From the National Corpus of Polish to the Polish Corpus Infrastructure

Maciej Ogrodniczuk

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

SLOVKO 2019

Bratislava, 25 October 2019

Agenda

Three main topics:

- NCP: The National Corpus of Polish (NKJP)
- PCI: The Polish Corpus Infrastructure (PIK)
- DCP: The Diachronic Corpus of Polish (NKDP)

The National Corpus of Polish

Narodowy Korpus Języka Polskiego (NKJP):

- resulted of a nationally funded project carried out between 2007 and 2011
- co-operation of 4 institutions previously involved in corpora collection:
  - Institute of Computer Science, Polish Academy of Sciences (Warsaw; coordinator: Adam Przepiórkowski)
  - Institute of Polish Language, Polish Academy of Sciences (Cracow; Rafał L. Górski)
  - University of Łódź (Barbara Lewandowska-Tomaszczyk, Piotr Pęzik),
  - PWN Scientific Publishers (Warsaw; Mirosław Bańko, Marek Łaziński — now Institute of Polish Language, University of Warsaw)

The National Corpus of Polish

Corpus in (3+2) numbers:

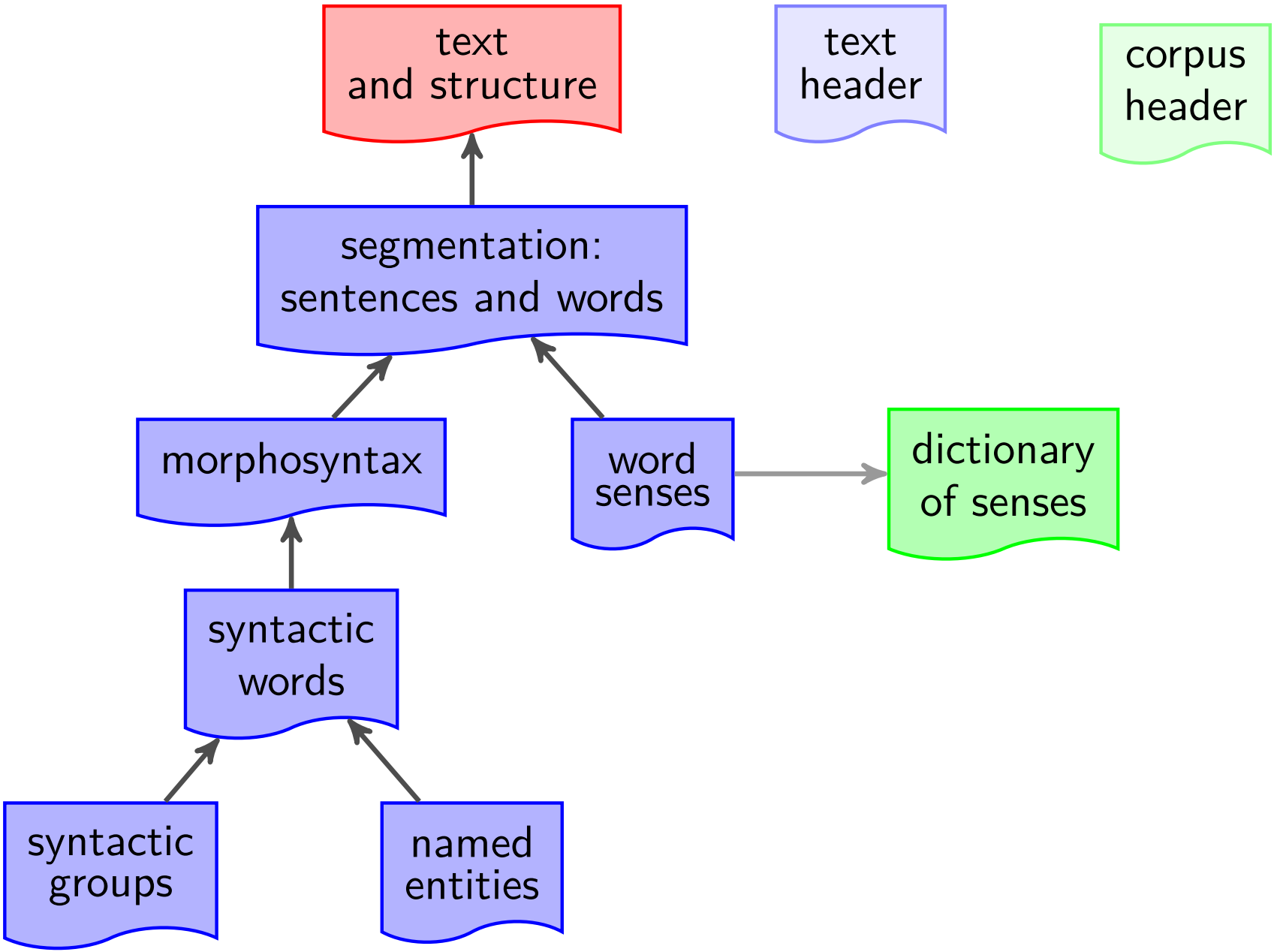
- 1.8B words in total
- balanced automatically-annotated part: 300M words
- balanced manually-annotated part: 1.2M words
- 'distributable' part: 100M words
- Wikipedia part: 140M words

The balanced NCP (NKJP300M)

Percentage of text types:

Daily newspapers	25.0%
Magazines	25.0%
Fiction literature	16.0%
Non-fiction literature	5.5%
Instructive writing and textbooks	5.5%
Spoken – conversational	5.0%
Internet non-interactive	3.5%
Internet interactive	3.5%
Misc. written	3.0%
Spoken from the media	2.0%
Quasi-spoken	2.0%
Academic writing and textbooks	2.0%
Journalistic books	1.0%
Unclassified written	1.0%

Annotation layers



Segmentation

Three levels:

- paragraph-level segmentation
- sentence-level segmentation
- token-level segmentation
  - segments no longer than space-to-space words
  - segments are continuous
  - segments don't overlap

The motivation for segments:

Gwizdalibyśmy. $\rightarrow$ Gwizdali|by|śmy|.
by|śmy gwizdali
długo|śmy gwizdali

Morphosyntax

Each segment carries information on its:

- lemma,
- grammatical class ($\approx$ POS),
- grammatical categories (case, gender etc.)

Morphosyntax

Several examples:

człowieka	`subst:sg:acc:m1`
	`subst:sg:gen:m1`
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`
ułożono	`imps:perf`

Syntactic words

Motivation:

- 'traditional' words (including analytical forms, reflective verbs etc.)
- with traditional categories, e.g. mood or tense (absent for segments)

Example:

Będę się bał jutro odezwać.

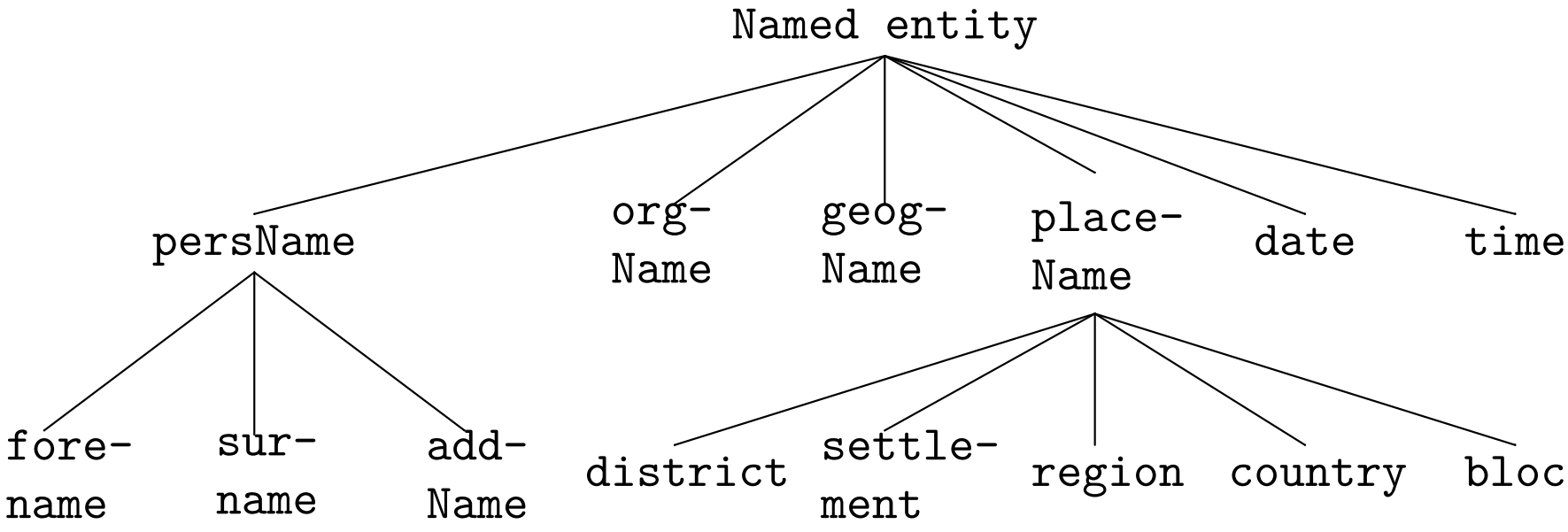
- *się bał*
- *Będę się bał* (nesting)
- *się odezwać.* (discontinuity, overlap)

Syntactic groups

Shallow description:

- typed groups: nominal, prepositional, ...
- may contain other syntactic groups and syntactic words
- marked syntactic and semantic heads
- no syntactic disambiguation
- no requirements of full parsing

Named entities



Named entities can:

- be nested (*Jan Kowalski*)
- be discontinuous (*Ocean wcale nie taki Spokojny*)
- overlap (*Ameryka Północna i Południowa*)

Word senses

First experiments with word sense disambiguation:

- 100 frequent and uncontroversially homonymous lexemes
- with grouped dictionary meanings (average 2–3 senses per word)

XML markup

```
<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:

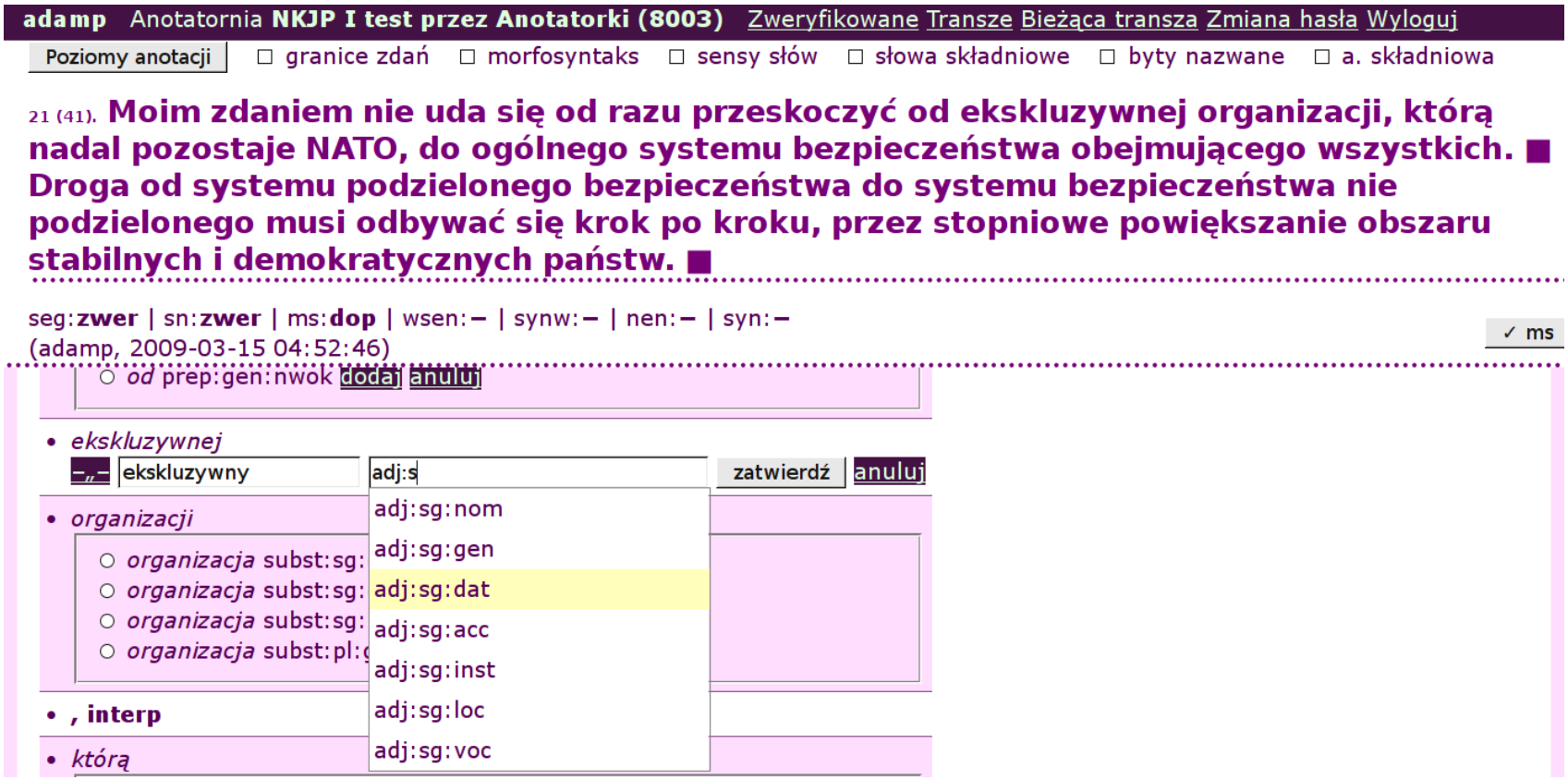
- 1 = [*word*; ORTH Bał się; BASE bać się; CTAG Verbfin; MSD sg:ter:m1:imperf:past:ind:aff:refl]
- 2 = [*word*; ORTH się odezwać; BASE odezwać się; CTAG Inf; MSD perf:aff:refl]

Annotation tools

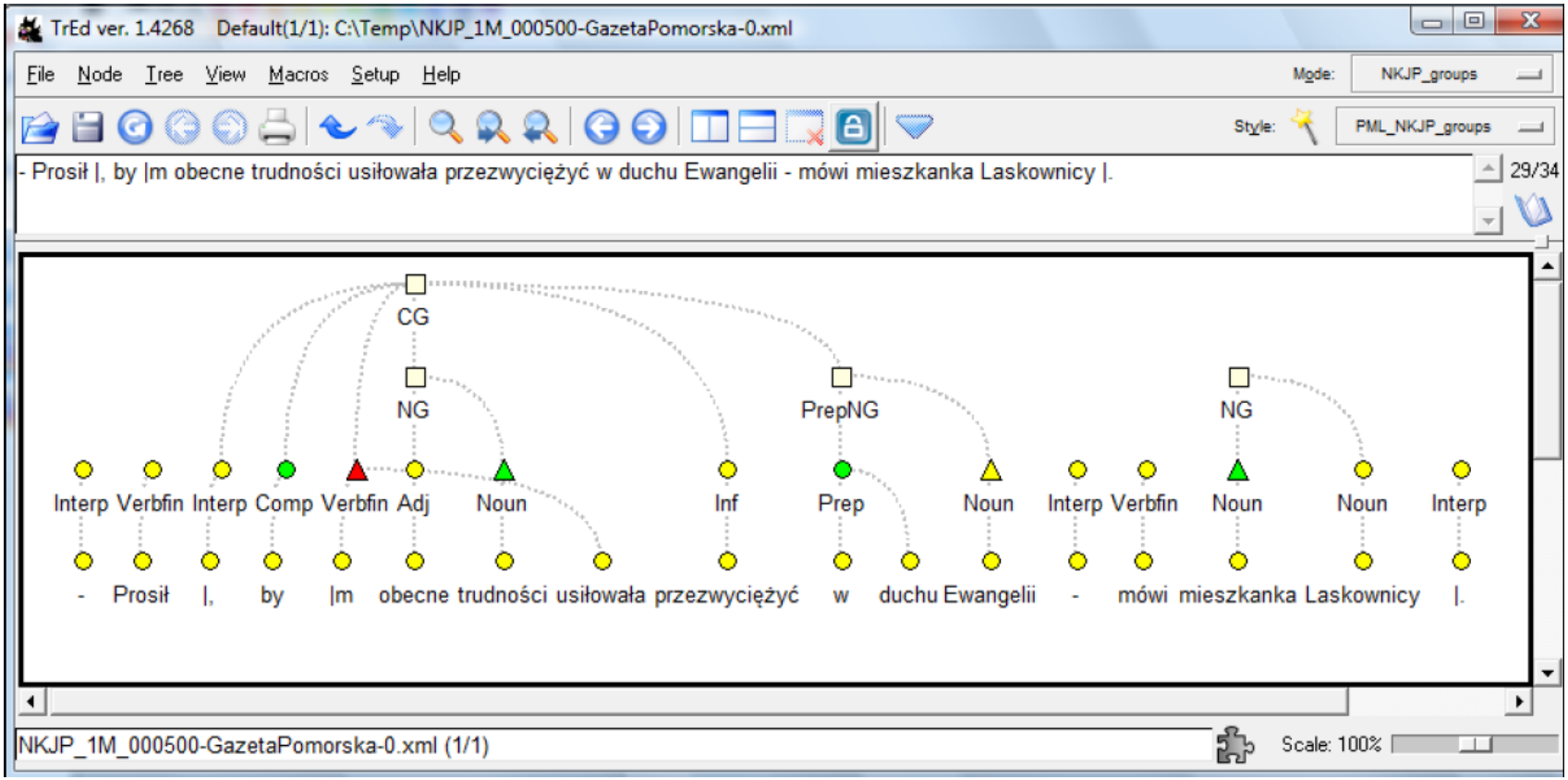
For manual annotation of NKJP1M:

- Anotatornia: segmentation, morphosyntax, word senses
- TrEd: syntactic words and groups, named entities

Anotatornia



TrEd


TrEd ver. 1.4268 Default(1/1): C:\Temp\NKJP_1M_000500-GazetaPomorska-0.xml
File Node Tree View Macros Setup Help
Mode: NKJP_groups
Style: PML_NKJP_groups
- Prosił |, by |m obecne trudności usiłowała przezwyciężyć w duchu Ewangelii - mówi mieszkanka Laskownicy |.
29/34
CG
NG
PrepNG
NG
Interp Verbfin Interp Comp Verbfin Adj Noun Inf Prep Noun Interp Verbfin Noun Noun Interp
- Prosił |, by |m obecne trudności usiłowała przezwyciężyć w duchu Ewangelii - mówi mieszkanka Laskownicy |.
NKJP_1M_000500-GazetaPomorska-0.xml (1/1)
Scale: 100%

Tools trained on NKJP1M

And used to automatically annotate full NCP:

- PANTERA disambiguating tagger
- NERF named entity recognizer
- WSDDE word sense disambiguating tool

NCP was a true achievement!

It found extremely diverse applications:

- it is still the main reference corpus in lexicography, applied linguistics, psycholinguistics and language modeling
- it has been used to boost the accuracy of natural language processing on various tasks
- it helped develop many tools and resources for Polish: disambiguating taggers, treebanks, coreference corpus, collocation databases, phraseological dictionary, valence dictionary
- is still the primary resource of linguistic research in Poland
- NCP search engines serve more than 1M distinct corpus user queries every year

But at the same time...

NCP is now truly outdated!

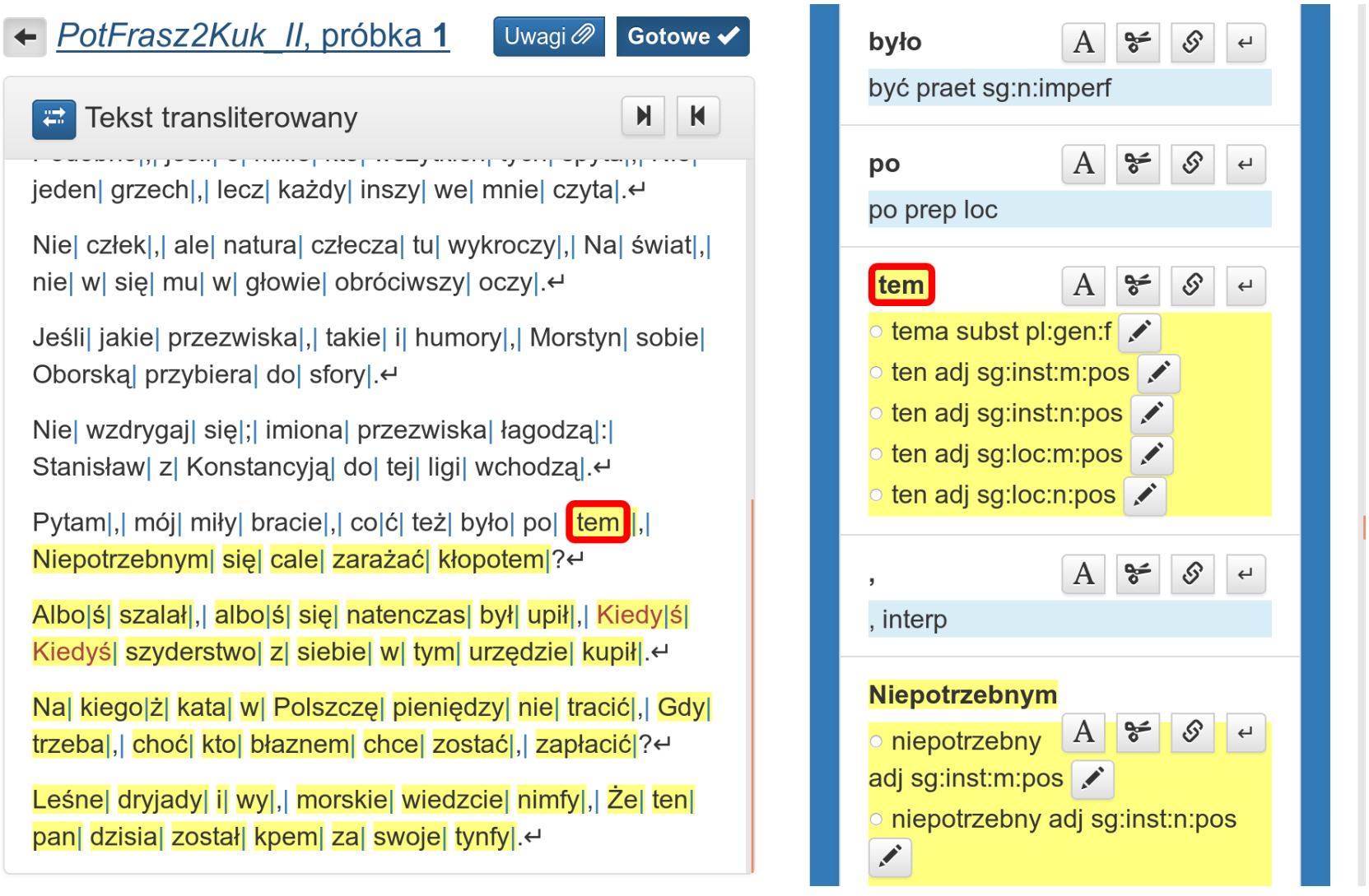
- it is a medium-sized corpus by modern standards!
- it does not cover modern lexical data or it occurs only in outdated contexts (*Emmanuel Macron*, *Donald Trump*, *Brexit*, *Instagram*, *fejk/fake*, *fanpage*, *selfie*)
- spoken data is low quality
- is TEI P5 really the optimal format?
- many nationally funded corpus projects creating data 'outside' NCP
- automatically annotated part is obsolete
- no funds for maintenance

Yet again...

Corpus researchers are really active in Poland:

- Chronofleks project provided a formal model of Polish inflection to represent historical changes, using new annotation environment (Anotatornia 2)
- Several corpora have been made available in the new MTAS-based corpus search engine
  - Electronic Corpus of 17th and 18th century Polish
  - Corpus of the 19th century Polish
  - NKJP1M
  - Polish Coreference Corpus
  - Polish Parliamentary Corpus

Anotatornia 2



Korpusomat

Corpus creation tool:

- a Web application automatically creating annotated and searchable corpora from documents provided by users
- technical knowledge-free data processing:
  - upload of user files or scraping data from a particular website
  - running automatic linguistic analysis
  - indexing and making the corpus available in MTAS
- what's new (vs. Poliqarp)?
  - new annotation layers (named entities), new toolset
  - querying across annotation layers
  - corpus statistics (frequency list, collocations, metadata-based graphs, term cloud)
  - corpus sharing (publicly or with specified users of the platform)

ZAPYTANIA DO KORPUSU 'LEM'

Zapytanie

[orth="A.*"][orth="M.*"] fullyalignedwith <ne="persName" />

Lp	Lewy kontekst	Rezultat	Prawy kontekst
1	skończyłem. Osobno zaprotokołowane zostało votum separatum doktora fizyki	**Archibalda** [Archibalda:subst:sg:gen:m1] **Messengera** [Messengera:subst:sg:gen:m1]	, który oświadczył, że to, co opowiedział Berton
2	Było to ostatnie już sito, ostatnia stacja odsiewu.	**Al** [Al:subst:sg:nom:m1] **Martin** [Martin:subst:sg:nom:m1]	, chłop, który wtedy, na Ziemi, wyglądał
3	, i w kolejnym barze wypiłem długiego drinka z	**Alfonsem** [Alfons:subst:sg:inst:m1] **Mauvinem** [Mauvinem:subst:sg:inst:m1]	z Agence France Press; po raz ostatni spróbowałem
4	dała rezultatu. Na koniec znany draper gwiazdowy, dzielny	**Ao** [Ao:subst:sg:nom:m1] **Murbras** [Murbras:subst:sg:nom:m1]	, wybrał się na Tairię samotrzeć z dwoma psami w
5	S. Totteles został, jak wiadomo, preceptorem młodego	**Aleksandra** [Aleksander:subst:sg:gen:m1] **Macedońskiego** [Macedoński:subst:sg:acc:m1]	. Winien był niedopatrzeń o koszmarnych skutkach, bo zawsze

STATYSTYKI KORPUSU 'LEM'

LISTA FREKWENCYJNA

	Lemat	Wystąpień
1	człowiek	7462
2	pan	6846
3	nic	5820
4	czas	4486
5	wszystko	3959
6	świat	3564
7	rzecz	3527
8	oko	2867
9	sposób	2742
10	ręka	2643

KOLOKACJE

Forma bazowa	Wystąpień	P-stwo (bias)	P-stwo
sprzężenie zwrotne	184	49.31	0.76
maszyna cyfrowa	114	9.32	0.21
gałka oczna	20	9.10	0.83
bańka mydlana	16	7.28	0.79
sieć neuronowa	80	7.16	0.22
równia pochyła	11	6.24	0.92
odruch warunkowy	30	6.15	0.40
nisza ekologiczna	21	5.80	0.51
instynkt samozachowawczy	23	5.80	0.47
ropa naftowa	14	5.78	0.70

SŁOWNICTWO CHARAKTERYSTYCZNE

Forma bazowa	C-value	Wystąpień
stan rzeczy	263.69	265
jedno słowo	193.00	194
druga strona	183.84	185
sprzężenie zwrotne	172.64	175
Stany Zjednoczone	128.67	130
dłuższa chwila	112.00	113
ułamek sekundy	106.56	108
maszyna cyfrowa	102.76	104
dzieło literackie	99.77	101
najwyższy stopień	90.63	92

Yet again...

Corpus researchers are really active in Poland:

- a number of historical corpora have been compiled
- several spoken corpora of Polish have been made available (enhanced NCP data made available in Spokes search engine, a large corpus documenting the dialect of Spisz with 2M words of transcripts, Corpus of Polish Teenage Talk)
- major parallel corpora were compiled (Polish-Russian, Polish-German, Polish-English, a Polish component of the International Comparable Corpus)

But still...

Polish is one of the few large European languages with an outdated national corpus!

The Polish Corpus Infrastructure (PIK)

Main goal:

- to create a unified platform for corpus-based studies of Polish
- covering present-day Polish starting from 1945
- with The National Corpus of Polish at its heart
- constantly updated
- of adequate quality
- federated with various existing corpora
- and covering a large genre-, channel- and register-balanced component
- and establish standards for the collection, processing and distribution of Polish corpus resources

The Polish Corpus Infrastructure

Implementation scope:

- implementing formats for representation of metadata, data and linguistic annotations
- extending the balanced segment of NCP with newest (post-2011) texts
- establishing a federation of Polish corpora
- providing tools for exploring and analysing the collection

The Polish Corpus Infrastructure

Access models:

- online access for end-users
- remote access for programmers
- full access to annotated subcorpora of samples
- full access to public domain resources
- full access to statistical and distributional models and other derivatives
- custom-made models

PCI: still many questions

And different ideas collected in the meantime:

- what is 'contemporary data'? since 1918? 1945? 1989?
- aren't we missing some data?
  - how about popular science texts, domain data, online data, spoken data...
  - electronic press or... blogs?
  - but maybe too much internet data is a curse?
  - too much legal data?
  - so many digital libraries out there!
  - monitoring internet data for corpora and lexicography
- should the corpus be balanced at all?
  - balanced wrt. time?
  - how about virtual corpora?
  - core corpus vs. literature for children, spoken, Internet, youth, historical, dialectal, multi-media, parallel (sub)corpora

PCI: still many questions

More ideas:

- licencing?
- is NCP really 'national'?
- NCP Lite:
  - ensuring the continuity of NCP: newest annotations, a "living" corpus
  - discussing the long-term development directions
  - grant funding mechanisms rewarding the transfer of the results of independent projects to NCP
- corpus as an institution
- commercial funding?

The National Diachronic Corpus of Polish

Aim of the project:

- to build an extensive, cross-sectional and linguistically enriched collection of Polish texts
- from (late) 14th to (the beginning of) 20th century
- using the existing resources and tools
- by federating existing corpora in a uniform technical implementation and a common additional layer of linguistic description

DCP: Disclaimers

Federation means that:

- existing resources can still exist and develop separately
- yet, from the user's point of view they can function as one coherent corpus of historical Polish

The National Diachronic Corpus of Polish

Integrated corpora:

- Corpus of Polish up to 1500 by the Institute of Polish Language, Polish Academy of Sciences (under construction)
- Corpus of 16th century Polish by the Institute for Literary Research, Polish Academy of Sciences
- Electronic Corpus of 17th and 18th century Polish texts (KORBA) by the Institute of Polish Language, Polish Academy of Sciences
- Corpus of 19th Century Polish (f19) by the University of Warsaw

The National Diachronic Corpus of Polish

Tasks:

- creating a common layer of linguistic description covering:
  - inflectional markers
  - principles of transliteration and transcription
  - metadata of each document
- two subcorpora:
  - manually-annotated with inflectional information
  - representative subcorpus
- providing technical compatibilty of all component corpora
  - (semi-)automatic transcription of the transliterated corpus
  - training of a disambiguating tagger for inflectional markup
  - making the corpus available in the federated search engine
- collection of a corpus of the years 1801–1918 (mostly from digital libraries)

To conclude

Getting back to our 3 main topics:

- NCP 2.0: The National Corpus of Polish (NKJP 2.0)
  $\rightarrow$ a large, representative synchronic corpus
- DCP: The Diachronic Corpus of Polish (NKDP)
  $\rightarrow$ an umbrella for diachronic data
- PCI: The Polish Corpus Infrastructure (PIK)
  $\rightarrow$ a vehicle to synchronize corpus initiatives in Poland

Thank you!

And all the corpus researchers in Poland and here!

- Let's promote infrastructural approach to national corpora
- Let's share research scenarios to react to what users need
- Let's exchange ideas among developers

References

- Andrzejczuk, A. (2010). *Narodowy Korpus Języka Polskiego — teoria i praktyka. Fakty, mity, potrzeby*. Legilingwistyka Porównawcza 3:133–141.
- Kieraś W., Kobyliński Ł., Ogrodniczuk M. (2018). *Korpusomat — a tool for creating searchable morphosyntactically tagged corpora*. Computational Methods in Science and Technology 24(1):21–27.
- Król, M., Gruszczyński W., Derwojedowa M., Górski R. L., Opaliński K., Potoniec P., Woliński M., Kieraś W., Eder M. (2019). *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*. Język Polski XCXIX(1):92–101.
- Ogrodniczuk M., Derwojedowa M., Łaziński M., Pęzik P. (2017). *Narodowy Korpus Języka Polskiego – co dalej?* Prace Filologiczne LXXI:237–245.
- Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. (2012, eds.). *Narodowy Korpus Języka Polskiego*. Warszawa, Wydawnictwo PWN.