Tekstversjon av Norsk aviskorpus (versjon 0.9)

Denne versjonen av tekstene fra Norsk aviskorpus er uferdig, og tekstene foreligger i tre ulike formater. I løpet av 2012 og 2013 vil tekstene blir ryddet opp i, og foreligge i et enhetlig xml-format. Korpuset er oppdatert per 28.12.2011.

Aviskorpuset blir tilgjengeliggjort for Språkbankens brukere, og kan kun benyttes til språkteknologisk forskning og utvikling. Brukere av korpuset har ikke lov til å videreformidle eller publisere noen del av tekstene, kun kunnskap og produkter utarbeidet med utgangspunkt i tekstene.

Ved spørsmål eller tilbakemeldinger angående dette korpuset, ta kontakt med Språkbanken på sprakbanken@nb.no.

Planer for oppgradering og oppdatering

Den ferdige versjonen vil ha en enhetlig katalogstruktur, materialet vil bli oppdatert med nyere tekster, og metadataformatet vil bli som i delkorpus 3 (se under).

Det foreligger noe materiale fra Gudbrandsølen Dagningen (GD), Morgenbladet (MB) og Vårt Land (VL). Dette materialet vil bli tilgjengelig når det er konvertert til egnet format.

Om Norsk aviskorpus

Ved det som i dag heter Uni Research AS har det siden 1998 blitt samlet inn et omfattende tekstmateriale bestående av norske avistekster. Databasen inneholder per 01.01.2012 omlag 1 milliard ord for bokmål, og 60 millioner ord for nynorsk, og er dermed den desidert største i sitt slag. Systemet innhenter automatisk store mengder tekst fra norske avisers nettsteder. Materialet vokser hver eneste dag. Av de 200.000-250.000 løpende ordene som daglig legges til i databasen, er 1000-1500 nyord. Samlet utgjør dette en verdifull kilde til informasjon om det norske språkets utvikling, nyorddanning, bruken av lånord og språklige bruksmønstre mer generelt.

Les mer om Norsk aviskorpus på prosjektets hjemmeside, http://avis.uib.no/.

1. Ti aviser fra 13.10.1998 til 03.04.2005

Dette delkorpuset inneholder de ti avisene (se under) som var med fra oppstarten av aviskorpus-prosjektet. Tekstene er samlet i 3 filer og dekker tidsrommet 13.10.1998-03.04.2005. Tekstene ligger i Corpus Workbench-format (et ord pr. linje, med skilletegn adskilt).

Eksempel:

```
<U #http://aftenbladet.no/kultur/article.jhtml?articleID=149472>
|
<B SA>
<A 03>
<M 01>
<D 17>
Historikeren
Olav
Nenseter
mener
at
det
var
et
benediktiner-kloster
i
Stavanger
på
1200-tallet
og
ikke
et
augustinerkloster
.
¶
```

Linjer som starter med "<" inneholder metadata. Følgende koder brukes for metadata:

- "U" angir url for den opprinnelige avisartikkelen
- "B" angir avis i form av en kode med to tegn (se under)
- "A" angir årstall (to siffer)
- "M" angir måned (to siffer)
- "D" angir dag (to siffer)

Videre markerer "¶" avsnitt/retur, mens "|" markerer start på tekst.

Delkorpuset inneholder følgende aviser:

- AA – Adresseavisen
- AP – Aftenposten
- BT – Bergens Tidende
- DA – Dagsavisen
- DB – Dagbladet
- DN – Dagens Næringsliv
- FV – Fædrelandsvennen
- NL – Nordlys
- SA – Stavanger Aftenblad
- VG – VG

2. Ti aviser fra 04.04.2005 til 28.12.2011

Tekstene i dette delkorpuset foreligger som én fil per avis (gzippet tar). Hver fil inneholder en katalog per år og en fil per dag. Filene er i et tekstformat før konvertering til Corpus WorkBench.

Eksempel:

```
##U #http://www.bt.no/bergenpuls/article358896>
##B BT>
##A 05>
##M 04>
##D 14>
a-ha i Frognerparken, Sissel på Festplassen¶ Hydro inviterer til gratis
folkefest både i Oslo og Bergen i forbindelse med 100-årsjubileumet. a-ha
blir hovedattraksjon i Frognerparken i Oslo, mens Sissel Kyrkjebø stiller
på Festplassen i Bergen.¶
```

Linjer som starter med "##" inneholder metadata. Følgende koder brukes for metadata:

- "U" angir url for den opprinnelige avisartikkelen
- "B" angir avis i form av en kode med to tegn (se under)
- "A" angir årstall (to siffer)
- "M" angir måned (to siffer)
- "D" angir dag (to siffer)

Videre markerer "¶" avsnitt/retur, mens "|" markerer start på tekst.

Delkorpuset inneholder følgende aviser:

- AA – Adresseavisen
- AP – Aftenposten
- BT – Bergens Tidende
- DA – Dagsavisen
- DB – Dagbladet
- DN – Dagens Næringsliv
- FV – Fædrelandsvennen
- NL – Nordlys
- SA – Stavanger Aftenblad
- VG – VG

3. Andre aviser

Tekstene i dette delkorpuset foreligger som én fil per avis. Hver fil inneholder to kataloger, NNO (nynorsk) og NOB (bokmål). I hver av disse katalogene er det én fil per artikkel. Materialet dekker i hovedsak tidsperioden 2006-2011, men for de fleste avisene foreligger det materiale tilbake til begynnelsen av 2000-tallet, fremskaffet gjennom arkivsøk.

Filene foreligger i xml-format:

Eksempel:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE document SYSTEM "http://gandalf.aksis.uib.no/aviskorpus/avis2.dtd">
<document>
<header>
<attribute name="file" value="NA/nno/2009/10/27/20091027-4668611.xml"/>
<attribute name="url" value="http://www.nationen.no/naring/article4668611.ece 233 2
3 1396"/>
<attribute name="source" value="NA"/>
<attribute name="date" value="2009-10-27 15:00"/>
<attribute name="author" value="Bjarne Bekkeheien Aase"/>
<attribute name="gender" value="M"/>
<attribute name="class1" value="naring"/>
<attribute name="class2" value=""/>
<attribute name="language" value="nno"/>
</header>
<body>
<div type="title" level="1">Riv seg laus frå Norske Felleskjøp</div>
<div type="ingress">Felleskjøpet Nordmøre og Romsdal skal heretter stå på eigne
bein på nett.</div>
<div type="author">BJARNE BEKKEHEIEN AASE</div>
<div type="date">Publisert 27.10.2009 kl 15:00 Oppdatert 27.10.2009 kl 19:56</div>
<div type="caption">Foto Arkivfoto Samarbeidet i Felleskjøp-familien raknar.</div>
<div type="text">
<p>Striden internt i Felleskjøp-familien held fram med uforminska styrke.</p>
<p>Rettsoppgjeret mellom Felleskjøpet Rogaland Agder (FKRA) og Felleskjøpet Agri
(FKA) om kornavtalen pågår i desse dagar for fullt i Eidsvoll tingrett.</p>
<p>Riv seg laus midt under rettssaka</p>
<p>Og som om ikkje det skulle vere nok går nå minstemann i familien, Felleskjøpet
Nordmøre og Romsdal, ut og riv seg laus frå overbygninga Norske Felleskjøp på nett
parallelt med at rettssaka pågår.</p>
<p>Etter at FKRA tidlegare i år trekte seg heilt ut av Norske Felleskjøp (NFK) i
kjølvatnet av kornstriden, sit FKA att med 96 prosent av NFK medan FKNR har fattige
4 prosent.</p>

…

<p>Nationen.no har fleire gonger prøvd å kome i kontakt med leiinga i FKNR for å få
ein kommentar til det som har skjedd, men det lukkast ikkje på tysdag.</p>
</div></body></document>
```

Følgende aviser foreligger på dette formatet:

- DT – Dag og Tid
- FI – Firda Media
- HD – Hallingdølen
- HO – Hordaland Bladdrift L/L
- KK – Klassekampen
- NA – Nationen
- SH – Sunnhordland
- SO – Sogn Avis
- SP – Sunnmørsposten
- VB – Vikebladet
- VT – Vest-Telemark Blad