# Schemat kodowania dla korpusów z wieloma poziomami anotacji lingwistycznej zgodny ze standardem TEI

Szymon Bemowski
Uniwersytet Warszawski
IPI PAN

12 października 2008

# XCES

XML Corpus Encoding Standard
http://www.xces.org/

- Konkretny format kodowania korpusów
- Poziomy anotacji: segmentacja i morfoskładnia
- Brak mechanizmów do wyrażania alternatyw i nieciągłości
- Brak wyczerpującej dokumentacji

# LAF

Linguistic Annotation Framework

- Specyfikacja abstrakcyjnego modelu danych
- Brak gotowego do użycia formatu
- Standard w trakcie powstawania

# TEI

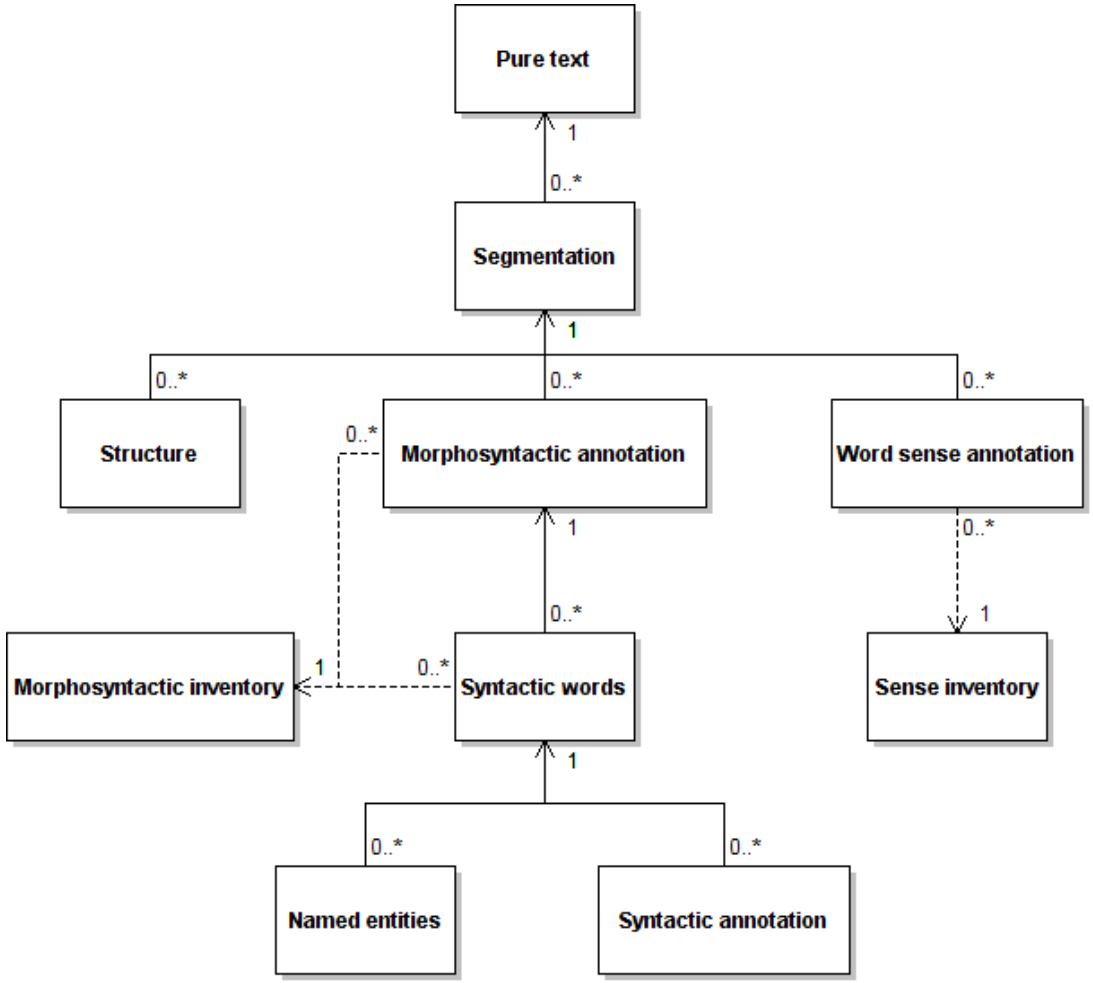
TEI Guidelines for Electronic Text Encoding and Interchange
http://www.tei-c.org/

- Zestaw sposobów kodowania zjawisk lingwistycznych
- Obszerna dokumentacja
- Konieczność tworzenia podzbioru dla korpusów
- Wiele możliwości wykorzystania dostępnych mechanizmów

# Ogólne założenia

- Gotowy do użycia format
- Wsparcie dla wielu poziomów anotacji
- Specyfikacja zależności pomiędzy poziomami
- Wykorzystanie *stand-off annotation*
- Wyrażalność zagnieżdżenia, alternatywy i nieciągłości
- Silna zgodność z TEI
- Elastyczność i rozszerzalność

# Poziomy anotacji



# Reprezentacja tekstów w korpusie

Reprezentacja pojedynczego tekstu - katalog zawierający:

- Plik z czystym tekstem (*Primary source*)
- Pliki z poziomami anotacji (przynajmniej jeden na poziom)
- Plik z nagłówkiem tekstu

Struktury danych wspólne dla całego korpusu:

- Nagłówek korpusu
- *Morphosyntactic inventory*
- *Sense inventory*

# Metaelementy TEI

```
<teiCorpus xmlns:xi="../XInclude">
  <xi:include xml:base="corpusHeader.xml"
              xpointer="#CH001" />
  <TEI>
    <xi:include xml:base="teiHeader.xml"
                xpointer="#H001" />
    <text>
      <body>
        <!-- ZAWARTOŚĆ POZIOMU -->
      </body>
    </text>
  </TEI>
</teiCorpus>
```

# Odwołania pomiędzy poziomami

```
<xi:include xml:id="i001"
  xml:base="lowerLevel.xml"
  xpointer="string-range(#b001,17,8)" />

<xi:include xml:id="i002"
  xml:base="lowerLevel.xml"
  xpointer="#t001" />

<xi:include xml:id="i003"
  xml:base="lowerLevel.xml"
  xpointer="range(#t003, #t010)" />
```

# Jednostki anotacji

```
<seg type="..." xml:id="...">
  <!-- ZAWARTOŚĆ JEDNOSTKI -->
  <!-- DANE LINGWISTYCZNE -->
<seg>
```

Wartości atrybutu `type`:

- `tok`, `synWord`, `synGroup`, `synUtt`, `morphInterp`, `semInterp`

Wyspecializowane elementy TEI:

- `<name>`, `<date>`, `<number>`
- `<div>`, `<p>`, `<s>`, `<head>`

# Jednostki anotacji - zawartość

```
<seg type="synWord" xml:id="st05">
  <xi:include
    xml:base="segmentation.xml"
    xpointer="#t02" />
  <seg type="synWord" xml:id="st06">
    <xi:include
      xml:base="segmentation.xml"
      xpointer="range(#t03,#t04)" />
  <seg>
  <xi:include
    xml:base="segmentation.xml"
    xpointer="range(#t05,#t06)" />
  <!-- DANE LINGWISTYCZNE -->
<seg>
```

# Jednostki anotacji - dane lingwistyczne

```
<seg type="synGroup" xml:id="g03">
  <xi:include
    xml:base="syntactic-words.xml"
    xpointer="range(#st02,#st05)" />
  <fs type="grpData">
    <f name="gType">
      <symbol value="NumG"/> </f>
    <f name="synHead">
      <string value="#st02"/> </f>
    <f name="semHead">
      <string value="#st04"/> </f>
  </fs>
<seg>
```

# Alternatywa

```
<choice xml:id="ch00" select="#br02">

  <!-- Gałąź 1: -->
  <seg type="branch" xml:id="br01">
    <seg type="tok" xml:id="t01"/>
      <xi:include xpointer="string-range(#b01,4,12)"/>
    </seg>
  </seg>

  <!-- Gałąź 2: -->
  <seg type="branch" xml:id="br02">
    <seg type="tok" xml:id="t02"/>
      <xi:include xpointer="string-range(#b01,4,7)"/>
    </seg>
    <seg type="tok" xml:id="t03"/>
      <xi:include xpointer="string-range(#b01,11,5)"/>
    </seg>
  </seg>
</choice>
```

# Nieciągłość

```
<!-- Jednostka agregująca: -->
<seg type="synWord" xml:id="sw01">
  <fs type="morphInterp">...</fs>
  <ptr type="parts" target="#sw02 #sw03" />
</seg>

<!-- I jednostka składowa: -->
<seg type="synWord" xml:id="sw02" part="Y"/>
  <ptr type="aggr" target="#sw01" />
  <xi:include xpointer="#md01"/>
</seg>

<!-- Fragment zewnętrzny: -->
<xi:include xpointer="#md02"/>

<!-- II jednostka składowa: -->
<seg type="synWord" xml:id="sw03" part="Y"/>
  <ptr type="aggr" target="#sw01" />
  <xi:include xpointer="#md03"/>
</seg>
```

# Czysty tekst

```
<body xml:id="b00">
  Rozdział, 1. <gap/>
  Janek z całą pewnością nie chciałby
  czarno-białego telewizora Panasonic!
</body>
```

# Segmentacja

```
<body xml:id="s01" xml:base="text.xml">

  <seg type="tok" xml:id="t01">    <!-- Rozdział -->
    <xi:include xpointer="string-range(#b00,0,8)" />
  </seg>

  <seg type="tok" xml:id="t02">    <!-- 1 -->
    <xi:include xpointer="string-range(#b00,9,1)" />
  </seg>

  <seg type="tok" xml:id="t03">    <!-- . -->
    <xi:include xpointer="string-range(#b00,10,1)" />
  </seg>

  <seg type="tok" xml:id="t04">    <!-- Janek -->
    <xi:include xpointer="string-range(#b00,18,5)" />
  </seg>

  <!-- Pozostałe segmenty ... -->
</body>
```

# Struktura

```
<front> ... </front>
<body xml:id="b02" xml:base="segm.xml">
  <div type="chapter" xml:id="div01">
    <head>    <!-- Rozdział 1. -->
      <xi:include xpointer="range(#t01,#t03)" />
    </head>
  <p>    <!-- Janek (...) Panasonic! -->
    <s> <xi:include xpointer="range(#t04,#t16)" /> </s>
  </p>
  </div>
</body>
<back> ... </back>
```

# Morfoskładnia

```
<seg type="morphInterp" xml:id="m05" />
  <xi:include xpointer="#t05" />
  <fs type="morphData">
   <f name="orth"> <string value="z"/> </f>
   <f name="interps">
    <vColl org="set">
     <fs type="lex">
       <f name="base"> <string value="z"/> </f>
       <f name="ctag"> <symbol value="prep"/> </f>
       <f name="msd" fVal="#gen" xml:base="morphInv.xml" />
       <f name="disamb"> <numeric value="0.3" /> </f>
     </fs>
     <fs type="lex"> ... </fs>
    </vColl>
   </f>
  </fs>
</seg>
```

# Wyrazy składniowe

```
<body xml:id="b04" xml:base="morph.xml">

  <!--  Janek z całą pewnością  -->
  <xi:include xpointer="range(#m04,#m07)" />

  <!--  nie chiałby  -->
  <seg type="synWord" xml:id="st01" />
    <xi:include xpointer="range(#m08,#m10)" />
    <fs type="morphData">...</fs>
  </seg>

  <!--  czarno|-|białego  -->
  <seg type="synWord" xml:id="st02" />
    <xi:include xpointer="range(#m11,#m13)" />
    <fs type="morphData">...</fs>
  </seg>

  <!--  telewizora Panasonic!  -->
  <xi:include xpointer="range(#m14,#m16)" />
</body>
```

# Składnia

```
<body xml:id="b05" xml:base="sWords.xml">
 <seg type="synUtt" xml:id="u01">
  <!-- poprzedzające znaczniki składniowe... -->

  <seg type="synGroup" xml:id="g04" >
    <xi:include xpointer="range(#st02, #m15)" />
    <fs type="grpData">
      <f name="gType"> <symbol value="NG"/> </f>
      <f name="synHead"> <string value="#m14"/> </f>
      <f name="semHead"> <string value="#m14"/> </f>
   </fs>
  </seg>

  <!-- pozostałe znaczniki składniowe... -->
 </seg>
</body>
```

# Byty nazwane

```
<body xml:id="b06" xml:base="sWords.xml">

  <!-- Janek -->
  <name type="person" xml:id="ne01" />
    <xi:include xpointer="#m04" />
  </name>

  <!-- z całą (...) telewizora -->
  <xi:include xpointer="range(#m05, #m14)" />

  <!-- Panasonic -->
  <name type="brand" xml:id="ne02" />
    <xi:include xpointer="#m15" />
  </name>

  <!-- wykrzyknik -->
  <xi:include xpointer="#m16" />
</body>
```

# Sensy słów

```
<seg type="semInterp" xml:id="sd01">
  <!-- telewizora -->
  <xi:include xpointer="#m14"/>
  <fs type="semData">
    <f name="sense"
       fVal="#si:telewizor:sense_1" />
  </fs>
</seg>
```

# Podsumowanie

Główne zalety zaproponowanego schematu:

- Gotowy do użycia format
- Wsparcie dla wielu poziomów anotacji
- Zgodność z TEI
- Rozszerzalmość

# Koniec

Dziękuję za uwagę.