The META-NET Language White Paper Series: Overview and Key Results

Bolette Sandford Pedersen

University of Copenhagen, Denmark

bspedersen@hum.ku.dk

A Strategy for Multilingual Europe
Brussels, Belgium, June 20/21, 2012
Language White Papers

- Key communication instruments to address decision makers and journalists.
- Cover all EU languages (30 volumes).
- Report on the state of a language (general, social, strategic and technological aspects) and the level of support through language technology.
- Inform target group about societal and technological problems and challenges as well as economic opportunities.
- Printed documents published by Springer; will be distributed by META-NET.
- PDF versions available for free.
30 Languages Covered

- Basque
- Bulgarian*
- Catalan
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*
- Galician
- German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*
- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Slovene*
- Spanish*
- Swedish*
- Croatian

* = Official EU language

http://www.meta-net.eu
THE SLOVAK LANGUAGE IN THE DIGITAL AGE

Mária Šimková
Radovan Garabík
Katarína Gajdošová
Michal Lackavík
Slavomír Ondrejovič
Jozef Juhár
Ján Genčí
Karol Furdaň
Helena Tvoríková
Jozef Ivanec

SLOVENSKÝ JAZYK V DIGITÁLNOM VEKU

Mária Šimková
Radovan Garabík
Katarína Gajdošová
Michal Lackavík
Slavomír Ondrejovič
Jozef Juhár
Ján Genčí
Karol Furdaň
Helena Tvoríková
Jozef Ivanec

META–NET

White Paper Series
Série bielych kníh

Springer
SLOVENSKÝ JAZYK V DIGITÁLNOM VEKU

1 Zhrnutie 1

2 Ohrozenie našich jazykov: Výzva pre jazykové technológie 3
   2.1 Jazykové hranice spomáhajú európsku informačnú spoločnosť 4
   2.2 Naše jazyky v ohrození 4
   2.3 Jazykové technológie sú kľúčovými technológiami 5
   2.4 Príležitosti pre jazykové technológie 5
   2.5 Výzvy pre jazykové technológie 6
   2.6 Osvetovanie si jazyka 6

3 Slovenčina v európskej informačnej spoločnosti 8
   3.1 Všeobecné štúdie 8
   3.2 Špecifická slovenčina 11
   3.3 Slovenčina na internete 12
   3.4 Slovenčina ako cudzí jazyk 13
   3.5 Slovenský národný korpus 15

4 Jazykové technológie na podporu slovenčiny 17
   4.1 Architektúra aplikácií 17
   4.2 Základné aplikáčné oblasti 19
   4.3 Ďalšie aplikáčné oblasti 27
   4.4 Jazykové technológie vo vzdelávaní 29
   4.5 Štúdium programov a iniciatív 29
   4.6 Dostupnosť nástrojov a zdrojov 30
   4.7 Pokovanie jazykov 32
   4.8 Závery 33

5 O META-NETe 37

THE SLOVAK 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 Slovak in the European Information Society 46
   3.1 General Facts 46
   3.2 Particularities of the Slovak Language 49
   3.3 Slovak on the Internet 51
   3.4 Slovak as a Foreign Language 51
   3.5 Slovak National Corpus 53

4 Language Technology Support for Slovak 55
   4.1 Application Architectures 55
   4.2 Core Application Areas 57
   4.3 Other Application Areas 65
   4.4 Language Technology in Education 66
   4.5 National Projects and Initiatives 67
   4.6 Availability of Tools and Resources 67
   4.7 Cross-language Comparison 70
   4.8 Conclusions 70

5 About META-NET 74
   A Zoznam literatúry – References 75
   B Členovia META-NETe – META-NET Members 81
   C Série bielych kníh META-NETe – The META-NET White Paper Series 85
A few Numbers …

- >160 national experts contributed as authors or co-authors (ca. 5 per language on average).
- >50 additional experts have contributed data and information.
- >8,000 copies will be printed and distributed by META-NET.
Cross-Lingual Ranking

- Journalists and politicians need simple and clear messages.
- In four application areas, each language is assigned to one of five clusters, ranging from excellent LT support to weak/no support:
  1. Machine Translation
  2. Speech Processing
  3. Text Analysis
  4. Resources
- Results finalised at a meeting in Berlin with representatives of all 30 languages (October 21/22, 2011).
<table>
<thead>
<tr>
<th>Level</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>excellent</td>
<td>English, French, Spanish, English</td>
</tr>
<tr>
<td>good</td>
<td>Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian, Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td>moderate</td>
<td>Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish, Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td>fragmentary</td>
<td>Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian</td>
</tr>
<tr>
<td>weak or no support</td>
<td>Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish, Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish</td>
</tr>
<tr>
<td>excellent</td>
<td>good</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>excellent</th>
<th>good</th>
<th>moderate</th>
<th>fragmentary</th>
<th>weak/no support</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>Basque, Bulgarian, Catalan, Czech, Danish, Estonian, Finnish, Galician, Greek, Hungarian, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene</td>
<td>Icelandic, Irish, Latvian, Lithuanian, Maltese</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish</td>
<td>Basque, Bulgarian, Catalan, Czech, Danish, Estonian, Finnish, Galician, Greek, Hungarian, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene</td>
<td>Icelandic, Irish, Latvian, Lithuanian, Maltese</td>
</tr>
</tbody>
</table>
Europe's Languages and LT

- English
- Dutch, French, German, Italian, Spanish
- Catalan, Dutch, Czech, Finnish, Hungarian, Polish, Portuguese, Swedish
- Basque, Bulgarian, Danish, Galician, Greek, Norwegian, Romanian, Slovak, Slovene
- Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian

*good support through Language Technology*  *weak or no support*
Key Results

- When it comes to Language Technology support, there are massive differences between Europe’s languages and technology areas.
- Language Technology support for English is ahead of any other language – but far from being perfect.
- For 16 of 30 languages, Language Technology support is only *fragmentary, very weak* or *non-existent*!
- We now have a list of gaps and needs for all 30 languages which need to be addressed in the years to come.

Ca. 20 of the 30 Language White Papers available online in their final versions (PDF).

First batch of printed copies (published by Springer) will be available soon.

New volumes upcoming (such as, “European Sign Languages in the Digital Age”).

Dissemination of printed volumes to politicians and journalists to start in July.
Thank you.

office@meta-net.eu

http://www.meta-net.eu
http://www.facebook.com/META.Alliance