META NET

White Paper Series Seria raportów

THE POLISH LANGUAGE IN THE DIGITAL AGE

JĘZYK POLSKI W ERZE CYFROWEJ

Marcin Miłkowski

Springer

White Paper Series Seria raportów

THE POLISH LANGUAGE IN THE DIGITAL AGE

JĘZYK POLSKI W ERZE CYFROWEJ

Marcin Miłkowski
Instytut Podstaw Informatyki PAN

Georg Rehm, Hans Uszkoreit
(redakcja, 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 978-3-642-30810-9 ISBN 978-3-642-30811-6 (eBook)
DOI 10.1007/978-3-642-30811-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012943360

Printed on acid-free paper

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

WSTĘP PREFACE

Poniższy raport jest częścią serii wydawniczej, której celem jest upowszechnianie wiedzy na temat technologii językowych i ich możliwych zastosowań.

Dostępność i wykorzystanie technologii językowych w Europie są różne w zależności od języka. Dlatego też działania, które należy podjąć, aby odpowiednio wspierać badania i rozwój technologii dla danego języka, są uzależnione od wielu czynników takich jak złożoność określonego systemu językowego i wielkość społeczności posługującej się tym językiem.

Członkowie META-NET, Sieci Doskonałości współfinansowanej przez Komisję Europejską, przeprowadzili analizę bieżącego stanu zasobów i technologii językowych dla 23 europejskich języków urzędowych oraz innych ważnych języków narodowych i regionalnych w Europie (s. 77). Wyniki tej analizy sugerują, że w przypadku każdego języka istnieje wiele istotnych braków. Bardziej szczegółowa, specjalistyczna analiza i ocena bieżącej sytuacji pozwoli na optymalne wykorzystanie dodatkowych badań.

Do sieci META-NET w listopadzie 2011 należały 54 ośrodki badawcze z 33 krajów, współpracujące z podmiotami komercyjnymi, agencjami rządowymi, przedstawicielami przemysłu, organizacjami badawczymi, producentami oprogramowania, dostawcami technologii i uczelniami europejskimi (s. 73). Wszyscy członkowie sieci tworzą wspólną wizję technologii językowych i zajmują się opracowaniem planów strategicznych, których realizacja pozwoli na uzupełnienie wykrytych braków technologicznych do 2020 r.

This white paper is part of a series that promotes knowledge about language technology and its potential. 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. 77). 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. 73). 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.

Autor tego opracowania dziękuje autorom raportu dotyczącego języka niemieckiego za zgodę na wykorzystanie materiałów niezależnych od języka [1].

Przekład na język polski: Anna Cichosz.

Opracowanie niniejszego raportu zostało sfinansowane w ramach siódmego programu ramowego oraz programu na rzecz wspierania polityki w zakresie technologii informacyjnych i komunikacyjnych Komisji Europejskiej w ramach umów T4ME (grant 249 119), CESAR (grant 271 022), META-NET4U (grant 270 893) i META-NORD (grant 270 899).

The author of this document is grateful to the authors of the White Paper on German for permission to re-use selected language-independent materials from their document [1].

Polish translation: Anna Cichosz

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).

SPIS TREŚCI CONTENTS

JĘZYK POLSKI W ERZE CYFROWEJ

THE POLISH LANGUAGE IN THE DIGITAL AGE

STRESZCZENIE

Informatyka zmienia nasze życie codzienne. Do pisania i redagowania tekstów, liczenia i wyszukiwania informacji używamy zwykle komputerów. Coraz bardziej służą nam one także do czytania, słuchania muzyki, przeglądania zdjęć i oglądania filmów. W kieszeniach nosimy małe komputery, za pomocą których prowadzimy rozmowy telefoniczne i piszemy e-maile. Są one źródłem informacji i rozrywki w dowolnym miejscu na świecie. Jak digitalizacja informacji, wiedzy i codziennej komunikacji wpływa na język? Czy nasz język zmieni się lub nawet zaniknie?

Jakie szanse przetrwania ma polszczyzna?

Wszystkie nasze komputery łączą się ze sobą w gęstniejącej sieci globalnej o coraz większych możliwościach. Dziewczyna z Ipanemy, celnik w Dorohusku i inżynier w Katmandu mogą rozmawiać ze znajomymi na Facebooku, ale prawdopodobnie nigdy nie spotykają się w społecznościach internetowych i na forach. Gdy chcą poradzić sobie z bólem ucha, wszyscy zajrzą do Wikipedii. Jednak nawet wtedy nie będą czytać tego samego artykułu. Kiedy na forach i czatach sieciowi obywatele Europy dyskutują na temat wpływu awarii jądrowej w Fukushimie na europejską politykę energetyczną, robią to w odseparowanych od siebie społecznościach językowych. Co łączy Internet, języki użytkowników nadal rozdzielają. Czy zawsze tak będzie?
Wiele spośród 6000 języków na świecie może nie przetrwać w zglobalizowanym cyfrowym społeczeństwie informacyjnym. Szacuje się, że co najmniej 2000 języków jest skazanych na wymarcie w nadchodzących dziesięcioleciach. Inne nadal będą odgrywać pewną rolę w rodzinach i życiu codziennym, ale nie w skali biznesu i środowisk akademickich.

Język polski, którym mówi ponad 40 milionów osób, ma dosyć dobrą pozycję w porównaniu do wielu języków. Istnieje duża liczba polskich kanałów telewizyjnych. Większość zaś filmów zagranicznych wyświetla się w wersjach z lektorem lub napisami w języku polskim. Wszystkie popularne pakiety oprogramowania zostały przetłumaczone na język polski i mimo wszelkich obaw o stopniową anglicyzację wydaje się, że w życiu codziennym Polacy wolą używać własnego języka. Istnieje jednak niebezpieczeństwo jego kompletnego zniknięcia z głównych dziedzin naszego życia. Nie chodzi o naukę, lotnictwo i globalne rynki finansowe, które faktycznie na całym świecie potrzebują *lingua franca*. Mamy na myśli wiele dziedzin życia, które są znacznie ważniejsze dla obywateli niż dla partnerów międzynarodowych – chodzi na przykład o politykę wewnętrzną, procedury administracyjne, prawo, kulturę i zakupy.

Status języka zależy nie tylko od liczby mówiących nim osób czy dostępnych w nim książek, programów komputerowych, filmów i stacji telewizyjnych, ale także od obecności języka w cyfrowej przestrzeni. Tutaj również polszczyzna jest w dosyć dobrej sytuacji. Polska Wikipedia jest jedną z największych na świecie, a domena .pl, mająca ponad 2 miliony zarejestrowanych poddomen, jest jedną z największych na świecie domen krajowych. (W USA bardzo niewiele stron internetowych faktycznie korzysta z domeny .us).

W dziedzinie technologii językowych polszczyzna dysponuje wieloma produktami, technologiami i zasobami. Istnieją aplikacje i narzędzia do syntezy mowy, jej rozpoznawania, korekty pisowni i gramatyki. Istnieje także wiele aplikacji do automatycznego tłumaczenia języka, mimo że często nie dają językowo i idiomatycznie poprawnych tłumaczeń, zwłaszcza gdy język polski jest językiem źródłowym. Wynika to głównie ze specyficznych cech języka polskiego.

Informatyka i komunikacja przygotowują się do kolejnej rewolucji.

Następna generacja techniki, po komputerach osobistych, sieci, miniaturyzacji, multimediach, urządzeniach przenośnych i przetwarzaniu „w chmurze", to oprogramowanie rozumiejące nie tylko wypowiedziane lub zapisane litery i dźwięki, ale całe słowa i zdania, a także znacznie lepiej służące użytkownikom, gdyż mówiące ich językiem i go znające. Prekursorskie są tutaj takie zjawiska, jak bezpłatne usługi internetowe Tłumacz Google, które tłumaczą między 57 językami, superkomputer Watson firmy IBM, który zdołał pokonać amerykańskiego mistrza w teleturnieju „Jeopardy", a także Siri, przenośny asystent firmy Apple, który potrafi reagować na polecenia głosowe i odpowiadać na pytania w języku angielskim, niemieckim, francuskim i japońskim. Ale już nie w języku polskim.

Następna generacja informatyki opanuje ludzki język w takim stopniu, że przy użyciu techniki ludzie będą mogli komunikować się we własnym języku. Urządzenia będą w stanie automatycznie znajdować najważniejsze wiadomości i informacje ze światowych zasobów wiedzy w odpowiedzi na proste w użyciu polecenia głosowe. Technika znająca język będzie w stanie tłumaczyć automatycznie lub pomagać tłumaczom, streszczać rozmowy i dokumenty, a także pomagać w nauce.

Następna generacja technik informatycznych i komunikacyjnych umożliwi robotom przemysłowym i usługowym (obecnie rozwijanym w laboratoriach badawczych) dobrze rozumieć, czego żądają ich użytkownicy, a następnie zdawać sprawę z realizacji tych żądań w języku naturalnym.

Ten poziom działania oznacza wyjście poza zestawy znaków i leksykony, korektory pisowni lub gramatyki oraz zasady wymowy. Technika musi odejść od uproszczonych podejść i zacząć modelowanie języka w sposób kompleksowy, biorąc pod uwagę składnię i semantykę, aby móc rozumieć kierunek pytań – a w ten sposób generować bogate i właściwe odpowiedzi.

Istnieje jednak coraz większa przepaść technologiczna między językiem polskim i angielskim. Europa utraciła kilka bardzo obiecujących innowacji technicznych na rzecz USA, gdzie jest większa ciągłość w strategicznym planowaniu badań i większe wsparcie finansowe dla wprowadzania nowej techniki na rynek. W wyścigu do innowacji technicznych dobry początek i wizjonerska koncepcja mogą zapewnić przewagę nad konkurencją tylko wtedy, jeśli rzeczywiście dotrze się na linię mety. Inaczej liczyć można co najwyżej na honorową wzmiankę w Wikipedii.

Każdy międzynarodowy konkurs technologiczny świadczy o tym, że wyniki automatycznej analizy języka angielskiego są znacznie lepsze niż dla polskiego, mimo że (albo właśnie dlatego), że metody analizy są podobne, jeśli nie identyczne. Odnosi się to do ekstrakcji informacji z tekstów, korekty gramatycznej, tłumaczenia maszynowego i bardzo wielu innych zastosowań.

Wielu badaczy uznaje, że opóźnienia rozwojowe biorą się stąd, iż od pięćdziesięciu lat metody i algorytmy lingwistyki komputerowej oraz badań nad aplikacjami językowymi skupiają się przede wszystkim na języku angielskim. Jednak inni sądzą, że język angielski z natury rzeczy lepiej nadaje się do przetwarzania komputerowego. Przy użyciu istniejących metod języki takie