The META-NET Whitepaper Series on European Languages

Andrejs Vasiljevs
Tilde, Latvia
andrejs@tilde.com

META-NET TEAM
Outline

- “Yet another study”?
- General Approach
- Preliminary Results
- Conclusions
Why Yet Another Survey?

- META-NET aims at a concerted effort to improve monolingual and multilingual LT support for all European languages.

- The degree to which LT is used in Europe varies from language to language depending on the commercial relevance of the language, the problems the language poses for automatic processing, and the research already devoted to it.

- So far, no one has ever evaluated the state of European languages in regards to LT support or their state in the digital information age.

http://www.meta-net.eu
The Language White Papers

- Survey of the state of the respective language in the digital society.
- Provide expert estimations about the current status and availability of language resources and technologies.
- Are meant to inform politicians, journalists and the public at large about societal and technological problems, challenges, and economic opportunities.

http://www.meta-net.eu
29 Languages Covered so far

- 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*
- Spanish*
- Swedish*

* = Official EU language

http://www.meta-net.eu
Structure of the White Papers

- Executive Summary
- Part 1: Introduction – A Risk for Our Languages and a Challenge for Language Technology
- Part 2: Language in the European Information Society
- Part 3: Language Technology Support for Language
- Part 4: About META-NET
- References

http://www.meta-net.eu
Assessing LT Support

- How to assess Language Technology support for a certain language?
- How to arrive at a result that can be communicated?
  - Count all existing tools and resources? => Does not result in a message.
  - Define quality criteria and perform a comparative evaluation? => Complicated, complex, time-consuming process, would take too long.
- For the White Papers, experts provided estimations condensed in a one table assessing core technology areas and resources such as:
  - Parsing, Information Retrieval, Machine Translation, Speech Recognition, Speech Synthesis, Reference Corpora, Language Models, Thesauri, etc.
- Assessment is done along criteria such as availability, quality, or coverage, maturity, sustainability and adaptability (see the example White Paper in your conference bag).

http://www.meta-net.eu
Preliminary Results
Overall Evaluation

- The tools & resources tables in the current papers allow meaningful interpretation per language (Where are gaps? What is there?).

- Calibration between languages has happened in smaller groups, but still needs to be performed across all languages (after META-FORUM 2011).

- Resultants for each tool and resource have been derived from the two central features quality and coverage, resulting in a big table:
Preliminary Results

- Overall ranking of the quality and coverage of tools and resources is plausible (see the top and bottom tools and technologies on the right).

- Scale:
  0 = non-existent
  6 = perfect

<table>
<thead>
<tr>
<th>Rank</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>Tokenization, Morphology</td>
</tr>
<tr>
<td>3.3</td>
<td>Speech Synthesis</td>
</tr>
<tr>
<td>2.7</td>
<td>Parsing</td>
</tr>
<tr>
<td>2.4</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>2.4</td>
<td>Speech Recognition</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Text Semantics</td>
</tr>
<tr>
<td>0.9</td>
<td>Language Generation</td>
</tr>
<tr>
<td>0.8</td>
<td>Advanced Discourse Processing</td>
</tr>
</tbody>
</table>

Scale:
0 = non-existent
6 = perfect

http://www.meta-net.eu
Preliminary Results

If one takes a value of at least 4 as threshold for practical usability of a technology/resource, then, e.g.,:

- 13 of the 30 languages lack sufficient support in speech synthesis
- 18 of the 30 languages lack sufficient support in parsing
- 26 of the 30 languages lack sufficient support in machine translation
### Example: MT Results

- For an example ranking and comparison,
  - a few values were manually adjusted and
  - normalization was carried out for each language, all technologies: \((x-m)/d\) with 
    - \(m\) mean and \(d\) standard deviation.

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Higher ranks</strong></td>
<td><strong>Medium ranks</strong></td>
<td><strong>Lower ranks</strong></td>
</tr>
<tr>
<td>Spanish, English, Latvian, Lithuanian, Maltese, Galician, Slovene, Hungarian, Catalan</td>
<td>Polish, Icelandic, French, Irish, Basque, Italian, Romanian, Bulgarian, Dutch, Portuguese, Estonian, Finnish</td>
<td>Croatian, German, Swedish, Czech, Norwegian, Danish, Serbian, Greek</td>
</tr>
</tbody>
</table>

*Includes romanian languages with active research.*

*Difficult languages and languages with little research.*

http://www.meta-net.eu
Example: Comparing the Situation for Parsing

<table>
<thead>
<tr>
<th>Cluster1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech, Dutch, English, German</td>
<td>Bulgarian, French, Norwegian, Polish, Portuguese, Spanish, Swedish</td>
<td>Basque, Catalan, Danish, Finnish, Galician, Hungarian, Irish, Italian, Romanian</td>
<td>Croatian, Estonian, Greek, Icelandic, Latvian, Lithuanian, Maltese, Serbian, Slovak, Slovene</td>
</tr>
</tbody>
</table>

For this example ranking, a few values were manually adjusted.
Example: Findings from the Nordic and Baltic White Papers
Example: Findings from the Nordic and Baltic White Papers

http://www.meta-net.eu
General Conclusions

- Speech processing and synthesis appears to be more mature than processing of written text. Advanced information access technology is in its infancy.

- Research was successful in designing particular high quality results in some areas, but many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programmes and initiatives are needed to standardize data and interchange formats.

- Most (very) large companies have stopped working in the area, leaving the field to SMEs, which can hardly attack an international market.
Thank you.

office@meta-net.eu

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