

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE
A MAGYAR NYELV A DIGITÁLIS KORBAN

Simon Eszter
Lendvai Piroska
Németh Géza
Olaszy Gábor
Vicsi Klára



White Paper Series

Fehér könyvek sorozat

THE
HUNGARIAN
LANGUAGE IN
THE DIGITAL
AGE

A MAGYAR
NYELV A
DIGITÁLIS
KORBAN

Simon Eszter MTA Nyelvtudományi Intézet

Lendvai Piroska MTA Nyelvtudományi Intézet

Németh Géza BME

Olaszy Gábor BME

Vicsi Klára BME

Georg Rehm, Hans Uszkoreit
(szerkesztők, 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 978-3-642-30378-4 ISBN 978-3-642-30379-1 (eBook)
DOI 10.1007/978-3-642-30379-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012945123

© 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)



ELŐSZÓ

PREFACE

Ez a fehér könyv egy sorozat részét képezi, amelynek célja, hogy felhívja a figyelmet a nyelvtechnológiára és az abban rejlő lehetőségekre. Elsősorban oktatókat, újságírókat, politikusokat és nyelvi közösségeket szólít meg. Az európai nyelvek nyelvtechnológiai feldolgozottsága és a nyelvtechnológia elterjedtsége meglehetősen eltérő. Ezért a nyelvtechnológia fejlődéséhez és a kutatás elősegítéséhez szükséges lépések is nyelvenként mások és mások, és olyan különféle tényezőkhöz kötődnek, mint az adott nyelv összetettsége, vagy a nyelvet használó közösség nagysága.

A META-NET, az Európai Bizottság által alapított hálózat felmérést végzett a rendelkezésre álló nyelvi erőforrásokról és technológiákról (lásd a 73. oldalt). Ez a felmérés a 23 hivatalos európai nyelv mellett egyéb nemzeti és regionális nyelvekre is kiterjed, és eredményei rámutatnak az egyes nyelvek terén fellelhető kutatási hiányosságokra. Egy, a jelenlegi helyzetet bemutató részletes szakértői elemzés és értékelés segíthet a további kutatások hatásának maximalizálásában.

A META-NET 33 ország 54 kutatóközpontjából áll (2011. novemberi helyzet szerint, lásd a 69. oldalt), akik a területtel foglalkozó vállalkozásokkal, kormányzati szervezetekkel, kutatószervezetekkel, szoftvercégekkel, szolgáltatókkal és európai egyetemekkel dolgoznak együtt. Egységes technológiai víziót alkotva egy olyan stratégiai kutatási terv létrehozásán dolgoznak, amelyben megfogalmazzák, hogyan tudnak a nyelvtechnológiai alkalmazások a kutatási hiányosságokon enyhíteni a 2020-ig terjedő időszakban.

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 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. 69). META-NET is working with stakeholders from economy (software companies, technology providers, 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.

A dokumentum szerzői köszönettel tartoznak a német fehér könyv szerzőinek azért, hogy engedélyezték a német változat egyes nyelvfüggetlen részeinek újrafelhasználását [1].

A fehér könyv megírását az Európai Bizottság 7. keretprogramja és ICT PSP programja támogatta a T4ME (szerződés szám: 249 119), a CESAR (szerződés szám: 271 022), a METANET4U (szerződés szám: 270 893) és a META-NORD (szerződés szám: 270 899) projekteken keresztül.

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 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).



TARTALOMJEGYZÉK TABLE OF CONTENTS

A MAGYAR NYELV A DIGITÁLIS KORBAN

1	Vezetői összefoglaló	1
2	Veszélyben a nyelveink: Kihívás a nyelvtechnológiának	4
2.1	Az európai információs társadalom gátjai: a nyelvi határok	5
2.2	Veszélyben a nyelveink	5
2.3	Nyelvtechnológia: egy kulcsfontosságú technológia	6
2.4	A nyelvtechnológia lehetőségei	6
2.5	A nyelvtechnológia kihívásai	7
2.6	Emberi és gépi nyelvelsajátítás	7
3	A magyar nyelv az európai információs társadalomban	10
3.1	Általános tények	10
3.2	A magyar nyelv különlegességei	10
3.3	Modernkori fejlődés	11
3.4	Nyelvművelés Magyarországon	12
3.5	A magyar nyelv az oktatásban	13
3.6	Nemzetközi vonatkozások	13
3.7	A magyar nyelv az interneten	14
4	Nyelvtechnológia magyarul	15
4.1	A nyelvtechnológiai alkalmazások felépítése	15
4.2	A fő alkalmazási területek	16
4.3	További alkalmazási területek	24
4.4	Nyelvtechnológia az oktatásban	26
4.5	Hazai projektek	27
4.6	Az eszközök és erőforrások elérhetősége	28
4.7	Nyelvek közötti összehasonlítás	28
4.8	Összegzés	30
5	A META-NET-ről	34

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE

1	Executive Summary	35
2	Languages at Risk: a Challenge for Language Technology	38
2.1	Language Borders Hold back the European Information Society	39
2.2	Our Languages at Risk	39
2.3	Language Technology is a Key Enabling Technology	40
2.4	Opportunities for Language Technology	40
2.5	Challenges Facing Language Technology	41
2.6	Language Acquisition in Humans and Machines	41
3	The Hungarian Language in the European Information Society	43
3.1	General Facts	43
3.2	Particularities of the Hungarian Language	43
3.3	Recent Developments	45
3.4	Official Language Protection in Hungary	45
3.5	Language in Education	46
3.6	International Aspects	46
3.7	Hungarian on the Internet	47
4	Language Technology Support for Hungarian	48
4.1	Application Architectures	48
4.2	Core Application Areas	49
4.3	Other Application Areas	57
4.4	Educational Programmes	58
4.5	National Projects and Initiatives	59
4.6	Availability of Tools and Resources	60
4.7	Cross-language comparison	60
4.8	Conclusions	62
5	About META-NET	66
A	Hivatkozások – References	67
B	META-NET tagok – META-NET Members	69
C	A META-NET fehér könyvek sorozat – The META-NET White Paper Series	73

VEZETŐI ÖSSZEFOGLALÓ

Az információs technológia jelentősen megváltoztatta mindennapi életünket. Jellemzően számítógépet használunk írásra, szerkesztésre, számolásra és információkeresésre, továbbá egyre inkább olvasásra, zenehallgatásra, fotó- és filmnézésre. Kis számítógépeket hordunk a zsebünkben, és használjuk őket telefonálásra, e-mailíráásra, információszerzésre és szórakozásra, bármerre is járunk. Milyen hatással van az információnak, a tudásnak és a mindennapi kommunikációnak ez a masszív digitalizálódása a nyelvünkre? Megváltozik a nyelvünk, vagy eltűnik? Számítógépeink egy nagy és erőteljes globális hálózat részeit képezik. A lány Ipanemában, a hivatalnok Budapesten és a mérnök Delhiben ugyanúgy tudnak csetelni a barátaikkal a Facebookon, de nem valószínű, hogy valaha is találkoznak egymással online közösségekben vagy fórumokon. Ha azon aggódnak, hogy hogy lehet kezelni a fülfájást, akkor mindannyian a Wikipédiát fogják megnézni, de nem ugyanazt a cikket fogják olvasni. Amikor európai netezők a fukushimai atomkatasztrófának az európai energetikai piacra gyakorolt hatásairól beszélgetnek a fórumokon, akkor ezt jellemzően nyelviileg szeparált közösségekben teszik. Amit az internet összeköt, azt a használók nyelvi korlátai még mindig szétválasztják. Ez vajon mindig így lesz? A világ 6000 nyelve közül sok nem fog életben maradni egy globalizált digitális információs társadalomban. Becslések szerint legalább 2000 nyelv kihalásra van ítélve a következő évszázadokban. Mások szerepet fognak ugyan játszani a családi és ismerősi körben, de a szélesebb üzleti és tudományos szférában nem. Milyenek a magyar nyelv esélyei a túlélésre?

Becslések szerint a magyar nyelvet összesen 13 millióan beszélik, ezzel a 12. helyen áll a legtöbb beszélővel rendelkező nyelvek listáján Európában. A Magyar Köztársaság államnyelve, ahol a 10 milliós lakosságnak kb. 97%-a magyar anyanyelvű. A szomszédos hét országban is találunk magyar nyelvű közösségeket, amelyek közül a legnagyobb a romániai, megközelítőleg másfél millió nyelvhasználóval. Ezen felül emigráns közösségek használják világszerte, elsősorban az Amerikai Egyesült Államokban, Kanadában és Izraelben.

A magyar nyelv szigeteket alkot Európában, ugyanis a legtöbb európai nyelv az indoeurópai nyelvcsaládba tartozik, de a magyar nem. A magyar finnugor nyelv, rokonai a finn, az észt és néhány más, Oroszországban élő népek által beszélt nyelv. A magyar nyelv a legtöbb beszélővel rendelkező nem indoeurópai nyelv Európában, de – ellentétben olyan világnyelvekkel, mint az angol vagy a kínai, és az olyan gyakran használt európai nyelvekkel, mint a német vagy a francia – a magyar nem játszik prominens szerepet nemzetközi szinten.

Sokan panaszkodnak Magyarországon az anglicizmusok egyre erősödő használata miatt, és attól tartanak, hogy a magyar nyelvet elárasztják az angol szavak és kifejezések. Ez a megközelítés félrevezető. A magyar nyelv már túlélte az új szavak hatását, melyeket különböző török nyelvekből vettünk át a honfoglalás előtti korban, és túlélte az erős szláv hatást is a Kárpát-medencében. Később az Oszmán Birodalom része volt az ország 150 évig, majd a Habsburg Birodalom ideje alatt a latin és a német nyelv hatása volt nagyon erős. Kedves kis magyar szavaink elvesztésének egy jó ellenszere, ha használjuk