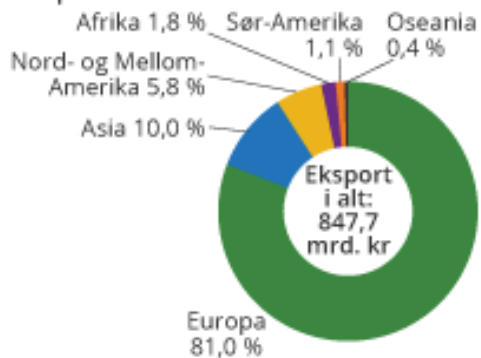


“Norsk i den digitale tidsalderen — Maskinoversettelse: hvordan fungerer det?”

Koenraad De Smedt (Universitetet i Bergen)

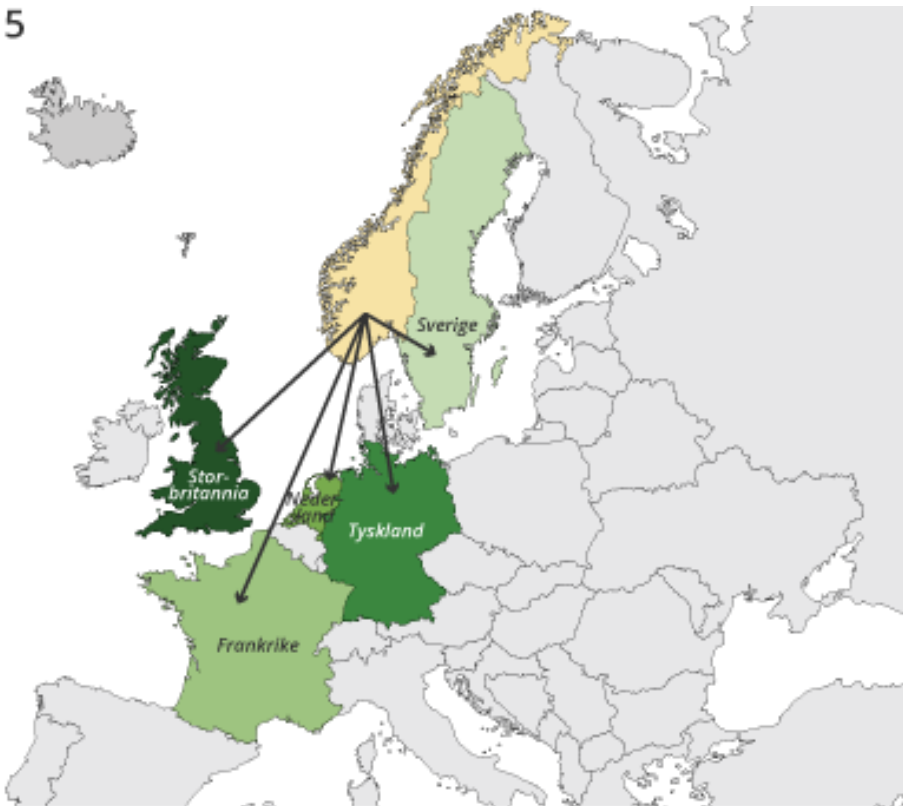
Figur 7. Eksport av varer. 2015

Eksport til ulike verdensdeler



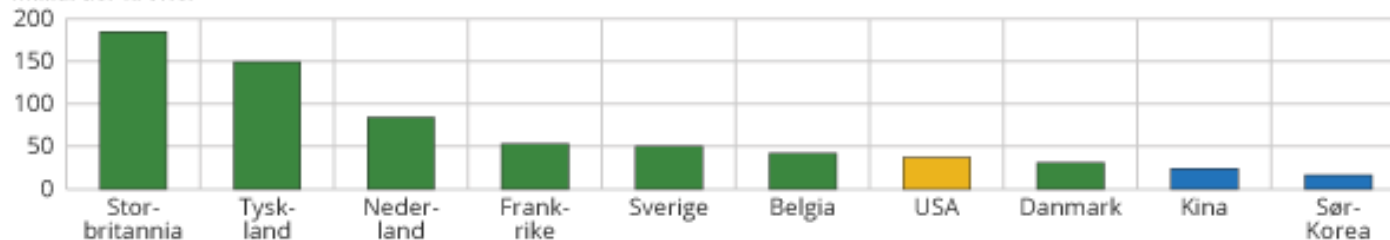
Eksport til de fem største handelspartnerne i Europa

Land	Mrd. kr
Storbritannia	184,8
Tyskland	149,0
Nederland	84,0
Frankrike	52,8
Sverige	50,4

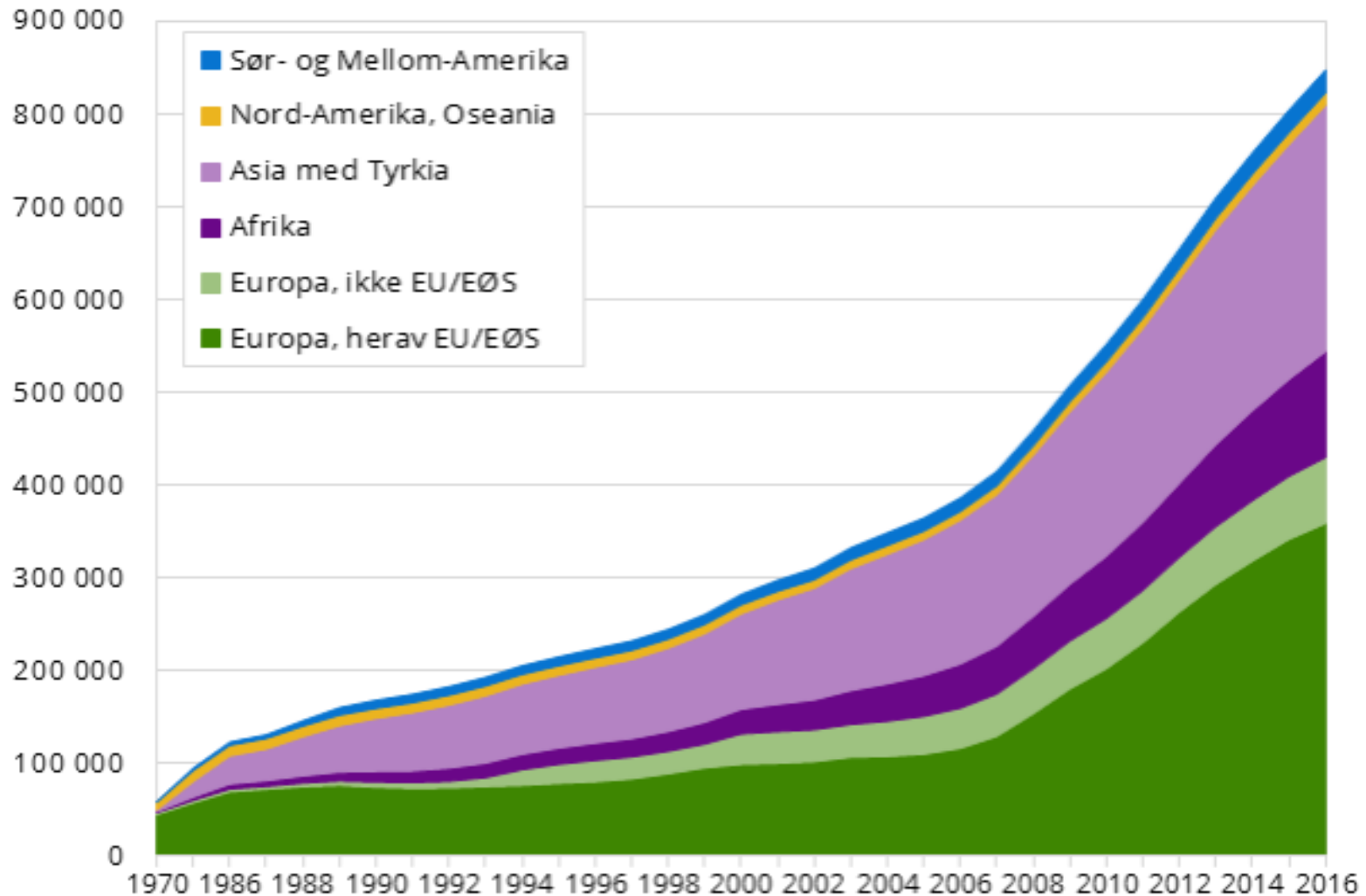


Eksport til Norges ti største handelspartnere på verdensbasis

Milliarder kroner



Innvandrere og innvandrerbakgrunn



- Engelsk kommer inn på viktige områder og i viktige sektorer: handel, industri, utdanning, osv.
- Norsk er stadig under press
- Flerspråklige IKT-tjenester som inkluderer norsk språk bør utvikles

Behov for kommunikasjon og tjenester



- Fremtidens digitale løsninger forutsetter digital språkbehandling
- Språklig og kulturell kontekst er viktig i utvikling av intelligente IKT-løsninger

Informasjonsteknologi som inkluderer modeller for språkbehandling, bl.a.

- oversettelse
- informasjonssøk
- sortering av epost
- dokumentklassifisering
- diktering (talegjenkjenning)
- opplesning (talesyntese)
- OSV.

Språkteknologi bygger alltid på informasjon om det språket som skal behandles

Den dominerende metoden er selvlærende systemer, som mates med store mengder språkdata som gjenspeiler det ønskede resultatet, f.eks:

- tekster med oversettelser
- tekster der det er angitt hva som er relevant
- epost som er sortert i spam og ikke-spam
- dokumenter som sortert i kategorier
- tale og transkripsjoner
- OSV.

Har man ikke store mengder språkdata, finnes det også regelbaserte teknologier. Disse krever likevel at man tester ut reglene på språkdata. Eksempler for bruk av regler:

- grammatikker
- analyse av sammensetninger
- finne setningsgrenser i tekst
- stavekontroll
- dialogstrategier

Det finnes også hybridsystemer

- Kan gjerne utvikles i internasjonalt samarbeid, men et norsk bidrag er avgjørende
- Språkteknologisk utvikling trenger tilgang til data for norsk (også minoritetsspråk og andre språk)
- Bør utvikles i en norsk kontekst (uttrykk, navn, terminologi)
- Ikke bare oversettelse fra og til engelsk, men fra og til flere språk som er viktige fra et norsk perspektiv
- Trenger samarbeid mellom forskere, industri, offentlig sektor og brukere

META NET

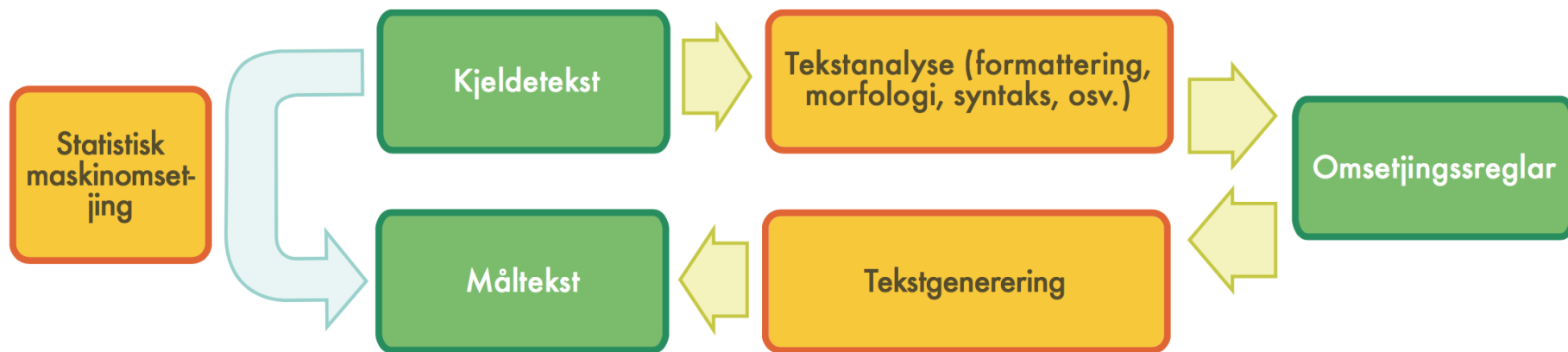
White Paper Series Kvitbokserie

THE NORSK
NORWEGIAN I DEN
LANGUAGE IN DIGITALE
THE DIGITAL TIDSALDEREN
AGE

NYNORSKVERSJON

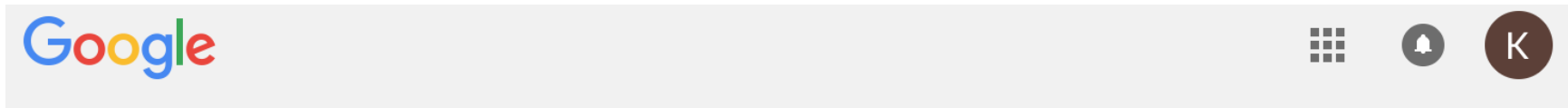
	Kvantitet	Tilgjengelegheit	Kvalitet	Dekningsgrad	Modenheit	Berekraft	Tilpassingsdyktigheit
Språkteknologi (verktøy, teknologiar og applikasjonar)							
Taleattkjenning	4	2	2	1	2	3	3
Talesyntese	3	2	3	2	3	3	3
Grammatisk analyse	4	4,5	4	4	4,5	4,5	5
Semantisk analyse	2	2	3,3	3	3,7	3,3	3,7
Tekstgenerering	1	4	4	3	5	4	5
Maskinomsetjing	4	4	2	2	3	5	3
Språkressursar (ressurs-, data- og kunnskapsbasar)							
Tekstkorpus	4,5	3,5	3,5	3	4	4,5	4
Talekorpus	5	4	3	5	4	5	5
Parallellkorpus	5	3	2	2	4	3	3
Leksikalske ressursar	2,5	2	2	2	2	2	2,5
Grammatikkar	2	4	5	3	4	5	3

Maskinoversettelse: hvordan fungerer det?



6: Maskinoversettelse (venstre: statistisk; høgre: regelbasert)

Hvorfor er maskinoversettelse vanskelig?



Translate

Turn off instant translation



English Dutch German Detect language ▾



Norwegian German Dutch ▾

Translate

pineapple chunks

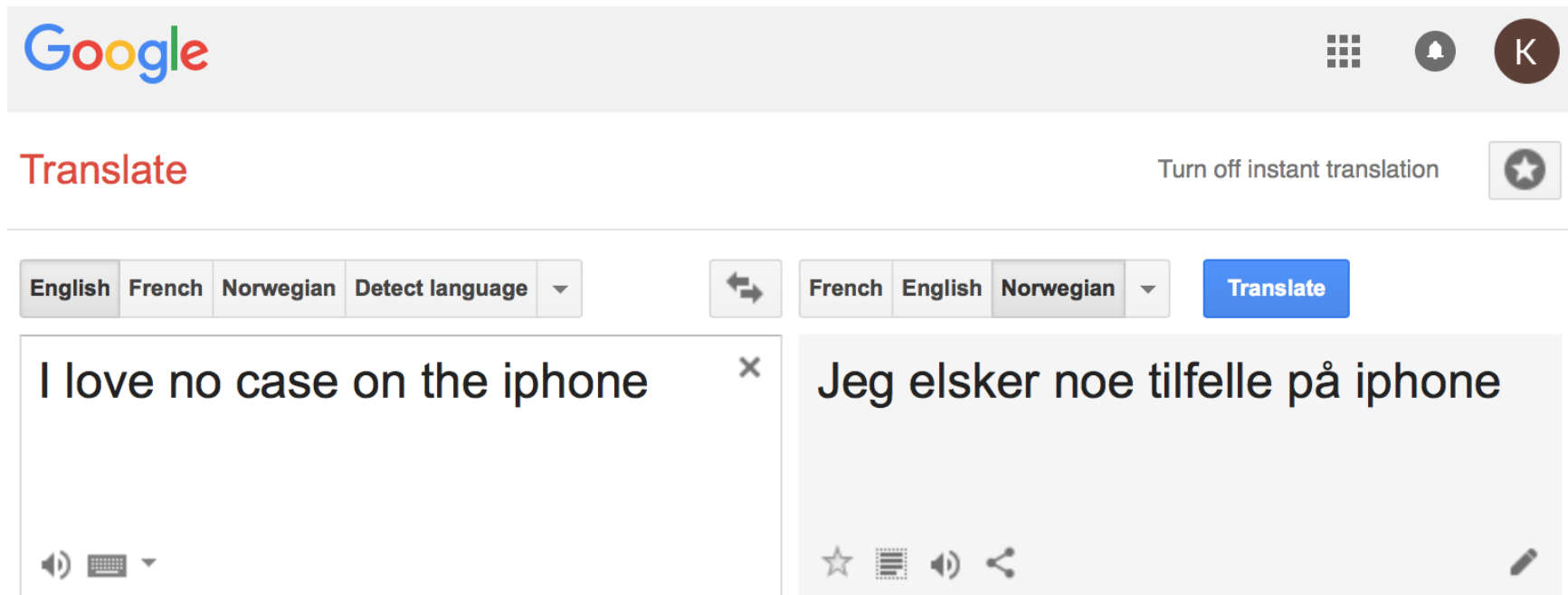
🔊 🗂️ ▾

ananas biter

☆ 🗂️ 🔊 🔄 ✎



Hvorfor er maskinoversettelse vanskelig?



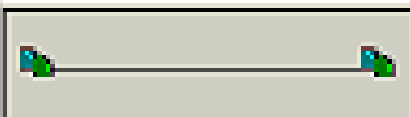







The screenshot shows the Google Translate web interface. At the top left is the Google logo. To the right are icons for a grid, a notification bell, and a user profile 'K'. Below the logo is the word 'Translate' in red. On the right side, there is a link 'Turn off instant translation' and a star icon. The main area has two language selection boxes. The left box shows 'English', 'French', 'Norwegian', and 'Detect language' with a dropdown arrow. The right box shows 'French', 'English', and 'Norwegian' with a dropdown arrow. A blue 'Translate' button is positioned between the two boxes. Below the left box is a text input field containing 'I love no case on the iphone' and a close 'x' icon. Below the right box is a text output field containing 'Jeg elsker noe tilfelle på iphone'. At the bottom of the left box are icons for a speaker and a keyboard. At the bottom of the right box are icons for a star, a list, a speaker, a share icon, and a pencil.

Hvorfor er maskinoversettelse vanskelig?



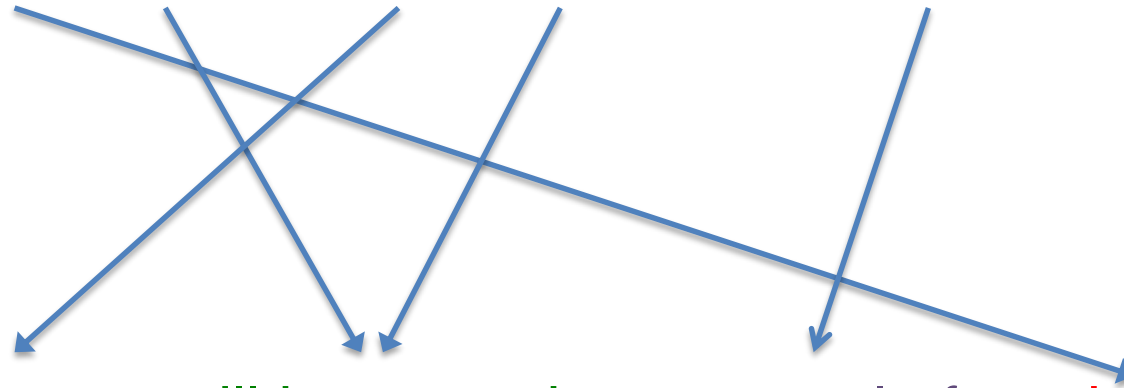
- Boken er **under** bordet
The book is **under** the table
- Publikumet sov **under** forelesningen
The audience slept **during** the lecture
- Bare et **under** kan redde Brann fra nedrykk
Only a **miracle** can save Brann from demotion
- **Skriv under** her
Sign here

Finn <fraser> heller enn ord

<p>This brings you to the calendar where you can begin to enter data.</p>		<p>Dedurch gelangen Sie zum Kalender, in dem alle weiteren Eingaben erfolgen.</p>
<p>After indicating your first cycle date in the dialog box, a calendar appears.</p>		<p>Nachdem Sie Ihren ersten Zyklus in das Dialogfenster eingegeben haben erscheint eine Kalenderansicht, die den Monat Ihres ersten Zyklus wiedergibt.</p>
<p>It indicates the month of your first cycle date.</p>		<p>Der Tag, den Sie als ersten Zyklustag eingegeben haben, wird in der rechten unteren Ecke des Kalenderfensters durch ^g1 und ^g2 dargestellt.</p>
<p>The day you indicated as the first cycle day will be marked with a ^g1 in the lower right hand corner and a ^g2.</p>		<p>Falls dies nicht der erste Tag Ihres ersten Zyklus ist, wählen Sie bitte "Datei", "Schließen", und "Nicht Speichern".</p>
<p>If this is not the day you want to indicate as the first cycle day, go to File, Close, and don't save it.</p>		<p>Führen Sie dann bitte erneut einen Doppelklick auf das CycleWatch Icon aus und korrigieren das Datum.</p>
<p>Then double-click on the CycleWatch icon, enter the correct date, and click on OK.</p>		<p>Um die Eingabe zu bestätigen klicken Sie bitte auf OK.</p>
<p>You will also see two numbers in each calendar day.</p>		<p>Es werden jeweils zwei Zahlen pro Kalendertag dargestellt.</p>
<p>The number in the upper right</p>		<p>In der rechten Ecke oben sehen</p>

Menu	Meny
Soupe à l'oignon	Løksuppe
Velouté de potiron	Gresskarsuppe
Goberge frit aux oignons	Seibiff med løk
Moules marinières	Blåsjell dampet i hvitvin
Foie de veau aux oignons	Kalvelever med løk
Tarte aux oignons confits et tomates	Pai med kandisert løk og tomater
Sélection de fromages	Utvalg av oster
Fondant au chocolat	Sjokoladefondant

I juni flytter Difis kurs inn på et nytt plattform



Difi's courses will be moved to a new platform in June


- Det finnes gode verktøy for utvikling av statistisk oversettelse (Moses, Jane)
- For hvert ordpar trengs minst 1 million ord tospråklige treningsdata
- I tillegg: énspråklige ordbøker, terminologi, navnelister, sammensetningsanalyse, osv.
- Det finnes oversettelser og andre ressurser for norsk–engelsk, men ikke for mange andre språkpar
- Automatisk preprosessering kan være nødvendig (f.eks. sammensetningsanalyse)
- Manuell postprosessering er nødvendig men lønnsom
- Evaluering (manuell eller automatisk) er nødvendig

Nasjonalbiblioteket Språkbankens ressurskatalog **BETA**

Enter search term Search Reset ☰ About the catalogue 🌐

Showing resources: 1 through 12 of total 42 Previous Next ▬ Språkbanken CLARINO

Text	Audio	Text	Tool
Tagged texts in Norwegian Bokmål from NBdigital (public domain material)	NB Tale - a basic acoustic phonetic speech database for Norwegian	N-grams from NBdigital	NB N-gram
[PUB] 07.03.2016	[PUB] 25.02.2016	[PUB] 24.02.2016	[PUB] 24.02.2016
Lexicon	Text	Text	Text
Norwegian Wordnet - Bokmål	Norwegian Newspaper Corpus	N-grams for Norwegian Bokmål (based on Norwegian news text)	Norwegian Acquis Communautaire
[PUB] 22.02.2016	[OTHER] 04.02.2016	[PUB] 04.02.2016	[OTHER] 03.02.2016


The CLARINO Bergen Centre offers:
 A repository to search and deposit language data
 Online services for treebanks and other corpora
 Online editing of CMDI metadata



Welcome to CLARINO Bergen Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines.

[Advanced Search](#)

Author	Subject	Language (ISO)
Gerstenberger, Cipri ... (21)	Bilingual Lexicon (9)	Northern Sami (7)
Parra Escartín, Carla (3)	South Saami (8)	Norwegian Bokmål (7)
Dione, Cheikh M. Bamba (1)	Machine-readable Dic ... (7)	Southern Sami (7)
Giellatekno and Divv ... (1)	Norwegian (7)	Kven Finnish (5)

- Mer språkdata er nødvendig for å utvikle norsk språkteknologi
- Maskinoversettelse for norsk trenger norske tekster som er parallellstilt med oversettelser – gjerne flere språk
- Språkdata bør være mest mulig tilgjengelig under åpne lisenser
- Samarbeid mellom forskere, industri, offentlig sektor og brukere er ønskelig
- META-FORUM 2016, Lissabon, 4.–5. juli 2016