META NET

White Paper Series Série Bílé knihy

THE CZECH LANGUAGE IN THE DIGITAL AGE

ČEŠTINA V DIGITÁLNÍM VĚKU

Ondřej Bojar
Silvie Cinková
Jan Hajič
Barbora Hladká
Vladislav Kuboň
Jiří Mírovský
Jarmila Panevová
Nino Peterek
Johanka Spoustová
Zdeněk Žabokrtský

Springer

White Paper Series Série Bílé knihy

THE CZECH LANGUAGE IN THE DIGITAL AGE

ČEŠTINA V DIGITÁLNÍM VĚKU

Ondřej Bojar Charles University in Prague
Silvie Cinková Charles University in Prague
Jan Hajič Charles University in Prague
Barbora Hladká Charles University in Prague
Vladislav Kuboň Charles University in Prague
Jiří Mírovský Charles University in Prague
Jarmila Panevová Charles University in Prague
Nino Peterek Charles University in Prague
Johanka Spoustová Charles University in Prague
Zdeněk Žabokrtský Charles University in Prague

Georg Rehm, Hans Uszkoreit
(editoři, 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 978-3-642-30705-8 ISBN 978-3-642-30706-5 (eBook)
DOI 10.1007/978-3-642-30706-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012942721

PŘEDMLUVA PREFACE

Tato Bílá kniha je součástí série, která podporuje znalosti jazykových technologií a jejich potenciál. Je určena pedagogům, novinářům, politikům, různým jazykovým komunitám a dalším. Dostupnost a využívání jazykových technologií se v Evropě u jednotlivých jazyků liší. V důsledku toho se pro každý jazyk liší také kroky, které je nutné podniknout pro další podporu výzkumu a vývoje jazykových technologií. Tyto plánované postupy závisí na mnoha faktorech, jako je složitost daného jazyka či velikost jeho komunity. META-NET (excelentní internetová síť) financovaný Evropskou komisí provedl analýzu současných jazykových zdrojů a technologií. Tato analýza se zaměřila na 23 oficiálních evropských jazyků a na další významné národní a regionální jazyky v Evropě. Výsledky analýzy naznačují, že ve výzkumu každého jazyka je značné množství mezer. Podrobnější expertní analýza a hodnocení současné situace přitom přispějí k maximalizaci účinku dalšího výzkumu a minimalizaci možných rizik. META-NET se skládá z 54 výzkumných center z 33 zemí, které pracují s podílníky z komerčních firem, vládních agentur, průmyslu, výzkumných organizací, softwarových firem, s poskytovateli technologií a evropských univerzit. Dohromady mají jednu společnou vizi – vyvíjejí strategický plán výzkumu, který ukazuje, jak aplikace jazykových technologií mohou do roku 2020 vyřešit případné mezery ve výzkumu.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, 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 for each language. 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. This 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 many significant research gaps for each language. A more detailed expert analysis and assessment of the current situation will help maximize the impact of additional research and minimize any risks. META-NET consists of 54 research centres from 33 countries that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

Autoři tohoto dokumentu děkují autorům Bílé knihy pro němčinu za povolení použít vybrané jazykově nezávislé části z jejich dokumentu [1]. Zároveň děkujeme za milou spolupráci kolegům Jan Cuřínovi, Evě Hajičové, Jirkovi Hanovi, Karlu Olivovi, Magdaleně Rysové, Magdě Ševčíkové, Ivanu Šmilauerovi a Danielu Zemanovi.

Práce na této Bílé knize byla financována 7. Rámcovým programem Evropské komise a Programem na podporu politiky informačních a komunikačních technologií (ICT Policy Support Programme of the European Commission) na základě smluv T4ME (grantové dohoda 249 119), CESAR (grantová dohoda 271 022), METANET4U (grantová dohoda 270 893) a META-NORD (grantová dohoda 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]. We also wish to thank our colleagues Jan Cuřín, Eva Hajičová, Jirka Hana, Karel Oliva, Magdalena Rysová, Magda Ševčíková, Ivan Šmilauer, Daniel Zeman for their nice cooperation.

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).

OBSAH CONTENTS

ČEŠTINA V DIGITÁLNÍM VĚKU

THE CZECH LANGUAGE IN THE DIGITAL AGE

SHRNUTÍ

Evropa se během posledních 60 let stala zřetelnou politickou a ekonomickou sítí, přesto je ale kulturně a jazykově stále velmi různorodá. Znamená to, že každodenní komunikace mezi evropskými občany (ať už přecházíme z portugalštiny do polštiny nebo z italštiny do islandštiny) i komunikace v oblasti podnikání a politiky se nevyhnutelně potýká s jazykovou bariérou.

Jazykové technologie staví mosty pro budoucnost Evropy.

Orgány EU utratí asi jednu miliardu eur ročně na překládání textů a tlumočení mluvené komunikace, aby řešily otázku mnohojazyčnosti. Musí to však být taková zátěž? Moderní jazykové technologie (language technology, LT) a lingvistický výzkum mohou významně přispět k bourání jazykových hranic. Když se jazykové technologie spojí s inteligentními zařízeními a aplikacemi, budou v budoucnosti schopné pomáhat Evropanům jednoduše komunikovat a obchodovat, i když nemluví společnou řečí. Česká ekonomika má na jednotném evropském trhu velkou výhodu. Přesto je možné, že jazykové bariéry způsobí např. zánik některých podniků, a to zejména jedná-li se o malé a střední podniky, které nemají finanční prostředky na zlepšení situace. Jedinou (i když nemyslitelnou) alternativou řešení otázky mnohojazyčné Evropy by bylo umožnit, aby jeden jazyk získal dominantní postavení a nakonec nahradil všechny ostatní. Bez technologické podpory, je zvládnutí 23 oficiálních jazyků členských států Evropské unie a dalších cca 60 evropských jazyků nepřekonatelná překážka pro občany našeho kontinentu, pro jejich ekonomiku, jejich politickou diskusi a vědecký pokrok. Řešením je vybudování klíčových technologií, které budou nabízet evropským subjektům velké výhody, a to nejen v rámci společného evropského trhu, ale i v obchodních vztazích se třetími zeměmi, zejména v nově se etablujících ekonomikách. Abychom dosáhli tohoto cíle a uchránili evropskou kulturní a jazykovou rozmanitost, musíme nejprve provést systematickou analýzu jazykových aspektů všech evropských jazyků a analýzu současného stavu podpory jazykových technologií. Pak budou moci jazykové technologie sloužit jako jedinečný most mezi evropskými jazyky. Nástroje pro automatický překlad a zpracování řeči, které jsou v současné době dostupné na trhu, ovšem stále ještě tohoto náročného cíle nedosahují. Dominantní subjekty v této oblasti jsou převážně soukromé podniky se sídlem v Severní Americe. Již na konci 70.

Jazykové technologie jako klíč k budoucnosti.

let si EU uvědomila nesmírný význam jazykových technologií jako nástroje k dosažení evropské jednoty a začala financovat první výzkumné projekty, např. EUROTRA. Ve stejné době začaly vznikat vnitrostátní projekty, které sice přinášely cenné výsledky, ale nikdy nevedly k evropské spolupráci. Ostatní mnohojazyčné komunity jako Indie (22 úředních jazyků) a Jihoafrická republika (11 úředních jazyků) naopak na rozdíl od tohoto vysoce selektivního financování nedávno vytvořily dlouhodobé národní programy pro jazykový výzkum a

technologický rozvoj. Dominantní subjekty v oblasti jazykových technologií se dnes spoléhají na nepřesné statistické postupy, které nevyužívají propracované jazykovědné metody a znalosti. Například automatický překlad vět funguje na principu porovnávání věty, kterou chceme automaticky přeložit, s tisíci jinými, které byly přeloženy lidmi. Kvalita výstupu do značné míry závisí na velikosti a kvalitě daného vzorku. Zatímco automatický překlad textu může u „velkých" jazyků s jednoduchou morfologickou strukturou dosáhnout přiměřené kvality, u složitějších jazyků nebo u jazyků s nižším počtem příkladového materiálu je tato statistická metoda odsouzena k neúspěchu. Evropská unie se proto rozhodla financovat projekty, jako je EuroMatrix, EuroMatrixPlus (fungující od roku 2006) a iTranslate4 (fungující od roku 2010), které provádějí základní a aplikovaný výzkum a snaží se vytvořit vysoce kvalitní jazykové technologie pro všechny evropské jazyky. Hlubší analýza struktury jazyků je jedinou možnou cestou, jak vytvářet aplikace, které fungují v rámci celé škály evropských jazyků dobře. Evropský výzkum dosáhl v této oblasti již řady úspěchů. Například překladatelské služby v Evropské unii nyní používají MOSES, open-source software pro strojový překlad, který byl vyvinut zejména prostřednictvím evropských výzkumných projektů. Spíše než stavět na výsledcích těchto projektů má Evropa tendenci pokračovat v izolované výzkumné činnosti jen s nepatrným vlivem na trh. Ekonomickou hodnotu počátečního úsilí lze vidět na počtu prodaných dceřiných společností. Např. společnost Trados (založena v roce 1984) byla v roce 2005 prodána společnosti SDL se sídlem ve Velké Británii.

Jazykové technologie pomáhají sjednotit Evropu.

Na základě dosud získaných poznatků se zdá, že dnešní „hybridní" jazykové technologie zahrnující hloubkové zpracování i statistické metody umožní překlenout propast mezi všemi evropskými jazyky. Jak tato série Bílých knih ukazuje, členské státy v Evropě se značně liší v ochotě a připravenosti řešit jazykové otázky. Velké rozdíly jsou také v oblasti výzkumu. Čeština patří mezi „menší" jazyky EU, a proto je zapotřebí nejprve provádět další specializované výzkumy, než pro ni budou jazykové technologie skutečně účinné a než budou moci sloužit pro každodenní použití. Dlouhodobým cílem projektu META-NET je představit kvalitní jazykové technologie pro všechny jazyky v EU. Tyto technologie pomohou evropským jazykům překonat dosavadní bariéry a navázat vzájemné spojení. To vyžaduje, aby všechny zúčastněné strany – v politice, výzkumu, podnikání i společnosti – spojily v budoucnosti své síly. Tento dokument doplňuje řadu dalších činností projektu META-NET (viz příloha). Aktuální informace, např. aktuální verzi plánů projektu META-NET [2] nebo strategický plán výzkumu (SRA), najdete na webových stránkách http://www.meta-net.eu.