

TEI/XCES for Corpus Encoding

Adam Przepiórkowski



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. J. K. Ordona 21, 01-237 Warszawa

NEERI'09

30 September 2009, Helsinki

National Corpus of Polish (NCP) project:

- a 3-year project (2008–2010),
- 4 institutions, coordinated by IPI PAN (Warsaw),
- <http://nkjp.pl/>,
- cf. also LREC 2008 paper,

Some features:

- based on all previous Polish corpora,
- extended to 1 billion words,
- multiple levels of annotation,
- 1 million word subcorpus annotated manually.

National Corpus of Polish (NCP) project:

- a 3-year project (2008–2010),
- 4 institutions, coordinated by IPI PAN (Warsaw),
- <http://nkjp.pl/>,
- cf. also LREC 2008 paper,

Some features:

- based on all previous Polish corpora,
- extended to 1 billion words,
- multiple levels of annotation,
- 1 million word subcorpus annotated manually.

General annotation:

- structure (chapters, sections, paragraphs),
- metadata.

Linguistic annotation:

- word-level and sentence-level segmentation,
- morphosyntactic annotation,
- syntactic words,
- syntactic groups,
- named entities,
- word sense annotation.

Need for stand-off annotation:

- alternative annotations at the same level,
- possible overlap of elements at different levels, etc.

General annotation:

- structure (chapters, sections, paragraphs),
- metadata.

Linguistic annotation:

- word-level and sentence-level segmentation,
- morphosyntactic annotation,
- syntactic words,
- syntactic groups,
- named entities,
- word sense annotation.

Need for stand-off annotation:

- alternative annotations at the same level,
- possible overlap of elements at different levels, etc.

General annotation:

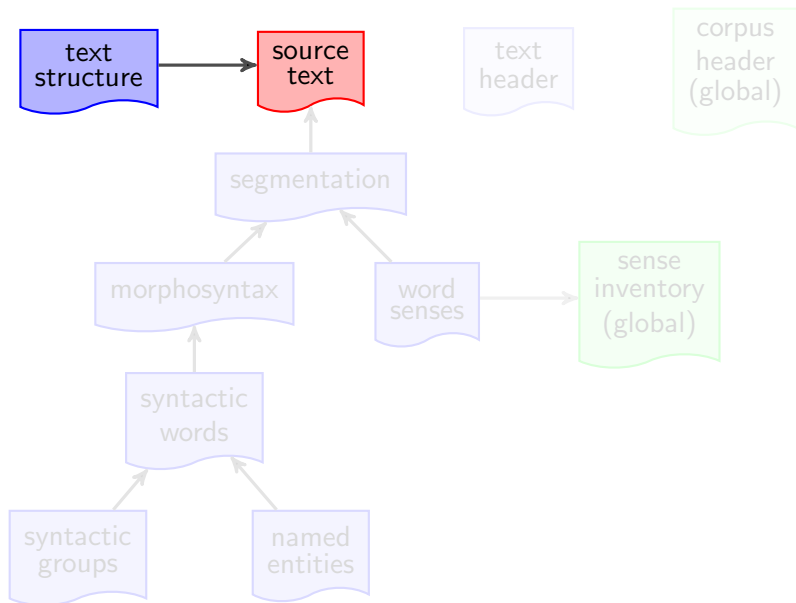
- structure (chapters, sections, paragraphs),
- metadata.

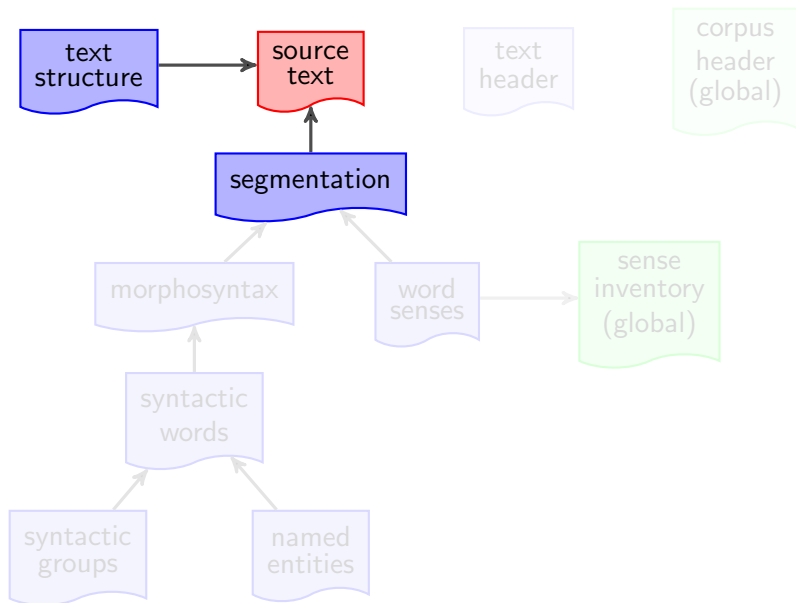
Linguistic annotation:

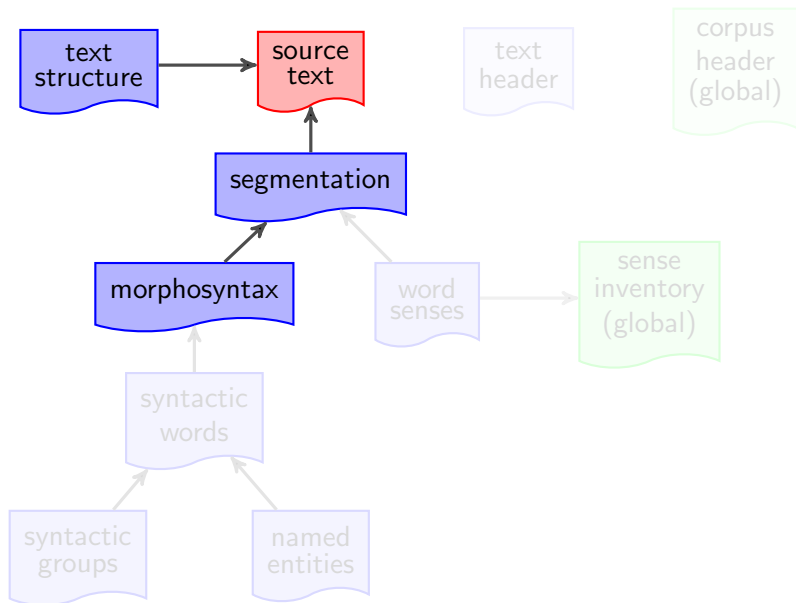
- word-level and sentence-level segmentation,
- morphosyntactic annotation,
- syntactic words,
- syntactic groups,
- named entities,
- word sense annotation.

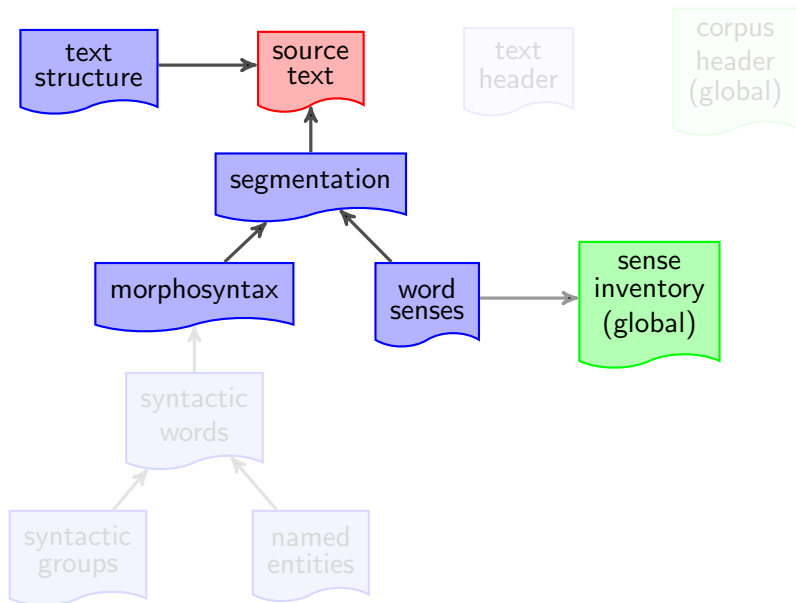
Need for stand-off annotation:

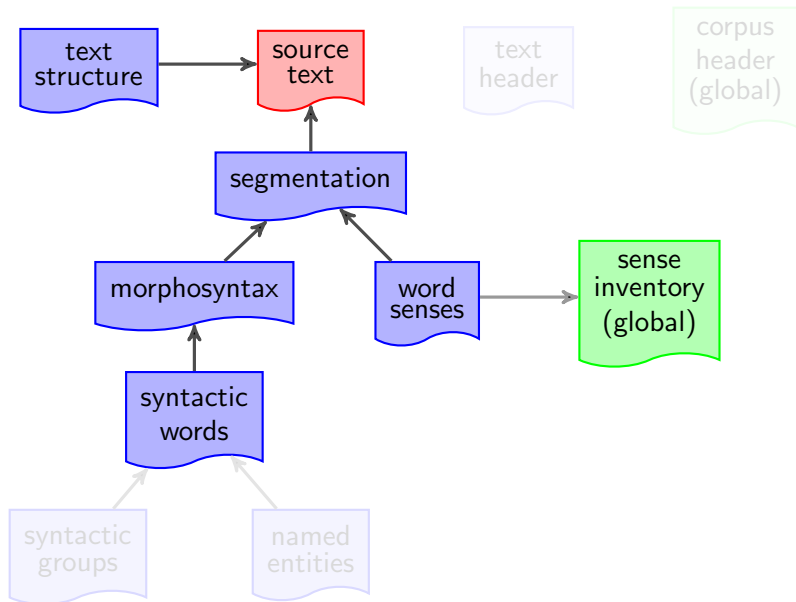
- alternative annotations at the same level,
- possible overlap of elements at different levels, etc.

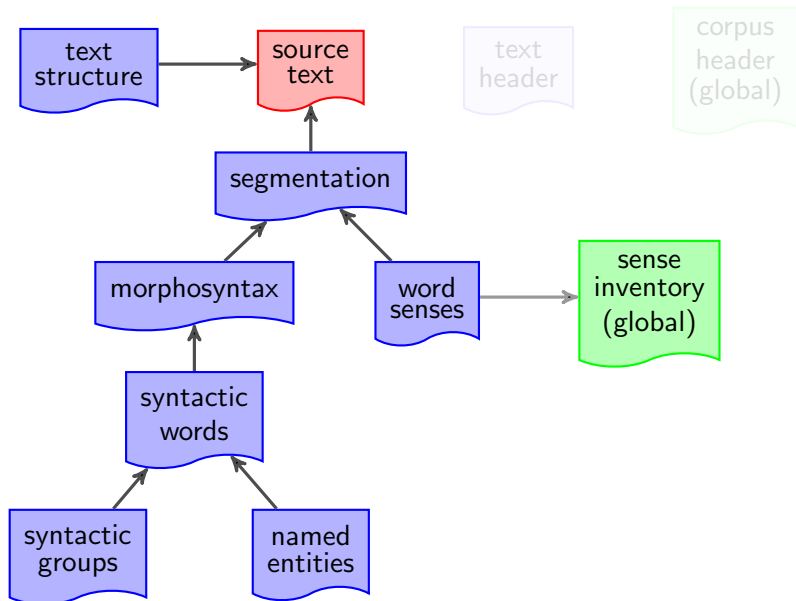


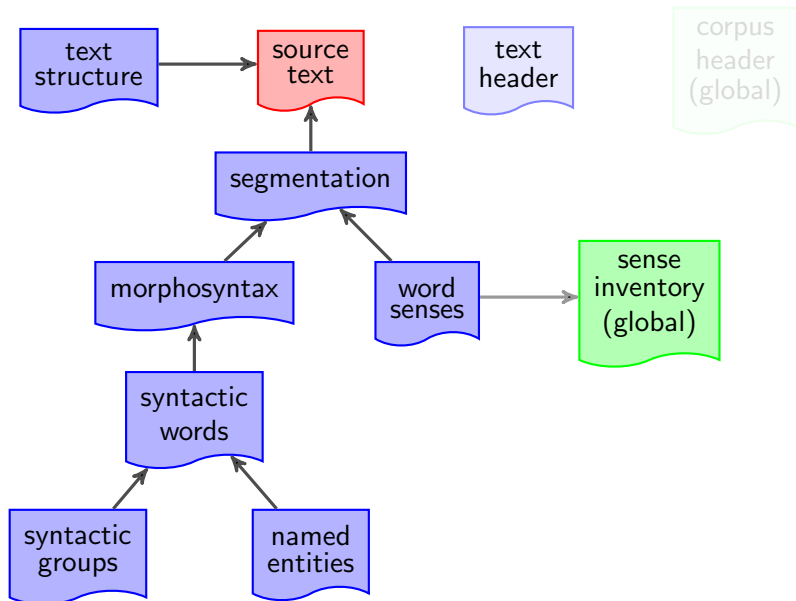


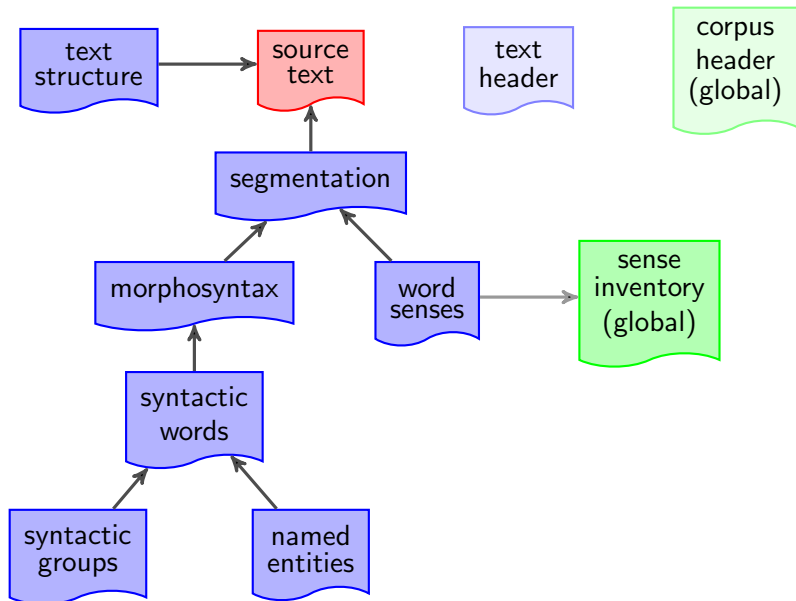












What encoding standards to use?

- **XCES**: XML Corpus Encoding Standard (Ide *et al.*, 2000),
- **TEI**: Text Encoding Initiative (Burnard and Bauman, 2008)
- ISO TC 37/SC 4 family — covered in the previous talk,
- other? — in the next talk.

For NCP:

- ISO TC 37/SC 4 family not sufficiently mature,
- seriously considered: **XCES** and **TEI**.

What encoding standards to use?

- **XCES**: XML Corpus Encoding Standard (Ide *et al.*, 2000),
- **TEI**: Text Encoding Initiative (Burnard and Bauman, 2008),
- ISO TC 37/SC 4 family — covered in the previous talk,
- other? — in the next talk.

For NCP:

- ISO TC 37/SC 4 family not sufficiently mature,
- seriously considered: **XCES** and **TEI**.

What encoding standards to use?

- **XCES**: XML Corpus Encoding Standard (Ide *et al.*, 2000),
- **TEI**: Text Encoding Initiative (Burnard and Bauman, 2008),
- ISO TC 37/SC 4 family — covered in the previous talk,
- other? — in the next talk.

For NCP:

- ISO TC 37/SC 4 family not sufficiently mature,
- seriously considered: **XCES** and **TEI**.

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- used in the IPI PAN Corpus of Polish,
- a *de facto* standard for corpus annotation,
- an instantiation and adaptation of TEI (P4).

Cons:

- no documentation (but documentation for CES),
- no specific annotation levels beyond morphosyntax,
- potential for confusion (DTD vs. XML Schema; no explicit information about the differences),
- later schemata more abstract (just use feature structures),
- general feature structure mechanisms not compliant with the ISO FSR standard,
- no mechanisms for discontinuity and alternatives,
- not compliant with TEI P5 (but there is a relevant promise on the XCES WWW page).

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

Pros:

- a *de facto* standard for text representation,
- well-documented,
- very rich,
- elements have well-defined semantics,
- alive and kicking,
- free (unlike ISO).

Cons:

- documentation daunting,
- large toolkit to select from — no single solution,
- few publicly available examples of TEI P5 corpora.

The general approach for NCP:

- use TEI P5,
- when a choice is available, choose solutions isomorphic with other standards.

More general thesis:

- for any reasonable corpus encoding standard,
- there is a solution isomophoric with it offered by TEI P5.

The general approach for NCP:

- use TEI P5,
- when a choice is available, choose solutions isomorphic with other standards.

More general thesis:

- for any reasonable corpus encoding standard,
- there is a solution isomophoric with it offered by TEI P5.

Various proposed and *de facto* standards for treebank XML encoding: TIGER-XML, SynAF, PAULA, etc. Isomorphic representations available in TEI P5.

```
<nt id="nt2"> <!-- TIGER-XML -->
  <edge label="head" idref="#t6"/>
  <edge label="nonhead" idref="#nt20"/>
  <edge label="nonhead" idref="#nt21"/>
</nt>

<struct id="syn2"> <!-- PAULA -->
  <rel id="rel3" type="head" xlink:href="tok.xml#t6"/>
  <rel id="rel4" type="nonhead" xlink:href="#syn20"/>
  <rel id="rel6" type="nonhead" xlink:href="#syn21"/>
</struct>

<seg xml:id="group2"> <!-- TEI -->
  <fs>...</fs>
  <ptr xml:id="ptr3" type="head" target="ann_morphosyntax.xml#seg6"/>
  <ptr xml:id="ptr4" type="nonhead" target="#group20"/>
  <ptr xml:id="ptr6" type="nonhead" target="#group21"/>
</seg>
```

Various proposed and *de facto* standards for treebank XML encoding: TIGER-XML, SynAF, PAULA, etc. Isomorphic representations available in TEI P5.

```
<nt id="nt2"> <!-- TIGER-XML -->
  <edge label="head" idref="#t6"/>
  <edge label="nonhead" idref="#nt20"/>
  <edge label="nonhead" idref="#nt21"/>
</nt>

<struct id="syn2"> <!-- PAULA -->
  <rel id="rel3" type="head" xlink:href="tok.xml#t6"/>
  <rel id="rel4" type="nonhead" xlink:href="#syn20"/>
  <rel id="rel6" type="nonhead" xlink:href="#syn21"/>
</struct>

<seg xml:id="group2"> <!-- TEI -->
  <fs>...</fs>
  <ptr xml:id="ptr3" type="head" target="ann_morphosyntax.xml#seg6"/>
  <ptr xml:id="ptr4" type="nonhead" target="#group20"/>
  <ptr xml:id="ptr6" type="nonhead" target="#group21"/>
</seg>
```

Various proposed and *de facto* standards for treebank XML encoding: TIGER-XML, SynAF, PAULA, etc. Isomorphic representations available in TEI P5.

```
<nt id="nt2"> <!-- TIGER-XML -->
  <edge label="head" idref="#t6"/>
  <edge label="nonhead" idref="#nt20"/>
  <edge label="nonhead" idref="#nt21"/>
</nt>

<struct id="syn2"> <!-- PAULA -->
  <rel id="rel3" type="head" xlink:href="tok.xml#t6"/>
  <rel id="rel4" type="nonhead" xlink:href="#syn20"/>
  <rel id="rel6" type="nonhead" xlink:href="#syn21"/>
</struct>

<seg xml:id="group2"> <!-- TEI -->
  <fs>...</fs>
  <ptr xml:id="ptr3" type="head" target="ann_morphosyntax.xml#seg6"/>
  <ptr xml:id="ptr4" type="nonhead" target="#group20"/>
  <ptr xml:id="ptr6" type="nonhead" target="#group21"/>
</seg>
```

Various proposed and *de facto* standards for treebank XML encoding: TIGER-XML, SynAF, PAULA, etc. Isomorphic representations available in TEI P5.

```
<nt id="nt2"> <!-- TIGER-XML -->
  <edge label="head" idref="#t6"/>
  <edge label="nonhead" idref="#nt20"/>
  <edge label="nonhead" idref="#nt21"/>
</nt>

<struct id="syn2"> <!-- PAULA -->
  <rel id="rel3" type="head" xlink:href="tok.xml#t6"/>
  <rel id="rel4" type="nonhead" xlink:href="#syn20"/>
  <rel id="rel6" type="nonhead" xlink:href="#syn21"/>
</struct>

<seg xml:id="group2"> <!-- TEI -->
  <fs>...</fs>
  <ptr xml:id="ptr3" type="head" target="ann_morphosyntax.xml#seg6"/>
  <ptr xml:id="ptr4" type="nonhead" target="#group20"/>
  <ptr xml:id="ptr6" type="nonhead" target="#group21"/>
</seg>
```

Based on the experiences of NCP:

- TEI is a very reasonable choice for encoding various kinds of corpora,
- what's missing is specific examples of TEI P5-encoded corpora.

Within NCP, coming soon (<http://nkjp.pl/>):

- articles showing how to encode various types of linguistic information in TEI P5,
- examples of texts encoded at various levels.

Based on the experiences of NCP:

- TEI is a very reasonable choice for encoding various kinds of corpora,
- what's missing is specific examples of TEI P5-encoded corpora.

Within NCP, coming soon (<http://nkjp.pl/>):

- articles showing how to encode various types of linguistic information in TEI P5,
- examples of texts encoded at various levels.

Thank you for your attention!

National Corpus of Polish

<http://nkjp.pl/>

Burnard, L. and Bauman, S., editors (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford.
<http://www.tei-c.org/Guidelines/P5/>.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 825–830, Athens. ELRA.