Towards the Annotation of Named Entities in the National Corpus of Polish

Agata Savary[*†]
Jakub Waszczuk[◇†]
Adam Przepiórkowski[†◇]

[*]Université François Rabelais Tours
[†]Institute of Computer Science, Polish Academy of Sciences
[◇]University of Warsaw, Poland

LREC'10, May 19-21, 2010

National Corpus of Polish (NKJP)

The project

- consortium: creators of big annotated corpora of Polish
- financed by the Polish Ministry of Science and Higher Education
- period: 2007-2010

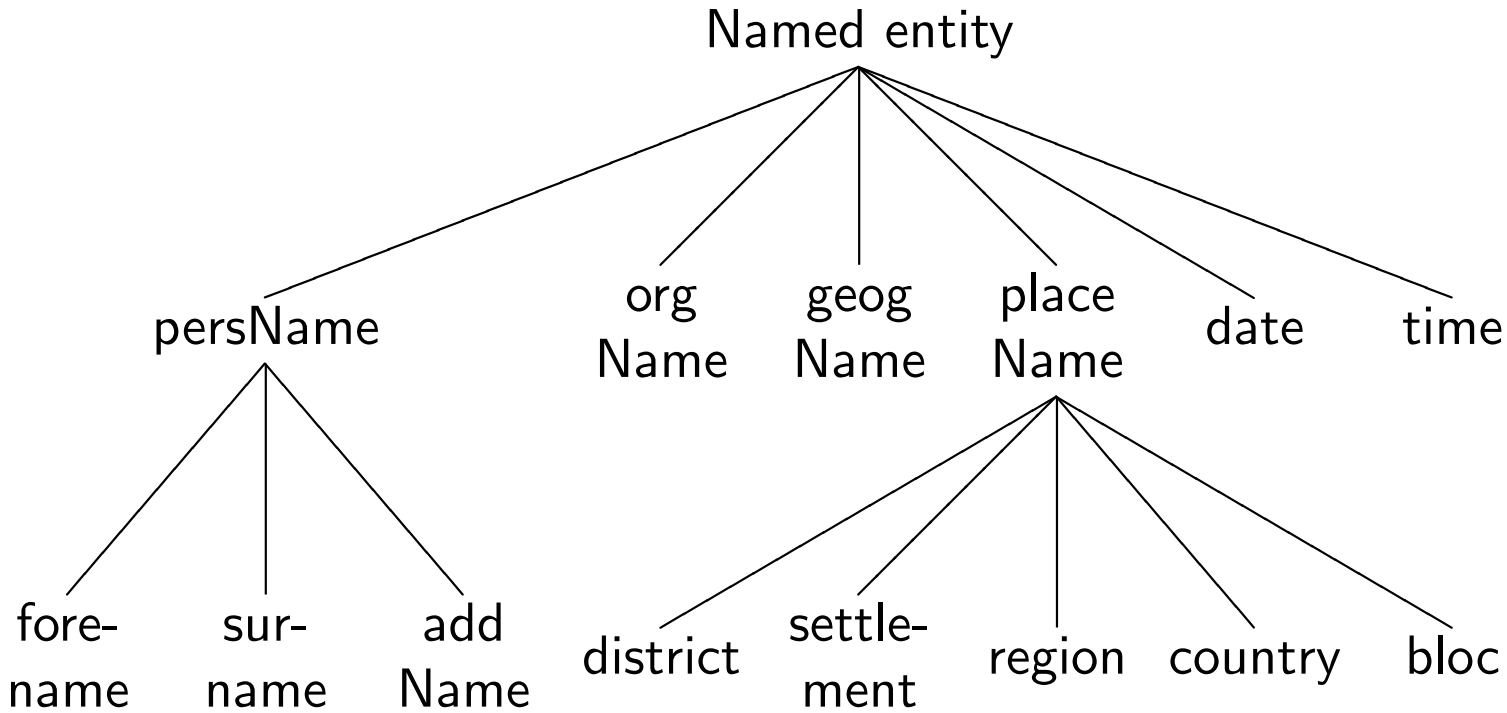
The aim: large national corpus of Polish

- 1 million words manually annotated, 1 billion words automatically annotated *(Przepiórkowski et al. LREC'2010)*
- representative
- balanced wrt. different genres *(Przepiórkowski et al. 2009)*
- associated to linguistic annotation tools

Annotation in NKJP

- stand-off
- TEI P5-conformant *(Przepiórkowski & Bański 2009)*
- multi-level
  * segmentation
  * morphosyntax *(Przepiórkowski & Murzynowski 2009)*
  * syntactic words (e.g. bał się)
  * syntactic groups *(Głowińska and Przepiórkowski 2010)*
  * **named entities**
  * word senses
- quality-ensured (double annotation + super-annotation)
- see other presentations in LREC'10: **W4** i **W20**

TEI P5-inspired hierarchy



- vertical hierarchy of **related names**
  - relational adjectives *poznański, ONZ-owski*
  - names of inhabitants and members *poznaniak, Grek, AK-owiec*
- not annotated: quantities, products, periods, events, titles, . . .

Annotation strategies

- Gramatically motivated lemma *(Piskorski et al. 2009)*

  Stanów Zjednoczonych → *Stany Zjednoczone*
- Sematically motivated derivation base

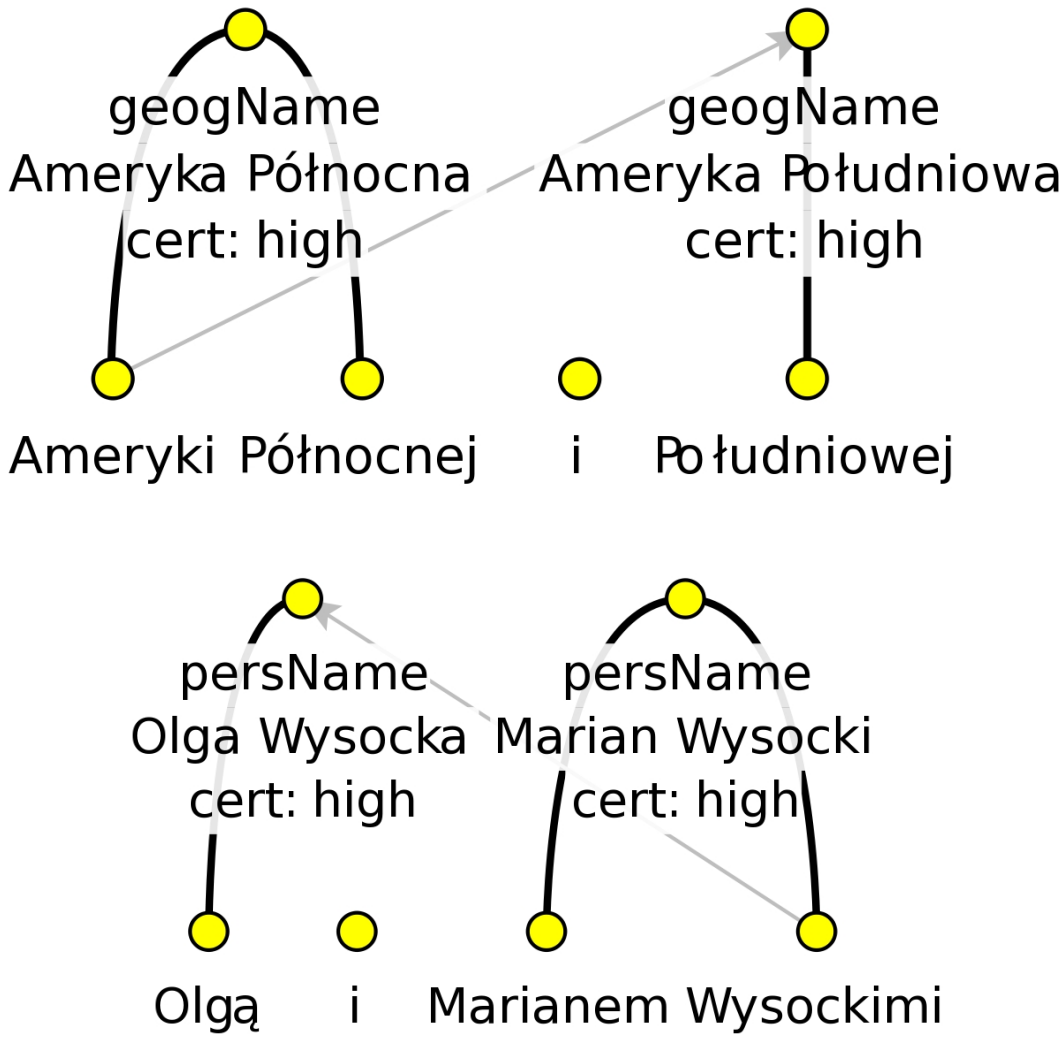
  amerykański → *Stany Zjednoczone*
- Embedded names annotated *(Galicia-Haro and Gelbukh 2009; Finkel and Manning 2009; Kravalová and Žabokrtský 2009)*

  $[[\textit{Tadeusz}]_{\text{forename}}[\textit{Kościuszko}]_{\text{surname}}]_{\text{persName}}$
- Discontinuous names

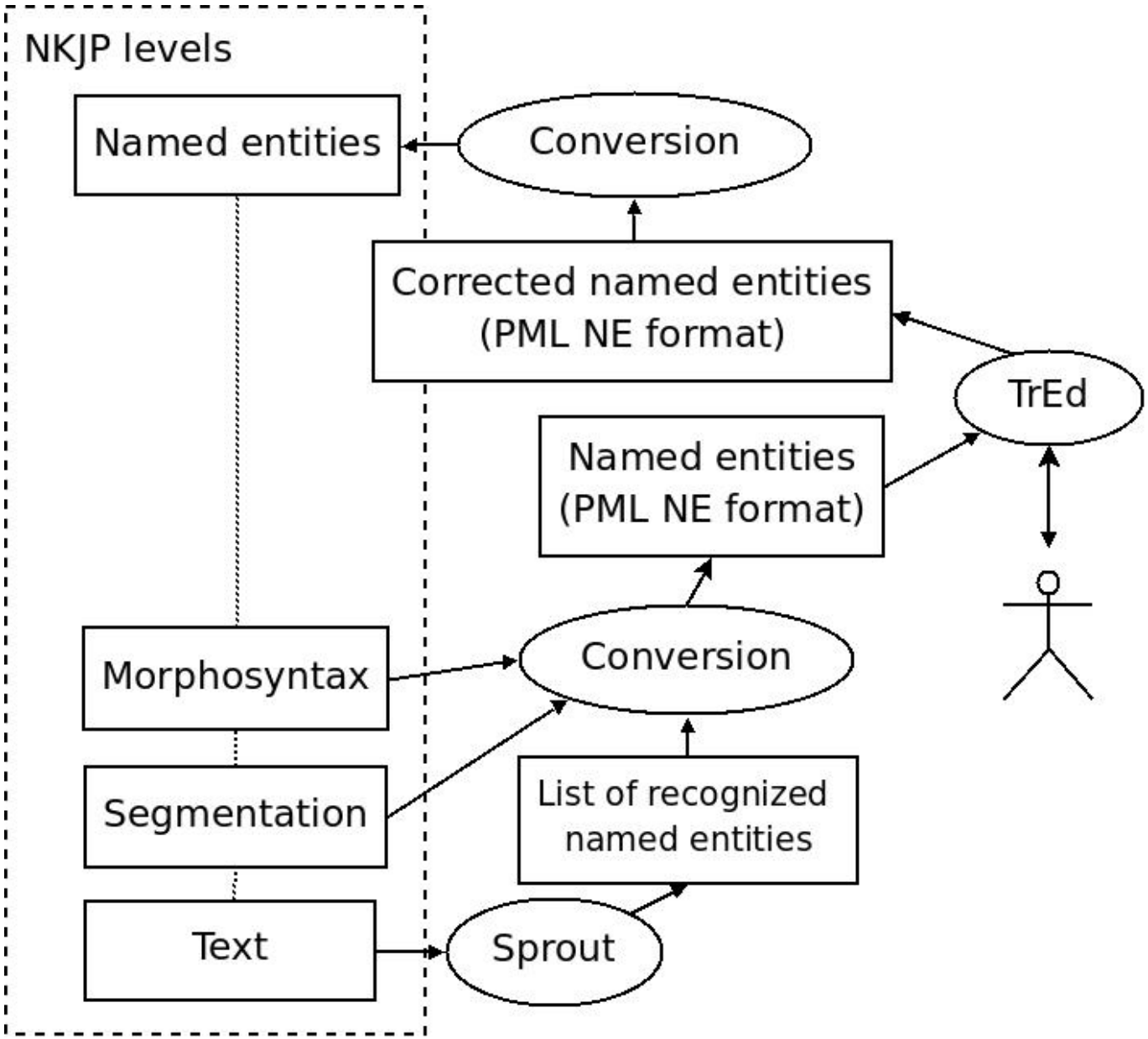
  Wydział Matematyczny ówczesnej Akademii Krakowskiej
- Coordinations separated *(Mazur and Dale 2009)*

  Ameryka Północna i Południowa

Coordinated names



Data Flow


NKJP levels
Named entities
Conversion
Corrected named entities
(PML NE format)
TrEd
Named entities
(PML NE format)
Morphosyntax
Conversion
Segmentation
List of recognized
named entities
Text
Sprout

Sprout – knowledge-based NE extraction

NLP platform

- fast gazetteer lookup *(Budiscak et al. 2009)*
- cascaded unification-based FST grammar parser
- output: feature-structures with user-defined types
- previous NE grammar for Polish *(Piskorski 2005)*

NKJP resources and rules *(Savary & Piskorski 2010)*

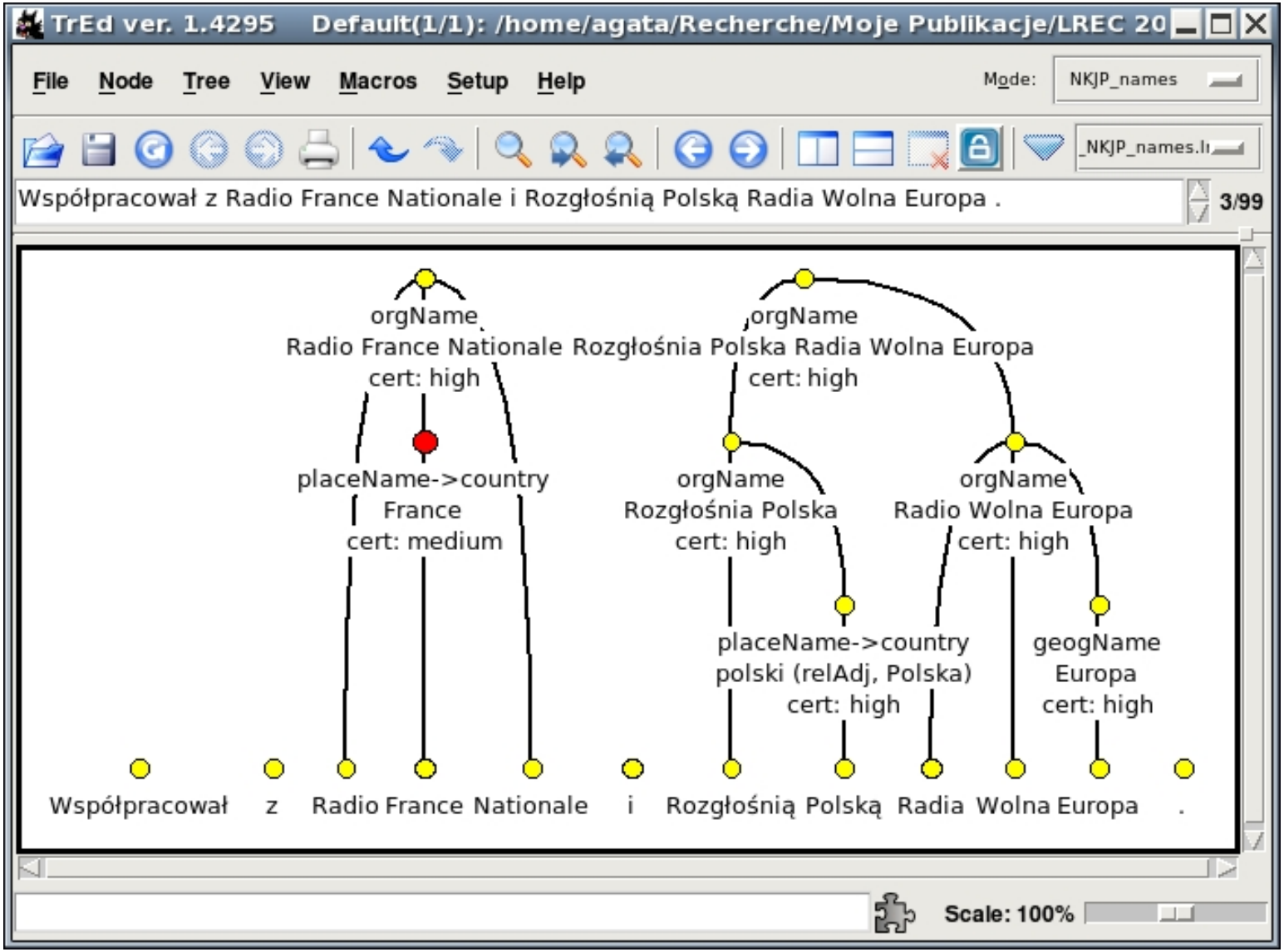
- gazetteer with 300,000 inflected forms
- 120 grammar rules
- precision: 0.88
- recall: 0.61

TrEd

Tree Editor *(Pajas & Štěpánek 2008)*

- manipulates tree structures (necessary for embedded, coordinated & discontinuous NEs)
- interoperable
- allows for stand-off multi-level annotations
- PML - open customizable XML abstract data format
- customizable GUI
- easy comparing two annotation
- reliable and well documented

TrEd interface



Conclusions and perspectives

What has been done

- State-of-the-art corpus methodology
- Advanced annotation strategies
- Annotator's and super-annotator's platform
- 18,000 annotated sentences (out of 75,000) until mid-May

To do

- 75% corpus to be annotated
- super-annotation
- machine-learning for 1 billion corpus

Further perspectives

- extending annotation to new categories (periods, events,. . .)
- corpus studies