Linguistic Annotation of the National Corpus of Polish

Adam Przepiórkowski

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

NKJP

FDSL 9

9 December 2011, Göttingen

National Corpus of Polish (12.2007–06.2011)

Pol. *Narodowy Korpus Języka Polskiego* (NKJP).

Institutions:

- Institute of Computer Science PAS (Warsaw; PI: Adam Przepiórkowski; coordinator) — cf. the IPI PAN Corpus of Polish;
- Institute of Polish Language PAS (Cracow; PI: Rafał Górski);
- PWN Publisher (Warsaw; PI: Marek Łaziński) — cf. the PWN Corpus of Polish;
- University of Łódź (PI: Barbara Lewandowska-Tomaszczyk) — cf. the PELCRA Corpus of Polish.

Some features:

- 1.5 billion words,
- 250-million-word balanced subcorpus,
- 1-million-word balanced subcorpus manually corrected,
- multiple levels of annotation.

National Corpus of Polish (12.2007–06.2011)

Pol. *Narodowy Korpus Języka Polskiego* (NKJP).

Institutions:

- Institute of Computer Science PAS (Warsaw; PI: Adam Przepiórkowski; coordinator) — cf. the IPI PAN Corpus of Polish;
- Institute of Polish Language PAS (Cracow; PI: Rafał Górski);
- PWN Publisher (Warsaw; PI: Marek Łaziński) — cf. the PWN Corpus of Polish;
- University of Łódź (PI: Barbara Lewandowska-Tomaszczyk) — cf. the PELCRA Corpus of Polish.

Some features:

- 1.5 billion words,
- 250-million-word balanced subcorpus,
- 1-million-word balanced subcorpus manually corrected,
- multiple levels of annotation.

National Corpus of Polish (12.2007–06.2011)

Pol. *Narodowy Korpus Języka Polskiego* (NKJP).

Institutions:

- Institute of Computer Science PAS (Warsaw; PI: Adam Przepiórkowski; coordinator) — cf. the IPI PAN Corpus of Polish;
- Institute of Polish Language PAS (Cracow; PI: Rafał Górski);
- PWN Publisher (Warsaw; PI: Marek Łaziński) — cf. the PWN Corpus of Polish;
- University of Łódź (PI: Barbara Lewandowska-Tomaszczyk) — cf. the PELCRA Corpus of Polish.

Some features:

- 1.5 billion words,
- 250-million-word balanced subcorpus,
- 1-million-word balanced subcorpus manually corrected,
- multiple levels of annotation.

National Corpus of Polish (12.2007–06.2011)

NKJP

Pol. *Narodowy Korpus Języka Polskiego* (NKJP).

Institutions:

- Institute of Computer Science PAS (Warsaw; PI: Adam Przepiórkowski; coordinator) — cf. the IPI PAN Corpus of Polish;
- Institute of Polish Language PAS (Cracow; PI: Rafał Górski);
- PWN Publisher (Warsaw; PI: Marek Łaziński) — cf. the PWN Corpus of Polish;
- University of Łódź (PI: Barbara Lewandowska-Tomaszczyk) — cf. the PELCRA Corpus of Polish.

Some features:

- 1.5 billion words,
- 250-million-word balanced subcorpus,
- 1-million-word balanced subcorpus manually corrected,
- multiple levels of annotation.

Balancing act

Aim for the **readership**-balanced subcorpora:

- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

Classification also according to channels (press, book, spoken, etc.) – see Górski and Łaziński 2011a,b in the forthcoming NKJP book.

Balancing act

Aim for the **readership**-balanced subcorpora:

- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

Classification also according to channels (press, book, spoken, etc.) – see Górski and Łaziński 2011a,b in the forthcoming NKJP book.

Balancing act

Aim for the **readership**-balanced subcorpora:

- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

Classification also according to channels (press, book, spoken, etc.) – see Górski and Łaziński 2011a,b in the forthcoming NKJP book.

Balancing act

Aim for the **readership**-balanced subcorpora:

- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

Classification also according to channels (press, book, spoken, etc.) – see Górski and Łaziński 2011a,b in the forthcoming NKJP book.

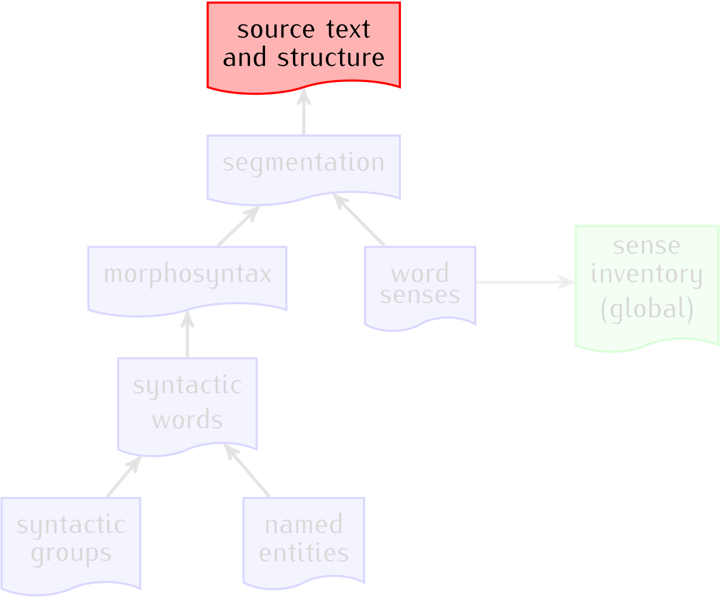
Balancing act

Aim for the **readership**-balanced subcorpora:

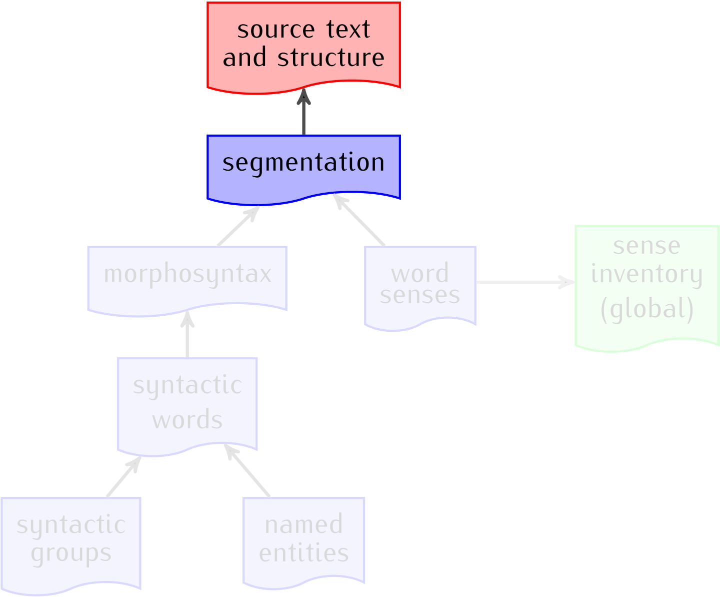
- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

Classification also according to channels (press, book, spoken, etc.) – see Górski and Łaziński 2011a,b in the forthcoming NKJP book.

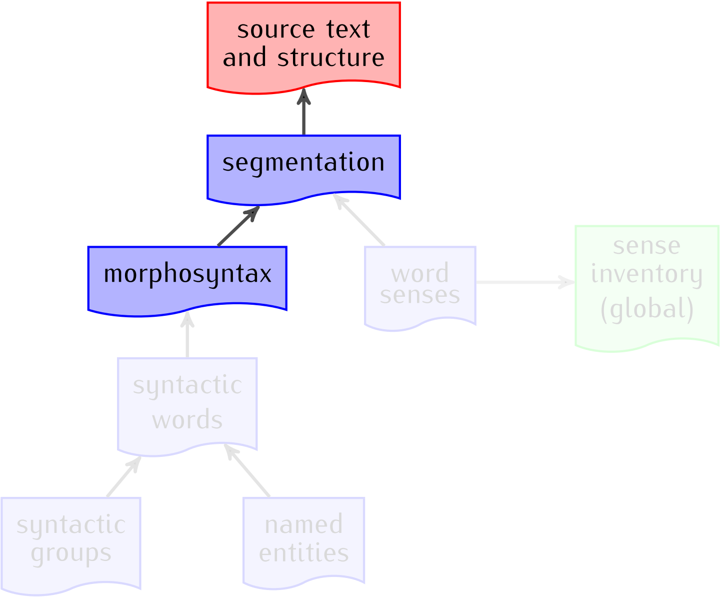
Levels of annotation



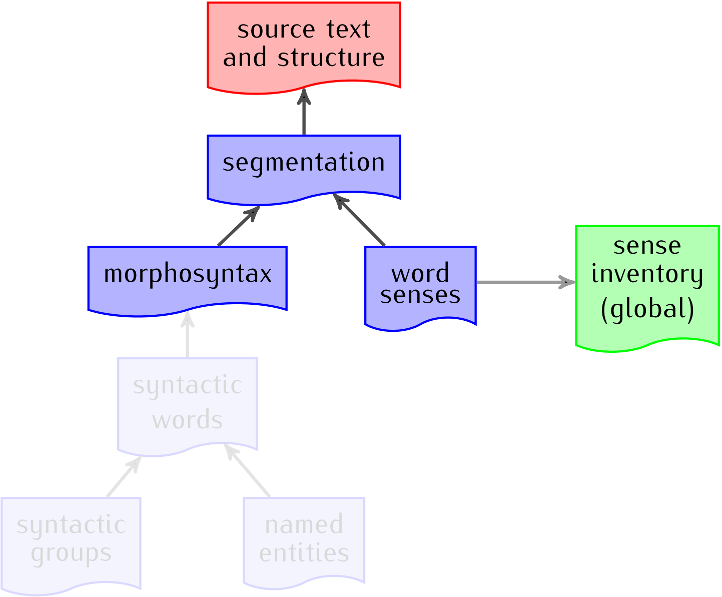
Levels of annotation



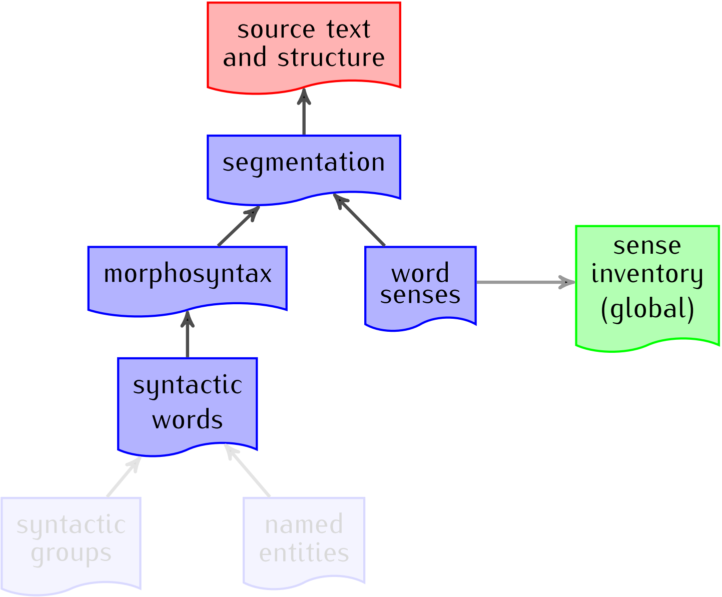
Levels of annotation



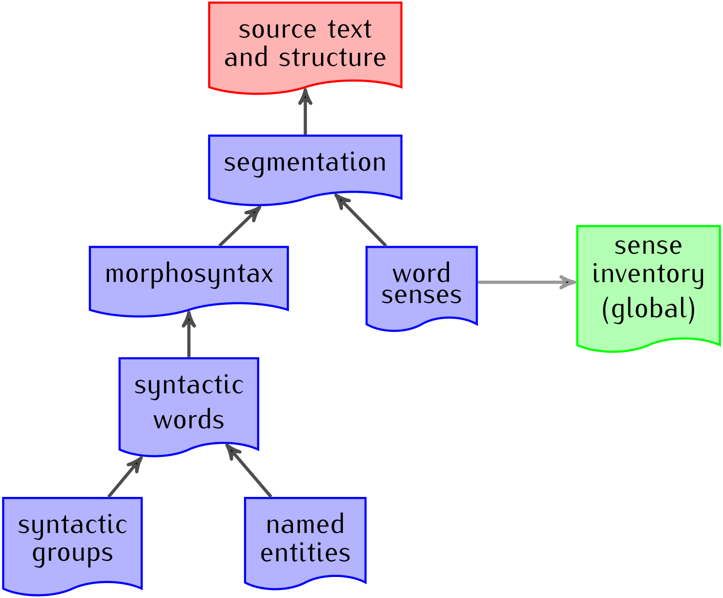
Levels of annotation



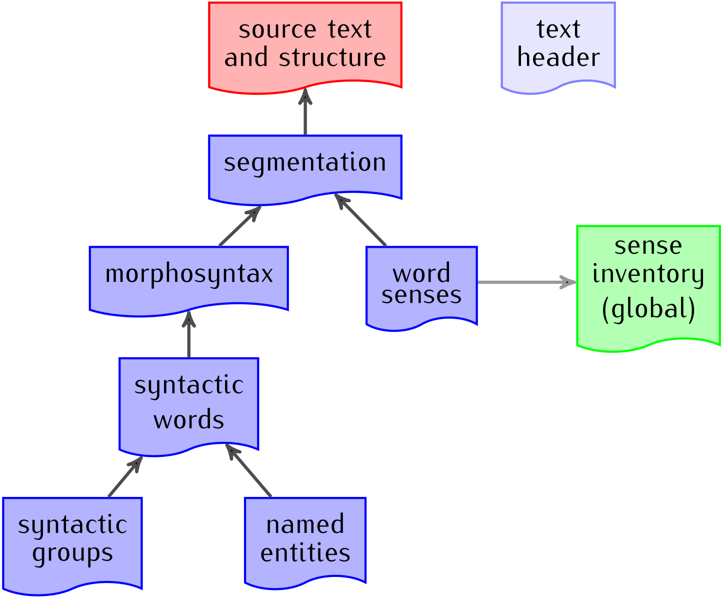
Levels of annotation



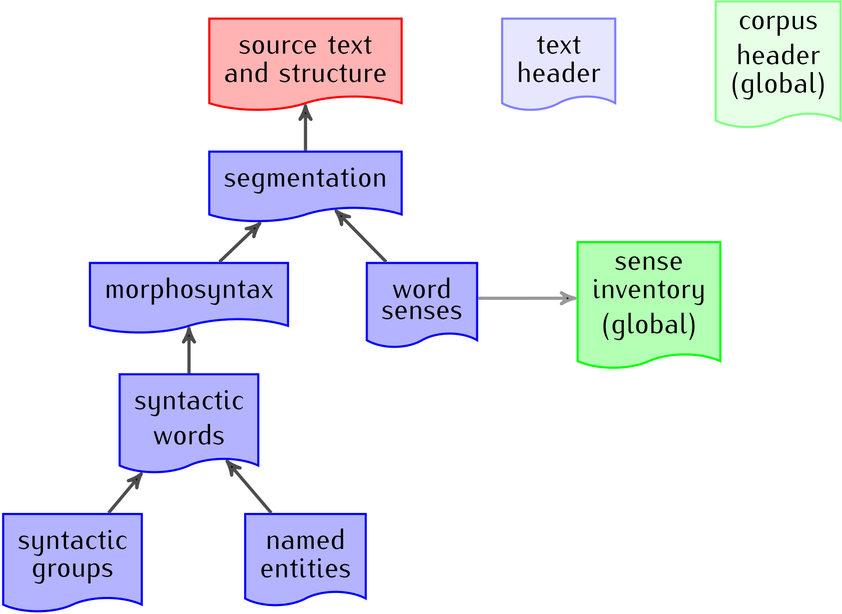
Levels of annotation



Levels of annotation



Levels of annotation



Segmentation 1

1. **Sentence-level segmentation**
   - mostly straightforward,
   - when in doubt, prefer longer sentences, e.g.:

‘Chodź	tu!’	– powiedział	Janek.
come	here	said	Janek

2. **Word-level segmentation**
   - segments are no longer than orthographic words,
   - they don’t overlap,
   - they are contiguous.

Word-level segments are assigned morphosyntactic tags, so the tagset depends on segmentation!

Segmentation 1

1. **Sentence-level segmentation**
   - mostly straightforward,
   - when in doubt, prefer longer sentences, e.g.:

‘Chodź	tu!’	–	powiedział	Janek.
come	here		said	Janek

2. **Word-level segmentation**
   - segments are no longer than orthographic words,
   - they don’t overlap,
   - they are contiguous.

Word-level segments are assigned morphosyntactic tags, so the tagset depends on segmentation!

Segmentation 1

1. **Sentence-level segmentation**
   - mostly straightforward,
   - when in doubt, prefer longer sentences, e.g.:

‘Chodź	tu!’	– powiedział	Janek.
come	here	said	Janek

2. **Word-level segmentation**
   - segments are no longer than orthographic words,
   - they don’t overlap,
   - they are contiguous.

Word-level segments are assigned morphosyntactic tags, so the tagset depends on segmentation!

Segmentation 2

Some consequences:

- 5 segments (including punctuation):

Będę	szedł	i	gwizdał.
will.1.SG	walk	and	whistle

- 4 segments:

Gwizdalibyśmy.
Gwizdali|by|śmy|.
we would (have) whistle(d)

because:

Długośmy	gwizdali.
long–1.PL	whistle.PAST

and:

byśmy	gwizdali
we would (have)	whistle(d)

Segmentation 2

Some consequences:

- 5 segments (including punctuation):

Będę	szedł	i	gwizdał.
will.1.SG	walk	and	whistle

- 4 segments:

Gwizdalibyśmy.
Gwizdali|by|śmy|.
we would (have) whistle(d)

because:

Długośmy	gwizdali.
long-1.PL	whistle.PAST

and:

byśmy	gwizdali
we would (have)	whistle(d)

Segmentation 2

Some consequences:

- 5 segments (including punctuation):

Będę	szedł	i	gwizdał.
will.1.SG	walk	and	whistle

- 4 segments:

Gwizdalibyśmy.
Gwizdali|by|śmy|.
we would (have) whistle(d)

because:

Długośmy	gwizdali.
long-1.PL	whistle.PAST

and:

byśmy	gwizdali
we would (have)	whistle(d)

Segmentation 2

Some consequences:

- 5 segments (including punctuation):

Będę	szedł	i	gwizdał.
will.1.SG	walk	and	whistle

- 4 segments:

Gwizdalibyśmy.
Gwizdali|by|śmy|.
we would (have) whistle(d)

because:

Długośmy	gwizdali.
long-1.PL	whistle.PAST

and:

byśmy	gwizdali
we would (have)	whistle(d)

Word Senses

Rather preliminary work:

- a little over 100 lexemes,
- all very frequent and homonymous,
- split into coarse-grained meanings.

Treated as empirical material for testing Word Sense Disambiguation systems, rather than as a full-fledged annotation level.

More: papers by Kopeć, Młodzki and Przepiórkowski in Slavicorp 1 publication (special issue of *Prace Filologiczne*, 2012) and in the forthcoming NKJP book.

Word Senses

Rather preliminary work:

- a little over 100 lexemes,
- all very frequent and homonymous,
- split into coarse-grained meanings.

Treated as empirical material for testing Word Sense Disambiguation systems, rather than as a full-fledged annotation level.

More: papers by Kopeć, Młodzki and Przepiórkowski in Slavicorp 1 publication (special issue of *Prace Filologiczne*, 2012) and in the forthcoming NKJP book.

Word Senses

Rather preliminary work:

- a little over 100 lexemes,
- all very frequent and homonymous,
- split into coarse-grained meanings.

Treated as empirical material for testing Word Sense Disambiguation systems, rather than as a full-fledged annotation level.

More: papers by Kopeć, Młodzki and Przepiórkowski in Slavicorp 1 publication (special issue of *Prace Filologiczne*, 2012) and in the forthcoming NKJP book.

Morphosyntax 1

Each segment annotated with:

- lemma,
- grammatical class (roughly, part of speech),
- appropriate grammatical categories (case, gender, etc.).

A conservative modification of the IPI PAN Tagset
(cf. Przepiórkowski 2004, 2009).

Preserved:

- flexemic approach to grammatical classes (36),
- detailed grammatical categories,
- grammatical approach to tagset, with (almost) no recourse to semantics or pragmatics.

All analyses preserved, the ones correct in the context marked as such.

Morphosyntax 1

Each segment annotated with:

- lemma,
- grammatical class (roughly, part of speech),
- appropriate grammatical categories (case, gender, etc.).

A conservative modification of the IPI PAN Tagset
(cf. Przepiórkowski 2004, 2009).

Preserved:

- flexemic approach to grammatical classes (36),
- detailed grammatical categories,
- grammatical approach to tagset, with (almost) no recourse to semantics or pragmatics.

All analyses preserved, the ones correct in the context marked as such.

Morphosyntax 1

Each segment annotated with:

- lemma,
- grammatical class (roughly, part of speech),
- appropriate grammatical categories (case, gender, etc.).

A conservative modification of the IPI PAN Tagset
(cf. Przepiórkowski 2004, 2009).

Preserved:

- flexemic approach to grammatical classes (36),
- detailed grammatical categories,
- grammatical approach to tagset, with (almost) no recourse to semantics or pragmatics.

All analyses preserved, the ones correct in the context marked as such.

Each segment annotated with:

- lemma,
- grammatical class (roughly, part of speech),
- appropriate grammatical categories (case, gender, etc.).

A conservative modification of the IPI PAN Tagset
(cf. Przepiórkowski 2004, 2009).

Preserved:

- flexemic approach to grammatical classes (36),
- detailed grammatical categories,
- grammatical approach to tagset, with (almost) no recourse to semantics or pragmatics.

All analyses preserved, the ones correct in the context marked as such.

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;
- postprepositionality: *niego* vs. *jego*, *go* vs.

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;
- postprepositionality: *niego* vs. *jego*, *go* vs. *-ń*;

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;
- postprepositionality: *niego* vs. *jego*, *go* vs. *-ń*;
- how about *profesory (przyszły)* vs. *profesorowie (przyszli)*?

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;
- postprepositionality: *niego* vs. *jego*, *go* vs. *–ń*;
- how about *profesory (przyszły)* vs. *profesorowie (przyszli)*?
  - here: two different grammatical classes (SUBST and DEPR),
  - an additional grammatical category also possible;

Morphosyntax 2

Some categories:

- gender: 5 values, i.e., `m1`, `m2`, `m3`, `n`, `f` (Mańczak, 1956);
- accentability: *go* vs. *jego*, *ci* vs. *cię*, etc.;
- postprepositionality: *niego* vs. *jego*, *go* vs. *-ń*;
- how about *profesory (przyszły)* vs. *profesorowie (przyszli)*?
  - here: two different grammatical classes (SUBST and DEPR),
  - an additional grammatical category also possible;
- accomodability: *dwaj (faceci)* vs. *dwóch (facetów)* in the subject position (i.e., both supposedly nominative).

Morphosyntax 3

Examples:

siano	`imps:imperf`
	`subst:sg:nom:n`
	`subst:sg:acc:n`
	`subst:sg:voc:n`
siane	`ppas:sg:nom:n:imperf:aff`
	`ppas:sg:acc:n:imperf:aff`
	`ppas:pl:nom:m2:imperf:aff`
	. . . (10 interpretations altogether)
śmy	`aglt:pl:pri:imperf:nwok`
jego	`ppron3:sg:gen:m1:ter:akc:npraep`
	`ppron3:sg:gen:m2:ter:akc:npraep`
	`ppron3:sg:gen:m3:ter:akc:npraep`
	`ppron3:sg:gen:n:ter:akc:npraep`
	`ppron3:sg:acc:m1:ter:akc:npraep`
	`ppron3:sg:acc:m2:ter:akc:npraep`
	`ppron3:sg:acc:m3:ter:akc:npraep`

Morphosyntax 3

Examples:

siano	`imps:imperf`
	`subst:sg:nom:n`
	`subst:sg:acc:n`
	`subst:sg:voc:n`

siane	`ppas:sg:nom:n:imperf:aff`
	`ppas:sg:acc:n:imperf:aff`
	`ppas:pl:nom:m2:imperf:aff`
	... (10 interpretations altogether)

śmy	`aglt:pl:pri:imperf:nwok`

jego	`ppron3:sg:gen:m1:ter:akc:npraep`
	`ppron3:sg:gen:m2:ter:akc:npraep`
	`ppron3:sg:gen:m3:ter:akc:npraep`
	`ppron3:sg:gen:n:ter:akc:npraep`
	`ppron3:sg:acc:m1:ter:akc:npraep`
	`ppron3:sg:acc:m2:ter:akc:npraep`
	`ppron3:sg:acc:m3:ter:akc:npraep`

Morphosyntax 3

Examples:

siano	`imps:imperf`
	`subst:sg:nom:n`
	`subst:sg:acc:n`
	`subst:sg:voc:n`

siane	`ppas:sg:nom:n:imperf:aff`
	`ppas:sg:acc:n:imperf:aff`
	`ppas:pl:nom:m2:imperf:aff`
	. . . (10 interpretations altogether)

śmy	`aglt:pl:pri:imperf:nwok`

jego	`ppron3:sg:gen:m1:ter:akc:npraep`
	`ppron3:sg:gen:m2:ter:akc:npraep`
	`ppron3:sg:gen:m3:ter:akc:npraep`
	`ppron3:sg:gen:n:ter:akc:npraep`
	`ppron3:sg:acc:m1:ter:akc:npraep`
	`ppron3:sg:acc:m2:ter:akc:npraep`
	`ppron3:sg:acc:m3:ter:akc:npraep`

Morphosyntax 3

Examples:

siano	`imps:imperf`
	`subst:sg:nom:n`
	`subst:sg:acc:n`
	`subst:sg:voc:n`
siane	`ppas:sg:nom:n:imperf:aff`
	`ppas:sg:acc:n:imperf:aff`
	`ppas:pl:nom:m2:imperf:aff`
	... (10 interpretations altogether)
śmy	`aglt:pl:pri:imperf:nwok`
jego	`ppron3:sg:gen:m1:ter:akc:npraep`
	`ppron3:sg:gen:m2:ter:akc:npraep`
	`ppron3:sg:gen:m3:ter:akc:npraep`
	`ppron3:sg:gen:n:ter:akc:npraep`
	`ppron3:sg:acc:m1:ter:akc:npraep`
	`ppron3:sg:acc:m2:ter:akc:npraep`
	`ppron3:sg:acc:m3:ter:akc:npraep`

Syntactic Words

Unlike word-level segments:

- may overlap,
- may be non-contiguous.

Motivation:

- “traditional” words (including analytic forms, reflexive verbs, etc.),
- mediation between fine-grained word-level segmentation and syntax proper.

For example, the two reflexive verbs in:

Bał	się	tam	odezwać.
feared.3.SG.PAST	Refl	there	answer.INF

Syntactic Words

Unlike word-level segments:

- may overlap,
- may be non-contiguous.

Motivation:

- “traditional” words (including analytic forms, reflexive verbs, etc.),
- mediation between fine-grained word-level segmentation and syntax proper.

For example, the two reflexive verbs in:

Bał	się	tam	odezwać.
feared.3.SG.PAST	Refl	there	answer.INF

Syntactic Words

Unlike word-level segments:

- may overlap,
- may be non-contiguous.

Motivation:

- “traditional” words (including analytic forms, reflexive verbs, etc.),
- mediation between fine-grained word-level segmentation and syntax proper.

For example, the two reflexive verbs in:

Bał	się	tam	odezwać.
feared.3.SG.PAST	Refl	there	answer.INF

Syntactic Groups

Shallow syntactic analysis, i.e.:

- recognition of basic NPs, PPs, etc.,
- without resolving attachment ambiguities,
- or necessarily constructing full trees (a separate treebank project is running now).

Syntactic groups:

- are annotated for syntactic and semantic heads,
- and group type.

More on syntactic words and groups: Głowińska and Przepiórkowski 2010 and papers by Katarzyna Głowińska in Slavicorp 1 publication and in the forthcoming NKJP book.

Syntactic Groups

Shallow syntactic analysis, i.e.:

- recognition of basic NPs, PPs, etc.,
- without resolving attachment ambiguities,
- or necessarily constructing full trees (a separate treebank project is running now).

Syntactic groups:

- are annotated for syntactic and semantic heads,
- and group type.

More on syntactic words and groups: Głowińska and Przepiórkowski 2010 and papers by Katarzyna Głowińska in Slavicorp 1 publication and in the forthcoming NKJP book.

Syntactic Groups

Shallow syntactic analysis, i.e.:

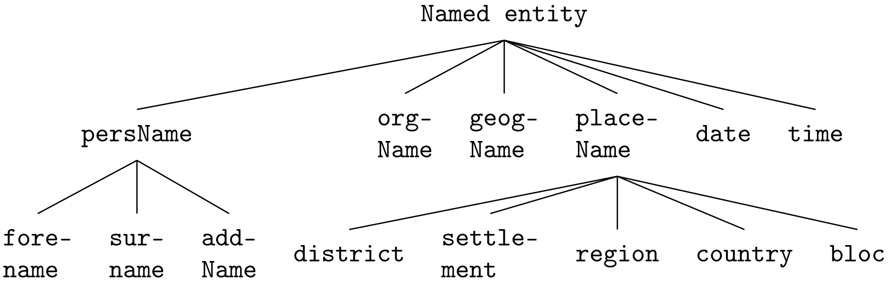
- recognition of basic NPs, PPs, etc.,
- without resolving attachment ambiguities,
- or necessarily constructing full trees (a separate treebank project is running now).

Syntactic groups:

- are annotated for syntactic and semantic heads,
- and group type.

More on syntactic words and groups: Głowińska and Przepiórkowski 2010 and papers by Katarzyna Głowińska in Slavicorp 1 publication and in the forthcoming NKJP book.

Named Entities

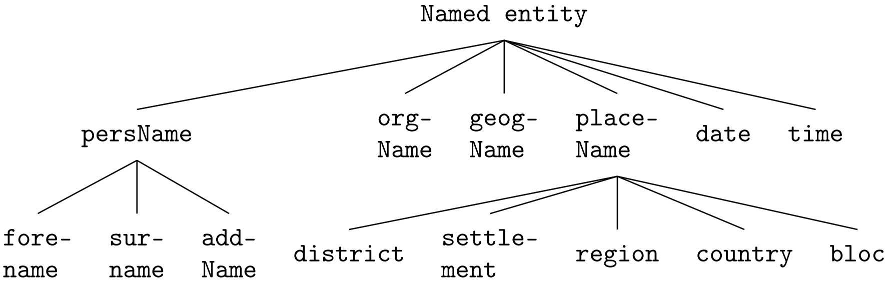


Named Entities may:

- be non-contiguous,
- overlap,
- be nested.

More on NEs: Savary *et al.* 2010 and papers co-authored by Agata Savary in the forthcoming NKJP book.

Named Entities

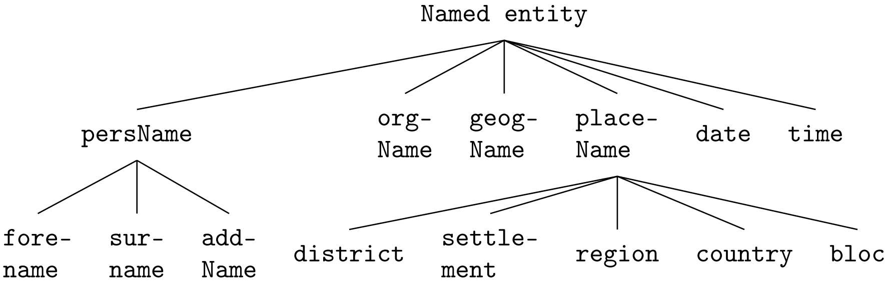


Named Entities may:

- be non-contiguous,
- overlap,
- be nested.

More on NEs: Savary *et al.* 2010 and papers co-authored by Agata Savary in the forthcoming NKJP book.

Named Entities



Named Entities may:

- be non-contiguous,
- overlap,
- be nested.

More on NEs: Savary *et al.* 2010 and papers co-authored by Agata Savary in the forthcoming NKJP book.

Manual annotation (morphosyntax) 1

Texts are morphologically analysed by Morfeusz (Woliński, 2006; Saloni *et al.*, 2007), which also provides word-level segmentation, but:

- Morfeusz assigns all possible morphosyntactic tags, to be disambiguated manually,
- there are omissions,
- there may be errors (even when only one tag assigned).

So, for each segment, the annotator:

- selects the right lemma and tag, if it's among those proposed,
- adds a new lemma and tag, otherwise.

Manual annotation (morphosyntax) 1

NKJP

Texts are morphologically analysed by Morfeusz (Woliński, 2006; Saloni *et al.*, 2007), which also provides word-level segmentation, but:

- Morfeusz assigns all possible morphosyntactic tags, to be disambiguated manually,
- there are omissions,
- there may be errors (even when only one tag assigned).

So, for each segment, the annotator:

- selects the right lemma and tag, if it's among those proposed,
- adds a new lemma and tag, otherwise.

Manual annotation (morphosyntax) 1

Texts are morphologically analysed by Morfeusz (Woliński, 2006; Saloni *et al.*, 2007), which also provides word-level segmentation, but:

- Morfeusz assigns all possible morphosyntactic tags, to be disambiguated manually,
- there are omissions,
- there may be errors (even when only one tag assigned).

So, for each segment, the annotator:

- selects the right lemma and tag, if it's among those proposed,
- adds a new lemma and tag, otherwise.

Manual annotation (morphosyntax) 1

Texts are morphologically analysed by Morfeusz (Woliński, 2006; Saloni *et al.*, 2007), which also provides word-level segmentation, but:

- Morfeusz assigns all possible morphosyntactic tags, to be disambiguated manually,
- there are omissions,
- there may be errors (even when only one tag assigned).

So, for each segment, the annotator:

- selects the right lemma and tag, if it's among those proposed,
- adds a new lemma and tag, otherwise.

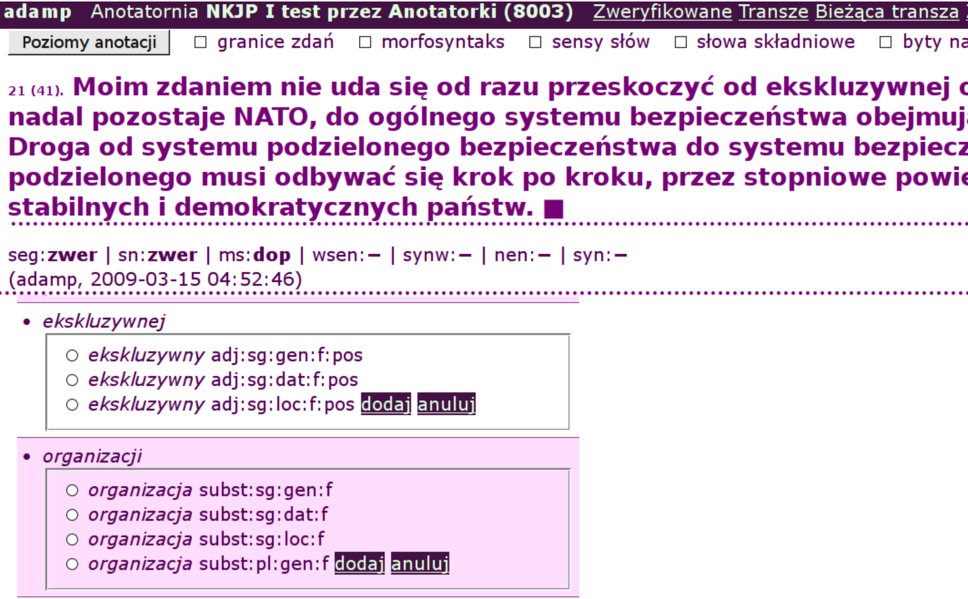
Manual annotation (morphosyntax) 1

Texts are morphologically analysed by Morfeusz (Woliński, 2006; Saloni *et al.*, 2007), which also provides word-level segmentation, but:

- Morfeusz assigns all possible morphosyntactic tags, to be disambiguated manually,
- there are omissions,
- there may be errors (even when only one tag assigned).

So, for each segment, the annotator:

- selects the right lemma and tag, if it's among those proposed,
- adds a new lemma and tag, otherwise.



adamp Anotatornia **NKJP I test przez Anotatorki (8003)** Zweryfikowane Transze Bieżąca transza

Poziomy anotacji □ granice zdań □ morfosyntaks □ sensy słów □ słowa składniowe □ byty na

21 (41). **Moim zdaniem nie uda się od razu przeskoczyć od ekskluzywnej o nadal pozostaje NATO, do ogólnego systemu bezpieczeństwa obejmuj Droga od systemu podzielonego bezpieczeństwa do systemu bezpiecz podzielonego musi odbywać się krok po kroku, przez stopniowe powie stabilnych i demokratycznych państw.** ■

seg:**zwer** | sn:**zwer** | ms:**dop** | wsen:**–** | synw:**–** | nen:**–** | syn:**–**
(adamp, 2009-03-15 04:52:46)

- *ekskluzywnej*
  - ○ *ekskluzywny* adj:sg:gen:f:pos
  - ○ *ekskluzywny* adj:sg:dat:f:pos
  - ○ *ekskluzywny* adj:sg:loc:f:pos dodaj anuluj
- *organizacji*
  - ○ *organizacja* subst:sg:gen:f
  - ○ *organizacja* subst:sg:dat:f
  - ○ *organizacja* subst:sg:loc:f
  - ○ *organizacja* subst:pl:gen:f dodaj anuluj

Manual annotation (morphosyntax) 2

At each level:

- annotation is introduced independently by two annotators,
- if they fully agree:
  - the annotation is accepted and
  - the paragraph is sent to the next level of annotation (by the same annotators),
- otherwise the annotators are informed:
  - about the existence and location of the discrepancy,
  - but not about the annotation of the other annotator,
- and they may change their own annotation or stick to it;
- after this round, if there still is a conflict:
  - it must be resolved by the super-annotator,
  - the annotatation guidelines should be modified accordingly.

Manual annotation (morphosyntax) 2

At each level:

- annotation is introduced independently by two annotators,
- if they fully agree:
  - the annotation is accepted and
  - the paragraph is sent to the next level of annotation (by the same annotators),
- otherwise the annotators are informed:
  - about the existence and location of the discrepancy,
  - but not about the annotation of the other annotator,
- and they may change their own annotation or stick to it;
- after this round, if there still is a conflict:
  - it must be resolved by the super-annotator,
  - the annotatation guidelines should be modified accordingly.

Manual annotation (morphosyntax) 2

At each level:

- annotation is introduced independently by two annotators,
- if they fully agree:
  - the annotation is accepted and
  - the paragraph is sent to the next level of annotation (by the same annotators),
- otherwise the annotators are informed:
  - about the existence and location of the discrepancy,
  - but not about the annotation of the other annotator,
- and they may change their own annotation or stick to it;
- after this round, if there still is a conflict:
  - it must be resolved by the super-annotator,
  - the annotatation guidelines should be modified accordingly.

Manual annotation (morphosyntax) 2

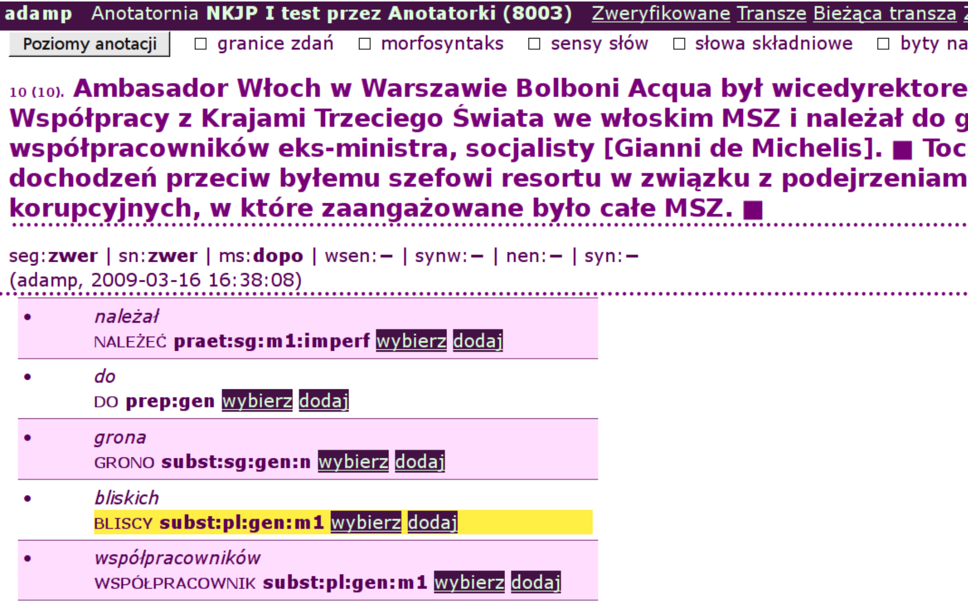
At each level:

- annotation is introduced independently by two annotators,
- if they fully agree:
  - the annotation is accepted and
  - the paragraph is sent to the next level of annotation (by the same annotators),
- otherwise the annotators are informed:
  - about the existence and location of the discrepancy,
  - but not about the annotation of the other annotator,
- and they may change their own annotation or stick to it;
- after this round, if there still is a conflict:
  - it must be resolved by the super-annotator,
  - the annotatation guidelines should be modified accordingly.

Manual annotation (morphosyntax) 2

At each level:

- annotation is introduced independently by two annotators,
- if they fully agree:
  - the annotation is accepted and
  - the paragraph is sent to the next level of annotation (by the same annotators),
- otherwise the annotators are informed:
  - about the existence and location of the discrepancy,
  - but not about the annotation of the other annotator,
- and they may change their own annotation or stick to it;
- after this round, if there still is a conflict:
  - it must be resolved by the super-annotator,
  - the annotatation guidelines should be modified accordingly.

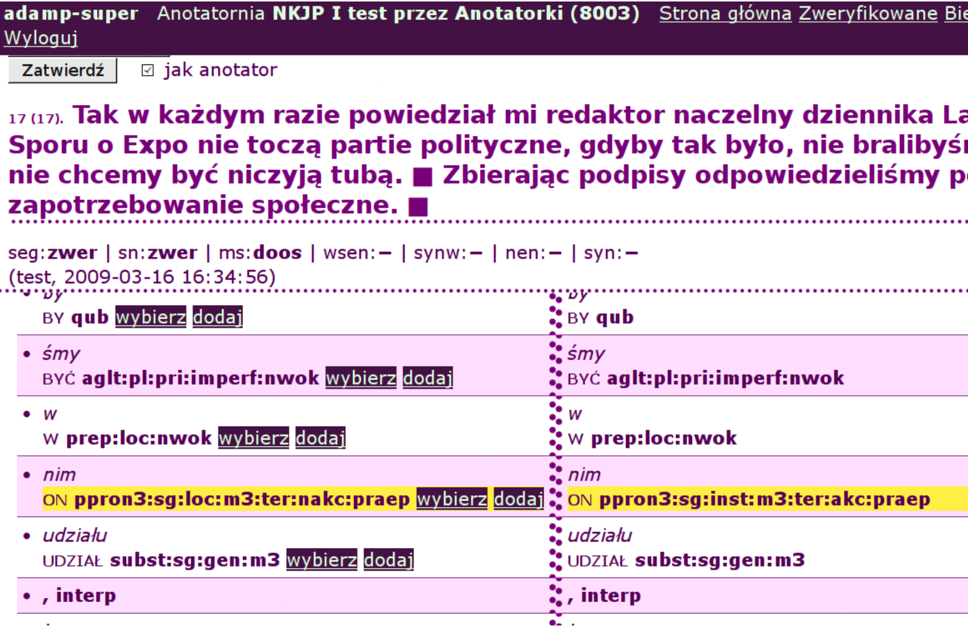


adamp Anotatornia **NKJP I test przez Anotatorki (8003)** Zweryfikowane Transze Bieżąca transza Z

10 (10). **Ambasador Włoch w Warszawie Bolboni Acqua był wicedyrektore Współpracy z Krajami Trzeciego Świata we włoskim MSZ i należał do g współpracowników eks-ministra, socjalisty [Gianni de Michelis]. ■ Toc dochodzeń przeciw byłemu szefowi resortu w związku z podejrzeniam korupcyjnych, w które zaangażowane było całe MSZ. ■**

seg:**zwer** | sn:**zwer** | ms:**dopo** | wsen:**–** | synw:**–** | nen:**–** | syn:**–**
(adamp, 2009-03-16 16:38:08)

- *należał*
  NALEŻEĆ **praet:sg:m1:imperf** wybierz dodaj
- *do*
  DO **prep:gen** wybierz dodaj
- *grona*
  GRONO **subst:sg:gen:n** wybierz dodaj
- *bliskich*
  BLISCY **subst:pl:gen:m1** wybierz dodaj
- *współpracowników*
  WSPÓŁPRACOWNIK **subst:pl:gen:m1** wybierz dodaj



adamp-super Anotatornia **NKJP I test przez Anotatorki (8003)** Strona główna Zweryfikowane Bie
Wyloguj

Zatwierdź ☑ jak anotator

17 (17). **Tak w każdym razie powiedział mi redaktor naczelny dziennika La**
Sporu o Expo nie toczą partie polityczne, gdyby tak było, nie bralibyś
nie chcemy być niczyją tubą. ■ Zbierając podpisy odpowiedzieliśmy p
zapotrzebowanie społeczne. ■

seg:**zwer** | sn:**zwer** | ms:**doos** | wsen:**–** | synw:**–** | nen:**–** | syn:**–**
(test, 2009-03-16 16:34:56)

• *by* BY **qub** wybierz dodaj	*by* BY **qub**
• *śmy* BYĆ **aglt:pl:pri:imperf:nwok** wybierz dodaj	*śmy* BYĆ **aglt:pl:pri:imperf:nwok**
• *w* W **prep:loc:nwok** wybierz dodaj	*w* W **prep:loc:nwok**
• *nim* ON **ppron3:sg:loc:m3:ter:nakc:praep** wybierz dodaj	*nim* ON **ppron3:sg:inst:m3:ter:akc:praep**
• *udziału* UDZIAŁ **subst:sg:gen:m3** wybierz dodaj	*udziału* UDZIAŁ **subst:sg:gen:m3**
• **, interp**	**, interp**

XML representation 1

NKJP

How to represent linguistic information in texts?

- XML (obviously),
- the Text Encoding Initiative P5 standard (Burnard and Bauman, 2008):
  - extensively documented (including semantics of elments),
  - very rich (also metadata),
- from a large toolbox offered by TEI, solutions most compatible with other standards selected,
- many publications, including: Przepiórkowski and Bański 2009,
- `http://nlp.ipipan.waw.pl/TEI4NKJP/`.

An example for *Bał się odezwać* at the level of syntactic words on the next slide.

XML representation 1

How to represent linguistic information in texts?

- XML (obviously),
- the Text Encoding Initiative P5 standard (Burnard and Bauman, 2008):
  - extensively documented (including semantics of elments),
  - very rich (also metadata),
- from a large toolbox offered by TEI, solutions most compatible with other standards selected,
- many publications, including: Przepiórkowski and Bański 2009,
- http://nlp.ipipan.waw.pl/TEI4NKJP/.

An example for *Bał się odezwać* at the level of syntactic words on the next slide.

XML representation 1

How to represent linguistic information in texts?

- XML (obviously),
- the Text Encoding Initiative P5 standard (Burnard and Bauman, 2008):
  - extensively documented (including semantics of elments),
  - very rich (also metadata),
- from a large toolbox offered by TEI, solutions most compatible with other standards selected,
- many publications, including: Przepiórkowski and Bański 2009,
- `http://nlp.ipipan.waw.pl/TEI4NKJP/`.

An example for *Bał się odezwać* at the level of syntactic words on the next slide.

XML representation 1

NKJP

How to represent linguistic information in texts?

- XML (obviously),
- the Text Encoding Initiative P5 standard (Burnard and Bauman, 2008):
  - extensively documented (including semantics of elments),
  - very rich (also metadata),
- from a large toolbox offered by TEI, solutions most compatible with other standards selected,
- many publications, including: Przepiórkowski and Bański 2009,
- `http://nlp.ipipan.waw.pl/TEI4NKJP/`.

An example for *Bał się odezwać* at the level of syntactic words on the next slide.

XML representation 1

How to represent linguistic information in texts?

- XML (obviously),
- the Text Encoding Initiative P5 standard (Burnard and Bauman, 2008):
  - extensively documented (including semantics of elments),
  - very rich (also metadata),
- from a large toolbox offered by TEI, solutions most compatible with other standards selected,
- many publications, including: Przepiórkowski and Bański 2009,
- `http://nlp.ipipan.waw.pl/TEI4NKJP/`.

An example for *Bał się odezwać* at the level of syntactic words on the next slide.

XML representation 2

```
<seg xml:id="word13">
 <fs>1</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg17"/> <!-- Bał -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
</seg>
<seg xml:id="word14">
 <fs>2</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
 <ptr target="ann_morphosyntax.xml#seg19"/> <!-- odezwać -->
</seg>
```

Where:

- $\boxed{1} = \begin{bmatrix} \textit{word} & \\ \textsc{orth} & \text{Bał się} \\ \textsc{base} & \text{bać się} \\ \textsc{ctag} & \text{Verbfin} \\ \textsc{msd} & \text{sg:ter:m1:imperf:past:ind:aff:refl} \end{bmatrix}$
- $\boxed{2} = \begin{bmatrix} \textit{word} & \\ \textsc{orth} & \text{się odezwać} \\ \textsc{base} & \text{odezwać się} \\ \textsc{ctag} & \text{Inf} \\ \textsc{msd} & \text{perf:aff:refl} \end{bmatrix}$

XML representation 2

```
<seg xml:id="word13">
 <fs>1</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg17"/> <!-- Bał -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
</seg>
<seg xml:id="word14">
 <fs>2</fs> <!-- (see below) -->
 <ptr target="ann_morphosyntax.xml#seg18"/> <!-- się -->
 <ptr target="ann_morphosyntax.xml#seg19"/> <!-- odezwać -->
</seg>
```

Where:

- $$\boxed{1} = \begin{bmatrix} \textit{word} & \\ \textsc{orth} & \text{Bał się} \\ \textsc{base} & \text{bać się} \\ \textsc{ctag} & \text{Verbfin} \\ \textsc{msd} & \text{sg:ter:m1:imperf:past:ind:aff:refl} \end{bmatrix}$$
- $$\boxed{2} = \begin{bmatrix} \textit{word} & \\ \textsc{orth} & \text{się odezwać} \\ \textsc{base} & \text{odezwać się} \\ \textsc{ctag} & \text{Inf} \\ \textsc{msd} & \text{perf:aff:refl} \end{bmatrix}$$

Some applications

NKJP

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
  - eponymous phrases – published (Czeszewski and Foremniak, 2011),
  - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on "Wörter des Tages";
- educational aid, solving crossword puzzles, and many more.

Some applications

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
  - eponymous phrases – published (Czeszewski and Foremniak, 2011),
  - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on "Wörter des Tages";
- educational aid, solving crossword puzzles, and many more.

Some applications

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
    - eponymous phrases – published (Czeszewski and Foremniak, 2011),
    - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on "Wörter des Tages";
- educational aid, solving crossword puzzles, and many more.

Some applications

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
  - eponymous phrases – published (Czeszewski and Foremniak, 2011),
  - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on "Wörter des Tages";
- educational aid, solving crossword puzzles, and many more.

Some applications

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
  - eponymous phrases – published (Czeszewski and Foremniak, 2011),
  - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on “Wörter des Tages”;
- educational aid, solving crossword puzzles, and many more.

Some applications

- 1-million-word manually annotated subcorpus: training and evaluating various tools (also LREC 2012 Language Library);
- new Polish dictionaries:
  - eponymous phrases – published (Czeszewski and Foremniak, 2011),
  - general dictionary – in preparation (Żmigrodzki *et al.*, 2007);
- theoretical linguistics (Dziwirek and Lewandowska-Tomaszczyk, 2010; Górski, 2008, 2011; Gębka-Wolak, 2011);
- translation studies (Pęzik, 2011a);
- *Words of the Day* (Łaziński and Andrzejczuk, 2011), based on “Wörter des Tages”;
- educational aid, solving crossword puzzles, and many more.

Access

A 1-million-word manually annotated balanced subcorpus of NKJP:

`http://clip.ipipan.waw.pl/LRT/`

Full corpus available at `http://nkjp.pl/` → EN → SEARCH THE CORPUS via 2 search engines:

- Poliqarp (Janus and Przepiórkowski, 2007):
  - very rich query syntax,
  - own binary format,
  - offers access to segmentation and morphosyntax;
- customized Apache Lucene with RDB (Pęzik, 2011b):
  - user friendly,
  - good collocation component,
  - no access to linguistic annotation.

Access

A 1-million-word manually annotated balanced subcorpus of NKJP:

`http://clip.ipipan.waw.pl/LRT/`

Full corpus available at `http://nkjp.pl/` → EN → SEARCH THE CORPUS via 2 search engines:

- Poliqarp (Janus and Przepiórkowski, 2007):
  - very rich query syntax,
  - own binary format,
  - offers access to segmentation and morphosyntax;
- customized Apache Lucene with RDB (Pęzik, 2011b):
  - user friendly,
  - good collocation component,
  - no access to linguistic annotation.

Access

A 1-million-word manually annotated balanced subcorpus of NKJP:

`http://clip.ipipan.waw.pl/LRT/`

Full corpus available at `http://nkjp.pl/` → EN → SEARCH THE CORPUS via 2 search engines:

- Poliqarp (Janus and Przepiórkowski, 2007):
  - very rich query syntax,
  - own binary format,
  - offers access to segmentation and morphosyntax;
- customized Apache Lucene with RDB (Pęzik, 2011b):
  - user friendly,
  - good collocation component,
  - no access to linguistic annotation.

Access

A 1-million-word manually annotated balanced subcorpus of NKJP:

`http://clip.ipipan.waw.pl/LRT/`

Full corpus available at `http://nkjp.pl/` → EN → SEARCH THE CORPUS via 2 search engines:

- Poliqarp (Janus and Przepiórkowski, 2007):
  - very rich query syntax,
  - own binary format,
  - offers access to segmentation and morphosyntax;
- customized Apache Lucene with RDB (Pęzik, 2011b):
  - user friendly,
  - good collocation component,
  - no access to linguistic annotation.

Some search examples

NKJP

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****…*

To ja ***jestem policja****.*

…ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza nad władzą****,* ***jestem principium wszelkiej władzy****,* ***jestem filar społeczeńskiego ładu i porządku****!*

Some search examples

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****…*

To ja ***jestem policja****.*

…ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza nad władzą, jestem principium wszelkiej władzy, jestem filar społeczeńskiego ładu i porządku****!*

Some search examples

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****…*

To ja ***jestem policja****.*

…ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza*** *nad władzą,* ***jestem principium*** *wszelkiej władzy,* ***jestem filar*** *społeczeńskiego ładu i porządku!*

Some search examples

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****...*

To ja ***jestem policja****.*

...ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza nad władzą****,* ***jestem principium wszelkiej władzy****,* ***jestem filar społeczeńskiego ładu i porządku****!*

Some search examples

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****…*

To ja ***jestem policja****.*

…ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza*** *nad władzą,* ***jestem principium*** *wszelkiej władzy,* ***jestem filar*** *społeczeńskiego ładu i porządku!*

Some search examples

```
[orth="wieczorem|rankiem] [pos=conj]
[pos=adj & case=acc]* [pos=subst & case=acc]
```

Trudno w to dziś uwierzyć, ale ten grysik jadłyśmy ***wieczorem i cały następny dzień****.*

```
[base=być & pos="fin" & pers="pri|sec"]
[pos=subst & case~~nom]
```

To prawda, że ***jesteś hycel****…*

To ja ***jestem policja****.*

…ale ja, Wincenty Korab Czartkowski, szlachcic i obywatel, ***jestem władza nad władzą****,* ***jestem principium wszelkiej władzy****,* ***jestem filar społeczeńskiego ładu i porządku****!*

Future (starts today) 1

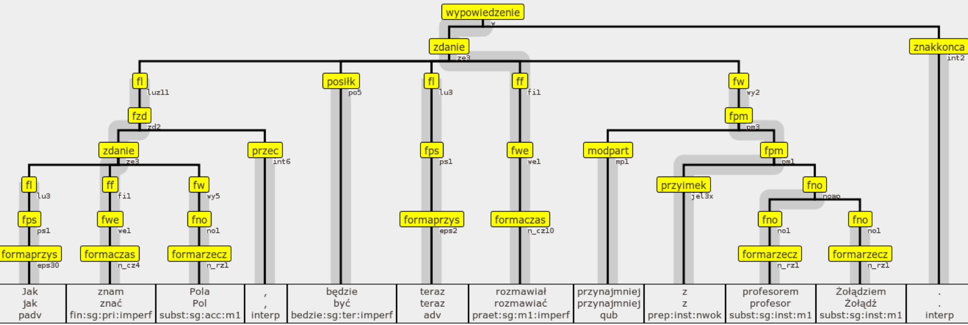
Some resources and research based on NKJP at IPI PAN:

- *Składnica* – the first Polish constituency treebank (`http://zil.ipipan.waw.pl/Składnica`);

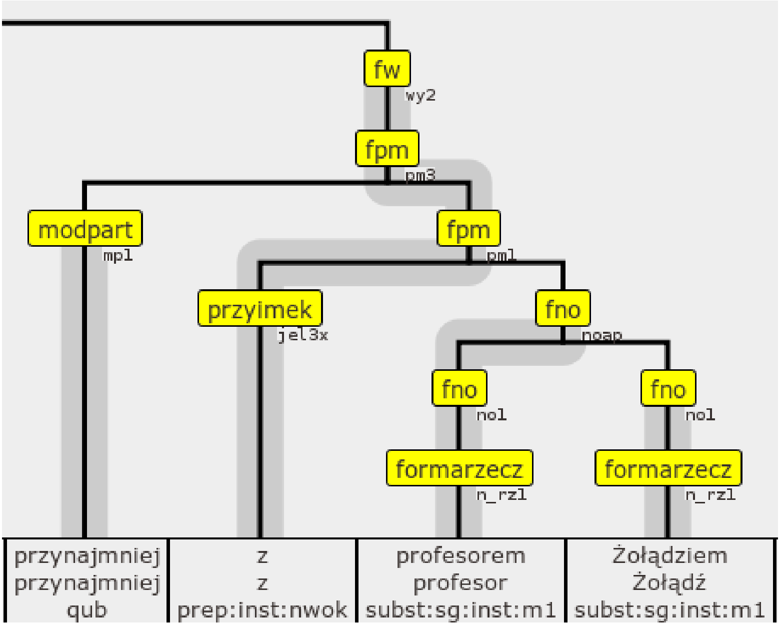
Future (starts today) 1

Some resources and research based on NKJP at IPI PAN:

- *Składnica* – the first Polish constituency treebank (`http://zil.ipipan.waw.pl/Składnica`);



Future (starts today) 2



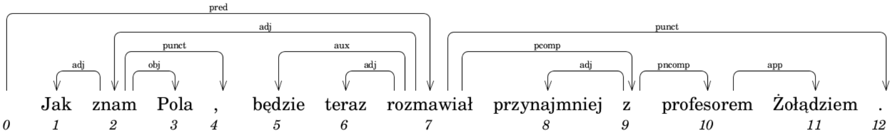
przynajmniej	z	profesorem	Żołądziem
przynajmniej	z	profesor	Żołądź
qub	prep:inst:nwok	subst:sg:inst:m1	subst:sg:inst:m1

Future (starts today) 3

- also a dependency version (same address);

Future (starts today) 3

- also a dependency version (same address);

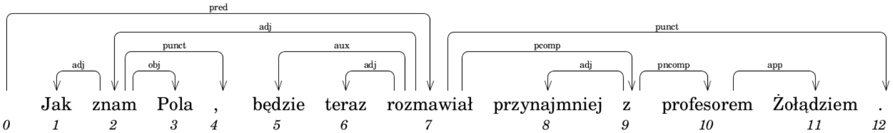


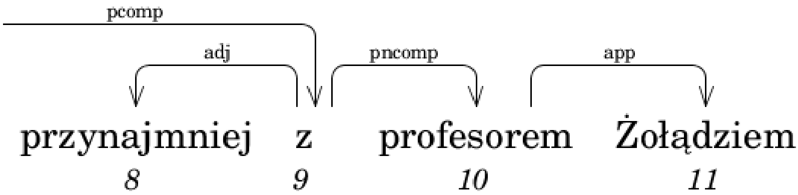
Jak znam Pola , będzie teraz rozmawiał przynajmniej z profesorem Żołądziem .

0 1 2 3 4 5 6 7 8 9 10 11 12

Future (starts today) 3

- also a dependency version (same address);


pred
adj
punct
adj
punct
aux
pcomp
obj
adj
adj
pncomp
app
Jak znam Pola , będzie teraz rozmawiał przynajmniej z profesorem Żołądziem .
0 1 2 3 4 5 6 7 8 9 10 11 12

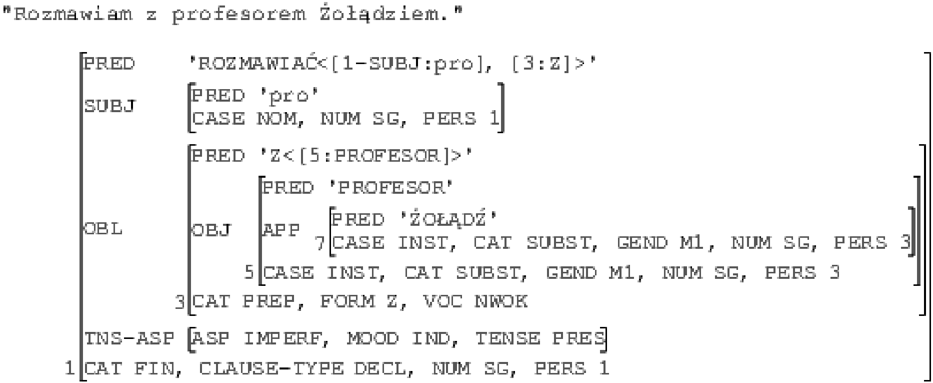


Future (starts today) 4

- a Lexical-Functional Grammar based on the constituency grammar, but with:
    - control and raising,
    - case assignment based on the structural / lexical case dichotomy,
    - negative concord, etc.

Future (starts today) 4

- a Lexical-Functional Grammar based on the constituency grammar, but with:
  - control and raising,
  - case assignment based on the structural / lexical case dichotomy,
  - negative concord, etc.



```
"Rozmawiam z profesorem Żołądziem."

  PRED     'ROZMAWIAĆ<[1-SUBJ:pro], [3:Z]>'
  SUBJ     [PRED 'pro'                  ]
           [CASE NOM, NUM SG, PERS 1    ]
           [PRED 'Z<[5:PROFESOR]>'
           [        PRED 'PROFESOR'
  OBL      [OBJ     APP   [PRED 'ŻOŁĄDŹ'
           [              7 CASE INST, CAT SUBST, GEND M1, NUM SG, PERS 3
           [      5 CASE INST, CAT SUBST, GEND M1, NUM SG, PERS 3
           3 CAT PREP, FORM Z, VOC NWOK
  TNS-ASP  [ASP IMPERF, MOOD IND, TENSE PRES]
1 CAT FIN, CLAUSE-TYPE DECL, NUM SG, PERS 1
```

Thank you for your attention!

`http://nkjp.pl/`

NKJP

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

Czeszewski, M. and Foremniak, K. (2011). *Ludzie i miejsca w języku. Słownik frazeologizmów eponimicznych*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Dziwirek, K. and Lewandowska-Tomaszczyk, B. (2010). *Complex Emotions and Grammatical Mismatches: A Contrastive Corpus-Based Study*. Mouton de Gruyter, Berlin.

Gębka-Wolak, M. (2011). *Pozycje składniowe frazy bezokolicznikowej we współczesnym zdaniu polskim*. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

Głowińska, K. and Przepiórkowski, A. (2010). The design of syntactic annotation levels in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Górski, R. L. (2008). *Diateza nacechowana w polszczyźnie. Studium korpusowe*. Lexis, Cracow.

Górski, R. L. (2011). Zastosowanie korpusów w badaniu gramatyki. In Przepiórkowski *et al.* (2011). Forthcoming.

Górski, R. L. and Łaziński, M. (2011a). Reprezentatywność i zrównoważenie korpusu. In Przepiórkowski *et al.* (2011). Forthcoming.

Górski, R. L. and Łaziński, M. (2011b). Typologia tekstów w NKJP. In Przepiórkowski *et al.* (2011). Forthcoming.

Janus, D. and Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague.

Łaziński, M. and Andrzejczuk, A. (2011). Słowa dnia. In Przepiórkowski *et al.* (2011). Forthcoming.

Mańczak, W. (1956). Ile jest rodzajów w polskim? *Język Polski*, **XXXVI**(2), 116–121.

Pęzik, P. (2011a). NKJP w warsztacie tłumacza. In Przepiórkowski *et al.* (2011). Forthcoming.

Pęzik, P. (2011b). Providing corpus feedback for translators with the PELCRA search engine for NKJP. In S. Goźdź-Roszkowski, editor, *Explorations across Languages and Corpora: PALC 2009*, pages 135–144, Frankfurt am Main. Peter Lang.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version.* Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, pages 138–144, Warsaw.

Przepiórkowski, A. and Bański, P. (2009). Which XML standards for multilevel corpus annotation? In Z. Vetulani, editor, *Proceedings of the 4th Language & Technology Conference*, pages 245–250, Poznań, Poland.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2011). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw. Forthcoming.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007). *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the annotation of named entities in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web*

Mining, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin.

Żmigrodzki, P., Bańko, M., Dunaj, B., and Przybylska, R. (2007). Koncepcja *Wielkiego słownika języka polskiego* — przybliżenie drugie. In P. Żmigrodzki and R. Przybylska, editors, *Nowe studia leksykograficzne*, pages 9–21. Lexis, Cracow.