ONTOQUERY

Januar 2000

Træning og brug af Brill-taggeren på danske tekster

Teknisk Rapport

af
Dorte Haltrup Hansen
Center for Sprogteknologi
dorte@cst.dk

1 INTRODUKTION

Træningen af Brill-taggeren er foretaget som en del af OntoQuery-projektet. OntoQuery er et tværfagligt forskningsprojekt med deltagere fra RUC, DTU, CST, HHK og SDU. Det har til formål at udvikle teorier og metoder til indholds-baseret informationssøgning, og i den forbindelse er taggingen en præproces i den sproglige analyse.

Denne rapport indeholder beskrivelser af :
1) principperne bag Brill-taggeren, 2) PAROLE-korpuset som taggeren er trænet på, 3) træningsprocessen, 4) resultater fra test af taggeren samt 5) en brugervejledning.

2 BRILL-TAGGEREN[1]

POS-tagging (part of speech tagging) i traditionel forstand vil sige at tildele morfosyntaktiske kategorier til ord i en tekst.
Brill-taggeren er en regel-baseret tagger der bliver trænet på et allerede tagget korpus,- fx et manuelt tagget korpus eller et semiautomatisk tagget korpus. Træningen foregår ved at taggeren automatisk lærer sig selv nogle regler hvorefter den er i stand til at tagge en ny og ukendt tekst.
Under træningen arbejdes med to versioner af samme korpus: den oprindelige taggede version samt en version hvor alle taggene er fjernet. Først tildeles ordene i det "nøgne" korpus et tilfældigt tag. Derefter ændres taggene vha. transformationer på en måde så den transformationsregel der får det "nøgne" korpus til at nærme sig det oprindelige, får en højere vægtning; mens de regler der får korpus til at fjerne sig fra det oprindelige, bliver smidt væk. På den måde opbygges lister af ordnede regler,- leksikalske regler og kontekstuelle regler. De leksikalske regler bruges til at analysere ukendte ord, mens de kontekstuelle regler bruges til at fjerne syntaktisk flertydighed.
Leksikalske regler kan se således ud:

```
s deletesuf 1 N_GEN 378.125989593482
N det fgoodright ADJ 292.581302071428
ede hassuf 3 V_PAST 277.815201465202
N t fdeletesuf 1 V_PARTC_PAST 253.34276092672
N 1 fchar NUM 246.966666666667
r deletesuf 1 V_PRES 223.077780170032
ige hassuf 3 ADJ 214
N ig fhassuf 2 ADJ 194.875
```

Den første regel betyder:

[1] Brill-taggeren er en softwarepakke der frit kan downloades fra: www.cs.jhu.edu/~brill/

”Hvis fjernelse af suffikset –s resulterer i et eksisterende ord, så ændr tagget (hvad det end er) til N_GEN”.

Anden regel siger:

” Ændr N til ADJ når det aktuelle ord (med tagget N) optræder umiddelbart til højre for ordet *det*”.

Tallene efter reglerne er en form for vægtning af reglerne.
Kontekstuelle regler kan se således ud:

```
UNIK UKONJ PREV1OR2WD ,
PRON_PERS PRON_DEMO NEXTTAG ADJ
PRON_DEMO PRON_PERS NEXTTAG V_PRES
PRÆP ADV NEXTTAG PRÆP
N V_INF PREVBIGRAM PRÆP UNIK
```

Første regel betyder:

” Ændr UNIK til UKONJ hvis et af de sidste to ord er et , (komma) ”.

Næste regel betyder:

” Ændr PRON_PERS til PRON_DEMO hvis det næste tag er ADJ”.[1]

3 PAROLE

Brill-taggeren er trænet på en delmængde af PAROLE-korpuset. Dette (fremover kaldet PAROLE) er et morfosyntaktisk annoteret korpus. Det indeholder 250.209 løbende ord fordelt på 16062 sætninger fra 1553 tekstuddrag. Teksterne, der ligger i SGML-format, stammer fra DSL og Den Danske Ordbog.
PAROLE er et EU-projekt der involverer 14 europæiske sprog. I projektet er der inden for hvert sprog opbygget korpora der følger en harmoniseret PAROLE-standard.
I den danske del af PAROLE er korpuset analyseret morfologisk med DAN-TWOL-algoritmen. Den giver en eller flere alternative analyser til hvert ord, hvorefter den korrekte analyse manuelt er markeret <**correct**!>. Fx:

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

[1] se også Bilag 1

"folke#ting" <*> N INT SG BEST NOM <**correct**!>

DAN-TWOLs tagsæt består af 338 forskellige analyser. I det endelige PAROLE-korpus er tagsættet transformeret til 151 tags der følger PAROLE-standarden[2]. Samme tekstbid som ovenfor ser i PAROLE-formatet ud på flg. måde:

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

,- hvor ordet efter *lemma* er tekstordets lemma, bogstaver og tegn efter *msd* er tekstordets tag og selve tekstordet står sidst før *</W>*.

BEARBEJDNING AF PAROLE-KORPUSET

For at kunne træne taggeren på korpuset, har det været nødvendigt at bearbejde det. Først er tagsættet reduceret fra 151 til 43 tags[3] efter Britt Kesons anvisning[4], da et mindre tagsæt giver bedre resultater. De nye tags består hovedsagelig af ordklassebetegnelser. For verbernes vedkommende indeholder tagget dog også tempus og modus. Det eneste ekstra træk jeg har valgt at taget med (ud over dem Keson har anvist), er *genitiv* (_GEN). Det er gjort med NP-genkendelse og NP-parsing i baghovedet. Jeg formoder at information af denne art kan afhjælpe flertydighed i afgrænsning af NPer.
Dernæst har jeg trukket ordform og kategori(tag) ud af PAROLE-korpuset og omformet korpus til det inputformat der kræves af taggeren. Dvs. en (hel-)sætning på hver linie, space omkring interpunktionstegn og tagget knyttet til ordet vha. /. Fx:

Samtlige/ADJ partier/N i/PRÆP Folketinget/N … ./TEGN

Også "ikke-sætninger" (dvs. størrelser der ikke er afsluttet af et punktum) er sat på separate linier. Det er fx:

Overskrifter: AFSLØRET/V_PARTC_PAST TILFÆLDIGT/ADJ
STØJSVAG/ADJ VENTILATOR/N

Signaturer osv.: Af/PRÆP Vibeke/EGEN Scheibel/EGEN
JP/EGEN
Tekst/N :/TEGN Torsten/EGEN Cilleborg/EGEN
Foto/N :/TEGN Flemming/EGEN Nielsen/EGEN

[2] se PAROLE-tagsættet i Bilag 2
[3] se Bilag 3
[4] se "Morfosyntaktisk Tagging af Danske Tekster" i *7. Møde om Udforskningen af Dansk Sprog*, 1999

Et ”problem” der endnu er uløst er anførselstegn efter et punktum. Da det er tvetydigt hvor de hører til, er de også sat på separate linier.
Man kan diskutere hvad man skal gøre ved fejl i korpus (tastefejl, stavefejl og skrivefejl). Man burde muligvis fjerne dem før træning af taggeren så de ikke slører en senere analyse. Det er ikke gjort her.

4 TRÆNING AF BRILL-TAGGEREN

Efter bearbejdning af korpus starter træningen. Først har jeg delt korpus i et træningskorpus (90%) og et testkorpus (10%). Det er selvfølgelig kun i situationer hvor man er interesseret i at teste taggerens performans, at man behøver et testkorpus. Er man ikke interesseret i at tjekke taggerens fejlprocent og fejltyper automatisk, kan man træne taggeren på hele korpus.
Derefter har jeg fulgt Brills anvisninger for træningen.[5] Træningen består i at der automatisk bliver lavet en række leksika og genereret transformationsregler. På en HP 9000/785 med HP–UX 10.20 (400 MHz) tager det ca. 30 timer at finde de 400 leksikalske regler og ca. 16 timer at finde de 350 kontekstuelle regler. Når taggeren er trænet én gang, er den klar til brug. Da tager det 15 sek. at tagge 28698 ord.

VEJEN TIL DE BEDSTE RESULTATER

Taggeren blev først trænet på træningskorpusets 261.904 ord og derefter testet på testkorpusets 28.698 ord. Det gav en fejlprocent på 3,86,- dvs. at 96,14 % af testkorpusets ord fik den rette analyse.[6] I et forsøg på at få fejlprocenten længere ned blev de tydeligste fejl kategoriseret[7]. Det viste sig at de to ”hovedsyndere” var ukendte ord (64% af fejlene) samt ord der begynder med stort bogstav (28% af fejlene).
For at afhjælpe problemet med de ukendte ord blev der lavet en træning som inkluderede Scarrie-ordbogen (166864 ordformer). Desværre gav det ikke nogen forbedring,- tværtimod. Fejlprocenten blev nu 3,9 (vs. tidligere 3,86 %). Grunden er nok at taggeren ikke er designet til at inkludere andre ord end dem den finder i korpus. Det er muligt at tilføje nye ord men kun i form af et nyt korpus.
Derfor var næste tiltag at alle store bogstaver efter punktum i korpus blev ændret til små bogstaver. Dette gav en lille forbedring,- fejlprocenten kom nu ned på 3,5; som desværre ikke var den forventede nedgang. Da over 28 % af fejlene stammede fra ord med stort begyndelsesbogstav, var det forventede resultat at fejlprocenten gik ned med ca. 1 %.
I et sidste forsøg blev alle store bogstaver ændret til små, hvorved fejlprocenten igen gik op til 3,9.

---

[5] se Bilag 4 og afsnittet ”Brugervejledning”
[6] Hvis man vil have en fuldstændig statistisk holdbar fejlrate, kan man lave en ”cross validation”,- dvs. dele korpuset i 10 dele, lave 10 træninger, teste på 10 forskellige bidder og sidst finde den gennemsnitlige fejlprocent.
[7] se Bilag 5

Altså:

1. forsøg		3,86 % fejl
2. forsøg	(træning m. Scarrie-ordbogen)	3,9 % fejl
3. forsøg	(stort bogstav efter punktum ændres til lille)	3,5 % fejl
4. forsøg	(alle store bogstaver til små)	3, 9 % fejl

Fejlene i det hidtil bedste forsøg (nr. 3) blev derfor kategoriseret. [8]
Det tyder på at det ikke er de store bogstaver i sig selv der giver problemer, men derimod egennavne generelt. 33,5 % af fejlene stammer fra problemer med egennavne,- enten skulle tagget have været EGEN (egennavn) ellers har ordet fejlagtigt fået tagget EGEN.
Man kan sige at en oplagt måde at forbedre taggingen på ville være at redigere i transformationsreglerne eller at tilføje helt nye regler. Ved at gøre det risikerer man dog at lave nye uoverskuelige fejl. Problemet er dels at reglerne er ordnede og dels at ingen fejl er fuldstændig systematiske. Derfor har jeg afholdt mig fra dette.

Forsøg på forbedringer er stoppet her. Ved en senere træning kunne man se på definitionen af egennavne i PAROLE-korpuset. Et navn er nemlig ikke altid et egennavn fordi det er stavet med stort i PAROLE. Hvis ordet fx findes som substativ, bliver det kaldt substantiv også selvom der i konteksten er tale om et navn. Fx:

Mogenstrup/EGEN Grusgrav/N
Det/PRON_DEMO kongelige/ADJ Teater/N
On/EGEN the/UL Air/EGEN

5 RESULTATER FRA DEN FÆRDIGTRÆNEDE TAGGER

Eric Brill skriver[9] at hans tagger giver 97,2 % korrekte analyser når den er trænet på 600.000 ord og alle ord i testmaterialet er kendt af taggeren. I dette forsøg fås (mindst) 98,5 % korrekte analyser når taggeren er trænet på 260.000 ord der alle er kendte af taggeren. Forskellen må bl.a. ligge i at der bruges forskellige tagsæt. Forsøget er ikke omtalt tidligere, da jeg ikke anser det for interessant ifht. taggerens fremtidige opgave hvor man ikke kan gå ud fra at alle ord er kendte. Videre skriver Brill at han får 96,5 % korrekte analyser når taggeren er trænet på 950.000 ord fra Penn Treebank, samt at 85 % af alle ukendte ord bliver gættet. I

[8] se Bilag 6

[9] I "Some Advances in Transformation-Based Part of Speech Tagging", Twelfth National Conference on Artificial Intelligence (AAAI-94)

dette forsøg er resultatet næsten det samme: 96,5 % af ordene får den korrekte analyse, dog bliver kun ca. 80 % af alle ukendte ord gættet.
Disse tal forudsætter at det materiale man vil have tagget, er korrekt "tokenized". Dvs. at hver "token" (ordform, tal, forkortelser, tegn, symboler o.lign.) er adskilt af et mellemrum. Hvis teksten ikke er korrekt "tokenized", kan fx tegn som skrives lige efter et ord give problemer. Hvis tegnet ikke er skilt fra ordet med et mellemrum, kan man ikke slå ordet op i ordbogen. Modsat hører tegn nogle gange med til ordet (fx forkortelsespunktummer) så ordet ikke findes i ordbogen uden det afsluttende tegn (punktummet).

ET EKSEMPEL

Som illustration for hvordan taggeren klarer en tekst fra et specifikt domæne hvor en del ord må forventes at være ukendte, har jeg tagget "A-vitamin-artiklen" fra Den Store Danske Encyklopædi.[10] Artiklen består af 378 ord (incl. interpunktionstegn) som er indtastet og tagget manuelt.[11] Samme artikel er derefter tagget med Brill-taggeren[12] hvorefter de to versioner er sammenlignet. 10 steder giver Brill-taggeren en forkert analyse[13] hvilket resulterer i en fejlprocent på 2,7,- dvs. 97,3 % af teksten er analyseret korrekt. Man skal nok ikke lægge for meget i at fejlprocenten er mindre end i det hidtil bedste forsøg. Testmaterialet er så lille at 3,8 ord kan få fejlraten til at svinge 1%. Derfor vil det kræve mere testning for at få et realistisk billede af hvordan taggeren opfører sig på et nyt domæne.

Nedenfor ses en sætning fra A-vitamin artiklen,- først uden tags, derefter tagget med Brill-taggeren og sidst er det forklaret hvad taggeren har gjort ved hver token:

- A-vitaminsyre (retinsyre , tretinoin) anvendes i cremer til medicinsk behandling af bumser (akne) .

- a-vitaminsyre/N (/TEGN retinsyre/N ,/TEGN tretinoin/N)/TEGN anvendes/V_INF i/PRÆP cremer/N til/PRÆP medicinsk/ADJ behandling/N af/PRÆP bumser/N (/TEGN akne/N)/TEGN ./TEGN

- √ a-vitaminsyre/N — ordet findes ikke i leksikon men taggeren gætter korrekt at det er et substantiv (N)
  √ (/TEGN — taggeren kender tegnet og giver der derfor det rette tag
  √ retinsyre/N — ordet findes ikke i leksikon men taggeren gætter korrekt at det er et substantiv (N)
  √ ,/TEGN — taggeren kender tegnet og giver der derfor det rette tag
  √ tretinoin/N — ordet findes ikke i leksikon men taggeren gætter korrekt at det er et substantiv (N)
  √)/TEGN — taggeren kender tegnet og giver der derfor det rette tag

[10] se Bilag 7
[11] se Bilag 8
[12] se Bilag 9
[13] se Bilag 10

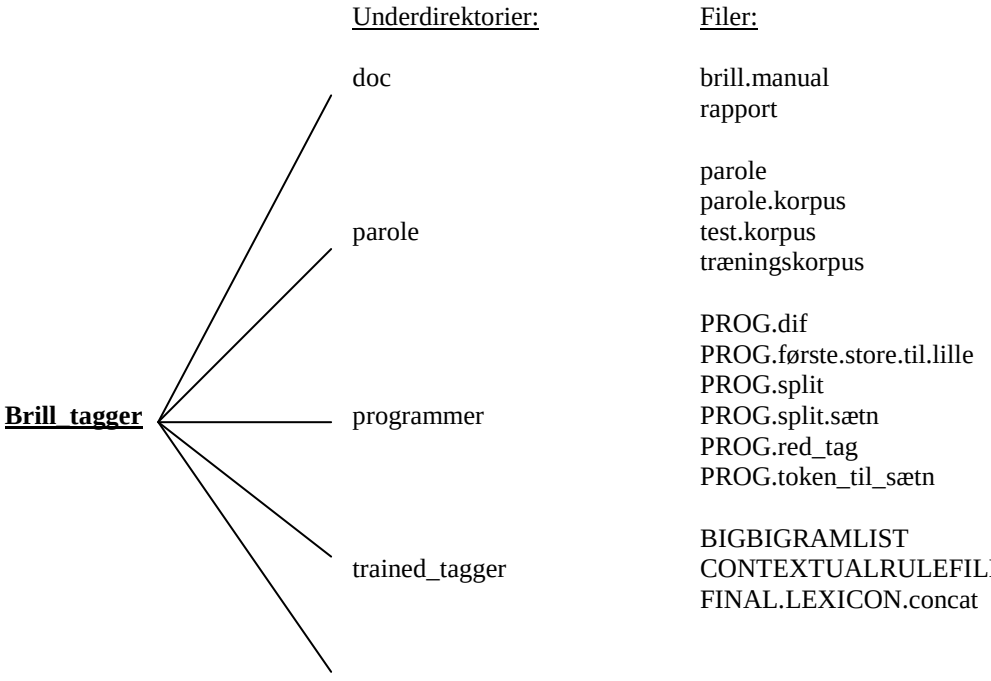
- anvendes/V_INF	i ordbogen findes kun infinitivformen at ordet derfor gives det forkerte tag
√ i/PRÆP	iflg. leksikon kan ordet have flg. tags: PRÆP, ADV, PERS_PRON, FORK, N. Ud af disse alternativer gættes den korrekte
√ cremer/N	taggeren kender ordet og giver derfor det rette tag
√ til/PRÆP	iflg. leksikon kan ordet have flg. tags: PRÆP, ADV, UKONJ. Ud af disse alternativer gættes den korrekte
√ medicinsk/ADJ	taggeren kender ordet og giver derfor det rette tag
√ behandling/N	taggeren kender ordet og giver derfor det rette tag
√ af/PRÆP	iflg. leksikon kan ordet have flg. tags: PRÆP, ADV. Ud af disse alternativer gættes den korrekte
√ bumser/N	iflg. leksikon kan ordet have flg. tags: V_PRES, N. Ud af disse alternativer gættes den korrekte
√ (/TEGN	taggeren kender tegnet og giver der derfor det rette tag
√ akne/N	ordet findes ikke i leksikon men taggeren gætter korrekt at det er et substantiv (N)
√)/TEGN	taggeren kender tegnet og giver der derfor det rette tag
√ ./TEGN	taggeren kender tegnet og giver der derfor det rette tag

Det er vigtigt at man ikke henledes til at tro at fejlen i eksempelsætningen og fejlene i Bilag 10 er repræsentative. Som tidligere sagt er dette testmateriale så lille at det ikke kan danne grundlag for eksplicitte regler til taggerens regelsæt.

6 BRUGERVEJLEDNING

Brill-taggeren downloades som en programpakke,- en *tar-fil*. Når den er pakket ud, ligger programfilerne i et net af direktorier. Det sted jeg refererer til som hjemmedirektoriet, er niveauet over ”RULE_BASED_TAGGER_V1.14”.

Direktoriet ”Brill_tagger” er organiseret på flg. måde:



LEXRULEOUTFILE

RULE_BASED_TAGGER_V1.14 …..

Rapporten her samt manualen fra Bilag 4 ligger i direktoriet ”doc”. Det originale PAROLE-korpus ”parole.korpus” samt de bearbejdede versioner af korpuset ligger i direktoriet ”parole”. De forskellige Perl-scripts til bla. at bearbejde korpuset med ligger i ”programmer”. Og selve den trænede tagger består af de sidste to direktorier ”trained_tagger” og ”RULE_BASED_TAGGER_V1.14”.

Dette afsnit er inddelt i 3 punkter: først en vejledning i brug af taggeren, derefter anvisning af hvad man gør hvis man vil lave ændringer eller forbedringer af taggeren, og sidst en udførlig beskrivelse af de forskellige faser i træningen af taggeren.

BRUG AF TAGGEREN

Som beskrevet tidligere skal det korpus eller den tekst man vil have tagget have et bestemt format: der skal være en (hel-)sætning på hver linie, og der skal være space omkring interpunktionstegn. Først skal teksten ”tokenizes” og bagefter bruges programmerne PROG.token.til.sætn og PROG.første.store.til.lille der ligger i under direktoriet ”programmer”.[14]
Derefter sættes taggeren i gang som anført i Bilag 4 under afsnittet ”Tagging med Brill-taggeren”:

> Stå i direktoriet RULE_BASED_TAGGER_V1.14/Bin_and_Data/.
> Start taggeren ved at skrive:
>
> ```
> tagger ../../trained_tagger/FINAL.LEXICON
>        ../../ trained_tagger/CORPUS
>        ../../trained_tagger/BIGBIGRAMLIST
>        ../../ trained_tagger/LEXRULEOUTFILE
>        ../../ trained_tagger/CONTEXT-RULEFILE
>      > ../../ trained_tagger/CORPUS-TAGGED
> ```
>
> INPUT = CORPUS det korpus der skal tagges
> (+ de 4 filer fra ”trained_tagger)
> OUTPUT = CORPUS-TAGGED samme korpus nu tagget

Der skal altså bruges et utagget korpus eller tekst (CORPUS) samt de fire filer fra underdirektoriet ”trained_tagger”. Outputet er et tagget korpus (CORPUS_TAGGED).
Det tager 15 sek. at tagge 28.698 ord.

[14] Se under afsnittet ”Bearbejdning af korpus” i filen ”doc/brill.manual” for vejledning i brug af programmerne.

ÆNDRINGER & FORBEDRINGER AF TAGGEREN

Der findes forskellige måder at ændre taggeren på. Måske ønsker man ikke samme tagsæt som det taggeren er trænet med nu. Det kunne være at man ønskede at have flere træk med eller (fx for verberne) at have færre med. Man går da ind i det oprindelige PAROLE-korpus (parole/parole.korpus) og laver taggene om. Det kan fx gøres med programmet ”PROG.red_tags” fra underdirektoriet programmer. Her betyder X det oprindelige PAROLE-tag og Y det tag man ønsker det lavet om til:

```
$linie =~ s/X/Y/;
```

Så køres programmerne PROG.split, PROG.split.sætn og PROG.første.store.til.lille fra underdirektoriet ”programmer” og til sidst træner man taggeren igen som anført i Bilag 4.

En anden mulighed ville være at ændre PAROLEs definition på egennavne. Igen ændres der i det oprindelige korpus hvorefter en ny træning foretages.

Brill skriver i sine README-filer at der er to måder at modificere taggeren på. Hvis man vil have tagget et nyt korpus, kan man tilføje korpusets ord som lister til taggeren. Formodningen er at kendskab til nye ords eksistens og omgivelser kan hjælpe taggeren. I mit forsøg gav det (desværre) ingen effekt. Det man i stedet kan gøre, er at tilføje oplysninger til ordbogen FINAL.LEXICON. Man kan tilføje nye ord med deres kategori (tag). Man skal blot være opmærksom på at et ord kan have flere forskellige tags. Ordbogen er organiseret så det mest frekvente tag står først. Fx:

```
så ADV V_PAST ADJ UKONJ
```

,- hvilket betyder at ordet *så* kan have 4 forskellige tags: ADV, V_PAST, ADJ, og UKONJ og at ADV er det tag der er hyppigst brugt.
I den trænede version af taggeren her er fx tilføjet tegn og symboler som taggeren ikke kendte, samt en række ord fra Scarrie-ordbogen.

I tilfælde hvor man ikke behøver en absolut analyse, kan man lave en ”n-best-tagging”. Da kobles et statistisk modul til som i tvivlstilfælde kan tildele flere analyser til ordene.

De lige nævnte forbedringer tage udgangspunkt i det eksisterende taggede PAROLE-korpus. En anden måde at forbedre taggeren på kunne være at lave en ”domænespecifik” tagger. Problemet er at man har brug for et præ-tagget korpus til at træne taggeren på. Ved at bruge den ”generelle” tagger på et ”domænespecifikt” materiale kan man opbygge et nyt tagget korpus. Dette skal så korrigeres hvorefter man kan træne taggeren på ny. På den måde opnår man en domænespecifik tagger. Det er dog temmelig tidskrævende at skulle gennemlæse og evt. korrigere et større tekstmateriale.

EN UDFØRLIG BRUGERVEJLEDNING I TRÆNING AF TAGGEREN

Under træningen af taggeren følges punkterne i Bilag 4 (eller i doc/brill.manual) slavisk. Punkterne er lavet så man kan bruge ”cut and paste”,- dvs. at man ikke behøver at indtaste alle kommandoer manuelt, og at man dermed undgår (irriterende) tastefejl.

I dette afsnit uddybes punkterne ”Træning af Brill-taggeren” fra Bilag 4 og der gives eksempler på format og indhold af de forskellige filer som taggeren genererer.

Først deles træningskorpuset i 2 lige store dele. Ud fra den første del læres de leksikalske regler, og fra den anden del læres en række kontekstuelle regler.

Derefter opbygges et leksikon over alle ordformer i træningsleksikonet (BIGWORDLIST). Det består af ordform og frekvens. Fx:

, 15768
. 12763
og 6653
i 6582
at 5654
er 4575
det 4568

Så laves en bigramliste over alle ordpar i træningskorpuset (BIGBIGRAMLIST). Fx fra sætningen:

”a-vitamin , fedtopløselig vitamin , der i kosten findes i flere former”

Fås:

a-vitamin ,
, fedtopløselig
fedtopløselig vitamin
vitamin ,
, der
der i
i kosten
kosten findes
findes i
i flere
flere former

Sidst laves et leksikon bestående af ordform, tag samt frekvens fra den første halvdel af træningskorpuset (SMALLWORDTAGLIST). Fx:

, TEGN 7819
. TEGN 6292
og SKONJ 3303

i PRÆP 3254
at UNIK 1582
er V_PRES 2331
det PRON_PERS 1588

Udfra disse leksika konstrueres en række leksikalske regler (LEXRULEOUTFILE). Fx:

s deletesuf 1 N_GEN 378.125989593482
N det fgoodright ADJ 292.581302071428
ede hassuf 3 V_PAST 277.815201465202
N t fdeletesuf 1 V_PARTC_PAST 253.34276092672
N 1 fchar NUM 246.966666666667
r deletesuf 1 V_PRES 223.077780170032
ige hassuf 3 ADJ 214
N ig fhassuf 2 ADJ 194.875

Den første regel betyder: ”Hvis fjernelse af suffikset –s resulterer i et eksisterende ord, så ændr tagget (hvad det end er) til N_GEN”. Anden regel siger: ” Ændr N til ADJ når det det aktuelle ord (med tagget N) optræder umiddelbart til højre for ordet *det*”.
De leksikalske regler bruges i taggingen til at gætte kategorien på ukendte ord.
Næste trin i træningen er opbygning af et leksikon bestående af alle ordformer med tags fra første del af træningskorpus (TRAINING.LEXICON).Fx:

Fyns EGEN_GEN
travtilskuers N_GEN
uld N
biskop N
festivalerne N
gennemførlige ADJ
Faurschou EGEN

Dernæst konstrueres et ”dummy-korpus”. Så vidt jeg kan forstå, er det 2. delkorpus som først får ”skrællet” alle tags af og derefter får tildelt tags udfra de opbyggede leksika og de leksikalske regler. Fx:

de hævder , at Ruslands vej til demokrati går gennem diktatur .

Bliver til:

de/PRON_DEMO hævder/V_PRES ,/TEGN at/UNIK Ruslands/EGEN_GEN vej/N til/PRÆP demokrati/N går/V_PRES gennem/PRÆP diktatur/N ./TEGN

Som man kan se laver taggeren fejl (allerede) her. ”*de/PRON_DEMO*” skulle have været ”*de/PRON_PERS*”. Jeg kan desværre ikke overskue hvilke konsekvenser det har.

Ud fra ”dummy-korpuset” og træningsleksikonet læres de kontekstuelle regler. Fx:

```
UNIK UKONJ PREV1OR2WD ,
PRON_PERS PRON_DEMO NEXTTAG ADJ
PRON_DEMO PRON_PERS NEXTTAG V_PRES
PRÆP ADV NEXTTAG PRÆP
N V_INF PREVBIGRAM PRÆP UNIK
```

Første regel betyder: ” Ændr UNIK til UKONJ hvis et af de sidste to ord er et , (komma) ”. Næste regel betyder: ” Ændr PRON_PERS til PRON_DEMO hvis det næste tag er ADJ”.

Så er taggeren færdigtrænet. [15]

7 KONKLUSION

Brill-taggeren er en overskuelig tagger der er meget let at træne og bruge. Dens processeringstid er hurtig hvilket gør den anvendelig til store tekstmængder. Det tager ca. 50 timer at træne den på en HP 9000/785 med HP-UX 10.20 (400 MHz) og 15 sek. at tagge 28.000 ord.
Endvidere har den en acceptabel lille fejlprocent (3,5%) når tagsættet er tilpas lille. Denne fejlprocent er fuld på højde med fejlraten for andre produkter på markedet. Taggeren har også en forholdsvis god evne til at gætte den morfosyntaktiske kategori på ukendte ord (80%),- en evne som selvfølgelig er nødvendig men ikke triviel.

[15] se Bilag 4 for en mere detaljeret brugervejledning

(eksempler på) LEKSIKALSKE REGLER bruges til at gætte kategorien på ukendte ord

---

X y fhassuf Z : ændre tag X til tag Z hvis ordet m. tagget X har suffixet y
y hassuf Z : ændre tag ? til tag Z hvis ordet med tagget ? har suffixet y

X y fgoodright Z : ændre tag X til tag Z hvis/når ordet optræder umiddelbart til højre for y
y goodright Z : ændre tag ? til tag Z (samme)
X y fgoodleft Z
y goodleft Z

y addsuf Z : ændre tag ? til tag Z hvis tilføjelse af suffixet y resultere i et ord
y char Z : ændre tag ? til tag Z hvis størrelsen y forekommer i ordet
y deletepref Z : ændre tag ? til tag Z hvis fjernelse af præfixet y resulterer i et ord

---

(eksempler på) KONTEKSTUELLE REGLER bruges til at finde den rette kategori

---

X Z WDNEXTTAG y W : ændre tag X til tag Z hvis nuværende ord er y og næste tag er W
X Z NEXT1OR2WD y : ændre tag X til tag Z hvis et af de næste 2 ord er y
X Z NEXTTAG Y : ændre tag X til tag Z hvis det næste tag er Y
X Z NEXT1OR2OR3TAG Y : ændre tag X til tag Z hvis et af de næste 3 tag er Y

X Z WDPREVTAG Y w : ændre tag X til tag Z hvis nuværende ord er w og sidste tag var Y
X Z PREVWD y : ændre tag X til tag Z hvis ordet før er y
X Z PREVTAG Y : ændre tag X til tag Z hvis tagget før er Y
X Z PREV1OR2OR3TAG Y : ændre tag X til tag Z hvis et af de 3 tidligere tags er Y

X Z PREVBIGRAM Y W : ændre tag X til tag Z ??? (Y og W er tags)
X Z LBIGRAM y w : ændre tag X til tag Z ??? (y og w er ord)

X Z SURROUNDTAG Y W : ændre tag X til tag Z ??? (Y og W er tags)
X Z CURWD y : ændre tag X til tag Z hvis ordet er y

* *dette er ikke en udtømmende liste. Brill har ikke selv forklaret reglerne (derfor forekommer ???)*

BILAG 2

Antal forekomster af de forskellige morfosyntaktiske analyser i PAROLE-korpuset

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

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
Interjectio		*lydord/udråbsor*			**I**	1

CatGram	**Attribute**	***(træk)***	**Value**	***(værdi)***	**Tag**	**Position**
n		*d*				
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
	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
	SsCatGra		Abbreviation	*forkortelse*	**A**	2

CatGram	Attribute	*(træk)*	Value	*(værdi)*	Tag	Position
	m					
			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

Den trænede taggers tag-sæt

Frekvens	Tag	Betydning	Eksempel
4119	NUM	cardinal adjektiv	to
6	NUM_GEN	cardinal adjektiv genitiv	2s
23327	ADJ	almindelig adjektiv	økonomiske
28	ADJ_GEN	almindelig adjektiv genitiv	radikales (fx. de radikales EF-ordfører)
419	NUM_ORD	ordinal adjektiv	3.
1	NUM_ORD_GEN	ordinal adjektiv genitiv	4.s (fx. Christian 4.s renæssanceslot)
9819	SKONJ	sideordnende konjunktion	og, men
5730	UKONJ	underordnende konjunktion	at, fordi, om, mens, ...
259	INTERJ	lydord/udråbsord	nå, ak, næh, ...
50660	N	appellativ	fremgangsmåde
1985	N_GEN	appellativ genitiv	husets
13188	EGEN	proprium	Eddie
1073	EGEN_GEN	proprium genitiv	Jugoslaviens
77	PRON_REC	reciprokt pronomen	hinanden
7	PRON_REC_GEN	reciprokt pronomen genitiv	hinandens
6939	PRON_DEMO	demonstrativ pronomen	den, det, de, denne, dette, disse
6	PRON_DEMO_GEN	demonstrativ pronomen genitiv	dennes, disses, ...
9490	PRON_UBST	ubestemt pronomen	en, et, noget, man, noget, nogen anden, andet, andre, .
22	PRON_UBST_GEN	ubestemt pronomen genitiv	ens, andres, ...
2433	PRON_POSS	possessivt pronomen	min, din , ...
13965	PRON_PERS	personligt pronomen	jeg, det, de, sig, ...
525	PRON_INTER.REL	interrogativt/relativt pronomen	hvem, hvad, hvilke, ...
29	PRON_INTER.REL_GEN	interrogativt/relativt pronomen genitiv	hvis
19017	ADV	generel adverbium	ud, ind, overfor
30927	PRÆP	præposition	i, ad, på
9037	UNIK	unik	som, er, at (...?)
9010	V_PAST	alm. verbum indikativ præteritum	gjorde, ville, lagde
19899	V_PRES	alm. verbum indikativ præsens	er, har, plejer
8627	V_INF	alm. verbum infinitiv	komme, æde, tage
56	V_GERUND	alm. verbum gerundium	medvirken, undren, stirren
437	V_IMP	alm. verbum imperativ	sæt, se,
6326	V_PARTC_PAST	alm. verbum participium præteritum	plantet, klippede, ansatte
667	V_PARTC_PRES	alm. verbum participium præsens	tilhørende, dinglende, manglende
71	V_MED_PAST	medial verbum indikativ præteritum	mislykkedes, syntes, lykkedes
220	V_MED_PRES	medial verbum indikativ præsens	findes, færdes, skændes
69	V_MED_INF	medial verbum infinitiv	enes, følges
16	V_MED_PARTC_PAST	medial verbum participium præteritum	lykkedes, enedes, skændtes

146	FORK	forkortelse	el., dr._scient._pol.
295	UL	udenlandsk ord	rock'n', roll, the
40391	TEGN	interpunktionstegn	" , . - ; : ! ? ( )
47	FORML	formler	1986:5, 16V, V8
169	SYMBOL	symboler	$, *, § , ...
1066	XX	andet	fejl mm

BILAG 3

TRÆNING AF BRILL-TAGGEREN

---

1) Del træningskorpus i 2 lige store dele:
cat “KORPUS” | perl RULE_BASED_TAGGER_V1.14/Utilities/
divide-in-two-rand.prl TAGGED-CORPUS TAGGED-CORPUS-2
(INPUT = “KORPUS” OUTPUT = TAGGED-CORPUS og TAGGED-CORPUS-2)

2) Fjern taggene fra korpus:
cat “KORPUS” | perl RULE_BASED_TAGGER_V1.14/Utilities/
tagged-to-untagged.prl > UNTAGGED-CORPUS
(INPUT = “KORPUS” OUTPUT = UNTAGGED-CORPUS)

3) Lav en liste over alle ord i det utaggede korpus:
cat UNTAGGED-CORPUS|perl RULE_BASED_TAGGER_V1.14/Utilities/
wordlist-make.prl |sort +1 -rn| awk "{print $1}" > BIGWORDLIST
(INPUT = UNTAGGED-CORPUS OUTPUT = BIGWORDLIST)

4) Lav en liste over alle ord m. tag fra 1. del af det taggede træningskorpus:
cat TAGGED-CORPUS|perl RULE_BASED_TAGGER_V1.14/Utilities/
word-tag-count.prl|sort +2 -rn > SMALLWORDTAGLIST
(INPUT = TAGGED-CORPUS OUTPUT = SMALLWORDTAGLIST)

5) Lav en liste over alle ordpar i det utaggede korpus:
cat UNTAGGED-CORPUS|perl RULE_BASED_TAGGER_V1.14/Utilities/
bigram-generate.prl|awk "{ print $1 $2 }" > BIGBIGRAMLIST

(INPUT = UNTAGGED-CORPUS OUTPUT = BIGBIGRAMLIST)

6) Nu kommer delen hvor de leksikalske regler skal læres:
perl RULE_BASED_TAGGER_V1.14/Learner_Code/unknown-lexical-learn.prl
BIGWORDLIST SMALLWORDTAGLIST BIGBIGRAMLIST 300 LEXRULEOUTFILE
(INPUT = BIGWORDLIST SMALLWORDTAGLIST BIGBIGRAMLIST
OUTPUT = LEXRULEOUTFILE)
OBS! husk at lave de initiale tags om fra NN/NNP til N/EGEN
i Laerner_Code/unknown-lexical-learn.prl
Laerner_Code/unknown-lexical-learn-continue.prlog
Tagger_Code/start-state-tagger.c

7) Kopier hele det taggede korpus over i en ny fil:
cp parole TAGGED-CORPUS-ENTIRE
(INPUT =parole OUTPUT = TAGGED-CORPUS-ENTIRE)

8) Lav et træningsleksikon:
cat TAGGED-CORPUS |perl RULE_BASED_TAGGER_V1.14/Utilities/
make-restricted-lexicon.prl > TRAINING.LEXICON

(INPUT = TAGGED-CORPUS OUTPUT = TRAINING.LEXICON)

9) Lav det leksikon som skal bruges til tagging når træningen er afsluttet:
cat parole.ny.FSTL |perl RULE_BASED_TAGGER_V1.14/Utilities/ make-restricted-lexicon.prl > FINAL.LEXICON
(INPUT = TAGGED-CORPUS-ENTIRE OUTPUT = FINAL.LEXICON)

10) Lav et utagget træningskorpus:
cat TAGGED-CORPUS-2| perl RULE_BASED_TAGGER_V1.14/Utilities/ tagged-to-untagged.prl > UNTAGGED-CORPUS-2
(INPUT = TAGGED-CORPUS-2 OUTPUT = UNTAGGED-CORPUS-2)

11) Her laves et "dummy-korpus" (korpus i initialt stadium??):
tagger ../../ trained_tagger/TRAINING.LEXICON ../../ trained_tagger/UNTAGGED-CORPUS-2
../../trained_tagger/BIGBIGRAMLIST ../../ trained_tagger/LEXRULEOUTFILE /dev/null -w ../../
trained_tagger/BIGWORDLIST -S > ../../ trained_tagger/DUMMY-TAGGED-CORPUS
(INPUT = TRAINING.LEXICON, UNTAGGED-CORPUS-2,
BIGBIGRAMLIST, LEXRULEOUTFILE
OUTPUT = DUMMY-TAGGED-CORPUS)

OBS! Vær opmærksom på at man SKAL stå nede i direktoriet hvor "tagger" ligger:
RULE_BASED_TAGGER_V1.14/Bin_and_Data

12) Nu skal de kontekstuelle regler laves:
RULE_BASED_TAGGER_V1.14/Bin_and_Data/contextual-rule-learn TAGGED-CORPUS-2
DUMMY-TAGGED-CORPUS CONTEXT-RULEFILE TRAINING.LEXICON
(INPUT = TAGGED-CORPUS-2, DUMMY-TAGGED-CORPUS, TRAINING.LEXICON
OUTPUT = CONTEXT-RULEFILE)

TAGGING MED BRILL-TAGGEREN

Stå i direktoriet RULE_BASED_TAGGER_V1.14/Bin_and_Data/.
Start taggeren:
tagger ../../ trained_tagger/FINAL.LEXICON ../../ trained_tagger/CORPUS ../../ trained_tag-ger/BIGBIGRAMLIST ../../ trained_tagger/LEXRULEOUTFILE trained_tagger/CONTEXT-RULEFILE > ../../ trained_tagger/CORPUS-TAGGED

INPUT = CORPUS det korpus der skal tagges dvs utagget
OUTPUT = CORPUS-TAGGED samme korpus nu tagget

Der er forskellige options man kan bruge:
-h :: help

-w wordlist :: provide an extra set of words beyond those in LEXICON.
See below.

-i filename :: writes intermediate results from start state tagger into filename

-s number :: processes the corpus to be tagged "number" lines at a time. This should be specified if memory problems result from trying to process too large a corpus at once. For example, on a Sparc 10 with 32 meg RAM, I usually process 250,000 words at a time. On a machine with 48 meg, I typically do 500,000 words. (note that "number" is the number of lines, not words). These numbers are just guidelines. You can test out what works best for you if you plan to tag large corpora.

-S :: use start state tagger only.

-F :: use final state tagger only. In this case, YOUR-CORPUS is a tagged corpus, whose taggings will be changed according to the final-state-tagger contextual rules. YOUR-CORPUS should be a tagged corpus ONLY when using this option.

The tagger writes to standard output.

TEST AF BRILL-TAGGEREN
---

1) Hvis taggeren skal testes, skal der bruges:
- et tagget test-korpus ("test")
- samme korpus i utagget form ("test.untagged")
Brug:
cat test | perl RULE_BASED_TAGGER_V1.14/Utilities/ tagged-to-untagged.prl > test.untagged
(INPUT = parole.test.FSTL OUTPUT = test.untagged)

2) Så tagges "test.untagged":
tagger ../../ trained_tagger/FINAL.LEXICON ../../ test.untagged ../../ trained_tagger/BIGBIGRAMLIST

../../ trained_tagger/LEXRULEOUTFILE ../../ trained_tagger/CONTEXT-RULEFILE
> ../../trained_tagger/test.tagged
(INPUT = test.untagged OUTPUT = test.tagged)
(Stå i direktoriet RULE_BASED_TAGGER_V1.14/Bin_and_Data/)

3) De 2 taggede korpora ("test" & "test.tagged") kan nu sammenlignes:
perl PROG.dif
(INPUT = test, test.tagged OUTPUT = test.dif)
(Stå i hjemmekataloget)

4) Recall og Precision udregnes ..... , fejl rettes og forbedringer laves

BILAG 5

FEJL pr. 1/11

Træningskorpus:	261904	ord
Testkorpus:	28698	ord
Antal fejl:	3,86	%
Fejl blandt kendte ord:	36	%
Fejl blandt ukendte ord:	64	%

1. 28,1% - 279 ord med mønstret [A-Z,Æ,Ø,Å][a-z,æ,ø,å]+\/
   dvs ord der begynder med stort
   - det er dels ord efter punktum
     (selvom ordformen stavet m. småt findes i leksikon)
   - dels egennavne som Parole ikke definerer som navne
     fx. "Mogenstrup/EGEN Grusgrav/N",
   - dels ord efter steder hvor der "burde" være et punktum,
     fx efter overskrifter, billedtekster osv.
   - samt andre (uforklarlige) fejl
     fx Mission/V_INF ist. Mission/N
     Poul-Erik/V_INF ist. Poul-Erik/EGEN

2. 7,3% - 73 ord med mønstret [A-Z,Æ,Ø,Å]+\/
   dvs. ord der består af flere store bogstaver
   - generelt kan taggeren ikke transformere store til små bogstaver

3. 6,7 %- 67 af fejlene hvor tagget skulle være "XX"
   dvs. der er tale om en fejl i korpus

4. 6,6%- 66 af ordene ender på -et
   skyldes evt. problemer med N vs. V_PARTC_PAST
5. 6,4%- 64 ord der ender på s
   dvs. problemer m genitiv eller evt.. problemer m V_PRES i korpus
6. 6,2%- 62 af ordene der ender på -er
   skyldes evt.. problemer med N vs. V_PRES
7. 3,9%- 39 af ordene ender på -ede
   skyldes evt. problemer med V_PAST vs. V_PARTC_PAST
8. 1,9%- 19 af ordene ender på -te
   skyldes evt. problemer med V_PAST vs. V_PARTC_PAST

BILAG 6

FEJLTYPER i det sidste (og bedste) forsøg (m. 1005 fejl = 3,5 %)

ANTAL FEJL

86	fejl i korpus (XX) dvs. taggeren giver et bud		
27	ord der KUN består af store bogstaver		(ud af 176 i testkorpus)
157	tagget skulle have været EGEN		(ud af 1546 i testkorpus)
	(af disse er 71 m. småt (ud af 1469 i testkorpus))		
63	ord der beg. m. stort og fejlagtigt tagget EGEN		
	(dvs. ikke alle ord m. stort er navne i Parole)		
117	ord der beg. m. småt og fejlagtigt tagget EGEN		
19	m. ordet "den"	PRON_DEMO, PRON_PERS	(ud af 359 i testkorpus)
18	m. ordet "det"	PRON_DEMO, PRON_PERS	(ud af 512 i testkorpus)
15	m. ordet "de"	PRON_DEMO, PRON_PERS	(ud af 302 i testkorpus)
16	m. ordet "så"	V_PAST, ADV, UKONJ, ADJ?, INTERJ	(ud af 241 i testkorpus)
22	m. ordet "for"	PRÆP, ADV, SKONJ	(ud af 345 i testkorpus)
126	m. tagget "PRÆP"		(ud af 3093 i testkorpus)
	(68 er fejlagtigt tagget PRÆP, resten skulle være PRÆP)		
57	m. suffixet "-s"		(ud af 1034 i testkorpus)
723	Fejl	dvs. 71,94 % af alle fejl	

”A-vitamin-artiklen” fra Den Store Danske Encyklopædi
indtastst manuelt

A-vitamin , fedtopløselig vitamin , der i kosten findes i flere former: retinol især i lever , fisk , mælk og æg; og som forstadium carotener , især i grøntsager og frugt med orange eller mørkegrønne farver , fx gulerødder , spinat , meloner og abrikoser .
Der findes omkring 50 forskellige carotener , hvoraf betacaroten er den vigtigste , og de kan alle omdannes i kroppen til retinol .
A-vitamin har en vigtig rolle i regulering af arveanlæggenes funktion i den enkelte celle (genekspression) og dermed i reguleringen af cellevækst .
Vitaminet er nødvendigt for dannelse af glykoproteiner , der er en vigtig del af slim , der dækker slimhinderne . A-vitamin indgår desuden i øjets synspigmenter og er nødvendigt , for at lys kan omdannes til synsindtryk .
Mangel på A-vitamin kan give natteblindhed og tørre slimhinder , der medfører øget modtagelighed for infektioner og , ved udtalt mangel , blindhed .
Leveren kan oplagre store mængder A-vitamin , og mangelsymptomer optræder først , hvis kosten i mange måneder har været uden A-vitamin .
I industrialiserede lande optræder mangel kun ved kronisk leversygdom og sygdomme , der medfører nedsadt fedtoptagelse fra tarmen .
Globalt er A-vitaminmangel et af de største ernæringsproblemer .
5 mio . mennesker får hvert år xerofthalmi (øjentørsot) og hos ½ mio . medfører det blindhed .
Lettere mangel , der medfører nedsadt modstandskraft over for infektioner , er meget hyppigere , og i mange ulande kan børnedødligheden reduceres med 25% , hvis børnene får dækket deres A-vitaminbehov .
For flere kræftformers vedkommende er det vist , at en stor indtagelse af grøntsager og frugt mindsker sygdomsrisikoen; årsagen er formodentlig bl .a . disse fødemidlers indhold af betacarotener .
For stor indtagelse af A-vitamin er skadelig og kan bl.a. give fostermisdannelser hos gravide .
Anbefalet daglig dosis er 800-1000 retinolækvivalenter for voksne .
KFMi
A-vitaminsyre (retinsyre , tretinoin) anvendes i cremer til medicinsk behandling af bumser (akne) .
Der kræves en nøje instruktion i brug af cremerne , da der optræder bivirkninger som svien , rødme og eksem .
I danmark er A-vitaminsyre ikke tilladt i kosmetik på grund af risikoen for bivirkninger .
Derivatet A-vitaminpalmitat anvendes i > antirynkecreme .

”A-vitamin-artiklen” fra Den Store Danske Encyklopædi
tagget manuelt

a-vitamin/N ,/TEGN fedtopløselig/ADJ vitamin/N ,/TEGN der/UNIK i/PRÆP kosten/N findes/V_MED_PRES i/PRÆP flere/ADJ former/N :/TEGN retinol/N især/ADV i/PRÆP lever/N ,/TEGN fisk/N ,/TEGN mælk/N og/SKONJ æg/N ;/TEGN og/SKONJ som/UNIK forstadium/N carotener/N ,/TEGN især/ADV i/PRÆP grøntsager/N og/SKONJ frugt/N med/PRÆP orange/ADJ eller/SKONJ mørkegrønne/ADJ farver/N ,/TEGN fx/ADV gulerødder/N ,/TEGN spinat/N ,/TEGN meloner/N og/SKONJ abrikoser/N ./TEGN
der/UNIK findes/V_MED_PRES omkring/PRÆP 50/NUM forskellige/ADJ carotener/N ,/TEGN hvoraf/ADV betacaroten/N er/V_PRES den/PRON_DEMO vigtigste/ADJ ,/TEGN og/SKONJ de/PRON_PERS kan/V_PRES alle/ADJ omdannes/V_INF i/PRÆP kroppen/N til/PRÆP retinol/N ./TEGN
a-vitamin/N har/V_PRES en/PRON_UBST vigtig/ADJ rolle/N i/PRÆP regulering/N af/PRÆP arveanlæggenes/N_GEN funktion/N i/PRÆP den/PRON_DEMO enkelte/ADJ celle/N (/TEGN genekspression/N)/TEGN og/SKONJ dermed/ADV i/PRÆP reguleringen/N af/PRÆP cellevækst/N ./TEGN
vitaminet/N er/V_PRES nødvendigt/ADJ for/PRÆP dannelse/N af/PRÆP glykoproteiner/N ,/TEGN der/UNIK er/V_PRES en/PRON_UBST vigtig/ADJ del/N af/PRÆP slim/N ,/TEGN der/UNIK dækker/V_PRES slimhinderne/N ./TEGN
a-vitamin/N indgår/V_PRES desuden/ADV i/PRÆP øjets/N_GEN synspigmenter/N og/SKONJ er/V_PRES nødvendigt/ADJ ,/TEGN for/PRÆP at/UKONJ lys/N kan/V_PRES omdannes/V_INF til/PRÆP synsindtryk/N ./TEGN
mangel/N på/PRÆP A-vitamin/N kan/V_PRES give/V_INF natteblindhed/N og/SKONJ tørre/ADJ slimhinder/N ,/TEGN der/UNIK medfører/V_PRES øget/V_PARTC_PAST modtagelighed/N for/PRÆP infektioner/N og/SKONJ ,/TEGN ved/PRÆP udtalt/V_PARTC_PAST mangel/N ,/TEGN blindhed/N ./TEGN
leveren/N kan/V_PRES oplagre/V_INF store/ADJ mængder/N A-vitamin/N ,/TEGN og/SKONJ mangelsymptomer/N optræder/V_PRES først/ADV ,/TEGN hvis/UKONJ kosten/N i/PRÆP mange/ADJ måneder/N har/V_PRES været/V_PARTC_PAST uden/PRÆP A-vitamin/N ./TEGN
i/PRÆP industrialiserede/V_PARTC_PAST lande/N optræder/V_PRES mangel/N kun/ADV ved/PRÆP kronisk/ADJ leversygdom/N og/SKONJ sygdomme/N ,/TEGN der/UNIK medfører/V_PRES nedsadt/V_PARTC_PAST fedtoptagelse/N fra/PRÆP tarmen/N ./TEGN
globalt/ADV er/V_PRES A-vitaminmangel/N et/PRON_UBST af/PRÆP de/PRON_DEMO største/ADJ ernæringsproblemer/N ./TEGN
5/NUM mio./N mennesker/N får/V_PRES hvert/PRON_UBST år/N xerofthalmi/N (/TEGN øjentørsot/N)/TEGN og/SKONJ hos/PRÆP ½/SYMBOL mio./N medfører/V_PRES det/PRON_PERS blindhed/N ./TEGN
lettere/ADJ mangel/N ,/TEGN der/UNIK medfører/V_PRES nedsadt/V_PARTC_PAST modstandskraft/N over/ADV for/PRÆP infektioner/N ,/TEGN er/V_PRES meget/ADJ hyppigere/ADJ ,/TEGN og/SKONJ i/PRÆP mange/ADJ ulande/N kan/V_PRES børnedødligheden/N reduceres/V_INF med/PRÆP 25/NUM %/SYMBOL ,/TEGN hvis/UKONJ børnene/N får/V_PRES dækket/V_PARTC_PAST deres/PRON_POSS A-vitaminbehov/N ./TEGN
for/PRÆP flere/ADJ kræftformers/N_GEN vedkommende/N er/V_PRES det/PRON_PERS vist/ADV ,/TEGN at/UKONJ en/PRON_UBST stor/ADJ indtagelse/N af/PRÆP grøntsager/N og/SKONJ frugt/N mindsker/V_PRES sygdomsrisikoen/N ;/TEGN årsagen/N er/V_PRES formodentlig/ADV bl.a./ADV disse/PRON_DEMO fødemidlers/N_GEN indhold/N af/PRÆP betacarotener/N ./TEGN
for/PRÆP stor/ADJ indtagelse/N af/PRÆP A-vitamin/N er/V_PRES skadelig/ADJ og/SKONJ kan/V_PRES bl.a./ADV give/V_INF fostermisdannelser/N hos/PRÆP gravide/ADJ ./TEGN
anbefalet/V_PARTC_PAST daglig/ADJ dosis/N er/V_PRES 800-1000/NUM retinolækvivalenter/N for/PRÆP voksne/ADJ ./TEGN
kfmi/EGEN
a-vitaminsyre/N (/TEGN retinsyre/N ,/TEGN tretinoin/N)/TEGN anvendes/V_PRES i/PRÆP cremer/N til/PRÆP medicinsk/ADJ behandling/N af/PRÆP bumser/N (/TEGN akne/N)/TEGN ./TEGN
der/UNIK kræves/V_PRES en/PRON_UBST nøje/ADJ instruktion/N i/PRÆP brug/N af/PRÆP cremerne/N ,/TEGN da/UKONJ der/UNIK optræder/V_PRES bivirkninger/N som/UNIK svien/V_GERUND ,/TEGN rødme/N og/SKONJ eksem/N ./TEGN
i/PRÆP Danmark/EGEN er/V_PRES A-vitaminsyre/N ikke/ADV tilladt/V_PARTC_PAST i/PRÆP kosmetik/N på/PRÆP grund/N af/PRÆP risikoen/N for/PRÆP bivirkninger/N ./TEGN
derivatet/N A-vitaminpalmitat/N anvendes/V_PRES i/PRÆP >/SYMBOL antirynkecreme/N ./TEGN

”A-vitamin-artiklen” fra Den Store Danske Encyklopædi
tagget af Brill-taggeren

a-vitamin/N ,/TEGN fedtopløselig/ADJ vitamin/N ,/TEGN der/UNIK i/PRÆP kosten/N findes/V_MED_PRES i/PRÆP flere/ADJ former/N :/TEGN retinol/N især/ADV i/PRÆP lever/N ,/TEGN fisk/N ,/TEGN mælk/N og/SKONJ æg/N ;/TEGN og/SKONJ som/UNIK forstadium/ADJ carotener/N ,/TEGN især/ADV i/PRÆP grøntsager/N og/SKONJ frugt/N med/PRÆP orange/ADJ eller/SKONJ mørkegrønne/ADJ farver/N ,/TEGN fx/ADV gulerødder/N ,/TEGN spinat/N ,/TEGN meloner/N og/SKONJ abrikoser/N ./TEGN
der/UNIK findes/V_MED_PRES omkring/PRÆP 50/NUM forskellige/ADJ carotener/N ,/TEGN hvoraf/ADV betacaroten/N er/V_PRES den/PRON_DEMO vigtigste/ADJ ,/TEGN og/SKONJ de/PRON_PERS kan/V_PRES alle/ADJ omdannes/V_INF i/PRÆP kroppen/N til/PRÆP retinol/N ./TEGN
a-vitamin/N har/V_PRES en/PRON_UBST vigtig/ADJ rolle/N i/PRÆP regulering/N af/PRÆP arveanlæggenes/N_GEN funktion/N i/PRÆP den/PRON_DEMO enkelte/ADJ celle/N (/TEGN genekspression/N)/TEGN og/SKONJ dermed/ADV i/PRÆP reguleringen/N af/PRÆP cellevækst/N ./TEGN
vitaminet/N er/V_PRES nødvendigt/ADJ for/PRÆP dannelse/N af/PRÆP glykoproteiner/N ,/TEGN der/UNIK er/V_PRES en/PRON_UBST vigtig/ADJ del/N af/PRÆP slim/N ,/TEGN der/UNIK dækker/V_PRES slimhinderne/N ./TEGN
a-vitamin/N indgår/V_PRES desuden/ADV i/PRÆP øjets/N_GEN synspigmenter/N og/SKONJ er/V_PRES nødvendigt/ADJ ,/TEGN for/PRÆP at/UKONJ lys/N kan/V_PRES omdannes/V_INF til/PRÆP synsindtryk/N ./TEGN
mangel/N på/PRÆP A-vitamin/N kan/V_PRES give/V_INF natteblindhed/N og/SKONJ tørre/ADJ slimhinder/N ,/TEGN der/UNIK medfører/V_PRES øget/V_PARTC_PAST modtagelighed/N for/PRÆP infektioner/N og/SKONJ ,/TEGN ved/PRÆP udtalt/V_PARTC_PAST mangel/N ,/TEGN blindhed/N ./TEGN
leveren/N kan/V_PRES oplagre/V_INF store/ADJ mængder/N A-vitamin/N ,/TEGN og/SKONJ mangelsymptomer/N optræder/V_PRES først/ADV ,/TEGN hvis/UKONJ kosten/N i/PRÆP mange/ADJ måneder/N har/V_PRES været/V_PARTC_PAST uden/PRÆP A-vitamin/N ./TEGN
i/PRÆP industrialiserede/V_PARTC_PAST lande/N optræder/V_PRES mangel/N kun/ADV ved/PRÆP kronisk/ADJ leversygdom/N og/SKONJ sygdomme/N ,/TEGN der/UNIK medfører/V_PRES nedsadt/V_PARTC_PAST fedtoptagelse/N fra/PRÆP tarmen/N ./TEGN
globalt/ADJ er/V_PRES A-vitaminmangel/N et/PRON_UBST af/PRÆP de/PRON_DEMO største/ADJ ernæringsproblemer/N ./TEGN
5/NUM mio./N mennesker/N får/V_PRES hvert/PRON_UBST år/N xerofthalmi/N (/TEGN øjentørsot/N)/TEGN og/SKONJ hos/PRÆP ½/ADJ mio./N medfører/V_PRES det/PRON_DEMO blindhed/N ./TEGN
lettere/ADJ mangel/N ,/TEGN der/UNIK medfører/V_PRES nedsadt/V_PARTC_PAST modstandskraft/N over/ADV for/PRÆP infektioner/N ,/TEGN er/V_PRES meget/ADJ hyppigere/ADJ ,/TEGN og/SKONJ i/PRÆP mange/ADJ ulande/N kan/V_PRES børnedødligheden/N reduceres/V_INF med/PRÆP 25/NUM %/SYMBOL ,/TEGN hvis/UKONJ børnene/N får/V_PRES dækket/V_PARTC_PAST deres/PRON_POSS A-vitaminbehov/N ./TEGN
for/PRÆP flere/ADJ kræftformers/N_GEN vedkommende/N er/V_PRES det/PRON_PERS vist/ADV ,/TEGN at/UKONJ en/PRON_UBST stor/ADJ indtagelse/N af/PRÆP grøntsager/N og/SKONJ frugt/N mindsker/V_PRES sygdomsrisikoen/N ;/TEGN årsagen/N er/V_PRES formodentlig/ADJ bl.a./ADV disse/PRON_DEMO fødemidlers/N_GEN indhold/N af/PRÆP betacarotener/N ./TEGN
for/PRÆP stor/ADJ indtagelse/N af/PRÆP A-vitamin/N er/V_PRES skadelig/ADJ og/SKONJ kan/V_PRES bl.a./ADV give/V_INF fostermisdannelser/N hos/PRÆP gravide/ADJ ./TEGN
anbefalet/V_PARTC_PAST daglig/ADJ dosis/N er/V_PRES 800-1000/NUM retinolækvivalenter/N for/PRÆP voksne/ADJ ./TEGN
kfmi/N
a-vitaminsyre/N (/TEGN retinsyre/N ,/TEGN tretinoin/N)/TEGN anvendes/V_INF i/PRÆP cremer/N til/PRÆP medicinsk/ADJ behandling/N af/PRÆP bumser/N (/TEGN akne/N)/TEGN ./TEGN
der/UNIK kræves/V_PRES en/PRON_UBST nøje/ADJ instruktion/N i/PRÆP brug/N af/PRÆP cremerne/N ,/TEGN da/UKONJ der/UNIK optræder/V_PRES bivirkninger/N som/UNIK svien/V_GERUND ,/TEGN rødme/N og/SKONJ eksem/N ./TEGN
i/PRÆP Danmark/EGEN er/V_PRES A-vitaminsyre/N ikke/ADV tilladt/V_PARTC_PAST i/PRÆP kosmetik/N på/PRÆP grund/N af/PRÆP risikoen/N for/PRÆP bivirkninger/N ./TEGN
derivatet/ADJ A-vitaminpalmitat/N anvendes/V_INF i/PRÆP >/ADJ antirynkecreme/N ./TEGN

Fejl i ”A-vitamin-artiklen” fra Den Store Danske Encyklopædi tagget af Brill-taggeren

10 fejl = 2,67 %

FEJL		RIGTIG
forstadium/ADJ	.....	forstadium/N
globalt/ADJ	.....	globalt/ADV
½/ADJ	.....	½/SYMBOL
det/PRON_DEMO	.....	det/PRON_PERS
formodentlig/ADJ	.....	formodentlig/ADV
kfmi/N	.....	kfmi/EGEN
anvendes/V_INF	.....	anvendes/V_PRES
derivatet/ADJ	.....	derivatet/N
anvendes/V_INF	.....	anvendes/V_PRES
>/ADJ	.....	>/SYMBOL

Fejlene i kontekst

a-vitamin/N ,/TEGN fedtopløselig/ADJ vitamin/N ,/TEGN der/UNIK i/PRÆP kosten/N findes/V_MED_PRES i/PRÆP flere/ADJ former/N :/TEGN retinol/N især/ADV i/PRÆP lever/V_PRES ,/TEGN fisk/N ,/TEGN mælk/N og/SKONJ æg/N ;/TEGN og/SKONJ som/UNIK ***forstadium***/ADJ carotener/N ,/TEGN især/ADV i/PRÆP grøntsager/N og/SKONJ frugt/N med/PRÆP orange/ADJ eller/SKONJ mørkegrønne/ADJ farver/N ,/TEGN fx/ADV gulerødder/N ,/TEGN spinat/N ,/TEGN meloner/N og/SKONJ abrikoser/N ./TEGN

globalt/ADJ er/V_PRES A-vitaminmangel/N et/PRON_UBST af/PRÆP de/PRON_DEMO største/ADJ ernæringsproblemer/N ./TEGN

5/NUM mio./N mennesker/N får/V_PRES hvert/PRON_UBST år/N xerofthalmi/N (/TEGN øjentørsot/N)/TEGN og/SKONJ hos/PRÆP **1/2**/ADJ mio./N medfører/V_PRES ***det***/PRON_DEMO blindhed/N ./TEGN

for/PRÆP flere/ADJ kræftformers/N_GEN vedkommende/N er/V_PRES det/PRON_PERS vist/V_PARTC_PAST ,/TEGN at/UKONJ en/PRON_UBST stor/ADJ indtagelse/N af/PRÆP grøntsager/N og/SKONJ frugt/N mindsker/V_PRES sygdomsrisikoen/N ;/TEGN årsagen/N er/V_PRES ***formodentlig***/ADJ bl.a./ADV disse/PRON_DEMO fødemidlers/N_GEN indhold/N af/PRÆP betacarotener/N ./TEGN

kfmi/N

a-vitaminsyre/N (/TEGN retinsyre/N ,/TEGN tretinoin/N)/TEGN ***anvendes***/V_PRES i/PRÆP cremer/N til/PRÆP medicinsk/ADJ behandling/N af/PRÆP bumser/N (/TEGN akne/N)/TEGN ./TEGN

derivatet/EGEN_GEN A-vitaminpalmitat/N ***anvendes***/V_PRES i/PRÆP >/ADJ antirynkecreme/N ./TEGN