Annotation d'entités nommées dans le Corpus National du Polonais

Agata Savary†, Adam Przepiórkowski*, Jakub Waszczuk*, Michał Lenart*

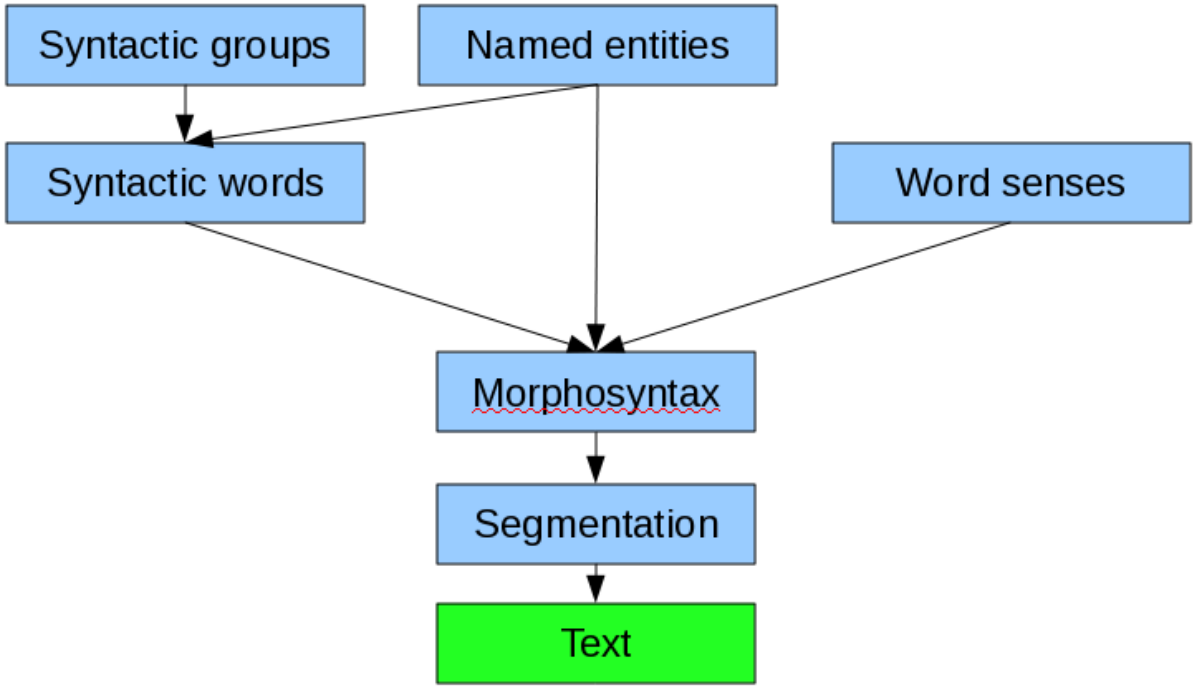
† Université François Rabelais Tours, Laboratoire d'Informatique, Blois, France
* Institute of Computer Science, Polish Academy of Sciences (IPIPAN), Warsaw, Poland (sabbatical in 2009–2010)

JIRC, November 17, 2011

National Corpus of Polish (NKJP)

- Coordinator: IPIPAN Warsaw
- The aim: a large national corpus of Polish
- 6 levels of linguistic annotation (segmentation, morphosyntax, syntactic words, syntactic groups, **named entities**, word senses)
- 1.5 billion words automatically annotated *(Przepiórkowski et al. LREC'2010)*
- 1 million words manually annotated (*gold standard*)
- representative and balanced *(Przepiórkowski et al. 2009)*
- stand-off annotation

Multi-level annotation

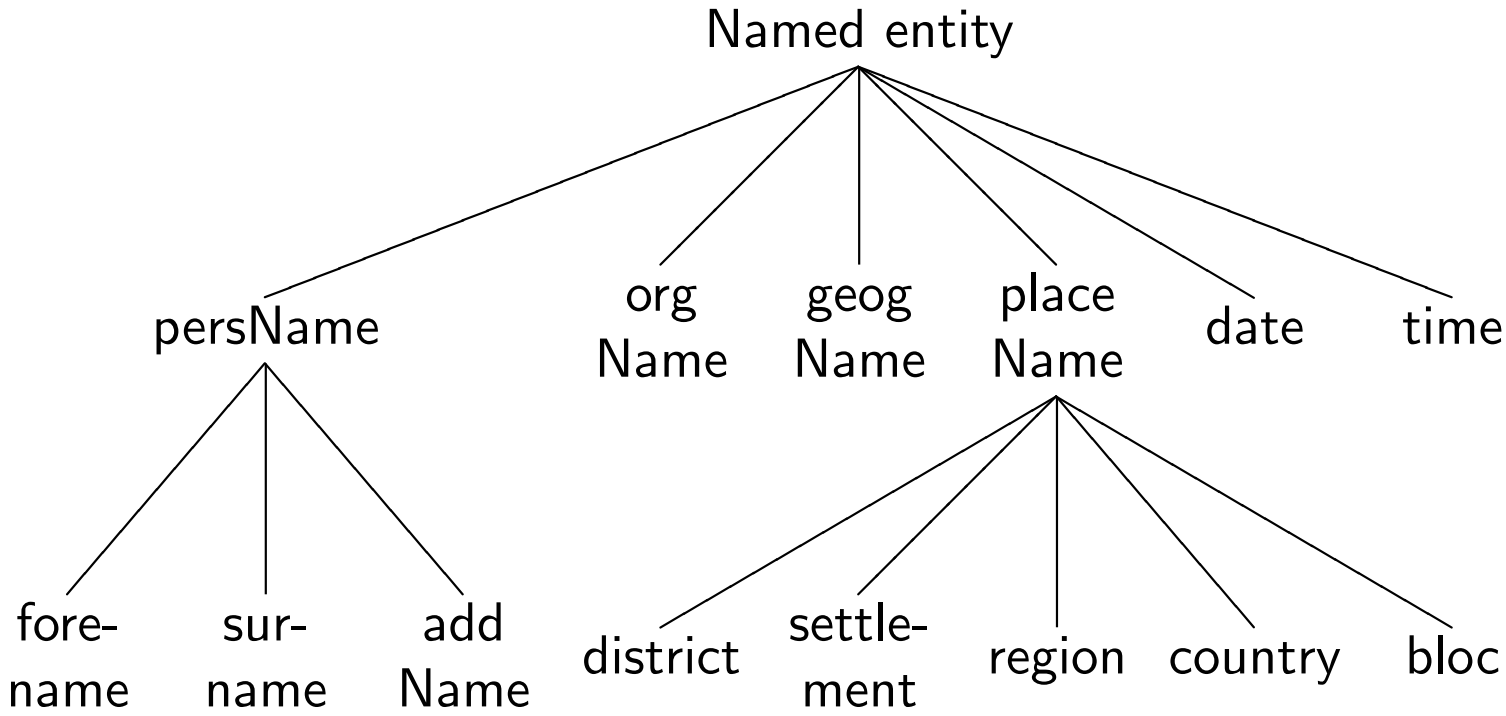


- 8 XML files for each text: text.xml, header.xml, ann_segmentation.xml, ann_morphosyntax.xml, ann_words.xml, ann_groups.xml, **ann_named.xml**, ann_senses.xml
- Complex integrity constraints (checked currently by *ad hoc* stripts)
- 2.6 GB in over 31,000 file in the gold standard (1,500 times as much in the total corpus)

TEI P5 standard format

```
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude" xmlns="http://www.tei-c.org/ns/1.0">
<xi:include href="NKJP_1M_header.xml"/>
<TEI>
  <xi:include href="header.xml"/>
  <text><body>
  <p xml:id="named_1-p" corresp="ann_words.xml#words_1-p">
    <s xml:id="named_1.34-s" corresp="ann_words.xml#words_1.34-s">
      <seg xml:id="named_1.34-s_n3">
        <fs type="named">
          <f name="derived">
            <fs type="derivation">
              <f name="derivType"><symbol value="relAdj"/></f>
              <f name="derivedFrom"><string>Irlandia</string></f>
            </fs>
          </f>
          <f name="ne_type"><symbol value="placeName"/></f>
          <f name="ne_subtype"><symbol value="country"/></f>
          <f name="orth"><string>Irlandzka</string></f>
          <f name="base"><string>irlandzki</string></f>
          <f name="certainty"><symbol value="high"/></f>
        </fs>
        <ptr target="ann_morphosyntax.xml#morph_1.1-seg"/>
      </seg>
    </s></p></body></text></TEI></teiCorpus>
```

Concept hierarchy for named entities



- vertical hierarchy of **related names**
  - relational adjectives *warszawski* 'Warsaw-related'
  - names of inhabitants and members *warszawiak* 'Warsaw inhabitant'

Annotation strategies

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

  Stanów Zjednoczonych → *Stany Zjednoczone* 'United Stated'
- Sematically motivated derivation base

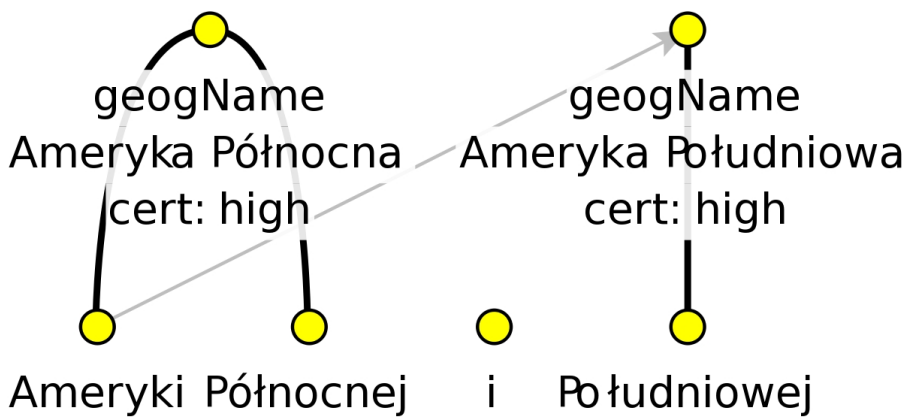
  warszawski, podwarszawski → *Warszawa*

  '(near-)Warsaw-related'

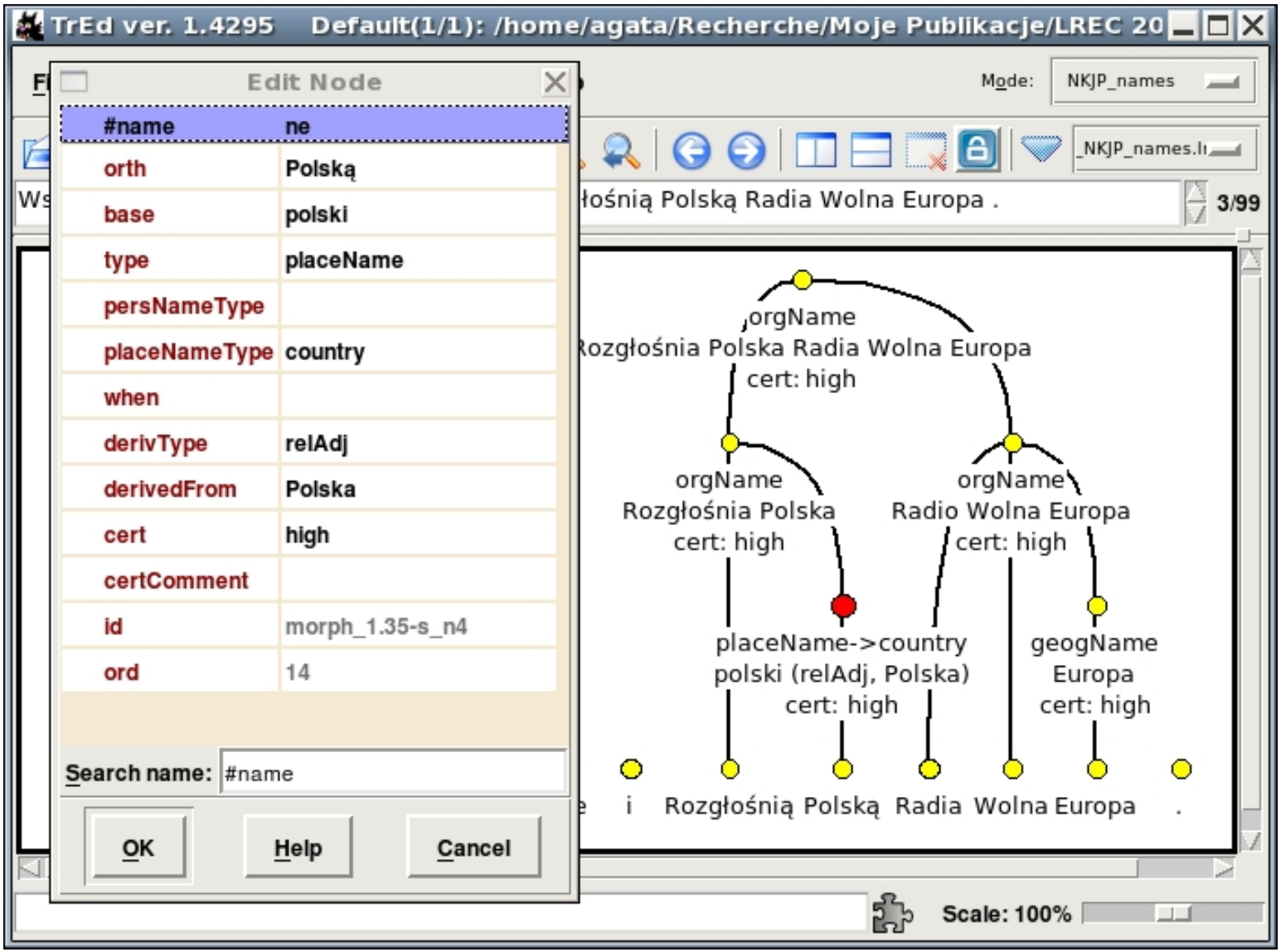
  amerykański → *Stany Zjednoczone* 'American → USA'
- Embedded names annotated *(Galicia-Haro and Gelbukh 2009; Finkel and Manning 2009; Kravalová and Žabokrtský 2009)*

  $[[\textit{Tadeusz}]_{\text{forename}}[\textit{Kościuszko}]_{\text{surname}}]_{\text{persName}}$
- Coordinated and discontinuous names *(Mazur and Dale 2009)*

  Ameryka Północna i Południowa 'North and South America'



TrEd – manual annotation tool



'Polish Broadcasting Station of the Free Europe Radio'

Concept hierarchy vs. linguistic fuzzyness

NE annotation – formal classification problem

- fixed scope (what we do and do not annotate)
- fixed hierarchy of concepts
- fixed annotation strategies

Named entities – linguistic objects with fuzzy properties

- controversial status of NEs
- fuzzy boundaries between categories
- fuzzy lexical and semantic relations

Examples of linguistic problems

- metonymy
- ellipsis
- relational adjectives
- extra-linguistic issues

Metonymy – case study: europejski ‘European’

Derivational base and its type

- *Europa* `geogName` ‘Europe’ (continent)
- *Europa* `orgName` ‘Europe’ (inhabitants/institution)
- *Unia Europejska* `placeName`→`bloc` ‘European Union’
- *Unia Europejska* `orgName` ‘European Union’
- *Parlament Europejski* `orgName` ‘European Parliament’
- *Komisja Europejska* `orgName` ‘European Commission’

Corpus examples: europejski ‘European’

- *europejski*: relAdj(orgName(*Parlament Europejski*))

  Wkrótce odbędą się kolejne europejskie wybory.

  ‘European elections will be held soon’

- *europejski*: relAdj(orgName)*Unia Europejska*))
  europejski: relAdj(placeName→bloc(*Unia Europejska*))
  europejski: relAdj(geogName(*Europa*))

  Państwo prawa, odpowiadające standardom państw europejskich.

  ‘State of law appropriate to standards of European countries’

- *europejski*: relAdj(geogName(*Europa*))
  europejski: relAdj(placeName→bloc(*Unia Europejska*))

  Spała, Wałcz czy Cetniewo są od dawna ośrodkami na najwyższym europejskim poziomie.

  ‘The resorts of Spała, Wałcz and Cetniewo have been up to the highest European standards for a long time’

Geopolitical, national and religious issues

- controversial status (country vs. region)

  Kosowo, Palestyna 'Kosovo, Palestine'

- simplistic relation between population and territory

  Ormianie $\rightarrow$ *Armenia* ? 'Armenians, Armenia'
  Żydzi, Cyganie $\rightarrow$? 'Jews, Romani'
  Arabowie $\rightarrow$? 'Arabs'

- religious groups as organizations

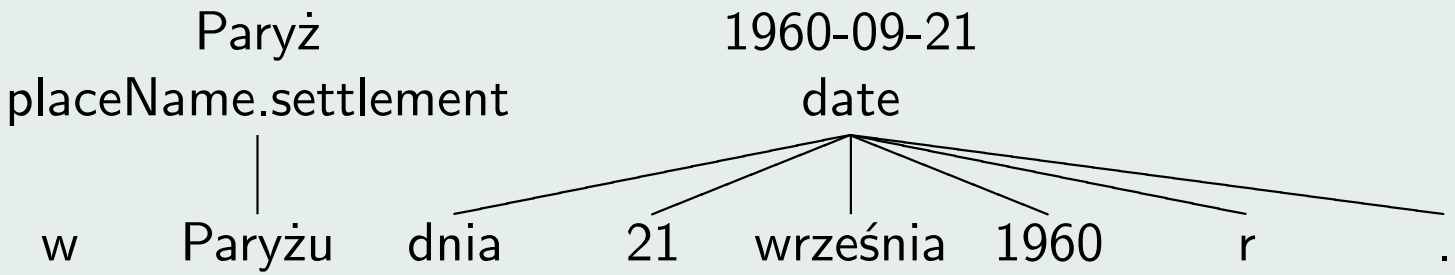
  buddyzm common name or NE ? 'buddhism'

- group name = member name

  $[\textit{karmelici}]_{orgName}$ vs. $[\textit{karmelici}]_{persDeriv(?)}$ 'Carmelites'

Automatic NE annotation as a tagging problem

Corpus example



'in Paris on September 21, 1960'

Annotations as IOB tags

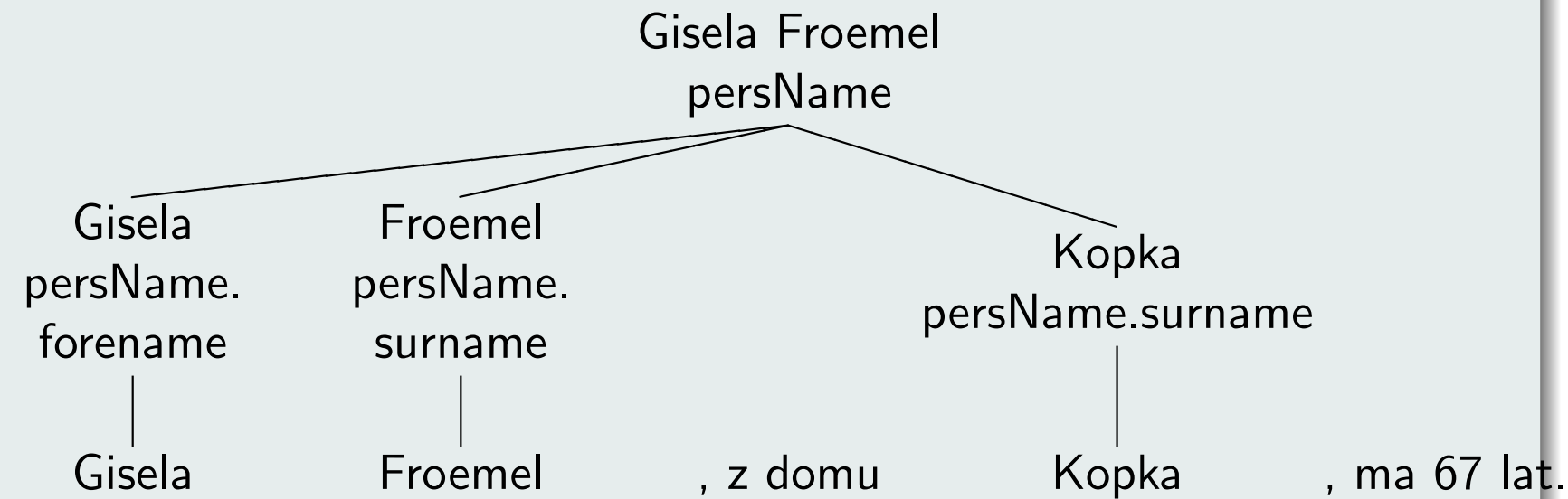
O, B-settlement, B-date, I-date, I-date, I-date, I-date, I-date

Post-processing heuristics

O, B-settlement, **I-date**, ... → O, B-settlement, **B-date**, ...

Non-contiguous annotations as tags

Corpus example



'Gisela Froemel (maiden name Kopka)'

Annotations as IOB tags (subtypes are omitted)

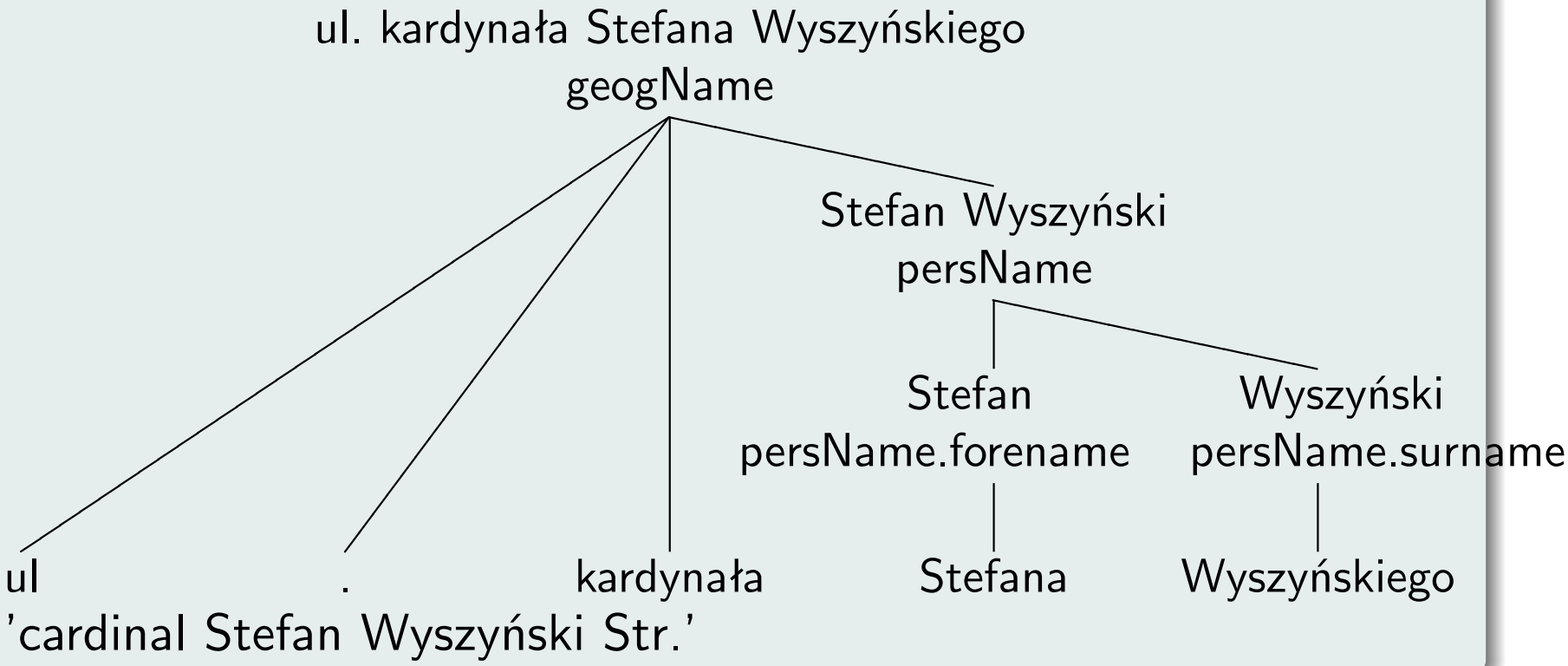
B-persName, I-persName, O, O, O, I-persName, O, O, O, O, O

Some post-processing heuristics non applicable

O, I-persName, ... $\not\rightarrow$ O, B-persName, ...

Embedded annotations as tags

Corpus example

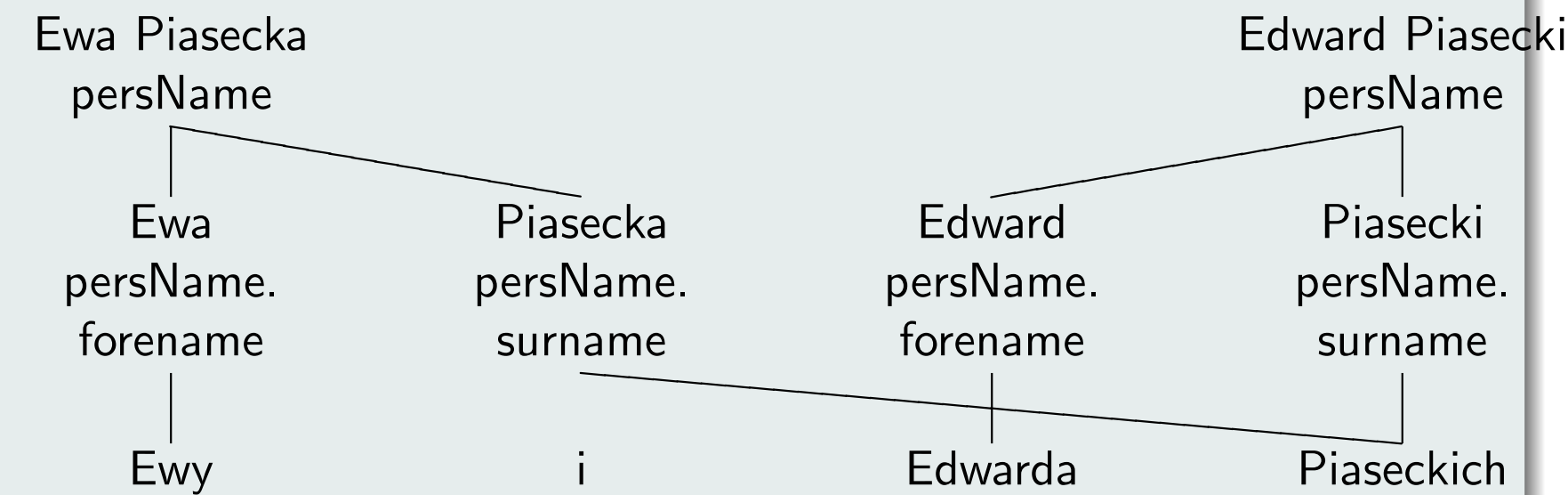


Joint label tagging *(Alex et al. 2007)*

B-geogName, I-geogName, I-geogname,
B-forename#B-persName#I-geogName,
B-surname#I-persName#I-geogName

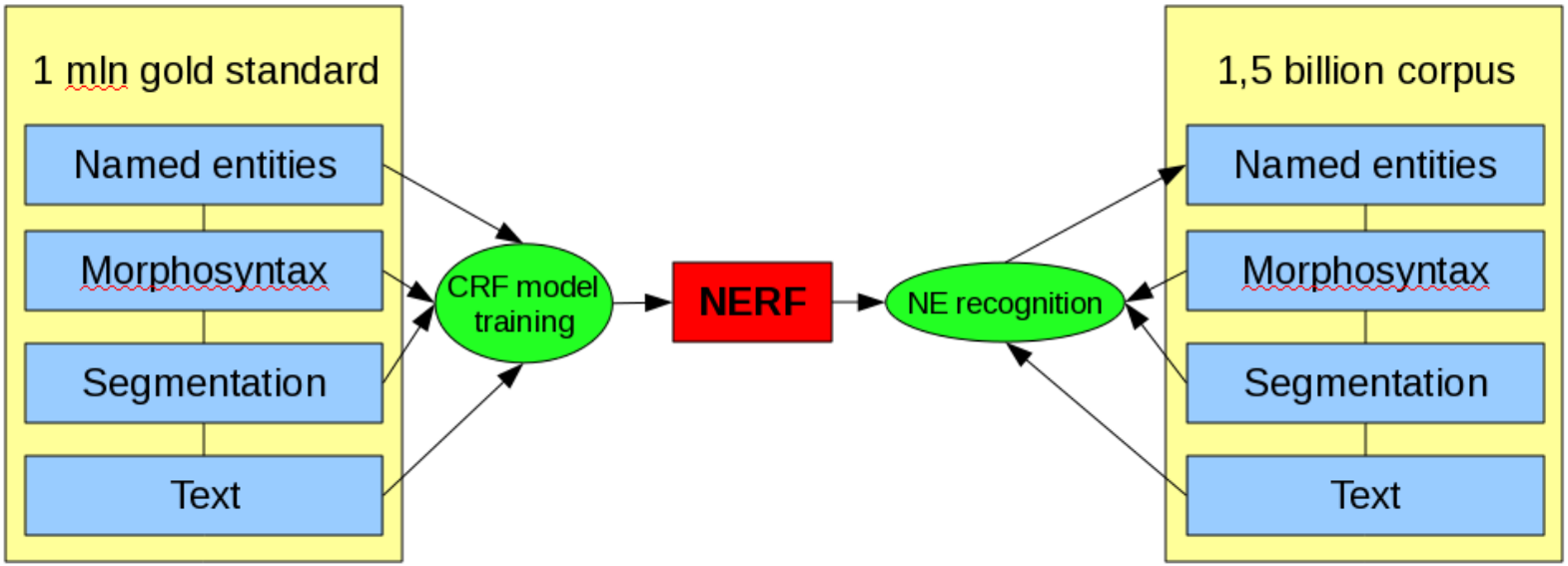
Overlapping annotations as tags

Corpus example



- indescribable by tags
- long-distance relations hard to process by sequential taggers
- 1% of all NE in the NKJP are overlapping

NERF – CRF-based baseline tagger



Basic feature set (for segments n and $n-1$)

- text form and lemma (e.g. *Giseli*, *Gisela*)
- prefixes and suffixes (*gis*, *eli*)
- parts of speech (*subst*) and grammatical tags (*sg:f:gen*)
- shape (*ullll*, *ddxddxdddd*←*17.11.2011*)

NERF – results

	Average		
	Precision	Recall	F-measure
Person names	0,86	0,80	0,83
Geographical and place names	0,83	0,71	0,77
Organisations	0,70	0,65	0,67
Temoral expressions	0,86	0,80	0,83
Derivations	0,87	0,68	0,77
Total	0,83	0,76	0,79

NERF – current development

Use of lexical resources

- Features based on NE *grammatical* gazetteers (Polish and foreign NEs inflect in Polish!)
- Approximate string matching of words against the lexical resources (if the inflection of a NE might be unknown)
- Using gazetteers of multi-word NEs

Towards a tree-based model (CRF-PCFG-like)

- a probabilistic context-free grammar chooses the most probable syntax tree (not only the sequence of labels)

Challenges

- annotating attributes with open values (lemmas and derivational bases)
- annotating oral data (frequent non contiguous names due to disfluencies)
- hybrid systems (rules + ML)
- taking more complex features into account
- integrating the level of syntactic groups (a NE often precisely overlaps a noun phrase)