META NET

White Paper Series Witboekserie

THE DUTCH LANGUAGE IN THE DIGITAL AGE

HET NEDERLANDS IN HET DIGITALE TIJDPERK

Jan Odijk

Springer

White Paper Series Witboekserie

THE DUTCH LANGUAGE IN THE DIGITAL AGE

HET NEDERLANDS IN HET DIGITALE TIJDPERK

Jan Odijk Universiteit Utrecht

Georg Rehm, Hans Uszkoreit
(redactie, 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 978-3-642-25977-7 ISBN 978-3-642-25978-4 (eBook)
DOI 10.1007/978-3-642-25978-4
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012940339

VOORWOORD PREFACE

Dit witboek maakt deel uit van een serie die kennis over taaltechnologie en het potentieel ervan bevordert. Het richt zich op journalisten, politici, taalgemeenschappen en anderen. De beschikbaarheid en het gebruik van taaltechnologie in Europa verschilt per taal. Daarom verschillen de acties die nodig zijn om ondersteuning van onderzoek en ontwikkeling van taaltechnologie te bevorderen eveneens per taal. De vereiste acties hangen af van veel factoren, zoals de complexiteit van een taal en de omvang van de taalgemeenschap.

META-NET, een 'Network of Excellence' gefinancierd door de Europese Commissie, heeft een analyse gemaakt van de huidige taalbronnen en -technologieën. Deze analyse richtte zich op de 23 officiële Europese talen en op andere belangrijke nationale en regionale talen in Europa. De resultaten van deze analyse suggereren dat er veel significante lacunes zijn voor iedere taal. Een gedetailleerdere expertanalyse en beoordeling van de huidige situatie zal ertoe bijdragen de impact van additioneel onderzoek te maximaliseren en risico's te verminderen.

META-NET bestaat tegenwoordig uit 54 onderzoekscentra in 33 landen (p. 75) die werken met belanghebbenden uit de economie (softwarebedrijven, technologieleveranciers en gebruikers), de overheid, onderzoek, niet-gouvernementele organisaties, het onderwijs, en taalgemeenschappen. Samen creëren zij een gemeenschappelijke technologievisie en ontwikkelen daarbij een strategische onderzoeksagenda die laat zien hoe taaltechnologische toepassingen lacunes in het onderzoek aan kunnen pakken tegen 2020.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, 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. 79). 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. This detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

META-NET currently consists of 54 research centres from 33 European countries (p. 75). META-NET is working with stakeholders from economy (software companies, technology providers and users), government, research, non-governmental organisations, education, and language communities in creating a common technology vision and strategic research agenda for multilingual Europe 2020.

De auteurs van dit document bedanken de auteurs van het taalwitboek voor het Duits [1] voor de toestemming om geselecteerd taalonafhankelijk materiaal uit hun witboek hier te hergebruiken. Verder wil de auteur Catia Cucchiarini (Nederlandse Taalunie), Walter Daelemans (Universiteit Antwerpen), Alice Dijkstra (NWO), Jean-Pierre Martens (Universiteit Gent), Jacomine Nortier (Universiteit Utrecht), Peter Spyns (Nederlandse Taalunie) en Remco van Veenendaal (TST-centrale) bedanken voor hun bijdragen aan het witboek.

De ontwikkeling van dit witboek is gefinancierd door het Zevende Kaderprogramma en het ondersteuningsprogramma voor ICT-beleid van de Europese Commissie onder de contracten T4ME (Toewijzingsovereenkomst 249119), CESAR (Toewijzingsovereenkomst 271022), METANET4U (Toewijzingsovereenkomst 270893) en META-NORD (Toewijzingsovereenkomst 270899).

The authors are grateful to the authors of the White Paper on German [1] for permission to re-use selected language-independent materials from their document. Furthermore, the author would like to thank Catia Cucchiarini (Dutch Language Union), Walter Daelemans (Antwerp University), Alice Dijkstra (NWO), Jean-Pierre Martens (Ghent University), Jacomine Nortier (Utrecht University), Peter Spyns (Dutch Language Union) and Remco van Veenendaal (HLT Agency) for their contributions to this white paper.

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).

INHOUDSOPGAVE TABLE OF CONTENTS

HET NEDERLANDS IN HET DIGITALE TIJDPERK

THE DUTCH LANGUAGE IN THE DIGITAL AGE

1

MANAGEMENTSAMENVATTING

Informatietechnologie verandert ons alledaagse leven. We gebruiken computers om te schrijven, te bewerken, te rekenen en om informatie te zoeken, en steeds meer om te lezen, naar muziek te luisteren, en om foto's en films te bekijken. We dragen kleine computers in onze zakken en gebruiken ze – waar we ook zijn – om op te bellen, e-mails te schrijven, informatie te verkrijgen en ons te onderhouden. Hoe beïnvloedt deze massale digitalisatie van informatie, kennis, en alledaagse communicatie onze taal? Zal onze taal veranderen of zelfs verdwijnen?

Al onze computers zijn met elkaar verbonden in een toenemend dicht en krachtig network. Het meisje in Ipanema, de douaneambtenaar in Venlo, en de ingenieur in Kathmandu kunnen allemaal chatten met hun vrienden op Facebook, maar ze zullen elkaar waarschijnlijk nooit in online gemeenschappen en forums ontmoeten. Als ze zich er zorgen over maken hoe oorpijn behandeld moet worden, zullen ze allemaal Wikipedia raadplegen om dit uit te zoeken, maar zelfs dan zullen ze niet hetzelfde artikel lezen. Wanneer de internettende burgers van Europa de effecten van het kernongeluk in Fukushima op het Europese energiebeleid bespreken in forums en chatsessies, doen ze dat in netjes gescheiden taalgemeenschappen. Wat het internet verbindt wordt nog steeds verdeeld door de talen van de gebruikers ervan. Zal het altijd zo zijn?

Veel van de 6000 talen van de wereld zullen niet overleven in een geglobaliseerde digitale informatiemaatschappij. Er wordt geschat dat minstens 2000 talen gedoemd zijn te verdwijnen in de komende decennia. Andere zullen een rol blijven spelen in families en buurtschappen, maar niet in de wijdere bedrijfs- en academische wereld. Wat zijn de overlevingskansen voor het Nederlands?

Met ongeveer 23 miljoen moedertaalsprekers is het Nederlands de achtste meest gesproken natuurlijke taal in de Europese Unie. Het is slechts een 'kleine' taal in vergelijking met de naburige talen Engels, Duits en Frans. De invloed van het Engels op het taalgebruik is significant, vooral onder jongeren. Het bedrijfsleven, zelfs wanneer het opereert in de Lage Landen (Nederland en Vlaanderen), gebruikt vaak Engels, vooral in multinationals. De communicatietaal in de wetenschap is het Engels. Hoger onderwijs wordt in toenemende mate in het Engels gegeven. Boekpublicaties in het Nederlands, films, en TV- en radioprogramma's in het Nederlands bestaan natuurlijk, maar de markt ervoor is nogal klein.

In de Europese Unie is het Nederlands een officiële taal, maar het Nederlands wordt nauwelijks in de Europese Unie gebruikt. De Nederlandse taal zal zeker niet helemaal verdwijnen, maar er is wel een reëel gevaar dat het gebruik van het Nederlands verdwijnt uit belangrijke gebieden van ons persoonlijke leven, in het bijzonder uit gebieden die te maken hebben met discussies over en beslissingen over beleidskwesties, administratieve procedures, de wetgeving, cultuur en het winkelen.

De status van een taal hangt niet alleen af van het aantal sprekers of het aantal boeken, films en Tv-stations in die taal, maar ook op de aanwezigheid van de taal in de digitale informatieruimte en in softwaretoepassingen. De Nederlandse Wikipedia is de op acht na grootste van de

wereld. Met ongeveer 1.24 miljoen internetdomeinen, is het topniveau landendomein .nl van Nederland de elfde landenextensie. Dat is niet slecht voor een klein land zeker aangezien het verder groeit. De hoeveelheid Nederlandstalige data op het web is natuurlijk heel klein in vergelijking tot het Engels en de taaldata van verschillende andere grotere talen zoals Duits en Frans. Dankzij het STEVIN-programma, dat het versterken van de Nederlandse taal expliciet als een van zijn doelstellingen had, doet het Nederlands het ook niet slecht wat betreft software voor de Nederlandse taal en wat betreft Nederlandstalige taalbronnen die nodig zijn om dergelijke software te ontwikkelen. Het speelt in dezelfde liga als het Frans en het Duits, maar loopt nog ver achter op het Engels.

De informatie- en communicatietechnologie bereidt zich nu voor op de volgende revolutie. Na persoonlijke computers, netwerken, miniaturisatie, multimedia, mobiele apparaten, en cloud-computing, zal de volgende generatie van technologie software bevatten die niet alleen maar gesproken klanken of geschreven letters begrijpt, maar hele woorden en zinnen, en die gebruikers veel beter ondersteunt omdat het hun taal spreekt, kent en begrijpt. Voorlopers van deze ontwikkeling zijn de gratis online dienst Google Translate, dat tussen 57 talen vertaalt, de Watson supercomputer van IBM die in staat was de kampioen van de Verenigde Staten in het spel "Jeopardy" te verslaan, en de mobiele assistent Siri van Apple voor de iPhone, die kan reageren op stemcommando's en vragen kan beantwoorden in het Engels, Duits, Frans en Japans.

De volgende generatie informatietechnologie zal natuurlijke taal zo goed beheersen dat menselijke gebruikers in staat zullen zijn te communiceren in hun eigen taal als ze de technologie gebruiken. Apparaten zullen op basis van makkelijk te gebruiken stemcommando's in staat zijn automatisch het belangrijkste nieuws en de belangrijkste informatie te vinden in de digitale kennisbank van de wereld. Van taaltechnologie voorziene software zal in staat zijn automatisch te vertalen of tolken bij te staan; om gesprekken en documenten samen te vatten; en om gebruikers te ondersteunen in leerscenario's. Bijvoorbeeld, het zal immigranten – zoals vereist door de regeringen van de Lage Landen – helpen de Nederlandse taal te leren en volledig te integreren in de cultuur van het land.

De volgende generatie informatietechnologie zal industriële en dienstenrobots (die momenteel in onderzoekslaboratoria ontwikkeld worden) in staat stellen op betrouwbare manier te interpreteren wat hun gebruikers hen willen laten doen om dan 'trots' over hun resultaten te rapporteren.

Dit prestatieniveau reikt ver uit boven simpele karakterverzamelingen en woordenboeken, spellingscontrole en uitspraakregels. De technologie moet ophouden met simplistische benaderingen en taal op een alomvattende manier modelleren, en daarbij syntaxis evenals semantiek in beschouwing nemen om de portee van vragen te begrijpen en rijke en relevante antwoorden te genereren.

Er is echter een gapend technologisch gat tussen het Engels en andere talen, inclusief het Nederlands, en dit gat wordt momenteel alleen maar groter. Commercieële bedrijven onderzoeken, ontwikkelen, verkopen en gebruiken taaltechnologie initieel voor het (Amerikaans) Engels, simpelweg omdat de interessantste markten zich in landen bevinden waar (Amerikaans) Engels gesproken wordt. De technologische voorlopers die boven genoemd zijn komen in enkele gevallen pas veel later beschikbaar voor het Nederlands, en in veel gevallen zelfs helemaal niet. Het Nederlands is bij deze ontwikkelingen nauwelijks in het zicht.

Internationale technologische competities laten gewoonlijk zien dat resultaten voor de automatische analyse van het Engels beter zijn dan die voor het Nederlands, alhoewel (of precies omdat) de analysemethodes gelijkaardig of zelfs identiek zijn. Dit geldt voor het