Polish LRTs: CESAR's story

Maciej Ogrodniczuk
Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
Maciej.Ogrodniczuk@ipipan.waw.pl
Adam.Przepiorkowski@ipipan.waw.pl

META-FORUM 2011: Solutions for Multilingual Europe
Budapest, Hungary, June 27/28
Research capabilities: 
bird's eye view

Most active NLP research centers in Poland:

- **Warsaw**: Institute of Computer Science, Polish Academy of Sciences:
  - morphosyntactic analysers
  - taggers
  - shallow parser and shallow Polish grammar
  - DCG-like efficient parser and Polish DCG grammar
  - syntactic valence dictionary of Polish
  - the IPI PAN Corpus of Polish
  - the National Corpus of Polish (with partners)
  - various corpus-related tools

http://www.meta-net.eu
Research capabilities: bird's eye view

Most active NLP research centers in Poland:

- **Warsaw**: Institute of Computer Science, Polish Academy of Sciences:
  - *Polish treebank*
  - *new taggers*
  - *wide coverage version of the DCG-like grammar with LFG extensions*
  - *more efficient shallow parser and larger shallow grammar of Polish*
  - *syntactico-semantic valence dictionary of Polish*
  - *WSD systems for Polish*
  - *NER systems for Polish*
  - *coreference resolution*

http://www.meta-net.eu
Research capabilities: bird's eye view

Most active NLP research centers in Poland:

- University of Łódź:
  - Polish-English parallel and comparable corpora (including web and spoken data components of National Corpus of Polish)
  - Information Retrieval solutions (web-data crawling, categorization and clustering)
  - spoken multimedia corpora of Polish (with time-alignement and discourse annotation)
  - social media text data analysis solutions
  - corpus-driven relational lexicons
  - graph-based language data visualisation
  - Information Extraction (semantic annotation for IE enhanced IR)

http://www.meta-net.eu
Research capabilities: bird's eye view

Most active NLP research centers in Poland:

- **Poznań** (Adam Mickiewicz University and Poznań University of Technology):
  - morphosyntactic dictionary
  - syntactic and semantic parsing
  - lexical semantics
  - speech analysis and synthesis
  - speaker recognition
  - commercial machine translation systems

- **Wrocław** University of Technology:
  - Polish WordNet
  - computational lexical semantics
  - tagging

http://www.meta-net.eu
Research capabilities: bird's eye view

Most active NLP research centers in Poland:

- **Kraków** (AGH University of Science and Technology and Jagiellonian University):
  - morphosyntactic dictionary
  - information extraction

- **Gdańsk** University of Technology:
  - finite state technologies
  - efficient shallow parsing

- University of **Warsaw**:
  - bilingual corpora
  - OCR of historical texts

- **Polish-Japanese Institute of Information Technology (Warsaw)**:
  - speech analysis and synthesis

http://www.meta-net.eu
Industry and open source

Commercial scene:
- bilingual dictionaries, spell- and grammar checkers: TiP
- search technologies: Netsprint, Szukacz, Carrot Search
- speech technologies: Ivona, PrimeSpeech, Skrybot
- machine translation: Poleng, Studio Gambit, Cafetran
- information retrieval: Institute of Media Monitoring
- semantic tools: Knowledge Hives

Community activities:
- Polish Wikipedia (810K articles)
- sjp.pl
- LanguageTool

http://www.meta-net.eu
Cross-border collaboration

Current projects:

- **ATLAS** – Applied Technology for Language-Aided CMS (ICT-PSP)
- **CESAR** – CEntral and South-east europeAn Resources (ICT-PSP)
  part of META-NET
- **CLARIN** – Common Language Resources and Technology Infrastructure (ESFRI infrastructure)
- **FLaReNet** – Fostering Language Resources Network (TN)

A few past projects:

- **LUNA** – Spoken Language Understanding in Multilingual Communication Systems (IST STREP)
- **LT4eL** – Language Technology for eLearning (IST STREP)
- A Treebank/Test-Suite of Polish Utterances (EU CRIT-2)

http://www.meta-net.eu
# Current status of LRTs

<table>
<thead>
<tr>
<th>Technology</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization, Morphology</td>
<td>5</td>
</tr>
<tr>
<td>Parsing</td>
<td>4</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>4</td>
</tr>
<tr>
<td>Speech Synthesis</td>
<td>4</td>
</tr>
<tr>
<td>Text Semantics</td>
<td>1</td>
</tr>
<tr>
<td>Advanced Discourse Processing</td>
<td>1</td>
</tr>
<tr>
<td>Language Generation</td>
<td>1</td>
</tr>
<tr>
<td>Summarization, QA</td>
<td>1</td>
</tr>
<tr>
<td>Dialogue Management</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resources</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Corpora</td>
<td>4</td>
</tr>
<tr>
<td>Syntax-Corpora</td>
<td>4</td>
</tr>
<tr>
<td>Parallel Corpora, TM</td>
<td>4</td>
</tr>
<tr>
<td>Lexicons, Terminologies</td>
<td>4</td>
</tr>
<tr>
<td>Thesauri, WordNets</td>
<td>4</td>
</tr>
<tr>
<td>Discourse-Corpora</td>
<td>1</td>
</tr>
<tr>
<td>Multimedia/multimodal data</td>
<td>1</td>
</tr>
<tr>
<td>Language Models</td>
<td>1</td>
</tr>
</tbody>
</table>
Current status of LRTs

<table>
<thead>
<tr>
<th></th>
<th>Quantity</th>
<th>Availability</th>
<th>Quality</th>
<th>Coverage</th>
<th>Maturity</th>
<th>Sustainability</th>
<th>Adaptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Resources</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3,5</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Recent advances

Impact of CESAR/META-NET:

- increasing **awareness**: [http://clip.ipipan.waw.pl](http://clip.ipipan.waw.pl) with information about Polish LRTs, key players, projects, courses

- increasing **coverage and scope of operation**:
  - development of underrepresented resources
  - cross-lingual usage of Polish LRTs

- increasing **availability, reusability and sustainability**:
  - liberating resources (SGJP, Morfologik)
  - clear licensing policies with open-source preference

http://www.meta-net.eu
Recent advances

Impact of CESAR/META-NET:

- increasing **quality**:
  - building specialized tools for quality improvement
  - larger scope of manual annotation
  - automata for error verification

- increasing **interoperability**:
  - adherence to standards
  - following FLaReNet, CLARIN and META-NET recommendations

http://www.meta-net.eu
To conclude...

- Various Polish resources and tools available or under intensive construction
- Many important research centres willing to co-operate and interested in applying expertise, tools and resources at the European level
- Active business and community players
- Expertise in the processing of Polish and, by extension, other “free word-order” morphologically rich languages
- Still, probably not enough participation in European projects

Visit [http://clip.ipipan.waw.pl](http://clip.ipipan.waw.pl)!
Thank you!

Computational Linguistics in Poland

Here you will find information on

- tools and resources,
- research centers,
- projects,
- courses

related to natural language processing of Polish.

Announcements about current events appear on the Polish Linguistic List.

http://www.meta-net.eu