

Hva er språkdata og hvordan kan vi få tak i dem?

Magnus Breder Birkenes
Språkbanken, Nasjonalbiblioteket

Hva er språkdata?

- “Alt er språk”, “Alt er tekst”
- Tekst
 - Hele tekstdokumenter, korpus (annotert, ikke-annotert) eller reduserte representasjoner av disse
 - Oversettelser (parallelltekst)
 - Ordbøker
 - Termlister
- Tale
 - (Store mengder) lydopptak av talt språk
 - Transkripsjoner av den samme talen: gjengivelse av lydene i skriftlig form
 - Uttaleleksikon
- Representativitet viktig (og problematisk)

Utvalgte bruksområder

- tekstdata:

- automatisk oversettelse mellom språk og målformer
- skrivestøtte: stavekontroll, grammatikkontroll, stilistiske tips
- samtaleroboter
- data mining: hente ut trekk i tekster
 - emneklassifisering/modellering
 - sentimentanalyse

- taledata:

- talegjenkjenning
- talesyntese

Hvor finner jeg språkdata for norsk?

- Språkbankens ressurskatalog:
<https://www.nb.no/sprakbanken/ressurskatalog/>
- Inneholder:
 - åpne språkdata fra Språkbanken ved Nasjonalbiblioteket (fortrinnsvis med svært liberale bruksbetingelser, CC-o)
 - data fra andre norske institusjoner som er medlem i CLARIN-samarbeidet (bruksbetingelser kan være strengere)

To eksempler

- Doffin: Parallellstilling og generering av oversettelsesminner
- Målfrid-korpuset: Massenedlasting og -prosessering av tekst fra statlige domener

Doffin

Doffin

- Doffin = nasjonal kunngjøringsdatabase for offentlige anskaffelser (hos DFØ):
<https://www.doffin.no/>
- Oversettelser til engelsk i TED Tenders Electronic Daily (TED): <https://ted.europa.eu>
- XML-dump av databasene
- Originaler og oversettelse forbundet gjennom dokument-identifikator

Søk kunngjøring med filter

Søk kjøperprofil

Oppdragsgivere

Leverandører

Brakerstøtte

Vis dokumenter

Se Ted kunngjøring

Kunngjøring

Anskaffelsesforskriften

Alminnelig kunngjøring av konkurranse

EØS-kunngjøring

Del I: Oppdragsgiver

I.1) Navn og adresser

Nasjonalbiblioteket
976029100
Henrik Ibsensgate 110
OSLO
0255
NO
Kontaktperson: Cyrille Nolin
E-post: cyrille.nolin@nb.no
NUTS-kode: NO - NORGE
Sted: hele Norge
Internetadresse(r):
Nettsted oppdragsgiver:
<https://permalink.mercell.com/74690480.aspx>
Nettsted Kjøperprofil: <http://www.nb.no/>

I.3) Kommunikasjon

Konkurransegrunnlaget er elektronisk tilgjengelig med gratis, direkte og ubegrenset tilgang på:
<https://permalink.mercell.com/74690480.aspx>
Tilleggsinformasjon finnes på en annen adresse:
Merrell Norge AS
Karihaugveien 89
Oslo
1086

Search TED for public tenders

[Advanced search](#) / [Expert search](#)

EUROPA > TED home > Display TED notice in current language

TED

TED SIMAP

TED eNotices

TED eTendering

041

OJ S current issue

2021

 [Release calendar](#)

[Check out our COVID-19 dedicated page for tenders related to medical equipment needs.](#)

[Brexit](#)

Supplies - 401940-2017

Original language

Data

Document family

 Share











11/10/2017 S195

I. II. III. IV. VI.

Norway-Oslo: Cinematographic film

2017/S 195-401940

Contract notice

Supplies

Legal Basis:

Directive 2014/24/EU

Section I: Contracting authority

I.1) **Name and addresses**


```

59      <CPV_CODE CODE="
60      92112000"/>
61    </CPV_ADDITIONAL>
62    <LOCATION code="071800"/>
63    <LOCATION code="010300"/>
64    <SHORT_DESCR>
65      <P>
        Nasjonalbiblioteket
        oppbevarer og
        konserverer
        35mm-filmer som
        skannes til DPX.
        Anbudet omfatter
        produksjon av
        DCP-filmer av nevnte
        DPX-filer og
        wave-filer.</P>
66      <P>Omfang, innhold
67      og krav til det som
68      skal kjøpes er
        inngående beskrevet
        i Appendix A
        -Specification
        requirements.</P>
69    </SHORT_DESCR>
70    <AC_PRICE/>
71    <VAL_OBJECT CURRENCY="NOK
72    ">60000000.00</VAL_OBJECT>
73    <DATE_START>2017-11-28</
    DATE_START>
74    <DATE_END>2019-11-28</
    DATE_END>
75    <RENEWAL/>
76    <RENEWAL_DESCR>
77      <P>Rammeavtalen skal
78      ha en varighet på 2
79      år med opsjon for
80      oppdragsgiver på
81      forlengelse på 1+1
82      år, slik at maksimal
83      varighet vil være 4
84      år.</P>
85      <P>Forlengelse skal

```

```

71      92112000"/>
72    </CPV_ADDITIONAL>
73    <n2016:NUTS CODE="N0071"
    />
74    <n2016:NUTS CODE="N0011"
    />
75    <SHORT_DESCR>
76      <P>The National
        Library stores and
        conserves 35mm films
        that are scanned to
        DPX. The tender is
        for the production
        of DCP films from
        the mentioned DPX
        files and wave files.
        </P>
77      <P>The extent,
78      content and
79      requirements for the
80      procurement are
81      described in detail
82      in Appendix A -
83      Specification
84      Requirements.</P>
85    </SHORT_DESCR>
86    <AC_PRICE/>
87    <VAL_OBJECT CURRENCY="NOK
88    ">60000000.00</VAL_OBJECT>
89    <DATE_START>2017-11-28</
    DATE_START>
90    <DATE_END>2019-11-28</
    DATE_END>
91    <RENEWAL/>
92    <RENEWAL_DESCR>
93      <P>The framework
        agreement will be
        valid for 2 years,
        with an option for
        the contracting
        authority to extend
        for 1+1 year, so
        that the maximum

```

Setningstokenisering

- Oppdeling av tekst i setninger
- Bruker ulike heuristikker:
 - Skilletegn: . ! ?
 - Stor/liten forbokstav
 - Forkortelser: *f.eks.*
 - Typiske ord setninger begynner med: *Jeg, det*
 - Kollokasjoner: *Johann S. Bach*
- Her: “Punkt Tokenizer” med norsk modell
- Trent på et større materiale av “gode setninger”

2017-677443_origianal.xml

2017-677443_translated.xml

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

<CPV_CODE CODE="92112000"/>

</CPV_ADDITIONAL>

<LOCATION code="071800"/>

<LOCATION code="010300"/>

<SHORT_DESCR>

<P>

Nasjonalbiblioteket oppbevarer og konserverer 35mm-filmer som skannes til DPX.

Anbudet omfatter produksjon av DCP-filmer av nevnte DPX-filer og wave-filer.</P>

<P>Umrang, innhold og krav til det som skal kjøpes er inngående beskrevet i Appendix A -Specification requirements.</P>

</SHORT_DESCR>

<AC_PRICE/>

<VAL_OBJECT CURRENCY="NOK">6000000.00</VAL_OBJECT>

<DATE_START>2017-11-28</DATE_START>

<DATE_END>2019-11-28</DATE_END>

<RENEWAL/>

<RENEWAL_DESCR>

<P>Rammeavtalen skal ha en varighet på 2 år med opsjon for oppdragsgiver på forlengelse på 1+1 år, slik at maksimal varighet vil være 4 år.</P>

<P>Forlengelse skal

71

72

73

74

75

76

77

78

79

80

81

82

83

84

</CPV_ADDITIONAL>

<n2016:NUTS CODE="N0071"/>

<n2016:NUTS CODE="N0011"/>

<SHORT_DESCR>

<P>The National Library stores and conserves 35mm films that are scanned to DPX. The tender is for the production of DCP films from the mentioned DPX files and wave files</P>

<P>The extent, content and requirements for the procurement are described in detail in Appendix A - Specification Requirements.</P>

</SHORT_DESCR>

<AC_PRICE/>

<VAL_OBJECT CURRENCY="NOK">6000000.00</VAL_OBJECT>

<DATE_START>2017-11-28</DATE_START>

<DATE_END>2019-11-28</DATE_END>

<RENEWAL/>

<RENEWAL_DESCR>

<P>The framework agreement will be valid for 2 years, with an option for the contracting authority to extend for 1+1 year, so that the maximum

ok, Line 5, Column 30

Tab Size: 4XML

1 Produksjon DCP-filmer
2 Produksjon av DCP-filmer.
3 Nasjonalbiblioteket oppbevarer og konserver
4 Anbudet omfatter produksjon av DCP-filmer a
5 Omfang, innhold og krav til det som skal kj
6 Rammeavtalen skal ha en varighet på 2 år me
7 Forlengelse skal gis til leverandøren senes
8 Forlengelse skal meddeles til leverandøren
9 Leverandøren skal være registrert i et fore
10 Vilklårene angitt i Appendix 1 - Terms of Co
11 Oslo
12 Eventuelle klager skal fremmes i tråd med g

1 Production of DCP films
2 Production of DCP films.
3 The National Library stores and conserves 3
4 The tender is for the production of DCP fil
5 The extent, content and requirements for th
6 The framework agreement will be valid for 2
7 The tenderer shall be notified of the exter
8 The tenderer shall be notified of the exter
9 Tenderers shall be registered in a company
10 The terms mentioned in Appendix 1 - Terms c
11 Oslo
12 Any appeals shall be filed in accordance wi

```
<tu>
196597   <tuv xml:lang="nob">
196598     <seg>Produksjon av DCP-filmer.</seg>
196599   </tuv>
196600   <tuv xml:lang="eng">
196601     <seg>Production of DCP films.</seg>
196602   </tuv>
196603 </tu>
196604 <tu>
196605   <tuv xml:lang="nob">
196606     <seg>Nasjonalbiblioteket oppbevarer og konserverer 35mm-filmer som skannes til
        DPX.</seg>
196607   </tuv>
196608   <tuv xml:lang="eng">
196609     <seg>The National Library stores and conserves 35mm films that are scanned to
        DPX.</seg>
196610   </tuv>
196611 </tu>
196612 <tu>
196613   <tuv xml:lang="nob">
196614     <seg>Anbudet omfatter produksjon av DCP-filmer av nevnte DPX-filer og
        wave-filer.</seg>
196615   </tuv>
196616   <tuv xml:lang="eng">
196617     <seg>The tender is for the production of DCP films from the mentioned DPX
        files and wave files.</seg>
196618   </tuv>
196619 </tu>
196620 <tu>
196621   <tuv xml:lang="nob">
196622     <seg>Omfang, innhold og krav til det som skal kjøpes er inngående
        beskrevet i Appendix A &#150; Specification requirements.</seg>
196623   </tuv>
196624   <tuv xml:lang="eng">
196625     <seg>The extent, content and requirements for the procurement are described in
        detail in Appendix A &#150; Specification Requirements.</seg>
```

Alignering

- en forholdsvis enkel jobb i dette tilfellet
 - alignert på dokumentnivå allerede
 - avsnitt i lik rekkefølge
- noen steder tilsvarer én setning på norsk flere setninger på engelsk og omvendt
- tilordning gjennom verktøyet HunAlign


```
<tu>
<tuv xml:lang="nob">
  <seg>Utstillingen "Vilje vakre Vega" vil vere
  hovedutstillingen ved verdensarvsenteret som
  skal formidle og fortolke områdes verdier som
  har gitt Vegaøyan en plass på Unescos
  verdensarvliste.</seg>
</tuv>
<tuv xml:lang="eng">
  <seg>The exhibit "Wild and beautiful Vega" shall
  be the main exhibit at the World Heritage
  Centre. ~~~ The exhibit shall communicate and
  interpret the area's values which have led to a
  place on the UNESCO's World Heritage list.</seg>
</tuv>
</tu>
```

Resultat

- 40.000 dokumentpar (original og oversettelse)
- 300.000 oversettelsespar (translation units) etter deduplisering
- datasettet er fritt tilgjengelig under:
<https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-63/>

Målfrid-korpuset

Målfrid

- samarbeidsprosjekt mellom Språkrådet og Nasjonalbiblioteket på oppdrag fra Kulturdepartementet
- mål: automatisere målformsrapportering (kravet om minst 25% nynorsk/bokmål i statlig informasjonsmateriale)
- middel: høsting av statlige nettsider, prosessering og automatisk språkdeteksjon av materiale

Høsting

- verktøy: *wget*
- rekursiv høsting ned til nivå 12 (det dypeste nivået vi fant under en prøveinnhøsting)
- kun tekstdokumenter (HTML, PDF, DOC)
- alle statlige domener med visse unntak
- alle undersider med unntak (f.eks. ikke metadataregistre)

Prosessering

- HTML: “Boilerplate removal”, Justext
 - navigasjonselementer og repeterende elementer, lenkelister fjernes
 - heuristikker. semantikken i HTML-elementer, stoppordtetthet, lenketetthet osv.
- PDF: full OCR med Google Vision API
- WORD/ODT: Python-biblioteket Textract
- alle dokumenter dedupliseres på domenenivå (sjekksum av redusert representasjon)



[> Forretningsplan](#)[> Om Bane NOR](#)[> Nasjonal transportplan](#)[> Organisasjon og ledelse](#)[> Eierskap og styring](#)[> Investor Relations](#)[> Jernbanereformen](#)[> Miljø](#)[> Etiske retningslinjer](#)[> Varslingskanal](#)[> Bane NORs historie](#)[> Arkiv Jernbaneverket](#)

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut 

Last ned dokumenter



Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf



Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken



Bane NOR SF

Postadresse

Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)

Schweigaards gate 33
0191 Oslo

Bane NOR kundesenter

Åpningstider

Mandag - fredag	08.00 - 19.00
Lørdag - søndag/helligdag	09.00 - 15.00

Chat med oss

Telefon

+47 477 70 008

Om nettstedet

[→ Nettstedskart](#)[→ Cookies \(informasjonskapsler\) på Bane NOR](#)[→ Personvernerklæring](#)

Ansvarlig redaktør

Konserndirektør Torild Lid Uribarri

Nettredaktør

- > Forretningsplan
- > **Om Bane NOR**
- > Nasjonal transportplan
- > Organisasjon og ledelse
- > Eierskap og styring
- > Investor Relations
- > Jernbanereformen
- > Miljø
- > Etiske retningslinjer
- > Varslingskanal
- > Bane NORs historie
- > Arkiv Jernbaneverket

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut

Last ned dokumenter



Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf



Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken



Bane NOR SF

Postadresse

Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)

Schweigaards gate 33
0191 Oslo

Bane NOR kundesenter

Åpningstider

Mandag - fredag	08.00 - 19.00
Lørdag - søndag/helligdag	09.00 - 15.00

Chat med oss

Telefon

1-800-457-70-000

Om nettstedet

- Nettstedskart
- Cookies (informasjonskapsler) på Bane NOR
- Personvernerklæring

Ansvarlig redaktør

Konserndirektør Torild Lid Uribarri

- > Forretningsplan
- > **Om Bane NOR**
- > Nasjonal transportplan
- > Organisasjon og ledelse
- > Eierskap og styring
- > Investor Relations
- > Jernbanereformen
- > Miljø
- > Etiske retningslinjer
- > Varslingskanal
- > Bane NORs historie
- > Arkiv Jernbaneverket

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut 

Last ned dokumenter



Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf



Brochure Bane NOR - We create the railway of the future.pdf

Publisert av: Bane NOR 12.06.2018

Del saken



Bane NOR SF

Postadresse

Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)

Schweigaards gate 33
0191 Oslo

Bane NOR kundesenter

Åpningstider

Mandag - fredag	08.00 - 19.00
Lørdag - søndag/helligdag	09.00 - 15.00

Chat med oss

Telefon

1-800-457-70-000

Om nettstedet

- Nettstedskart
- Cookies (informasjonskapsler) på Bane NOR
- Personvernerklæring

Ansvarlig redaktør

Konserndirektør Torild Lid Uribarri

- > Forretningsplan
- > Om Bane NOR
- > Nasjonal transportplan
- > Organisasjon og ledelse
- > Eierskap og styring
- > Investor Relations
- > Jernbanereformen
- > Miljø
- > Etiske retningslinjer
- > Varslingskanal
- > Bane NORs historie
- > Arkiv Jernbaneverket

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt

S SøkerherSøk

Meny

Forside

Om oss

Om Bane NOR

Om Bane NOR

Bane NOR er et statlig foretak med ansvar for den nasjonale jernbaneinfrastrukturen.

Bane NORs formål er å sørge for tilgjengelig jernbaneinfrastruktur og effektive og brukervennlige tjenester, inkludert knutepunkts- og godsterminalutvikling.

Bane NOR har ansvaret for planlegging, utbygging, forvaltning, drift og vedlikehold av det nasjonale jernbanenettet, trafikkstyring og forvaltning og utvikling av jernbaneeiendom. Bane NOR har det operative koordineringsansvaret for sikkerhetsarbeidet og operativt ansvar for samordning av beredskap og krisehåndtering.

Bane NOR har om lag 3 400 ansatte og har hovedkontor i Oslo.

Bane NOR SF er 100 prosent eid av staten og er underlagt Samferdselsdepartementet.

Skriv ut

Last ned dokumenter

- Brosjyre Bane NOR - Vi skaper fremtidens jernbane.pdf
- Brochure Bane NOR - We create the railway of the future.pdf

Bane NOR SF

Postadresse
Postboks 4350
2308 Hamar

Hovedkontor (besøksadresse)
Schweigaards gate 33
0191 Oslo

final class	good
cotext-free class	good
heading	False
length (in characters)	157
number of characters within links	0
link density	0.000
number of words	18
number of stopwords	6
stopword density	0.333
html.body.main.div.div.article.div.p	

VEDTEKTER FOR STATENS LÅNEKASSE FOR UTDANNING

Fastsatt av Kunnskapsdepartementet den 15.januar 2016 i medhold av lov 3. juni 2005 nr. 37 om utdanningsstøtte.

1 Formål

Statens lånekasse for utdanning (Lånekassen) skal forvalte utdanningsstøtten i samsvar med bestemmelsene gitt i eller i medhold av utdanningsstøtteloven. Rammene for virksomheten til Lånekassen fastsettes for øvrig av overordnet myndighet.

Utfyllende bestemmelser om virksomhetens formål framgår av Hovedinstruks for Lånekassen.

2 Lånekassens ledelse og oppgaver

Lånekassen er et ordinært forvaltningsorgan som ledes av eget styre og tilsatt administrerende direktør.

Styret er Lånekassens øverste organ og er ansvarlig for den samlede virksomheten. Styret skal blant annet påse at Lånekassen drives i samsvar med regelverket, samt styringssignaler og retningslinjer gitt av overordnede myndigheter. Styret skal sikre at fastsatte mål oppnås og at ressursbruken er effektiv. Styret har også ansvar for at det foreligger vurderinger som identifiserer de viktigste risikofaktorene knyttet til virksomheten og at det foreligger planer for oppfølging av disse faktorene.

```
<div class='ocr_carea' id='block_2' title='bbox 100 153 169 161'>
  <p class='ocr_par' dir='ltr' id='par_2_0' title='bbox 100 153 169 161'>
    <span class='ocr_line' id='line_2_0_0' title='bbox 100 153 169 161; baseline 0 0'>
      <span class='ocrx_word' id='word_2_0_0' title='bbox 100 153 108 161'><i>/span>
      <span class='ocrx_word' id='word_2_0_1' title='bbox 114 153 161 161'>Formål</span>
    </span>
    <span class='ocr_line' id='line_2_0_1' title='bbox 114 153 169 161; baseline 0 0'>
      </span></p></div>

<div class='ocr_carea' id='block_3' title='bbox 100 174 201 21'>
  <p class='ocr_par' dir='ltr' id='par_3_0' title='bbox 100 174 201 21'>
    <span class='ocr_line' id='line_3_0_0' title='bbox 100 174 201 21; baseline 0 0'>
      <span class='ocrx_word' id='word_3_0_0' title='bbox 100 174 148 182'>Statens</span>
      <span class='ocrx_word' id='word_3_0_1' title='bbox 154 174 182 182'>>lånekasse</span>
      <span class='ocrx_word' id='word_3_0_2' title='bbox 182 174 212 182'>>for</span>
      <span class='ocrx_word' id='word_3_0_3' title='bbox 212 174 242 182'>>utdanning</span>
      <span class='ocrx_word' id='word_3_0_4' title='bbox 242 174 282 182'>>(Lånekassen)</span>
      <span class='ocrx_word' id='word_3_0_5' title='bbox 282 174 312 182'>>skal</span>
      <span class='ocrx_word' id='word_3_0_6' title='bbox 312 174 342 182'>>forvalte</span>
      <span class='ocrx_word' id='word_3_0_7' title='bbox 342 174 372 182'>>utdanningsstøtten</span>
      <span class='ocrx_word' id='word_3_0_8' title='bbox 372 174 402 182'>>i</span>
      <span class='ocrx_word' id='word_3_0_9' title='bbox 402 174 432 182'>>sansvar</span>
      <span class='ocrx_word' id='word_3_0_10' title='bbox 432 174 462 182'>>med</span>
    </span>
    <span class='ocr_line' id='line_3_0_1' title='bbox 462 174 492 182; baseline 0 0'>
      <span class='ocrx_word' id='word_3_0_11' title='bbox 462 174 492 182'>>bestemmelsene</span>
      <span class='ocrx_word' id='word_3_0_12' title='bbox 492 174 522 182'>>gitt</span>
      <span class='ocrx_word' id='word_3_0_13' title='bbox 522 174 552 182'>>i</span>
      <span class='ocrx_word' id='word_3_0_14' title='bbox 552 174 582 182'>>eller</span>
      <span class='ocrx_word' id='word_3_0_15' title='bbox 582 174 612 182'>>i</span>
      <span class='ocrx_word' id='word_3_0_16' title='bbox 612 174 642 182'>>medhold</span>
      <span class='ocrx_word' id='word_3_0_17' title='bbox 642 174 672 182'>>av</span>
      <span class='ocrx_word' id='word_3_0_18' title='bbox 672 174 702 182'>>utdanningsstøtteloven.</span>
      <span class='ocrx_word' id='word_3_0_19' title='bbox 702 174 732 182'>>Rammene</span>
      <span class='ocrx_word' id='word_3_0_20' title='bbox 732 174 762 182'>>for</span>
      <span class='ocrx_word' id='word_3_0_21' title='bbox 762 174 792 182'>>virksomheten</span>
      <span class='ocrx_word' id='word_3_0_22' title='bbox 792 174 822 182'>>til</span>
    </span>
    <span class='ocr_line' id='line_3_0_2' title='bbox 822 174 852 182; baseline 0 0'>
      <span class='ocrx_word' id='word_3_0_23' title='bbox 822 174 852 182'>>Lånekassen</span>
      <span class='ocrx_word' id='word_3_0_24' title='bbox 852 174 882 182'>>fastsettes</span>
      <span class='ocrx_word' id='word_3_0_25' title='bbox 882 174 912 182'>>for</span>
      <span class='ocrx_word' id='word_3_0_26' title='bbox 912 174 942 182'>>øvrig</span>
      <span class='ocrx_word' id='word_3_0_27' title='bbox 942 174 972 182'>>av</span>
      <span class='ocrx_word' id='word_3_0_28' title='bbox 972 174 1000 182'>>overordnet</span>
      <span class='ocrx_word' id='word_3_0_29' title='bbox 1000 174 1000 182'>>myndighet.</span>
    </span>
    <span class='ocr_line' id='line_3_0_3' title='bbox 852 174 882 182; baseline 0 0'>
      </span></p></div>
```

Språkdeteksjon

- Modeller for bokmål, nynorsk, samiske språk og andre språk (utviklet av UIT/Giellatekno)
- Frekvens av bokstavsekvenser og enkeltord
- Algoritme: Out-of-place-ranking (Cavnar/Trenkle 1994)
- Implementasjon: Textcat
- Klassifikasjon på dokument- og avsnittnivå

Språkdeteksjon på avsnitt

↑↓	paragraph_nr ↑↓	line ↑↓	lang ↑↓	tokens ↑↓
0	0	VEDTEKTER FOR STATENS LÅNEKASSE FOR UTDANNING		6
1	1	Fastsatt av Kunnskapsdepartementet den 15.januar 2016 i medhold av lov 3. juni 2005 nr. 37 om utdanningsstøtte.	nob	17
2	2	1 Formål		2
3	3	Statens lånekasse for utdanning (Lånekassen) skal forvalte utdanningsstøtten i samsvar med bestemmelsene gitt i eller i medhold av utdanningsstøtteloven. Rammene for virksomheten til Lånekassen fastsettes for øvrig av overordnet myndighet.	nob	30
4	4	Utfyllende bestemmelser om virksomhetens formål framgår av Hovedinstruks for Lånekassen.		10
5	5	2 Lånekassens ledelse og oppgaver		5
6	6	Lånekassen er et ordinært forvaltningsorgan som ledes av eget styre og tilsatt administrerende direktør.	nob	14
7	7	Styret er Lånekassens øverste organ og er ansvarlig for den samlede virksomheten. Styret skal blant annet påse at Lånekassen drives i samsvar med regelverket, samt styringssignaler og retningslinjer gitt av overordnede myndigheter. Styret skal sikre at fastsatte mål oppnås og at ressursbruken er effektiv. Styret har også ansvar for at det foreligger vurderinger som identifiserer de viktigste risikofaktorene knyttet til virksomheten og at det foreligger planer for oppfølging av disse faktorene.	nob	71
8	8	Styret skal fastsette Lånekassens strategi og sørge for at denne blir realisert. Styret skal behandle forslag fra Lånekassen om vesentlige endringer i utdanningsstøtteordningen, samt øvrige viktige saker for Lånekassens virksomhet.	nob	30
9	9	Styret ansetter administrerende direktør på åremål. Betingelsene for ansettelsesforholdet fastsettes av Kunnskapsdepartementet.	nob	12

Korpus

- 4,3 milliarder løpende ord (deduplisert)
 - 3,2 milliarder på bokmål
 - 765 millioner på engelsk
 - 289 millioner på nynorsk
 - 5,8 millioner på nordsamisk
- Muligens den største frie tekstressursen for norsk
 - Common Crawl (OSCAR):
 - Bokmål: 804 millioner
 - Nynorsk: 9,4 millioner

Emnemodellering: Helsedir.

Topic 0: kommune, pasient, helse-, pårørend, besøk, omsorgstjeneste, tjeneste, behov, ansette, helsepersonell, råd, anbefaling, sykehjem, kommunal, kompetanse, beboer, nødvendig, omsorgsinstitusjone, vurdere, informasjon, institusjon, smitte, sikre, fastlege, oppfølging

Topic 1: prosent, pasient, helsevern, psykisk, antall, tertial, poliklinisk, aktivitet, nedgang, konsultasjon, periode, endring, sykehus, kontakt, ventetid, døgnopphold, andel, datum, sammenligne, voksen, fastlege, måned, talle, somatisk, behandling

Topic 2: metode, behandling, metodevurdering, pasient, legemiddel, legemiddelverk, drøfting, innspill, metodevurderinge, beslutning, kostnad, effekt, klinisk, studium, vurdering, forslag, al, metodevarsel, rapport, hurtig, fagdirektør, dokumentasjon, oppdrag, indikasjon, fullstendig

Topic 3: barn, skole, barnehage, elev, ung, veileder, ansette, råd, tiltak, skolehelsetjeneste, smitte, sykdom, tjeneste, avstand, gravid, behov, ungdom, kohort, møte, anbefaling, foresette, faglig, helsestasjons-, drift, ivareta

Topic 4: person, tiltak, kommune, smitte, karantene, reise, arrangement, avstand, land, test, oppdrag, covid, teste, testing, utbrudd, uke, nærkontakt, vurdering, unntak, hurtigtest, symptom, innreisekarante, område, antall, tilfelle

Emnemodellering: Helsedir.

01.

https://www.helsedirektoratet.no/rapporter/tiltak-pa-skole-og-barnehageområdet-under-koronautbruddet-varen-2020/Tiltak%20p%C3%A5%20skole-%20og%20barnehageomr%C3%A5det%20under%20koronautbruddet%20v%C3%A5ren%202020.pdf/_attachment/inline/1811d255-4e8d-4c8f-8433-c97c002df1c0:85099ff1d203787e87babbacfd6ee4bddfc9658/Tiltak%20p%C3%A5%20skole-%20og%20barnehageomr%C3%A5det%20under%20koronautbruddet%20v%C3%A5ren%202020.pdf

02. <https://www.helsedirektoratet.no/veiledere/smittevern-for-skoletrinn-1-7-covid-19/bakgrunn>

03. <https://www.helsedirektoratet.no/veiledere/smittevern-i-ungdomsskole-og-videregaende-skole-covid-19/bakgrunn>

04.

https://www.helsedirektoratet.no/tema/beredskap-og-krisehandtering/koronavirus/anbefalinger-og-beslutninger/Presentasjon%20rapport%20ekspertgruppe%203.%20april%202020.pdf/_attachment/inline/c4f51efd-ac8c-46ea-8a89-91e086f3e693:c1d87b7cc09d4c11612355f9b2978729297f61a6/Presentasjon%20rapport%20ekspertgruppe%203.%20april%202020.pdf

05. <https://www.helsedirektoratet.no/veiledere/covid-19-smittevern-i-barnehager/bakgrunn>

06.

<https://www.helsedirektoratet.no/veiledere/smittevern-for-skoletrinn-1-7-covid-19/er-det-noen-barn-og-ansatte-det-ma-tas-spesielle-hensyn-til>

07.

<https://www.helsedirektoratet.no/veiledere/covid-19-smittevern-i-barnehager/er-det-barn-og-ansatte-det-ma-tas-spesielle-hensyn-til>

08. <https://www.helsedirektoratet.no/nyheter/kunnskap-om-barn-og-koronasmitte>

09.

<https://www.helsedirektoratet.no/veiledere/smittevern-i-ungdomsskole-og-videregaende-skole-covid-19/er-det-noen-ungdom-og-ansatte-det-ma-tas-spesielle-hensyn-til>

10. <https://www.helsedirektoratet.no/nyheter/oppdaterer-veiledere-anbefalinger-og-rad-om-barn-og-unge>



Veien videre

Avleveringsløsning for språkdata

- Språkbanken ønsker å etablere en digital avleveringsløsning for språkdata
- Skal gjøre det mulig for statlige institusjoner og private virksomheter å laste opp (anonymisert) materiale som kan publiseres i Språkbankens ressurskatalog

Takk for oppmerksomheten!