META NET

White Paper Series Kvitbokserie

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

NORSK I DEN DIGITALE TIDSALDEREN

NYNORSKVERSJON

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

Springer

White Paper Series Kvitbokserie

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

NORSK I DEN DIGITALE TIDSALDEREN

NYNORSKVERSJON

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

Georg Rehm, Hans Uszkoreit
(Redaktørar, 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 978-3-642-31432-2 ISBN 978-3-642-31433-9 (eBook)
DOI 10.1007/978-3-642-31433-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012941133

FORORD PREFACE

Dette dokumentet er del av ein serie som skal fremje kunnskap om språkteknologiens status og potensiale. Målgruppa er journalistar, politikarar, språkbrukarar, lærarar og andre interesserte. Tilgangen til, og nytta av, språkteknologi i Europa varierer frå språk til språk. Difor vil òg naudsynte tiltak for å støtte forsking og utvikling av språkteknologi vere ulike for kvart språk. Kva for tiltak som er naudsynte, avheng av fleire faktorar, til dømes kompleksiteten i eit gjeve språk og mengda språkbrukarar.

Forskingsnettverket META-NET, eit *Network of Excellence* finansiert av Europakommisjonen, presenterer i denne serien (jf. s. 81) analysen sin av eksisterande språkressursar og teknologiar for dei 23 offisielle EU-språka og andre nasjonale og regionale språk i Europa – mellom dei norsk. Resultata av denne analysen tyder på at det er betydelege hol i forsking og utvikling for alle språka. Denne detaljerte ekspertanalysen av den noverande situasjonen i denne serien vil vonleg bidra til å maksimere effekten av ny forsking.

Per november 2011 består META-NET av 54 forskingsinstitusjonar i 33 land (jf. s. 77) som samarbeider med kommersielle aktørar (IT-føretak, utviklarar og brukarar), offentlege etatar, ikkje-statlege organisasjonar, representantar for språksamfunn og universitet. I samarbeid med desse samfunnsrepresentantane er målet å skape ein felles teknologivisjon og å utvikle ein strategisk forskingsagenda for eit fleirspråkleg Europa innan 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.

Forfattarane av denne rapporten takkar forfattarane av rapporten for tysk språk for løyve til å gjenbruke utvalt språkuavhengig material frå dokumentet deira [1]. Forfattarane takkar òg Gisle Andersen, Torbjørg Breivik, Helge Dyvik, Kristin Hagen, Torbjørn Nordgård, Torbjørn Svendsen og Trond Trosterud for verdifulle bidrag og kommentarar.

Arbeidet med denne utgreiinga er finansiert av det sjuande rammeprogrammet og Den europeiske kommisjonens ICT Policy Support program, gjennom kontraktane 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).

INNHALD CONTENTS

NORSK I DEN DIGITALE TIDSALDEREN

THE NORWEGIAN LANGUAGE IN THE DIGITAL AGE

1

SAMANDRAG

Informasjonsteknologi påverkar kvardagen vår. Vi brukar datamaskiner når vi skriv, redigerer, reknar ut, søkjer etter informasjon, og i aukande grad også når vi les, høyrer på musikk, kikkar på bilete og ser på film. Vi har med oss små datamaskiner i lomma og brukar desse til å ringe, skrive e-post, innhente informasjon og til å underhalde oss sjølve kvar vi enn er. Men på kva måte verkar denne utstrakte digitaliseringa av informasjon, kunnskap og dagleg kommunikasjon inn på språket vårt? Vil språket vårt endre seg eller til og med forsvinne? Kva er sjansane for at norsk språk vil bestå?

Mange av dei 6000 språka som finst i verda i dag vil ikkje overleve i det globaliserte digitale informasjonssamfunnet. Ein reknar med at minst 2000 språk kjem til å forsvinne dei kommande tiåra. Andre vil framleis spele ei rolle i privatsfæren og lokalsamfunnet, men ikkje i det breiare offentlege liv som næringsliv og akademia. Statusen til eit språk avheng ikkje berre av talet på brukarar eller kor mange bøker, filmar og TV-stasjonar som nyttar språket, men også av i kor stor grad språket gjer seg gjeldande i den digitale verkelegheita og blir brukt i programvareapplikasjonar.

I denne samanhengen slit norsk framleis med veksesmerter. I byrjinga av det tjueførste hundreåret eksisterte norsk språkteknologi berre i svært liten skala. Det fanst eit relativt godt system for omsetjing frå bokmål og nynorsk, der var stavekontroll, og det fanst også eit lite dialogsystem som svarer på spørsmål, medan folk flest lo av den dårlege kvaliteten til dei første talegjenkjenningsprogramma. Eit ambisiøst industrielt initiativ til utvikling av språkteknologi på Voss mislykkast. Innan høgare utdanning fanst det program for språkteknologi og datalingvistikk, og det eksisterte forsking på desse felta, men det mangla språkressursar og språkverktøy.

Biletet endra seg då Forskingsrådet tok initiativ til eit språkteknologiprogram i 2002, med sikte på å utvikle ny kunnskap og nødvendige verktøy. Programmet resulterte i fleire prosjekt som skapte ny kompetanse og eit betre grunnlag for norsk språkteknologi. Dei største prosjekta i dette språkteknologiprogrammet leverte eit tekst-til-tale-system og ein demonstrator for omsetjing av høg kvalitet frå norsk til engelsk.

Etter Stortingsmeldinga frå 2008 [2], og vedtaket av denne meldinga i Stortinget, vart ei fritt tilgjengeleg samling av norske språkteknologiske ressursar, *Språkbanken*, etablert i 2010. Språkbanken er no i gong med å byggje opp og distribuere norske språkdata, ei oppgåve som lenge har vore etterspurd innan forsking og utvikling. Dersom dette arbeidet blir halde ved like, vil det utgjere ei uvurderleg investering i framtida til det norske språket.

Trass ei betydeleg utvikling innan norsk språkteknologi det siste tiåret viser denne rapporten at det enno berre er for basisverktøy og -ressursar at situasjonen er nokolunde tilfredsstillande. Når det gjeld meir avanserte applikasjonar, finst det framleis svært få verktøy og ressursar for norsk, og vi har framleis langt igjen før norsk språk er sikra ei framtid som fullverdig aktør i det moderne – og framtidige – europeiske språksamfunnet.

Informasjons- og kommunikasjonsteknologien førebur seg no til neste teknologirevolusjon. I kjølvatnet av per-

sonlege datamaskiner, nettverk, stadig mindre og lettare komponentar, multimedia, mobile einingar og databehandling i digitale skyer, vil den neste generasjonen teknologi bestå av programvare som ikkje berre forstår talte og skrivne bokstavar og lydar, men også heile ord og setningar, og som støttar brukaren betre enn dagens teknologi, fordi han snakkar, kjenner og forstår språket deira. Forløparar i denne utviklinga er IBM si superdatamaskin Watson, som sigra over USA-meisteren i kunnskapsspelet "Jeopardy", og Apple sin mobilassistent Siri for iPhone, som responderer på språkkommandoar og kan svare på spørsmål på engelsk, tysk, fransk og japansk. Eit norsk taleattkjenningssystem for iPhone er tilgjengeleg, men det er framleis mykje mindre påliteleg enn det tilsvarande engelske systemet.

Språkbrukarar kommuniserer allereie ved hjelp av teknologien som er utvikla for deira språk. Etter kvart vil teknologiske innretningar, som respons på enkle talekommandoar, vere i stand til å hente dei viktigaste nyhenda og informasjonen frå den globale digitale kunnskapsbasen. Språkbasert teknologi vil kunne omsetje automatisk eller fungere som støtte for tolkar, lage samandrag av samtaler og dokument og vere eit hjelpemiddel i læringssituasjonar. Språkteknologi vil til dømes kunne hjelpe innvandrarar med å lære norsk, og dermed også med integrering i det norske samfunnet.

Informasjons- og kommunikasjonsteknologi vil gjere industrielle robotar og tenesterobotar (som i dag er under utvikling i forskingslaboratoria) i stand til å forstå kva brukaren ønskjer at dei skal gjere og til å rapportere om oppgåvene dei har utført. Eit slikt prestasjonsnivå strekkjer seg langt ut over enkle bokstavlister og leksika, stavekontrollar og uttalereglar. Skal språkteknologi kunne tolke spørsmål og levere utfyllande og relevante svar, må han bevege seg frå basale tilnærmingar til eit meir altomfattande perspektiv, der språkmodelleringa tek omsyn til syntaks så vel som semantikk.

Ikkje alle europeiske språk er like godt førebudde til ei slik framtid. Denne rapporten presenterer ei evaluering av graden av språkteknologistøtte for 30 europeiske språk, basert på fire kjerneområde: maskinomsetjing, taleprosessering, tekstanalyse og, til sist, basisressursar som er naudsynte for å kunne byggje språkteknologiske applikasjonar. Språka vart delte inn i fem klynger etter nivå, og ikkje overraskande hamna norsk i botnklynga, og i enkelte tilfelle i klynga over, for alle typar verktøy og ressursar. Norsk ligg langt etter større språk som til dømes tysk og fransk. Men ikkje ein gong desse språka klarer å nå opp til kvaliteten og dekningsgraden til samanliknbare ressursar og verktøy for engelsk, som er det klart leiande språket på nesten alle felt innan språkteknologi.

I St.meld. nr. 48 [3] konstaterer ein at språkteknologifeltet kan verte "ein av dei fremste arenaene der kampen om norsk språk og kultur vil utspela seg i tida framover" (kap. 12.9, s. 196). Kva må vi så gjere for å sikre norsk språk ei framtid i informasjonssamfunnet? I 2002 anslo ei ekspertgruppe skipa av myndigheitene at det vil krevje ei investering på 20 millionar kroner *kvart år* dei første fem åra [4]. Sjølv om Språkbanken no er etablert og verksam, er det eit faktum at dei årlege investeringane så langt har utgjort berre ein brøkdel av estimert behov. Det skulle difor ikkje komme som noka overrasking at norsk språkteknologi framleis heng att i tidleg barndom. Kommersielt er fem millionar språkbrukarar for få til aleine å forsvare ei kostbar utvikling av nye produkt. Norsk IT-industri, og spesielt store og mellomstore bedrifter, kan ikkje sjølve ta kostnadene ved å byggje opp store språkressursar og verktøy for norsk. Framleis offentleg støtte er nødvendig for å sikre at eksisterande verktøy og opparbeidd kunnskap og erfaring hos forskarar og bedrifter skal bli utnytta til fulle.

Norsk språk er ikkje umiddelbart trua av den engelske dominansen innan språkteknologi. Dette kan likevel endre seg drastisk når den nye generasjonen teknolo-