N-gram 1-6, bokmål

Disse n-grammene er laget av Knut Hofland ved Uni Research AS, med utgangspunkt i tekstene som er samlet inn til Norsk aviskorpus (avis.uib.no) og deler av tekstmaterialet som ble samlet inn til tekstkorpuset til Nordisk språkteknologi (NST).

N-grammene kan benyttes fritt til språkteknologisk forskning og utvikling. Brukere oppfordres til å informere Språkbanken om hvilken nytte de har av denne ressursen.

Eventuelle spørsmål og tilbakemeldinger kan rettes til sprakbanken@nb.no.

1. Innhold

N-grammene er basert på følgende materiale:

Kilde	Tidsperiode	Størrelse*	Proveniens
Bergens Tidende	1990-1999	107	Tekstkorpuset til NST
NTB	1985-1998	121	Tekstkorpuset til NST
Aftenposten	1984-1998	282	Tekstkorpuset til NST
Nettaviser	1998-2011	665	Norsk aviskorpus[1]
Totalt	1985-2011	1175	

* Millioner ord

Materialet er delt inn i to deler etter proveniens, henholdsvis NSTs tekstkorpus (510 millioner ord) og Norsk aviskorpus (665 millioner ord). Det er også laget en sammenflettet del for hele materialet samlet (1175 millioner ord). Se filoversikten under for hvordan dette er organisert.

Generelt om ryddingen i tekstmaterialet

- <s> og </s> markerer setningsgrenser.
- Skilletegn er separert med blanke.
- Første ord i setning er erstattet med liten bokstav dersom ordet finnes skrevet slik i samme årgang av den enkelte avis. Ellers er ordene beholdt som de er skrevet.
- Setninger som ikke er avsluttet med stort skilletegn (.!?;:) er filtrert vekk (overskrifter o.l.).
- Setninger med mye tall (ofte resultatlister), oppramsinger o.l. er også filtrert vekk.
- For å filtrere vekk dubletter er setningene fra web-materialet sortert alfabetisk og kun ett eksemplar av hver setning er beholdt.
- Filene er sortert på Linux med LC_ALL=POSIX

[1] Følgende nettaviser inngår i Norsk aviskorpus: Adresseavisen, Aftenposten, Bergens Tidende, Dagbladet, Dag og Tid, Dagsavisen, Dagens Næringsliv, Firda Media, Fædrelandsvennen, Gudbrandsdølen Dagningen, Hallingdølen, Hordaland Bladdrift L/L, Klassekampen, Morgenbladet, Nationen, Nordlys, Sogn Avis, Stavanger Aftenblad, Sunnhordland, Sunnmørsposten, Vest-Telemark Blad, VG, Vikebladet, Vårt Land.

Antall linjer (= antall n-grammer)

	Sammenflettet versjon	Norsk aviskorpus	NSTs tekstkorpus
Ord:	~1.175.000.000	~665.000.000	~510.000.000
1-gram:	7.940.352	4.864.069	4.705.778
2-gram:	91.603.227	56.879.521	51.276.339
3-gram:	339.372.238	203.817.157	176.258.808
4-gram:	637.019.897	372.241.506	309.861.244
5-gram:	834.937.627	479.950.475	387.278.530
6-gram:	911.038.924	518.832.476	410.495.429

2. Filoversikt

2.1. Sammenflettet versjon, basert på 1175 millioner ord med løpende tekst

- Filarkivet **ngram_nob_1000.zip** (48 KB, utpakket 144 KB) inneholder de 1000 mest frekvente 1-grammer, 2-grammer, 3-grammer, 4-grammer, 5-grammer og 6-grammer, til sammen seks tekstfiler:

  ngram1-1-topp1000.txt
  ngram2-1-topp1000.txt
  ngram3-1-topp1000.txt
  ngram4-1-topp1000.txt
  ngram5-1-topp1000.txt
  ngram6-1-topp1000.txt

- Filarkivet **1gram_nob_abc.zip** (33 MB, utpakket 179 MB) inneholder en tekstfil med en liste over alle ord (1-grammer) i materialet. Ordene er listet alfabetisk.

- Filarkivet **1gram_nob_f1_abc.zip** (14 MB, utpakket 72 MB) inneholder en tekstfil med en liste over alle ord (1-grammer) i materialet med frekvens større enn 1. Ordene er listet alfabetisk.

- Filarkivet **1gram_nob_f1_freq.zip** (14 MB, utpakket 72 MB) inneholder en tekstfil med en liste over alle ord (1-grammer) i materialet med frekvens større enn 1. Ordene er listet etter fallende frekvens.

- Filarkivet **ngram_nob.tar.gz** (27 GB, utpakket 128 GB) inneholder hele n-gram-samlinga, og inneholder følgende filer, alle ren tekst:

Filnavn	**Innhold**
ngram1.srt	Alle 1-grammene, alfabetisk liste
ngram1-1.frk	1-gram, frekvenssortert (fallende), frekvens > 1
ngram1-1.srt	1-gram, alfabetisk liste, frekvens > 1
ngram1-1-topp1000.txt	De 1000 mest frekvente 1-grammene

Filnavn	Innhold
ngram2.srt	Alle 2-grammene, alfabetisk liste
ngram2-1.frk	2-gram, frekvenssortert (fallende), frekvens > 1
ngram2-1.srt	2-gram, alfabetisk liste, frekvens > 1
ngram2-1-topp1000.txt	De 1000 mest frekvente 2-grammene
ngram3.srt	Alle 3-grammene, alfabetisk liste
ngram3-1.frk	3-gram, frekvenssortert (fallende), frekvens > 1
ngram3-1.srt	3-gram, alfabetisk liste, frekvens > 1
ngram3-1-topp1000.txt	De 1000 mest frekvente 3-grammene
ngram4.srt	Alle 4-grammene, alfabetisk liste
ngram4-1.frk	4-gram, frekvenssortert (fallende), frekvens > 1
ngram4-1.srt	4-gram, alfabetisk liste, frekvens > 1
ngram4-1-topp1000.txt	De 1000 mest frekvente 4-grammene
ngram5.srt	Alle 5-grammene, alfabetisk liste
ngram5-1.frk	5-gram, frekvenssortert (fallende), frekvens > 1
ngram5-1.srt	5-gram, alfabetisk liste, frekvens > 1
ngram5-1-topp1000.txt	De 1000 mest frekvente 5-grammene
ngram6.srt	Alle 6-gramma, alfabetisk liste
ngram6-1.frk	6-gram, frekvenssortert (fallende), frekvens > 1
ngram6-1.srt	6-gram, alfabetisk liste, frekvens > 1
ngram6-1-topp1000.txt	De 1000 mest frekvente 6-grammene

2.2. Norsk aviskorpus, basert på 665 millioner ord med løpende tekst

- Filarkivet **ngram_nob_avis_1000.zip** (48 KB, utpakket 144 KB) inneholder de 1000 mest frekvente 1-grammer, 2-grammer, 3-grammer, 4-grammer, 5-grammer og 6-grammer, til sammen seks tekstfiler:

  ngram1-1-topp1000.txt
  ngram2-1-topp1000.txt
  ngram3-1-topp1000.txt
  ngram4-1-topp1000.txt
  ngram5-1-topp1000.txt
  ngram6-1-topp1000.txt

- Filarkivet **ngram_nob_avis.tar.gz** (16 GB, utpakket 73 GB) inneholder hele n-gram-samlinga, med følgende filer, alle ren tekst:

Filnavn	Innhold
ngram1.srt	Alle 1-grammene, alfabetisk liste
ngram1-1.frk	1-gram, frekvenssortert (fallende), frekvens > 1
ngram1-1.srt	1-gram, alfabetisk liste, frekvens > 1
ngram1-1-topp1000.txt	De 1000 mest frekvente 1-grammene

Filnavn	Innhold
ngram2.srt	Alle 2-grammene, alfabetisk liste
ngram2-1.frk	2-gram, frekvenssortert (fallende), frekvens > 1
ngram2-1.srt	2-gram, alfabetisk liste, frekvens > 1
ngram2-1-topp1000.txt	De 1000 mest frekvente 2-grammene
ngram3.srt	Alle 3-grammene, alfabetisk liste
ngram3-1.frk	3-gram, frekvenssortert (fallende), frekvens > 1
ngram3-1.srt	3-gram, alfabetisk liste, frekvens > 1
ngram3-1-topp1000.txt	De 1000 mest frekvente 3-grammene
ngram4.srt	Alle 4-grammene, alfabetisk liste
ngram4-1.frk	4-gram, frekvenssortert (fallende), frekvens > 1
ngram4-1.srt	4-gram, alfabetisk liste, frekvens > 1
ngram4-1-topp1000.txt	De 1000 mest frekvente 4-grammene
ngram5.srt	Alle 5-grammene, alfabetisk liste
ngram5-1.frk	5-gram, frekvenssortert (fallende), frekvens > 1
ngram5-1.srt	5-gram, alfabetisk liste, frekvens > 1
ngram5-1-topp1000.txt	De 1000 mest frekvente 5-grammene
ngram6.srt	Alle 6-gramma, alfabetisk liste
ngram6-1.frk	6-gram, frekvenssortert (fallende), frekvens > 1
ngram6-1.srt	6-gram, alfabetisk liste, frekvens > 1
ngram6-1-topp1000.txt	De 1000 mest frekvente 6-grammene

2.3. NSTs tekstkorpus, basert på 510 millioner ord med løpende tekst

- Filarkivet **ngram_nob_nst_1000.zip** (47 KB, utpakket 145 KB) inneholder de 1000 mest frekvente 1-grammer, 2-grammer, 3-grammer, 4-grammer, 5-grammer og 6-grammer, til sammen seks tekstfiler:

  ngram1-1-topp1000.txt
  ngram2-1-topp1000.txt
  ngram3-1-topp1000.txt
  ngram4-1-topp1000.txt
  ngram5-1-topp1000.txt
  ngram6-1-topp1000.txt

- Filarkivet **ngram_nob_nst.tar.gz** (13 GB, utpakket 61 GB) inneholder hele n-gram-samlinga, med følgende filer, alle ren tekst:

Filnavn	Innhold
ngram1.srt	Alle 1-grammene, alfabetisk liste
ngram1-1.frk	1-gram, frekvenssortert (fallende), frekvens > 1
ngram1-1.srt	1-gram, alfabetisk liste, frekvens > 1
ngram1-1-topp1000.txt	De 1000 mest frekvente 1-grammene

Filnavn	Innhold
ngram2.srt	Alle 2-grammene, alfabetisk liste
ngram2-1.frk	2-gram, frekvenssortert (fallende), frekvens > 1
ngram2-1.srt	2-gram, alfabetisk liste, frekvens > 1
ngram2-1-topp1000.txt	De 1000 mest frekvente 2-grammene
ngram3.srt	Alle 3-grammene, alfabetisk liste
ngram3-1.frk	3-gram, frekvenssortert (fallende), frekvens > 1
ngram3-1.srt	3-gram, alfabetisk liste, frekvens > 1
ngram3-1-topp1000.txt	De 1000 mest frekvente 3-grammene
ngram4.srt	Alle 4-grammene, alfabetisk liste
ngram4-1.frk	4-gram, frekvenssortert (fallende), frekvens > 1
ngram4-1.srt	4-gram, alfabetisk liste, frekvens > 1
ngram4-1-topp1000.txt	De 1000 mest frekvente 4-grammene
ngram5.srt	Alle 5-grammene, alfabetisk liste
ngram5-1.frk	5-gram, frekvenssortert (fallende), frekvens > 1
ngram5-1.srt	5-gram, alfabetisk liste, frekvens > 1
ngram5-1-topp1000.txt	De 1000 mest frekvente 5-grammene
ngram6.srt	Alle 6-gramma, alfabetisk liste
ngram6-1.frk	6-gram, frekvenssortert (fallende), frekvens > 1
ngram6-1.srt	6-gram, alfabetisk liste, frekvens > 1
ngram6-1-topp1000.txt	De 1000 mest frekvente 6-grammene