N-gram 1-6, nynorsk

Desse n-gramma er laga av Knut Hofland ved Uni Research AS, med utgangspunkt i tekstene som er samla inn til Norsk aviskorpus (avis.uib.no), og det som fanst av nynorske tekster i tekstkorpuset til Nordisk språkteknologi holding AS, som vart overført til Nasjonalbiblioteket i 2011.

N-gramma kan nyttast fritt. Brukarar vert oppfordra til å informere Språkbanken om kva nytte dei har av denne ressursen.

Eventuelle spørsmål og attendemeldingar kan rettast til sprakbanken@nb.no.

1. Innhald

N-gramma er baserte på følgjande materiale av nynorsk nyhendetekst, primært samla inn frå Internett-utgåvene til avisene:

Avis	**Tidsperiode**	**Storleik***	**Proveniens**
Bergens Tidende	1990-1999	12.8	Tekstkorpuset til NST
Bergens Tidende, Stavanger Aftenblad; innslag frå Adresseavisen, Aftenposten, Dagbladet	1998-2011	7.9	Norsk aviskorpus, opphavleg 10 aviser
Dag og Tid, Firda, Hallingdølen, Hordaland, Klassekampen, Nationen, Sunnhordland, Sogns Avis, Sunnmørsposten, Vikebladet og Vest-Telemark Blad	2000-2011	38.6	Tillegget til Norsk aviskorpus
Totalt	1990-2011	59.3	

* Millionar ord

- <s> og </s> markerer setningsgrenser.
- Skiljeteikn er separerte med blanke.
- Setningar som ikkje er avslutta med stort skiljeteikn (.!?;:) er filtrerte vekk (overskrifter o.l.).
- Setningar med mykje tal (ofte resultatlister), oppramsingar o.l. er filtrerte vekk.
- Filene er sorterte på Linux med LC_ALL=POSIX.

2. Filoversikt

- Filarkivet **ngram_nno_1000.zip** (42 KB, pakka ut 143 KB) inneheld dei 1000 mest frekvente 1-gram, 2-gram, 3-gram, 4-gram, 5-gram og 6-gram, til saman seks tekstfiler:

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

- Filarkivet **1gram_nno_abc.zip** (5 MB, pakka ut 24 MB) inneheld fila **1gram_nno_abc.srt** (rein tekst). Dette er ei liste over alle ord (1-gram) i materialet. Orda er lista alfabetisk.

- Filarkivet **1gram_nno_f1_abc.zip** (2 MB, pakka ut 10 MB) inneheld fila **1gram_nno_f1_abc.srt** (rein tekst). Dette er ei liste over alle ord (1-gram) i materialet med frekvens større enn 1. Orda er lista alfabetisk.

- Filarkivet **1gram_nno_f1_freq.zip** (2 MB, pakka ut 10 MB) inneheld fila **1gram_nno_f1_frq.srt** (rein tekst). Dette er ei liste over alle ord (1-gram) i materialet med frekvens større enn 1. Orda er lista etter fallande frekvens.

- Filarkivet **ngram_nno.tar** (1.8 GB, pakka ut 7.8 GB) inneheld følgjande filer, alle reine tekstfiler:

Filnamn	**Innhald**
ngram1.srt	Alle 1-gramma, alfabetisk liste
ngram1-1.frk	1-gram, frekvenssortert (fallande), frekvens > 1
ngram1-1.srt	1-gram, alfabetisk liste, frekvens > 1
ngram1-1-topp1000.txt	Dei 1000 mest frekvente 1-gramma
ngram2.srt	Alle 2-gramma, alfabetisk liste
ngram2-1.frk	2-gram, frekvenssortert (fallande), frekvens > 1
ngram2-1.srt	2-gram, alfabetisk liste, frekvens > 1
ngram2-1-topp1000.txt	Dei 1000 mest frekvente 2-gramma
ngram3.srt	Alle 3-gramma, alfabetisk liste
ngram3-1.frk	3-gram, frekvenssortert (fallande), frekvens > 1
ngram3-1.srt	3-gram, alfabetisk liste, frekvens > 1
ngram3-1-topp1000.txt	Dei 1000 mest frekvente 3-gramma
ngram4.srt	Alle 4-gramma, alfabetisk liste
ngram4-1.frk	4-gram, frekvenssortert (fallande), frekvens > 1
ngram4-1.srt	4-gram, alfabetisk liste, frekvens > 1
ngram4-1-topp1000.txt	Dei 1000 mest frekvente 4-gramma
ngram5.srt	Alle 5-gramma, alfabetisk liste
ngram5-1.frk	5-gram, frekvenssortert (fallande), frekvens > 1
ngram5-1.srt	5-gram, alfabetisk liste, frekvens > 1
ngram5-1-topp1000.txt	Dei 1000 mest frekvente 5-gramma
ngram6.srt	Alle 6-gramma, alfabetisk liste
ngram6-1.frk	6-gram, frekvenssortert (fallande), frekvens > 1
ngram6-1.srt	6-gram, alfabetisk liste, frekvens > 1
ngram6-1-topp1000.txt	Dei 1000 mest frekvente 6-gramma