CLARIN-PL – a Polish Language Technology Infrastructure for the Users

Maciej Piasecki
Wrocław University of Technology
G4.19 Research Group
maciej.piasecki@pwr.wroc.pl
2014-11-22
Users make problems

Users make all software systems imperfect.

However if a software system is not used, it does not exist.

Who can use language technology?
Basic Notions

- **Language Technology (LT)**
  - language resources and tools
  - **robust** in terms of quality and coverage
  - **multipurpose**
  - **component based**

- **Language Technology Infrastructure**
  - a software framework (architecture or platform)
  - for combining language tools with language resources into **processing chains** (or pipelines)
  - the defined processing chains are next applied to language data sources
  - **interoperability**, also with the external systems
LT in Humanities and Social Sciences: Barriers

- **Physical** – language tools and resources are not accessible in Internet
- **Informational** – descriptions are not available or there is no means for searching
- **Technological** – lack of commonly accepted standards for LT, lack of a common platform, varieties of technological solutions, insufficient users’ computers
- Related to **knowledge** – the use of LT requires programming skills or knowledge from the area of natural language engineering
- **Legal** – licences for language resources and tools (LRTs) limit their applications
LTI for H&SS: Lowering Barriers

- **CLARIN**
  - ERIC consortium of several countries
  - member countries contribute parts of the LTI

- **CLARIN Mission**
  - Lowering the barriers for LT in Humanities & Social Sciences (H&SS)
  - integration of different LT components into one interoperable system
  - one sign on and one login into the distributed infrastructure
  - common standards
  - common licences and promotion of the open access
  - installation-free, web-based user interface
Different ways to LTI

- Bottom-up
  - a collected offer approach
  - based on linking together the already existing Language Resources and Tools
  - focused on accessibility, technical interoperability and processing chains

- Top-down
  - based on user-centred design paradigm
  - research applications for H&SS are a starting point

- Bi-directional
  - linking of Language Resources and Tools
  - combined with the development of research applications
Bi-directional LTI development

- **Idea**
  - development of the necessary elements
    - a distributed network infrastructure
    - basic LT processing chain
  - combined with user-centred approach based on the development of research applications

- **Characteristic features**
  - a metaphor of the Agile-like light weight software designing method
  - close co-operation with key users from the H&SS domain
  - application development stimulates the construction of technical fundamentals
  - inspirations and identification of the further user needs
Polish scientific consortium

- Wrocław University of Technology, G4.19 Research Group
- Institute of Computer Science, Polish Academy of Science
- Polish-Japanese Institute of Information Technology, Chair of Multimedia
- University of Łódź, PELCRA group at Chair of English Language and Applied Linguistics
- Institute of Slavic Studies, Polish Academy of Science
- Wrocław University

Goal: implementation of the Polish part of the CLARIN ERIC LTI

Generously financed by the Polish Ministry of Science and Higher Education (about 4 millions Euro for three years)

An example of the bi-directional approach
CLARIN-PL structure

- Context
  - many basic LRTs for Polish were still lacking at the start of CLARIN-PL
  - Deeper technological barrier
    - e.g. the lack of a robust dependency parser for Polish

- Pillars:
  - CLARIN-PL Language Technology Centre
    - www.clarin-pl.eu
    - the Polish node of the CLARIN distributed infrastructure
  - Complete set of basic LRTs for Polish
  - Research applications for H&SS – first created for key users and selected H&SS sub-domains.
CLARIN-PL Language Technology Centre: bottom-up

- B-type centre, located in Wrocław University of Technology
  - based on modified D-Space system from Lindat (Czech CLARIN)
- Distributed authorisation
  - linked to the national identity federation
  - one sign-on, one login
- Proper repository system supporting
  - persistent identifiers for resources and tools,
  - CMDI meta-data format
- Interface for Federated Content Search
  - On meta-data and content of corpora
- Depositing service for researchers from H&SS
  - focused on LRTs
  - adherence to all CLARIN specifications about standards and protocols
- Web Services for LRTs:
  - the basic processing chain of Polish
  - Flexible composition of the specialised processing chains
  - SOAP & REST interfaces
- An active K-type centre in several areas
### NLP Web Services for Polish

#### Home | Morpho-WS | Tagger-WS | Chunker-WS | ChunkRel-WS | NER-WS | Serel-WS
---|---|---|---|---|---|---
| **Plain text** | W XIX wieku na wyspie mieszkał sultan Abdullah z Peraku, zesłany tu przez Brytyjczyków. | **Morpho-WS** | Performs tokenisation and morphological analysis of Polish text using Morfeusz SGJP and MACA toolkit. For full tagging (no ambiguity) please use Tagger-WS. | **Tagger-WS** | Performs tokenisation and morphosyntactic tagging of Polish text using WCRFT tagger. | **Chunker-WS** | Recognises boundaries of noun phrases, verb phrases and adjective phrases in Polish text. | **ChunkRel-WS** | Prototype service for recognition of inter-chunk syntactic relations (object, subject, copula). | **NER-WS** | Recognises 56 fine-grained categories of proper names in Polish texts. | **http://nlp.pwr.edu.pl/synat** |

LRTs and LRT chains can be useful …
  - if the required tools and resources exist,
  - and, they are robust!

What is the minimal set of LRTs?
What kind of LRTs can be called robust?
  - automated applications in H&SS seem to require high quality of language tools and mostly large coverage of resource

BLARK – The Basic Language Resource Kit
  - “the minimal set of language resources that is necessary to do any precompetitive research and education at all” (Krauwer, 2003) and also basic processing chains
  - possible reference point to compare LRTs for different languages
CLARIN-PL: language resources

- Good starting point, e.g.
  - a huge National Corpus of Polish (1 billion tokens)
  - plWordNet 2.0 – a very large wordnet for Polish
  - Korpus Politechniki Wrocławskiej – an open Polish corpus with rich annotation

- Main goals
  - completing the construction of selected resources
  - building bi-lingual resources and specialised corpora facilitating the envisaged needs of H&SS

- Bilingual resources crucial for interoperability
  - Large number of language pairs vs limited funds
  - Priority given to Polish-English resources
CLARIN-PL: selected resources in development

- plWordNet 3.0
  - a comprehensive description of the Polish lexico-semantic system (~200,000 lemmas, ~280,000 senses)
  - mapping to enWordNet – an expanded Princeton WordNet 3.1
- A large lexicon of the Multi-word Expressions
  - described with the minimal constraints on their lexico-syntactic structures
  - linked to plWordNet
- NELexicon 2.0 - ~2.5 million distinct PNs, semantically classified
- Dynamic lexicons – tools for automated expansion of the manual core
- A large semantic valency lexicon for Polish predicative lexical units
- Corpora:
  - a transcribed training-testing Polish speech corpus, conversational corpus
  - parallel corpora, historical Polish corpus of text news…
- Several systems for searching text and speech corpora
1. Segmentation into tokens and sentences
2. Morphological analysis
3. Morphological guessing of unknown words (both without context and context sensitive)
4. Morpho-syntactic tagging
5. Word Sense Disambiguation
6. Chunker and shallow syntactic parser
7. Named Entity Recognition and disambiguation
8. Co-reference and anaphora resolution
9. Temporal expression recognition
10. Semantic relation recognition
11. Event recognition
12. Shallow semantic parser
13. Deep syntactic parser with disambiguated output: dependency and constituent
14. Deep semantic parser
CLARIN-PL: language tools

- A generic set of morpho-syntactic tools for Polish that can be adapted to a domain specified by the user
- Tools for the extraction of the semantic-pragmatic information from documents and collections of documents, e.g.
  - keywords,
  - semantic relations between text fragments
  - and text summaries
- Web Services will be provided for all LRTs and systems
  - already implemented: segmentation, morphological analysis, tagging, chunking, Named Entity Recognition, and WSD
  - accessible via REST or SOAP and described by CMDI
Web Service test for Named Entity Recognition

Tools:

Submit new text:

Input text: Nie żyje Thomas Duncan - pierwszy pacjent, u którego w USA wykryto zakażenie wirusem ebola. Choroba u pochodzącego z Liberii Thomasa Duncana rozwinięła się już po jego przybyciu do Ameryki.

Upload file: Nie wybrano pliku.

Input format: text

Output format: ccl

Submit

The text was submitted. Use the form below to check the status or get the result.

Check status/Show result:

Request ID: b1803aaa0ca4e6f71d6220a37299eac5826e2438fd2f611bb1885fb1a1223

Check status     Show result     Download

Status/result:

Move the mouse over a word to see its morphological analysis.

Nie żyje Thomas Duncan - pierwszy pacjent, u którego w USA wykryto zakażenie wirusem ebola. Choroba u pochodzącego z Liberii Thomasa Duncana rozwinięła się już po jego przybyciu do Ameryki.
Bi-directional - top-down part: selection of applications

- **Criteria**
  - to cover a maximal variety of research areas
  - but also to co-operate first with the most active users
  - matching the available LT for Polish
  - a few application but broadening our understanding of the domain

- **First applications**
  - **Spokes** – a search system for the corpus of conversational data (users from inside of CLARIN-PL)
  - A system for collecting Polish text corpora from the Web
  - A open textometric and stylometric system focused on Polish
  - Semantic text classification for sociology
  - **Literary Map**
System for collecting Polish text corpora from the Web

- Requests from users revealed gaps in the available technology
- Existing corpus building systems were too sensitive to text encoding errors found in the web
- A system for collecting Polish text corpora from the Web had to be constructed:
  - based on solutions developed in Masaryk University in Brno
  - applies morphological analysis to detect texts including larger number of errors
  - Supports semi-automated extraction of texts from blogs
Open textometric and stylometric system

- Several textometric and stylometric tools available
- But not designed for languages of rich inflection like Polish
- Enabling the use of features defined on any level of the linguistic structure:
  - from the level of word forms
  - up to the level of the semantic-pragmatic structures.
- Re-use of several existing components, e.g. Stylo
- Available as Web Application and a Web Service
- Stylometric techniques appear to be applicable in many tasks of H&SS
  - sociology (characteristic features that are for different subgroups), political studies (similarity and differences between political parties), literary studies …
Semantic text classification for sociology

- Users: Collegium Civitas, Warsaw
- Initially:
  - Text document classification according to the manually annotated examples
- Finally:
  - Whole system from corpus gathering to tuning machine learning methods for the semantic classification of text snippets
Semantic text classification for sociology

1. Corpus building
2. Pre-processing
   - Text segmentation utilising the original structure
   - Morpho-syntactic tagging, parsing
3. Automated sample selection
   - Collection distribution
   - Clustering – different techniques
4. Manual annotation
   - Abstract definitions of semantic classes
   - Availability of open annotation editors
5. Training classifiers
6. Analysis of the results
   - Error estimation
GeTClasS – Generalised Text Classification for Sociology

Visualisation of the text classifier generated by Generalised Text Classification for Sociology.

**Wygeneruj nowe zapytanie**

**Tekst do analizy:** Cytryckim o hipotetycznym wsparciu przez NBP budżetu państwa kilka miesięcy przed wyborami, które może wygrać PIS. Belka w zamian za wsparcie stawia warunek dymisji ówczesnego ministra finansów Jacka Rostowskiego oraz nowelizacji ustawy o banku centralnym. Tygodnik twierdzi, że do rozmowy doszło w lipcu 2013 roku.

**Model:** Model socjologiczny

Analizuj

Tekst został zatwierdzony. Użyj poniższego formularza, aby sprawdzić wynik.

**Rezultat**

**Kod zapytania:** 95d35db06babb0592cb07982b55c975201c58ca7

Pokaż rezultat

Pobierz plik wyników

**Rezultat**

<table>
<thead>
<tr>
<th>CECHY</th>
<th>DEFINICJE</th>
</tr>
</thead>
<tbody>
<tr>
<td>obsz_edu</td>
<td>edukacja</td>
</tr>
<tr>
<td>obsz_rel</td>
<td>religia</td>
</tr>
<tr>
<td>obsz_spol</td>
<td>społeczeństwo</td>
</tr>
<tr>
<td>obsz_szt</td>
<td>sztuka</td>
</tr>
</tbody>
</table>
Literary Map

- **Users**
  Digital Humanities Centre of The Institute of Literary Research PAS)

- **Idea**
  to identify all geographical names in the literary text (or a corpus) and map them onto the geographical map

- **Technical requirements**
  - Named Entity Recognition combined with geo-location
  - PNs recognised in text must be grouped into expression recognised by Google
  - Recognition of semantic relations between non-spational PNs and locations

- Parallel research on the method and its applications
Literary Map

Tekst do analizy


Rozpoznane obiekty

Dawid Podsiadło
Polska
Puławy
Factora
Chełm
Lublin
Kraków
Conclusions

- Application of LT to the research in Humanities & Social Sciences seem to be much more challenging than in commercial systems!
- LT for Polish achieved a stage in which valuable support can be provided for research applications
- Bi-directional approach combines
  - development of the basic, universal set of language tools and resources
  - with inspirations from the research applications
- Error monitoring and management in LT-based applications is required
Thank you very much for your attention!

www.clarin-pl.eu

Supported by the Polish Ministry of Science and Higher Education [CLARIN-PL] and the EU’s 7FP under grant agreement no 316097 [ENGINE]