

THE DANISH LANGUAGE IN
THE DIGITAL AGE

DET DANSKE SPROG I DEN
DIGITALE TIDSALDER

Bolette Sandford Pedersen
Jürgen Wedekind
Steen Bøhm-Andersen
Peter Juel Henriksen
Sanne Hoffensetz-Andresen
Sabine Kirchmeier-Andersen
Jens Otto Kjærum
Louise Bie Larsen
Bente Maegaard
Sanni Nimb
Jens-Erik Rasmussen
Peter Revsbech
Hanne Erdman Thomsen

White Paper Series

Hvidbogsserie

THE DANISH
LANGUAGE IN
THE DIGITAL
AGE

DET DANSKE
SPROG I DEN
DIGITALE
TIDSALDER

Bolette S. Pedersen Københavns Universitet

Jürgen Wedekind Københavns Universitet

Steen Bøhm-Andersen Ankiro

Peter J. Henrichsen Copenhagen Business School

Sanne Hoffensetz-Andresen ordbogen.com

Sabine Kirchmeier-Andersen Dansk Sprognævn

Jens Otto Kjærum Prolog Development Center

Louise Bie Larsen Ankiro

Bente Maegaard Københavns Universitet

Sanni Nimb Det Danske Sprog- og Litteraturselskab

Jens-Erik Rasmussen Mikro Værkstedet

Peter Revsbech ordbogen.com

Hanne E. Thomsen Copenhagen Business School

Georg Rehm, Hans Uszkoreit

(udgivere, 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 978-3-642-30626-6 ISBN 978-3-642-30627-3 (eBook)
DOI 10.1007/978-3-642-30627-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940338

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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



FORORD

PREFACE

Denne rapport er en del af en hvidbogsserie som omhandler sprogteknologi og dets potentiale. Den henvender sig til undervisere, journalister, politikere og til sprogsamfundet generelt. Det varierer fra sprog til sprog hvor meget sprogteknologi der er tilgængeligt, og hvor meget det bliver brugt. Derfor er det også forskelligt fra land til land hvilken indsats der er behov for. De nødvendige tiltag afhænger af mange forskellige faktorer, såsom et givent sprogs kompleksitet og størrelsen af det samfund hvor det tales.

META-NET, som er et Network of Excellence finansieret af EU-Kommissionen, har gennemført en undersøgelse af sprogresurser og sprogteknologier i de europæiske lande. Undersøgelsen har fokuseret på de 23 officielle europæiske sprog så vel som andre vigtige nationale og regionale sprog i Europa. Undersøgelsen viser at der stadig er lang vej igen for de fleste sprogs vedkommende. En grundig ekspertanalyse og vurdering af den nuværende situation kan hjælpe med til at højne effekten af mere forskning og minimere risikoen for fejlinvesteringer.

META-NET består af 54 forskningscentre fra 33 lande (se s. 69) som arbejder med interessenter fra virksomheder, ministerier, forskningsinstitutioner og europæiske universiteter. Ved at udvikle en strategisk dagsorden for forskning inden for det sprogteknologiske område, arbejdes der mod en fælles vision for hvordan manglerne kan udbedres inden 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 differ. 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. 73). The analysis focuses 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 future research.

As of November 2011, META-NET consists of 54 research centres in 33 European countries (p. 69). 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 af dette dokument er forfatterne til sprograpporten for tysk taknemmelige for tilladelsen til at genbruge udvalgt sproguafhængigt materiale fra deres dokument [1].

Udarbejdelsen af denne sprograpport er blevet finansieret af EU's 7. rammeprogram og ICT Policy Support Programme under kontrakterne T4ME (kontrakt nr. 249119), CESAR (kontrakt nr. 271022), METANET4U (kontrakt nr. 270893) og META-NORD (kontrakt nr. 270899). Vi takker Lina Henriksen og Sussi Olsen for den danske oversættelse af den engelske version.

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].

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 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899). We thank Lina Henriksen and Sussi Olsen for the Danish translation of the English version.



INDHOLD CONTENTS

DET DANSKE SPROG I DEN DIGITALE TIDSALDER

1	Resumé	1
2	En risiko for sproget og en udfordring for sprogteknologien	4
2.1	Sproggrænserne hæmmer det europæiske informationssamfund	5
2.2	EU-sprog i fare	5
2.3	Sprogteknologi er en nøgleteknologi	6
2.4	Sprogteknologiens muligheder	6
2.5	Sprogteknologiens udfordringer	7
2.6	Menneskers og maskiners indlæring af sprog	7
3	Dansk i det europæiske informationssamfund	9
3.1	Generelle fakta	9
3.2	Særlige karakteristika for dansk	9
3.3	Den seneste udvikling	10
3.4	Sprogpleje i Danmark	11
3.5	Sproget i uddannelsen	12
3.6	Internationale aspekter	12
3.7	Danmark på internettet	13
4	Sprogteknologisk støtte til dansk	14
4.1	Systemarkitektur	14
4.2	Centrale udviklingsområder	15
4.3	Andre anvendelsesområder	22
4.4	Sprogteknologi i forskning og uddannelse	24
4.5	Nationale programmer og tiltag	26
4.6	Værktøjer og resurser	26
4.7	Sammenligning på tværs af sprog	28
4.8	Konklusioner	29
5	Om META-NET	32

THE DANISH LANGUAGE IN THE DIGITAL AGE

1	Executive Summary	33
2	Languages at Risk: a Challenge for Language Technology	36
2.1	Language Borders Hold back the European Information Society	37
2.2	Our Languages at Risk	37
2.3	Language Technology is a Key Enabling Technology	38
2.4	Opportunities for Language Technology	38
2.5	Challenges Facing Language Technology	39
2.6	Language Acquisition in Humans and Machines	39
3	The Danish Language in the European Information Society	41
3.1	General Facts	41
3.2	Particularities of the Danish Language	41
3.3	Recent Developments	42
3.4	Official Language Protection in Denmark	43
3.5	Language in Education	44
3.6	International Aspects	45
3.7	Danish on the Internet	45
4	Language Technology Support for Danish	47
4.1	Application Architectures	47
4.2	Core Application Areas	48
4.3	Other Application Areas	55
4.4	Language Technology in Research and Education	57
4.5	National Projects and Initiatives	58
4.6	Availability of Tools and Resources	59
4.7	Cross-language comparison	60
4.8	Conclusions	61
5	About META-NET	65
A	Referencer – References	67
B	META-NET Medlemmer – META-NET Members	69
C	META-NET-Hvidbogsserien – The META-NET White Paper Series	73

RESUMÉ

Informationsteknologien forandrer vores hverdag. Vi bruger computeren når vi skriver, læser, hører musik og ser billeder og film. Vi har computere i lommestørrelse som vi bruger til telefonopkald, e-mails, informationsøgning og underholdning, uanset hvor vi er. Men hvordan påvirkes sproget af denne massive digitalisering af information, viden og kommunikation? Vil vores sprog forandre sig eller måske endda forsvinde?

Alle vores computere er forbundet i et globalt netværk som hele tiden bliver stærkere. Pigen i Ipanema, toldofficeren i Padborg og ingeniøren i Katmandu kan chatte med venner på Facebook, men det er ikke særligt sandsynligt at de nogensinde mødes i online fora. Hvis de gerne vil vide hvordan man behandler ørepine, vil de alle tjekke Wikipedia for at lære mere om emnet, men ikke engang dér vil de læse den samme artikel. Når Europas internetbrugere i forskellige chatrum diskuterer Fukushima-atomulykkens indvirkning på europæisk energipolitik, gør de det i klart adskilte sprogfællesskaber. Hvad internettet forbinder, holdes stadig adskilt af brugernes sprog. Vil det altid være sådan?

Mange af verdens 6000 sprog vil ikke overleve i et globaliseret, digitaliseret informationssamfund. Man regner med at mindst 2000 sprog vil uddø i de kommende årtier. Andre vil stadig spille en rolle i familier og inden for mindre geografiske områder, men måske ikke i finansverdenen og i den akademiske verden. Hvad er chancerne for det danske sprogs overlevelse?

Ca. 5 mio. har dansk som modersmål, så dansk må anses for at være et relativt lille sprog, i hvert fald sammenlignet med flere andre EU-sprog. I lighed med andre in-

dustrialiserede lande er vores hverdag i høj grad påvirket af det engelske sprog. Store internationale virksomheder bruger i stigende grad engelsk som deres virksomhedssprog, og engelsk er ved at blive *lingua franca* inden for højere uddannelser, ligesom det er inden for videnskab og teknologi hvor det har haft den rolle i lang tid.

Man hører ofte kritik af den støt stigende brug af anglicismer, og nogle mennesker frygter ligefrem at det danske sprog er ved at blive gennemsyret af engelske ord og udtryk. Men danske ord og udtryk kan man kun bevare ved rent faktisk ved at bruge dem – ofte og bevidst; lingvistisk polemik om udenlandsk indflydelse og statslig regulering hjælper som regel ikke. Vores største bekymring bør dog ikke være den gradvise anglisering af vores sprog, men snarere at dansk kan forsvinde ud af store dele af vores liv. Videnskab, luftfart og de globale finansmarkeder har reelt brug for et verdensomspændende *lingua franca*, men vi bør værne om vores eget sprog inden for områder som primært angår landets borgere, fx national politik, administrative procedurer, love, kultur og handel.

Et sprogs status afhænger ikke kun af det antal af mennesker, bøger, film og tv-stationer, der bruger det, men også af at det findes i det digitale informationsrum og bruges i softwareprogrammer. Her er det danske sprog temmelig godt placeret: mange internationale softwareprodukter findes i danske versioner, det danske Wikipedia er i vækst, og med mere end 1 million internetdomæner registreret i 2011 er dansk godt repræsenteret på webben set i forhold til befolkningens størrelse.

Men inden for sprogteknologien mangler det danske sprog både værktøjer, teknologier og resurser for at