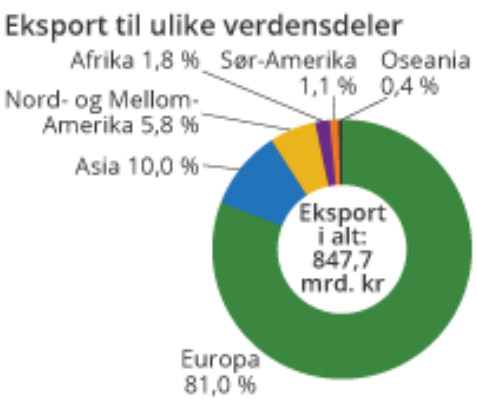
European Language Resource Coordination
Connecting Europe Facility

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

Koenraad De Smedt (Universitetet i Bergen)

Internasjonal handel

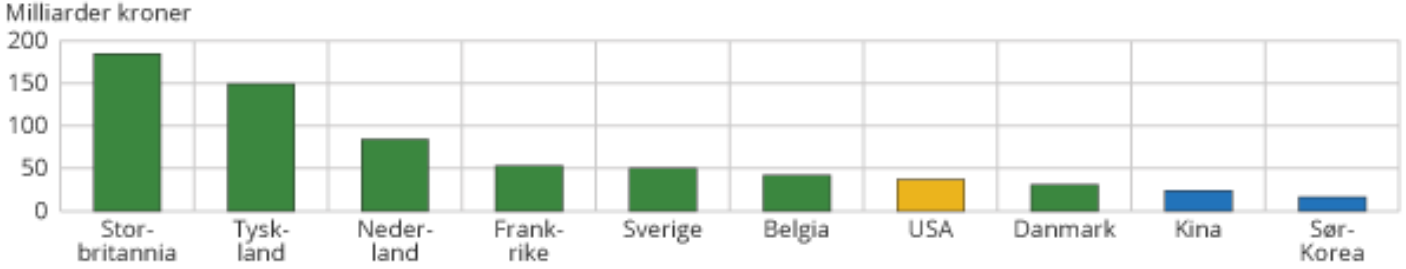
Figur 7. Eksport av varer. 2015



Eksport til de fem største handelspartnerne i Europa

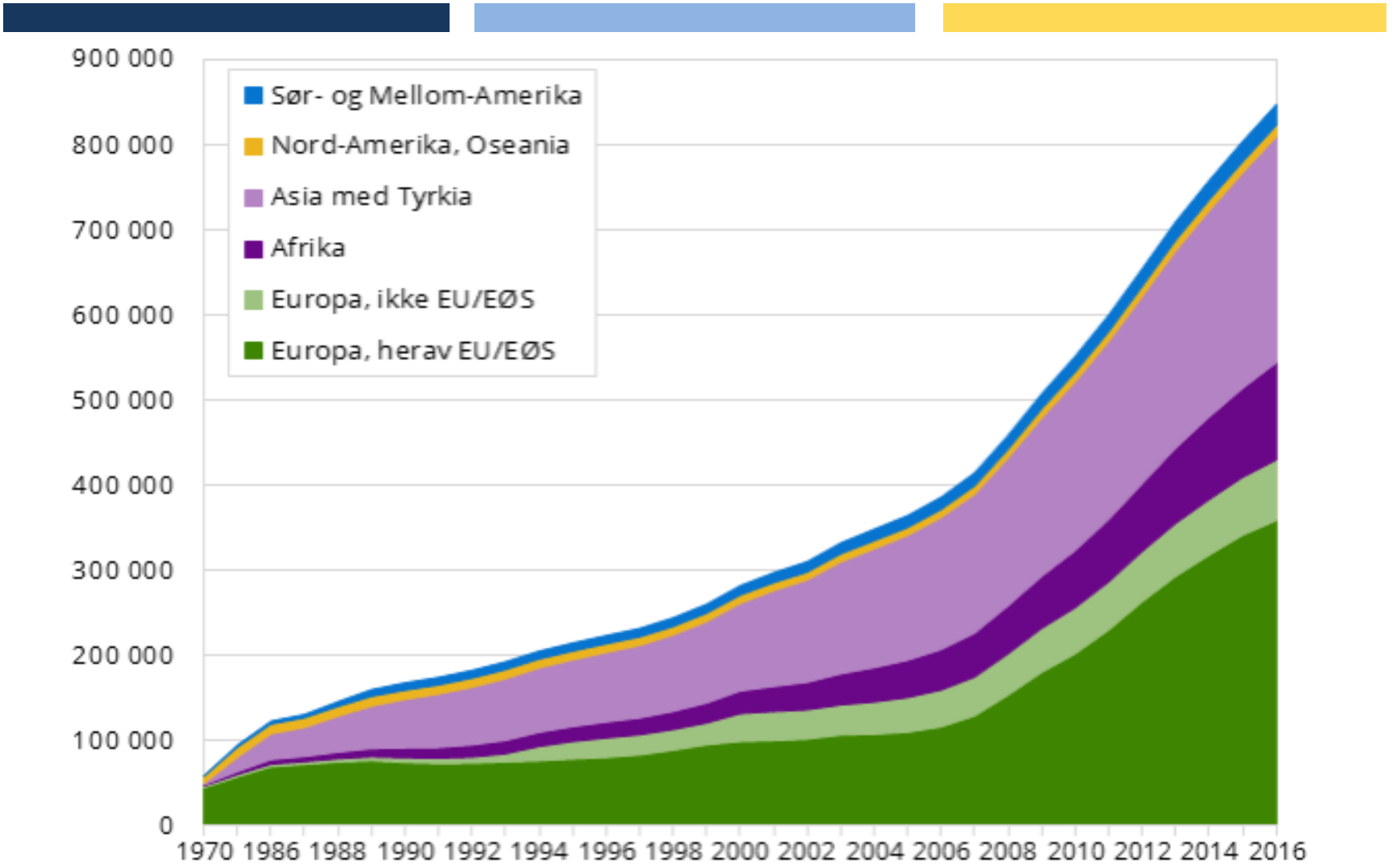
	Mrd. kr
Storbritannia	184,8
Tyskland	149,0
Nederland	84,0
Frankrike	52,8
Sverige	50,4

Eksport til Norges ti største handelspartnere på verdensbasis



Kilde: Statistisk sentralbyrå.

Innvandrere og innvandrerbakgrunn



Norsk i den digitale tidsalderen

- 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

European Language Resource Coordination
Connecting Europe Facility

Behov for kommunikasjon og tjenester

Norsk i den digitale tidsalderen

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

Hva er språkteknologi?

Informasjonsteknologi som inkluderer modeller for språkbehandling, bl.a.

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

Selvlærende systemer

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.

Regelbaserte systemer

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

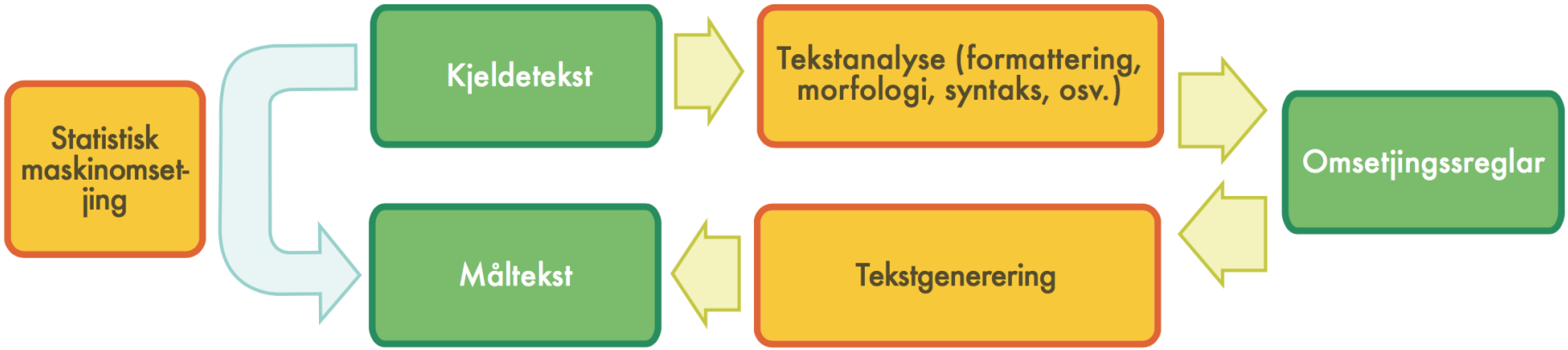
Utvikling av språkteknologi for norsk

- 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

Hvitbokserie om 31 språk (2012)
www.meta-net.eu

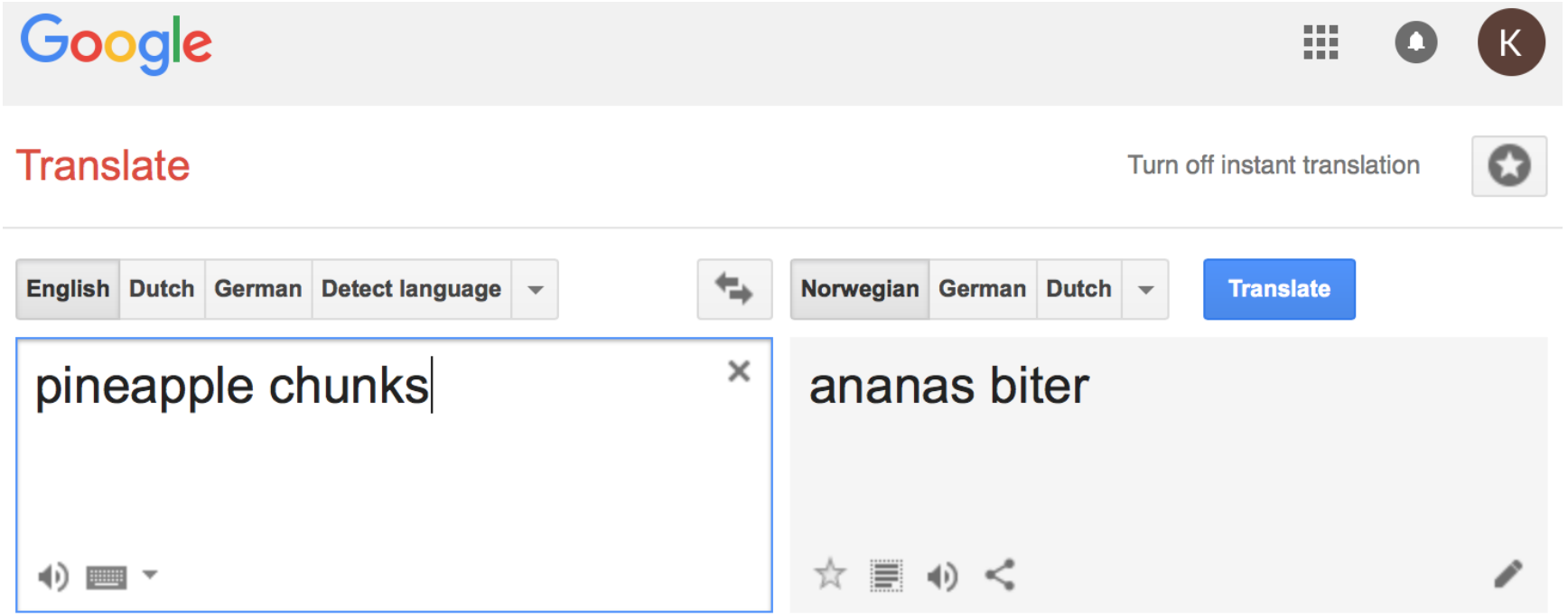
	Kvantitet	Tilgjengelegheit	Kvalitet	Dekningsgrad	Modenheit	Berekraft	Tilpassingsdyktigheit
Språkteknologi (verktøy, teknologiar og applikasjonar)							
Taleattkjenning	4	2	2	1	2	3	3
Talesyntese	3	2	3	2	3	3	3
Grammatisk analyse	4	4,5	4	4	4,5	4,5	5
Semantisk analyse	2	2	3,3	3	3,7	3,3	3,7
Tekstgenerering	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: Maskinomsetjing (venstre: statistisk; høgre: regelbasert)

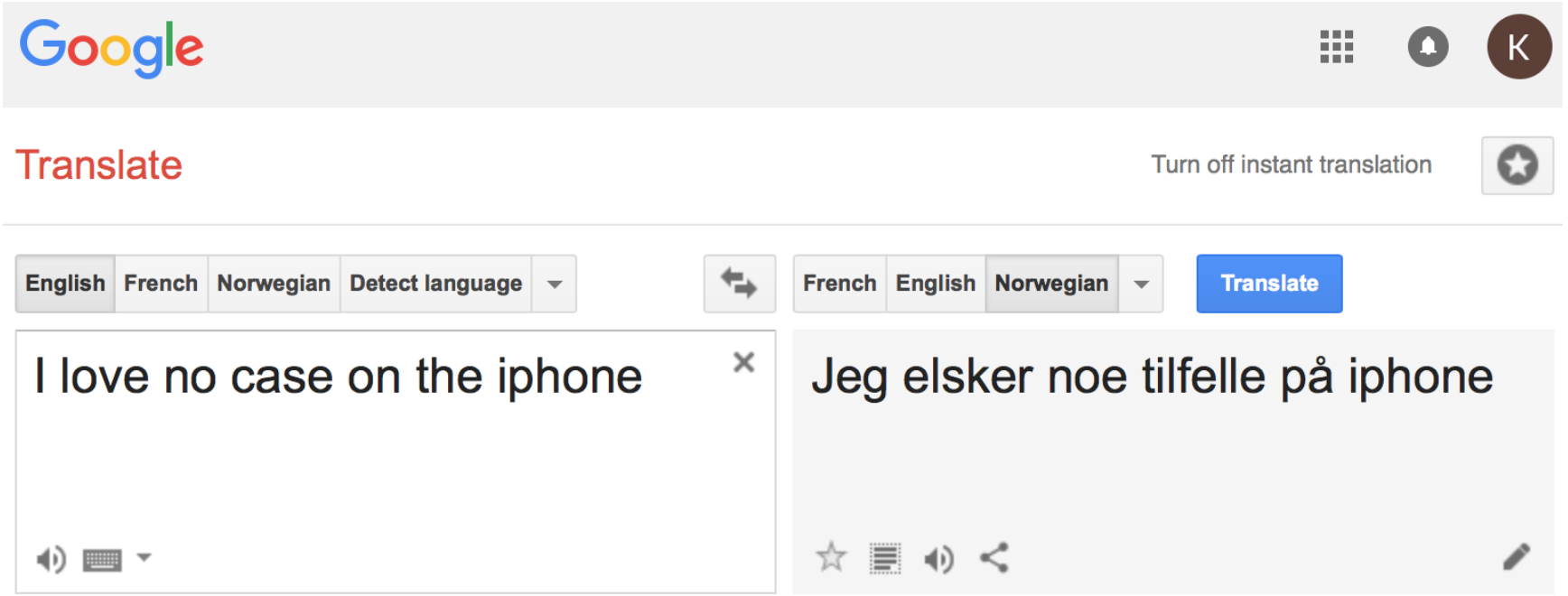
Hvorfor er maskinoversettelse vanskelig?



European Language Resource Coordination
Connecting Europe Facility

Farlig!

Hvorfor er maskinoversettelse vanskelig?

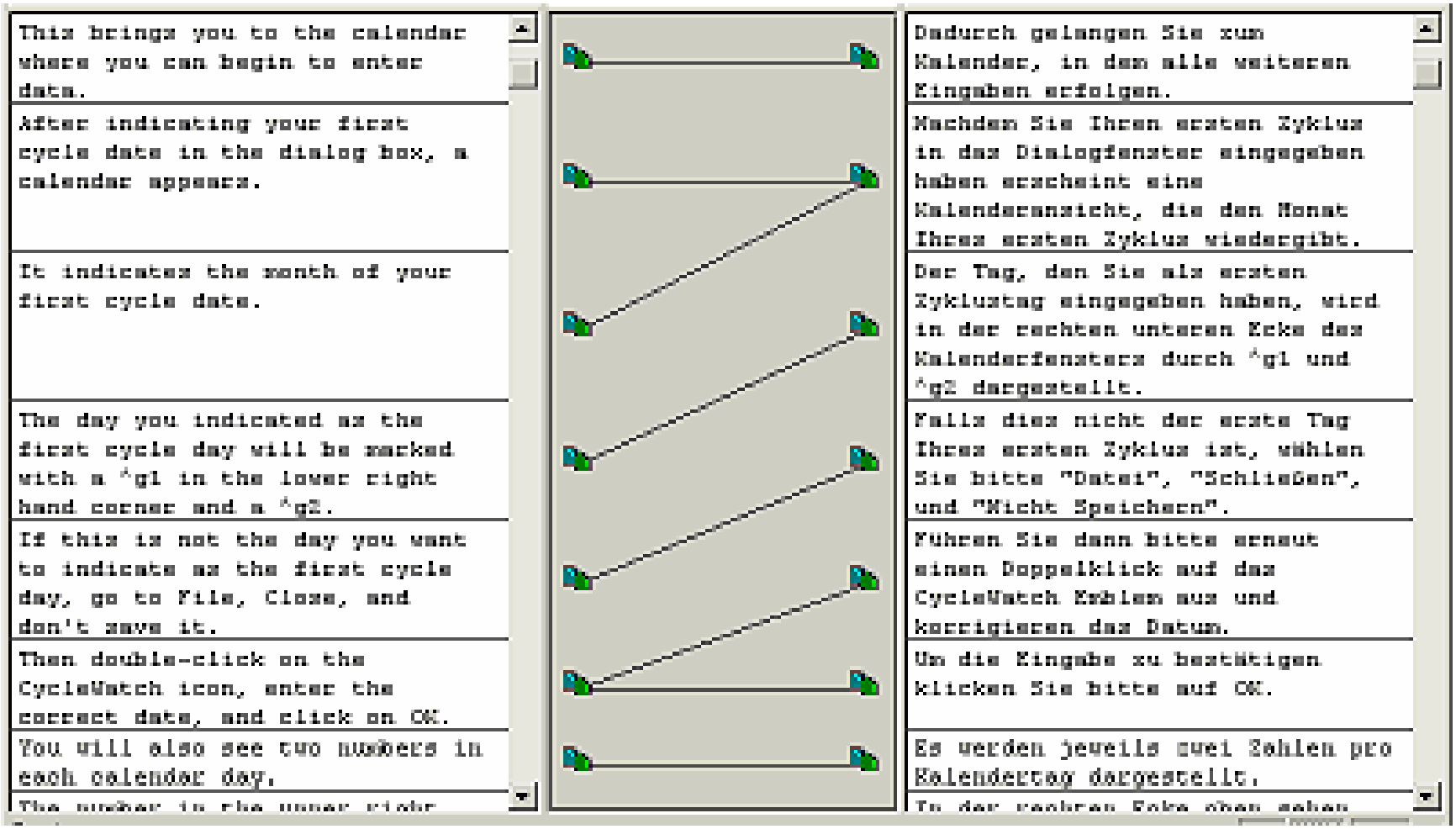


Hvorfor er maskinoversettelse vanskelig?

- 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

Grensesnitt for parallellstilling (Trados)

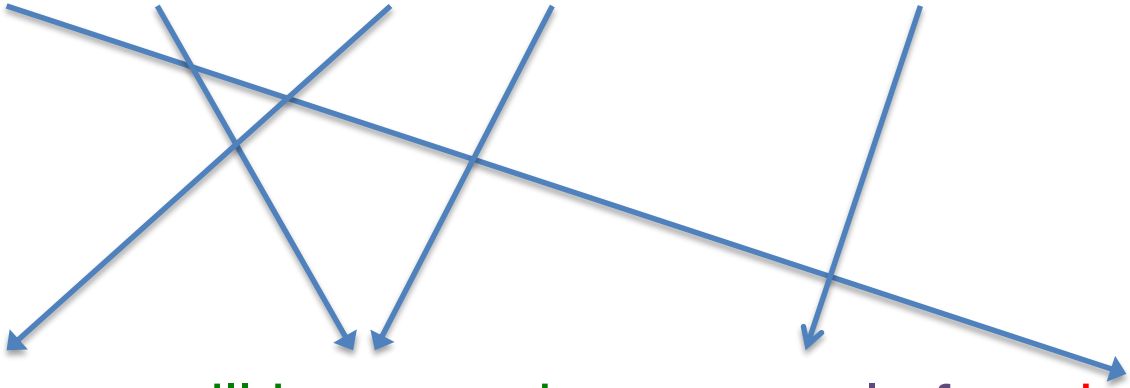


Fraser i parallellstilt tekst

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

Koblinger mellom fraser og oversettelser

I juni flytter Difis kurs inn på et nytt plattform



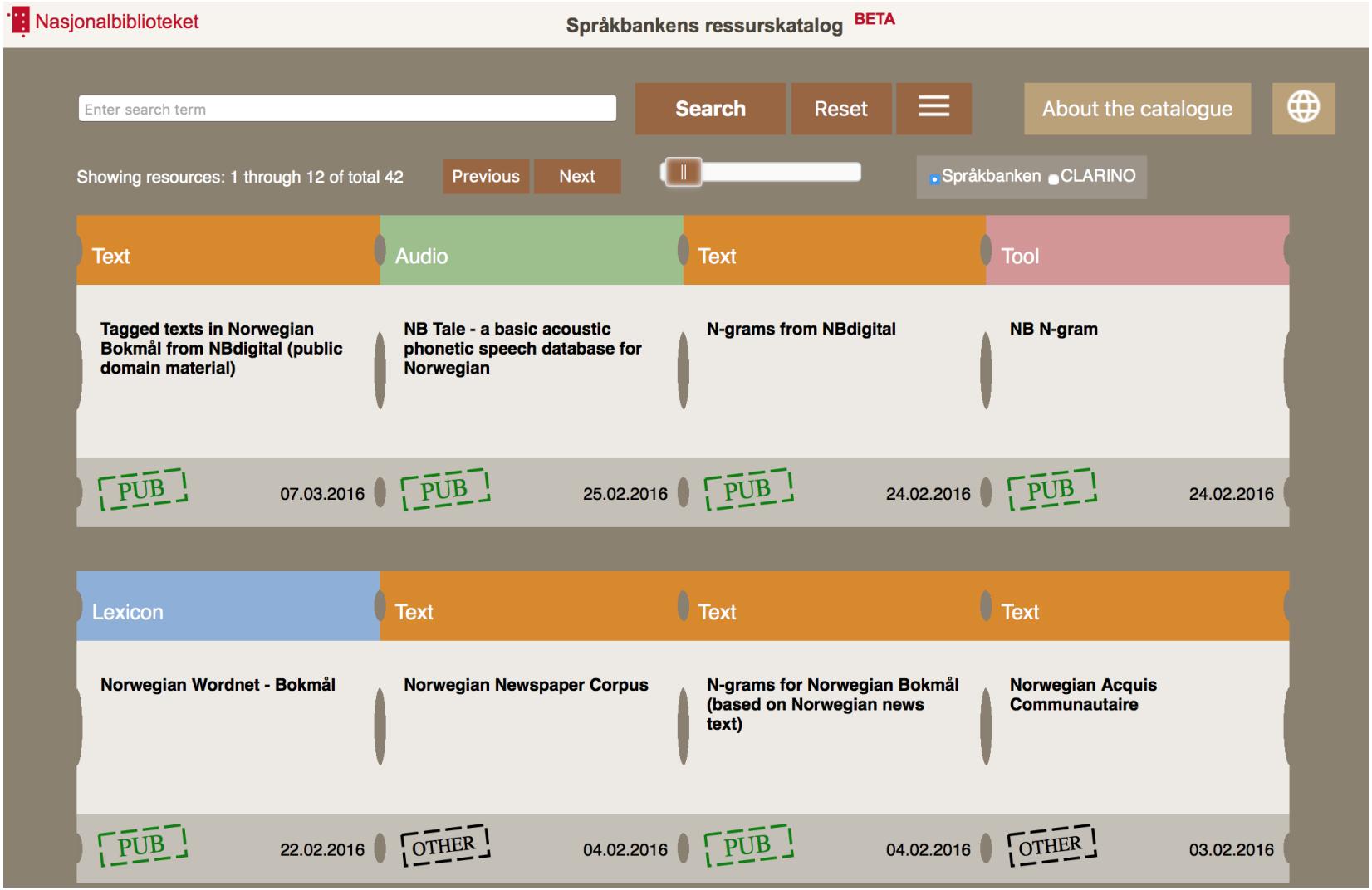
Difi's courses will be moved to a new platform in June

Verktøy og data

- 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 preprossessering kan være nødvendig (f.eks. sammensetningsanalyse)
- Manuell postprosessering er nødvendig men lønnsom
- Evaluering (manuell eller automatisk) er nødvendig

European Language Resource Coordination
Connecting Europe Facility

Språkressurser i Norge: Språkbanken



European Language Resource Coordination
Connecting Europe Facility

Språkressurser i Norge: CLARINO

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

CLARINO

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.

Search

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)

Konklusjoner

- 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