META NET

White Paper Series Hvitbokserie

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

NORSK I DEN DIGITALE TIDSALDEREN

BOKMÅLSVERSJON

Koenraad De Smedt
Gunn Inger Lyse
Anje Müller Gjesdal
Gyri S. Losnegaard

Springer

White Paper Series Hvitbokserie

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

NORSK I DEN DIGITALE TIDSALDEREN

BOKMÅLSVERSJON

Koenraad De Smedt UIB
Gunn Inger Lyse UIB
Anje Müller Gjesdal UIB
Gyri S. Losnegaard UIB

Georg Rehm, Hans Uszkoreit
(Redaktører, editors)

Editors
Georg Rehm
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: georg.rehm@dfki.de

Hans Uszkoreit
DFKI
Alt-Moabit 91c
Berlin 10559
Germany
e-mail: hans.uszkoreit@dfki.de

ISSN 2194-1416 ISSN 2194-1424 (electronic)
ISBN 978-3-642-31388-2 ISBN 978-3-642-31389-9 (eBook)
DOI 10.1007/978-3-642-31389-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940568

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

FORORD PREFACE

Dette dokumentet er del av en serie som skal fremme kunnskap om språkteknologiens status og potensiale. Målgruppen er journalister, politikere, språkbrukere, lærere og andre interesserte. Tilgjengeligheten og bruken av språkteknologi i Europa varierer fra språk til språk. Derfor vil også nødvendige tiltak for å støtte forskning og utvikling av språkteknologi være forskjellige for hvert språk. Hvilke tiltak som er nødvendige avhenger av flere faktorer, for eksempel kompleksiteten i et gitt språk og antall språkbrukere.

Forskningsnettverket META-NET, et *Network of Excellence* finansiert av Europakommisjonen, presenterer i denne serien (jf. s. 81) sin analyse av eksisterende språkressurser og teknologier for de 23 offisielle EU-språkene og andre nasjonale og regionale språk i Europa – deriblant norsk. Resultatene av denne analysen tyder på at det er betydelige hull i forskning og utvikling for alle språkene. Denne detaljerte ekspertanalysen av den nåværende situasjonen i denne serien vil forhåpentlig bidra til å maksimere effekten av ny forskning.

Per november 2011 består META-NET av 54 forskningsinstitusjoner i 33 land (jf. s. 77) som samarbeider med kommersielle aktører (IT-bedrifter, utviklere og brukere), offentlige etater, ikke-statlige organisasjoner, representanter for språksamfunn og universiteter. I samarbeid med disse samfunnsrepresentantene er målet å skape en felles teknologivisjon og å utvikle en strategisk forskningsagenda for flerspråklighet i Europa innen år 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differs. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 81). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of November 2011, META-NET consists of 54 research centres from 33 European countries (p. 77). META-NET is working with stakeholders from economy (Software companies, technology providers and users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

Forfatterne av denne teksten takker forfatterne av hvitboken for tysk for tillatelsen til å gjenbruke visse språkuavhengige materialer fra deres tekst [1]. Forfatterne takker også Gisle Andersen, Torbjørg Breivik, Helge Dyvik, Kristin Hagen, Torbjørn Nordgård, Torbjørn Svendsen og Trond Trosterud for verdifulle bidrag og kommentarer.

Arbeidet med denne utredningen er finansiert av det sjuende rammeprogrammet og Den europeiske kommisjonens ICT Policy Support program, gjennom kontraktene T4ME (tildelingsavtale 249 119), CESAR (tildelingsavtale 271 022), METANET4U (tildelingsavtale 270 893) og META-NORD (tildelingsavtale 270 899).

The authors of this document are grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1]. They also wish to thank Gisle Andersen, Torbjørg Breivik, Helge Dyvik, Kristin Hagen, Torbjørn Nordgård, Torbjørn Svendsen and Trond Trosterud for valuable contributions and comments.

The development of this White Paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

INNHOLD CONTENTS

NORSK I DEN DIGITALE TIDSALDEREN

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

1

SAMMENDRAG

Informasjonsteknologi påvirker hverdagen vår. Vi bruker datamaskiner når vi skriver, redigerer, regner ut, søker etter informasjon, og i økende grad også når vi leser, hører på musikk, ser på bilder og ser film. Vi har med oss små datamaskiner i lomma og bruker disse til å ringe, skrive e-post, innhente informasjon og til å underholde oss selv hvor vi enn er. Men hvilken innvirkning har denne utstrakte digitaliseringen av informasjon, kunnskap og daglig kommunikasjon på språket vårt? Vil språket vårt endre seg eller til og med forsvinne? Hva er sjansene for at norsk språk vil bestå?

Mange av de 6000 språkene som finnes i verden i dag vil ikke overleve i det globaliserte digitale informasjonssamfunnet. En regner med at minst 2000 språk kommer til å forsvinne de kommende tiårene. Andre vil fremdeles spille en rolle i privatsfæren og lokalsamfunnet, men ikke i det bredere offentlige liv som næringsliv og akademia. Statusen til et språk avhenger ikke bare av tallet på brukere eller hvor mange bøker, filmer og TV-stasjoner som benytter språket, men også av i hvilken grad språket gjør seg gjeldende i den digitale virkeligheten og brukes i programvareapplikasjoner.

I denne sammenhengen sliter norsk fremdeles med voksesmerter. I begynnelsen av det tjueførste århundret eksisterte norsk språkteknologi bare i svært liten skala. Det fantes et relativt godt system for oversettelse fra bokmål og nynorsk, der var stavekontroll, og det fantes også et lite dialogsystem som svarer på spørsmål, mens folk flest lo av den dårlige kvaliteten til de første talegjenkjenningsprogrammene. Et ambisiøst industrielt initiativ til språkteknologiutvikling på Voss mislyktes. Innen høyere utdanning fantes det program for språkteknologi og datalingvistikk, og det eksisterte forskning på disse feltene, men det manglet språkressurser og språkverktøy.

Bildet endret seg da forskningsrådet tok initiativ til et språkteknologiprogram i 2002, med sikte på å utvikle ny kunnskap og nødvendige verktøy. Programmet resulterte i flere prosjekt som skapte ny kompetanse og et bedre grunnlag for norsk språkteknologi. De største prosjektene i dette språkteknologiprogrammet leverte et tekst-til-tale-system og en demonstrator for oversettelse av høy kvalitet fra norsk til engelsk.

Etter Stortingsmeldingen fra 2008 [2], og vedtaket av denne meldingen i Stortinget, ble en fritt tilgjengelig samling av norske språkteknologiske ressurser, *Språkbanken*, etablert i 2010. Språkbanken er nå i gang med å bygge opp og distribuere norske språkdata, en oppgave som lenge har vært etterspurt innen forskning og utvikling. Dersom dette arbeidet blir opprettholdt, vil det utgjøre en uvurderlig investering i norsk språks fremtid.

På tross av en betydelig utvikling innen norsk språkteknologi det siste tiåret viser denne rapporten at det ennå bare er for basisverktøy og -ressurser at situasjonen er noenlunde tilfredsstillende. Når det gjelder mer avanserte applikasjoner, finnes det fremdeles svært få verktøy og ressurser for norsk, og vi har fremdeles langt igjen før norsk språk er sikret en fremtid som fullverdig aktør i det moderne – og framtidige – europeiske språksamfunnet.

Informasjons- og kommunikasjonsteknologien forbereder seg nå til neste teknologirevolusjon. I kjølvannet av

personlige datamaskiner, nettverk, stadig mindre og lettere komponenter, multimedia, mobile enheter og databehandling i digitale skyer, vil den neste generasjonen teknologi bestå av programvare som ikke bare forstår talte og skrevne bokstaver og lyder, men også hele ord og setninger, og som støtter brukeren bedre enn dagens teknologi, fordi den snakker, kjenner og forstår språket deres. Forløpere i denne utviklingen er IBMs superdatamaskin Watson, som slo USA-mesteren i kunnskapsspillet "Jeopardy", og Apples mobilassistent Siri for iPhone, som responderer på språkkommandoer og kan svare på spørsmål på engelsk, tysk, fransk og japansk. Et norsk talegjenkjenningssystem for iPhone er tilgjengelig, men det er fremdeles mye mindre pålitelig enn det tilsvarende engelske systemet.

Språkbrukere kommuniserer allerede ved hjelp av teknologien som er utviklet for deres språk. Etter hvert vil teknologiske innretninger, som respons på enkle talekommandoer, være i stand til å hente de viktigste nyhetene og informasjonen fra den globale digitale kunnskapsbasen. Språkbasert teknologi vil kunne oversette automatisk eller fungere som støtte for tolker, lage sammendrag av samtaler og dokumenter og være et hjelpemiddel i læringssituasjoner. Språkteknologi vil for eksempel kunne hjelpe innvandrere med å lære norsk, og dermed også med integrering i det norske samfunnet.

Informasjons- og kommunikasjonsteknologi vil gjøre industrielle roboter og tjenesteroboter (som i dag er under utvikling i forskningslaboratorier) i stand til å forstå hva brukeren ønsker at de skal gjøre og å rapportere om oppgavene de har utført. Et slikt prestasjonsnivå strekker seg langt ut over enkle bokstavlister og leksikon, stavekontroller og uttaleregler. Skal språkteknologi kunne tolke spørsmål og levere utfyllende og relevante svar, må den bevege seg fra basale tilnærminger til et mer altomfattende perspektiv, hvor språkmodelleringen tar hensyn til syntaks så vel som semantikk.

Ikke alle europeiske språk er like godt forberedt til en slik fremtid. Denne rapporten presenterer en evaluering av graden av språkteknologistøtte for 30 europeiske språk, basert på fire kjerneområder: maskinoversettelse, taleprosessering, tekstanalyse og, til sist, basisressurser som er nødvendige for å kunne bygge språkteknologiske applikasjoner. Språkene ble delt inn i fem klynger etter nivå, og ikke overraskende havnet norsk i bunnklyngen, og i enkelte tilfeller i klyngen over, for alle typer verktøy og ressurser. Norsk ligger langt etter større språk som for eksempel tysk og fransk. Men heller ikke disse språkene klarer å nå opp til kvaliteten og dekningsgraden til sammenlignbare ressurser og verktøy for engelsk, som er det klart ledende språket på nesten alle felter innen språkteknologi.

I St.meld. nr. 48 [3] konstaterer en at språkteknologifeltet kan bli "en av de fremste arenaene der kampen om norsk språk og kultur vil utspille seg i tiden fremover" (kap. 12.9, s. 196). Hva må vi så gjøre for å sikre norsk språk en fremtid i informasjonssamfunnet? I 2002 anslo en ekspertgruppe på oppdrag fra myndighetene at det vil kreve en investering på 20 millioner kroner *hvert år* de første fem årene [4]. Selv om Språkbanken nå er etablert og virksom, er det et faktum at de årlige investeringene så langt har utgjort bare en brøkdel av estimert behov. Det skulle derfor ikke komme som noe overraskelse at norsk språkteknologi fremdeles henger igjen i tidlig barndom. Kommersielt er fem millioner språkbrukere for få til alene å forsvare en kostbar utvikling av nye produkter. Norsk IT-industri, og spesielt store og mellomstore bedrifter, kan ikke alene ta kostnadene ved å bygge opp store språkressurser og verktøy for norsk. Fortsatt offentlig støtte er derfor nødvendig for å sikre at eksisterende verktøy og den opparbeidede kunnskapen og erfaringen hos forskere og bedrifter skal bli utnyttet til fulle.

Norsk språk er ikke umiddelbart truet av den engelske dominansen innen språkteknologi. Det kan likevel endre seg drastisk når den nye generasjonen tekno-