

THE GREEK LANGUAGE IN
THE DIGITAL AGE

Η ΕΛΛΗΝΙΚΗ
ΓΛΩΣΣΑ ΣΤΗΝ
ΨΗΦΙΑΚΗ
ΕΠΟΧΗ

Maria Gavrilidou
Maria Koutsombogera
Anastasios Patrikakos
Stelios Piperidis



White Paper Series

Σειρά Λευκών Βιβλίων

THE GREEK
LANGUAGE IN
THE DIGITAL
AGE

Η ΕΛΛΗΝΙΚΗ
ΓΛΩΣΣΑ ΣΤΗΝ
ΨΗΦΙΑΚΗ
ΕΠΟΧΗ

Maria Gavrilidou R. C. "Athena", ILSP

Maria Koutsombogera R. C. "Athena", ILSP

Anastasios Patrikakos R. C. "Athena"

Stelios Piperidis R. C. "Athena", ILSP

Georg Rehm, Hans Uszkoreit

(επιμελητές, 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 978-3-642-28935-4 ISBN 978-3-642-28936-1 (eBook)
DOI 10.1007/978-3-642-28936-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012945121

© 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)



ΠΡΟΟΙΜΙΟ

Η παρούσα Λευκή Βίβλος εντάσσεται σε μια σειρά από παρόμοιες ενημερωτικές αναφορές σχετικά με τη γλωσσική τεχνολογία και τις δυνατότητές της. Απευθύνεται σε εκπαιδευτικούς, δημοσιογράφους, πολιτικούς, γλωσσικές κοινότητες και άλλους φορείς. Η διαθεσιμότητα και η χρήση γλωσσικής τεχνολογίας στην Ευρώπη ποικίλλει από γλώσσα σε γλώσσα. Κατά συνέπεια, οι δράσεις που απαιτούνται για την περαιτέρω στήριξη της έρευνας και της ανάπτυξης γλωσσικών τεχνολογιών επίσης διαφέρουν για κάθε γλώσσα. Οι απαιτούμενες δράσεις εξαρτώνται από πολλούς παράγοντες, όπως είναι η πολυπλοκότητα μιας γλώσσας και το μέγεθος της κοινότητάς της.

Το META-NET, ένα Δίκτυο Αριστείας που χρηματοδοτείται από την Ευρωπαϊκή Επιτροπή, διεξήγαγε μια έρευνα των υπάρχοντων γλωσσικών πόρων και τεχνολογιών στη σειρά αυτή των Λευκών Βιβλίων (σελ. 79). Αυτή η έρευνα επικεντρώθηκε στις 23 επίσημες ευρωπαϊκές γλώσσες, καθώς και σε άλλες σημαντικές εθνικές και περιφερειακές γλώσσες στην Ευρώπη. Τα αποτελέσματα αυτής της ανάλυσης δείχνουν ότι υπάρχουν πολλά σημαντικά ερευνητικά κενά σε κάθε γλώσσα. Η λεπτομερής ανάλυση και εκτίμηση της τρέχουσας κατάστασης από εμπειρογνώμονες θα βοηθήσει πιθανές μελλοντικές έρευνες.

Το META-NET απαρτίζεται πλέον από 54 ερευνητικά κέντρα από 33 χώρες (σελ. 75), και συνεργάζεται με φορείς που κυμαίνονται από εμπορικές επιχειρήσεις, δημόσιες υπηρεσίες, τη βιομηχανία, ερευνητικά ιδρύματα, εταιρείες ανάπτυξης λογισμικού, μέχρι εταιρείες παροχής τεχνολογίας και ευρωπαϊκά πανεπιστήμια. Το META-NET, σε συνεργασία με τους φορείς αυτούς, υλοποιεί ένα κοινό όραμα για την τεχνολογία και αναπτύσσει μια στρατηγική ατζέντα για την πολύγλωσση Ευρώπη του 2020.

PREFACE

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. 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. 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. 75). 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.

Οι συντάκτες του κειμένου αυτού θα ήθελαν να ευχαριστήσουν τους συγγραφείς της γερμανικής Λευκής Βίβλου για την άδεια χρήσης επιλεγμένων εισαγωγικών χωρίων από το κείμενό τους [1].

Η κατάρτιση αυτής της Λευκής Βίβλου χρηματοδοτήθηκε από το 7ο Πρόγραμμα Πλαίσιο και το Πρόγραμμα “Υποστήριξη της Πολιτικής για τις ΤΠΕ” της Ευρωπαϊκής Επιτροπής, με τα Έργα T4ME (Αρ. Σύμβασης: 249 119), CESAR (Αρ. Σύμβασης: 271 022), METANET4U (Αρ. Σύμβασης: 270 893) και META-NORD (Αρ. Σύμβασης: 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].

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).



ΠΕΡΙΕΧΟΜΕΝΑ CONTENTS

Η ΕΛΛΗΝΙΚΗ ΓΛΩΣΣΑ ΣΤΗΝ ΨΗΦΙΑΚΗ ΕΠΟΧΗ

1	Περίληψη	1
2	Γλώσσες σε κίνδυνο: μια πρόκληση για τη Γλωσσική Τεχνολογία	3
2.1	Γλωσσικά σύνορα: εμπόδιο στην Ευρωπαϊκή Κοινωνία της Πληροφορίας	4
2.2	Οι γλώσσες μας κινδυνεύουν	4
2.3	Γλωσσική Τεχνολογία: μια βασική τεχνολογία προσβασιμότητας	5
2.4	Ευκαιρίες για τη Γλωσσική Τεχνολογία	6
2.5	Προκλήσεις που αντιμετωπίζει η Γλωσσική Τεχνολογία	7
2.6	Κατάκτηση της γλώσσας από ανθρώπους και μηχανήματα	7
3	Τα Ελληνικά στην ευρωπαϊκή κοινωνία της πληροφορίας	9
3.1	Γενικά δεδομένα	9
3.2	Το ελληνικό αλφάβητο	10
3.3	Ιδιαιτερότητες της ελληνικής γλώσσας	11
3.4	Πρόσφατες εξελίξεις	12
3.5	Γλωσσική πολιτική στην Ελλάδα	13
3.6	Η γλώσσα στην Εκπαίδευση	14
3.7	Η διεθνής διάσταση	15
3.8	Τα Ελληνικά στο Διαδίκτυο	15
4	Η Γλωσσική Τεχνολογία για τα Ελληνικά	17
4.1	Αρχιτεκτονικές Εφαρμογών	17
4.2	Βασικά πεδία εφαρμογών	19
4.3	Άλλα πεδία εφαρμογών	26
4.4	Εκπαιδευτικά Προγράμματα	28
4.5	Εθνικά προγράμματα και πρωτοβουλίες	29
4.6	Ο ιδιωτικός τομέας	29
4.7	Διαθεσιμότητα εργαλείων και πόρων	30
4.8	Διαγλωσσική σύγκριση	32
4.9	Συμπεράσματα	33
5	Σχετικά με το META-NET	37

THE GREEK LANGUAGE IN THE DIGITAL AGE

1	Executive Summary	39
2	Languages at Risk: a Challenge for Language Technology	41
2.1	Language Borders Hold back the European Information Society	42
2.2	Our Languages at Risk	42
2.3	Language Technology is a Key Enabling Technology	42
2.4	Opportunities for Language Technology	43
2.5	Challenges Facing Language Technology	44
2.6	Language Acquisition in Humans and Machines	44
3	The Greek Language in the European Information Society	46
3.1	General Facts	46
3.2	The Greek Alphabet	46
3.3	Particularities of the Greek Language	47
3.4	Recent Developments	49
3.5	Language Policy in Greece	50
3.6	Language in Education	51
3.7	International Aspects	51
3.8	Greek on the Internet	52
4	Language Technology Support for Greek	54
4.1	Application Architectures	54
4.2	Core Application Areas	55
4.3	Other Application Areas	61
4.4	Educational Programmes	63
4.5	National Projects and Initiatives	64
4.6	The Private Sector	64
4.7	Availability of Tools and Resources	65
4.8	Cross-language comparison	66
4.9	Conclusions	67
5	About META-NET	71
A	Παραπομπές – References	73
B	Μέλη του META-NET – META-NET Members	75
C	Σειρά Λευκών Βιβλίων του META-NET – The META-NET White Paper Series	79

ΠΕΡΙΛΗΨΗ

Τα τελευταία 60 χρόνια, παρόλο που η Ευρώπη έχει γίνει μια διακριτή πολιτική και οικονομική οντότητα, εντούτοις παρουσιάζει έντονη πολιτισμική και γλωσσική ποικιλότητα. Αυτό σημαίνει ότι, από τα πορτογαλικά έως τα πολωνικά και από τα ιταλικά έως τα ισλανδικά, η επικοινωνία μεταξύ των ευρωπαίων πολιτών σε καθημερινό αλλά και σε επιχειρηματικό και πολιτικό επίπεδο παρεμποδίζεται αναπόφευκτα από γλωσσικούς φραγμούς. Οι οργανισμοί της Ευρωπαϊκής Ένωσης δαπανούν ετησίως περίπου ένα δισ ευρώ για τη διατήρηση της πολιτικής της πολυγλωσσίας, δηλαδή τη μετάφραση κειμένων και τη διερμηνεία της προφορικής επικοινωνίας. Ωστόσο, γιατί πρέπει αυτό να αποτελεί επιβάρυνση; Η σύγχρονη γλωσσική τεχνολογία και η γλωσσολογική έρευνα μπορούν να συνεισφέρουν σημαντικά στην κατάργηση των γλωσσικών φραγμών. Ο συνδυασμός της γλωσσικής τεχνολογίας με έξυπνες συσκευές και εφαρμογές θα παρέχει μελλοντικά στους Ευρωπαίους τη δυνατότητα συνομιλίας και επιχειρηματικών συναλλαγών ακόμη και αν δεν μιλούν την ίδια γλώσσα.

Η Γλωσσική Τεχνολογία χτίζει γέφυρες για το μέλλον της Ευρώπης.

Οι γλωσσικοί φραγμοί θέτουν εμπόδια στην ανάπτυξη των επιχειρήσεων, κυρίως των μικρομεσαίων, οι οποίες δεν διαθέτουν τα οικονομικά μέσα για να αντιστρέψουν την κατάσταση. Η μόνη (αδιανόητη) εναλλακτική λύση θα ήταν η υιοθέτηση μίας μόνο γλώσσας, η οποία θα είχε κυρίαρχη θέση και τελικά θα αντικαθιστούσε όλες τις άλλες γλώσσες. Χωρίς όμως τεχνολογική υποστήριξη, η αντιμετώπιση των 23 επίσημων γλωσσών των κρατών μελών της Ευρωπαϊκής Ένωσης καθώς και των 60 περίπου άλλων ευρωπαϊ-

κών γλωσσών αποτελεί αξεπέραστο εμπόδιο για τους ευρωπαίους πολίτες καθώς και για την οικονομία, την πολιτική διαβούλευση και την επιστημονική πρόοδο της Ευρώπης. Η λύση εντοπίζεται στην ανάπτυξη βασικών τεχνολογιών, οι οποίες θα προσφέρουν στους ευρωπαϊκούς φορείς σημαντικά πλεονεκτήματα όχι μόνο εντός της ευρωπαϊκής κοινής αγοράς, αλλά και στις εμπορικές σχέσεις με τρίτες χώρες, κυρίως με τις αναδυόμενες οικονομίες. Η επίτευξη του στόχου αυτού και η διατήρηση της πολιτισμικής και γλωσσικής ποικιλότητας της Ευρώπης προϋποθέτουν τη διεξαγωγή συστηματικής ανάλυσης των ιδιαιτεροτήτων όλων των ευρωπαϊκών γλωσσών καθώς και του επιπέδου ανάπτυξης υποστηρικτικής γλωσσικής τεχνολογίας για καθεμιά από αυτές.

Η Γλωσσική Τεχνολογία είναι το κλειδί για το μέλλον.

Τα εργαλεία αυτόματης μετάφρασης και επεξεργασίας φωνής που διατίθενται στο εμπόριο απέχουν ακόμη αρκετά από αυτόν τον φιλόδοξο στόχο. Οι κυρίαρχοι παίκτες στο χώρο αυτό είναι κατεξοχήν ιδιωτικές κερδοσκοπικές εταιρίες με έδρα τη Βόρεια Αμερική. Ήδη από τα τέλη του 1970 η ΕΕ αντιλήφθηκε τη σπουδαιότητα της γλωσσικής τεχνολογίας στην πορεία προς την ευρωπαϊκή ενοποίηση και ξεκίνησε τη χρηματοδότηση των πρώτων της ερευνητικών προγραμμάτων, όπως το EUROTRA. Παράλληλα, συστάθηκαν εθνικά έργα τα οποία, αν και παρήγαγαν σημαντικά αποτελέσματα δεν οδήγησαν ποτέ σε συντονισμένες ευρωπαϊκές ενέργειες. Σε αντίθεση προς αυτές τις προσπάθειες επιλεκτικής χρηματοδότησης, άλλες πολύγλωσσες κοινωνίες όπως η Ινδία (22 επίσημες γλώσσες) και η

Ν. Αφρική (11 επίσημες γλώσσες) έχουν οργανώσει μακροπρόθεσμα εθνικά προγράμματα γλωσσικής έρευνας και τεχνολογικής ανάπτυξης.

Οι κυρίαρχοι παίκτες στο χώρο της γλωσσικής τεχνολογίας σήμερα βασίζονται σε μη ακριβείς στατιστικές προσεγγίσεις οι οποίες δεν αξιοποιούν γλωσσολογικές μεθόδους και γνώση. Για παράδειγμα, οι προτάσεις που μεταφράζονται αυτόματα προκύπτουν από τη σύγκριση μιας νέας πρότασης με χιλιάδες προτάσεις που έχουν προηγουμένως μεταφραστεί από ανθρώπους. Η ποιότητα του αποτελέσματος εξαρτάται σε μεγάλο βαθμό από το μέγεθος και την ποιότητα του διαθέσιμου σώματος κειμένων. Αν και η αυτόματη μετάφραση απλών προτάσεων σε γλώσσες με επαρκή όγκο διαθέσιμων πόρων μπορεί να επιτύχει ικανοποιητικά αποτελέσματα, οι επιφανειακές στατιστικές μέθοδοι τέτοιου τύπου είναι καταδικασμένες να αποτύχουν σε περιπτώσεις γλωσσών με πολύ μικρότερο σώμα δεδομένων ή σε περιπτώσεις προτάσεων με πολύπλοκες δομές.

Η Ευρωπαϊκή Ένωση αποφάσισε να χρηματοδοτήσει έργα όπως το EuroMatrix και το EuroMatrix+ (από το 2006) και το iTranslate4 (από το 2010), στο πλαίσιο των οποίων διεξάγεται βασική και εφαρμοσμένη έρευνα και παράγονται πόροι που εξασφαλίζουν λύσεις γλωσσικής τεχνολογίας υψηλής ποιότητας για όλες τις ευρωπαϊκές γλώσσες. Η ανάλυση των βαθύτερων δομικών ιδιοτήτων των γλωσσών είναι η μόνη διέξοδος, αν ο στόχος είναι η ανάπτυξη εφαρμογών υψηλής απόδοσης για το συνολικό εύρος των ευρωπαϊκών γλωσσών. Η ευρωπαϊκή έρευνα στο χώρο αυτό έχει ήδη σημειώσει αρκετές επιτυχίες. Για παράδειγμα, οι μεταφραστικές υπηρεσίες της Ευρώπης χρησιμοποιούν πλέον το Moses, λογισμικό ανοιχτού κώδικα για την αυτόματη μετάφραση, το οποίο αναπτύχθηκε κυρίως στο πλαίσιο ευρωπαϊκών ερευνητικών προγραμμάτων. Αντί να αξιοποιεί τα αποτελέσματα των ευρωπαϊκών της προγραμμάτων, η Ευρώπη έχει την τάση να επιδιώκει μεμονωμένες ερευνητικές δράσεις με περιορισμένο αντίκτυπο στην αγορά. Η οικο-

νομική αξία ακόμη και των πρώιμων προσπαθειών είναι εμφανής και στην περίπτωση των τεχνολογικών, όπως της εταιρίας Trados, η οποία ιδρύθηκε το 1984 και αγοράστηκε από την βρετανική SDL το 2005.

Η Γλωσσική Τεχνολογία συμβάλλει στην ενοποίηση της Ευρώπης.

Κρίνοντας από τα αποτελέσματα των εξελίξεων στο χώρο, φαίνεται πως η σημερινή “υβριδική” γλωσσική τεχνολογία που συνδυάζει τη γλωσσική επεξεργασία με στατιστικές μεθόδους είναι σε θέση να γεφυρώσει το χάσμα μεταξύ των ευρωπαϊκών γλωσσών. Όπως φαίνεται και από την παρούσα σειρά των Λευκών Βιβλίων, υπάρχουν σημαντικές διαφορές μεταξύ των κρατών μελών της Ευρώπης ως προς την ετοιμότητα της γλωσσικής τεχνολογίας και το επίπεδο της έρευνας. Αν και ο χώρος της γλωσσικής τεχνολογίας στην Ελλάδα έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια, απαιτείται περαιτέρω έρευνα και ανάπτυξη για την επίτευξη πραγματικά αποτελεσματικών λύσεων γλωσσικής τεχνολογίας για καθημερινή χρήση.

Μακροπρόθεσμο στόχο του META-NET αποτελεί η ανάπτυξη γλωσσικής τεχνολογίας υψηλής ποιότητας για όλες τις γλώσσες με στόχο την επίτευξη της πολιτικής και κοινωνικής ενοποίησης μέσω της πολιτισμικής ποικιλότητας. Η τεχνολογία θα βοηθήσει στην κατάργηση των υπαρχόντων φραγμών και θα δημιουργήσει γέφυρες μεταξύ των γλωσσών της Ευρώπης. Ο στόχος αυτός απαιτεί την συνένωση των προσπαθειών όλων των φορέων της πολιτικής, της έρευνας, των επιχειρήσεων και της κοινωνίας.

Η συλλογή των εν λόγω Λευκών Βιβλίων αποτελεί μια από τις στρατηγικές δραστηριότητες που έχει αναλάβει το META-NET. Περαιτέρω ενημέρωση σχετικά με τις εκδόσεις των σχετικών εγγράφων [2], συμπεριλαμβανομένης της Στρατηγικής Ατζέντας για την Έρευνα, βρίσκεται στον ιστότοπο του META-NET: <http://www.meta-net.eu>.