META NET

White Paper Series Weißbuch-Serie

THE GERMAN LANGUAGE IN THE DIGITAL AGE

DIE DEUTSCHE SPRACHE IM DIGITALEN ZEITALTER

Aljoscha Burchardt
Markus Egg
Kathrin Eichler
Brigitte Krenn
Jörn Kreutel
Annette Leßmöllmann
Georg Rehm
Manfred Stede
Hans Uszkoreit
Martin Volk

Springer

White Paper Series Weißbuch-Serie

THE GERMAN LANGUAGE IN THE DIGITAL AGE

DIE DEUTSCHE SPRACHE IM DIGITALEN ZEITALTER

Aljoscha Burchardt DFKI
Markus Egg Humboldt-Universität zu Berlin
Kathrin Eichler DFKI
Brigitte Krenn ÖFAI
Jörn Kreutel FH Brandenburg
Annette Leßmöllmann Hochschule Darmstadt
Georg Rehm DFKI
Manfred Stede Universität Potsdam
Hans Uszkoreit Universität des Saarlandes, DFKI
Martin Volk Universität Zürich

Georg Rehm, Hans Uszkoreit
(Herausgeber, 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 978-3-642-27165-6 ISBN 978-3-642-27166-3 (eBook)
DOI 10.1007/978-3-642-27166-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012947391

Printed on acid-free paper

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

VORWORT PREFACE

Dieses Weißbuch gehört zu einer Serie, die Wissen über Sprachtechnologie und deren Potenzial vermitteln soll und sich insbesondere an Journalisten, Politiker, Sprachgemeinschaften und Lehrende richtet. Die derzeitige Verfügbarkeit und Nutzung von Sprachtechnologie in Europa variiert stark je nach Sprache. Folglich müssen auch die notwendigen Maßnahmen für die künftige Unterstützung dieses Bereiches durch Forschung und Entwicklung variieren. Die Maßnahmen hängen von zahlreichen Faktoren ab, z. B. der Komplexität einer Sprache und der Anzahl ihrer Sprecher.

META-NET, ein von der Europäischen Kommission gefördertes Spitzenforschungsnetzwerk, hat die aktuellen Sprachressourcen und -technologien in der vorliegenden Weißbuch-Serie analysiert (siehe S. 81). Die Analyse umfasst die 23 europäischen Amtssprachen sowie weitere wichtige nationale und regionale Sprachen Europas. Die Ergebnisse zeigen, dass für alle Sprachen beträchtliche Defizite in der technologischen Unterstützung sowie signifikante Forschungslücken existieren. Die ausführliche Expertenanalyse und Bewertung der aktuellen Situation dient dazu, die Wirksamkeit künftiger Forschung zu maximieren.

META-NET besteht aus 54 Forschungszentren in 33 Ländern (siehe S. 77), die mit Interessensvertretern aus Wirtschaft (Softwareunternehmen, Technologieanbietern, Nutzern), Verwaltung, NGOs, Sprachgemeinschaften und europäischen Universitäten zusammenarbeiten. Zusammen mit diesen Gruppen entwickelt META-NET eine gemeinsame Vision für das Technologiegebiet und eine strategische Forschungsagenda für das mehrsprachige Europa 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. 81). 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 to maximise the impact of future research.

META-NET consists of 54 research centres in 33 European countries [1] (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.

Die Erstellung dieses Weißbuchs wurde mit Mitteln aus dem Siebten Rahmenprogramm und dem Programm zur Unterstützung der Politik für Informations- und Kommunikationstechnologien der Europäischen Kommission im Rahmen der Verträge T4ME (Finanzhilfevereinbarung 249 119), CESAR (Finanzhilfevereinbarung 271 022), METANET4U (Finanzhilfevereinbarung 270 893) und META-NORD (Finanzhilfevereinbarung 270 899) finanziert.

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).

INHALTSVERZEICHNIS TABLE OF CONTENTS

DIE DEUTSCHE SPRACHE IM DIGITALEN ZEITALTER

THE GERMAN LANGUAGE IN THE DIGITAL AGE

ZUSAMMENFASSUNG

Die Sprache ist ein zentraler Bestandteil unserer Kultur und unserer individuellen Identität. Sie ist einem ständigen Wandel unterworfen, der ganz besonders von Entwicklungen der Kulturtechniken und dem Kontakt mit anderen Sprachen bestimmt wird. Noch nie in der Geschichte der Menschheit haben sich Technologien des täglichen Lebens und der Austausch zwischen den Völkern so radikal verändert wie in unserer Zeit. Wir sind die erste Generation, die Computer ganz normal im Alltag zum Schreiben, zum Lesen, zum Rechnen und zur Informationssuche einsetzt, aber auch zum Musikhören und zum Betrachten von Fotos und Filmen. In der Tasche tragen wir kleine Computer mit uns herum, mit denen wir telefonieren und korrespondieren, uns informieren und amüsieren, und das an jedem beliebigen Ort. Welche Bedeutung hat die Digitalisierung von Informationen, Wissen und Kommunikation für unsere Sprache? Wird sie sich ändern, wird sie irgendwann verschwinden?

Unsere Computer sind miteinander verbunden, ihr weltumspannendes Netz wird immer dichter und mächtiger. Das Girl in Ipanema, der Zöllner in Lindau und die Ingenieurin in Katmandu treffen ihre Freunde auf Facebook und dennoch werden sie sich in den Communitys und Foren untereinander nie begegnen. Wenn sie ein Ohrenschmerz beunruhigt, suchen sie in der Wikipedia nach möglichen Ursachen und lesen dennoch nicht denselben Artikel. Wenn die europäischen Netzbürger in Foren und Chaträumen die Konsequenzen des Fukushima-Unglücks auf die europäische Energiepolitik diskutieren, dann nach Sprachgemeinschaften getrennt. Was das Netz verbindet, trennen immer noch die Sprachen seiner Benutzer. Muss das so bleiben?

Viele unserer 6000 Sprachen werden in der Informationsgesellschaft des digitalen Zeitalters nicht überleben; man geht derzeit von mindestens 2000 Sprachen aus, die dem Untergang geweiht sind. Andere werden zwar eine Funktion in Familie und Nachbarschaft behalten, aber nicht im Geschäftsleben oder auf der Universität. Wie stehen die Chancen für das Deutsche?

Die deutsche Sprache ist mit ihren fast 100 Millionen Sprechern im internationalen Vergleich gar nicht so schlecht aufgestellt. Es gibt eine große Zahl öffentlich-rechtlicher Fernsehsender mit deutschsprachigem Programm (Deutschland: 23, Österreich: 6, Schweiz: 4) sowie mehr als 50 private Free-TV-Sender. Die meisten internationalen Spielfilme werden nach wie vor für das Deutsche synchronisiert. Auch der lange schon totgesagte Buch- und Zeitschriftenmarkt ist auf hohem Niveau stabil.

Trotz eines international starken Rückgangs der Rolle des Deutschen ist die Sprache in Europa immer noch die am zweithäufigsten erlernte Fremdsprache. Auf Grund der besonderen Vorgeschichte der europäischen Integration im 20. Jahrhundert haben Deutschland und Österreich bisher nicht konsequent den Status für die deutsche Sprache in der Politik und Verwaltung der EU eingefordert, der ihr gemäß der Zahl ihrer Sprecher und der Unionsverträge zustehen würde. Diese Zurückhaltung brachte den deutschsprachigen Ländern nicht etwa Nachteile, sondern hat zu der seit Jahren beobachteten Imageverbesserung beigetragen.

Vielfach wird aber in den deutschsprachigen Ländern die schleichende Anglisierung des Deutschen prophezeit und beklagt. Hier gibt unsere Studie Entwarnung. Die deutsche Sprache hat die Unterwanderung durch die lateinischen und griechischen Begriffe der Wissenschaftssprachen überlebt wie auch die Französisierung im 18. und frühen 19. Jahrhundert. Dem Opfern schöner Wörter und Wendungen des Deutschen tritt man am besten durch deren häufige und bewusste Verwendung entgegen und nicht durch Polemik und Verordnungen. Unsere Hauptsorge sollte nicht einer schleichenden Anglisierung unserer Sprache gelten, sondern der Verdrängung des Deutschen in wichtigen Bereichen des Lebens. Hier sind nicht die Bereiche gemeint, die einer weltweiten lingua franca bedürfen, wie z. B. der Wissenschaft, der Luftfahrt und der globalen Finanzmärkte, sondern die Vielzahl von Lebensbereichen, in denen die Nähe zum Bürger wichtiger ist als die zu internationalen Partnern, wie Innenpolitik, Verwaltung, Kultur und Einzelhandel.

Der Status einer Sprache hängt nicht nur von der Zahl der Sprecher ab und der Menge von Büchern, Filmen und Fernsehsendern, sondern immer mehr auch von der Präsenz der Sprache im digitalen Informationsraum und den verfügbaren Softwareprodukten für diese Sprache. Auch hier stehen die Zeichen für das Deutsche nicht so schlecht. Alle wesentlichen internationalen Softwareprodukte sind in deutschen Versionen verfügbar. Die deutsche Wikipedia ist die zweitgrößte Wikipedia nach der englischen. Mit mehr als 14 Millionen Einträgen ist die Domäne .de das größte Länderkürzel der Welt.

Sprachtechnologisch ist das Deutsche derzeit auch gut durch Produkte und Ressourcen versorgt. Es gibt Programme für die Sprachsynthese, Spracherkennung, Rechtschreibkorrektur und die Grammatiküberprüfung. Die vielen Produkte zur maschinellen Übersetzung können allerdings bislang selten sprachlich korrekte Übersetzungen erzeugen, besonders bei Übersetzungen ins Deutsche, was zu einem großen Teil an den Eigenschaften unserer Sprache liegt.

Die Informationstechnologie steht vor einer Revolution. Nach Personalisierung, Vernetzung, Miniaturisierung, Multimedialisierung, Mobilisierung und Cloud-Computing kündigt sich eine Technologiegeneration an, die dem Menschen aufs Wort gehorcht und ihn durch Kenntnis seiner Sprache besser in Information und Kommunikation unterstützt. Vorboten dieser Entwicklung sind der Übersetzungsdienst Google Translate, der zwischen 57 Sprachen übersetzt, IBMs Supercomputer Watson, der als Quizchampion die amerikanischen Jeopardy-Meister schlug, und Apples mobiler Assistent Siri, der auf dem iPhone Fragen beantwortet.

Die nächste IT-Generation wird die menschliche Sprache beherrschen – zumindest soweit, dass der Mensch mit der Technologie in seiner Sprache kommunizieren und die Technologie automatisch die wichtigsten Informationen aus dem digitalen Wissen der Welt ziehen kann. Die sprachfähige Technik wird verlässlich übersetzen und dolmetschen können, sie wird Gespräche und Texte zusammenfassen und sie wird beim Lernen helfen. Zum Beispiel wird sie die Zuwanderer, die die deutschsprachigen Länder weiterhin benötigen werden, beim Erlernen der deutschen Sprache und bei der kulturellen Integration unterstützen.

Für diese Stufe der Technologie reichen aber Zeichensätze und Wörterbücher nicht aus, auch nicht Rechtschreibkorrektur und Ausspracheregeln. Wenn es um die Modellierung von Sprachverstehen geht und um die automatische Generierung von richtigen Fragen und Antworten, muss die Technologie die Sprache umfassender modellieren und über die Syntax hinaus bis zur Semantik vorstoßen.

Hier klafft nun aber ein Abstand zwischen dem Englischen und dem Deutschen, der derzeit größer wird und nicht kleiner. Nach erfolgreichen Forschungsanstrengungen in den achtziger und neunziger Jahren ist