

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



available
now

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*

*under
construction;
available
in 1-2 years*

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-alignment 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)*

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

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

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

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)

Current status of LRTs

Technology	Median
Tokenization, Morphology	5
Parsing	4
Information Retrieval	4
Speech Synthesis	4
...	
Text Semantics	1
Advanced Discourse Processing	1
Language Generation	1
Summarization, QA	1
Dialogue Management	1

Resources	Median
Reference Corpora	4
Syntax-Corpora	4
Parallel Corpora, TM	4
Lexicons, Terminologies	4
Thesauri, WordNets	4
...	
Discourse-Corpora	1
Multimedia/multimodal data	1
Language Models	1

Current status of LRTs

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Technology	1	2	3	2	1	2	3
Resources	2	1	4	3,5	2	2	2
	2	2	3	3	2	2	2

Recent advances

Impact of CESAR/META-NET:

- ❑ increasing **awareness**: <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

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

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>!

Thank you!

The screenshot shows a browser window with the address bar containing `http://clip.ipipan.waw.pl`. The page title is "Home - CLIP". The main content area features the "CLIP" logo on the left and a search bar on the right with "Titles" and "Text" buttons. Below the search bar is a "Login" link. A dark blue navigation bar contains "Home" on the left and "History" and "Actions" on the right. The main content area has a "Menu" sidebar on the left with links to "Tools and resources", "Research centers", "Projects", and "Courses". The "Wiki" section includes "Find page", "Recent changes", and "Help". The main heading is "Computational Linguistics in Poland". Below it, the text reads: "Here you will find information on" followed by a bulleted list: "tools and resources", "research centers", "projects", and "courses". The text continues: "related to natural language processing of Polish." and "Announcements about current events appear on the [Polish Linguistic List](#)."