimi'en..
<W lemma="krimi" msd="NCCSU==D">krimi'en</W>

Når apostrofen anvendes ved udeladelsen af et eller flere bogstaver (ifølge reglerne i RO96, §6.7), angives det fuldstændige lemma stadig i SGML-lemma-attributtet i PAROLE-korpusset (jf. eksemplerne nedenfor). Udeladelse af bogstaver uden anvendelse af apostrof accepteres dog også i nogle tilfælde i PAROLE-korpusset, samt et mindre antal andre hyppige stavevarianter, der ikke er opført i RO96.

..jo længere vi kører mod øst, jo goldere bli'r landskabet ..
<W lemma="blive" msd="VADR=----A-">bli'r</W>
..selv om vi fra starten havde i baghovedet, at vi burde ku' trille den kamp hjem..
<W lemma="kunne" msd="VAF-=----A-">ku'</W>
..hvis du skal have klar besked, blir du nødt til at tale med din læge. Du ska ha foretaget en underlivsundersøgelse..
<W lemma="blive" msd="VADR=----A-">blir</W>
<W lemma="skulle" msd="VADR=----A-">ska</W>
<W lemma="have" msd="VAF-=----A-">ha</W>
.."Hva!" råbte de som med én mund, "arrangerer du ikke keglerne når du er keglerejser?..
<W lemma="hvad" msd="PT-[CN]SU--U">Hva</W>
..det er sgu nok en invalidevogn lissom sidst, griner Sally og tager sin smøgæske frem..
<W lemma="lissom" msd="RGU">lissom</W>
..der kan være tale om brok eller vandsvulst. Men der kan osse være tale om en ondartet svulst..
<W lemma="osse" msd="RGU">osse</W>

6.3.3 Forkortelser

I PAROLE-korpusset accepteres forkortelser i (næsten) alle deres varianter — dvs. med og uden forkortelsespunktum(mer) eller apostrof (ved genitiv) — også selvom de afviger fra reglerne i RO96, §6.1, §14, §21.3 og §42 samt eksemplerne i Eriksen og Hamburger, 1988.

..varetægtsfængsling, isolation og andre indgreb (såsom ransagning, aflytning, anvendelse af politiagenter o.s.v.)..
<W lemma="og_så_videre" msd="RGU">o.s.v.</W>
..gummilisterne fryser fast til metallet og revner, når du bruger vold for at åbne døre, motorklap osv. ..
<W lemma="og_så_videre" msd="RGU">osv.</W>
..som sender 16 kanalers vellyd (jazz, klassisk, underholdning, kultur osv) ud sammen med TV-kanalerne..
<W lemma="og_så_videre" msd="RGU">osv</W>

6.3.4 Sammenskrivning

I modsætning til ukorrekt særskrivning af ord (jf. afsnit 6.2.2.1 om særskriving) er nogle ukorrekte sammenskrivning af ord blevet accepteret i PAROLE-korpusset. Dette er dels pga. deres forholdsvis større hyppighed i korpusteksterne[51], men også dels fordi mange af dem allerede problemfrit er indført i DAN-TWOLs leksikon. Følgende 18 ukorrekt sammenskrevne ordformer er således blevet accepteret i PAROLE-korpustektsterne: *afsted, altfor, altimens, fornylig, forresten, ialt, idag, igang, igår, ihvertfald, imorgen, iorden, iøvrigt, ombord, omend, overbord, tilsidst* og *tilstede.*

..partiet er med ialt 3,8 procent af stemmerne nede på syv mandater..
<W lemma="ialt" msd="RGU">ialt</W>
..eleverne er meget kritiske idag. De gider ikke læse jammerbøger..
<W lemma="idag" msd="RGU">idag</W>
..han er sikkert flintret afsted på en cykel ti minutter efter, at han var landet..
<W lemma="afsted" msd="RGU">afsted</W>
..Forresten burde I ikke opholde jer på sporet, det er forbudt..
<W lemma="forresten" msd="RGU">Forresten</W>

I denne forbindelse skal det igen nævnes, at visse sammenskrivninger af sammensatte præpositioner (**adverbium + præposition**) også er blevet accepteret, selvom præpositionen har en styrelse i sætningen (jf. afsnit 5.4.1 om præpositioner). Følgende "sammenskrevne" præpositioner er således blevet accepteret i PAROLE-korpusset: *indenfor, udenfor, henad, nedover, opad, overfor, udfra* samt *udover.*

..men, hvis der ikke straks gribes ind overfor denne praksis, vil eksperterne ikke udelukke en mulig epidemi..
<W lemma="overfor" msd="SP">overfor</W>
..de store "trailer-rock" tournéer er blevet et begreb indenfor musikindustrien..
<W lemma="indenfor" msd="SP">indenfor</W>

6.3.5 Udenlandske sted- og indbyggerbetegnelser

[51] Samt deres hyppighed i DDO's tekstkorpus. Antallet af forekomster af ovenstående ordformer i DDO's tekstkorpus (på ca. 40 mio. løbende ord) fordeler sig som vist nedenfor. Der tages dog forbehold for nedenstående tal, da de særskrevne ordformer naturligvis kan optræde i andre sammenhæng i DDO's korpus, f.eks. *han vil gøre alt for hende* og *for resten af befolkningen er det godt.* Dette er bl.a. årsagen til, at f.eks. *alt for* og *for resten* ikke samles på forhånd af DAN-TWOL-tokeniseren (jf. note 19).

afsted/1.411 af sted/2.801 (33,5%)	altfor/151 alt for/5.747 (2,6%)	altimens/30 alt imens/185 (14%)
fornylig/349 for nylig/1.401 (19,9%)	forresten/618 for resten/724 (46,1%)	ialt/971 i alt/4.155 (18,9%)
idag/1.190 i dag/23.078 (4,9%)	igang/1.119 i gang/7.577 (12,9%)	igår/205 i går/9.259 (2,2%)
ihvertfald/329 i hvert fald/10.996 (2,9%)	imorgen/143 i morgen/3.655 (3,8%)	iorden/26 i orden/2.213 (1,1%)
iøvrigt/1.668 i øvrigt/6.589 (20,2%)	ombord/447 om bord/1.042 (30%)	omend/716 om end/444 (61,7%)
overbord/35 over bord/101 (25,7%)	tilsidst/199 til sidst/4.206 (4,5%)	tilstede/212 til stede/1.695 (11,1%)

På trods af stavenormeringerne i RO96 er lidt variation i stavning af udenlandske landenavne og indbyggerbetegnelser blevet accepteret i PAROLE-korpusset:

..de tre nye selvstændige nationer, Estland, Letland og Lithauen..
<W lemma="Lithauen" msd="NP--U==-">Lithauen</W>
..de tidligere Sovjetrepublikker Uzbekistan og Kazakstan..
<W lemma="Uzbekistan" msd="NP—U==-">Uzbekistan</W>
<W lemma="Kazakstan" msd="NP—U==-">Kazakstan</W>
..fra en yderst primitiv start i slutningen af 1700-årene i Sydtyskland og det tjekkiske Böhmen..
<W lemma="Böhmen" msd="NP--U==-">Böhmen</W>
..derudover var der en række passagerer, blandt andet en amerikaner og tre malayer..
<W lemma="malay" msd="NCCPU==I">malayer</W>
..at forhindre en hvilkensomhelst cambodianer i at organisere et tribunal..
<W lemma="cambodianer" msd="NCCSU==I">cambodianer</W>
..den czechoslovakiske præsident Vaclav Havel..
<W lemma="czechoslovakisk" msd="ANP[CN]SU=DU">czechoslovakiske</W>
..at fragte den ene af sine joller til den katalanske hovedstad..
<W lemma="katalansk" msd="ANP[CN]SU=DU">katalanske</W>

6.3.6 Danske ord af udenlandsk oprindelse

Til sidst skal det nævnes, at PAROLE-korpusset også indeholder et par ukorrekte danske stavevarianter, der snarere tilnærmer sig en oprindelig, udenlandsk stavemåde end følger stavenormeringerne i RO96 — oftest optræder de i mere eller mindre "fagsproglige" tekster. I nogle tilfælde, især hvor der også fandtes andre hyppige belæg for disse stavevarianter i DDO's korpus, er disse stavevarianter også blevet accepteret under korpustaggingen.

..**anvendelsen af alternative drivmidler i biler, såsom bio-ethanol, methanol og rapsolie..**
<W lemma="bio-ethanol" msd="NCNSU==I">bio-ethanol</W>
<W lemma="methanol" msd="NCNSU==I">methanol</W>
..filmen, der altså er en engelsk-dansk co-produktion bliver afviklet på en blanding af engelsk og dansk..
<W lemma="co-produktion" msd="NCCSU==I">co-produktion</W>
..bl.a. kan der fra asfalt udvaskes phenoler og andre giftige organiske forbindelser..
<W lemma="phenol" msd="NCNPU==I">phenoler</W>
..**i henhold til den omtalte artikel har ingen tv-stationer hidtil kunnet sende fodbold i UEFA-regie..**
<W lemma="UEFA-regie" msd="NCNSU==I">UEFA-regie</W>
..det er iøvrigt slet ikke selve tarmen der er betændt men blot et lille vedhæng som kaldes appendix..
<W lemma="appendix" msd="NCNSU==I">appendix</W>
..mygblomst er en sjælden orchide. Den vokser kun i stærkt kalkholdige moser..
<W lemma="orchide" msd="NCCSU==I">orchide</W>
..på de fabrikker vi har i dag, omdannes råvarer til nye produkter. Soyabønner bliver til salatolie..
<W lemma="soyabønne" msd="NCCPU==I">Soyabønner</W>
..Pludselig genkendte han temaet. Det var Begin the Beguine. Glenn Millers schlager fra krigen..
<W lemma="schlager" msd="NCCSU==I">schlager</W>

7. Litteraturliste

Allan, Robin, Philip Holmes og Tom Lundskær-Nielsen (1995), *Danish – A Comprehensive Grammar*, Routledge.

Arndt, Hans (1996), *Grammatisk Analyse*, Institut for Lingvistik, Aarhus Universitet.

Becker-Christensen, Christian et al (1996), *Politikens Store Nye Nudansk Ordbog*, Politikens Forlag.

Becker-Christensen, Christian og Peter Widell (1995), *Politikens Nudansk Grammatik*, Politikens Forlag.

Bilgram, Thomas og Hans Arndt (1993), "Automatisk morfologisk analyse af danske tekster", i Mette Kunøe og Erik Vive Larsen (red.) *4. Møde om Udforskning af Dansk Sprog*, Aarhus Universitet.

Bilgram, Thomas (1994), *Computerstyret analyse af dansk*, specialeopgave, Institut for Lingvistik, Aarhus Universitet.

Bilgram, Thomas og Britt Keson (1998), "The Construction of a Danish Tagged Corpus", i *NODALIDA '98 Proceedings*.

Braasch, Anna og Ole Norling-Christensen (1997), "En trækbaseret beskrivelse af dansk bøjningsmorfologi", i Tom Brøndsted & Inger Lytje (red.) *Sprog og Multimedier 1997*, Aalborg Universitetsforlag.

Dansk Sprognævn (1986), *Retskrivningsordbogen*, (RO86), Gyldendal Forlag.

Dansk Sprognævn (1996), *Retskrivningsordbogen*, (RO96), Aschehoug Forlag, 2. udgave.

Diderichsen, Paul (1968), *Elementær Dansk Grammatik*, Gyldendal, 3. udgave, 3.oplag.

Ejerhed, Eva et al (1992), *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*, Department of Linguistics, University of Umeå.

Eriksen, Jørgen og Arne Hamburger (1988), *Forkortelser i hverdagen*, Dansk Sprognævn, Gyldeldal.

Galberg Jacobsen, Henrik og Henrik Skyum-Nielsen (1996), *Dansk Sprog*, Det Schønbergske Forlag.

Galberg Jacobsen, Henrik (red.) (1996), *Grammatisk talt*, Dansk Sprognævns skrifter.

Hansen, Erik (1975), "Noget om Intet" i *At Færdes i Sproget*, Gyldendal, s. 113 - 137.

Hansen, Erik (1992), *Dæmonernes Port*, Hans Reitzels Forlag A/S.

Hansen, Erik (1998), "Fra Stort til småt", kronik i *Politiken* (22. marts 1998).

Ide, N., D. Durand, G. Priest-Dorman og J. Veronis (1995), *MULTEXT: Corpus Encoding Standard*, LRE Project 62-050, CNRS.

Jensen, Per Anker (1985), *Principper for grammatisk analyse*, Nyt Nordisk Forlag, Arnold Busck.

Juul-Jensen, H., J. Ernst-Hansen, Holger Hansen og Holger Sandvad (1919), *Ordbog over det danske sprog*, Gyldendalske Boghandel, Nordisk Forlag.

Karlsson, F., J. Voutilainen, J. Heikillö and A. Anttila (red.) (1994), *Constraint Grammar: a Language independent System for Parsing Unrestricted Text*, Berlin/New York, Mouton de Gruyter.

Keson, Britt (1999), "Morfosyntaktisk tagging af danske tekster", i *7. Møde om Udforskning af Dansk Sprog*, Aarhus Universitet.

Koskenniemi, Kimmo (1983), *Two-level morphology. A General Computational Model for Word-form Production and Generation*. Publication No. 11, Dept. of Linguistics, University of Helsinki.

LE-PAROLE (http://www-tei.uic.edu/orgs/tei/app/le02.hmtl).

LE-PAROLE Project Summary (http://www2.echo.lu/langeng/en/le2/le-parole/summary.html).

Monachini, M. og Nicoletta Calzolari (1996), *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, EAGLES Document EAG-LSG/IP-T4.6/CSG-T3.2, EAGLES.

Norling-Christensen, Ole (1996), *Design and Composition of Reusable Harmonzed Written Language Reference Corpora for European Languages* MLAP PAROLE 63-386 WP 4.1.1, Det Danske Sprog- og Litteraturselskab (DSL).

Norling-Christensen, Ole og Jørg Asmussen (1999), "The Corpus of the Danish Dictionary", i *Lexikos* (8), Afrilex, Stellenbosch.

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

Sperberg-McQueen, C.M., og L. Burnard (1994), *Guidelines for Electronic Text Encoding and Interchange* (TEI 3), Chicago, ACH, ACL, ALLC.

Vinterberg, Hermann, og C.A. Bodelsen (1990), *Dansk-engelsk Ordbog*, Gyldendal, 3. udgave.

Volz, Norbert og Susanne Lenz (1996), *Multilingual Corpus Tagset Specifications*, MLAP PAROLE 63 386 WP 4.1.4a, IDS, Mannheim.

Wynne, Martin (1996), *A Post-Editor's Guide to CLAWS7 Tagging*, (http://www.comp.lancs.ac.uk/computing/users/eiamjw/claws/claws7.html).

8. Appendiks

8.1 Fordeling af tekstord og ordtyper på ordklasser

PAROLE ORDKLASSER	tekstord ('tokens')	ordtyper ('types')[52]
SUBSTANTIVER (N)	**66.906**	**24.465**
appellativer (NC)	52.645	18.960
proprier (NP)	14.261	5.505
ADJEKTIVER (A)	**27.900**	**6.292**
'almindelige' adjektiver (AN)	23.355	5.340
kardinale adjektiver (AC)	4.125	895
ordinale adjektiver (AO)	420	57
KONJUNKTIONER (C)	**15.549**	**50**
underordningskonjunktioner (CS)	5.730	41
sideordningskonjunktioner (CC)	9.819	9
INTERJEKTIONER (I)	**259**	**54**
PRONOMINER (P)	**33.493**	**95**
reciprokke pronominer (PC)	84	2
demonstrative pronominer (PD)	6.945	15
ubestemte pronominer (PI)	9.512	29
possessive pronominer (PO)	2.433	20
personlige pronominer (PP)	13.965	21
interrogative pronominer (PT)	554	8
ADVERBIER (RG)	**19.017**	**498**
PRÆPOSITIONER (SP)	**30.927**	**69**
UNIQUE (U)	**9.037**	**3**
VERBER (V)	**45.398**	**6.069**
indikativ (VAD/VED)	29.200	2.320
infinitiv (VAF/VEF)	8.696	1.538
gerundium (VAG)	56	41
imperativ (VAM)	437	144
præteritum participium (VAPA/VEPA)	6.342	1.686
præsens participium (VAPR)	667	340
RESIDUAL (X)	**42.114**	**1.185**
forkortelser (XA)	146	62
udenlandske ord (XF)	295	201
interpunktionstegn (XP)	40.391	13
formler (XR)	47	40
symboler (XS)	169	7
tekstfejl mm. (XX)	1.066	862
I ALT (uden interpunktionstegn)	**250.209**	**38.767**
I ALT (med interpunktionstegn)	**290.600**	**38.780**

[52] Her henviser ordtyper (eller 'types') til antallet af grafisk forskellige tekstord (fordelt på ordklasserne). I denne tabel er de normaliseret mht. store/små bogstaver, men ikke mht. eventuelle bindestreger og accenttegn.

8.2 Fortegnelse over samtlige værdier i det danske PAROLE-tagsæt

CatGram	Attribute	*(træk)*	Value	*(værdi)*	Tag	Position
Adjective		*adjektiv*			**A**	1
	SsCatGram		Cardinal	*kardinal*	**C**	2
			Normal	*almindelig*	**N**	2
			Ordinal	*ordinal*	**O**	2
	Degree	*grad*	Positive	*positiv*	**P**	3
			Comparative	*komparativ*	**C**	3
			Superlative	*superlativ*	**S**	3
			Absolute Superl.	*absolut superlativ*	**A**	3
	Gender	*genus*	Common	*fælleskøn*	**C**	4
			Neuter	*intetkøn*	**N**	4
	Number	*numerus*	Singular	*singularis*	**S**	5
			Plural	*pluralis*	**P**	5
	Case	*kasus*	Unmarked	*umarkeret for kasus*	**U**	6
			Genitive	*genitiv*	**G**	6
	Definiteness	*bestemthed*	Definite	*bestemt*	**D**	8
			Indefinite	*ubestemt*	**I**	8
	Use	*transkategorisering*	Adverbial Use	*adverbiel anvendelse*	**R**	9
			Unmarked	*umarkeret for anvendelse*	**U**	9
Adposition		*adposition*			**S**	1
	SsCatGram	*præposition*	Preposition	*præposition*	**P**	2
Adverb		*adverbium*			**R**	1
	SsCatGram		General	*generel*	**G**	2
	Degree	*grad*	Positive	*positiv*	**P**	3
			Comparative	*komparativ*	**C**	3
			Superlative	*superlativ*	**S**	3
			Absolute Superl.	*absolut superlativ*	**A**	3
			Unmarked	*umarkeret for komparation*	**U**	3
Conjunction		*konjunktion*			**C**	1
	SsCatGram		Coordinative	*sideordnende*	**C**	2
			Subordinative	*underordnende*	**S**	2
Interjection		*lydord/udråbsord*			**I**	1
Noun		*substantiv*			**N**	1
	SsCatGram		Proper	*proprium*	**P**	2
			Common	*appellativ*	**C**	2
	Gender	*genus*	Common	*fælleskøn*	**C**	3
			Neuter	*intetkøn*	**N**	3
	Number	*numerus*	Singular	*singularis*	**S**	4
			Plural	*pluralis*	**P**	4
	Case	*kasus*	Unmarked	*umarkeret for kasus*	**U**	5
			Genitive	*genitiv*	**G**	5
	Definiteness	*bestemthed*	Definite	*bestemt*	**D**	8
			Indefinite	*ubestemt*	**I**	8
Pronoun		*pronomen*			**P**	1
	SsCatGram		Personal	*personligt*	**P**	2
			Demonstrative	*demonstrativt*	**D**	2
			Indefinite	*ubestemt*	**I**	2
			Interrog./relative	*interrogativt/relativt*	**T**	2
			Reciprocal	*reciprokt*	**C**	2
			Possessive	*possessivt*	**O**	2

CatGram	Attribute	*(træk)*	Value	*(værdi)*	Tag	Position
	Person	*person*	First	*første*	**1**	3
			Second	*anden*	**2**	3
			Third	*tredje*	**3**	3
	Gender	*genus*	Common	*fælleskøn*	**C**	4
			Neuter	*intetkøn*	**N**	4
	Number	*numerus*	Singular	*singularis*	**S**	5
			Plural	*pluralis*	**P**	5
	Case	*kasus*	Nominative	*nominativ*	**N**	6
			Genitive	*genitiv*	**G**	6
			Unmarked	*umarkeret for kasus*	**U**	6
	Possessor	*ejernumerus*	Singular	*singularis*	**S**	7
			Plural	*pluralis*	**P**	7
	Reflexive	*refleksivitet*	Yes	*ja*	**Y**	8
			No	*nej*	**N**	8
	Register	*stilleje*	Formal	*formel*	**F**	9
			Obsolete	*forældet*	**O**	9
			Polite	*høflig*	**P**	9
			Unmarked	*umarkeret for stilleje*	**U**	9
Residual		*residual*			**X**	1
	SsCatGram		Abbreviation	*forkortelse*	**A**	2
			Foreign Word	*udenlandske ord*	**F**	2
			Punctuation	*interpunktionstegn*	**P**	2
			Formulae	*formler*	**R**	2
			Symbol	*symboler*	**S**	2
			Other	*andet*	**X**	2
Unique		*unik*			**U**	1
Verb		*verbum*			**V**	1
	SsCatGram		Main	*'almindeligt' verbum*	**A**	2
			Medial	*medial*	**E**	2
	Mood	*modus*	Indicative	*indikativ*	**D**	3
			Imperative	*imperativ*	**M**	3
			Infinitive	*infinitivform*	**F**	3
			Gerund	*gerundium*	**G**	3
			Participle	*participium*	**P**	3
	Tense	*tempus*	Present	*præsens*	**R**	4
			Past	*præteritum*	**A**	4
	Number	*numerus*	Singular	*singularis*	**S**	6
			Plural	*pluralis*	**P**	6
	Gender	*genus*	Common	*fælleskøn*	**C**	7
			Neuter	*intetkøn*	**N**	7
	Definiteness	*bestemthed*	Definite	*bestemt*	**D**	8
			Indefinite	*ubestemt*	**I**	8
	Use	*transkategorisering*	Adjectival Use	*adjektivisk anvendelse*	**A**	9
			Adverbial Use	*adverbiel anvendelse*	**R**	9
			Unmarked	*umarkeret for anvendelse*	**U**	9

8.3 Antal forekomster af de forskellige morfosyntaktiske analyser i PAROLE-korpusset

Antal	Adjektiver (A)
6	AC---G=--
4119	AC---U=--
3	ANA---=-R
5	ANA[CN][SP]U=DU
583	ANC---=-R
259	ANC[CN]PU=[DI]U
133	ANC[CN]SU=IU
2	ANC[CN][SP]G=[DI]U
650	ANC[CN][SP]U=[DI]U
3716	ANP---=-R
2789	ANPCSU=IU
140	ANPCSU=[DI]U
1464	ANPNSU=IU
202	ANPNSU=[DI]U
25	ANP[CN]PG=[DI]U
4894	ANP[CN]PU=[DI]U
8	ANP[CN]SG=DU
3293	ANP[CN]SU=DU
1969	ANP[CN]SU=IU
155	ANP[CN]SU=[DI]U
1	ANP[CN][SP]G=[DI]U
2225	ANP[CN][SP]U=[DI]U
289	ANS---=-R
91	ANS[CN]PU=DU
5	ANS[CN]PU=[DI]U
21	ANS[CN]SU=DU
6	ANS[CN]SU=IU
395	ANS[CN][SP]U=DU
32	ANS[CN][SP]U=[DI]U
1	AO---G=--
419	AO---U=--

Antal	Konjunktioner (C)
9819	CC
5730	CS

Antal	Interjektioner (I)
259	I=

Antal	Substantiver (N)
190	NCCPG==D
154	NCCPG==I
1997	NCCPU==D
7932	NCCPU==I
23	NCCPU==[DI]
709	NCCSG==D
271	NCCSG==I
7500	NCCSU==D
17899	NCCSU==I
44	NCNPG==D
98	NCNPG==I
569	NCNPU==D
3758	NCNPU==I
376	NCNSG==D
137	NCNSG==I
3565	NCNSU==D
7094	NCNSU==I
3	NC[CN]PU==D
26	NC[CN]PU==I
79	NC[CN]SU==I
6	NC[CN][SP]G==[DI]
47	NC[CN][SP]U==I
168	NC[CN][SP]U==[DI]
1073	NP--G==-
13188	NP--U==-

Antal	Pronominer (P)
7	PC--PG---
77	PC--PU---
5	PD-CSG--U
1	PD-CSU--O
2747	PD-CSU--U
1633	PD-NSU--U
1	PD-[CN]PG--U
2251	PD-[CN]PU--U
307	PD-[CN][SP]U--U
10	PI-CSG--U
5240	PI-CSU--U
1011	PI-C[SP]N--U
2646	PI-NSU--U
11	PI-[CN]PG--U
4	PI-[CN]PU--O
589	PI-[CN]PU--U
1	PI-[CN][SP]G--U
43	PO1CSUPNF
200	PO1CSUSNU
22	PO1NSUPNF
84	PO1NSUSNU
63	PO1[CN]PUPNF
55	PO1[CN]PUSNU
134	PO1[CN][SP]UPNU
50	PO2CSUSNU
12	PO2NSUSNU
4	PO2[CN]PUSNU
14	PO2[CN][SP]UPNU
20	PO2[CN][SP]U[SP]NP
480	PO3CSUSYU
220	PO3NSUSYU
152	PO3[CN]PUSYU
318	PO3[CN][SP]UPNU
562	PO3[CN][SP]USNU
1224	PP1CPN-NU
240	PP1CPU-[YN]U
1645	PP1CSN-NU
310	PP1CSU-[YN]U
52	PP2CPN-NU
17	PP2CPU-[YN]U
397	PP2CSN-NU
87	PP2CSU-[YN]U
78	PP2C[SP]N-NP
18	PP2C[SP]U-[YN]P
2655	PP3CSN-NU
1104	PP3CSU-NU
3623	PP3NSU-NU
1055	PP3[CN]PN-NU
326	PP3[CN]PU-NU
1134	PP3[CN][SP]U-YU
20	PT-CSU--U
54	PT-C[SP]U--U
56	PT-NSU--U
25	PT-[CN]PU--U
370	PT-[CN]SU--U
29	PT-[CN][SP]G--U

Antal	Adverbier (R)
4	RGA
59	RGC
134	RGP
16	RGS
18804	RGU

Antal	Præpositioner (S)
30927	SP

Antal	Unique (U)
9037	U=

Antal	Verber (V)
8976	VADA=----A-
34	VADA=----P-
19127	VADR=----A-
772	VADR=----P-
7999	VAF-=----A-
628	VAF-=----P-
56	VAG-=SCI--U
437	VAM-=------
2	VAPA=---R--
4	VAPA=P[CN][DI]A-G
372	VAPA=P[CN][DI]A-U
15	VAPA=SCDA-U
5	VAPA=S[CN]DA-G
201	VAPA=S[CN]DA-U
220	VAPA=S[CN]IA-U
5507	VAPA=S[CN]I[ARU]-U
12	VAPR=---R--
456	VAPR=[SP][CN][DI]A-U
199	VAPR=[SP][CN][DI][ARU]-U
71	VEDA=----A-
220	VEDR=----A-
69	VEF-=----A-
16	VEPA=[SP][CN][DI][ARU]-U

Antal	Residual (X)
146	XA
295	XF
40391	XP
47	XR
169	XS
1067	XX

8.4 Fortegnelse over koderne til korpusteksternes klassifikation ifølge medium, genre og emne

8.4.1 Medium

Kode	Beskrivelse	Antal tekster
P.M1	bog	147
P.M2	dagblad	1208
P.M3.1	fagblad	29
P.M3.2	distriktsblad	9
P.M3.3	blad	106
P.M3.4	tidsskrift	3
P.M4.1	korrespondance	
P.M4.2	elektronisk medium (radio/tv)	21
P.M4.3	småtryk	30
P.M4.4	håndskrevent	
P.M4.5	tale	
P.M4.6	maskinskrevent	
I ALT		1553

8.4.2 Genre

Kode	Beskrivelse	Antal tekster	Kode	Beskrivelse	Antal tekster
P.G1	ukendt genre	817	P.G6.11	leder	19
P.G2.1	annonce/brochure		P.G6.12	liste	
P.G2.2	tryksag		P.G6.13	meddelelse	1
P.G2.3	foromtale	1	P.G6.14	notits	27
P.G2.4	katalog		P.G6.15	nyhedsartikel	7
P.G2.5	reklame		P.G6.16	nyhedsudsendelse	16
P.G2.6	reklametryksag		P.G6.17	petitstof	2
P.G2.7	tilbudsavis		P.G6.18	plakat	
P.G3.1	debat	77	P.G6.19	redaktionelt	2
P.G3.2	dialog		P.G6.20	redegørelse/referat	20
P.G3.3	folketingstale		P.G6.21	reportage	11
P.G3.4	interview/gruppeinterview	27	P.G6.22	skilt	
P.G3.5	kulturdebat		P.G7.1	bibliografi	
P.G3.6	læserbrev	8	P.G7.2	brevkasse	9
P.G3.7	diskussion/gruppediskussion/ samtale/gruppesamtale/ klassesamtale/telefonsamtale		P.G7.3	brevkassespørgsmål	2
P.G4.1	anmeldelse	76	P.G7.4	brevkassesvar	
P.G4.2	artikel	116	P.G7.5	fagbog	10
P.G4.3	causerie	1	P.G7.6	håndbog	2
P.G4.4	citat		P.G7.7	lærebog	13
P.G4.5	essay	2	P.G7.8	monografi	4
P.G4.6	feature		P.G7.9	opskrift/undervisning	
P.G4.7	klumme	27	P.G7.10	opslagsværk	3
P.G4.8	kronik	11	P.G7.11	rejsebog	2
P.G4.9	program/aktualitetsprogram/ magasin	12	P.G7.12	skolebog	7
P.G5	underholdning		P.G7.13	manual/vejledning	2
P.G5.1	børneside		P.G8.1	aforisme	1
P.G5.2	digt		P.G8.2	biografi	5
P.G5.3	dramatik		P.G8.3	festtale	
P.G5.4	eventyr	2	P.G8.4	kalender	
P.G5.5	film		P.G8.5	levnedsbeskrivelse	
P.G5.6	fortælling	8	P.G8.6	navne	22
P.G5.7	horoskop	2	P.G8.7	nekrolog	3
P.G5.8	humor/vittighed		P.G8.8	opgave	2
P.G5.9	lejlighedsvise		P.G8.9	portrætartikel	6
P.G5.10	novelle	26	P.G8.10	prædiken	2
P.G5.11	roman	84	P.G8.11	rapport	
P.G5.12	sangtekst		P.G8.12	selvbiografi	9
P.G5.13	serie	1	P.G8.13	skolestil	
P.G5.14	spil		P.G9.1	bekendtgørelse	1
P.G5.15	tegneserie		P.G9.2	betænkning	
P.G5.16	ungdomsside		P.G9.3	blanket	
P.G6.1	baggrundsartikel	22	P.G9.4	cirkulære	
P.G6.2	besked		P.G9.5	dokument	
P.G6.3	billedtekst		P.G9.6	forretningsbrev	
P.G6.4	dokumentarprogram		P.G9.7	kontrakt	
P.G6.5	enquete		P.G9.8	lov	3
P.G6.6	foredrag/forelæsning/ oplæg/oplæsning		P.G9.9	retsreferat	
P.G6.7	forbrugertips	1	P.G9.10	skrivelse	
P.G6.8	foreningsnyt		P.G10.1	brev	2
P.G6.9	folder/informationshæfte/ pjece	8	P.G10.2	dagbog	1
P.G6.10	information	8	I ALT		1553

8.4.3 Emne

Kode	Beskrivelse	Antal tekster	Kode	Beskrivelse	Antal tekster
P.T1	ukendt emne	617	P.T6.14	musik og dans	67
P.T2.1	økonomi	42	P.T6.15	litteratur	26
P.T2.2	erhvervsliv	32	P.T6.16	sprog	4
P.T2.3	landbrug	2	P.T7.1	transportmidler	29
P.T2.4	fiskeri og jagt	4	P.T7.2	bolig	15
P.T2.5	handel	11	P.T7.3	mad	15
P.T2.6	reklame	3	P.T7.4	tøj	9
P.T2.7	industri	2	P.T7.5	fjernsyn	30
P.T2.8	håndværk	5	P.T7.6	sport	106
P.T2.9	byggeri	2	P.T7.7	leg og spil	3
P.T3.1	folkloristik		P.T7.8	fritid	7
P.T3.2	geografi	2	P.T8.1	edb	5
P.T3.3	rejser	18	P.T8.2	almen naturvidenskab	3
P.T3.4	antropologi/ etnologi/-grafi	1	P.T8.3	matematik	1
P.T4.1	psykologi	7	P.T8.4	astronomi	4
P.T4.2	familie	17	P.T8.5	fysik	4
P.T4.3	pædagogik	2	P.T8.6	kemi	2
P.T4.4	helbred	58	P.T8.7	geologi	
P.T4.5	sex og samliv	19	P.T8.8	biologi	12
P.T5.1	historie	10	P.T8.9	botanik	6
P.T5.2	personlalhistorie	10	P.T8.10	zoologi	8
P.T6.1	bogvæsen	2	P.T8.11	teknik	5
P.T6.2	filosofi		P.T8.12	miljø	15
P.T6.3	kommunikation	4	P.T9.1	videnskab	2
P.T6.4	religion	6	P.T9.2	samfund	18
P.T6.5	undervisning/uddannelse	18	P.T9.3	politik	68
P.T6.6	presse	3	P.T9.4	EU	22
P.T6.7	kunst	10	P.T9.5	jura	8
P.T6.8	kultur	7	P.T9.6	kriminalitet	50
P.T6.9	arkitektur	2	P.T9.7	social forsog	16
P.T6.10	foto	2	P.T9.8	militær	11
P.T6.11	teater	17	P.T9.9	forbrugerstof	2
P.T6.12	film	33	P.T9.10	trafik	9
P.T6.13	radio	3	I ALT		1553

8.5 Samlet oversigt over flerordsforbindelser i PAROLE-korpusset

I følgende flerordsforbindelser anvendes en understregning (_) til at samle ordformerne i korpusteksten (samt i lemmaformen). Understregninger kan dog også forekomme i lemmaformer til andre ordformer i korpusteksterne (som f.eks. forkortelser).

8.5.1 Gruppesammensætninger (jf. afsnit 4.2.2)

"Bog_til_tiden"-princippet
"Dirty_Harry"-filmen
"Skør_med_Klatten"-programmerne
100_Pfennig-frimærke
120_meter-bakken
2._december-notatet
20_års-perioden
200_meter-tid
22_procent-skatten
7_pct.-gebyret
Andy_Rouse-teamet
B_&_W-direktør
Baltica_Holding-aktierne
big_band-arrangement
big_band-plade
Bjørn_Wiinblad-fad
BMW_700-serie
Bonnie_og_Clyde-agtigt
cand._merc.-eksamen
Caroline_Amalie-lunden
Charta_77-folkene
Chet_Baker-tone
Christian_den_Fjerde-statuette
Cole_Porter-cabaret
Dansk_Idræts-Forbund
Davis_Cup-opgør
Davis_Cup-opgøret
De_Beers-topfolk
Elton_John-backing
Europa_Cup-deltagelse
Europa_Cup-runde
Europa_Cup-turneringen
Europa_Cup-turneringer
fast_food-generationen
Formel_1-veteran
Grand_Prix-deltagelser
Gruppe_A-serie
Hanne_Boel-koncert
Helle_Stangerup-udgivelse
hip_hop-genren
Holiday_Inn-hotellet
IHF_Cup-kamp
Iron_Maiden-turneer
joint_venture-aftalen
joint_venture-alliancen
joint_venture-selskaber
Jysk_Sengetøj-koncernen
Le_Mans-race
Le_Mans-racet
low_budget-præg
M_17-holdet
Magnum_44-revolver
Michael_Jackson-plade
new_age-musik
objekter/ting_der_bliver_levende-syndromet
Rich._Strauss-opera
Røde_Kro-løjer
San_Camillo-hospitalet
Sanam_Luang-pladsen
tax-free_shop
tidsrejser_contra_paradokser-syndromet
TV_2-AKTIER
TV_2-bestyrelsen
TV_2-Nyhederne
TV_2-programmet
Woody_Allen-film
World_Cup-takterne
World_Cup-turneringen
YOM_KIPPUR-HELT
Yom_Kippur-krigen
zweite_Kanalton-knappen
Øde_ø-syndromet
Østre_Gasværks-grunden

8.5.2 Forkortelser (jf. afsnit 5.8.1)

bl._a.
cand._jur.
cand._mag.
cand._merc.
cand._polyt.
cand._psych.
cand._theolen
dr._med.
dr._phil.
dr._scient._pol.
dr._theol.
lic._jur.
lic._techn.

8.5.3 Faste ordforbindelser (jf. afsnit 4.2.1)

a_la_carte
af_og_til
af_sted
all_right
alt_imens
blandt_andet
blandt_andre
bortset_fra
café_au_lait
en_bloc
en_suite
for_nylig
for_tiden
for_øvrigt
i_aften
i_alt
i_bunkevis
i_dag
i_det_hele_taget
i_eftermiddag
i_fjor
i_forkøbet
i_forvejen
i_gang
i_går
i_hvert_fald
i_læssevis
i_morgen
i_månedsvis
i_nat
i_timevis
i_tusindvis
i_ugevis
i_øvrigt
i_år
i_årevis
ikke_desto_mindre
med_hensyn_til
mere_eller_mindre

om_bord
om_end
om_muligt
over_bord
på_grund_af
på_ny
selv_om
simpelt_hen
som_om
stort_set
til_sidst

8.5.4 Fossilerede dativer/genitiver (jf. afsnit 4.2.3)

af_syne
i_aftes
i_eftermiddags
i_fredags
i_går_aftes
i_går_eftermiddags
i_går_morges
i_hænde
i_live
i_lørdags
i_mandags
i_onsdags
i_sinde
i_søndags
i_tide
i_tirsdags
i_torsdags
med_rette
nu_til_dags
på_fode
på_tide
til_bords
til_bunds
til_døde
til_fods
til_fulde
til_gode
til_huse
til_lands
til_orde
til_rette
til_rors
til_stede
til_tops
til_tåls
til_veje
til_vejrs
til_vægs

8.5.5 Andet 1½

B_1903-Silkeborg (tekstfejl)
Boing_747-fragtmaskiner (tekstfejl)
charge_d'affairs (tekstfejl)
Gatorade-Chateau_d'Ax (proprium)